TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

VİDEO VERİ SETLERİ İLE İNSAN EYLEMİ TANIMA YAKLAŞIMLARINA YÖNELİK ALAN ARAŞTIRMASI

YÜKSEK LİSANS TEZİ

Duygu Selin AK

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. Tansel ÖZYER

OCAK 2021

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntı yapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığını bildiririm.

Duygu Selin AK

ÖZET

Yüksek Lisans Tezi

VİDEO VERİ SETLERİ İLE İNSAN EYLEMİ TANIMA YAKLAŞIMLARINA YÖNELİK ALAN ARAŞTIRMASI

Duygu Selin AK

TOBB Ekonomi ve Teknoloji Üniveritesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Doç.Dr. Tansel ÖZYER

Tarih: Ocak 2021

İnsan eylem tanıma, insanların hareketlerinin makine öğrenmesi ve derin öğrenme metotları kullanılarak tahmin edilmesidir. Son yıllarda makine öğrenmesi ve derin öğrenme yöntemlerine artan ilgiyle birlikte, insan hareketlerinin tespiti konusu da gelişmektedir. İlk olarak duruk görüntüler üzerinden çıkarılan eylem tespitleri teknolojinin gelişmesiyle beraber videolar üzerinde ve hatta canlı akışlarda bile gerçekleştirilmeye başlanmıştır. Günlük hayatta da artık fazlasıyla görmeye başladığımız öğrenme tabanlı yöntemlerden biri de insan eylem tanıma yöntemleri olmaktadır. Hırsızların, kriminal suç işleyecek kişilerin veya tehlikeli aktiviteleri gerçekleştirecek bireylerin önceden tahmin edilmesi, yaya aktivitelerinin trafikteki öngörülerinde ve diğer birçok alanda insan eylem tanıma yöntemleri aktif olarak kullanılmaya başlanmıştır. Bununla birlikte, insan eylemlerini tanıma konusu öğrenme yöntemlerinin gelişmesiyle hem hız ve doğru tanıma performanslarının artmasıyla hem de pratik yöntemlerin gelişmesiyle kullanım alanları da genişlemektedir. Bu çalışmada, insan eylem tanıma konusundaki farklı yöntemleri ile dikkat çeken on beş farklı makale ele alınmış ve her biri detaylı olarak incelenerek bir araştırma hazırlanmıştır. Bu araştırmada yöntemleri bakımından insan eylem tanıma konusuna yeni bir bakış açısı kazandıran yaklaşımlar incelenecektir. Araştırma boyunca ele

alınan tüm makaleler videolar üzerinde tanıma işlemlerini gerçekleştirmektedir. Tüm bu tanıma işlemleri bir taksonomiye göre kategorilendirilmiştir ve beş ana kategori oluşturulmuştur. Bu kategoriler; ağ tabanlı yaklaşımlar, hareket tabanlı yaklaşımlar, çoklu örnek öğrenme tabanlı yaklaşımlar, sözlük tabanlı yaklaşımlar ve histogram tabanlı yaklaşımlardır. Bu kategorilendirmeye göre tüm makaleler incelenmiştir. İncelenen makalelerin her birine ait yöntemlerin açıklanması, geliştirme aşamaları, hangi ihtiyaçtan ortaya çıktığı, veri setleri üzerindeki çalışmalar ve elde edilen doğruluk sonuçları detaylandırılmıştır. Bununla birlikte, makalelerde kullanılan günlük hayat, trafikteki araçlar, spor ve uçangözlerden elde edilen videoları içeren veri setleri de ele alınmış, araştırmadan esinlenerek yeni araştırmacılara bir fikir kazandırmak amacıyla, her bir veri seti incelenerek bir karşılaştırma tablosu oluşturulmuştur. Veri setlerini kullanan makalelerin eğitim ve test ayrımlarını içeren tablolar da dâhil edilmiştir. Bu araştırma ile detayları verilen tüm makaleler ve veri setleri gelecekte yapılacak olan çalışmalara bir referans olacaktır.

Anahtar Kelimeler: İnsan eylem tanıma, Video analizi, Ağ tabanlı eylem tanıma, Hareket tabanlı eylem tanıma, Sözlük tabanlı eylem tanıma, Histogram tabanlı eylem tanıma, Çoklu örnek öğrenme tabanlı eylem tanıma.

ABSTRACT

Master of Science

HUMAN ACTION RECOGNITION APPROACHES

WITH VIDEO DATASETS – A SURVEY

Duygu Selin AK

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Assoc. Prof. Tansel ÖZYER

Date: January 2021

Human action recognition is the prediction of people's movements using machine learning and deep learning methods. With the increasing interest in machine learning and deep learning methods in recent years, the issue of detection of human movements has also been developing. With the development of technology, action detections, which were first extracted from static images, started to be performed on videos and even in live streams. Human action recognition methods are one of the learning-based methods we have started to see in daily life. Prediction of thieves, criminals or individuals who will carry out dangerous activities has started to be actively used in traffic predictions of pedestrian activities and in many other areas. Furthermore, with the development of learning methods on the subject of human action recognition, the areas of use are expanding with the increase of speed and correct recognition performances and the development of practical methods. In this study, fifteen different articles drawing attention with their different methods on Human action recognition are discussed and a research is prepared by examining each one in detail. In this research, approaches that give a new perspective to human action recognition in terms of methods will be examined. All the articles discussed throughout the research carry out recognition on the videos. All these recognition processes are categorized by a

taxonomy and five main categories are created. These categories are; network-based approaches, motion-based approaches, multiple instance learning based approaches, dictionary-based approaches and histogram-based approaches. All articles were examined according to this categorization. Explanation of the methods for each of the articles examined, the stages of development, the need arising from it, the studies on the data sets and the accuracy results obtained were detailed. In addition, datasets containing videos from daily life, vehicles in traffic, sports and drone videos were also considered and a comparison table was created by examining each dataset in order to gain an idea for new researchers. Tables containing training and test separations of articles using data sets are also included. This research will be a reference to future studies of all articles and data sets detailed.

Keywords: Human action recognition, Video analysis, Network based action recognition, Motion based action recognition, Dictionary based action recognition, Histogram based action recognition, Multiple Instance Learning (MIL) based action recognition.

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla beni yönlendiren hocam Doç. Dr. Tansel ÖZYER'e, kıymetli tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü öğretim üyelerine ve destekleriyle her zaman yanımda olan aileme ve arkadaşlarıma çok teşekkür ederim.

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

RESİM LİSTESİ

Sayfa

KISALTMALAR

FPS : Saniye başına düşen çerçeve
MIL : Çoklu Örnek Öğrenme (Multiple Instance Learning)
CUHK : Hong Kong Çin Üniversitesi (The Chinese University of Hong Kong)
CNN : Evrişimli sinir ağı
ReLU : Doğrultulmuş Doğrusal Birim (Rectified Linear Unit)
LSTM : Uzun-Kısa Vadeli Hafıza (Long-Short Term Memory)
SSIM : Yapısal benzerlik endeksi
GAN : Çekişmeli üretici ağlar
MSE : Ortalama hata karesi
PSNR : En yüksek sinyal gürültü oranı
GPU : Grafik İşleme Birimi (Graphics Processing Unit)
HMDB : İnsan Metabolom Veri Tabanı (Human Metabolom Database)
CAD : Cornell Eylem Veri Seti (Cornell Action Dataset)
DML : Ayrık Çok Görevli Öğrenme
RNN : Tekrarlamalı sinir ağları
RGB : Kırmızı yeşil mavi
VGG : Görsel Geometri Grubu (Visual Geometry Group)
RGB-D : Kırmızı yeşil mavi - derinlik
HoG : Histogram of Oriented Gradients
HoF : Histogram of Optical Flow
PCA : Temel Bileşenler Analizi (Pricipal Component Analysis)
NMF : Non-Negative Matrix Factorization
STIP : Spatio Temporal Interest Point
tIoU : Time Intersection Over Union
UCLA : Kaliforniya Üniversitesi (University of California, Los Angeles)
MVRM : Multirate Visual Recurrent Model
KTH : Kraliyet Teknoloji Enstitüsü
1-NN : En yakın birinci komşu
UCF : Florida Merkez Üniversitesi
DPC : Dense Predictive Coding
C3D : Üç boyutlu evrişimli ağlar
XDC : Cross-Modal Deep Clustering
GRNN : Geçitli tekrarlamalı sinir ağı
SGN : Semantics-guided Neural Network
NTU : Nottingham Trent Üniversitesi
GWR : Grow When Required
ASR : Otomatik konuşma tanıma
VQ : Yöney Nicemleme (Vector Quantization)
BERT : Bidirectional Encoding Representation from Transformers
UCF-ARG : University of Central Florida-Aerial camera, Rooftop camera and Ground camera
UTD-MHAD : University of Texas at Dallas – Multimodal Human Action Dataset

AdaScan : Adaptive Scan Pooling
FASTER : Feature Aggregation for Spatio Temporal Redundancy
KL : Kullback Leibler
JS : Jensen-Shannon
WCGAN-GP : Conditional Wasserstein General Adversarial Network
AVA : Atomic Visual Action
mAP : Ortalama kesinlik
SVM : Destek vektör makinesi
IDT+FV : Short-Term Motion Along With Appearance Because Of Presence Of The HoG

1. GİRİŞ

Son on yılda birçok araştırmaya katılan insan eylemlerinin tanınması, günlük yaşamın farklı alanlarındaki sorunların çözülmesine yardımcı olmaktadır. Günümüzde sağlık, güvenlik, robotik veya oyun alanlarında insan eylemi tanıma algoritmalarını giderek daha fazla kullanıyoruz. Ancak, tüm bu uygulamaları geliştirmek için dikkat edilmesi gereken nokta, tanıma konusunu anlamaktır. Tanıma, farklı sensörlerin yardımıyla alınan görüntülerdeki eylemlere veya etkinliklere bir etiket verme işlemidir. Ayrıca, tanıma işlemleri farklı hareket seviyelerinde gerçekleştirilir. Bunlar eylem ilkelleri, eylemler ve faaliyetlerdir. En temel seviye eylem ilkeleridir ve bunlar atom düzeyinde değerlendirilen hareketlerdir. Eylemler eylem ilkelinden oluşur ve genellikle tüm vücut hareketlerini temsil eder. En üst düzeydeki faaliyetler bir dizi eylem serisinden oluşur ve gerçekleştirilen tüm eylemleri temsil eder.

Örneğin; basketbol oynamak bir etkinlikken, "oynamak" bir eylemdir ve "sol kol yukarı" bir eylem ilkesidir. Tanıma işlemleri farklı öğrenme yöntemleri kullanılarak gerçekleştirilir. Aynı şekilde, tüm bu yöntemlerin ortak noktası genellikle özellik çıkarma, eylem öğrenme ve sınıflandırma, eylem tanıma ve kesimleme içermesidir.

Araştırmamız kapsamında son dönemde önem kazanan ve popülerleşen faaliyet ve eylemlerin sınıflandırılması ve tahmin edilmesi konusunda farklı çalışmalar ele alınmıştır. Her birinde farklı yöntemler incelenerek, eylem tanımaya yönelik on beş farklı yaklaşımın özellikleri, benzerlikleri ve farklılıkları beş ana başlık altında tartışılmıştır. Tüm başlıklar, eylem tanıma yöntemlerine göre ayrıştırılmıştır. Elealınan eylem tanıma yaklaşımları, ürettikleri çıktılara göre de ayrışmaktadır. Burada üç sınıf altında çıktı üretilmektedir: Verilen girdiden bir sonraki çerçevenin tahmini [1,13,18,16], verilen girdiden önceki ve sonraki çerçevenin tahmini [14] ve yalnızca verilen girdinin eyleminin tahmini (diğer iki kategoriye girmeyen tüm ele alınan yaklaşımlar). Eylem tanıma yöntemlerinin yanı sıra, sınıflandırma ve tanıma yöntemlerinde sıklıkla seçilen bu kapsamda uygun video ve görüntü veri setleri incelenmiştir. Seçilen yirmi beş farklı veri seti için detaylı çalışmalar yapılmıştır. Veri

setlerinin özellikleri, video içerikleri, sayıları, kullanım şekilleri ve çözünürlükleri gibi birçok ayrıntı araştırmamızda ele alınmaktadır.

Veri setleri ve yöntemler birbirleri ile karşılaştırılmış, taksonomi bilgileri üretilmiş, yöntemlerin doğruluk oranları incelenmiş ve hangi yöntemin hangi koşullarda çalıştığı açıklanmıştır. Son olarak, yöntemleri kullanırken karşılaşılan zorluklardan bahsedilmiştir.

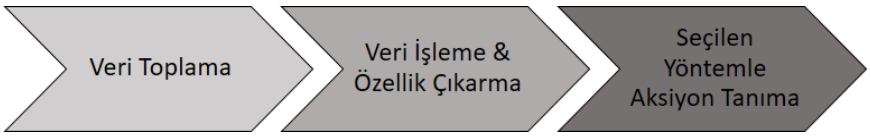
Bu araştırmada incelenen yöntemler ve veri setleri ile eylem tanıma üzerine çalışmak isteyen araştırmacıların sıklıkla tercih ettikleri veri setleri hakkında yeterli bilgiye tek kaynaktan erişmeleri ve araştırmalarına uygun veri setlerini seçmeleri için destek sağlanması amaçlanmaktadır. Ayrıca araştırmacıların tartışılan farklı öğrenme temelli yöntemler ile farklı açılardan ele almaları ve öğrenme yöntemleri arasında karşılaştırmalar yaparak araştırmaları için doğru yöntemi seçmelerinde yol gösterici olmaları amaçlanmaktadır.

Tezin ana hatları şu şekildedir: İlk bölüm tez ile ilgili giriş bölümüdür ve genel çerçevede tez çalışması ile ilgili bilgiler verilmiştir. İkinci bölümde tezin anlaşılması için gerekli görülen bazı temel terimler yer almaktadır. Üçüncü bölümde tez boyunca ele alınacak olan on beş farklı makalenin kullandığı veri setleri açıklanmıştır. Dördüncü bölümde gerçekleştirilen çalışmaların yöntemleri ele alınmıştır. Beşinci bölümde ele alınan makaleler ile ilgili nicel analizlerin sonuçları paylaşılmıştır. Altıncı bölümde ele alınan makalelerin karşılaştırmalı değerlendirme sonuçları yer almaktadır. Son olarak gelecekte yapılabilecek çalışmalar ve sonuçların paylaşılması ile tez sona ermektedir.

2. TEKNİK ALTYAPI BİLGİLERİ

2.1 İnsan Eylem Tanıma

İnsan eyleminin tanınması, kişinin veya nesnenin belirli hareketlerini inceleyerek ana sensör girdilerinden bir eylemin tanınma işlemidir. Video tabanlı insan eylem tanıma sistemleri; kameralar, uçangözler veya radarlar gibi girdilerden elde edilen videolar üzerinde çalışmalar gerçekleştiren tanıma sistemleridir. Bu girdileri almak ve işlemek oldukça zor ve karmaşıktır. Bu bağlamda, farklı yöntemlerle çeşitli yaklaşımlar üretilebilir. Genel anlamda yöntemler, veri toplama, ön işleme ve özellik çıkarma ve tercih edilen öğrenme yöntemlerine uygun olarak etkinliklerin tanınması olmak üzere Şekil 1.1'deki 3 ana süreçten oluşur.



Şekil 1.1 : İnsan eylem tanıma işleminin ana süreçleri.

2.2 Gözetimli ve Gözetimsiz Öğrenme

Seçilen veriler, insan eylemi tanıma süreçlerinde kullanılacak öğrenme yönteminin belirlenmesinde de önemlidir. Gözetimsiz öğrenme, etiketlenmemiş veriler aracılığıyla bilinmeyen bir yapıyı tahmin etmek için kullanılan bir makine öğrenmesi tekniğidir. Bu öğrenme yönteminde, girdi verilerinin hangi sınıfa ait olduğu belirsizdir. Elde edilen etiketlenmemiş verileri kümeleme yöntemi ile gruplayarak öğrenme sürecini gerçekleştirir. Bu öğrenme biçiminde, bu grupların hangi sınıfa ait oldukları bilinmemektedir, ancak belirli bir örneğin hangi gruba ait olduğu bilgisi elde edilir. Gözetimli öğrenme ise etiketli eğitim verileri ile istenen sonuçlar arasında eşleşen bir

öğrenme yöntemidir. Başka bir deyişle, giriş değerinin hangi çıktıda üretildiği bilinen veriler eğitim verisi olarak kullanılır ve yeni verilen girdinin eşleşme sonucunu tahmin eder. Eğitim verileri hem girdilerden hem de çıktılardan oluşur. Fonksiyon, sınıflandırma veya regresyon algoritmaları ile belirlenebilir.

Gözetimli öğrenmenin en zor kısmı eğitim verileri oluşturmaktır. Eğitim verileri ve makine öğrenmesi yöntemleri kullanılarak bir işlev oluşturulur. Bu işlev, gelen yeni verileri tahmin etmeye çalışır. Bu nedenle, gözetimli öğrenmede verilerin hazırlanması hem zaman alır hem de verilerin çok dikkatli bir şekilde hazırlanması son derece önemlidir. Yanlış etiketlenmiş veriler çıktı olarak yanlış sonuçlar üretebilir.

Gözetimsiz öğrenmede etiketli eğitim verisi yoktur. Gözetimsiz öğrenmenin içerdiği birçok algoritmayı kullanarak kümeleri öğrenmek mümkündür. Bu bölümdeki algoritmalar verileri gruplamaya ve yeni verileri en uygun gruba atamaya çalışır. Eğitim verisi olmadığı için uygulaması kolaydır. Ancak zor problemlerde çok iyi sonuçlar vermeyebilir.

İnsan eylemini tanıma konusunda gözetimli öğrenme ile ilgili birçok çalışma bulunmaktadır. Bu çalışmalarda kullanılan yöntemlerden farklı olarak, tanımaya yönelik gözetimsiz öğrenme yöntemlerinin kullanılması son zamanlarda gündemdedir. Bu bağlamda, insan eylemini tanımada gözetimsiz öğrenmenin kullanımına ilişkin dört farklı makale incelemesi yapılmıştır.

[88] 'de kodlama işlemi MVRM ile bir video klibin farklı aralıkları alınarak gerçekleştirilir. Bu yöntem, video karelerinde geçmiş ve şimdiki zaman ile şimdi ve gelecek arasındaki zamansal bağlamı kullanarak videolardaki hareket hızı farkıyla başa çıkmayı kolaylaştırır. Bu yöntemde iki farklı görev ele alınır. Birincisi karmaşık olay algılama, diğeri ise video altyazı oluşturmadır. Her iki görev için de benzer yöntemlerle karşılaştırılmıştır.

[89] 'da Kinect sensörü yardımıyla toplanan etiketsiz iskelet verileri kullanılarak bir eylem tanıma yöntemi geliştirilmiştir. Bu yöntemde, Kinect yardımı ile işaretlenen vücut anahtar noktaları toplanır ve benzer hareketleri içeren iskelet verileri gözetimsiz öğrenme yöntemleri kullanılarak gruplandırılır. Geliştirilen bu sistem Tahmin Et &

Kümele olarak adlandırılır. Üç farklı veri seti kullanılarak elde edilen doğruluk değerleri arasında %84,9'luk başarı elde edilmiştir.

İncelenen başka bir makalede, kelime torbaları yönteminin görsel verileri içeren insan eylemini tanımaya uyarlanmasıyla geliştirilen gözetimsiz öğrenme yöntemi tartışılmıştır [90]. Bu makalede, her eylemi içeren durum için özellik çıkarma işlemi gerçekleştirilmiştir. Yerel özellikler kaldırıldıktan sonra tüm görsel kelimeleri içeren bir sözlük oluşturulmuştur. Bu sözlüğü kullanarak bir sınıflandırıcı veya olasılıksal model geliştirilmiş ve sonuçlar elde edilmiştir. KTH veri seti üzerinde çalışan model sonucunda 1-NN kullanılarak %83,3 tanıma oranı elde edilmiştir.

[91] ’de ise, LSTM kodlayıcı - kod çözücü modeli kullanılarak gözetimsiz bir öğrenme yöntemi geliştirilmiştir. Görüntü pikselleri yamaları ve önceden eğitilmiş evrişimli ağ kullanılarak çıkarılan video karelerinin yüksek seviyeli temsilleri ("algılar") girdi olarak kullanılmıştır. UCF-101 ve HMDB-51 veri setlerinin kullanıldığı bu çalışmada, UCF-101 veri setinde %84,3 doğruluk elde edilmiştir.

Gözetimli öğrenme kapsamında, insan eylemini tanıma konusu birçok makalede tartışılmıştır. Araştırma boyunca incelenen makalelerde ayrıntılı bilgi bulunabilir.

Gözetimli ve gözetimsiz yöntemlere ek olarak, son zamanlarda kendi kendini gözeten yöntemlerin önem kazandığını görüyoruz. Kendi kendini gözeten öğrenme yöntemi, gözetimsiz öğrenmenin bir alt kümesi olmasına rağmen, verilen etiketlenmemiş verilerin kendi kendini etiketlemesini gerçekleştirir. Kendi kendini gözeten farklı yöntemlerden seçilen birkaç makalenin incelemesi aşağıdaki paragraflarda yer almaktadır.

[92]'de videolardan elde edilen uzamsal-zamansal yerleştirmeleri kullanarak kendi kendini gözeten bir öğrenme yönteminin geliştirilmesi amaçlanmıştır. Bu yöntemde üç aşama takip edilmiştir. İlk olarak, öğrenme süreci DPC çerçevesi ile videolar üzerinde gerçekleştirilmiştir. Daha sonra, öğretim programı eğitim şeması, zamansal bağımlılığı azaltarak gelecekteki tahminde daha fazla tahmin yapmak için önerilmiştir. Son olarak, Kinetics-400 veri kümesini kullanılarak, ilk adımda bahsedilen DPC çerçevesi ile eğitildiği gibi, ardından işlem tanımada kullanılmak üzere ince ayarlar

gerçekleştirilmiştir. Sonuç olarak, %75,7 ile çalışmanın en yüksek doğruluk oranı elde edilmiştir.

Ele alınan bir diğer yöntemde, videolardan karıştırılan kliplerin kendi kendini gözeten öğrenme yöntemi ile zamansal ve uzamsal temsilini kronolojik sıralama (kliplerin sırasını belirleyerek) gerçekleştirmeyi amaçlamaktadır [93]. Yöntemde, özellik çıkarımı için üç boyutlu evrişimli sinir ağı kullanılır, çıkarılan özellikler video kliplerin sırasını tahmin etmek için kullanılır. Temsiller bir sinir ağı ile hesaplanır. Ayrıca, eylem tanıma modelini ayarlamak için en yakın komşu yöntemi kullanılır. Sonuç olarak, geliştirilen yöntem klip sırası tahmininde %64'ün üzerinde doğruluk elde etmiştir.

[94]'te video gösterimleri için uzamsal-zamansal özellikleri öğrenmek için kendi kendini gözeten öğrenme yöntemi kullanılmaktadır. Görsel özellikleri öğrenmek için hem uzamsal hem de zamansal bağlamlarda hareket ve görünüm istatistikleri ile birlikte öğrenme süreci gerçekleştirilir. Ayrıca hem zamansal hem de uzamsal bağlamlarda elde edilen desenlerden renk, yön, renk çeşitliliği ve hızlı hareket bölgesi gibi istatistiksel bilgiler elde edilir. Geliştirilen bu yöntem C3D ile doğrulanmış ve makalede kıyaslanan diğer yöntemlere göre en yüksek doğruluk elde edilmiştir.

[95]'te, bir videodaki ses ve görüntü arasındaki ilişkiye dayanarak, diğerinin (örneğin ses) anlamını birinin (ör. video) varlığından tahmin etmek için kendi kendini gözeten bir yöntem geliştirilmiştir. XDC adı verilen kendi kendini gözeten öğrenme yöntemi, sesi kümelemek ve diğerini de anlamlandırmak için kullanılır. Bu yöntemin son derece ileri teknoloji bir yöntem olduğu belirtilmekle birlikte birden fazla karşılaştırma yapılan makalede en yüksek doğruluk değeri %95,5 olarak belirtilmiştir.

Kendi kendini gözeten bir eylem sınıflandırma yöntemi olan VideoBERT, insan dilindeki kelimeleri kullanarak videolardaki üst düzey nesneleri ve olayları tanımaya odaklanır [130]. Bu bağlamda, metni konuşmadan dönüştürmek için otomatik konuşma tanıma ile, önceden eğitilmiş video sınıflandırma modellerinden düşük seviyeli uzamsal-zamansal görsel özellikler elde etmek için VQ ve BERT [128] ayrık belirteç dizileri üzerinde ortak dağılımları öğrenmek için model kullanılır. Başka bir deyişle, ortaya çıkan görsel kelimelerin konuşulan kelime çiftlerini öğrenmek için

BERT kullanılmıştır. YouCookII [129] veri kümesinde ilk tahmin edilen sonucun doğruluğu %3,2 ve ilk tahmin edilen nesne doğruluğu %13,1'dir.

Başka bir kendi kendini gözeten video-metin temsil yöntemi olan ActBERT, eşleştirilmiş video dizileri ve metin açıklamalarından küresel ve yerel görsel ipuçlarını ortaya çıkarmayı amaçlamaktadır [131]. Hem küresel hem de yerel görsel sinyaller anlamsal akışla etkileşim halindedir. Bu model, derin bağlam bilgisinden yararlanır ve video-metin birleştirme modellemesi için ilişkileri çıkarır. Video yakalama sonuçları incelendiğinde, YouCookII [129] veri seti kullanıldığında, bazı standart değerlendirme ölçümlerinde VideoBERT [130] 'den daha iyi sonuçlar elde ettiği görülmüştür.

Kendi kendini denetleyen video metin gösterme yöntemlerinde, videolarla eşleşen ek açıklamaları el ile ekleyerek veri kümesini oluşturmak zaman alacaktır ve zor bir süreçtir. Bu zorluğu gidermek için oluşturulan HowTo100M [132] veri kümesi, bilgilendirici içeriğe sahip milyonlarca video içerir. Ayrıca, ek açıklamaları el ile eklemeyi gerektirmez. HowTo100M veri kümesi kullanılarak geliştirilen video-metin yöntemi de sıklıkla tercih edilen YouCookII [129] verileriyle eğitilmiştir. Video erişimi, YouCookII [129] veri kümesinde %4,2 sonuç elde ederken, HowTo100M veri kümesinde daha iyi performans göstermiştir ve %6,1 sonuç elde edilmiştir.

2.3 Sinir Ağları

Sinir ağı tabanlı yöntemler genellikle insan eylemini tanıma süreçlerinin gerçekleştirilmesini kolaylaştırmak ve performanslarını arttırmak için seçilir. Sinir ağı, insan beynine benzer bir çalışma şekline sahip katmanlardan oluşan bir mimaridir. Çok sayıda parametre ve değişken kullanma kabiliyeti, esnekliği, hata toleransı ve doğrusal olmayan problemlere yeteneği sayesinde karmaşık problemlerde sıklıkla tercih edilir.

Son derece büyük video verileriyle yapılan insan eylemi tanıma yöntemlerinde özellik çıkarma, en az öğrenme süreci kadar zaman ve kaynak kullanır. Bu nedenle, sinir ağı yöntemleri, seçilen yöntemleri optimize etmek için sıklıkla kullanılmaktadır. Sinir ağı yöntemleri, insan eylemini tanıma kapsamında yüksek başarı elde edebilen, karmaşık sorunlara uyum sağlayan ve sıklıkla tercih edilen yöntemler haline gelmektedir. Bu

araştırmada, evrişimli sinir ağı ve tekrarlamalı sinir ağı olarak başlıca yöntemlerden bahsedilmektedir.

GRNN yöntemi kullanılarak geliştirilen hibrit modelde insan eylemi tanıma için yeni bir model önerilmiştir [96]. Bu modelde, zor bir problem olan öznitelik çıkarımı seçilen yöntemle gerçekleştirilir ve insan eylemleri tahmin edilir. Standart RNN'e ek olarak, RNN'deki değişkenlerin sayısını ve gizli birimlerdeki parametreleri azaltmak için her RNN düğümüne geçitli tekrarlamalı ünitenin bir bellek hücresi eklenmiştir. Yeni sinir ağı yaklaşımı ile yapılan video sınıflandırmasında KTH veri setinde %93 doğruluk elde edilmiştir.

Optik akış ve evrişimli sinir ağı kullanılarak geliştirilen makalede ise, insan hareketlerinin zamansal bilgileriyle uyumlu uzamsal bilgi optik akışla elde edilir ve bu hareketlerin uzamsal bilgileri hem zamansal hem de eş zamanlı olarak ayrıştırmak için iki akışlı bir CNN yapısına girdi verilir [97]. Ardından senaryo oluşturmak için CNN'nin özellik çıkarma yöntemi kullanılır. Çıkarılan özellikler, transfer öğrenmesi kullanılarak anlamlı hale getirilir. Elde edilen sonuçlarda, iki akışlı modelin ortalama sınıflandırma doğruluk değeri %88,31 olarak verilmiştir.

Bir sinir ağına dayalı olarak geliştirilen başka bir çalışmada, SGN kullanılarak iskelet tabanlı bir eylem tanıma yöntemi geliştirilmiştir [112]. Geliştirilen bu modelde, iki anlamsal kavram içeren ortak düzey ve çerçeve düzeyinde modül bulunmaktadır. Böylelikle SGN ile zamansal ve uzamsal korelasyonların ortak ve çerçeve düzeyinde çıkarılması işlemi gerçekleştirilmiştir. Modelin NTU60 veri setinde %90,6 doğruluk elde edilmiştir.

