



**VİDEOLARIN KURGUSU VE ÖZETLENMESİ İÇİN DERİN ÖĞRENME
İLE HAREKET ANALİZİ**

Ahmet Samet HALICI

**YÜKSEK LİSANS TEZİ
ELEKTRİK ELEKTRONİK MÜHENDİSLİĞİ ANA BİLİM DALI**

**GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ**

ŞUBAT 2022

ETİK BEYAN

Gazi Üniversitesi Fen Bilimleri Enstitüsü Tez Yazım Kurallarına uygun olarak hazırladığım bu tez çalışmada;

- Tez içinde sunduğum verileri, bilgileri ve dokümanları akademik ve etik kurallar çerçevesinde elde ettiğimi,
- Tüm bilgi, belge, değerlendirme ve sonuçları bilimsel etik ve ahlak kurallarına uygun olarak sunduğumu,
- Tez çalışmada yararlandığım eserlerin tümüne uygun atıfta bulunarak kaynak gösterdiğimi,
- Kullanılan verilerde herhangi bir değişiklik yapmadığımı,
- Bu tezde sunduğum çalışmanın özgün olduğunu,

bildirir, aksi bir durumda aleyhime doğabilecek tüm hak kayıplarını kabullendiğimi beyan ederim.

Ahmet Samet HALICI

21/02/2022

VİDEOLARIN KURGUSU VE ÖZETLENMESİ İÇİN DERİN ÖĞRENME İLE HAREKET ANALİZİ

(Yüksek Lisans Tezi)

Ahmet Samet HALICI

GAZİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Şubat 2022

ÖZET

Sosyal medya platformlarında düzenlenmiş video paylaşımlarının ve içerik tüketimlerinin hızla artması ile video düzenleme üzerinde yoğun olarak çalışılan aktif bir konu haline gelmiştir. Bu tez çalışmasında videoların özetlenebilmesi için videodaki kişilerin hareketlerinin poz tahmini algoritmalarının çıktılarını kullanan bir takip yöntemi geliştirilerek anlamlandırılması ve en önemli sahnelerin parçalanarak bir arka plan müziğine göre birleştirilmesi amaçlanmıştır. Literatürdeki çeşitli poz tahmini algoritmalarının geliştirilen poz takibi yöntemine etkisi ile farklı ön işleme adımlarının poz takibi başarımına katkıları PoseTrack veri kümesi kullanılarak incelenmiştir. Kişi takibinin sonucunda elde edilen zamana bağlı eklem noktası dizilerini kullanarak derin öğrenme yöntemleri ile hareketleri tanıma işlemi gerçekleştirilmiştir. Ayrıca video içeriğindeki muhtemel başrol oyuncularını tespit edilerek karakterlerin davranışlarına ve belirlenen kurguya göre en değerli sahnelerin belirlenmesi için analizler yapılmıştır. Tüm bu çalışmaların yapılabilmesi için insansız hava aracı ile kaydedilen 396 037 görüntüden oluşan yeni bir video veri kümesi meydana getirilmiştir. Veri kümesi ile ilgili detaylara [posedrone.com](https://www.posedrone.com) sayfasından erişilebilir. Oluşturulan veri kümesi üzerinde 7 hareket sınıfı için 33 083 adet kişinin eklem noktalarının etiketlenmesi yapılmıştır. Görüntülerin ham hali ve OpenPose kullanılarak etiketlenen doğru veriler [posedrone.com](https://www.posedrone.com) sayfasında yayınlanmıştır. Poz tahmini yöntemleri için geliştirilen ön işleme adımları; PoseTrack veri kümesi üzerinde MOTA değerinde DCPose için %21.2, OpenPose için %22.6 iyileştirme sağlamıştır. Hareket tanıma için tasarlanan modele veri kalitesini artırmak için uygulanan işlemlerinin de olumlu katkı sağladığı gösterilmiştir. Video özetleme için poz tahmini, poz takibi, hareket tanıma ve en değerli sahne algoritmalarının birbirinden bağımsız olarak güncellenebildiği bir alt yapı meydana getirilmiştir. Müzik sinyalindeki vuruşların zamanları tespit edilerek sahneler arası geçiş yapan yeni bir video özetleme yöntemi geliştirilmiştir. Literatürde bir ilk olan hem hareket tanıma hem de video özetleme alanında birlikte kullanılabilecek insansız hava aracı ile kaydedilen veri kümesi yayınlanmıştır.

Bilim Kodu : 90541

Anahtar Kelimeler : Video özetleme, poz tahmini, gerçek zamanlı poz takibi, iskelet tabanlı eylem tanıma

Sayfa Adedi : 83

Danışman : Dr. Öğr. Üyesi Ayşe DEMİRHAN

ACTION ANALYSIS WITH DEEP LEARNING FOR EDITING AND SUMMARIZING VIDEOS

(M. Sc. Thesis)

Ahmet Samet HALICI

GAZİ UNIVERSITY

GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCES

February 2022

ABSTRACT

Video editing has become an active topic that is being studied intensively with the rapid increase in video shares and content consumption organized on social media platforms. In this thesis study, in order to convert the videos into clips, it is aimed to make sense of the actions of the people in the video by developing a tracking method using the outputs of pose estimation algorithms and to combine the most important scenes. The effects of various pose estimation algorithms in the literature on the developed pose tracking method and the contribution of different preprocessing steps to the pose tracking performance were examined using the PoseTrack dataset. Recognition of actions was carried out using deep learning methods and using time dependent keypoint sequences obtained as a result of pose tracking. In addition, possible lead actors in the video content were identified and analyzes were made to determine the most valuable scenes according to the behaviors of the characters and the determined fiction. In order to carry out all these studies, a new video dataset consisting of 396 037 images recorded with by unmanned aerial vehicle has been created. Details about the dataset are available at posedrone.com. On the created dataset, the joint points of 33 083 people were labeled for 7 action classes. The preprocessing steps developed for pose estimation methods provided an improvement of 21.2% for DCPose and 22.6% for OpenPose in the value of MOTA on the PoseTrack dataset. It has been shown that the processes applied to increase the data quality of the model designed for action recognition also make a positive contribution. For video summarization, an infrastructure has been created in which pose estimation, pose tracking, action recognition and most valuable scene algorithms can be updated independently of each other. A new video summarization method has been developed that switches between scenes by deciphering the times of beats in the music signal. A set of data recorded by an unmanned aerial vehicle that can be used together in the field of both action recognition and video summarization, which is the first in the literature, has been published.

Science Code : 90541

Key Words : Video summarization, pose estimation, real-time pose tracking, skeleton based action recognition

Page Number : 83

Supervisor : Assist. Prof. Dr. Ayşe DEMİRHAN

TEŞEKKÜR

Çalışmalarım boyunca değerli yardım ve katkılarıyla ben yönlendiren Hocam Dr. Ayşe DEMİRHAN'a, kıymetli bilgi, tecrübe ve ekipmanlarından faydalandığım iş arkadaşlarım Orhan AYRAN, Barış SATAR ve Tarık UMUTLU'ya teşekkür ederim. Desteğini hiçbir zaman esirgemeyen eşim Rüya HALICI'ya çok teşekkür ederim. Öğretilelerle dolu bir hayat yaşamamı sağlayan ve emeklerinin ödenmesi mümkün olmayan ağabeyime, anneme ve babama saygı ve sevgilerimi sunarım. Ayrıca bu tez çalışması, G.Ü. Bilimsel Araştırma Projeleri tarafından FGA-2021-7113 kodu ile desteklenmiştir. Destek programının tasarlanmasında emeği geçen herkese de teşekkür ederim.

İÇİNDEKİLER

	Sayfa
ÖZET	iv
ABSTRACT.....	v
TEŞEKKÜR.....	vi
İÇİNDEKİLER	vii
ÇİZELGELERİN LİSTESİ.....	ix
ŞEKİLLERİN LİSTESİ.....	x
SİMGELER VE KISALTMALAR.....	xiii
1. GİRİŞ.....	1
2. MATERYAL ve METHOD	11
2.1. Derin Öğrenme	11
2.1.1. Makine öğrenmesi	11
2.1.2. Veri ön işleme	13
2.1.3 Yapay sinir ağı	14
2.1.4 Özyinelemeli sinir ağı	18
2.2. PoseTrack Veri Kümesi	19
2.2.1. Performans ölçütleri	20
2.3. Poz Tahmini	22
2.3.1. OpenPose yöntemi	22
2.3.2. DCPose yöntemi.....	23
2.4. Poz Takibi Ön İşleme Adımları	23
2.5. Poz Takibi	27
2.5.1. Kalman filtresi.....	27
2.5.2. Veri ilişkilendirme	31

	Sayfa
2.6. Hareket Tanıma	32
2.6.1. Veri etiketleme	35
2.6.2. Veri hazırlama işlemleri	38
2.6.3. Derin öğrenme modeli.....	39
2.6.4. Yapay hareket alt sınıfları	41
2.6.5. Hedef hareket sınıfları	42
2.6.6. Hareket sınıflarının birleştirilmesi	44
2.7. Müzik İşleme ve Vuruş Tespit Algoritması	44
2.8. Video Özetleme	48
2.8.1. Video özetleme için ön işlem adımları.....	51
2.8.2. Bellman denklemi ile videonun keşfedilmesi	53
2.8.3. Videoların parçalanması.....	53
2.8.4. Videoların birleştirilmesi	55
2.9. PoseDrone Veri Kümesi.....	56
3. DENEYSEL SONUÇLAR.....	61
3.1. Poz Tahmini ve Takibi ile İlgili Çalışmalar.....	61
3.2. Hareket Tanıma ile İlgili Çalışmalar	63
3.3. Müzik Sinyalinde Vuruş Tespit Çalışmaları.....	66
3.4. Video Özetleme ile İlgili Çalışmalar.....	67
4. SONUÇ VE ÖNERİLER.....	71
KAYNAKLAR	75
ÖZGEÇMİŞ	83

ÇİZELGELERİN LİSTESİ

Çizelge	Sayfa
Çizelge 2.1. PoseTrack veri kümesini oluşturmada kullanılan kaynaklar ve videoların özellikleri.....	19
Çizelge 2.2. Her bir hareket sınıfı için elde edilen veri etiketi sayıları.....	39
Çizelge 2.3. Yapay hareket alt sınıfları.....	41
Çizelge 3.1. Ön işleme adımlarının poz takibi performansına etkisi	61
Çizelge 3.2. Her bir video için görüntü önem değerlerinin hesaplanması için oluşturulan tablodan 9 görüntülük bir örnek.....	67