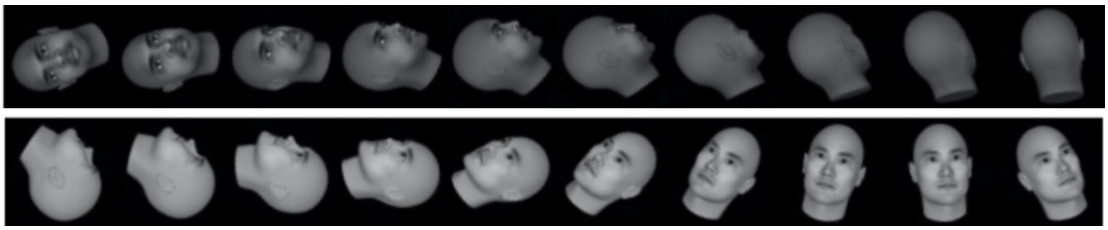
Son olarak, insan eylemlerini derinlik bilgisi ve RGB görüntülerle öğrenmek ve sınıflandırmak için büyüyen kendi kendini gözeten ağların çeşitleri kullanılmaktadır [113]. Bu makalede, GWR modeli gibi girdi dağılımına göre kendi kendine büyüyen ve küçülen ağ modelleri arasında farklı çalışmalarla karşılaştırmalar yapılmıştır.

3. MAKALELERDE KULLANILAN VERİ SETLERİ

Çalışma kapsamında seçilen makalelerin kullandığı veri setleri ve genellikle insan eylem tanıma konusunda tercih edilen verisetleri bu başlık altında incelenmiştir. Ele alınan tüm verisetlerinin detaylı karşılaştırılmasının yer aldığı tablo EK 1'de yer almaktadır.

3.1 FaceGen Veri Seti

FaceGen, [1]'deki makalede kullanılan veri setidir ve makalenin yazarları tarafından türetilmiştir. Veri seti FaceGen adı verilen bir yazılım paketi kullanılarak üretilmiştir. Bu yazılım paketi 1998 yılında Singular Inc. tarafından geliştirilmiştir [2]. Yazılım farklı açılarda üç boyutlu (3D) yüz şekilleri oluşturur. Yazılım tarafından yüz şekilleri türetildiğinden istenilen sayıda üretilebilir ve veri setindeki şekiller ve veri sayısı değişiklik gösterebilir. Makalede kullanılan veri setinde ise 25 farklı yüz Z ekseninde 7 ve X ekseninde 8 farklı açıyla elde edilmiştir. Bu sebeple; veri seti her zaman sabit olmamaktadır. Buna karşın, temel noktamız [1] olduğundan veri setinin içindeki yüz şekillerinin sayısını bu makaleden baz almaktayız. Ele alınan makalede, 4000 resim içeren bir seri üzerinde regresyon eğitilmiş, 500'ü doğrulama ve 1000'i de test için kullanılmıştır. Resim 3.1'de örnek veriler yer almaktadır.

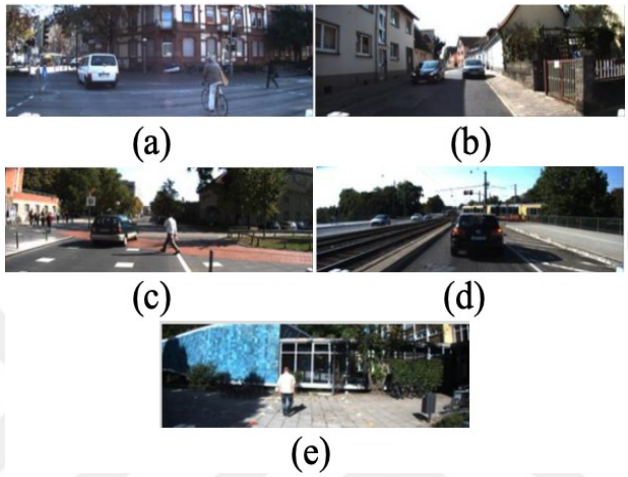


Resim 3.1: FaceGen veri setinde ele alınan bazı örnek yüz şekilleri [1].

3.2 KITTI Veri Seti

KITTI veri seti Karlsruhe Teknoloji Enstitüsü (KIT) ve Chicago'daki Toyota Teknolojik Enstitüsü (TTI-C) tarafınfan bir proje için üretilmiştir [3]. Bu veri seti bir araç içerisine yerleştirilmiş kameradan elde edilen görüntüleri içermektedir. Kameradan elde edilen bu görüntüler 6 farklı kategoriye ayrılmıştır. Bunlar: Şehir, Yerleşim Bölgesi, Yol, Kampüs, İnsan ve Kalibrasyon'dur. Kalibrasyon çalışmaya

dahil edilmeyen bazı test sahnelerini barındırmaktadır. Tüm videoların çözünürlükleri 1392x512 piksel ve yaklaşık olarak 10 FPS'dir. Resim 3.2'de kategorilere ait ekran görüntüleri bulunmaktadır. Veri seti toplamda 156 adet video içerir, kategorilere düşen video sayısı ise; Şehir kategorisi için 28, Yerleşim Bölgesi kategorisi için 21, Yol kategorisi için 12, Kampüs kategorisi için 10, İnsan kategorisi için 80 ve Kalibrasyon kategorisi için 5 videodur. Bazı videolar birden fazla insan ve aracı sahnelemektedir.



Resim 3.2: KITTI veri seti içerisindeki videolardan alınmış bazı ekran görüntüleri. Kategoriler; (a) şehir, (b) yerleşim bölgesi, (c) kampüs, (d) yol, (e) insan şeklindedir [3].

3.3 CalTech Yaya (Pedestrian) Veri Seti

CalTech yaya veri seti, KITTI veri setine benzer olarak Los Angeles'da araç içi kameradan elde edilen videoları içeren bir veri setidir [4]. Veri setinin geliştirilme amacı yol üzerinde yer alan insanların tespit etmektir. Toplamda 250.000 kare içeren yaklaşık 10 saate denk gelen video içermektedir ve videolarda 2300'e yakın farklı insan barındırmakla birlikte aynı videodan birden fazla insan da görmek mümkündür. Videoların çözünürlükleri 640x480'dir. Resim 3.3'de örnek kareler bulunmaktadır.

Resim 3.3: CalTech Yaya Veri Seti'nden alınan bazı örnekler [4].

3.4 TV İnsan Eylem Etkileşimi (TV Human Action Interaction) Veri Seti

TV İnsan Eylem Etkileşimi veri seti Oxford Üniversitesi'nin Görsel Geometri Grubu tarafından geliştirilmiştir [5]. 20 farkı TV programından elde edilen 300 video barındırır. Videolar 4 farklı kategoriye ayrılmıştır; el sıkışmak, sarılmak, çak beşlik ve öpmek. Eylem kategorilerine ait örnekler Resim 3.4'te yer almaktadır. Buna karşın veri setindeki bazı videolar herhangi bir kategoriye ait değildir. Ayrıca; videolardaki her kare için bir ek açıklama (annotation) bilgisi bulunmaktadır.

Resim 3.4: TV İnsan Eylem Etkileşimi veri setindeki kategorilerden bazı örnekler [5].

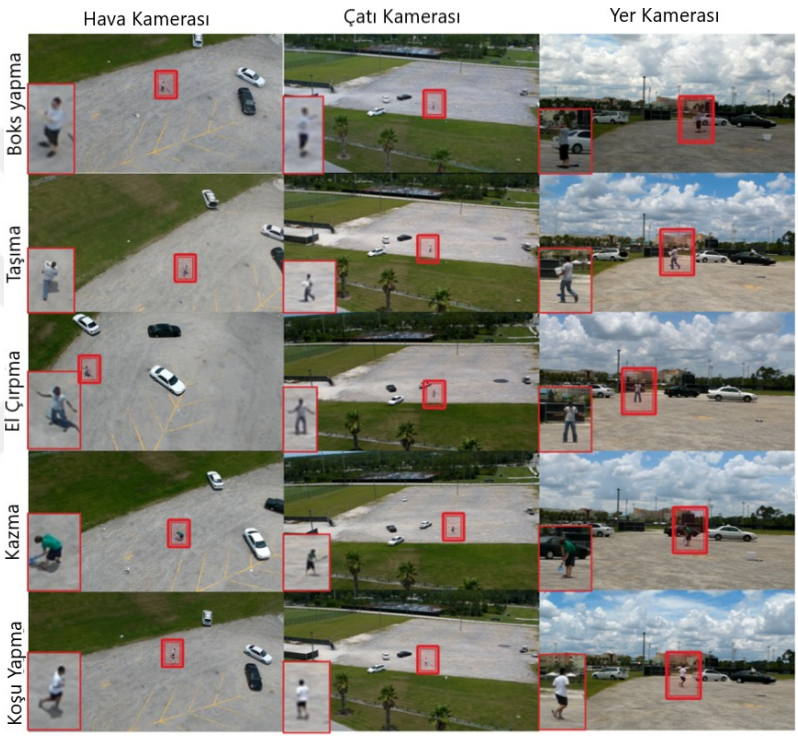
3.5 KTH Veri Seti

KTH veri seti, insan eylem tanıma işlemleri için oluşturulmuş, 6 farklı kategori içeren bir veri setidir [6]. Kategoriler: Yürüme, Koşu yapmak, Koşma, Boks yapma, El sallamak ve Alkışlamak'tır. Tüm kategorilerin 25 farklı kişi ile 4 farklı arka planda çekilmiş videoları bulunmaktadır. Veri seti 2391 seri içerir ve videolar homojen arka planda 25 FPS'dir. Bununla birlikte, videoların çözünürlükleri 120x160 iken video süreleri yaklaşık 4 saniyedir ve bir videoda yalnızca bir kişi bulunur. Resim 3.5'te 6 kategorinin örnek kareleri bulunmaktadır.

Resim 3.5: KTH veri setindeki 6 kategori için örnek kareler [6].

3.6 UCF – ARG Veri Seti

UCF-ARG veri seti, Central Florida Üniversitesi tarafından geliştirilmiş çoklu görüş içeren bir veri setidir [7]. Videolar Kingfisher Aerostat helyum balonu üzerindeki kameralardan, zeminde ve çatıdaki kameralardan elde edilmiştir. Videolar 10 farklı aksiyona hizmet eder: Boks yapma, Taşıma, El çırpma, Kazma, Koşu yapma, Açık-kapalı bagaj, Koşma, Fırlatma, Yürüme ve Sallanma. Resim3.6'da beş eyleme ait görüntüler bulunur. Video çözünürlükleri 1920x1080'dir ve videolar 60 FPS'tir.



Resim 3.6: UCF – ARG veri seti için bazı örnek görüntüler [7].

3.7 YouTube Anten (YouTube Aerial) Veri Seti

Veri setindeki makale yazarları tarafından toplanmıştır. Veri seti YouTube'daki uçangöz videolarından elde edilmiştir [8]. Videoların eylemleri ise, sıkça tercih edilen UCF-101 [9] veri setinin aksiyonlarından seçilmiştir Bunlar; bisiklet sürme, yamaç dalışı, golf yapma, at binme, kano sporu, koşma, kaykay yapmak, sörf yapma, yüzme ve yürümedir. Her aksiyon için 50 video bulunmaktadır. Ayrıcai videolar farklı yüksekliklerde ve kamera hareketlerinde çekilmiştir.

3.8 Weizmann Veri Seti

Veri seti Gorelick, Lena, et al. tarafından geliştirilmiştir ve 10 farklı aksiyon içermektedir: Koşma, Yürüme, Atlama, Çift Ayakla İleri Atlama, Çift Ayakla Yere Atlama, Dörtnala gitme, Çift elle dalgalanma, Tek elle dalgalanma, Bükme [10]. Eylemleri gerçekleştiren 9 farklı aktör bulunmaktadır ve toplam video sayısı 90'dır. Örnek görüntüler Resim 3.7'de verilmiştir.

Resim 3.7: Weizmann veri seti örnekleri [10].

3.9 HMDB51 Veri Seti

Veri seti 51 aksiyona kategorilendirilmiş 7000 farklı video klip içermektedir. Her kategori minimum 101 video klibi barındırır. HMDB veri seti YouTube, Google ve Prelinger arşivi gibi farklı platformlardan alınan videolardan oluşturulmuştur [11]. Eylem kategorileri 5 ana kategoride ayrıştırılmıştır. Bunlar; gülme ve konuşma gibi genel yüz eylemleri; sigara içme gibi obje hareketleriyle Yüz aksiyonları; yürüme, kalkma, dalma, itme gibi genel vücut hareketleri; saç tarama, topa vurma gibi obje etkileşimiyle vücut hareketleri; sarılma, el sıkışma gibi insan etkileşimiyle vücut hareketleri şeklindedir. HMDB veri seti diğer veri setlerine kıyasla daha karmaşıktır. Bunun sebebi ise, karmaşık arka planlar bulunması, video kalitesinin düşük olması ve farklı kategorilerin birbirine benzemesidir.

3.10 THUMOS Veri Seti

THUMOS 2014 ve THUMOS 2015 yarışmalarında kullanılmak üzere üretilmiş ve webden alınmış videolardan oluşmaktadır [55,56]. Bu çalışmada hem 2015 hem de 2014 yılındaki yarışmada toplanan veri setleri farklı çalışmalarda kullanılmıştır. 2014 yılındaki yarışmadan alınan verilerde YouTube'dan alınmış yaklaşık 3 dakikalik 20 farklı spor eylemlerini içeren videolar bulunurken, 2015 yılındaki ise webden alınmış yaklaşık 400 saatlik öğretici videolar ve spor videoları bulunmaktadır.

3.11 CUHK Avenue Veri Seti

CUHK Avenue veri seti CUHK kampüsünde kaydedilmiş 30652 kare içerir. Bu veri setinin 15328 karesi eğitim verisi olarak ayrılmış geri kalanı ise test verisi olarak ayrılmıştır. Toplamda 16 farkı eğitim ve 21 test verisi bulunmaktadır. Genellikle anormal durumların tespiti için kullanılan bu veri setinde 47 tane anormal olay içermektedir.

3.12 ShanghaiTech Kampüs Veri Seti

Shanghai Tech Üniversitesi tarafından hazırlanan veri seti anormal olayların tespiti için geliştirilmiştir. Diğer anormal durum tespiti yapan veri setlerinden farkı videoların farklı kamera açılarından çekilmesi ve videolarda farklı hava koşullarının gözlemlenebilmesidir. Videolarda 13 farklı sahne yer alır ve toplamda 317398 kare içeren 130 farklı anormal durum içeren video barındırır.

3.13 CAD-60 ve CAD-120 Veri Seti

CAD-60 ve CAD-120 veri seti, Cornell Üniversitesi'nin Robot Öğrenme Laboratuvarı'nda geliştirilmiştir [59]. Videolar Microsoft firmasının Kinect sensörü kullanılarak kaydedilmiştir. CAD-60 veri seti 12 farklı eylemi içeren toplamda 60 videodan oluşur. Bu eylemlere örnek olarak; diş fırçalama, yemek pişirme, telefonda konuşma verilebilir. Diğer yandan, CAD-120 veri seti 10 yüksek seviyeli aktiveler (obje temizleme, ilaç içme gibi), 10 alt aktivite etiketi (yemek yemek, yer değiştirmek gibi), ve 12 obje bağlayıcı etiket (içilebilir, kapatılabilir gibi) içeren 120 uzun günlük

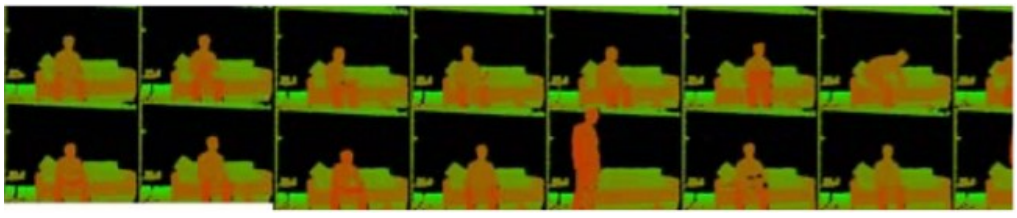
aktivite videosu içerir. Her iki veri setine ait karelerin bazıları Resim 3.8 ve Resim 3.9'da görülmektedir.

Resim 3.8: CAD – 60 veri seti örnek kareleri [59].

Resim 3.9: CAD – 120 veri seti örnek kareleri [59].

3.14 MSRDailyActivity3D Veri Seti

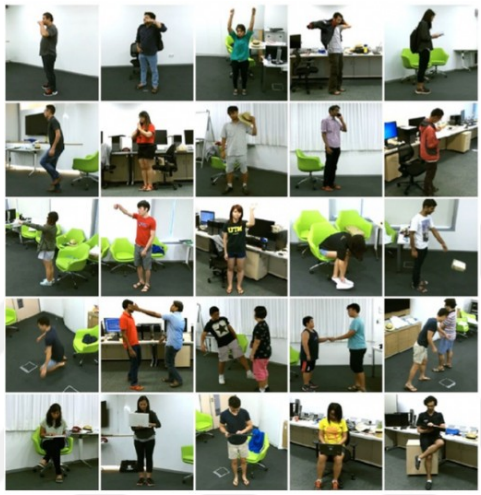
MSRDailyActivity3D veri seti, 16 farklı eylem için 320 eylem örneği içerir [60]. Eylemlere; kitap okuma, kağıt fırlatma örnek olarak verilebilir. Videolar Kinect sensöründen elde edilmiştir. Veri setine ait kareler Resim3.10'da örneklendirilmiştir. Veri setinin en önemli dezavantajı sabit kamera açısından elde edilmiş videolar içermesi ve az sayıda örnek bulundurmasıdır.



Resim 3.10: MSRDailyActivity3D veri setinden alınmış örnekler [60].

3.15 NTURGB+D Veri Seti

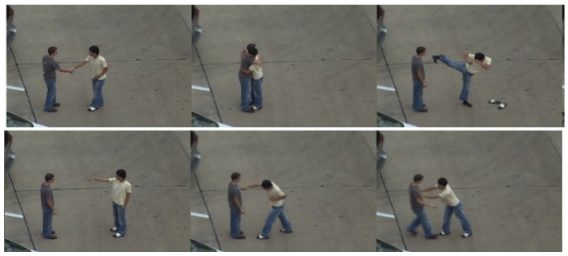
NTURGB+D veri seti, 56.880 eylem örneği ve 60 farklı eylem sınıfı içermektedir [61]. Tüm videolar 3 farklı Microsoft Kinect V2 kameralarından elde edilmiştir. Örneklerin çözünürlükleri 1920x1080 olmakla birlikte her videoda 3 boyutlu iskelet verisi bulunur. Videolardaki aksiyonları 40 farklı kişi gerçekleştirmiştir ve Resim 3.11'deki bazı eylemler şu şekildedir; su içme, neşelenme, top zıplatma vb.



Resim 3.11: NTURGB+D veri setinden alınmış bazı eylem videolarının örnek kareleri [61].

3.16 UT – Interaction Veri Seti

UT – Interaction veri seti 6 farklı insanlar arası etkileşimi gösteren 120 video içermektedir [62]. Bu insanlar arası etkileşim; el sıkışma, itme, sarılma, yumruk atma ve tekme atma şeklindedir ve örnekler Resim3.12'de bulunur. Her video yaklaşık 1 dakika uzunluğundadır ve videolarda 15 farklı kıyafetli kişi bulunmaktadır. Video çözünürlükleri 720x480'dir ve 30 FPS'dir.



Resim 3.12: 6 farklı etkileşimi gösteren örnek kareler [62].

3.17 UCF – 101 Veri Seti

Central Florida Üniversitesi tarafından geliştirilen 101 kategori içeren UCF-101 veri seti büyük bir video veri tabanıdır [9]. YouTube platformundan alınmış 13320 video içerir. Veri setindeki 101 kategori daha genel 5 kategoriye ayrıştırılmıştır. Bunlar; insan – obje etkileşimi, yalnızca vücut hareketleri, insan – insan etkileşimi, müzikal enstrüman çalma ve spordur. Veri seti çok sayıda video içerdiğinden farklı kamera açılarında çekilmiş videolar, aynı aksiyonun farklı videoları ve farklı arka planları içeren videolar bulunmaktadır. Bu sebeple çok fazla sayıda çalışma tarafından tercih edilmektedir. Bazı kategoriler Resim 3.13'de verilmiştir.

Resim 3.13: UCF – 101 veri setinden bazı kategori örnekleri [9].

3.18 Hollywood2 Veri Seti

12 farklı eylem içeren Hollywood2 veri seti, genellikle insan eylemini tanıma için seçilen bir veri setidir [104]. 3669 farklı video içermektedir. Videolar 69 film arasından seçilerek oluşturulmuştur ve Resim 3.14'te bazı örnek kareler bulunur. Eylemler: telefona cevap verme, araba kullanma, yemek yeme, kavga etme, el sıkışma, öpüşme vb. şeklindedir. Videolar saniyede 24 kare çerçeve içermektedir.

Resim 3.14: Hollywood2 veri kümesindeki filmlerden bazı anlık görüntüler [104].

3.19 MSR Action 3D Veri Seti

Derinlik kamerası tarafından yakalanan 20 farklı eylem kategorisi içeren bir veri setidir [105]. Videolar 640x480 çözünürlüğe ve saniyede 15 kareye sahiptir. Veri kümesinde 4020 video vardır. Videolar şu eylemleri içermektedir: Daire çizme, el çırpma, iki el sallama, boks, eğilme, öne tekme, yan tekme, koşu, tenis vb. Wanqing Li tarafından Microsoft Research Redmond'da oluşturulmuştur. Resim 3.15'te örnekler yer almaktadır.

Resim 3.15: MSR Action 3D veri setinden örnek bir görüntü dizisi [105].

3.20 Northwestern UCLA Veri Seti

UCLA'da oluşturulan ve birden çok Kinect kamerasından yakalanan derinlik, RGB ve iskelet verilerini içeren veri setidir [106]. 10 eylem ve 10 oyuncu içerir; tek elle toplama, iki el ile kaldırma, çöpü atma, dolaşma, oturma, ayağa kalkma, giyme, çıkarma, fırlatma, taşıma.

3.21 Olimpik Sporlar Veri Seti

Olimpik Sporlar veri kümesi Stanford Görü Labaratuvar'ında oluşturulmuştur [107]. YouTube'da farklı spor videoları toplanarak elde edilmiştir. Veri setinde yüksek atlama, cirit, sıçrama tahtası gibi 16 farklı spor kategorisi bulunmaktadır ve eylemlerden bazısı Resim 3.16'da bulunur.

Resim 3.16: Basketbol turnuvası, bowling, tenis servisi ve platformdan aksiyon örnekleri [107].

3.22 Kinetics Veri Seti

Kinetics veri kümesi üç farklı türden oluşur: Kinetics 400, Kinetics 600 ve Kinetics 700 [108]. Veri kümelerinin ayrılma noktası, içerdikleri videoların sayısıdır. Geliştirilen ilk veri seti olan Kinetics 400, her sınıfta 400 sınıf ve minimum 400 video içerir. Veri kümesi, YouTube'dan alınan profesyonel olmayan videolardan (dağınıklık, sarsıntı / hareket durumları dahil) oluşturulmuştur. UCF-101 ve HMDB-51 verileri, insan eylemi tanıma için kullanılan en ayrıntılı veri kümeleri arasındadır [9,11]. Ancak, artık modellerin geliştirilmesi için yeterli sınıf içermemektedir. Bu nedenle Kinetics 400, bu veri kümelerinden esinlenmiştir. Videolarda tekil eylemler, insan-insan etkileşimleri ve insan-nesne etkileşim eylemleri yer almaktadır.

3.23 UTD-MHAD Veri Seti

Texas Üniversitesi'nde geliştirilen UTD-MHAD veri seti 27 farklı eylem içermektedir [109]. Resim 3.17’de örnek eylemler bulunur. Videolar, Kinect kameraları ve giyilebilir sensörlerin yardımıyla elde edilmiştir. 861 farklı videoyu içeren bu veri

setinde 640x480 ve 320x240 olmak üzere iki farklı çözünürlükte videolar yer almakta ve videoların saniyedeki kare sayısının 30 olduğu belirtilmektedir.

Resim 3.17: Sırasıyla Kol çapraz, Basketbol atışı, X Çiz, Daire çiz (saat yönünde), Daire çiz (saat yönünün tersine) eylemlerinin örnek kareleri [109].

3.24 Something-Something V2 Veri Seti

Something-Something, veri setinde 220.847 farklı video içeren çok büyük bir veri setidir [110]. İnsanların günlük yaşamlarında kullandıkları nesnelerle elde edilen videoları içerir. 174 farklı eylem içeren ve Resim 3.18'de örnekleri bulunan bu veri

seti Twentybn tarafından oluşturuldu. Videoların çözünürlüğü 320x240 olmasına rağmen ortalama video süresi 4.03 saniyedir.

Resim 3.18: Videolardan örnek kareler [110].

3.25 Charades

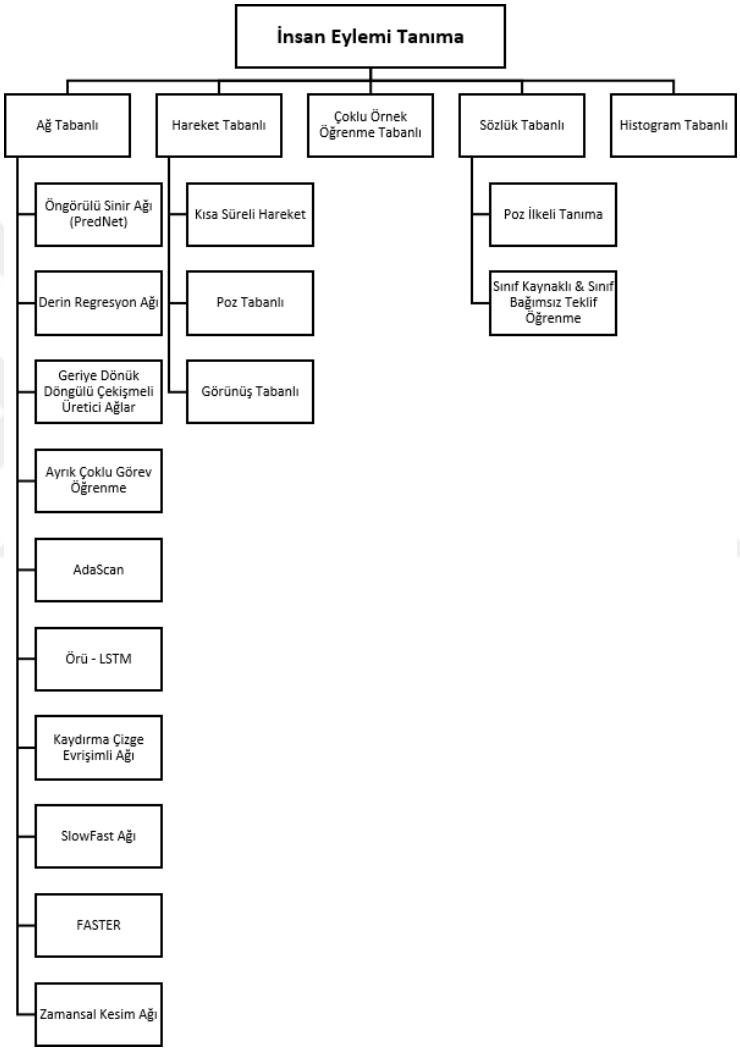
Charades veri seti, Allen Institute for AI'nın Perceptual Reasoning an Interaction Research ekibi tarafından geliştirilmiştir [111]. Amazon Mechanical Turk üzerinden iç mekan aktivitelerini içeren videolar toplanmıştır. 267 kullanıcı tarafından kaydedilen videolar ile 9848 video içeren bir veri seti oluşturulmuştur. Çözünürlük ve FPS değerlerinin değiştiği videolarda ortalama video süresi 30,1 saniyedir.

4. YAKLAŞIMLAR VE METOTLAR

İnsan eylemlerini tanıma günümüzde en popüler konulardan biridir. Bu konu yıllar içerisinde gelişmeye devam etmektedir. İnsan eylemini tanıma işlemleri birçok farklı yöntem kullanılarak gerçekleştirilebilir. Kullanılan yöntemler, hedef sorunun ve

verilerin çözümüne odaklanarak geliştirilmiştir. Tanınacak faaliyetlerin karmaşıklığı, yaklaşımların kullandığı yöntemleri etkiler. Bu araştırma kapsamında ele alınacak çalışmalar yöntemlere göre sınıflandırılmış ve taksonomi diyagramı verilmiştir.

Araştırma çalışması boyunca ele alınacak makaleler yukarıda verilen Şekil 4.1'deki taksonomiye göre kategorilendirilecek ve bu başlıklar altında incelenecektir.Araştırma kapsamındaki tüm makaleler Çizelge 4.1 verilmektedir.



Şekil 4.1: Araştırma boyunca ele alınan makalelerin taksonomisi.

Çizelge 4.1 : Araştırma kapsamında incelenecek makaleler.