ŞEKİLLERİN LİSTESİ

Şekil	Sayfa
Şekil 1.1. Tez çalışmasının genel blok diyagramı	2
Şekil 1.2. Geliştirilen poz takibi yönteminin yapısı.....	4
Şekil 1.3. Hareket tanıma için eklem noktaları arasında belirlenen bazı ilişkiler.....	5
Şekil 1.4. Tez çalışması için kaydedilen veri kümesi PoseDrone'dan bir görüntü.....	7
Şekil 2.1. Makine öğrenmesinin üç farklı alt başlığı	11
Şekil 2.2. Denetimli öğrenmenin genel yapısı	12
Şekil 2.3. n girişli bir nöron	14
Şekil 2.4. (a) Sigmoid fonksiyonu, (b) Tanh fonksiyonu, (c) ReLU fonksiyonu	15
Şekil 2.5. İki katmanlı bir sinir ağı	17
Şekil 2.6. Standart bir UKSB hücresi	19
Şekil 2.7. PoseTrack veri kümesi örnek görüntüleri ile elde edilen poz tahminleri	20
Şekil 2.8. OpenPose'un genel işlem adımları	22
Şekil 2.9. DCPose yönteminin genel akış diyagramı.....	23
Şekil 2.10. Poz tahmini algoritmalarının çıktılarındaki aykırı değerlerin tespit edilmesi ve düzeltilmesi adımı.....	24
Şekil 2.11. Poz tahmininden elde edilen eklem noktaları ile baş bölgesinin bulunması	25
Şekil 2.12. Baş pozisyonlarına göre sadece poz tahmini ile hesaplanan baş bölgeleri ve tespit edilmesi gereken baş bölgeleri	26
Şekil 2.13. Kalman filtresi akış diyagramı.....	28
Şekil 2.14. Bir kişinin video boyunca merkez noktasının sütun değerlerinin gerçek verilerle poz tahmini ve poz takibi çıktılarının karşılaştırılması	30
Şekil 2.15. PoseC3D'nin hareket tanıma için kullandığı model ve genel yapısı.....	33
Şekil 2.16. Eklem noktaları arasında hesaplanan açılı ilişkileri.....	34
Şekil 2.17. Veri etiketleme adımının blok diyagramı	36

Şekil	Sayfa
Şekil 2.18. PoseDrone veri kümesinden diz çökme hareketinin etiketlendiği sahneden bir görüntü.....	37
Şekil 2.19. 33 bin adet veri etiketleme işlemi yapılırken doldurulan dosya kağıtları.....	38
Şekil 2.20. Tek bir UKSB katmanı ile sınıflandırma.....	40
Şekil 2.21. El ele tutuşma ve yürüyüş sınıflarının örnek gösterimi	42
Şekil 2.22. Diz çökme hareket sınıfı örnek gösterimi.....	43
Şekil 2.23. Dans hareket sınıfı örnek gösterimi	43
Şekil 2.24. Sabit duruş hareket sınıfı örnek gösterimi	44
Şekil 2.25. Etkinlik videosunda eş zamanlı kaydedilen yerdeki sol kamera görüntüsü .	45
Şekil 2.26. Farklı frekans bantlarına bölünmüş örnek bir müzik sinyali	46
Şekil 2.27. Örnek müzik sinyalinin FFT'si ile Yarım Hanning Penceresinin FFT'sinin çarpılması sonucu elde edilen sinyal.....	47
Şekil 2.28. Farkı alınan ve yarım dalga doğrultucu ile doğrultulan örnek bir müzik sinyali.....	47
Şekil 2.29. İçerik tabanlı video özetleme yaklaşımlarının alt dalları.....	49
Şekil 2.30. Video özetleme yaklaşımlarının genel yapısı	49
Şekil 2.31. Görüntünün merkez bölgesi ve başrol oyuncularının ilişkisi	52
Şekil 2.32. Video parçalama adımının girdi ve çıktıları	54
Şekil 2.33. Örnek bir müzik sinyalindeki vuruş zamanlarına göre farklı sahneler arası geçişlerin gerçekleştirilme örneği	55
Şekil 2.34. PoseDrone veri kümesini kaydederken kullanılan ekipmanlar	57
Şekil 2.35. Etkinlik videosunda eş zamanlı kaydedilen yerdeki sol kamera görüntüsü .	58
Şekil 2.36. Etkinlik videosunda eş zamanlı kaydedilen insansız hava aracının kamera görüntüsü	58
Şekil 2.37. Etkinlik videosunda eş zamanlı kaydedilen yerdeki sağ kamera görüntüsü.	59
Şekil 3.1. Örnek birer videoda OpenPose ile videonun (a) 22. ve (b) 40. karelerinde, DCPose ile videonun (c) 53. ve (d) 77. karelerinde bulunan kişilerin takip sonuçları.....	62

Şekil	Sayfa
Şekil 3.2. Ön işleme adımlarının her bir videonun poz takip süresinin videodaki görüntü sayısına bölünmesi ile elde edilen ortalama işlem sürelerine etkisi ..	63
Şekil 3.3. Diz çökme hareketi için etiketlenmiş tüm kişilerin sağ ve sol dizinde meydana gelen açı değerlerinin filtrelenmiş ve filtrelenmemiş değerleri.....	64
Şekil 3.4. Sağ ve sol dizdeki açı değerlerinin filtrelenmemiş ve filtrelenmiş dağılımı ..	65
Şekil 3.5. Sağ ayak ve sağ diz ile sağ diz ve sağ bel arasındaki uzunlukların filtrelenmemiş ve filtrelenmiş gösterimleri.....	66
Şekil 3.6. Bir müzik sinyalindeki vuruş zamanlarının tespitinin gösterimi.....	66
Şekil 3.7. Çizelge 3.2’de değerleri paylaşılan bir video sahnesinin görüntüleri.....	68
Şekil 3.8. 2088 adet görüntüden oluşan bir videonun görüntülerinin önem puanları.....	69
Şekil 3.9. Bir etkinlikteki bazı sahnelerin en değerli görüntüleri	69
Şekil 3.10. Derin öğrenme modelinin eğitimi.....	70

SİMGELER VE KISALTMALAR

Bu çalışmada kullanılmış simgeler ve kısaltmalar, açıklamaları ile birlikte aşağıda sunulmuştur.