Çalışma	Yazarlar	Yayımcı	Yayın Tarihi	Alıntılanma Sayısı

Deep Predictive Coding Networks For Video Prediction And Unsupervised Learning [1]	William Lotter, Gabriel Kreiman & David Cox	International Conference on Learning Representations (ICLR)	2017	532
Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos [12]	Fabian Caba Heilbron, Juan Carlos Niebles, Bernard Ghanem	The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	2016	188
Anticipating Visual Representations from Unlabeled Video [13]	Carl Vondrick, Hamed Pirsiavash, Antonio Torralba	The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	2016	261
Human Action Recognition in Drone Videos Using a Few Aerial Training Examples [8]	Waqas Sultani, Mubarak Shah	IEEE Robotics and Automation Letters with International Conference on Robotics and Automation (ICRA) option	2020 (under review)	2
Predicting Future Frames using Retrospective Cycle GAN [14]	Yong-Hoon Kwon, Min-Gyu Park	The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	2019	32
A New Hybrid Architecture for Human Activity Recognition from RGB-D Videos [15]	Srijan DasEmail, Monique Thonnat, Kaustubh Sakhalkar,Michal K operski, Francois Bremond	International Conference on Multimedia Modeling	2019	10
Human Activity Prediction: Early Recognition of Ongoing Activities From Streaming Videos [16]	M. S. Ryoo	International Conference on Computer Vision	2011	523
AdaScan: Adaptive Scan Pooling in Deep convolutional Neural Networks for Human Action Recognition in videos [17]	Amlan Kar, Nishant Rai, Karan Sikka, Gaurav Sharma	The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	2017	110
Lattice Long Short-Term Memory for human Action Recognition [18]	Lin Sun, Kui Jia, Kevin Chen, Dit-Yan Yeung, Bertram E. Shi, Silvio Savarese	The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	2017	100
Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning [19]	Saad Ali, Mubarak Shah	IEEE Transactions on Pattern Analysis and Machine Intelligence	2010	539
Pose Primitive Based Human Action Recognition in Videos or Still Images [20]	Vaclav Hlavac, Christian Thurau	The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)	2008	363
Skeleton-Based Action Recognition With Shift Graph Convolutional Network [120]	Cheng, K., Zhang, Y., He, X., Chen, W., Cheng, J., & Lu, H.	IEEE/CVF Conference on Computer Vision and Pattern Recognition	2020	15

Çizelge 4.1 : Araştırma kapsamında incelenecek makaleler.

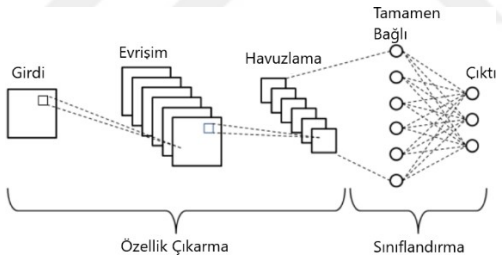
FASTER Recurrent Networks for Efficient Video Classification [121]	Linchao Zhu, Laura Sevilla-Lara, Du Tran, Matt Feiszli, Yi Yang, Heng Wang	IEEE/CVF Conference on Computer Vision and Pattern Recognition	2020	5

Slowfast networks for video recognition [125]	Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, Kaiming He	IEEE international conference on computer vision	2019	405
Temporal segment networks for action recognition in videos [127]	Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, Luc Van Gool	IEEE transactions on pattern analysis and machine intelligence	2018	172

4.1 Ağ Tabanlı Tanıma Yöntemleri

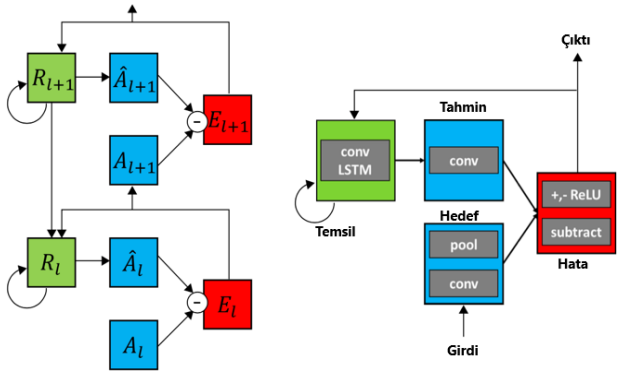
4.1.1 Öngörülü Sinir Ağı (Prednet)

Sinir Ağı yapıları son zamanlarda karmaşık problemleri çözmek için kullanılan yöntemlerdir. Bu yöntemler insan beyninin yapısına benzetilir. Derin öğrenme için kullanılan derin sinir ağlarından biri olan Evrişimli Sinir Ağları, birden fazla gizli katmana sahip sinir ağları gibi doğrusal olmayan işlemlerin çok düzeylerinden oluşur. Amaç, daha düşük seviye özellikleri kullanarak daha üst seviye özellikler oluşturarak özellik hiyerarşisini öğrenmektir. Bu başlık altında, derin sinir ağlarına dayalı yöntemler tartışılacaktır. Şekil 4.2 basit bir sinir ağı verilmiştir.



Şekil 4.2: Basit bir Evrişimli Sinir Ağı diyagramı [27].

Günümüzde, çoğu insan eylemi tanıma yöntemi gözetimli öğrenme ile çözülmektedir. Ancak; buradaki önemli nokta tanıma sürecini gözetimsiz bir yöntemle analiz etmektir. Bu bağlamda makalenin yazarları, gözetimsiz yöntemler kullanılarak gelecekteki karelerin tahmini için bir Öngörülü Sinir Ağı (PredNet) geliştirmiştir [1] ve mimarisi Şekil 4.3'te yer almaktadır. Sinirbilim araştırmalarından esinlenerek keşfettikleri bu yöntem, bir dizi tekrarlanan kümelenmiş modülden oluşmaktadır. Bu modüller; modüle girdi olarak verilen verilerden yerel tahminler üretir. Ardından, tahmin değeri gerçek giriş değerinden çıkarılır ve sonuç bir sonraki katmana gönderilir.



Şekil 4.3: Öngörülü Sinir Ağı (PredNet) mimarisi [1].

Ağdaki her modül 4 ana bölümden oluşur: A_l giriş evrişimli bir katmandır, R_l tekrarlamalı bir temsil katmanıdır, $\hat{A}_l$ bir tahmin katmanıdır ve E_l bir hata temsilidir.

R_l, Tahmin yapan tekrarlamalı evrişimli bir ağdır. $\hat{A}_l$, giriş katmanının ne olacağını izler ve A_l sonraki karede ne olacağını tutar. E_l, temsil hatası olarak A_l ve $\hat{A}_l$ arasındaki farkı verir, ardından E_l temsil hatasını pozitif veya negatif olarak böler.

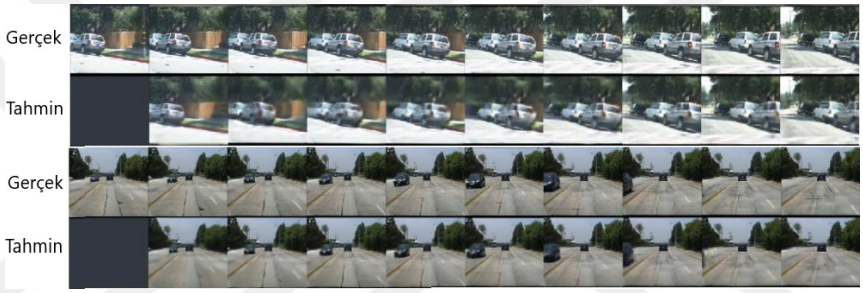
Ağ mimarisi 1999'da yapılan bir çalışmadan esinlenmiştir [21]. Modern yöntemler kullanılarak yeniden formüle edilmiş ve uçtan uca gradyan inişi (gradient descent) ile eğitilmiştir. Aynı zamanda ağda yerleşik bir kayıp fonksiyonu oluşturulmuştur.

Mimari video verilerine odaklanmıştır (görüntü dizisi). Hesaplama için ReLU aktivasyonu ve maksimum havuzlama kullanılmaktadır. Ayrıca, evrişimli LSTM üniteleri, nöronları temsil etmek için kullanılır [22]. Model, faaliyetlerin hata değerlerinin ağırlıklı toplamını en aza indirecek şekilde eğitilmiştir. PredNet algoritmasının hesaplanmasına göre durum güncellemeleri 2 geçişle gerçekleştirilir. R_l Durumları yukarıdan aşağıya geçişte hesaplanır. Ardından, tahminler, hatalar ve daha yüksek düzeydeki hedefler ileri geçişle hesaplanır.

Modelin çalışması 2 veri seti üzerinde gerçekleştirilmiştir. İlk olarak, PredNet çerçevesinin genel temsilini anlamak için FaceGen yöntemiyle üretilen dönen yüzlerden oluşan veri setiyle test edilmiştir. Daha sonra, KITTI veri seti videolarında işlemler yapılmıştır. Bu araştırma çalışması kapsamında video veri seti KITTI sonuçlarından bahsedilmiştir. Ağın gerçekten güçlü bir temsil öğrendiğini görmek için CalTech Yaya veri seti ile test işlemleri gerçekleştirilmiştir. Test sonuçlarında;

kameralı araç dönerken çekilen zor video senaryolarında bile mantıksal sonuçların üretildiği görülmektedir. Tahminler girdi olarak verildiğinde ve tekrarlı yinelemeler yapıldığında PredNet'in hem tek bir gelecek kareyi hem de çoklu kareyi tahmini yapabileceği belirtilmektedir.

PredNet modelleri diğer modellere göre iyi sonuçlar verdiği gözlemlenmiştir. Bu başarının sadece hiper parametrelerin seçiminden kaynaklanmadığını kanıtlamak için, katmanlar, filtre boyutları ve katman başına filtre sayısından oluşan 4 farklı hiper parametre seti ile yeniden modellenmiştir. Tüm 4 setten elde edilen sonuçlarda, PredNet sadece alt tabakadaki $L_{\#}$ kayıp değeri kullanılarak hesaplanmıştır, MSE $3.13x10^{\$3}$ ve SSIM 0.884 olarak hesaplanmıştır ve performansı karşılaştırılan 3 yöntemden daha yüksektir ve tahminler Resim 4.1'deki gibidir.



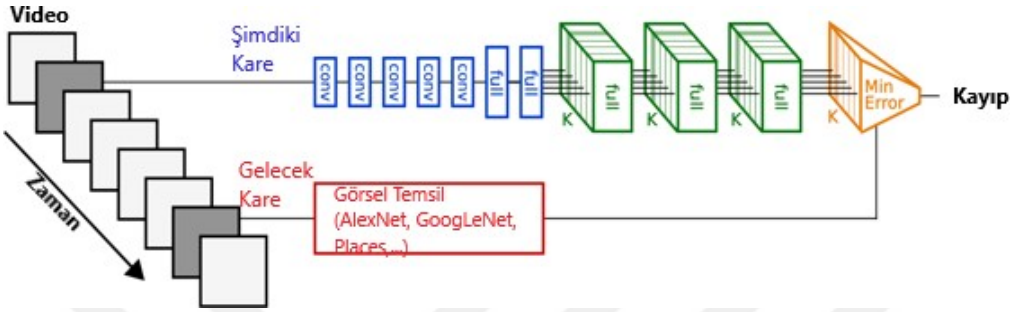
Resim 4.1: PredNet ile örnek tahmin sonuçları [1].

4.1.2 Derin Regresyon Ağı

Eylemleri ve nesneleri oluşmadan önce tahmin etmek, bilgisayarla öğrenme alanında zor bir sorun olarak kabul edilir. Problemin çözümünde kullanılan büyük ölçekli videolarda, eylem tanıma ve akıllı yaklaşımlar hakkında bilgi gerektirir. Ele alınan makale, etiketlenmemiş videodan görsel temsiller öngören yeni bir çerçeve sunmaktadır [13]. Bu yöntem, yaklaşımı nedeniyle diğer yöntemlerden farklı olduğunu gösterir. Diğer yöntemlerin aksine, piksel tabanlı bir tahmin gerçekleştirmez ve gözetimli öğrenme yöntemleri kullanmamaktadır. Başka bir deyişle, her piksel için işlenmek üzere uzun zaman kaybetmezler, bunun yerine daha üst düzey bir konsept sunarlar ve etiketli kategorilerle tahmin etmezler.

Çalışmadaki ana fikir, etiketlenmemiş videoları tahmin için kullanmaktır. Bu nedenle, görsel sunumları öngörmek için kendi kendini gözeten öğrenmeyi kullanırlar. Bu adımdaki amaç, video *i*'nin *t* zamanındaki $x^{\&}_{t}$ karesinden, $x^{\&}_{t\text{'}(}$ karesini tahmin

etmektir. Tahmin gösterimleri için derin regresyon ağ modelinin kullanılmasını önerilmektedir. Geliştirilen yöntemde etiketler için veri gerekmediğinden çok sayıda eğitim verisi elde edilmiştir. Böylece; veri karmaşıklığı artırılarak model karmaşıklığı artırılabilmekte ve büyük veriler olasılıksal geçişli inişi ile verimli bir şekilde eğitilebilmektedir.



Şekil 4.4: Derin regresyon ağının mimarisi [13].

Yöntemin mimarisi AlexNet mimarisinden esinlenilerek geliştirilmiştir [23, 24] ve Şekil 4.4'te görselleştirilmiştir. AlexNet mimarisinden farklı olarak, kayıp fonksiyonu içerir ve 3 tam bağlı katman daha içerir.

Öte yandan, geliştirme sırasında önemli kabul edilen ve tahmini etkileyen bir faktör çok modlu çıktılardır. Çok modlu çıktı, bir videoda birden fazla makul gelecek olduğunda oluşur. Derin regresyon ağı, çok modlu çıktılar üretmek için genişletilmiştir. Bu durumda; bir giriş karesi için K çıkışı mümkün olduğunda, K ağının bir karışımı eğitilir. Her karışım gelecekteki modlardan birini tahmin etmek için kullanılır. Ancak; burada birden fazla derin regresyon ağını eğitmek için 2 farklı zorlukla karşılaşılmıştır. İlk olarak, çok modlu çıktılar nadiren bulunmaktadır. İkinci olarak, hangi K karışımının hangi kareye karşılık geldiği bilinmemektedir. Sonuçta her iki sorunun da üstesinden gelmek için gizli değişken yardımıyla karışım ataması yapılır.

Derin regresyon ağı, TV İnsan Eylem Etkileşimi veri kümesi [5] ve THUMOS 2015 veri kümesi [56] gibi çevrimdışı videolarla eğitilmiştir, ancak akış videoları ile eğitim vermenin de mümkün olduğu belirtilmiştir. Bu durum nispeten farklıdır; çünkü sürekli öğrenme akıştaki bir karenin bilgisini saklamaksızın gerçekleştirilir.

Etiketlenmemiş video, ağda kullanıldığından bu videoları etiketleme işlemi gerçekleştirilmelidir. Bu işlem için, hedef görevden nispeten küçük örnekler seçilir ve bu örnekler kategorilendirme için kullanılır. Standart tanıma algoritmaları kategori tahmini için uygulanabilir, çünkü tahmin işlemleri en son teknoloji tanıma sistemleri ile aynı yöntemler kullanılarak gerçekleştirilir. Tahmini gösterim elde edildiktensonra, tanıma algoritmasını uygulamak için 2 farklı strateji geliştirilmiştir. Birincisi, tahmini gösterime uygulamaktır, kategorinin standart özelliklerine sahip eğitimli bir görsel sınıflandırıcı kullanılır. İkincisi, regresyondaki yapısal hatalara uyum sağlayan öngörülen temsil üzerine eğitilmiş görsel sınıflandırıcıdır.

Çıkarım sırasında model, gelecekteki çoklu temsilleri tahmin eder. Kategori sınıflandırıcılar, tahmin edilen her bir gösterim için geçerlidir ve hangi kategorilerin mümkün olduğunu gösteren bir dağıtım elde edilir. Bu dağılımları marjinalleştirerek, ağ en olası kategorinin hangisi olduğuna karar verir.

Sistemin uygulama detayları, çalışmada kullanılan yöntemler kadar önem kazanmaktadır. Ağ mimarisi 5 evrişimli ve 5 tam bağlı katman içerir. Ağ genelinde ReLU doğrusal olmayan etkinleştirme kullanılmıştır. Evrişimli kısım AlexNet [23] mimarisini takip etmektedir. Tam olarak bağlanmış 5 katmanın her birinde 4096 gizli katman bulunur. K ağları, parametreleri ayırabilir veya paylaşabilir. Bu kapsamda deneylerde bağımsız parametreleri azaltmak için bir strateji uygulanmıştır: 5 evrişimli ve ilk 2 gizli katman için bunlar her karışıma bağlanır. Son 3 tam bağlı katman için; gizli birim araya eklenir. Bu işlem bir kez yapılır ve öğrenme sırasında değişmez. Ağlar olasılıksal gradyan iniş eğitilir. Tüm bu işlemler sırasında ise Tesla K40 GPU kullanılmış ve ağ Caffee'de uygulanmıştır [25].

Derin regresyon ağının tek bir kareden 1 saniyenin geleceğini tahmin etme tahmin doğruluğu %43,6 ∓ 4,8 olarak elde edilmiştir. Bir diğer yandan, çoklu tahminler göz önünde bulundurulmuştur. Gelecek açıkça görülemediğinde, birden fazla tahmin yapılmaktadır. Makalede, sadece önerilen yöntemle geleceğe yönelik tahminler yapmanın yanı sıra, 12 farklı gönüllüden tahminlerde bulunmaları istenmiş ve gönüllüler eğitim setine çalıştırılmış daha sonra gönüllülerden test setinde tahminler yapmaları istenmiştir. İnsan doğruluğu %71 oranında doğruluk sağlamıştır. Fakat; birden fazla tahmin yapılmıştır, çünkü insanın hem problemi hem de doğası inkâr edilemez. Burada, çoğunluk oyu ile (%85) doğruluk kabul edilmiştir. Örnek tahminler Resim 4.2'de yer alır.

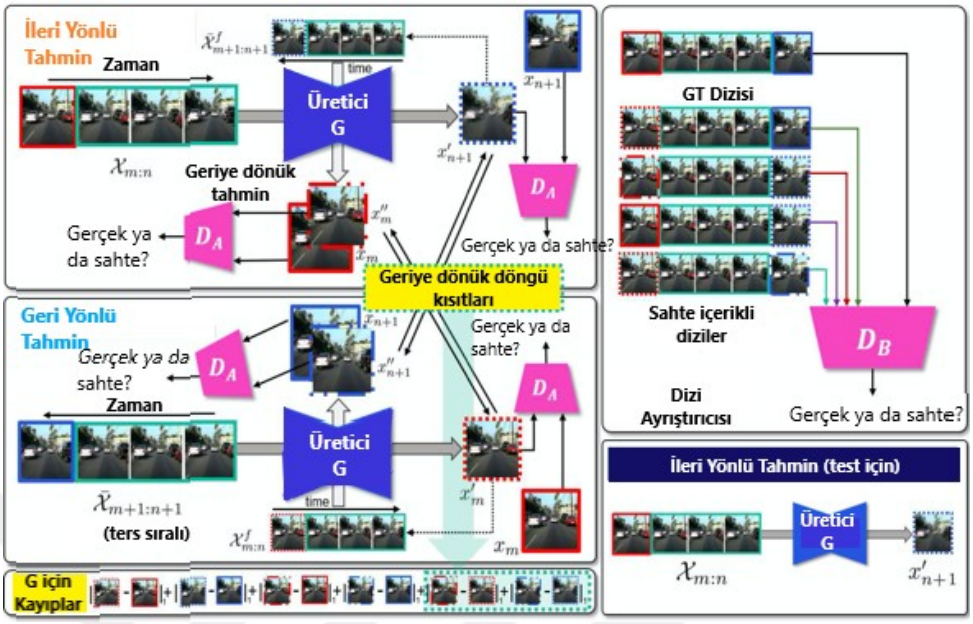
Makalede yapılan eylem tahminine ek olarak, nesnelerin tahminleri de yapılır. Bu bağlamda, nesnelerin tahmini için benmerkezci (egocentric) videolarda ağın eğitimi için kullanılan veri setinin %75'i kendi kendini gözeten öğrenme için ayrılmıştır. Geri kalanı ile, nesnenin kategori etiketleri, “leave-one-out” yöntemi kullanılarak elde edilmiştir. Çerçevede birden fazla nesne varsa, ortalama tahmin nesne görünmeden 5 saniye önce değerlendirilir ve bu durum “leave-one-out” ayarlarına göre ortalama bir değer olduğu durumda geçerlidir. Yazarların nesne tahmini sırasındaki hipotezi, nesnelerin arka planının nesnelerin ne olduğunu tahmin etmede yararlı olacağıdır. Nesne tahmini için ortalama kesinlik değeri 10.1 verilmiştir.

Resim 4.2: Modelden elde edilen bazı tahmin sonuçları [13].

4.1.3 Geçmişe Dönük Çevrimli Çekişmeli Üretici Ağlar Yaklaşımı

Gelecekteki analiz çalışmaları gün geçtikçe artmıştır. Artan çalışmalara bağlı olarak, tahminlerin performansı giderek artmaktadır. Ancak; performans arttıkça bulanık tahmin kareleri belirmektedir. Bu çalışmada, hem geçmiş hem de gelecek kareleri

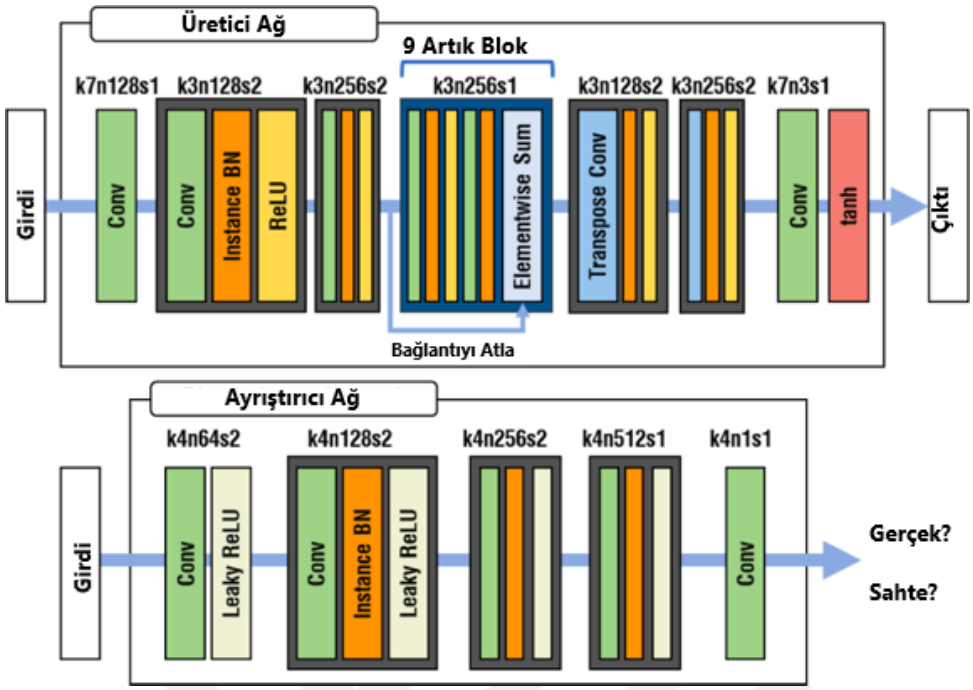
öngörecek tek bir üreteç (generator) eğitmek amaçlanır [14]. Önerilen kare 1 üreteç ve 2 ayırtaçtan oluşur. Bunlar; kare ve dizi ayırtaçlarıdır.



Şekil 4.5: Metodun genel işleyişi [14].

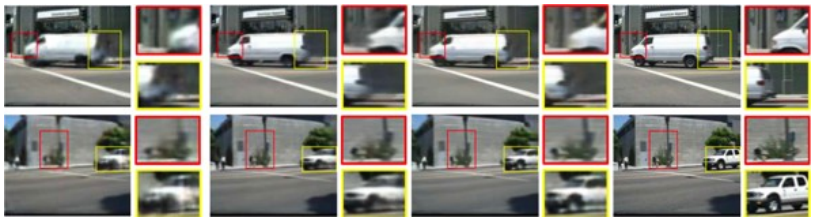
Girdi sahte olsa bile, üretici hem geçmişi hem de sonraki kareyi tahmin eder. Ayrıca; kare ayırtaç sahte çerçeveleri ayrı ayrı ayırt edebilirken, dizi ayırtaç ise dizinin sahte kareler içerip içermediğine karar verir. Genel işleyiş Şekil 4.5'de bulunur.

[37] 'den esinlenen ağ mimarisi, Şekil 4.6'da yer almaktadır. Buradaki önemli fark, üreteçin gelecekteki kare tahmini için girdi olarak birden fazla görüntü almasıdır. Üreteç ağı 4 evrişimli katman, 9 artık (residual) blok ve 2 devrik evrişimli katmandan oluşur. Ayırıcı, sızdıran ReLU 5 evrişimli katmana sahiptir. Aynı zamanda; ağ yapısı, kare ve dizi ayırıcılarındaki girdi görüntülerinin sayısı dışında aynıdır. Ayrıca; örnek normalizasyonu [38] şeması girdi ve çıktı katmanları hariç tüm üreteç ve ayırtaç katmanlarında kullanılır.



Şekil 4.6: Yaklaşımın ağ mimarisi [14].

Eğitim operasyonları KITTI [3] veri seti ile yapıldığı için, Geçmişe Dönük Çevrimli GAN'ın performansını değerlendirmek amacıyla, CalTech Pedestrian veri setini [4] ölçmek için PredNet [1] protokolleri takip edilmiştir. Nicel analiz için 3 ölçüm tartışılmıştır. Bunlar; MSE, SSIM ve PSNR'dir. MSE için düşük değerler, SSIM ve PSNR için yüksek değerler iyi kabul edilir. Araştırmalar 2 farklı veri seti üzerinde gerçekleştirilmiştir. İlk olarak, CalTech Pedestrian veri seti için sonuçlar MSE için 1.61, PSNR için 29.2 ve SSIM için 0.919'dur. CalTech Pedestrian veri seti [4], kameranın hızlı hareketi nedeniyle zorlayıcı görünmektedir, örnek sonuçlar Resim 4.3'de verilmiştir. Bu nedenle hatalar daha yüksektir. İkinci olarak, UCF - 101 veri kümesi için sonuçlar MSE için 1.37, PSNR için 35.0 ve SSIM için 0.94'tür.



Resim 4.3: Örnek sonuçların video kareleri [14].

Öte yandan, girdi çerçevesi sayısının duyarlılık değerleri de değerlendirilmiştir ve Çizelge 4.2'de bulunmaktadır. Giriş dizilerinin optimal uzunluğunun PSNR için 4 ve SSIM için 6 olduğu belirtilmektedir. Bu konuda önemli bir fark gözlenmemiştir. Ancak; dikkat edilmesi gereken nokta 2 fotoğrafın 8 fotoğraftan daha iyi sonuç vermesidir. 2 resmin yeterli olduğu durumlar sadece eğitim verileri yeterli olduğunda geçerlidir.

Çizelge 4.2: Değişen girdi resim çerçevelerine göre PSNR ve SSIM değerleri.

Resim Sayısı	2	4	6	8	10
PSNR	29.167	29.222	29.006	28.940	29.009
SSIM	0.9193	0.9189	0.9208	0.9197	0.9189

Makalede önerilen yaklaşımda tartışılan bir diğer olgu da çok adımlı tahmindir. Videolarda geleceğini tahmin etmek için çok adımlı tahmin kullanılmıştır. İlk verilen giriş dizisinden sonraki kare tahmin edilir, daha sonra dizinin son 3 karesi ve tahmin edilen kare birleştirilir ve yeni bir dizi oluşturulur. Elde edilen bu yeni dizi ile yeni bir kare tahmin edilmektedir. Bu işlem, tahmini kare sayısı belirlenene kadar devam eder. Yukarıdaki çok adımlı tahmin ile ilgili sonuç, kare sayısı arttıkça hata sayısının artmasıdır.

Son olarak, farklı ortamlarda ve yöntemlerde kullanılan adımların etkisini görmek için bir çalışma yapılmıştır. Elde edilen sonuçlarda herhangi bir modülün bulunmadığı, performansı düşürdüğü gözlemlenmiştir. Geçmişe dönük bir tahminin bulunmaması, kayıp terimlerinin geri tahminle ilgili olduğu ve eğitim sırasında ortadan kaldırıldığı anlamına gelir. Bu, giriş görüntülerinin sayısını yarıya indirir.

4.1.4 Ayrık Çok Görevli Öğrenme

Gelişen görüntü teknolojileri ile uçangözler hayatımızda yer almaya başladı. Bu bağlamda, birçok alanda kullanılan uçangözlerden görüntü almak önemli hale gelmiştir. Eylem tanıma için yeni bir bakış açısı sağlayan bu çalışmada, uçangözlerden kaydedilen videolar üzerinde bir yaklaşım geliştirilmiştir [8]. Seçilen video kaynağı, tanımada çok yeni olduğundan hazır bir veri kümesi bulmak son derece zordur. Bu çalışmada, sınırlı sayıda uçangöz videosu kullanılarak bir eylem tanıma işlemi gerçekleştirilmiştir.

Yazarlar sınırlı video sayısını artırmak için bazı yöntemler izlemişlerdir. GAN [32] kullanılarak gerçekçi görünümlü sahte videolar üretilebilir. Ancak; bu videoların kalitesi, tanıma için yeterli kabul edilmez. Buna rağmen; son çalışmalar GAN ile sahte özelliklerin elde edilebileceğini göstermektedir. GAN, 2 farklı ağdan oluşur. Bunlar; üreteç ve ayırtaçtır. Üreteç gerçek verileri tahmin etmeye çalışır ve gerçekçi görünümlü özellikler üretir, sahte özellikleri sağlam bir şekilde sınıflandırmak için ayırtaçı kandırmaya çalışır. "Vanilla-GAN" ile karşılaştırıldığında hem üreteç hem de ayırtaç koşullu GAN'daki farklı bilgilere odaklanır. Bu bilgiler video etiketleri veya özellik bilgileri olabilir.

Çalışma kapsamında GAN'ın kullanım amacı; yerdeki kameralardan kaydedilen videolardan gerçek özellikler elde etmek ve havadaki görüntülerden sahte özellikler üretmektir. İşlevi optimize etmek için KL veya LS sapması kullanılır. Bunun sebebi; Gerçek ve sahte veri dağılımları arasındaki farkı azaltmaktır. Ancak; KL ve LS ayrışmasının bazı sınırlamaları vardır. Bu sınırlamalardan biri; mesafe arttıkça, ıraksaklık eğimi azalır, böylece üretici hiçbir şey öğrenemez. Bu duruma bir çözüm bulmak için "Wasserstein GAN" tercih edilir. Kullanılan Wasserstein mesafesinin herhangi bir noktada daha yumuşak bir eğimi vardır. Wasserstein mesafesini daha pürüzsüz hale getirmek için, gradyan ceza kaybında 1-Lipschitz kısıtı kullanılır [33]. Dikkate alınması gereken bir nokta; GAN gerçekçi özellikler üretebilir ancak bu özelliklerin sınıflandırma için uygun olacağını garanti etmez.