Simgeler

Açıklamalar

w	Ağırlık gösterimi
x	Giriş değeri
y	Çıkış değeri
σ	Sigmoid fonksiyonu
l_2	Bir tür kayıp fonksiyonu
t	Zaman

Kısaltmalar

Açıklamalar

2B	İki Boyutlu
3B	Üç Boyutlu
3B Mask R-CNN	Üç Boyutlu Maske Bölgesel Evrişimli Sinir Ağı
col	Sütun değeri
DCPose	Deep Dual Consecutive Pose (Derin Çift Ardışık Poz)
DVS	Dakikadaki Vuruş Sayısı
enb	En büyük
enk	En küçük
ESA	Evrişimli Sinir Ağı
Eş.	Eşitlik
FFT	Fast Fourier Transform
GEA	Grafik Evrişimli Ağlar
HRNet	High-Resolution Network (Yüksek Çözünürlüklü Ağ)
json	JavaScript Object Notation (bir tür dosya formatı)
KEYK	Küresel En Yakın Komşu
LIDAR	Laser Imaging Detection and Ranging

Kısaltmalar**Açıklamalar****MOTA**

Çok kişili nesne takibi doğruluğu

MOTP

Çok kişili nesne takibi hassasiyeti

ÖSA

Özyinelemeli Sinir Ağı

ReLU

Rectified Linear Unit (Düzeltilmiş Doğrusal Birim)

RGB

Kırmızı Yeşil Mavi

row

Satır değeri

Tanh

Tanjant Hiperbolik

TUM

Münih Teknik Üniversitesi

UKSB

Uzun Kısa Süreli Bellek

VATIC

Video Annotation Tool (Video Etiketleme Aracı)

VidCon

VidCon isimli çevrimiçi video konferans

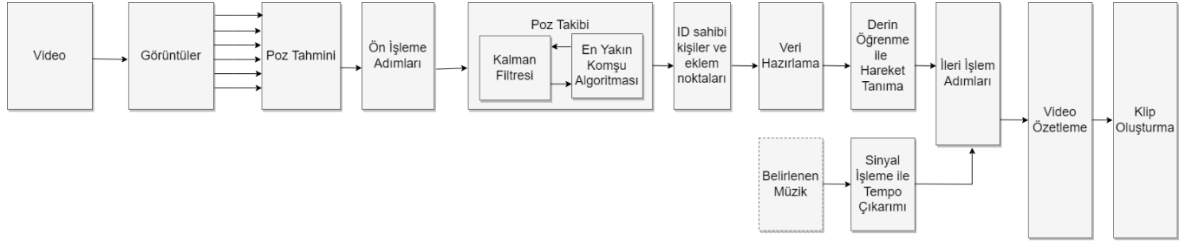
YZ

Yapay Zeka

1. GİRİŞ

1960'lı yıllarda yapay zekâ alanında ilk gelişmeler ve akademik çalışmalar ortaya çıkmaya başlamıştır. Bazı araştırmacılar yapay zekâ alanının önünün açık olduğuna inanıyor ve insan kadar zeki makinelerin ortaya çıkmasını 1990 yıllarını geçmeyeceğini düşünüyorlardı. Fakat günümüzde birçok bilgisayarlı görü probleminde halen istenen başarılar elde edilememiştir. Bilgisayarlı görü problemleri; kameraların bakış açıları, üretimden kaynaklı kalibrasyon değerlerinin farklılıkları, aydınlatma koşullarındaki değişkenlik gibi farklılıklar nedeniyle zordur. İnsan hareketlerini videoları kullanarak analiz etmek; arka plan görüntüsündeki hızlı değişimler, renk değerlerinin farklılıkları gibi nedenlerle zor olduğu için daha dayanıklı çözümlere ihtiyaç duyulmaktadır. Bu dayanıklı çözümler; video kayıt cihazlarındaki gelişmelere bağlı olarak kaydedilen görüntülerdeki bilgi detaylarının artması, işlem gücü yüksek donanımların geliştirilmesi ve yapay zekâ alanındaki algoritmaların gelişmesi ile daha mümkün hale gelmiştir. Son yıllarda, internet kullanımındaki bant genişliğinin artmasına da bağlı olarak sosyal medya ve video paylaşım platformlarına yüklenen video sayıları ivmelenerek artmaktadır. Video içerik tüketimi ve üretiminin hızlanmasına bağlı olarak bu verilerin işlenmesi konusunda da doğan ihtiyaçla birlikte yeni çalışmalar kendini göstermektedir. Video içeriklerinden hareket analizi talep gören bir konu haline gelmiştir. İnsan içerikli videolarda kişilerin poz tahmini, poz takibi ve eylemlerinin tanınması akıllı video sistemlerinin en önemli bileşenlerindendir.

Bu tez çalışmasında, öncelikle derin öğrenme hakkında kısa bilgiler verilerek poz tahmini, poz takibi, hareket tanıma ve video özetleme yöntemleri incelenmiştir. Farklı poz tahmini algoritmalarını kullanan gerçek zamanlı bir poz takibi yöntemi geliştirilmiştir. Belirlenen poz tahmini yöntemleri için geliştirilen ön işleme adımlarının doğruluk değerine ve işlem sürelerine etkisi incelenmiştir. Poz tahmini yöntemlerinin çıktısı üzerinde veri kalitesini artırarak hareket tanıma yapabilen bir yöntem meydana getirilmiştir. Tüm bu bilgileri kullanarak videodaki karakterlerin başrol olma olasılıkları hesaplanmıştır. Hareket tanıma ve müzik sinyallerine dayalı olarak gerçek hayat problemlerine çözüm olabilecek yeni bir video özetleme algoritması önerilmektedir. Bu çalışma için insansız hava aracı kullanarak etkinlik videoları kaydedilmiştir. Oluşturulan yeni video veri kümesi ilk defa tez çalışması ile herkese açık olarak yayınlanmaktadır. Bu veri kümesi hem hareket tanıma hem de video özetleme alanında kullanılabilmesi yönüyle literatürde bir ilk olma özelliğini taşımaktadır. Tez çalışmasının genel yapısı Şekil 1.1'de gösterildiği gibidir.



Şekil 1.1. Tez çalışmasının genel blok diyagramı

Çok kişili poz takibi insan hareketlerinin tanınması, video özetlemede önemli sahnelerin çıkartılması ve video içeriklerinin sınıflandırılmasında kullanılan önemli bir problemidir [1, 2]. Poz takibi, insan eklem noktalarının yörüngelerinin doğru tahmini, insan eylemi tanıma, insan etkileşimini anlama, hareket yakalama ve animasyonlar için faydalıdır. Özellikle kameraların rastgele hareketler yaparken kaydettiği videolarda çok kişili poz takibi oldukça zordur. Çünkü tekil kameralarda kişiler birbirinin önünden geçerken arkada kalan kişiler geçici olarak hiç görünmeyebilir. Görünmediği görüntülerde de o kişilerin geçici olarak takibi yapılabilmektedir. Poz takibi iki ana adımda gerçekleşir. Birinci adımda, kişilerin eklem noktalarının tespit edildiği poz tahmini yapılır. Bu adım bir sonraki adımda kullanılacak ölçümleri elde etmek için gereklidir. İkinci adımda, görüntü içinde tespit edilen çok sayıdaki eklem noktasının ait olduğu kişiye atanması ve zamana bağlı olarak o kişinin videoda bulunduğu sahne boyunca izlenmesi gerçekleştirilir [3].

Kamera ile çekilen bir görüntüde insan iskeletindeki eklem noktalarının çıkartılması işlemi poz tahmini olarak isimlendirilir. Poz tahmininde, görüntüdeki kişilerin önceden belirlenen eklem noktalarının piksel değerleriyle birlikte güvenilirlik değerinin bulunması amaçlanır. Güvenilirlik değeri 0-1 arasında değişir. Poz tahminin bazı uzuvların görüntüde olmadığı ya da kişinin bazı uzuvlarının önüne başka bir nesnenin geçtiği durumlarda da iyi çalışabilmesi amacıyla farklı yaklaşımlar geliştirilmiştir. Bu yaklaşımlar aşağıdan yukarı yaklaşım ve yukarıdan aşağı yaklaşım olarak isimlendirilir. Aşağıdan yukarı yaklaşımlar önce tüm vücut eklemlerini parça temelli olarak bulur ve bulunan eklem noktasının kime ait olduğunu belirleyerek o kişiye eklem noktası ataması yapar. Yukarıdan aşağı yaklaşımlar ise önce kişileri sınırlandıran bölgeyi bir dikdörtgen alan biçiminde tespit eder ve daha sonra bu dikdörtgen bölgenin içinde yer alan eklem noktalarını bulmaya çalışır [4-6].

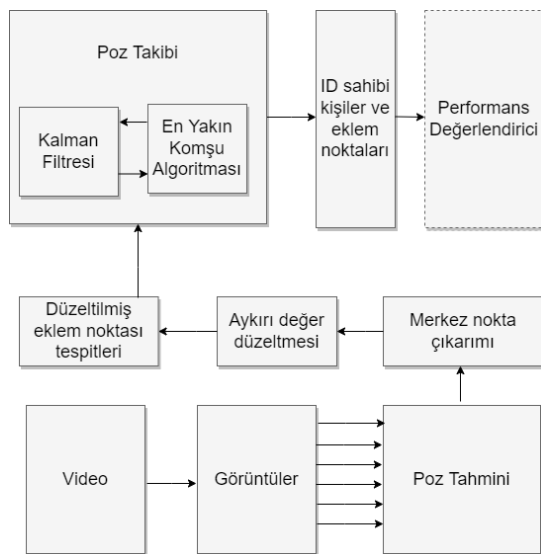
Poz tahmini adımında genellikle evrişimli sinir ağları kullanılmaktadır [7]. Aşağıdan yukarı yaklaşımını kullanan çalışmalara örnek olarak Cao ve diğerleri [8] tarafından geliştirilen OpenPose gösterilebilir. Bu çalışmada eklem noktalarının ısı haritalarını ve eklemlerin

aralarındaki ikili ilişkileri (parça yakınlık alanları) tahmin etmek için bir model oluşturulmuştur [8]. Eklem noktaları, aralarındaki parça yakınlık alanlarının ikili eşleşmelerine göre Hungarian algoritması [9] ile gruplandırılır. Hesaplama performansı, görüntüde tespit edilen kişi sayısından bağımsızdır. OpenPose çok kişili poz tahminlerinde gerçek zamanlı olarak çalışabilir. Ladicky ve diğerleri [10] yönlü gradyanlar histogramı olarak adlandırılan öznetelik çıkarımı ile insan uzuvlarının bölütlemesini ve uzuvların pozisyonlarını birlikte tahmin eden bir algoritma önermiştir. Yukarıdan aşağı yaklaşımı kullanan Sun ve diğerleri [6] HRNet olarak da bilinen farklı çözünürlükteki temsilleri paralel olarak işleyen ve bilgi paylaşımı yapan bir ağ yapısı kurarak çok ölçekli bir birleştirme uygulamıştır. HRNet eklem noktalarını potansiyel ısı haritası ile belirleyen başarılı ve önde gelen bir yöntem olarak kullanılmaktadır. Wang ve diğerleri [5] tarafından önerilen ve yukarıdan aşağı çalışan yöntem diğer yöntemlerden farklı olarak kişileri tespit eden ağdan bağımsız olarak tespit edilmeyen kişilerin de eklem noktalarını tahmin edebilir. Bu çalışmada tespit edilen kişileri çevreleyen kutunun içinde başka kişilerin de olabileceği ve bir kişinin diğerinin önüne geçtiği için tespit edilememiş olabileceği durumlar da göz önüne alınarak klip izleme ağı olarak isimlendirilen bir ağ yapısı ile tahmin işlemi yapılmıştır. Ardışık görüntü karelerinde üst üste gelen tespitleri ilişkilendirerek tespit edilmeyen kişiler üzerinde de poz tahmini yapabilmeyi öneren bir çalışmadır [5]. Liu ve diğerleri [4] geleneksel görüntü tabanlı yaklaşımların video kareleri arasında zamansal bağımlılığı ve geometrik tutarlılığı göz ardı ettiğini öne sürerek DCPose olarak bilinen yeni bir yapı önermiştir. DCPose ardışık görevler icra eden üç modül içermektedir. İlk modül ardışık üç video karesi üzerinde bulunan eklem noktalarını birleştirir. İkinci modül mevcut görüntü ve ardışık görüntüler arasındaki poz artıklarını verimli bir şekilde elde etmek için geçici mesafeleri kullanarak çerçeveler arası eklem noktalarının farklarını hesaplar. Üçüncü modül ise eklem noktalarının ısı haritalarının yeniden örneklenmesi için farklı genişleme oranlarına sahip beş paralel evrişim katmanından oluşan bir poz düzeltme ağı ile son çıktıları üretir.