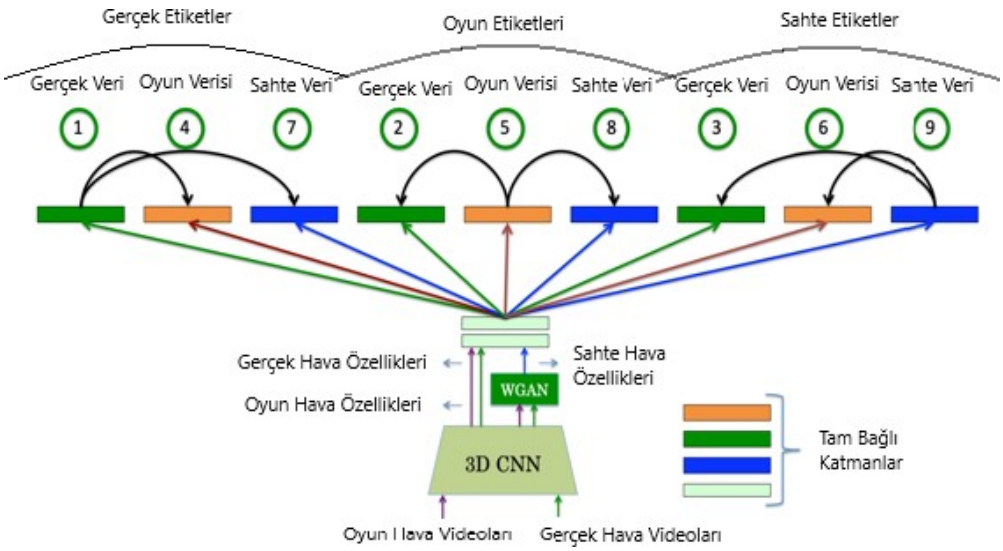
Soft-max sınıflandırıcılar bu durumu çözmek için gerçek hava örnekleri ile eğitilir. WCGAN-GP'yi ayırt edici özellikler üretmeye zorlamak için, sahte hava örnekleri kullanılarak hesaplanan sınıflandırıcı kaybı kullanılır.

Son çalışmalar, yeniden inşa etme kaybının, GAN kaybına ek olarak üretilen numunelerin kalitesini artırdığını göstermektedir. Buradaki yöntemde, yeniden inşa kaybı anten (aerial) ve zemin (ground) için çift videoları gerektirir. Yine de bu çiftleri bulmak yeterince zordur. Buna rağmen üreteç, toplanması nispeten kolay olan oyun videolarından seçimler yaparak (hem anten hem de zemin için çiftler) yeniden inşa kaybı ile geliştirilmiştir. Bir diğer deyişle, üreteç her bir zemin özelliğine karşılık gelen anten özelliğini de ayrıca üretmelidir.

Tüm bu işlemleri gerçekleştirmek için hesaplanan görsel özellikler 3 boyutlu çoklu fiber ağ ile yapılır. Çoklu fiber ağ, karmaşık sinir ağlarını daha küçük ve daha az ağırlıklı ağlara bölerek hesaplama işlemini gerçekleştirdiğinden diğer özellik ağlarından daha hızlı çalışır.

Özellik çıkarma işlemi tamamlandıktan sonra sınıflandırma işlemi başlatılır. Burada, sınıflandırma için kullanılan yöntem Ayrık Çok Görevli Öğrenmedir. Çoklu görev öğrenme, birden fazla görevi öğrenerek modelin genelleme yeteneklerini artırmayı amaçlamaktadır. Bu metot aynı zamanda eşzamanlı nesne algılama, kesimleme ve iskelet poz tahmini gibi birçok alanda kullanılır. Çok görevli öğrenmenin bir dezavantajı; aynı veriler için birden fazla etiket gerektirir. Ancak; çoğu veri kümesinde bu tür veriler bulunmaz. Bu duruma bir çözüm olarak; ayrışık çok örnekli çerçeve (disjoint multi-instance framework) geliştirilmiştir. Bu sayede, derin ağın genelleştirilmesini geliştirmek için farklı veri setleri tercih edilebilir.

Bu çalışma kapsamında; çok yönlü öğrenme için hem oyun eylem veri seti hem de sahte anten video veri seti kullanılmıştır. İki veri kümesinin farklı olduğu ve aynı eylem sınıflarını içermediği varsayılmaktadır. Eylem sınıflandırması 3 farklı veri ile gerçekleştirilmiştir. Bunlar; gerçek, sahte ve oyun videolarıdır. İlk olarak, gerçek ve oyun videoları 3 boyutlu evrişimli sinir ağı kullanılarak hesaplanır [34]. Daha sonra, yukarıda açıklanan GAN yöntemi ile sahte anten özellikleri elde edilir. 3 görev arasında paylaşılan 2 tam bağlı katman ve her göreve ayrılmış 1 tam bağlı katman vardır.



Şekil 4.7: Ayrık Çok Görevli Öğrenme yönteminin mimarisi [8].

Yukarıdaki Şekil 4.7 göz önüne alındığında, 1, 5 ve 9 yer gerçeği etiketleri ile eğitilmiştir. 4 ve 7, oyun ve sahte veriler için gerçek etiketleri tahmin eder. 2 ve 8, gerçek ve sahte veriler için oyun etiketlerini tahmin eder. 3 ve 6, gerçek ve oyun verileri için sahte etiketleri tahmin eder. Çok görevli çerçeve gerçek, sahte ve oyun verileri için eğitilmiştir. İlk olarak; kayıp 1, 2 ve 3 için birkaç gerçek anten videosu ile hesaplanır. Buna karşın; 1 gerçek veri etiketi, 2 oyun veri etiketi ve 3 sahte veri etiketi tahmin edilir. Temel gerçek etiketleri gerçek videolar için olsa da gerçek videolar için oyun ve sahte eylem etiketleri yoktur. Çünkü; sorunun doğasında bir eşitsizlik vardır. 5 ve 9'un tahmini, 2 ve 3 sınıflandırma kaybı için temel doğruluk etiketleri olarak kabul edilir.

Sonradan; 4, 5 ve 6 kayıpları hesaplanır ve ağ oyun videolarıyla eğitilir. 4 gerçek veri etiketlerini, 5 oyun veri etiketlerini ve 6 sahte veri etiketlerini tahmin eder. Oyun hava videoları ağa giriş olarak verilir ve sadece 5 tane zemin gerçeği oyun etiketine sahiptir. 1 ve 9 tahmini, 4 ve 6 sınıflandırma kaybını hesaplamak için kullanılır. Tüm bu işlemler birkaç kez uygulanır ve doğrulama verilerinde ince ayarlamalar yapılır.

Tüm sınıflandırma sonuçları gerçek anten videolarında gerçekleştirilmiştir. Aşağıdaki Çizelge 4.3'te, ilk satır yer kameralarından eğitim alırken elde edilen sonuçları ifade eder. UCF-ARG veri kümesi için yer videoları şu anda mevcuttur [7]. Diğer yandan, YouTube anten veri kümesi için yer videosu yoktur [8]. Bu nedenle; UCF-101 veri kümesinden seçilen 8 eylemin videoları, YouTube anten veri kümesi için yer videoları olarak seçilmiştir.

İkinci satır, oyun videolarının ayrık çoklu görev yöntemiyle eğitiminin sonucunu, son satır ise hem oyun hem de sahte videoların eğitimde kullanıldığı durumun sonucunu ifade eder. İki veri seti karşılaştırılırken, UCF-ARG'nin daha düşük doğruluk oranına sahip olduğu gözlenmiştir. Bu durumun nedeni olarak; ayrımı zorlaştıran arka plan ve küçük aktör boyutları gösterilir.

Sadece zemin videolarını eğitmek yerine, oyun ve sahte anten örneklerinin doğru eklenmesinin doğruluk oranını arttırdığı ve farklı eylemler arasındaki karmaşıklığı azalttığı belirtilmektedir. Bahsedilmesi gereken bir diğer nokta, ayrık çok görevli öğrenmenin ince ayarını yapılmadan sadece 5 anten videosu ile hesaplanmıştır. Doğruluk değerleri Çizelge 4.3'de verilmektedir. Çalışma sonuçlarından elde edilen bir başka sonuç, ayrık çok görevli öğrenmenin ince ayardan daha iyi sonuçlar vermesidir. Son sütundaki tüm anten videolarını kullanarak ayrık çok görevli öğrenme anlamına gelir.

Çizelge 4.3: Farklı durumlarda elde edilen doğruluk sonuçları [8].

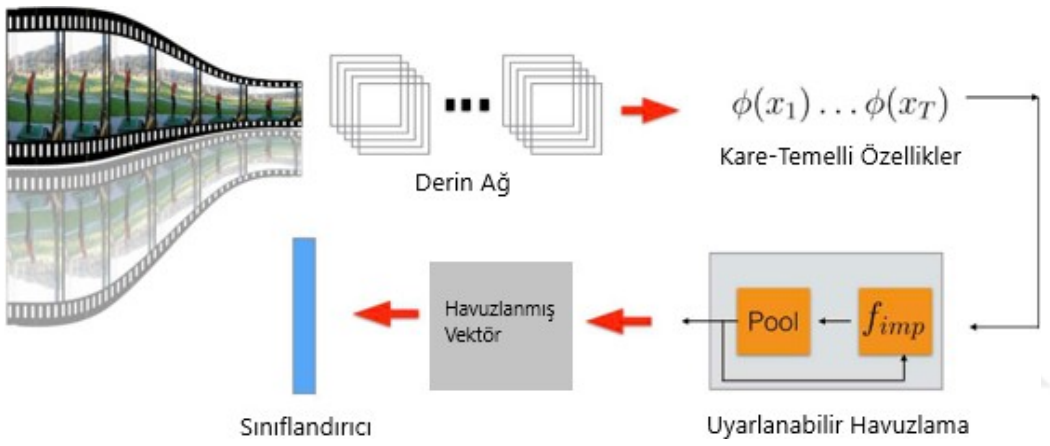
	5 Hava Videosu	Ayrık Çok Görevli Öğrenme (DML)	Tüm Hava Videoları
UCF-ARG	17.3	29.8	32.5
YouTube-Aerial	60.8	65.0	68.3

Sahte verilerin üretiminde oyun verilerinin önemini karşılaştırmak için, yeniden inşa kaybı olmadan oyunun uygulanmasıyla elde edilen sonuçlar, YouTube anten veri seti üzerinde verilmiştir. Yeniden inşa olmadan önceki değer 61.92 iken, yeniden inşa sonrası 63.33'e yükselmiştir. Burada dikkat edilmesi gereken nokta ağ, ayrık çok görevli öğrenmeyi kullanmadan yalnızca sahte anten videoları ile eğitilir.

4.1.5 AdaScan Yöntemi

İnternetteki videoların sayısının artmasıyla videoların anlambilimini anlamak gittikçe zorlaşmaktadır. AdaScan, insan aksiyonu tanıma işlemlerini farklı bir perspektiften gerçekleştirecek havuzlama yöntemini kullanarak bu amaç için geliştirilmiş bir yöntemdir [17].

AdaScan, uyarlanabilir havuzlama (Adaptive Pooling) adlı havuzlama modülü ile özelleştirilmiş bir evrişimli sinir ağıdır. Bir videoyu tarar ve toplanan son vektörü oluşturmak için seçilen karelerin özelliklerini dinamik olarak havuzlar.



Şekil 4.8: AdaScan yönteminin mimari diyagramı [17].

Model mimarisi Şekil 4.8'de yer alır. Model, 3 ardışık parçadan oluşur. 3 bölüm sırasıyla aşağıdaki hedeflere sahiptir. Birincisi, videonun her karesinden özelliklerin çıkarılmasından ve sonuçların $\emptyset(x_t) \in R^{)\#e+}$ vektöründe tutulmasından sorumlu olan özellik çıkarmadır. İkincisi, uyarlanabilir havuzlama, yalnızca son görev için farklı olan karelerden bilgi toplayarak ve geri kalan kareleri yok sayarak çerçevenin özelliklerini birleştirir. Üçüncüsü ise, uyarlanabilir havuzlama yöntemi, dinamik havuzlama işlemini yapar. Videoda geçici bir tarama gerçekleştirir, havuzdaki vektörü ve mevcut karenin özellik vektörünü dikkate alarak havuzlama işlemlerini gerçekleştirir. Her çerçeve için ayrımcı bir önem göz önünde bulundurulur. Video karesi ile pozitif ilintisi varsa, ayrımcı önem yüksek olarak kabul edilir ve video karesi ile negatif ilintisi varsa düşük olarak kabul edilir.

Bu tanım; Çoklu Örnek Öğrenme (MIL) tanımlı yöntemlerde ayrımcılık kavramına benzemektedir. Buna rağmen; MIL esaslı ağırlıklandırmanın “one-hot” vektör yönteminden farklı olarak, bir videoda birden fazla kareye odaklanabileceği belirtilmektedir. Son parça etiket tahminidir.

AdaScan uygulaması [73]'ten alınmıştır ve iki akışlı ağda 16 zamansal ve uzamsal VGG ağı bulunmaktadır. [64] 'deki çok ölçekli kırma tekniği eğitim verilerini arttırmak için kullanılır. Uyarlanabilir havuzlama katmanı bileşenlerine de [74] 'deki önerilerle atanmıştır. Uzamsal ağ, UCF-101 [9] eğitimi ile ImageNet [63] üzerinde eğitilmiş VGG-16 modeli [35] ile başlatılırken, UCF-101 [9] ile zamansal ağın eğitimi için kıvrımlı katmanlar 16000 yinelenen anlık görüntü ile [64] başlatılmıştır. HMDB-51 [11] veri setinin eğitimi için UCF-101 [9] veri setinin eğitiminden elde edilen evrişimli katman ağırlıkları hem uzamsal hem de zamansal ağ için kullanılmıştır.

Bu arada, uyarlanabilir havuzlama modülünün rastgele yeniden atanmasının UCF-101'deki [9] ağırlıkların kullanılmasından daha iyi sonuçlar verdiği gözlenmiştir. HMDB-51 [11] ataması ile daha da kötü sonuçlar elde edilmiştir. Buna rağmen; ImageNet'teki [63] modelin daha iyi performans gösterdiği belirtilmektedir. Bu durumun nedeni; video sınıflandırması için kullanılan karelerin, eylem sınıflarına eklenen alakasız kareler nedeniyle daha az genel özellik katmasıdır.

AdaScan, farklı yöntemlerle karşılaştırıldığında, uzamsal ağ için %79,1 doğruluk ve zamansal ağ için %81,7 doğruluk ile en iyi sonucu verdiği gözlemlenmiştir.

4.1.1 Kafes – LSTM ($L^2/01$) Yöntemi

Video setlerinden gerçekleştirilen eylem tanıma işlemleri, eylem modelleri ile öğrenilir. Bu süreçler kısa süreli hareketin modellenmesinde genellikle CNN ve RNN'ler, özellikle LSTM olarak görünür. Bu çalışmada, LSTM'nin uzun vadeli hareketlerini tahmin etmek amacıyla, ayrı lokasyon yerleri için öğrenilen bellek hücrelerinin bağımsız gizli durum geçişleri ile genişletilmiş, yöntem için Kafes - LSTM (L^2STM) seçilmiştir [18]. Burada uygulanan modeller ve algoritmalar başlıklar altında açıklanacaktır.

LSTM geliştirmeye karar verme durumunda, RNN ile ilgili bazı sorunlar vardır. Bu nedenle; RNN'de problemi yeniden tanımlamak önemlidir. Bu işlem, önceden türetilmiş 2 boyut filtreli giriş görüntülerinden bir özellik eşleminin çıkarılması açısından görüntü işlemeye benzer. 2 boyut filtrelerin öğrenilmesi de literatüre dahil edilmiştir [51]. Öğrenilen filtre boyutları genellikle 5x5'tir. Resimler yerine video serilerine uygulandığında kısa süreli hareket içeren eğitim videolarına uyarlanmış filtrelerle öğrenme işlemi gerçekleştirilir. Ancak; çoğu videoda uzun vadeli hareketleri karakterize etmek, çoğu video işleme ve uygulamada önemlidir. Uzun vadeli hareket modellerini doğrudan aynı filtreyle öğrenmek büyük işlemlere sebep olacaktır. Bu durum, modelin karmaşıklığını öğrenmenin zorluğu nedeniyle artırır. Bu karmaşıklık sorununu çözmek için kısa uzunluktaki hareket filtrelerinin öğrenme süreci uygulanır. Bu filtreler, sırasıyla her t (t zamanı temsil etmektedir) değeri için uygulanır. Her şeye rağmen; bu işlem RNN'lerde (LSTM'lerde) tam bir avantaj sağlamak için yetersizdir. Bu yüzden; RNN'ler (LSTM'ler) için video öğrenme yöntemini incelemek gerekir. RNN'ler, zamanla değişen durumları öğrenmek için gizli bir katman veya bellek hücresi kullanır. RNN sınırlarını daha iyi sunmak için, öncülerden elde edilen sonuçlar aynı olduğundan, LSTM'nin aksine RNN tanımı doğrusal olmayan bir şekilde uyarlanır.

Çalışma, karmaşık hareket modellerini karakterize etme kapasitesini artırmaya çalışmaktadır. Bu bağlamda, geleneksel özellik alanını yerel parçalara bölme ve her parça için haritalama yöntemi takip edilmektedir. Lokal hücrelere bölünme, uzamsal alandaki konumlarına göre gerçekleştirilir. L^2STM, önceki adımda bahsedilen sorunlara bir çözüm olarak geliştirilmiştir.

Hesaplamadan kaynaklanan sınırlamalar nedeniyle, sisteme tüm ardışık girdileri vermek ve eğitmek yerine videoları uçtan uca eğitmek için örnekleme gereklidir. Aşağıdaki resimde, farklı örnekleme yöntemlerinin aynı video dizilerinde farklı bölümler ürettiği gözlenmektedir ve bu örneklemeler Resim 4.4'te bulunur. Bu adımda, tekrar eden ağların uzun ve kısa sunumları daha mantıklı bir şekilde öğrenmesi için yeni bir örnekleme yöntemi geliştirilmiştir.

Resim 4.4: Farklı örnekleme metotlarından elde edilen bölümler [18].

Kafes-LSTM mimarisini daha açık bir şekilde anlatmak amacıyla sözlü olarak ifade edilmelidir. CNN'den özellik haritası kullanmak yerine, bellek hücresi, giriş / unut ve çıkış geçitlerini göstermek için sıra RGB ve optik akış görüntüleri kullanılmıştır.

Yük üst düşüm, LSTM'nin karmaşık hareket modellerini öğrenmek ve uzun süreli bağımlılık problemini geliştirmek için örü tekrarlayan bellek oluşturmak için hücre belleğine uygulanır.

RGB girişlerinin tahmin yeteneğini artırmak ve daha fazlasını telafi etmek amacıyla, optik akışın ek bir yöntem olarak beslendiği iki akışlı çerçeve benimsenir. Her videodan örneklenen karelerden oluşan klipler, her seferinde üst düzey özellik haritalarını çıkarmak için önceden belirlenmiş uzamsal ve zamansal ağlardan beslenir. Geleneksel iki akış çerçevesiyle karşılaştırıldığında, bu modül, bellek hücresi için tekrarlayan bir dikkat maskesi oluşturan paylaşılan giriş / unut ağ geçitlerini öğrenmek için RGB ve akış bilgilerini aynı anda besler. Burada dikkat edilmesi gereken nokta; RGB arasında yalnızca giriş / unut geçitleri paylaşılır. Diğer bileşenler bağımsız olarak öğrenilir.

Söz konusu çok modlu öğrenme prosedürü, giriş / unut geçitlerinin hem görünümden hem de dinamik bilgilerden faydalanmasını sağlar. Öğrenilmiş tekrarlanan giriş / unut geçitlerinin dikkat maskesi, bellek hücresinden gelişen dinamikleri düzenlerken, bellek hücresinin giriş ve çıkış dinamiklerini de kontrol eder. Giriş / unut geçitlerinin düzenli hale getirilmesinden sonra, çıkışın çok karmaşık dinamikleri yakalayan hareket özelliklerine sahip olması beklenmektedir.

Her bir bellek hücresi ve çıkış geçidi, her modaliteye göre özellikleri optimize etmek üzere bağımsız olarak öğrenilir. Son tahmin, RGB dizisi ve akışından elde edilen çıktıların ağırlıklı ortalamasıdır. Video sırası, en yüksek olasılığa sahip eylem kategorisine göre tanınır.

RGB ve optik akış görüntülerinin özelliklerini çıkarmak için, VGG-16'nın [35] 13 kıvrımlı katman içeren bir kısmı CNN olarak seçilir. Uzamsal ve zamansal ağlar ImageNet [63] ile önceden eğitilmiştir. Bununla birlikte; L²STM sıfırdan eğitilir. Optik akış ağının girişi yığılmış optik akış görüntüleridir. Uygulama, 8 kaydedilmemiş zaman adımını içerir. Tüm bu işlemler aşağıdaki adımlarla gerçekleştirilir; ilk CNN ağırlıkları sabittir ve sadece L²STM eğitilir. Doğrulama videolarındaki L²STM'nin doğruluğu arttığında, ağ daha düşük bir öğrenme hızına ayarlanır. Renk seğirme [23], yatay çevirme, veri kırpma için rastgele kırpma uzamsal ağ için kullanılır. Test sırasında daha fazla video klip beslenir ve uzun ya da kısa tahminleri birleştirmek için farklı adımlarla farklı uzunluk dizilerinden farklı olasılıklar elde edilir.

Videolardan klipler sisteme beslenir ve L²STM 'de farklı zaman adımlarının puanları alınır ve bu puanların finalinin ortalaması ile nihai puan alınır. Son olarak, bu yaklaşımın akış sırayla verildiğinde gerçek zamanlı videolarda kullanılmaya uygun olduğu belirtilmektedir.

Önerilen yaklaşım özellikle karmaşık hareketler içeren videolar için geliştirilmiştir. UCF-101 [9] veri seti; insan-insan etkileşimi, insan-nesne etkileşimi ve spor olarak 3 farklı başlık altında toplanmıştır. İnsan-insan etkileşimi için performans değeri %86,7, insan-nesne etkileşimi için performans değeri %95,4'tür ve vücut hareketi için performans değeri %88,6'dır. L²STM'nin karmaşık hareketlerde daha iyi sonuçlar verdiği belirtilmektedir.

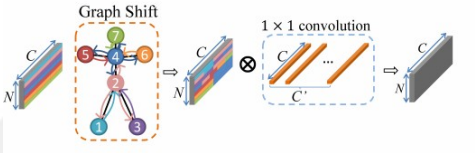
Öte yandan L²STM, LSTM benzeri mimarilere kıyaslanmıştır. Burada farklı verisetleri olmasına rağmen, L²STM diğer yöntemlere göre daha iyi sonuçlar verdiği gözlemlenmiştir. UCF-101 [9] veri kümesi için doğruluk değerleri %93,6 ve HMDB51 [11] veri kümesi için doğruluk değerleri %66,2'dir.

4.1.2 Kaydırma Çizge Evrişimli Ağı (Shift-GCN) Yöntemi

Giderek daha ilginç hale gelen iskelet verileri, eylem tanımada da önem kazanmıştır. Bu bağlamda tanıma işleminde kullanılan çizge evrişimli ağlar iskelet verilerle

gerçekleştirilir. Ancak, çizge evrişimli ağ yöntemleri genellikle karmaşık ve esnek olmayan çözümler sunar. Makale boyunca tartışılan Shift-GCN bu karmaşıklığı azaltmak için geliştirilmiştir [120]. Shift-GCN, Shift-CNN'den [114,115,116] esinlenerek geliştirilmiştir ve uzamsal kaydırma çizge evrişimi ve zamansal kaydırma çizge evrişimini içerir.

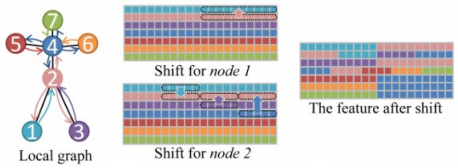
Uzamsal kaydırma çizge evrişimi, bir kaydırma çizge operasyonu ve noktasal evrişim içerir. Görselleştirmesi Şekil 4. 9'de verilmiştir. Kaydırma çizge işleminin ana fikri, özellikleri mevcut evrişimli düğümden komşu düğümlere kaydırmaktır. Makalede iki farklı kaydırma çizgesi çalışma yöntemi önerilmiştir. Bunlar yerel kaydırma çizge evrişimi ve yerel olmayan kaydırma çizge evrişimidir.



Şekil 4.9: Uzamsal kaydırma çizge evrişimi [120].

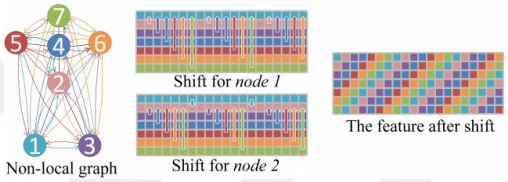
Yerel kaydırma çizge evrişimi için, daha önce iskelet veri kümelerinde tanımlanan insan vücudunun fiziksel yapısı ve alıcı alanları belirlenir. Bu yapıda vücut fiziksel grafiğinin komşu düğümleri arasında kaydırma çizge işlemi gerçekleştirilir. Gövde bağlantıları ve CNN özellikleri sıralanamaz, farklı sayıda farklı düğüm vardır.

Şekil 4. 10 yerel kaydırma çizge çalışmasını görselleştirmek için incelenebilir. Yöntemde bir düğüm, komşu numarası + 1 bölümüne bölünmüştür. Örneğin, düğüm 1'in 1 komşusu vardır ve 2 bölüme ayrılmıştır. Bu parçalardan biri düğüm 1'in özelliği iken 2. parça düğüm 2'ye kaydırılmıştır. Düğüm 2, 3 komşusu olduğu için 4 parçaya bölünmüştür. 1 parça 2. düğüm özelliğidir. Diğer kısımlar, düğüm 1, 3 ve 4'e kaydırılır. Kaydırılmış durumda, her düğüm, alıcı alanından bilgi taşır. Yerel kaydırma çizge evrişimi, yerel kaydırma çizge işlemini noktasal evrişim ile birleştirerek elde edilir.



Şekil 4.10: Yerel çizge uzaysal kaydırma işlemi [120].

Yerel olmayan kayma çizge evrişiminde, her düğümün alıcı alanlarının tüm çizgeyi kapsayacak şekilde yapılması işlemi gerçekleştirilir. Bu işleme yerel olmayan kaydırma çizgesi işlemi denir. Yerel olmayan kaydırmadan sonra, özellikler bir spiral şeklini alır. Bu sayede tüm düğümlerden bilgi akışı sağlanır. Görselleştirme Şekil 4. 11'a dahil edilmiştir. Lokal olmayan kaydırma çizgesi işlemini noktasal evrişim ile birleştirerek, yerel olmayan kaydırma çizge evrişimi elde edilir. Yerel olmayan kaydırma çizgesi evrişiminde, farklı düğümler arasındaki bağlantı aynıdır. Ancak insan iskeletindeki eklemler arasındaki önem farklıdır. Bu nedenle, uyarlanabilir yerel olmayan kaydırma mekanizması geliştirilmiştir. Bu yöntem, 3 bitişik matris kullanarak iskelet ilişkilerini modeller.



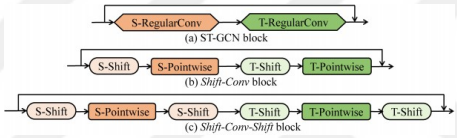
Şekil 4.11: Yerel olmayan çizge uzaysal kaydırma işlemi [120].

İskelet çerçeveleri uzamsal kaydırma çizgesi evrişimi ile modellendikten sonra, iskelet dizisini modellemek için zamansal kaydırma çizge evrişimi kullanılır ve iki farklı türü bulunur. Bunlar; saf zamansal kaydırmama çizge evrişimi ve uyarlamalı zamansal kaydırma çizge evrişimidir.

Saf zamansal kaydırma çizge evrişiminde, çizgenin zamansal kısmı, zamansal boyutun ardışık çerçevelerini birleştirerek elde edilir. Kanallar eşit olarak bölümlere ayrılmıştır ve her bölüm sırasıyla bir zamansal kaymaya sahiptir. Kaydırılan kanallar kesilir ve boş kalan kanallara 0 değeri atanır. Kaydırma işleminden sonra, her çerçeve komşu çerçeveden bilgi alır. Zamansal kaydırma çizgesi operasyonu ile zamansal nokta bazlı evrişim birleştirilerek saf bir zamansal kaydırma çizge evrişimi elde edilir. Saf zamansal kaydırma çizge evrişiminin hesaplama maliyeti, normal zamansal kaydırma çizge evrişiminden 9 kat daha azdır.

Saf zamansal kaydırma çizge evrişiminde, zamansal kayma mesafesi el ile atanır. Bu durum iki farklı dezavantaja neden olur. Birincisi, farklı katmanların video sınıflandırmasında ayrı zamansal alıcı alanlara ihtiyaç duymasıdır. İkincisi, farklı veri kümelerinin farklı zamansal alıcı alanlar gerektirmesidir [117]. Bu durum, saf zamansal kayma çizge evrişiminin genelleme yeteneğini sınırlar. Bu dezavantajların çözümü için uyarlanabilir zamansal kaydırma çizge evrişim çözümü önerilmiştir.

Bu çalışmada, Uzamsal-zamansal Kayma Çizge Evrişimsel Ağı, ST-GCN [118] ile aynı temel yapı kullanılarak oluşturulmuştur. ST-GCN, 1 giriş bloğuna ve 9 artık bloğa dayanır. Her blok düzenli bir uzamsal evrişim ve düzenli bir zamansal evrişim içerir. Düzenli uzaysal evrişim yerine, çalışmada geliştirilen uzamsal evrişim yerleştirilir ve düzenli zamansal evrişim yerine zamansal evrişim değiştirilir. Bu birleştirme işleminin de iki türü vardır. Bu yöntemlerin görselleştirmeleri Şekil 4. 12'de bulunmaktadır.