Poz takibi adımımda çeşitli yöntemler mevcuttur. Poz takibi çalışmaları genellikle poz tahminlerini zaman içinde birbiri ile yeni bir ağ yapısı ile ilişkilendirerek yapmaya çalışmaktadır [5, 11-14]. Yukarıdan aşağıya yöntemler [13, 14] genel olarak kişilerin eklem noktalarını sınırlayan bölgeler arasında zamansal grafikler oluşturarak ve bölgelerdeki özellikleri kullanarak ilişkilendirme yapar. Aşağıdan yukarı yöntemler [11, 12] ise tespit edilen eklemler arasında uzamsal-zamansal grafikler oluşturarak yakınlık alanları ile ilişkilendirme yapar. Raaj ve diğerleri [12] çıkarım sırasında geçmiş bilgileri de dikkate

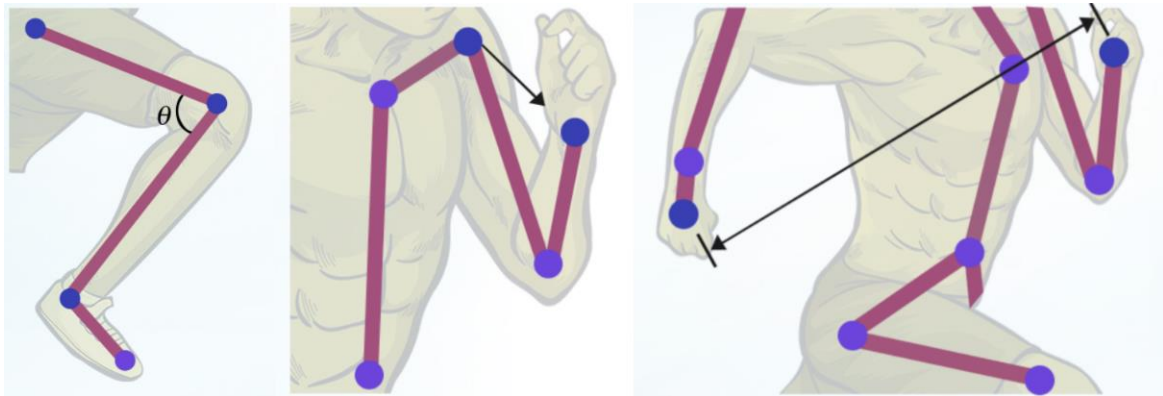
olarak sadece o anki görüntüyü kullanır çünkü gerçek zamanlı çalışabilmesi amaçlanmıştır. Her görüntüde parça yakınlık alanları, eklem noktaları ve geçici yakınlık alanları arasında bağlantı oluşturan bir ağ yapısı ile kişi tespitlerini ilişkilendirir [12]. SimpleBaseline yöntemi önce her bir karede bağımsız olarak bir kişi detektörü çalıştırır ve ardından algılamalarını pahalı optik akış kullanarak zamansal benzerliğin tanımlandığı bir grafikte ilişkilendirir [14]. Detect-and-Track yöntemi ise kısa bir video klipte insanların eklem noktalarını tespit etmek için bir 3D Mask R-CNN yaklaşımı kullanır ve ardından tespit edilen sınırlayıcı bölgelerin konumlarını karşılaştırarak ardışık klipleri birbirine bağlar [13]. Wang ve diğerleri [5] ise bir görüntüde tespit edilen sınırlayıcı kutuları ardışık görüntü karelerine yayarak gözden kaçabilecek kişileri aramayı önermektedir.

Poz takibi yöntemleri arasında performansı yüksek ve maliyeti düşük yöntemi bulmak önemli bir problemdir. Bu çalışmada bu problemin çözümü için gerçek zamanlı çalışan düşük maliyetli bir izleyici Kalman filtresi ve küresel en yakın komşu (KEYK) kullanılarak geliştirilmiş ve OpenPose [8] ve DCPose [4] poz tahmini yöntemlerini kullanarak değerlendirilmiştir. Ön işlem adımlarının sistemin performansına etkisi araştırılmıştır. Aşağıdan yukarı bir yaklaşım olan OpenPose literatürdeki düşük işlem gücü ihtiyacına göre en başarılı yöntemdir. Yukarıdan aşağı bir yaklaşım olan DCPose ise daha çok işlem gücüne ihtiyaç duymakla birlikte PoseTrack veri kümesi üzerinde en başarılı sonuçları veren poz tahmini yöntemlerinden biridir. Geliştirilen poz takibi yönteminin yapısı Şekil 1.2’de verilmiştir.



Şekil 1.2. Geliştirilen poz takibi yönteminin yapısı

İnsan hareketlerini görüntüler üzerinden tanıma aktif olarak araştırılan bir konudur. Bu konudaki çalışmalarda genellikle derinlik [15, 16], RGB [17, 18] ve eklem noktaları [19, 20] olmak üzere üç ana bilgi kullanılmıştır. Özellikle RGB verileri ESA'nın getirdiği tespit yetenekleri ve kolaylıkları sayesinde yaygınlaşmıştır. Ayrıca 3B hareket tanıma sistemleri, yüksek hassasiyetli kamera dizileri kullanılarak 3B eklem konumlarını çıkartır. Bu sistemler zengin bilgi içeriğinden kaynaklı olarak avantajlı durumlara sahiptir. Maliyeti düşük tekil kameralarla kaydedilen 2B görüntüleri kullanarak yapılan hareket tanıma işlemleri hem donanım maliyeti açısından hem de kullanılabilirlik açısından daha uygundur. Çünkü günümüzde internette, TV yayınlarında paylaşılan ya da izlenen videolar tekil kamera kullanılarak kaydedilmektedir. Kinect gibi 3B güvenilir iskeletler oluşturan algılayıcılarla hareket tanıma işlemi video özetleme probleminin çözümü için uygun değildir. Bu nedenle tekli kamera algılayıcılarında da başarılı sonuçlar elde edilmesi ihtiyacı bulunmaktadır. Bu nedenle bu çalışmada RGB ve derinlik yerine 2B eklem noktalarını en verimli şekilde kullanarak hareket tanıma problemi incelenmiştir. Eklem noktaları kamera bakış açısının değiştiği durumlarda da tespit edilebildiği için hareketli kameraların kaydettiği görüntülere dayanıklıdır ve standart sapma değeri daha düşüktür. İlgisi olmayan bilgilerin karışmasını önlediği için aynı zamanda filtrelenmiş veri elde ettiği düşünülebilir. Yao ve diğerleri [21], iskelet tabanlı özelliklerin görünüm tabanlı özelliklerden daha iyi performans gösterdiğini TUM Kitchen [22] veri kümesini ve sınıflandırıcıyı kullanarak kanıtlamıştır. Özyinelemeli Sinir Ağı (ÖSA) ile yapılan güncel araştırmalar görüntü dizilerinin işlenmesi üzerinde önemli etki oluşturmuştur [25-27]. İnsan eklem noktalarına dayalı çok katmanlı ÖSA yapıları ile hareket tanıma konusunda başarılı sonuçlar elde edilmiştir [19, 20, 23, 24]. Mevcut çalışmalarda girdiler eklem noktalarının görüntüdeki piksel konumları ile sınırlıdır.



Şekil 1.3. Hareket tanıma için eklem noktaları arasında belirlenen bazı ilişkiler

Bu tez çalışmasında eklem noktalarının piksel konumları ile birbirleri arasındaki geometrik ilişkileri hesaplayarak Uzun-Kısa Süreli Bellek'in (UKSB) kaliteli verilerle beslenmesi önerilmektedir. Çünkü insanlar hareket ederken fizyolojik olarak kaslarını hareket ettirir ve çeşitli eklem noktaları arasındaki açılar değişmektedir. Buna dayanarak; belirli eklem noktaları arasındaki geometrik ilişkiler tanımlanmaktadır. Bu ilişkilerden bazıları Şekil 1.3'te gösterilmektedir. El, kol ve bacakların yapabileceği hareketler görüntü koordinat sistemine göre 7 farklı açı bilgisi hesaplanarak modellenmiştir. Açı değerlerindeki değişime bağlı olarak hareketlerin tanınması UKSB yapısındaki gradyan işlemleri sebebiyle başarı sağlayacağı sezgisel olarak keşfedilmiştir. Geliştirilen yöntemin avantajı, basit geometrik özelliklerin eklem noktalarının piksel konumlarından daha iyi performans göstermesidir. Bu özelliklerin anlaşılması ve analizi kolaydır. Sonuç olarak, iskelet tabanlı eylem tanıma için geometrik özellikleri hesaplayan ve UKSB modelini kullanan basit ve başarılı bir algoritma geliştirilmiştir. Du ve diğerleri [19] kolların bacakların ve gövdenin ilişkilerini ÖSA ile modellemektedir ve ardışık eklemler arasındaki ilişkileri dikkate almamaktadır. Zhu ve diğerleri [20] ardışık olmayan uzuvlar arasındaki ilişkileri kullanarak UKSB ile sınıflandırma için mesafeye bağlı düzenleme terimi ekler. Bu yöntemler vücut bölümleri arasındaki ilişkileri başarılı bir şekilde inceler ancak sınıflandırma için hedef hareket sınıfları dışında hareketlerin tespiti için esneklik sağlamaz. Mevcut çalışmalar eklem noktaları arasındaki geometrik ilişkileri kullanarak belirli hareketleri tanımaya yönelik başarılı çalışmalardır. Bu çalışmada, önceki başarılı çalışmalardan ilham alarak ardışık eklemler arasındaki açıları kullanarak eklemlerin açılıp kapanmasını ve birbirlerine uzaklıklarını temel alan farklı bir yöntem geliştirilmiştir. Doğrudan hareketleri sınıflandırmak yerine hareket ettiğinden emin olunan uzuvları sınıflandırarak bu hareketlerin meydana getirebileceği insan hareket sınıfı eşleştirmesi ile hareketleri tanıyabilecek ara bir iyileştirme adımı önerilmektedir. Bunlar alt hareket sınıfları ve hedef hareket sınıfları olarak isimlendirilmektedir. Örneğin; diz çöken bir kişi çeşitli hareketler yapıyor olabilir. Bu diz çökme hareketini, evlilik teklifi anı ile ilişkilendirmek videonun kurgusu ile ilgilidir. Bu nedenle mevcut insan hareketlerinin tamamının sınıflandırılabilmesi için yeni bir alt yapı meydana getirilmektedir.

Ayrıca bu çalışma için hazırlanan *PoseDrone* isimli veri kümesinde, iskelet tabanlı hareket tanıma için ilk aşama olarak 5 farklı hareket sınıfına odaklanılmıştır. Bunlar; yürüyüş, diz çökme, sarılma, el ele tutuşma ve sabit duruş sınıflarıdır. Veri kümesinden örnek görüntüler Şekil 1.4'te verilmiştir. Hareketlerin etiketlenmesi için bu çalışmaya has bir yöntemle

verilerin etiketlenme işlemini kolaylaştıran bir alt yapı geliştirilmiştir. Poz takibinin getirdiği kolaylık sayesinde her görüntü karesini tek tek etiketleme ihtiyacı ortadan kalkmıştır. Veri kümesi, 5 farklı kamera kullanılarak 1920x1080 piksel boyutlarında kaydedilmiştir. *PoseDrone* veri kümesi hakkındaki detaylı bilgiler Bölüm 2.9'da verilmiştir. Güncel detaylar posedrone.com adresinden de takip edilebilmektedir.