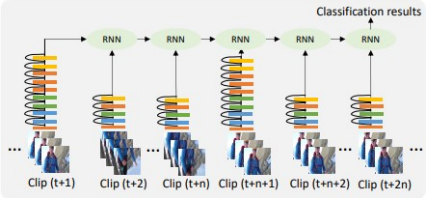
Şekil 4.12: İki tip birleştirme süreci [120].

Geliştirilen yöntemde performans sonuçları NTU RGB + D [61], NTU-120 RGB + D [61] ve Northwestern UCLA [106] veri setleri kullanılarak elde edilmiştir. Çalışma kapsamında karşılaştırılan çoğu yöntem, çoklu akış füzyon stratejilerini kullanır. Doğru bir karşılaştırma için [119]'daki 4 akışlı aynı akış füzyon stratejisi benimsenmiştir. 4. akıştaki "kemik hareketi akışı" girişi ile Shift-GCN, NTU RGB + D veri kümesinde %96,5, Northwestern UCLA veri kümesinde %94,6 ve NTU-120 RGB + D veri kümesinde %85,9 doğruluğa sahiptir. Bu değerler, karşılaştırılan yöntemlerin en başarılısıdır. Ek olarak, bu sonuçlar diğer yöntemlere kıyasla daha az hesaplama maliyeti ile elde edilmiştir.

4.1.3 FASTER Yöntemi

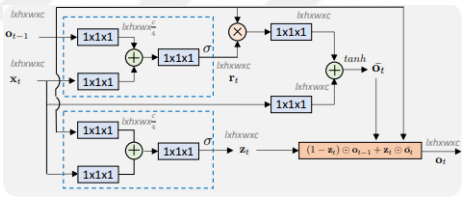
Standart video sınıflandırma süreçlerinde videolar küçük parçalara bölünür ve her klip bağımsız olarak değerlendirilir. Bununla birlikte, zamansal yapıdan bağımsız olarak benzer kliplerin işlenmesi, hesaplama maliyetini artıran faktörlerden biridir. Bu duruma bir çözüm olarak, FASTER [121] yöntemi geliştirilmiştir. Bu bağlamda, FAST-GRU adı verilen farklı temsillerin karışımını toplamak için tasarlanmış bir ağ

önerilmektedir. Çalışmada savunulan durum, birbirine yakın çerçevelerin benzerliği dikkate alındığında, bu çerçevelerin her birinin işlenmesinin fazlalığa neden olmasıdır. Şekil 4.13'de FASTER mimarisinin bir görüntüsü bulunmaktadır.



Şekil 4.13: FASTER mimarisi [121].

FASTER, verilen her kareyi işlemek yerine, eylemin ayrıntılarını içeren bir model ve zaman içinde değişen sahneleri yakalayan bir modelin kombinasyonundan oluşur. Tekrardan kaçınarak tüm videoyu düşük maliyetle kapsamayı amaçlar. FASTER çerçevesine ek olarak, FAST-GRU adlı bir RNN mimari tasarımı hazırlanmıştır. Bu ağ, farklı kliplerin modellerini bir araya getirmekten sorumludur. Ayrıca, FAST-GRU'nun diğer popüler RNN yapılarına göre daha uzun bir öğrenme süreci gerçekleştirdiği belirtilmektedir. FAST-GRU mimarisi Şekil 4. 14'de verilmiştir.



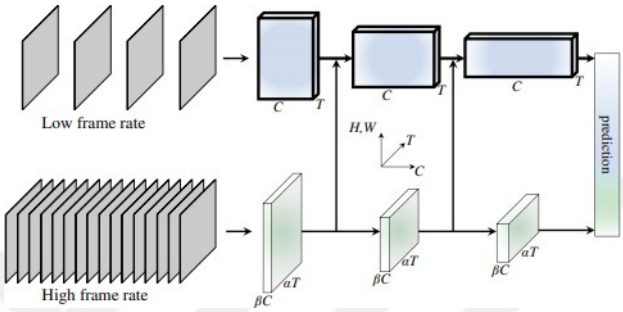
Şekil 4.14: FAST-GRU mimarisi [121].

Kinetics [108] veri seti FASTER'i test etmek için seçilmiştir. Kinetics'e ek olarak, UCF-101 [9] ve HMDB-51 [11] veri kümeleri için sonuçlar oluşturulmuştur. Kinetics kullanılan testlerde klip uzunluğu 8.16 ve 32 kare olarak seçilmiştir. En yüksek doğruluk değeri 32 karede %74,5 olarak elde edilmiştir.

Kinetics veri setinde karşılaştırılan son teknoloji yöntemler göz önüne alındığında, en yüksek doğruluk değerinin yine FASTER 32 karede olduğu görülmüştür. Aynı şekilde UCF-101 ve HMDB-51 ile yapılan karşılaştırmalarda sırasıyla %96,6 ve %75,7 doğruluk değerleri elde edilmiştir.

4.1.4 SlowFast Ağı Yöntemi

Videolarda eylem tanıma olarak iki tür bilgi vardır. Bu bilgiler, hızlı değişen ve yavaş değişen verilerden oluşur. Örneğin, atlama, koşma veya yürüme gibi eylemler hızla değişirken, eylemi gerçekleştiren kişi veya kişinin özellikleri yavaş değişir. Bu ayrıma dayanarak, SlowFast [125] modeli geliştirilmiştir. Bu model Şekil 4. 15'de yer almaktadır.



Şekil 4.15: SlowFast ağı mimarisi [125].

Modelin bir dalı, düşük kare hızlarında ve yavaş yenileme hızında anlamsal bilgileri yakalarken, diğer dal ise hızlı yenileme hızı ve yüksek zamansal çözünürlükle hızlı değişen eylemleri yakalar. Bu iki kol, yanal bağlantılarla birleştirilir. Hızlı dalda, zamansal havuzlamaya gerek yoktur. Çünkü tüm ara katmanlarda yüksek kare hızında çalışabilir ve zamansal devamlılık sağlanır. Yavaş dalda, uzamsal alan ve anlambilim üzerinde daha fazla odaklanma vardır.

SlowFast yöntemi, Kinetics-400 [122], Kinetics-600 [123], Charades [111] ve AVA [124] veri kümeleri ile hesaplanmıştır. Kinetics-400 ile elde edilen sonuçlarda karşılaştırılan modeller arasında en yüksek doğruluk değeri %79,8'dir. Kinetics-600'deki en iyi SlowFast konfigürasyonu ile doğruluk %81,8'e yükselmiştir. Charades'de %45,2'lik bir doğruluk gözlemlenirken AVA verileri ile yapılan test sonuçlarında ortalama kesinlik (mAP) değeri 28,2 olarak elde edilmiştir.

4.1.5 Zamansal Kesim Ağı Yöntemi

Evrişimli sinir ağları, video tabanlı eylem tanımada sıklıkla tercih edilen ve geliştirilen yöntemlerdir. Ne yazık ki bu çalışma, el yapımı özelliklere kıyasla video tabanlı eylem

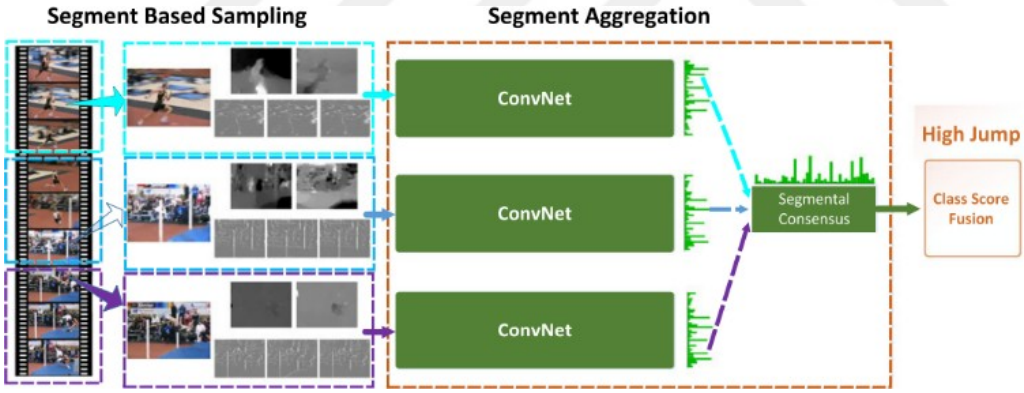
tanımada hala sınırlı bir gelişme olduğunu göstermektedir. Bu çalışma, evrişimli sinir ağlarında eylem tanıma ile ilgili 3 probleme odaklanmaktadır.

a. Uzun menzilli zamansal yapıları yakalayan video temsillerini etkili bir şekilde öğrenmek nasıl mümkün olabilir?

b. Kesilmemiş videolarda daha gerçekçi ayarlar için öğrenilmiş evrişimli sinir ağlarından nasıl yararlanılır?

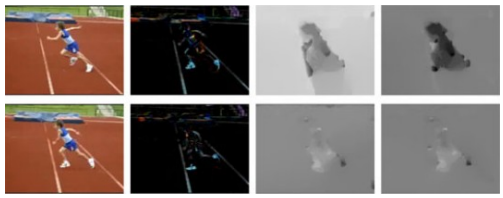
c. Sınırlı sayıda eğitim örneği verildiğinde, evrişimli sinir ağları büyük ölçekli verilere nasıl etkili bir şekilde uygulanabilir?

Zamansal Kesim Ağı, uzun menzilli zamansal yapıları yakalamak için geliştirilmiş ve Şekil 4. 16'de görselleştirilmiştir [127]. Buradaki varsayım, FASTER'e [121] benzer şekilde, ardışık çerçevelerin fazlalık olarak değerlendirileceği ve daha ayrık çerçeveler ele alınarak daha yüksek bir performans sonucunun elde edileceğidir.



Şekil 4.16: Zamansal kesim ağı mimarisi [127].

Diğer bir konu olan sınırlı eğitim vakası için çapraz modalite başlatma stratejisi geliştirilmiştir. Bu strateji ile öğrenilen temsiller, RGB modalitesinden optik akış ve RGB farkı gibi diğer modalitelere aktarılır. Ayrıca, 4 farklı girdi modeli kullanılarak deneysel çalışmalar gerçekleştirilmiştir. Bunlar; tek RGB görüntüsü, yığınlanmış RGB farkı, yığınlanmış optik akış alanı ve yığınlanmış çarpık optik akış. Girişler arasındaki farklar Resim 4. 5'tedir. Bu koşullar altında RGB ve RGB farkını birleştirmenin en iyi gerçek zamanlı eylem tanıma sistemi olduğu iddia edilmektedir.



Resim 4.5: Giriş modaliteleri sırasıyla RGB görüntüleri, RGB farkı, optik akış alanı (x, y yönleri) ve çarpık optik akış alanı (x, y yönleri) [127].

Model HMDB-51 [11], UCF-101 [9], THUMOS14 [55], ActivityNet [126] ve Kinetics400 [122], veri setlerinde uygulanmıştır. Birden fazla yöntemle karşılaştırıldığında doğruluk sonuçları sırasıyla %94,9, %80,1, %89,6 ve %75,7'dir.

4.2 Hareket Tabanlı Yaklaşımlar

Makine öğrenmesi ve bilgisayarla görü alanlarında, eylem tanıma popüler olmaya devam etmektedir. Bu makalede, insanların günlük yaşam aktiviteleri hakkında RGB-D videolarında etkili eylem tanıma yapacak bir mimari sunulmaktadır [15]. Bu yaklaşımda, videolardaki eylemlerin özelliklerinin gruplandırılmasının tanıma sürecinde son derece önemli bir faktör olduğu söylenmektedir.

İşlemlerin; hareketler, duruş açıları, fazla hareket içermeyen eylemler (örneğin, telefonda mesaj yazma) veya hareketi anlamak için bir zaman dilimi gerektiren eylemler gibi benzer özelliklerle nasıl gruplandırıldığı önemlidir. Bu nedenle; yapılacak eylemler 3 ana grupta incelenir.

a) Görünüm Modelleme: Evrişimli sinir ağından alınan uzamsal eylem durumlarının modelidir.

b) Kısa Dönemli Hareket: Optik akışla hesaplanan kısa vadeli hareketlerin bir grubudur.

c) Poz Tabanlı Hareket: İnsan vücudunun 3 boyutlu duruşuna dayanan poz ve zamansal gelişim ile uğraşan, tekrarlamalı bir sinir ağıdır.

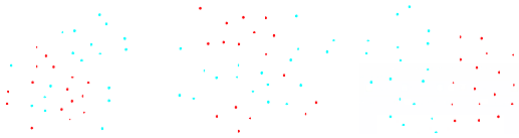
Bu çalışmada; görünüme dayalı özelliklerin önemini yüksektir. Her bir eylem için ortalama özellik sayıları elde edilmiş ve görünümün doğruluk oranları ile kısa vadeli hareket temelli sınıflamalar karşılaştırılmıştır [39]. Diğer yandan; kısa vadeli ve poz bazlı eylemler için tanınma doğruluğu karşılaştırılmıştır. HoG, LSTM kullanılarak elde edilen yoğun yörünge ile adil bir karşılaştırma yapmak için kullanılmamıştır.

Burada asıl önemli nokta, hangi görsel grupla çalışıldığından bağımsız olarak bir avantaj bulunmaktadır ve daha başarılı hale getirilebilir. Eğer eylemi karakterize eden özellikler için geç füzyon yöntemleri yerine erken füzyon tercih edilirse bu durum gerçekleşir. Bu bağlamda, eylemleri en uygun düzeyde birleştirmek için 2 aşamalı bir füzyon stratejisi geliştirilmiştir.

İlk füzyon (erken füzyon), özellikleri dengeli bir şekilde birleştirerek eylemlerin çoğunu özelliklerle karakterize etmeyi amaçlamaktadır. Geç füzyon ise belirli eylemleri belirli özelliklerle karakterize eden özellikleri seçmeyi amaçlamaktadır.

Erken füzyon için görünüm (F_1) ve kısa süreli hareket (F_2) genellikle yüksek korelasyona sahip oldukları için birleştirilir. Birleşik sürüm $F_x = [F_1, F_2]$ ile ifade edilir. Geç füzyon için poz temelli hareket daha önemlidir. Çünkü; önceki özellikler bu özelliği tamamlar. Pozlardan elde edilen zamansal bilgiler her eylem için farklı değildir. Bu nedenle, bu bilgileri erken füzyonda kullanmak sınıflandırıcıda gürültüye neden olur. Yazma veya telefonda konuşma gibi eylemlerde zamansal bilgiler önemli olmayabilir. Bir diğer deyişle, görünüş ve kısa vadeli hareket alanlarında ortak bir özellik olarak, zamansal vektörler, ayrımcılığa herhangi bir fayda sağlamaz. Bu nedenle, füzyonun modalitelerin bireysel gücüne odaklandığı geç füzyon stratejisini kullanarak, poz tabanlı hareketi kaynaştırır (F_3).

Makalede kullanılan veri setleri CAD- 60 [59], CAD- 120 [59], MSRDailyActivity3D [60] ve NTURGB + D [61]'dir. Nitel sonuçlar t-SNE [65] aracı yardımıyla grafiklendirilir. Şekil 4.17'de kısa süreli, görünüş tabanlı, kısa süreli ve görünüş tabanlı yöntemlerinin sırasıyla "içme" ve "oturma" eylemleri için birlikte kullanıldığı durum gösterilmektedir. Birlikte kullanılan kısa süreli ve görünüş tabanlı yöntemlerin burada gözlemlenmesi, daha net bir ayrım sağlar.



(a) (b) (c)

Şekil 4.17: (a) kısa vadeyi, (b) görünüşü ve (c) (a) ve (b) [15] 'in kombinasyonunu belirtir.

Görünüm, kısa süreli ve poz tabanlı hareket sonuçları CAD-60 [59], CAD-120 [59] ve MSRDailyActivity3D [60] veri kümeleri ile verilmiş ve CAD- 60 veri kümesi performans değeri %98,53, CAD- 120 veri kümesi için performans değeri %87,90'dır ve önerilen füzyonda MSRDailyActivity3D veri kümesi performans değeri %97,81'dir. Yazarlar tarafından yöntemlerden elde edilen sonuçların veri setlerine bağlı olduğu gözlenmiştir.

Ek olarak, her veri kümesinin başarılı olduğu yöntemler görülebilir. CAD-60 veri kümesi görünüşte daha başarılı, CAD-120 kısa vadede daha başarılı ve MSRDailyActivity3D veri kümesi poz tabanlı hareket yöntemlerinde daha başarılı olduğu gözlemlenmiştir. Bununla birlikte, iki seviyeli füzyonun tüm özelliklerde avantajlı olduğu belirtilmiştir.

Bir diğer yandan, benzer ve ayrıştırılması zor görünen eylemler için, eylem çifti modülünün eylem ve başarı yöntemi incelenmiştir.

Çizelge 4.4'te, tüm veri kümeleri için ikili sınıflandırıcı kullanma konusundaki doğruluk değerlerini göstermektedir. İşlem çifti modülü, doğrusal bir SVM olan ikili sınıflandırıcı tarafından ayrı olarak sınıflandırılan ve karıştırılan eylemleri izler. CAD -120 için, IDT + FV (görünüm boyunca kısa süreli hareket), karışık eylem çiftlerini %100 doğrulukla ayrıştırır. Bu modülün dezavantajı, çapraz doğrulama kümesine bağlı olmasıdır. Bu işlem çifti modülünün CAD-60 ve MSRDailyActivity3D veri kümelerini etkilemediği belirtilir.

Makalenin temel aldığı bazı benzer çalışmalarla, aynı veri kümeleri üzerinde karşılaştırmalar yapılmıştır. Bu bağlamda doğruluk değerleri CAD-60 için %98,52, CAD-120 için %94,40, MSRDailyActivity3D için %97,81 ve NTURGB + D için %87,09 olarak ifade edilmiştir.

Çizelge 4.4: Tüm veri setleri ile ikili sınıflandırıcı için doğruluk değerleri [15].

Veri Seti	İkili Sınıflandırma Öncesi Doğruluk (%)	İkili Sınıflandırma Sonrası Doğruluk (%)
CAD-60	%98,52	%98,52
CAD-120	%87,90	%94,40
MSR3D	%97,81	%97,81
NTURGB+D	%84,95	%87.09

4.3 Çoklu Örnek Öğrenme (MIL) Tabanlı Yaklaşımlar

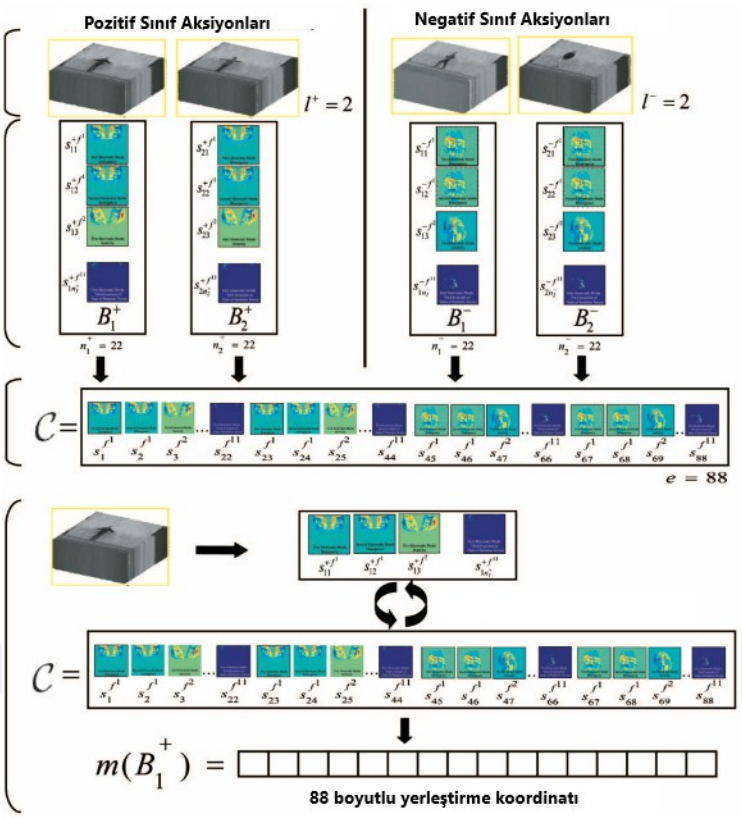
Bu çalışma [19], her örneğin "örnekler" olarak adlandırılan bazı özellik vektörleri ile temsil edildiği Çoklu Örnek Öğrenme yöntemini kullanarak öğrenme gerçekleştirir. Yöntemde kinematik mod tabanlı temsil işlemleri gerçekleştirilir. Kinematik mod ise, eylemi temsil eden bir örnek olarak tanımlanır. Buradaki yöntemde, her eylem kinematik modları temsil edecek şekilde dönüştürülür. Bu yüzden; her video kinematik modların bir koleksiyonu olarak temsil edilir. Videonun işlem etiketi, aynı zamanda koleksiyonda bir etiket olarak seçilir. Bu durumda amaç, eylemin kinematik mod tabanlı temsilini öğrenerek koleksiyonun eylemi temsil edip etmediğini öğrenmektir. Bu işlem kinematik modda veya koleksiyonları örnek tabanlı özellik alanına yerleştirerek gerçekleştirilir ve bu alandaki koleksiyonun (torbanın) koordinatlarını sınıflandırma için kullanarak yapılır. Yerleştirme sürecindeki ana fikir, eğitim setindeki her kinematik mod, çantayı temsil eden bir özellik veya özellik olabilir. Bu

işlemler [40]'dan ilham alınmıştır ve aşağıdaki matematiksel ifade ile açıklanmaktadır. Ayrıca mimarinin diyagramı Şekil 4.18'de verilmiştir.

Diyelim ki $B^{\blacksquare}_{\&} = \{s^{\blacksquare 0}_{\&1}{}^{1}, s^{\blacksquare 0}_{\&2}{}^{1}, s^{\blacksquare 0}_{\&2}{}^{2}, \ldots, s^{\blacksquare 0}_{\&1}{}^{k}, \ldots\}$ i. pozitif çanta olsun ve $s^{\blacksquare 0}_{\&1}{}^{k}$ ise $B_{\&}$'deki j. kinematik modu temsil etsin. A^k k=(1,…,11) olduğu yerlerde kinematik modları türeten kinematik özellik olarak kullanılsın ve $n^{\blacksquare}_{\&}$ is$e\, B^{\blacksquare}_{\&}$'deki toplam kinematik mod sayısı olsun. $n^{\blacksquare}_{\&}$ ayrıca $B^{\blacksquare}_{\&}$'ya karşılık gelen tüm kinematik özellikler boyunca korunan toplam mod sayısına (özvektörler) eşittir. Benzer şekilde , $B^{\$}_{\&} = \{s^{\$0}_{\&1}{}^{k}, s^{\$0}_{\&2}{}^{k}, \ldots\}$ i. negatif torbayı temsil eder. Toplam kinematik modların sayısı da $n^{\$}_{\&}$'dir. Ayrıca, eğitim setindeki pozitif torbalar $l^{\blacksquare}$, negatif torbalar $l^{\$}$ olarak ifade edilir.

Önceki paragraflarda belirtildiği gibi, eğitim setindeki her kinematik mod, torbanın örnek tabanlı temsilini türetmek için bir özellik veya öznitelik gibi davranır. Bu nedenle, tüm kinematik modlar (tüm torbalarda) bir C setine hizalanır ve tekrar endekslenir. Bu indekslemede kinematik modlar $s^{0}_{e}{}^{k}$ ile ifade edilir, burada e = $\{1,\ldots,(\sum_{\&41}^{l^{\%}} n^{\blacksquare}_{\&} + \sum_{\&41}^{l^{\&}} n^{\$}_{\&})\}$

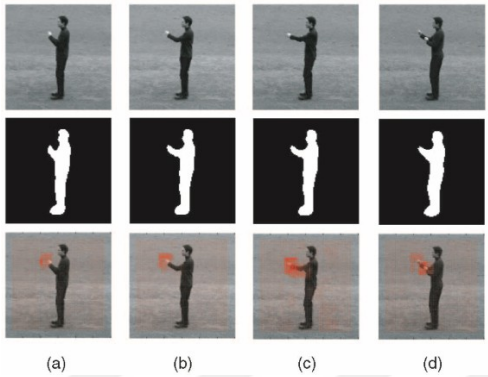
Her eylem videosu bir e-boyutlu vektöre eşlenir. m ($B_{\&}$) hesaplaması Şekil 4.18'de yer almaktadır. Son olarak; En yakın komşu sınıflandırıcı, farklı eylemlerin torbasının m ($B_{\&}$) örnek uzay koordinatları tarafından öğrenilir.



Şekil 4.18: Yaklaşımın mimari diyagramı [19].

Makalede tartışılan çalışmalar iki farklı veri seti üzerinde gerçekleştirilmiştir. İlk olarak, Weizmann veri setinde [10] operasyonlar gerçekleştirilmiştir. Weizmann veri kümesinde 90 video bulunmaktadır, ancak her parça bir eylem döngüsü içerecek şekilde videoların bölümlere ayrılmasıyla eylem sayısı elde edilmiştir. Döngü uzunluğu eyleme bağlı olarak seçilmiş ve net olmayan aktörler içeren video parçaları atılmıştır. Bu şekilde bir ayrım yapılarak toplam 180 video elde edilmiştir. Daha sonra, video arka planlarını kaldırmak için 100x100 piksel sınırlayıcı kutularoluşturulmuştur. Sınırlayıcı kutuları toplamanın amacı, eylemin uzay-zaman hacminiüretmektir ve daha sonra bu kutular optik akışları hesaplamak için kullanılmıştır. Optikakışlar kinematik özelliklerin ve baskın kinematik modların hesaplanmasında rol oynamaktadır. İlk olarak, eğitim setindeki baskın kinematik modlar hesaplanır. Öte yandan, eğitim videolarındaki koordinatların en yakın komşu sınıflandırıcıyı öğrenmek için kullanıldığı belirtilmektedir.

İkinci olarak, işlemler KTH veri kümesinde [6] gerçekleştirilmiştir. Algoritma öğrenme ve tanıma için kişi lokalizasyonu gerektirdiğinden, kişi kesimleme işlemi [71]'de önerilen yöntemle gerçekleştirilir. Kesimleme, “level-set-based frameworks” ile bölgenin konturunun arka planını en aza indirme işlemidir. Her çerçeve için kenarlık boyunca konturlar yapılır. Resim 4. 6 bu işlemi göstermektedir.



Resim 4.6: İlk satır giriş dizisini temsil eder. İkinci sıra, segmentasyondan sonraki kareler ve üçüncü sıra ise kişinin bulunduğu bir bölge olan optik akış alanlarıdır [19].

Aktör yerelleştirildikten sonra, çevresine bir sınırlayıcı kutu eklenir ve optik akış hesaplanır. Hesaplanan akış alanı, önceden belirlenmiş 70x70 piksel boyutuyla senkronize edilir. Daha sonra uzay-zaman hacmi oluşturmak için biriktirilir. Uzay-zaman hacmi kinematik modu ve özelliği bulmak için kullanılır. Girdi verileri [72]'de eğitim ve test ayrımı şeklinde gerçekleştirilir ve her eylem için tek bir döngü kullanılır. Her deney için, kinematik mod tabanlı özellik alanı, eğitim örneklerinin kinematik modları ile oluşturulmuştur.

Test, “leave-one-out” ayarlarıyla gerçekleştirilmiştir. Her özellik için 1 kinematik mod kullanılırken, 10 eylem için ortalama %80,3 doğruluk elde edilirken, kinematik mod sayısındaki artış da doğruluk artışında etkili olmuştur.

Farklı çalışma durumlarının bir sonucu olarak, her özellik için 4 kinematik mod kullanılarak en iyi durumun elde edildiği belirtilmektedir. Karmaşık yapıların, farklı eylemler arasında ayrım yapmak için daha fazla temsil sağlayarak algoritma başarısını arttırdığı gösterilmiştir. Yöntemin ayrımcılıkta güçlük çektiği eylemler, benzerliğin yüksek olduğu eylemlerdir. Örneğin; "run" ve "skip" eylemleri genellikle birbirinden ayrılmaz. Bunun sebebi, her iki eylemde de aktörlerin hız ve bacak hareketlerinin benzer olduğu belirtilmiştir.

Kinematik özellikler yönteminin, geleneksel optik akış yönteminden daha iyi çalıştığı gözlenmiştir. Kinematik özellikler yönteminin, optik mod yöntemine göre 1 ila 5 mod arasında değişen tüm sonuçlarda daha yüksek doğruluk değerleri ürettiği gözlenmiştir.

Yöntemin işleyişine ek olarak, ölçek durumları değiştiğinde yöntemin yanıtını da ölçülmüştür. Bu nedenle, ölçek değişikliğinin yöntem üzerindeki etkisini görmek için 100x100 sınırlama kutuları 50x50 piksel olarak güncellenmiş ve pencere boyutu 8x8 piksel olduğunda optik akış blok tabanlı korelasyon ile yeniden hesaplanmıştır.

Aynı test yöntemleri kullanılarak, optik akış sonraki kinematik özellikler için 100x100'e yeniden boyutlandırılmış ve sonuçlarda bir düşüş gözlemlenmiştir (%91,3 ortalama doğruluk). Bunun nedeni, önemli vücut parçalarının kaybı olarak belirtilmektedir. Öte yandan, sınırlayıcı kutu 200x200'e yükseltildiğinde ve optik akış 32x32 piksel değerinde boyutlandırıldığında, sonuçlar ortalama %95,2 doğruluk değerine sahip orijinal duruma benzer. Bunun nedeni, kutunun boyutunun, belirli bir değerden sonra vücut parçalarının detaylarını elde etme üzerinde hiçbir etkisi olmamasıdır.

Öte yandan, KTH veri seti [6] kullanılarak yapılan test işlemlerinde, en çok karıştırılan eylemlerin "koşu yapma" ve "koşma" olduğu belirlenmiştir.

Baskın dinamikler incelendiğinde, her özelliğin hızındaki küçük değişiklikler PCA tarafından dikkate alınmaz. Algoritmanın performansını etkileyen önemli bir faktör optik akışın kalitesi olarak ifade edilir. KTH veri kümesindeki kareler arasında bulanıklaşma, akışın hesaplanmasında zorluk yaratır.

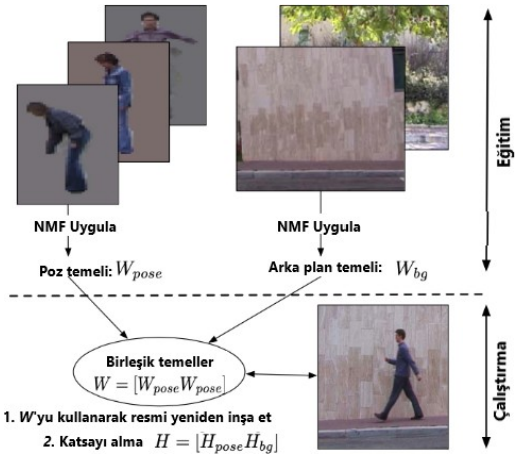
Son olarak, performans karşılaştırması KTH veri seti kullanan diğer yöntemlerle yapılmıştır. Diğer yöntemlerle karşılaştırıldığında ortalama doğruluk değeri %87,7 olarak bulunmuştur.

Sonuç olarak, daha fazla özellik genel performans artışına yardımcı olmaktadır. Tüm kinematik özellikler kullanılarak en iyi sonuç %95,75'tir. Bunun, her bir özellik için tamamlayıcı bilgi sağlamak ve daha ayrımcı sonuçlar üretmek için yararlı olduğu söylenmektedir. Son olarak bu durumun, birbirleriyle karışan eylemlerin ayırt edilmesinde de faydalı olduğu belirtilmektedir.

4.4 Sözlük Tabanlı Yaklaşımlar

4.4.1 Poz Ilkeli Tanıma Yöntemi

Bu çalışma, tamamen poz tabanlı bir eylem tanıma yöntemi sunmaktadır. Poz temsili pozlamada önemlidir [20]. Poz eşleştirmedeki en zor kısım, karmaşık arka plan ve pozun birleşimidir. Bunun sebebi, arka plan genellikle vücudun bir parçası olarak kabul edilmesidir. [41] ve [42] 'den esinlenerek poz tanıma, poz ilkeleri kümesiyle eşleştirilerek gerçekleştirilir. Pozları tanımlamak için HOG tanımlayıcıları kullanılır. Standart HOG, NMF tabanlı gradyan histogramları ile genişleyerek bahsedilen arka plan ve eklem karmaşasının giderilmesi amaçlanmaktadır. NMF tabanlı poz gösterimi, negatif olmayan matris çarpanlarına ayırma ile temiz bir arka plan kullanan bir dizi uygulama yoluyla öğrenilir. Negatif olmayan bir veri matrisi V'ye NMF uygulanması, hem W hem de H'nin negatif olmamasıyla sınırlanan bir V ≈ WH faktörüne yol açar. W belirli bir baz vektörünü ve H katsayıları temsil eder. Sonuç olarak; V, W ve H kullanılarak yeniden oluşturulur. Resim 4.7'de işlemlerin görselleştirilmiş hali bulunmaktadır.



Resim 4.7: Poz İlkeli Eylem Tanıma için NMF uygulaması [20].

PCA ve benzer tekniklerin aksine, NMF sadece gözler, kulaklar ve burun gibi parçaları tespit edebilir. Ancak; parça bazlı sabitleme her zaman yapılamaz. NMF'nin bu özelliği mevcut çalışmaya dahil edilmemiştir. Yaklaşık çarpanlarına ayırma V ≈ WH değerini bulmak için çarpımsal güncelleme kuralı [43] kullanılır. Burada daha fazla poz türü bulmaya çalışılmaktadır. Geçişli görüntü yalnızca poz tabanından (G_{pose}) yeniden yapılandırılırsa, arka planın istenmeyen bir durum olarak yeniden yapılandırıldığı gözlenir. Bu nedenle; 2 modifikasyon işlemi uygulanır. Her şeyden önce, poz içeren tüm görüntü (W) alınır ve arka plan görüntüsünden hesaplanan G_{b9} eklenir. Böylece V = [$G_{pose}G_{b9}$][$H_{pose}H_{b9}$] olarak ifade edilir.

Egzersiz sırasında poz ve arka plan bağımsız olarak öğrenilir. G_{pose}, resimdeki pozun anlamlı kısımlarını ifade eder. Poz eşleştirme ve aktivite tanıma için 40 ila 80 baz vektörün yeterli olduğu belirtilmektedir. Yeni V^{new} görüntüsünden poz hesaplaması için, ilk adımda H^{new} W = [$\mathbf{G}_{pose}\mathbf{G}_{b9}$]. 'ye göre hesaplanır. Bu nedenle; W'yi sabit tutmak için standart bir yinelemeli algoritma kullanılır. Böylece, ağırlıklar birleştirilir, V^{new}'ın en iyi açıklaması $\mathbf{G}_{b9}$ ve $\mathbf{G}_{pose}$ kullanımı altında yapılır. NMF'ye sıkı ağırlık eklenmesi nedeniyle, bazı tanımlayıcı parçalar için genellikle G_{b9} veya G_{pose} ağırlıkları arka plan ağırlıkları kullanılır. Poz ve arka plan ağırlıkları ayrı katsayılar olarak üretilir ve $H^{new} = [H^{new}_{b9}\ H^{new}_{pose}]$ olarak ifade edilir. Arka plan etkili bir şekilde NMF temelli yeniden oluşturma ile ön plandan ayrıştırılır. Az sayıda eğitim örneği ve yetersiz sayıda baz nedeniyle, G_{b9} doğru modelleme sonuçları üretemez. Buna rağmen; G_{b9} poz tabanlarının yanlış birleşmesini azaltır. G_{b9} Yalnızca arka plan görünümlerinin yaklaşık olarak modellenmesini sağlarken, G_{pose} pozları doğru şekilde modellemeye dahil olur.

Ortaya çıkan H_{pose} katsayıları örnek pozlar veya poz ilkelleri ile karşılaştırılabilir. Poz ilkelleri, eğitim dizisi setinden ve karşılık gelen $H_{pose}'dan$ elde edilir. Öklid mesafesi ve "Standard Agglomerative Clustering" [44] kümeleme için kullanılır. Genel olarak, 30 ila 80 poz ilkel doğru eylem tanıma sonuçları vermek için yeterlidir.En sonunda, poz ilkel için minimum Öklid mesafeli bir indeks alınır, bu da benzer eğitim pozları kullanılarak yeni bir resim oluşturulduğu anlamına gelir.

Poz tabanlı eylem tahmini yapmanın yanı sıra, makale ayrıca insan tespiti de içerir. Standart kayar pencere yaklaşımı için insan algılama prosedürleri uygulanır. Ölçek değişiklikleriyle başa çıkmak için, detektör pencere boyutu ve bağımsız detektör sonuçları değiştirilir. NMF plakaları [$G_{pose}G_{b9}$], gradyan boyama için 2 alternatif yöntem sunar. [45] 'e benzer şekilde, insan tespit işlemleri de incelenmiştir.

Çalışma kapsamında yapılan eylem tanımadan esinlenilerek [46], bu referans histogram temelli bir yaklaşımı temsil etmektedir. Çalışmanın aksine, ilham verici referans daha ilkel eylemlere odaklanmıştır ve önceden tanımlanmış olaylar yerine kümelenmiş pozlar ele alınmıştır. Temel yaklaşım poz ilkellerini sıralayarak karmaşık eylemler oluşturmaktır. Ayrıca, n-gram ile ilkel pozun zamansal içeriğinin alt sırasını ve poz başına verilen bilgi derecesini gerçekleştirmenin önemli olduğu belirtilmektedir.

Eylem tanıma Poz Histogram Sınıflandırması, Poz İlkel Ağırlıkları ve Lokal Zamansal Bağlam adımlarında gerçekleştirilir.

Poz Histogram Sınıflandırma adımında, poz ilkellerini doğrudan sırayla analiz etmek yerine, odak özelleştirilmiş pozlar oluşturma durumudur. Sınıflandırılması histogram karşılaştırması ile yapılır. Burada her histogram, bir özne tarafından gerçekleştirilen belirli bir aktiviteye karşılık gelir. Sınırlı egzersiz histogram kaynağı nedeniyle, 1-NN sınıflandırıcısı kullanılmıştır. Histogram toplama [47], gelecekteki araştırmalar ve daha büyük veriler için de kullanılabilir. Histogram karşılaştırması için KL ıraksaklık sorgusu histogramın bölünmesini cezalandırmaz [47]. Bu durum, değişken uzunluktaki poz ilkelerinin tanınması için önemlidir. Burada her histogram bölmesi, poza karşılık gelen tam bir görüntünün alanına karşılık gelir. İlginç bir şekilde, KL-ıraksama hareketsiz görüntüleri kullanarak hareket sınıflarını sezgisel olarak tanımak için yararlıdır. Bu çerçeve hem görüntü dizisinde hem de hareketsiz görüntülerde etkinlik tahmini için kullanılabilir.

Poz ilkel ağırlıklandırmalar adımında, belirli pozlar davranış hakkında diğer tüm pozlardan daha fazla bilgi verebilir. Örneğin; yalnızca ayakta duran bir kişiden alınacak sonucun bir zımbanın hareketinden daha fazla olasılığı vardır. Poz ilkelleri tam bir davranışın bir parçası olduğu için, yöntemi uygulamanın [48] 'de bir yüz sınıflandırmasından daha kolay olduğu belirtilmektedir.

Lokal Zamansal Bağlam adımında, eylemleri yalnızca içerikle tanımlamak doğru değildir. Olayları içeren ardışık pozlar, aktivitenin tanımlanmasında büyük rol oynar. Bunun sebebi; sıralanmamış pozlara dayalı etkinlik tahmini yanlış sonuçlara yönlendirilebilir. Bu bağlamda yazarlar tarafından yerel zamansal bağlamın dahil edilmesinin yararlı olacağı incelenmiştir. Bu bağlamda n-gram ifadelerle takip yapılır [46]. n-gram n uzunluğunun bir alt dizisini sağlar.

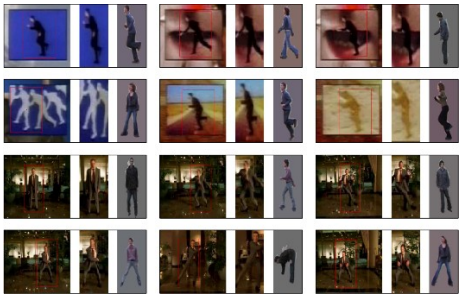
Bu çalışmada, sadece eğitim sırasında gözlenen alt diziler tartışılmıştır. Bu nedenle; gerçek sayı çok daha düşük olma eğilimindedir. Poz ilkel dizilerini n-gram dizilere dönüştürmek için değişiklik gerekmez. Poz ilkellerinin histogramını hesaplamak yerine, n-gram numunelerinin histogramları hesaplanır. Temel ilkel poz, her kare için bir tanımlayıcı ile aynı kalır. En iyi sonuçlar alt sekanslama uzunluğuna göre 2 ve 3-gram olarak elde edilir.

Önerilen yaklaşımda eylem tanıma, yöntemde kullanılan farklı uygulama türleri için sonuçlanır. En iyi sonuç 30 poz ilkel sayısı ile elde edilmiştir. Buradaki sınıflandırma, her bir alt dizi için tanınan eylem sınıflarına uygulanan çoğunluk oylama şemasına dayanmaktadır. En kötü eylem tanıma sonucunun özellikle "tek bir yere atlamak" (jumping in one place) eyleminde gözlendiği belirtilmiştir.

Diğer yandan, Weizmann veri seti ile elde edilen hareketsiz görüntülerin sınıflandırma sonuçları ise, veri seti için en iyi ortalama hassasiyet sonucu, 110 poz ilkeli ile %70,4'tür.

Buna ek olarak, yöntem Weizmann aksiyon tanıma karşılaştırma seti ile yapılan son yaklaşımlarla karşılaştırıldığında [10], bu yöntem %94,40 olan en yüksek doğruluk değerine sahiptir ve hareketsiz görüntüler için değer %70,4'tür.

Son olarak, yaklaşımın kısıtlamalarını ele almak ve sonuçlara ulaşmak için sadece poz eşleştirme süreci odaklanılmaktadır. Eğitim için Weizmann veri seti kullanılmıştır. Test için iki farklı şarkının klipleri kullanılmıştır. Bu klipler; "A Road to Nowhere" ve "Weapon of Choice" dur. Her iki klipe de insan tespiti uygulanmıştır. "A Road to Nowhere" için orijinal resim boyutu nedeniyle, bir alan insanın değişen arka planıyla kesilmiştir. Test setindeki pozlar Weizmann ile tam olarak eşleşmemektedir. Ayrıca; arka planların NMF'nin arka plan tabanlarından tamamen farklı olduğu belirtilmektedir. Sonuçlar Resim 4.8'de yer almaktadır.



Resim 4.8: Müzik video klipleri için eşleşen sonuçlar [20].

4.4.2 Sınıf Kaynaklı & Sınıf Bağımsız Teklif Öğrenme

Büyük ölçekli video analiz süreçlerini gerçekleştiren yöntemlerin çoğu, insan eyleminin tanınması için kısa vadeli videolarla ilgilidir. Kısa vadeli videolarda bile büyük ölçekli videoları işlemek ve işlemek hala zor bir konudur. Bu makale, gelişmemiş videolarda eylemlerle zamansal segmentler bulmaya izin veren bir yöntemi açıklamaktadır [12].

Büyük ölçekli ve pratik senaryolarda, bir eylem teklif yöntemi 2 farklı durumla şekillenir. İlk olarak, geçici yöntemlerin tanımlanması, kodlanması ve puanlanmasında teklif yöntemi etkili olmalıdır. İkinci olarak, teklif yöntemi ilgili faaliyetleri ayrıştırmalı, diğer bir deyişle, bu etkinlik sınıflarını gösteren görsel bilgiler içeren geçici bölümler getirmelidir.

İçerik tabanlı herhangi bir bilgi olmadan uzun bir videoyu örneklemek son derece hızlıdır. Ancak; bu strateji genellikle istenen eylem sınıfına sahip parçaları getirmede başarılı olmaz. Öte yandan, literatürdeki en başarılı eylem tanıma yöntemleri uygulanabilir değildir.

Kesilmiş eylem tanıma yöntemlerinin çoğu için; yoğun yörünge çıkarma ve Fisher vektörü ile kodlama bir standart haline gelmiştir. Fakat; bu yöntemler büyük ölçekli videolarda etkinlik önerileri almak için son derece yavaş çalışır.

Bu çalışmada, yukarıda belirtilen teklif yöntemlerini şekillendiren 2 farklı durum arasında ayarlanabilir bir dengeye sahip başarılı bir yöntem geliştirilmiştir.

Bu yöntem, verilen eğitim seti videolarından uzamsal ve zamansal görünümü yakalayan özellikleri çıkarır. Etkinlik sınıfı kümesi için, ayrıştırıcı bilgilerini kodlayan evrensel bir sözlük öğrenilir ve oluşturulur. Sözlük, sözlük alındıktan sonra görünmeyen video dizilerindeki geçici parçaları kodlamak için kullanılır. Bu işlemi gerçekleştirmek için çok sayıda aday zamansal segment üretilir. En son zamansal eylem önerilerini elde etmek için, bu segmentler sözlük tarafından ne kadar iyi temsil edildiklerine göre sıralanır. Aralarında en iyi temsil edilen geri alınır.

Yöntem adayların ilk seti ile başlar. Adaylar, alınan teklifler seçilerek oluşturulur. Aday teklifler oluşturmak için farklı uzunluktaki zamansal segmentler giriş video dizisinden örneklenir. Bu örnekleme zaman içinde eşit olarak gerçekleştirilir, ancak zamansal olarak yerel eylem sınıfları içeren bir eğitim setinden derlenen bir dağılımdan bir örnek alınır. Bu işlemi yaparken; önceden tanımlanacak ve girişvideosu zamanında çakışacak birkaç teklif adayına ayrılmıştır.

Özellikleri ayıklamak için, hesaplama verimliliği ve ayrıştırma gücü bir değiş tokuş olduğu için STIP'ler [26] kullanılmıştır. Böylece, özellik yukarıda belirtilen teklif adaylarından kaldırılmıştır. Uzamsal ve zamansal görünümü karakterize etmek için, her STIP noktası HoG ve HoF kullanılarak kodlanır. STIP'ler daha önce farklı eylem tanıma çalışmaları için kullanılmıştır [6]. Ayrıca, aday seti paralel hale getirilerek çıkarma ve temsil işlemleri hızlandırılmıştır.

Teklifleri temsil etmeyi öğrenmek için, [28] ve [29] 'dan esinlenilen bu adımlarda, x teklif adayındaki her STIP özelliği seyrek bir sözlük seti kullanılarak doğrusal olarak temsil edilebilir. Bu asıl adaylar, açıklamalı etkinlik örneklerine sahip geniş video setinden elde edilir. Ek olarak; aynı asıl adaydaki farklı STIP özelliklerinin bağımsız değil ortak noktalar olması beklenir. Böylece iki farklı tasarım yöntemi ortaya çıkmaktadır.

a) Sınıftan Bağımsız Teklif Öğrenimi: Seyrek bir sözlük kullanılarak; aynı asıl adayda toplu olarak STIP özelliklerini temsil edebilecek eksiksiz bir sözlük elde edilmeye çalışılmıştır. Buradaki odak noktası; gözetimli herhangi bir bilgiyi temsil eder ve bunlardan bağımsızdır. Özellik çıkarıldıktan sonra, x hesaplanır, daha sonra sözlük öğrenme problemi çözülür. Ortak temsil şeması çok amaçlı öğrenmenin bir şeklidir. Ancak bu çalışmada, her STIP tek bir görev olarak kabul edilmektedir.

b) Sınıf Kaynaklı Teklif Öğrenimi: Bir önceki a maddesindekine benzer birsözlük elde edilir. Aynı zamanda, ayrı faaliyet sınıflarına bir teklif temsili dahiletme yolunu açar. Her ikisi de uygulanabilir, ancak bu öğeden daha başarılı sonuçlar elde edilmiştir. Önceki çalışmalar da benzer sonuçlara varmıştır [23, 30]. Gözetimli bilgi ile gerçekleştirilir. Amaç, eğitim setinde denetimli bilgiler içeren evrensel bir sözlük oluşturmaktır.

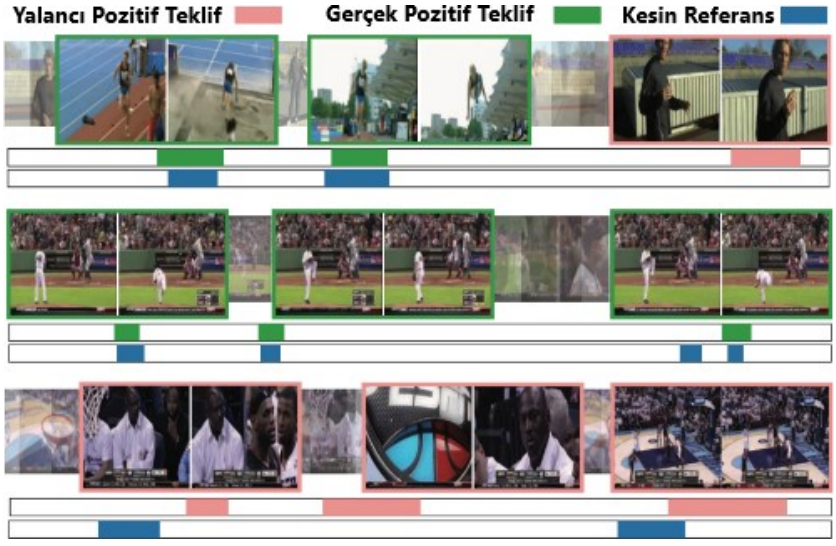
Teklifleri almak için amaçları, eğitim setindeki etkinlik örneklerine benzer etkinlikler içeren görünmeyen girdi videolarından etkinlik teklifleri getirmektir. Girdi videosu birkaç teklif adayına bölünür ve her teklif adayı kodlanır.

Her bir yöntemle üretilen sonuçların kalitesini hesaplamak için, her bir teklif ile temel hakikat zamansal ek açıklamaları arasındaki çakışma hesaplanır. Bu hesaplama tIoU ile hesaplanır. Bu yönteme göre; belirlenen eşiğin üzerindeki değerler gerçek pozitif olarak belirlenir.

İyi bir eylem önerisi algoritması, mümkün olduğunca doğru etkinlik sonuçlarını üretmeye çalışır. Bu amaçla, zamansal önerilerin kalitesi 2 farklı şekilde analiz edilmektedir. İlk olarak kaliteyi Önerilen Yerelleştirme Kalitesi (Proposed Localization Quality) yöntemi ile analiz edilmiştir. Bu bağlamda, her bir yöntemden sabit sayıda teklif alınır ve tIoU eşik değerlerine getirilen teklifler zemin gerçeği ile karşılaştırılır ve geri çağırma değeri ölçülür. Eşik değeri 0,8 olarak seçildiğinde, sınıfa bağlı yöntemin hatırlama değeri yaklaşık %80'dir. Sınıftan bağımsız yöntemin, sınıfa bağlı yönteme kıyasla düşük bir geri çağırma değeri gösterdiği ve sınıf sağlama özelliklerinin kullanılmamasının yöntemin başarısını etkilediği gözlenmiştir. İkinci olarak kaliteyi Teklif Sıralama Kalitesi (Proposal Ranking Quality) yöntemi ile analiz edilmiştir. Bu kapsamda, sadece sınırlı sayıda teklif verildiğinde, yöntemin geri çağırma değeri hesaplanmaktadır. Bunun için, her yöntemden üretilen yüksek rütbeli teklifler seçilir ve geri çağırma farklı eşik değerlerinde hesaplanır. [75]'te belirtilen ortalama geri çağırma önlemleri, eşik değerlerindeki performansı özetler ve algılama performansıyla ilişkilidir. Önerilen sınıf kaynaklı yöntemin diğer yöntemlere göre daha iyi dereceler belirlediği açıktır. Az sayıda öneri alındığında, yüksek bir hatırlama değeri verilir. Bu durum; zamana duyarlı algılama görevlerinde önem kazanır.

Önerilen yönteme ilişkin bir diğer önemli nokta, kayan pencere yöntemleriyle eylem yükü sınıflandırma süreçlerinin yükünü azaltmak için kullanılan teklif üretim yöntemleridir. Bu nedenle, yöntemlerin verimliliği son derece önemlidir. Verimlilik analizi için işlem hızı THUMOS 14 veri kümesinde [55] ölçülür ve rakip yaklaşımlarla karşılaştırılır. Videoların ortalama süresi 180 saniyedir. Önerilen yöntemin diğerlerine göre daha hızlı sonuç verdiği gözlemlenmiştir.

Makalede geliştirilen zamansal teklif yönteminin temel amacı, uzun ve gelişmemiş videolarda insan faaliyetlerinin tespitini iyileştirmektir. İşlemler, bu amaçla eylem tanıma yöntemine geliştirilen yöntem eklenerek gerçekleştirilmiştir. Her şeyden önce, eylemler her veri setindeki eğitim seti ile kesilmiş eylem örnekleri ile eğitilmiştir. Test adımında, giriş videosu, zamansal teklifler üretmek için her yöntemle işlenmiştir. Daha sonra her bir zamansal teklife eğitimli eylem sınıflandırıcıları uygulanmıştır. Tespit performansı mAP ile ölçülür. Nihai eylem algılama performansını ölçmek için her bir teklifin üretim yöntemi ile aynı işlemler gerçekleştirilir. Elde edilen zamansal eylem saptama, THUMOS 14 veri kümesi ile sonuçlanır ve 0.5 eşiğinde mAP değeri %13,5'tir. Diğer yöntemlere kıyasla önemli bir hesaplama performansı gösterir. Resim 4.9'da teklif örneklerinin sonuçları verilmektedir.



Resim 4.9: Yöntemin başarılı ve başarısız eşleştirme örnekleri [12].

4.5 Histogram Tabanlı Yaklaşım

İnsan aksiyon tanıma yöntemlerine yeni bir bakış açısı kazandırmak için, tehlikeli veya cezai durumların ortaya çıkmadan önce önlenmesine yönelik bir yaklaşım düşünülmektedir. Bu çalışmada, tümlevsel kelime torbası (integral bag-of-words) adı verilen bir yaklaşım geliştirilmiştir [16]. Savunulan yöntemin en büyük farkı, video akışlarında devam eden etkinliklerin durumunu analiz etmek için tasarlanmıştır.

Çalışmada 2 farklı tahmin yöntemi tartışılmıştır. Bunlardan biri tümlevsel kelime torbası, diğeri ise dinamik kelime torbası yöntemidir. Tahmin işlemleri bu iki yöntem ile gerçekleştirir.

a) Tümlevsel Kelime Torbası Tahmini: Özellikler ve görsel kelimeler oluşturmak için, tahmin sırasında kullanılmak üzere seçilen özellikler 3D uzay-zaman kapsamından yararlanır. Uzay-zaman özelliği çıkarıcı videodaki belirgin hareket değişikliklerini algılar ve hareketi temsil etmek için tanımlayıcılar alır [6, 49]. Resim çerçevelerini 3 boyutlu XYT zaman ekseni boyunca birleştirerek özellik ayıklayan bir videoyu birleştirir ve 3 boyutlu hacim düzeltme eklerini belirgin hareket değişiklikleriyle yerleştirir. Her bir yerel yama için tanımlayıcılar, yamadaki geçişleri özetleyerek hesaplanır.

   İlk olarak, yerel özellikler elde edilir, yöntem bu özellikleri görünüşe göre çoklu temsili tiplere göre kümeler. Buna "görsel kelimeler" denir ve bir dizi özellik olarak adlandırılır. K-means kümeleme algoritması, örnek videolardan çıkarılan özelliklerden görsel kelimeler oluşturmak için kullanılır. Hepsinden sonra, videolardan elde edilen her özellik k görsel kelimelerinden birine ait olur.

   Tümlevsel Kelime Torbası insan eylemlerini temsil etmek için tümlevsel histogramlar oluşturan olasılıklı eylem tahmin yaklaşımıdır. T uzunluğunda bir videonun O gözlemi ile eylemi sürdüren eylem tahmini için; A_p eylem olası tüm ilerleme seviyesi d için P (O | A_p, d) hesaplanmalıdır. Buraya; etkinlik olasılıklarını hesaplayan etkili bir metodolojiden, her bir faaliyetin görsel kelimelerin ayrılmaz bir histogramı olarak modellenmesiyle bahsedilir.

Tümlevsel Kelime Torbası, özellik histogramlarına dayalı olarak P (O | A_p, d) olasılıklarını hesaplayan ve devam eden eylemlerin olanaklarından mantıklı bir yaklaşımdır. Buradaki fikir, O videosu ile (A_p, d) eylem modelini histogram temsili ile karşılaştırarak aralarındaki benzerliği bulmaya dayanmaktadır. Histogram temsillerinin avantajı, çeşitli ölçeklerde gürültülü gözlemlerle başa çıkarak belirtilir. Tüm olası (A_p, d) için, bu yaklaşım etkinliğin histogramını hesaplar ve test videosunun histogramı ile karşılaştırılır.

Bu özellik histogram bir dizi k histogram kutusudur. Burada; k görsel kelime sayısını temsil eder. Belirli bir gözlem videosu için, uzamsal-zamansal konumlar yoksayılır ve her bir histogramın bölmesi, aynı tür çıkarılan özellikleri sayar. Bir (A_p, d) aktivite modelinin histogramının gösterimi; eğitim videolarındaki histogramların ortalaması, d-çerçevesinden sonraki zaman aralığından sonraki kareler eklenerek hesaplanır. Başka bir deyişle, A_p eylemi d karesine (A_p, d) ilerlerse, histogram modeli, her bir bölmeye karşılık gelen görsel kelimenin oluşum sayısını gösterir. (A_p, d) olasılıklarını etkili bir şekilde hesaplamak için, her eylem bir tümlevsel histogram oluşturularak filtrelenir. Videonun tümlevsel histogramı, bir dizi özellik histogramı olarak tanımlanır. Bu histogram yöntemi, mekansal tümlevsel histogramın geçici bir versiyonu olarak düşünülebilir [50].

Aslında, tümlevsel histogram, gözlem süresi arttıkça histogram değerlerinin nasıl değiştiğini açıklayan bir zaman fonksiyonudur. Tümlevsel histogram, eylemlerin tüm eğitim videoları için hesaplanır ve faaliyeti temsil etmek için ortalama tümlevsel histogram kullanılır. Buradaki fikir; faaliyet devam ederken görsel kelimelerde gözlenen değişiklikler takip edilmektedir.

Yerleşik tümlevsel histogram, insan faaliyetlerini tahmin etme sürecini aktive eder. Etkinliklerin tümlevsel histogramının Gauss dağılımı ile modellenmesinde eşit bir varyans vardır.

b) Dinamik Kelime Torbası Tahmini: Dinamik Kelime Torbası yöntemi bu başlık altında açıklanacaktır. Dinamik Kelime Torbası'nın Tümlevsel Kelime Torbası'ndan farkı, bütünleşmiş kelime çantası faaliyetlerinin devam eden durumlarını analiz ederek bir tahmin yapması ve çıkarılan özellikler arasındaki geçici ilişkileri göz ardı etmesidir.

Dinamik Kelime Torbası, gürültü gözlemlerinin üstesinden gelmek için kelime torbası avantajlarını kullanırken, insan faaliyetlerinin düzenli doğasını dikkate alan yeni bir etkinlik tanıma yaklaşımıdır.

Bir etkinlik videosu, insan duruşlarını tanımlayan sıralı resimlerden oluşur ve tanımadan kaldırılan uzamsal-zamansal özellikler tarafından gösterilen sıralı yapı dikkate alınmalıdır. Ayrıca, öğrenilen aktivite modeli tarafından elde edilen video gözleminin "posterior" olasılığını ölçer. Bunun avantajı P (O | A_p, d) olasılığı faaliyetlerin sıralı yapıları dikkate alınarak hesaplanmaktadır. Önceki gözlemlerin hesaplanmış olasılığından yararlanarak ve tümgözlemlerin olasılığını güncellediği belirtilmektedir. Bu artımlı olasılık hesaplaması yalnızca artımlı gözlemler için etkili aktivite tahmini sağlamakla kalmaz, aynı zamanda gözlemlerin aktivite modeliyle sırayla eşleştiği zamansal kısıtlamayı da belirtir.

Yöntem, eylem modeli ve gözlenen dizi arasında yapısal bir benzerlik bulmak için bölme işlemini gerçekleştirir. Dikkat edilmesi gereken bir nokta, eylem modeli segment süresinin (Δd) yeni gözlem segmentiyle eşleştirilebilmesi için dinamikleri seçmek ve benzerlik mesafelerini tekrar tekrar hesaplamak için en uygun segment çiftlerinin bulunması gerektiğidir. Segment olasılığı P P ($O^{\Delta t}$ | A_p, Δd), histogram gösterimleri karşılaştırılarak hesaplanır. Buna göre; kelime torbası paradigması aralık segmentleriyle eşleşmek için uygulanırken, segmentlerin kendileri yinelemeli etkinlik tahmini formülasyonuna göre düzenlenir.

Tümlevsel histogram kullanılarak video segment eşleştirmesi yapma: Dinamik kelime çantası aralık segmentleri arasındaki benzerliği hesaplamak için tümlevsel histogramı kullanır. Tümlevsel histogramlar mümkün olan her (Δd,Δt) için verimli yapı sağlar. Son olarak, aktiviteyi tahmin etmek için Maksimum a posteriori (MAP) sınıflandırıcısı kullanıldı.

Yaklaşıma göre, iki yöntem uygulanmıştır; bu yöntemler, yaklaşım başlığında tarif edildiği gibi, tümlevsel kelime torbası ve dinamik kelime torbasıdır. Sistemde mekansal-zamansal özellikler küboid özellik tanımlayıcıları olarak seçilir [49]. Bütünleşmiş histogramlar, test videolarındaki eylemleri tanımak için eğitim videoları üzerine inşa edilmiştir. Bu konfigürasyonlarla yapılan testler sonucunda Dinamik Kelime Torbası ve Tümlevsel Kelime Torbası'nın karşılaştırılan diğer yöntemlerden daha yüksek doğruluk değerlerine sahip olduğu görülebilir. Bu, kaba kuvvet yöntemlerinden daha verimli çalışır. Her doğruluk değeri en yüksek performansı veren optimum parametreler ve görsel kelimelerle gerçekleştirildiğinde, Dinamik Kelime Torbası sonuçları %85 ve Tümlevsel Kelime Torbası sonuçları tam videolarla %81,7'dir.

5. NİCEL ANALİZ SONUÇLARI

Bölüm 5 altında on beş farklı makale ayrıntılı olarak açıklandıktan sonra, bu makaleler arasındaki farklılıkları, benzerlikleri, performans oranlarını ve yöntemleri daha net bir şekilde incelemek için karşılaştırmalar yapılmıştır. Bu bölümde nicel analizler yapılmış veri tabanları ve yaklaşımların analizi verilmiştir. Bu bağlamda, analiz sonuçları iki alt başlık altında paylaşılacaktır.

5.1 Eylem Tanıma Verisetleri Analizi

Eylem tanıma yaklaşımlarının başarısı üzerinde önemli etkisi olan veri kümeleri incelenmiştir. Bu veri kümelerinin hangi yaklaşımda kullanıldığı ve nasıl bölümlendikleri hakkında tablolar oluşturulmuştur. Yaklaşımlarda, video sayımlarında, video sürelerinde, çözünürlüklerde ve saniyede kare sayısında kullanılan veri kümelerinin karmaşıklığı, tanıma başarısını etkileyen önemli faktörlerdir. Bu bilgilerin tamamı ve karşılaştırması Bölüm 3'te ve EK 1'de açıklanmaktadır.

Veri setlerinin yaklaşımlar üzerinde eğitim ve testlere ayrılması sürecinde oranların ve kategorilerin doğru şekilde tahsis edilmesi, uygunluk durumunu önlemeye, tespit edilecek kategorilerin doğru bir şekilde belirlenmesine yardımcı olur ve sonuçta tanıma başarısını artırır. Bu bağlamda, araştırma sırasında ele alınan çalışmalarda kullanılan veri kümelerinin ayrışma tablosu ve ayrışma konfigürasyonları Çizelge 5.1 ve EK 2'de verilmektedir.

Çizelge 5. 1: Çalışmalarda kullanılan eğitim ve test veri setleri listesi.

Çalışma	Eğitim Veri Seti	Test Veri Seti
Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning [1]	KITTI Veri Seti [3]	CalTech Pedestrian Veri Seti [4]
Anticipating Visual Representations from Unlabeled Video [13]	TV Human Interaction Veri Seti [5]	THUMOS 2015 [56] TV Human Interaction Veri Seti [5]
Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos [12]	KTH Veri Seti [6]	THUMOS 2014 Detection Challenge Veri Seti [55]

Çizelge 5. 2: Çalışmalarda kullanılan eğitim ve test veri setleri listesi. (devam)

Human Action Recognition in Drone Videos Using a Few Aerial Training Examples [8]	UCF-ARG [7] Youtube Aerial [8]	UCF-ARG [7] Youtube Aerial [8]
AdaScan: Adaptive Scan Pooling in Deep convolutional Neural Networks for Human Action Recognition in Videos [47]	UCF-101 [9] HMDB-51 [11]	UCF-101 [9] HMDB-51 [11]
Predicting Future Frames using Retrospective Cycle GAN [14]	KITTI Veri Seti [3]	CalTech Pedestrian Veri Seti [4] UCF-101 [9] CUHK Avenue [57] ShanghaiTech Campus Veri Seti [58]
Pose Primitive Based Human Action Recognition in Videos or Still Images [20]	Weizmann Veri Seti [10]	Weizmann Veri Seti [10] KTH Veri Seti [6]
A New Hybrid Architecture for Human Activity Recognition from RGB-D Videos [15]	CAD-60 [59] CAD-120 [59] MSRDailyActivity3D [60] NTURGB+D [61]	CAD-60 [59] CAD-120 [59] MSRDailyActivity 3D [60] NTURGB+D [61]
Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos [16]	UT-Interaction Veri Seti [62]	UT-Interaction Veri Seti [62]
Lattice Long Short-Term Memory for human Action Recognition [18]	UCF-101 [9] HMDB-51 [11]	UCF-101 [9] HMDB-51 [11]
Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning [19]	Weizmann Veri Seti [10] KTH Veri Seti [6]	Weizmann Veri Seti [10] KTH Veri Seti [6]
Skeleton-Based Action Recognition With Shift Graph Convolutional Network [120]	NTURGB+D [61] NTU-120 RGB+D [61] Northwestern UCLA [106]	NTURGB+D [61] NTU-120 RGB+D [61] Northwestern UCLA [106]
FASTER Recurrent Networks for Efficient Video Classification [121]	Kinetics [108]	UCF-101 [9] HMDB-51 [11] Kinetics [108]
Slowfast networks for video recognition [125]	Kinetics - 400 [122]	Kinetics - 400 [122] Kinetics - 600 [123] Charades [111] AVA [124]

Temporal segment networks for action recognition in videos [127]	UCF-101 [9] HMDB-51 [11] ActivityNet [126]	UCF-101 [9] HMDB-51 [11] ActivityNet [126]

5.2 Eylem Tanıma Yöntemleri Analizleri

Araştırma sırasında incelenen makalelerin eylem tanıma sonuçları Çizelge 5.2'de verilmektedir. Tüm sonuçlar orijinal çalışmalardan alınmıştır. Tablodaki tüm sonuçlar, en iyi sonucu veren yapılandırma değerlerinden elde edilir. Bazı makalelerde doğruluk değerleri verilirken, bazılarında MSE, SSIM ve mAP değerleri verilmiştir.

SSIM'in yüksek değerlerde olmasını beklerken MSE'nin mümkün olduğunca düşük olması hedeflenmektedir.

Tüm bu sonuçlara ek olarak, sonuçları etkileyen bazı özel durumlar da vardır. [8] 'de test işlemleri iki veri seti ile gerçekleştirilmiş ve bu iki veri setinin sonuçları arasında büyük bir fark gözlenmiştir. Bunun, UCF-ARG veri kümelerindeki videoların arka planının tespit edilmesini zorlaştırdığı ve küçük boyutlu aktörlerin olduğu gerçeğinden kaynaklanmaktadır.

[17]'deki AdaScan yöntemi mekansal ve zamansal olmak üzere iki farklı ağ yapısına sahiptir. Bu bağlamda, tablodaki sonuçlar ağ yapısının en iyi olduğu sonuçlara karşılık gelmektedir. Bu çalışmada, iki farklı veri kümesinde gerçekleştirilen test operasyonlarının sonuçları arasında büyük farklılıklar gözlenmiştir. Bunun nedeni HMDB51 veri kümelerinin UCF-101'den daha az eğitim videosu içermesi, eylemlerin

birbirinden farklı olması ve UCF-101'in HMDB51 ile karşılaştırıldığında daha belirgin kategoriler içermesidir.

[20] 'de testler, hem Weizmann hem de KTH veri tabanları ile gerçekleştirilmiştir. Ancak, makalede yalnızca Weizmann veritabanının genel doğruluk sonuçları paylaşılmaktadır. Diğer çalışmaların tüm sonuçları tabloya dahil edilmiştir.

Çizelge 5.2: Yaklaşımların tanıma sonuçları.

Veri Seti	Kategori	Çalışma	Yöntem	Doğruluk Oranı (%)	MSE (%)	SSIM (%)	mAP (%)
TV Human Interaction Veri Seti	Ağ Tabanlı	[13]	Derin Regresyon Ağı	43.3 ± 4.7	-	-	-
THUMOS 2015 Veri Seti				43.6 ± 4.8	-	-	-
THUMOS 2014 Veri Seti	Sözlük Tabanlı	[12]	Sınıf Kaynaklı & Sınıf Bağımsız Teklif Öğrenme	-	-	-	13,5
	Ağ Tabanlı	[127]	Zamansal Kesim Ağı	80,1	-	-	-
UCF- ARG Veri Seti	Ağ Tabanlı	[8]	Ayrık Çok Görevli Öğrenme	32,5	-	-	-
YouTube Aerial Veri Seti				68,3	-	-	-
UCF101 Veri Seti	Ağ Tabanlı	[17]	AdaScan	83,4	-	-	-
	Ağ Tabanlı	[14]	Geçmişe Dönük Çevrimli Çekişmeli Üretici Ağlar	-	1,37	0,94	-
	Ağ Tabanlı	[18]	Kafes-LSTM	93,6	-	-	-
	Ağ Tabanlı	[121]	FASTER	96,9	-	-	-
	Ağ Tabanlı	[127]	Zamansal Kesim Ağı	94,9	-	-	-
HMDB51 Veri Seti	Ağ Tabanlı	[17]	AdaScan	49,2	-	-	-
	Ağ Tabanlı	[18]	Kafes-LSTM	66,2	-	-	-
	Ağ Tabanlı	[121]	FASTER	75,7	-	-	-
	Ağ Tabanlı	[127]	Zamansal Kesim Ağı	71	-	-	-
CalTech Pedestrian Veri Seti	Ağ Tabanlı	[1]	Öngörülü Sinir Ağı (PredNet)	-	$3.13x\ 10^{-3}$	0,884	-
	Ağ Tabanlı	[14]	Geçmişe Dönük Çevrimli Çekişmeli Üretici Ağlar	-	1,61	0,919	-
CUHK Avenue Veri Seti	Ağ Tabanlı	[14]	Geçmişe Dönük Çevrimli Çekişmeli Üretici Ağlar	-	0,39	0,98	-
ShanghaiTech Veri Seti				-	0,64	0,97	-
Weizmann Veri Seti	Histogram Tabanlı	[20]	Poz İlkeli Tanıma	94,4	-	-	-
	Çoklu Örnek Öğrenme Tabanlı	[19]	Kinematik Özellikler ile Çoklu Örnek Öğrenme	95,75	-	-	-
CAD- 60 Veri Seti	Hareket Tabanlı	[15]		98,52	-	-	-

CAD- 120 Veri Seti			İki Seviyeli Füzyon Stratejisi	94,4	-	-	-
MSRDailyActivity3D Veri Seti				97,81	-	-	-
NTURGB+D Veri Seti				92,2	-	-	-
	Ağ Tabanlı	[120]	Kaydırma Çizge Evrişimli Ağı	96,5	-	-	-
UT Interaction Veri Seti	Histogram Tabanlı	[16]	Tümlevsel Kelime Torbası	81,7	-	-	-
			Dinamik Kelime Torbası	85	-	-	-
KTH Veri Seti	Çoklu Örnek Öğrenme Tabanlı	[19]	Kinematik Özellikler ile Çoklu Örnek Öğrenme	87,7	-	-	-
NTU-120 RGB+D Veri Seti	Ağ Tabanlı	[120]	Kaydırma Çizge Evrişimli Ağı	85,9	-	-	-
Northwestern UCLA			Kaydırma Çizge Evrişimli Ağı	94,6	-	-	-
Kinetics [108]	Ağ Tabanlı	[121]	FASTER	75,3	-	-	-
Kinetics- 400 [122]	Ağ Tabanlı	[125]	SlowFast Ağı	79,8	-	-	-
Kinetics- 600 [123]			SlowFast Ağı	81,8	-	-	-
Charades [111]	Ağ Tabanlı	[125]	SlowFast Ağı	-	-	-	45,2
AVA [124]	Ağ Tabanlı	[125]	SlowFast Ağı	-	-	-	28,2
ActivityNet [126]			Zamansal Kesim Ağı	89,6	-	-	-

6. KIYASLAMA SONUÇLARI

Araştırma çalışması sırasında hedeflenen bir başka araştırma, aynı veri seti ile ele alınan tüm makaleleri çalıştırmak ve doğruluk sonuçlarını elde etmek ve karşılaştırmaktı. Ancak, alan araştırmamız kapsamındaki tüm makalelerin çalışmaya hazır kodlarının paylaşılmaması nedeniyle, istenen doğruluk değerlerine ulaşılamamıştır. Öte yandan, araştırmada incelenen beş makalenin kodlarının paylaşıldığı ve bu yaklaşımları karşılaştırmak için farklı makalelerin kodlarının alındığı gözlenmiştir. Kodları paylaşan makaleler sırasıyla PredNet [1], AdaScan [17], Sınıftan Kaynaklı ve Sınıftan Bağımsız Öğrenme [12], SlowFast Ağı [125], Zamansal Kesim Ağı [127]'dır.Bu metotlara ek olarak, benzer action recognition yöntemlerinden 4 farklı yöntem karşılaştırma için seçilmiştir. Bu yöntemler detaylı olarak açıklanmayacaktır. Yalnızca belirtilen konfigürasyon ve hiperparametreler ile çalıştırılmış ve sonuçlar elde edilmiştir. Yapılan bu karşılaştırmalarda Python dili kullanılmış ve hepsi için uygun olan veri seti olarak UCF-101 veri seti seçilmiştir. Çalıştırılan yöntemlerin tamamı Linux işletim sistemi üzerinde ve Nvidia RTX 2060

GPU ile gerçekleştirilmiştir. Çizelge 5.3'te uygulanan yöntemlerin karşılaştırma sonuçları dahil edilmiştir.

Burada değinilmesi gereken bir nokta ise, [12]'deki makalenin kodları çalıştırılamamıştır. Bu durumun sebebi kodların çalıştırılması için gereken özellik çıkarma yöntemi paylaşılmamıştır. Yalnızca daha önceden eğitilmiş veriler ile çalıştırılmasına müsaade edilmektedir. Diğer yandan [85]'de ise hali hazırda bulunan işletim sistemimizden eski bir versiyon kullanılmıştır. Bununla birlikte, Torch için gerekli kurulumlar gerçekleştirilirken Nvidia RTX mimarisi ile uyumlu olmadığı gözlemlenmiştir. Kodların yalnızca GTX mimarisinde çalıştırılması uygundur. Sonuç olarak, bu iki makale yöntemleri çalıştırılamamış, uygun koşullar tespit edildiği için bilgilendirmenin faydalı olacağı düşünülmüştür.

Çizelge 5.3: Yaklaşımların kıyaslama sonuçları.

Çalışma	İşletim Sistemi	Dil	Gereksinimler	Doğruluk (%)	MSE
Deep predictive coding networks for video prediction and unsupervised learning [1]	Ubuntu 16.04	Python 3.6	Keras=2.2.40 Tensorflow-gpu=1.6 scipy=1.1.0 requests bs4 numpy imageio hickle matplotlib	-	0.004503
Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. [12]	Ubuntu 16.04	Python 2.7	numpy h5py scikit-learn pandas joblib spams	-	-
Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. [17]	Ubuntu 16.04	Python 2.7	numpy skimage sk-video tensorflow	91.6	-

Learning spatiotemporal features with 3d convolutional networks. [84]	Ubuntu 16.04	Python 3.6	numpy=1.16.4 Keras=2.2.4 pipe=1.5.0 moviepy=1.1.0 mPyPl=0.0.3.7 scikit-learn=0.21.3	78	-
TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition [85]	Ubuntu 14.04	-	Torch Cuda & cuDNN torch-pastalog	-	-
Quo vadis, action recognition? a new model and the kinetics dataset [86]	Ubuntu 16.04	Python 3.6	tensorflow numpy sonnet	84.4	-
Towards good practices for very deep two-stream convnets [87]	Ubuntu 16.04	Python 3.5	Cuda 8.0 OpenCV 3 dense_flow	SN 85.60* TN 85.71	-
Slowfast networks for video recognition [125]	Ubuntu 16.04	Python 3.6	tensorflow=1.12 pillow=5.1.0 ffmpeg opencv-python	55.4	-
Temporal segment networks for action recognition in videos [127]	Ubuntu 16.04	Python 3.6	Caffe OpenCV 2.4.13 dense_flow	94.8	-

7. GELECEKTEKİ ÇALIŞMALAR

İnsan eylemi tanıma yöntemleri gittikçe daha popüler hale gelirken, kullanılan yöntemler artık çoğunlukla derin sinir ağına dayanmaktadır. Derin sinir ağı tabanlı mimarilerle daha yüksek doğruluk değerleri elde edildiğini görüyoruz. Burada önemli bir nokta, derin sinir ağ tabanlı yöntemler kullanmak yerine, birden çok yöntemin melez yapılar oluşturmak için harmanlandığını dikkate almak gerekir. Ayrıca; insan eyleminin tanınmasının günlük hayatta kullanılacağı noktalarda, gerçek zamanlı video işleme son derece önemli hale gelir. Gerçek video işleme yöntemleri ile, anormal

durumların tespiti ve suç tespiti gibi kritik konularda yararlı olacaktır. Bu nedenle, geliştirilen yöntemler hız, kamera dönüş açıları ve algılama için gerekli donanım gibi performansta önemli değişiklikler yapmalıdır.

8. SONUÇ

Birçok farklı alanda kullanılan insan eylem tanıma özelliğinin videolara uygulanması karmaşık bir sorun olarak tanımlanmaktadır. Bu karmaşık soruna bir perspektif kazandırmak için kapsamlı bir genel bakış sunduk. Ele alınan 15 farklı makalenin yöntemlerini inceledik ve bu yöntemleri detaylı olarak açıkladık. Her bir makalenin yöntemini kapsayacak şekilde bir taksonomi oluşturduk ve bu taksonomide beş ana yöntem kategorisi oluşturduk. Oluşturulan bu beş ana kategoride ele alınan yaklaşımları analiz ettik.

Nispeten başarılı sayılacak yöntemleri göz önünde bulundurarak mimarilerine yakından baktık ve bu yöntemlerin kullandığı veri setlerini ayrıntılı olarak ele aldık. Her bir veri setindeki videoları inceleyerek veri setinin orijinal sayfalarında da yer almayan çözünürlük, FPS ve video süreleri gibi bilgilere ulaştık ve tüm bu bilgilerin karşılaştırılacağı detaylı bir tablo oluşturduk. Bu sayede araştırmada ele alınan veri setleri ile geliştirilecek yeni yöntemlerin temelini oluşturduk. Ayrıca, yaklaşımların ve veri kümelerinin nicel analizini yaptık. Analiz sonuçları da tablolar halinde detaylandırarak okuyucular için bir bakış açısı sağladık. Son olarak, alan araştırmamızı çalışmamızı insan eylemlerini tanıma problemini geliştirmek için bazı açık araştırma alanlarına değinerek sonlandırdık.

KAYNAKLAR

[1] **Lotter, W., Kreiman, G., & Cox, D.** (2016). Deep predictive coding networks for video prediction and unsupervised learning. arXiv preprint arXiv:1605.08104

[2] **Singular Inversions, Inc. FaceGen**. http://facegen.com.

[3] **Geiger, A., Lenz, P., Stiller, C., & Urtasun, R.** (2013). Vision meets robotics: The kitti dataset. The International Journal of Robotics Research, 32(11), 1231-1237.

[4] **Dollár, P., Wojek, C., Schiele, B., & Perona, P.** (2009, June). Pedestrian detection: A benchmark. In 2009 IEEE Conference on Computer Vision and Pattern Recognition (pp. 304-311). IEEE.

[5] "TV Human Action Interaction Dataset", http://www.robots.ox.ac.uk/~alonso/tv_human_interactions.html

[6] **Schuldt, C., Laptev, I., & Caputo, B.** (2004, August). Recognizing human actions: a local SVM approach. In Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004. (Vol. 3, pp. 32-36). IEEE.
[7] **"Ucf-arg data set,"** https://www.crcv.ucf.edu/data/UCF-ARG.php, accessed: 2020-09-8.
[8] **Sultani, W., & Shah, M.** (2019). Human Action Recognition in Drone Videos using a Few Aerial Training Examples. arXiv preprint arXiv:1910.10027.
[9] **Soomro, K., Zamir, A. R., & Shah, M.** (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402.
[10] **Gorelick, L., Blank, M., Shechtman, E., Irani, M., & Basri, R.** (2007). Actions as space-time shapes. IEEE transactions on pattern analysis and machine intelligence, 29(12), 2247-2253.
[11] **Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T.** (2011, November). HMDB: a large video database for human motion recognition. In 2011 International Conference on Computer Vision (pp. 2556-2563). IEEE.
[12] **Caba Heilbron, F., Carlos Niebles, J., & Ghanem, B.** (2016). Fast temporal activity proposals for efficient detection of human actions in untrimmed videos. In Proceedings of the IEEE conference on computer vision and pattern recognition(pp. 1914-1923).
[13] **Vondrick, C., Pirsiavash, H., & Torralba, A.** (2016). Anticipating visual representations from unlabeled video. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 98-106).
[14] **Kwon, Y. H., & Park, M. G.** (2019). Predicting future frames using retrospective cycle gan. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition(pp. 1811-1820).
[15] **Das, S., Thonnat, M., Sakhalkar, K., Koperski, M., Bremond, F., & Francesca, G.** (2019, January). A new hybrid architecture for human activity recognition from rgb-d videos. In International Conference on Multimedia Modeling (pp. 493-505). Springer, Cham.
[16] **Ryoo, M. S.** (2011, November). Human activity prediction: Early recognition of ongoing activities from streaming videos. In 2011 International Conference on Computer Vision (pp. 1036-1043). IEEE.
[17] **Kar, A., Rai, N., Sikka, K., & Sharma, G.** (2017). Adascan: Adaptive scan pooling in deep convolutional neural networks for human action recognition in videos. In Proceedings of the IEEE conference on computer vision and pattern recognition(pp. 3376-3385).
[18] **Sun, L., Jia, K., Chen, K., Yeung, D. Y., Shi, B. E., & Savarese, S.** (2017). Lattice long short-term memory for human action recognition. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2147-2156).
[19] **Ali, S., & Shah, M.** (2008). Human action recognition in videos using kinematic features and multiple instance learning. IEEE transactions on pattern analysis and machine intelligence, 32(2), 288-303.
[20] **Thurau, C., & Hlavác, V.** (2008, June). Pose primitive based human action recognition in videos or still images. In 2008 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-8). IEEE.
[21] **Rao, R. P., & Ballard, D. H.** (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nature neuroscience, 2(1), 79-87.
[22] **Hochreiter, S., & Schmidhuber, J.** (1997). Long short-term memory. Neural computation, 9(8), 1735-1780.

[23] **Krizhevsky, A., Sutskever, I., & Hinton, G. E.** (2012). Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems (pp. 1097-1105).
[24] **Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., & Oliva, A.** (2014). Learning deep features for scene recognition using places database. In Advances in neural information processing systems (pp. 487-495).
[25] **Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... & Darrell, T.** (2014, November). Caffe: Convolutional architecture for fast feature embedding. In Proceedings of the 22nd ACM international conference on Multimedia (pp. 675-678).
[26] **Laptev, I.** (2005). On space-time interest points. International journal of computer vision, 64(2-3), 107-123.
[27] **Rhee, E. J.** (2018). A Deep Learning Approach for Classification of Cloud Image Patches on Small Datasets. Journal of information and communication convergence engineering, 16(3), 173-178.
[28] **Yang, J., Yu, K., Gong, Y., & Huang, T.** (2009, June). Linear spatial pyramid matching using sparse coding for image classification. In 2009 IEEE Conference on computer vision and pattern recognition (pp. 1794-1801). IEEE.
[29] **Guha, T., & Ward, R. K.** (2011). Learning sparse representations for human action recognition. IEEE transactions on pattern analysis and machine intelligence, 34(8), 1576-1588.
[30] **Jiu, M., Wolf, C., Garcia, C., & Baskurt, A.** (2012). Supervised learning and codebook optimization for bag-of-words models. Cognitive Computation, 4(4), 409-419.
[31] **Zhao, B., & Xing, E. P.** (2014). Quasi real-time summarization for consumer videos. In Proceedings of the IEEE conference on computer vision and pattern recognition(pp. 2513-2520).
[32] **Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y.** (2014). Generative adversarial nets. In Advances in neural information processing systems (pp. 2672-2680).
[33] **Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C.** (2017). Improved training of wasserstein gans. In Advances in neural information processing systems(pp. 5767-5777).
[34] **Chen, Y., Kalantidis, Y., Li, J., Yan, S., & Feng, J.** (2018). Multi-fiber networks for video recognition. In Proceedings of the european conference on computer vision (ECCV) (pp. 352-367).
[35] **Simonyan, K., & Zisserman, A.** (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
[36] **Mao, X., Li, Q., Xie, H., Lau, R. Y., Wang, Z., & Paul Smolley, S.** (2017). Least squares generative adversarial networks. In Proceedings of the IEEE International Conference on Computer Vision (pp. 2794-2802).
[37] **Johnson, J., Alahi, A., & Fei-Fei, L.** (2016, October). Perceptual losses for real-time style transfer and super-resolution. In European conference on computer vision (pp. 694-711). Springer, Cham.
[38] **Ulyanov, D., Vedaldi, A., & Lempitsky, V.** (2016). Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022.
[39] **Koperski, M.** (2017). Human action recognition in videos with local representation (Doctoral dissertation).

[40] **Chen, Y., Bi, J., & Wang, J. Z.** (2006). MILES: Multiple-instance learning via embedded instance selection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(12), 1931-1947.
[41] **Lu, W. L., & Little, J. J.** (2006, June). Simultaneous tracking and action recognition using the pca-hog descriptor. In The 3rd Canadian Conference on Computer and Robot Vision (CRV'06) (pp. 6-6). IEEE.
[42] **Zhang, L., Wu, B., & Nevatia, R.** (2007, October). Detection and tracking of multiple humans with extensive pose articulation. In 2007 IEEE 11th International Conference on Computer Vision (pp. 1-8). IEEE.
[43] **Lee, D. D., & Seung, H. S.** (2001). Algorithms for non-negative matrix factorization. In Advances in neural information processing systems (pp. 556-562).
[44] **Ward Jr, J. H.** (1963). Hierarchical grouping to optimize an objective function. Journal of the American statistical association, 58(301), 236-244.
[45] **Bissacco, Alessandro, Ming-Hsuan Yang, and Stefano Soatto.** (2007). "Detecting humans via their pose." Advances in Neural Information Processing Systems.
[46] **Hamid, R., Johnson, A., Batta, S., Bobick, A., Isbell, C., & Coleman, G.** (2005, June). Detection and explanation of anomalous activities: Representing activities as bags of event n-grams. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)(Vol. 1, pp. 1031-1038). IEEE.
[47] **Schroff, F., Criminisi, A., & Zisserman, A.** (2006). Single-histogram class models for image segmentation. In Computer Vision, Graphics and Image Processing (pp. 82-93). Springer, Berlin, Heidelberg.
[48] **Ullman, S., Vidal-Naquet, M., & Sali, E.** (2002). Visual features of intermediate complexity and their use in classification. Nature neuroscience, 5(7), 682-687.
[49] **Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S.** (2005, October). 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). IEEE.
[50] **Porikli, F.** (2005, June). Integral histogram: A fast way to extract histograms in cartesian spaces. In 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 1, pp. 829-836). IEEE.
[51] **Elad, M., & Aharon, M.** (2006). Image denoising via sparse and redundant representations over learned dictionaries. IEEE Transactions on Imageprocessing, 15(12), 3736-3745.
[52] Xingjian, S. H. I., Chen, Z., Wang, H., Yeung, D. Y., Wong, W. K., & Woo, W. C. (2015). Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In Advances in neural information processing systems (pp. 802-810).
[53] **Ioffe, S., & Szegedy, C.** (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprintarXiv:1502.03167.
[54] **Sun, L., Jia, K., Yeung, D. Y., & Shi, B. E.** (2015). Human action recognition using factorized spatio-temporal convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4597-4605).
[55] **Jiang, Yu-Gang, et al.** (2014) THUMOS challenge: Action recognition with a large number of classes.
[56] **Gorban, Alex, et al.** (2015) THUMOS challenge: Action recognition with a large number of classes.

[57] **Lu, C., Shi, J., & Jia, J.** (2013). Abnormal event detection at 150 fps in matlab. In Proceedings of the IEEE international conference on computer vision (pp. 2720-2727).
[58] **Liu, W., Luo, W., Lian, D., & Gao, S.** (2018). Future frame prediction for anomaly detection–a new baseline. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6536-6545).
[59] **Sung, J., Ponce, C., Selman, B., & Saxena, A.** (2012, May). Unstructured human activity detection from rgbd images. In 2012 IEEE international conference on robotics and automation (pp. 842-849). IEEE.
[60] **Wang, J., Liu, Z., Wu, Y., & Yuan, J.** (2012, June). Mining actionlet ensemble for action recognition with depth cameras. In 2012 IEEE Conference on Computer Vision and Pattern Recognition (pp. 1290-1297). IEEE.
[61] **Shahroudy, A., Liu, J., Ng, T. T., & Wang, G.** (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 1010-1019).
[62] **Ryoo, Michael S., and J. K. Aggarwal.** (2010) UT-interaction dataset, ICPR contest on semantic description of human activities (SDHA). IEEE International Conference on Pattern Recognition Workshops.
[63] **Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L.** (2009, June)**.** Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248-255). IEEE.
[64] **Wang, L., Xiong, Y., Wang, Z., & Qiao, Y.** (2015). Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159.
[65] **Maaten, L. V. D., & Hinton, G.** (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579-2605.
[66] **Kingma, D. P., & Ba, J.** (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.
[67] **Byeon, W., Wang, Q., Kumar Srivastava, R., & Koumoutsakos, P.** (2018). Contextvp: Fully context-aware video prediction. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 753-769).
[68] **Zach, C., Pock, T., & Bischof, H.** (2007, September). A duality based approach for realtime TV-L 1 optical flow. In Joint pattern recognition symposium (pp. 214-223). Springer, Berlin, Heidelberg.
[69] **Patron-Perez, A., Marszalek, M., Zisserman, A., & Reid, I. D.** (2010, August). High Five: Recognising human interactions in TV shows. In BMVC (Vol. 1, No. 2, p. 33).
[70] **Sadanand, S., & Corso, J. J.** (2012, June). Action bank: A high-level representation of activity in video. In 2012 IEEE Conference on Computer Vision and Pattern Recognition(pp. 1234-1241). IEEE.
[71] **Yilmaz, A., Li, X., & Shah, M.** (2004). Contour-based object tracking with occlusion handling in video acquired using mobile cameras. IEEE Transactions on pattern analysis and machine intelligence, 26(11), 1531-1536.
[72] **2009,**http://www.nada.kth.se/cvap/actions/00sequences.txt
[73] **Simonyan, K., & Zisserman, A.** (2014). Two-stream convolutional networks for action recognition in videos. In Advances in neural information processing systems (pp. 568-576).
[74] **Glorot, X., & Bengio, Y.** (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256).

[75] **Hosang, J., Benenson, R., Dollár, P., & Schiele, B.** (2015). What makes for effective detection proposals?. IEEE transactions on pattern analysis and machine intelligence, 38(4), 814-830.
[76] **Poppe, R.** (2010). A survey on vision-based human action recognition. Image and vision computing, 28(6), 976-990
[77] **Zhu, F., Shao, L., Xie, J., & Fang, Y.** (2016). From handcrafted to learned representations for human action recognition: A survey. Image and Vision Computing, 55, 42-52.
[78] **Weinland, D., Ronfard, R., & Boyer, E.** (2011). A survey of vision-based methods for action representation, segmentation and recognition. Computer vision and image understanding, 115(2), 224-241.
[79] **Chaquet, J. M., Carmona, E. J., & Fernández-Caballero, A.** (2013). A survey of video datasets for human action and activity recognition. Computer Vision and Image Understanding, 117(6), 633-659.
[80] **Cheng, G., Wan, Y., Saudagar, A. N., Namuduri, K., & Buckles, B. P.** (2015). Advances in human action recognition: A survey. arXiv preprint arXiv:1501.05964.
[81] **Aggarwal, J. K., & Ryoo, M. S.** (2011). Human activity analysis: A review. ACM Computing Surveys (CSUR), 43(3), 1-43.
[82] **Dhamsania, C. J., & Ratanpara, T. V.** (2016, November). A survey on Human action recognition from videos. In 2016 Online International Conference on Green Engineering and Technologies (IC-GET) (pp. 1-5). IEEE.
[83] **Herath, S., Harandi, M., & Porikli, F.** (2017). Going deeper into action recognition: A survey. Image and vision computing, 60, 4-21.
[84] **Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M.** (2015). Learning spatiotemporal features with 3d convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 4489-4497).
[85] **Ma, C. Y., Chen, M. H., Kira, Z., & AlRegib, G.** (2019). TS-LSTM and temporal-inception: Exploiting spatiotemporal dynamics for activity recognition. Signal Processing: Image Communication, 71, 76-87.
[86] **Carreira, J., & Zisserman, A. (2017).** Quo vadis, action recognition? a new model and the kinetics dataset. In proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 6299-6308).
[87] **Wang, L., Xiong, Y., Wang, Z., & Qiao, Y.** (2015). Towards good practices for very deep two-stream convnets. arXiv preprint arXiv:1507.02159
[88] **Zhu, Linchao, Zhongwen Xu, and Yi Yang.** "Bidirectional multirate reconstruction for temporal modeling in videos." Proceedings of the IEEE conference on computer vision and pattern recognition. 2017.
[89] **Su, Kun, Xiulong Liu, and Eli Shlizerman.** "Predict & cluster: Unsupervised skeleton based action recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[90] **Kejun, Wang, and P. Popoola Oluwatoyin.** "Ant-based clustering of visual-words for unsupervised human action recognition." 2010 Second World Congress on Nature and Biologically Inspired Computing (NaBIC). IEEE, 2010.
[91] **Srivastava, Nitish, Elman Mansimov, and Ruslan Salakhudinov.** "Unsupervised learning of video representations using lstms." International conference on machine learning. 2015.
[92] **Han, Tengda, Weidi Xie, and Andrew Zisserman.** "Video representation learning by dense predictive coding." Proceedings of the IEEE International Conference on Computer Vision Workshops. 2019.

[93] **Xu, Dejing, et al.** "Self-supervised spatiotemporal learning via video clip order prediction." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[94] **Wang, Jiangliu, et al.** "Self-supervised spatio-temporal representation learning for videos by predicting motion and appearance statistics." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[95] **Alwassel, Humam, et al**. "Self-supervised learning by cross-modal audio-video clustering." arXiv preprint arXiv:1911.12667 (2019).
[96] **Jaouedi, Neziha, Noureddine Boujnah, and Med Salim Bouhlel.** "A new hybrid deep learning model for human action recognition." Journal of King Saud University-Computer and Information Sciences 32.4 (2020): 447-453.
[97] **Xiong, Qianqian, et al.** "Transferable two-stream convolutional neural network for human action recognition." Journal of Manufacturing Systems (2020).
[98] **Majumder, Sharmin, and Nasser Kehtarnavaz.** "Vision and Inertial Sensing Fusion for Human Action Recognition: A Review." IEEE Sensors Journal (2020).
[99] **Zhang, Hong-Bo, et al.** "A comprehensive survey of vision-based human action recognition methods." Sensors 19.5 (2019): 1005.
[100] **Dang, L. Minh, et al.** "Sensor-based and vision-based human activity recognition: A comprehensive survey." Pattern Recognition 108 (2020): 107561.
[101] **Singh, Tej, and Dinesh Kumar Vishwakarma.** "Video benchmarks of human action datasets: a review." Artificial Intelligence Review 52.2 (2019): 1107-1154.
[102] **Kong, Yu, and Yun Fu.** "Human action recognition and prediction: A survey." arXiv preprint arXiv:1806.11230 (2018).
[103] **Guo, Guodong, and Alice Lai**. "A survey on still image based human action recognition." Pattern Recognition 47.10 (2014): 3343-3361.
[104] **Marszalek, Marcin, Ivan Laptev, and Cordelia Schmid.** "Actions in context." 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009.
[105] **Li, Wanqing, Zhengyou Zhang, and Zicheng Liu.** "Action recognition based on a bag of 3d points." 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops. IEEE, 2010.
[106] **Wang, Jiang, et al.** "Cross-view action modeling, learning and recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2014.
[107] **Niebles, Juan Carlos, Chih-Wei Chen, and Li Fei-Fei.** "Modeling temporal structure of decomposable motion segments for activity classification." European conference on computer vision. Springer, Berlin, Heidelberg, 2010.
[108] **Kay, Will, et al.** "The kinetics human action video dataset." arXiv preprint arXiv:1705.06950 (2017).
[109] **Chen, Chen, Roozbeh Jafari, and Nasser Kehtarnavaz.** "UTD-MHAD: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor." 2015 IEEE International conference on image processing (ICIP). IEEE, 2015.
[110] **Goyal, Raghav, et al.** "The" Something Something" Video Database for Learning and Evaluating Visual Common Sense." ICCV. Vol. 1. No. 4. 2017.
[111] **Sigurdsson, Gunnar A., et al.** "Hollywood in homes: Crowdsourcing data collection for activity understanding." European Conference on Computer Vision. Springer, Cham, 2016.

[112] **Zhang, Pengfei, et al.** "Semantics-Guided Neural Networks for Efficient Skeleton-Based Human Action Recognition." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[113] **Parisi, German I.** "Human Action Recognition and Assessment via Deep Neural Network Self-Organization." arXiv preprint arXiv:2001.05837 (2020).
[114] **Jeon, Yunho, and Junmo Kim.** "Constructing fast network through deconstruction of convolution." Advances in Neural Information Processing Systems. 2018.
[115] **Wu, Bichen, et al.** "Shift: A zero flop, zero parameter alternative to spatial convolutions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[116] **Zhong, Huasong, et al.** "Shift-based primitives for efficient convolutional neural networks." arXiv preprint arXiv:1809.08458 (2018).
[117] **Li, Chao, et al.** "Collaborative spatiotemporal feature learning for video action recognition." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[118] **Yan, Sijie, Yuanjun Xiong, and Dahua Lin.** "Spatial temporal graph convolutional networks for skeleton-based action recognition." arXiv preprint arXiv:1801.07455 (2018).
[119] **Shi, Lei, et al.** "Skeleton-based action recognition with directed graph neural networks." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019.
[120] **Cheng, Ke, et al.** "Skeleton-Based Action Recognition With Shift Graph Convolutional Network." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[121] **Zhu, Linchao, et al.** "FASTER Recurrent Networks for Efficient Video Classification." AAAI. 2020.
[122] **Kay, Will, et al.** "The kinetics human action video dataset." arXiv preprint arXiv:1705.06950 (2017).
[123] **Carreira, Joao, et al.** "A short note about kinetics-600." arXiv preprint arXiv:1808.01340 (2018).
[124] **Gu, Chunhui, et al.** "Ava: A video dataset of spatio-temporally localized atomic visual actions." Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018.
[125] **Feichtenhofer, Christoph, et al.** "Slowfast networks for video recognition." Proceedings of the IEEE international conference on computer vision. 2019.
[126] **Caba Heilbron, Fabian, et al**. "Activitynet: A large-scale video benchmark for human activity understanding." Proceedings of the ieee conference on computer vision and pattern recognition. 2015.
[127] **Wang, Limin, et al.** "Temporal segment networks for action recognition in videos." IEEE transactions on pattern analysis and machine intelligence 41.11 (2018): 2740-2755.d
[128] **Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.** BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805, 2018.
[129] **Luowei Zhou, Chenliang Xu, and Jason J Corso.** Towards automatic learning of procedures from web instructional videos. In AAAI, 2018.
[130] **Sun, Chen, et al.** "Videobert: A joint model for video and language representation learning." Proceedings of the IEEE International Conference on Computer Vision. 2019.

[131] **Zhu, Linchao, and Yi Yang.** "ActBERT: Learning Global-Local Video-Text Representations." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2020.
[132] **Miech, Antoine, et al.** "Howto100m: Learning a text-video embedding by watching hundred million narrated video clips." Proceedings of the IEEE international conference on computer vision. 2019.

EKLER

EK 1: Veri setlerinin detayları ile birlikte karşılaştırılması.

EK 2: Çalışmalarda kullanılan eğitim ve test setlerinin dağılımı.

EK 3: PredNet algoritmasının örnek sonuçları.

EK 4: Makale [86] yönteminin örnek sonuçları

EK 1

Çizelge Ek.1: Veri setlerinin detayları ile birlikte karşılaştırılması.

Veri Seti	Kaynak	Oluşturan	Video Sayısı	Çözünürlük	Fps	Kategoriler	Kategori Sayısı	Özne Sayısı	Video Süresi
FaceGen [1][2]	FaceGen Yazılım Paketiyle elde edilen yüz şekilleri	Singular Inversions ve William Lotter, Gabriel Kreiman ve David Cox	-	-	-	z ekseninde 7 ve x ekseninde 8 farklı şekilde elde edilen yüz şekillleri	56 (yönlendirme)	25 (yüz sayısı)	-
KITTI [3]	Araçlara sabitlenmiş kamera	Karlsruhe Teknoloji Enstitüsü (KIT) ve Toyota Teknoloji Enstitüsü (TTI-C)	156	1392x512	10	Şehir, Yerleşim yeri, Yol, Kampüs, İnsan	5	-	-
CalTech Pedestrian [4]	Los Angeles'da sürülen araca sabitlenmiş kamera	Piotr Dollar, Christian Wojek, Bernt Schiele ; Pietro Perona	-	640x480	-	-	-	2300	1s - 60s
TV Human Action Interaction [5]	20 farklı TV programından elde edilmiş videolar	Oxford Üniversitesi Görsel Geometri Grubu.	300	Variable	-	El Sıkışma, Sarılma, Beşlik Çakma ve Öpüşme	4	-	1-27s
KTH [6]	Kapalı ve açık alan videoları	Christian Schuldt, Ivan Laptev ve Barbara Caputo	600	120x160	25	Yürüme, Koşu yapma, Koşma, Boks yapma, El Sallama ve El Çırpma	6	25	4s
UCF - ARG [7]	Kingfisher Aerostat helyum balonuna sabitlenmiş kamera, yer kamerası ve çatı kamerası	Central Florida Üniversitesi	1440	1920x1080	60	Boks yapma, Taşıma, El Çırpma, Kazma, Koşu yapma, Bagaj aç-kapa, Koşma, Fırlatma, Yürüme ve Sallanma	10	12	1s - 104s
Youtube - Aerial [8]	YouTube'dan alınmış uçangöz videoları	Waqas Sultani ve Mubarak Shah	500	-	-	Bisiklet sürme, Golf, At Binme, Koşma, Sörf yapma, Yürüme, Yüzme, Yamaç dalışı, Kano sporu, Kaykay yapma	10	-	-

Çizelge Ek.1: Veri setlerinin detayları ile birlikte karşılaştırılması. (devam)

Weizmann [10]	Arka planı olan açık alanda çekilmiş videolar	Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani and Ronen Basri	90	180x144	50	Koşma, Yürüme, Zıplama, Bükme, Sallanma	10	9	~3s
HMDB51 [11]	Youtube, Google ve Prelinger arşivinden elde edilen videolar	H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre	7000	Değişken	30	Genel yüz aksiyonları, Obje etkileşimli yüz aksiyonları, Genel vücut hareketleri, Obje etkileşimli vücut hareketleri, İnsan etkileşimli vücut hareketleri	51	-	2-3s
UCF-101 [9]	Youtube'daki gerçekçi videolar	Khurram Soomro, Amir Roshan Zamir ve Mubarak Shah	13320	320x240	25	İnsan-Obje Etkileşimi, Yalnızca Vücut Hareketleri, İnsan-İnsan Etkileşimleri, Müzik Enstrümanı Çalma, Spor	101	-	1.06 - 71.04 s
CUHK Avenue [57]	Aynı lokasyonda hareketli obje videoları	Hong Kong Çin Üniversitesi'nden Cewu Lu, Jianping Shi, Jiaya Jia	37	640x360	25	-	-	-	1 - 60s
Shanghai Tech Campus [58]	Karmaşık ışık koşulları ve kamera açılarıyla elde edilmiş videolar	Shanghai Tech Üniversitesi	130	-	-	-	-	-	-
CAD - 60 [59]	Kinect sensörüyle elde edilmiş videolar	Cornell Üniversitesi Robot Öğrenme Laboratuvarı	60	320x240	-	Ağız çalkalama, Diş fırçalama, lens takma, telefonda konuşma, su içme, ilaç kutusu açma, yemek yapma, Rinsing Mouth, Brushing Teeth, koltukta oturma, yazı yazma, bilgisayarda çalışma	12	4	-
CAD - 120 [59]	Kinect sensörüyle elde edilmiş günlük yaşam videoları	Cornell Üniversitesi Robot Öğrenme Laboratuvarı	120	640x480	-	Mısır gevreği hazırlama, ilaç içme, obje istifleme, objeyi alma, mikrodalgada yemek, obje temizleme, yemek alma, obje hizalama, yemek yeme	10	4	-

Çizelge Ek.1: Veri setlerinin detayları ile birlikte karşılaştırılması. (devam)

MSR Daily Activity 3D [60]	Kinect sensörüyle elde edilmiş oturma odası videoları	Microsoft	320	-	-	İçme, Yeme, Kitap okuma, Telefon çalma, Laptop kullanma, Yazı yazma, Süpürge kullanma, Sevinme, Kâğıt fırlatma, gitar çalma, Oyun oynama, Yürüme, Oturma, Kalkma	16	-	-
NTURGB+D [61]	3 farklı Microsoft Kinect V2 kamerasıyla elde edilen videolar.	Amir Shahroudy, Jun Liu, Tian-Tsong Ng, Gang Wang	56880	1920x1080	30	Günlük aksiyonlar	60	40	~1-10s
UT - Interaction [62]	Park alanındaki insanlar arası etkileşim içeren videolar	Ryoo, M. S. ve Aggarwal, J. K.	120	720x480	30	Tokalaşma, İtme, Tekme atma, Sarılma, Yumruk atma, Doğrulma	6	15	~60s
Hollywood2 [104]	Filmlerden alınan videolar	I. Laptev, M. Marszalek, C. Schmid ve B. Rozenfeld	3669	-	24	Telefona cevap verme, araç sürme, yemek yeme, kavga etme, arabadan inme, el sıkışma, sarılma, öpüşme, koşma, oturma, kalkma,ayakta durma	12	-	-
MSR Action 3D [105]	Derinlik haritası da dahil olmak üzere uzuv hareketlerini içeren bir veri seti	Li, W.; Zhang, Z.; Liu, Z.	4020	640x480	15	Yüksek kol dalgası, yatay kol dalgası, çekiç, elle yakalama, ileri yumruk, yüksek atış, x çek, kene çek, daire çiz, el çırpma, iki el dalgası, yandan boks, eğil, ileri tekme, yandan tekme, koşu, tenis salıncak, tenis servisi, golf salıncak, alma ve atma	20	7	-
Northwestern UCLA [106]	Kinect kameralarından alınan RGB, derinlik ve insan iskeleti verileri	Wang, J.; Nie, X.; Xia, Y.; Wu, Y.; Zhu, S.C	-	-	-	Tek elle toplama, iki eliyle kaldırma, çöpü atma, dolaşma, oturma, ayağa kalkma, takma, çıkarma, atma, taşıma	10	10	-
Olympic Sports [107]	Sporcuların farklı spor videoları	Niebles, J.C.; Chen, C.W.; Li, F.F	800	-	-	Yüksek atlama, uzun atlama, üçlü atlama, sırıkla atlama, basketbol dizilişi, bowling, tenis servisi, platform, çekiç, cirit, gülle atma, sıçrama tahtası, koparma	16	-	-

Çizelge Ek.1: Veri setlerinin detayları ile birlikte karşılaştırılması. (devam)

Kinetics [108]	YouTube'dan insan-nesne ve insan-insan etkileşimi videoları	Deepmind	650000	Değişken	Değişken	Seçilen veri seti sürümüne göre 400/600/700 farklı eylem	400/600/700	-	~10s
UTD-MHAD [109]	Kinect kamera ve giyilebilir atalet sensöründen alınan videolar	Chen Chen, Roozbeh Jafari, ve Nasser Kehtarnavaz	861	640x480 ve 320x240	30	Sağ kol sola kaydırma, sağ kol sağa kaydırma, iki el ön alkış, basketbol şut, sağ el çekme x, bowling (sağ el), ön boks, tenis sağ el forehand vuruş, yerinde koşu vb.	27	8	-
Something - Something V2 [110]	Yoğun etiketli videolar	Twentybn	220847	320x240	-	Günlük nesnelerle önceden tanımlanmış temel eylemler gerçekleştiren insanlar	174	-	~4.03 s
Charades [111]	Amazon Mechanical Turk aracılığıyla toplanan videolar	Perceptual Reasoning an Interaction Research team of Allen Institute for AI	9848	Değişken	Değişken	Günlük iç mekân aktiviteleri	157	267	~30.1s

EK 2

Çizelge Ek.2: Çalışmalarda kullanılan eğitim ve test setlerinin dağılımı.

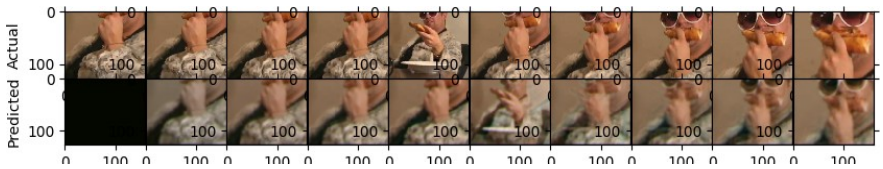
Çalışma	Veri Seti	Eğitim Videolarının Sayısı	Doğrulama Videolarının Sayısı	Test Videolarının Sayısı
Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning [1]	KITTI Veri Seti [3]	57	4	0
	CalTech Pedestrian Veri Seti [4]	0	0	Tüm Videolar
Anticipating Visual Representations from Unlabeled Video [13]	THUMOS 2015 Veri Seti [56]	25 - Katlamalı Çapraz Doğrulama		
	TV Human Interaction Veri Seti [5]			
Fast Temporal Activity Proposals for Efficient Detection of Human Actions in Untrimmed Videos [12]	KTH Veri Seti [6]	200	0	0
	THUMOS 2014 Detection Challenge Veri Seti [55]	0	0	213
Human Action Recognition in Drone Videos Using a Few Aerial Training Examples [8]	UCF-ARG * [7]	~28	~4	~14
	Youtube Aerial [8]	30	5	15
AdaScan: Adaptive Scan Pooling in Deep convolutional Neural Networks for Human Action Recognition in Videos [17]	UCF-101 [9]	VGG-16'dan Uzamsal ağ [35] model ImageNet [63] üzerinde eğitilmiş, 6 dönem çalışır. Temporal ağ eğitimi, 2 dönem çalışan 16000 yineleme anlık görüntüsü ile [64] evrişimli katmanları başlatır.		
	HMDB-51 [11]	Hem Uzamsal hem de Zamansal Ağ, 6 dönem çalışan, eğitimli UCF-101 ağından evrişimli katman ağırlıklarını başlatır.		
Predicting Future Frames using Retrospective Cycle GAN [14]	KITTI Veri Seti [3]	41000	0	4100
	CalTech Pedestrian Veri Seti [4]	-	-	
	UCF-101 [9]	11988	-	1332
	CUHK Avenue [57]	-	-	-
	ShanghaiTech Campus Veri Seti [58]	-	-	-
Pose Primitive Based Human Action Recognition in Videos or Still Images [20]	Weizmann Veri Seti [10]	Leave - One - Out Çapraz Doğrulama (eğitim videoları arka planları yok sayılıyor.)		
	KTH Veri Seti [6]	0	0	36 (her kategoriden 6 örnek)

Çizelge Ek.2: Çalışmalarda kullanılan eğitim ve test setlerinin dağılımı. (devam)

A New Hybrid Architecture for Human Activity Recognition from RGB-D Videos [15]	CAD-60 [59]	Tüm veri setlerindeki ayırma, leave-one-person-out şemasına göre veya veri setinde belirtildiği şekilde gerçekleştirildi.		
	CAD-120 [59]			
	MSRDailyActivity 3D [60]			
	NTURGB+D [61]			
Human Activity Prediction: Early Recognition of Ongoing Activities from Streaming Videos [16]	UT-Interaction Veri Seti [62]	10 - Katlamalı Çapraz Doğrulama		
Lattice Long Short-Term Memory for human Action Recognition [18]	HMDB-51 [11]	Veri Seti, uygulanacak yönteme göre 3 farklı bölünme durumuna sahiptir.		
	UCF-101 [9]			
Human Action Recognition in Videos Using Kinematic Features and Multiple Instance Learning [19]	Weizmann Veri Seti [10]	Leave - One - Actor - Out Ayarı		
	KTH Veri Seti [6]			
Skeleton-Based Action Recognition With Shift Graph Convolutional Network [120]	NTURGB+D [61]	20 denekten seçilen eğitim verileri, diğer 20 denekten seçilen test verileri.		
	NTU-120 RGB+D [61]	53 denekten seçilen eğitim verileri, diğer 53 denekten seçilen test verileri.		
	Northwestern UCLA [106]	İlk iki kameradan seçilen eğitim verileri için örnekler, diğer kamera test verileri olarak kullanıldı.		
FASTER Recurrent Networks for Efficient Video Classification [121]	Kinetics [108]	-	-	-
	UCF-101 [9]	Kinetics ile önceden eğitilmiş veriler kullanılmıştır.		
	HMDB-51 [11]	Kinetics ile önceden eğitilmiş veriler kullanılmıştır.		
Slowfast networks for video recognition [125]	Kinetics - 400 [122]	~240k	20k	0
	Kinetics - 600 [123]	~392k (eğitim için kullanılmamıştır.)	30k	0
	Charades [111]	~9.8k (eğitim için kullanılmamıştır.)	1.8k	0
	AVA [124]	211k (eğitim için kullanılmamıştır.)	57k	0
Temporal segment networks for action recognition in videos [127]	UCF-101 [9]	13320	1010	1575
	HMDB-51 [11]	-	-	-
	ActivityNet [126]	4819	2383	2480
	THUMOS 2014 Detection Challenge Veri Seti [55]	-	-	-

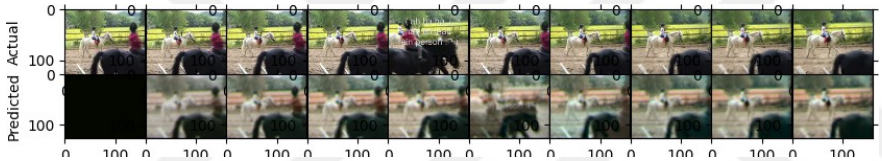
EK 3

Kıyaslama sonuçlarını elde ettiğimiz bazı makalelerin çıktıları aşağıdaki resimlerde mevcuttur. PredNet [1] yönteminin bazı tahmin sonuçları verilmiştir.



Şekil Ek.1: PredNet algoritmasında "yemek yeme" aksiyonunun tahmini.

Şekil Ek.2: PredNet algoritmasında "içecek içme" aksiyonunun tahmini.



Şekil Ek.3: PredNet algoritmasında "at binme" aksiyonunun tahmini. Burada iki farklı at olmasına rağmen atlar karıştırılmadan doğru tahmin edilmiştir.

Şekil Ek.4: PredNet algoritmasında "saç kurutma" aksiyonunun tahmini.

Şekil Ek.5: PredNet algoritmasında "öpmek" aksiyonunun tahmini. Burada son kare yanlış tahmin edilmiştir.

EK 4

[86] 'daki yöntemle, 3 farklı videoya tahmin işlemi örnekleri ekledik. Her video için en yüksek olasılığa sahip 5 tahmin algoritma tarafından verilmektedir. Bebeğin emeklemesi ve kriket oynamasıyla ilgili sonuçlar son derece tatmin edici olsa da at yarışı için aynı durumu söyleyemeyiz. Bu nedenle, at yarışı eylemi için ilk 5 tahminde çok farklı tahminler üreten bu modelin hatalı olarak tespit edilen bir eylemini gösterdik.

Çizelge Ek.3: Kriket oynama eyleminin tahmin sonuçları. (v_CricketShot_g04_c02.avi)

Olasılıklar	Eylemler
%97.77	Kriket oyanama
%0.71	Kaykay kayma
%0.56	Robot dansı
%0.56	Patenle kayma
%0.13	Golf egzersizi

Çizelge Ek.4: Bebek emeklemesi eyleminin tahmin sonuçları.(v_BabyCrawling_g01_c01.avi)

Olasılıklar	Eylemler
%99.99	Bebek emeklemesi
%0.01	Kahakaha atma
%0.00	Kafa sallama
%0.00	Araba itme
%0.00	Bebek uyanması

Çizelge Ek.5: At yarışı eyleminin tahmin sonuçları. (v_HorseRace_g02_c04.avi)

Olasılıklar	Eylemler
%44.23	Koşu yapmak
%29.26	Yürüyüş yapmak
%9.43	Gangnam Style dansı
%8.29	At sürme veya at ile yürüme
%2.66	Yumruk atmak