Şekil 1.4. Tez çalışması için kaydedilen veri kümesi PoseDrone'dan bir görüntü

İnternet, sosyal medya ve video kayıt cihazlarındaki gelişmelerle birlikte video işleme alanı çeşitli zorluklarla karşı karşıya kalmıştır. Video işleme; videodaki anormal olay tespitleri, insan davranışları gibi birçok veri, sahnelerin analiz edilmesi için kullanılmaktadır. Bu sahneler; genellikle halka açık etkinlikler, ürün ya da marka tanıtım videoları, çeşitli spor ve dans faaliyetlerinde kaydedilmiş olabilir. Bilindiği gibi video, görüntü ve ses bilgilerinin zaman dizisi olarak birleştirilmiş halidir. Son zamanlarda çalışmalar videolardaki faydalı bilgileri çıkarabilecek tekniklerin arayışı içindedir. Bu çalışmada görüntü bilgisi kullanılarak bilgi çıkarımına odaklanılmıştır. Genellikle çalışmalar depolama alanları problemine çözüm getirmek ve video içinde arama yapabilmek için videoları özetlemeye çalışmaktadır [28]. Bu çalışmada farklı olarak sanatsal bir yaklaşımla videoların son kullanıcı tarafından tüketilebilir olması noktasında teknik olarak nelerin yapılabileceği konusunda çeşitli yaklaşımlar ortaya konmaktadır. Çünkü genellikle videolar çok kişi tarafından izlenebilir olabilmesi amacıyla kaydedilmekte ve düzenlenmektedir.

Video özetleme yaklaşımları genellikle sahne tabanlı ve içerik tabanlı olmak üzere 2 ana kategoride incelenmektedir. Değişen sahneler içeren bir video, kısa video kliplerin birleşmesinden meydana gelir. Bu durumda, aynı sahneleri temsil eden video dizileri belirlenerek özetleme gerçekleştirilir [29-35]. Çıkarılan öz nitelikler, uygun kümeleme yöntemleriyle ana karenin seçilmesini sağlar. Bu yaklaşımda anlamsız görüntü karelerinin kaldırılması için bazı işlemler uygulanması gerekebilir. İçerik tabanlı yöntemlerde ise videonun içeriğine bağlı olarak; harekete ve olaylara dayalı olarak özetleme yapılabilir. İçerik tabanlı video özetleme, hareket tanıma gibi ön işleme adımlarının sonuçlarını kullanır [36-41]. Video özetlemede, videonun anlaşılması ve önemli görüntülerin sıralamalarının sınıflandırılması en zor problemdir. Videoların sınıflandırılması nesne türlerine de bağlı olduğu için zorluk artmaktadır. Bu çalışmada insan içerikli etkinlik videolarının içerik tabanlı olarak özetlenmesine odaklanılmıştır.

Video özetleme; hızlı inceleme için uzun orijinal videoları sıkıştırma tekniği [42] olarak tanımlansa da bu çalışmada video özetleme, uzun videoların izleyiciler tarafından tüketilebilir halinin sanatsal bir yaklaşımla kısaltılması olarak ele alınmıştır. Sanatsal yaklaşımı düzenlenen videolarda kurguların belirlenmesi ile ilişkilendirilmektedir. Örneğin; yayın programları bir uzman tarafından standart kamera bakış açıları, anlatım veya üst üste bindirilen metinler gibi düzenleme kurallarına göre kaydedilir ve montajlanır [43]. Bu yaklaşımlara dayanarak, video konularının ve kurgularının bir uzman tarafından bilindiğini varsayarak özetleme yöntemine odaklanılmaktadır. Poz tahmini ve takibi sonuçlarına bağlı olarak başrol oyuncularının tahmini, kişiler arası video boyunca Öklid mesafesi takibi, muhtemel başrol oyuncular arasındaki kesişim hesaplamalarını kullanarak oyuncuların kameraya göre birbirinin önünü kapatıp kapatmadığı gibi bilgileri kullanarak sahnelerin kurguyla ilişkilendirilmesi gerçekleştirilir. Sahnelerin uzunluklarının ve diğer sahneye geçiş sürelerinin belirlenmesi ise oluşturulacak video klibin arka planına eklenecek müziğin temposuna bağlı olarak değişmektedir. Seçilen müziğin vuruş zamanları belirlenir. Sahnenin anlamına göre bazen her vuruşta bazen iki vuruşta bir bazen de dört vuruşta bir olmak üzere üç farklı seçim yapılarak parça video sürelerindeki değişim yapılmaktadır. Çalışmanın asıl amacı müzik sinyallerinden tempo bulunması değildir. Sahne sürelerini belirleyebilmek için gerçek hayatta kullanılan bir video düzenleme yöntemi olan bu yaklaşımın kullanılması uygun görülmüştür.

Çalışmanın devamı aşağıdaki gibi tasarlanmıştır: İkinci bölümde, çalışmada kullanılan veri kümesi, ön işleme adımları, poz takibi sistemini geliştirmede kullanılan teknikler olan Kalman filtresi ve KEYK ile performans değerlendirme ölçütleri açıklanmıştır. Hareket tanıma yöntemleri, belirlenen müziğin temposunun çıkarılması ve video özetleme yöntemleri incelenmiştir. Ayrıca bu çalışma için oluşturulan yeni veri kümesinin detaylarından bahsedilmiştir. Üçüncü bölümde, çalışmadan elde edilen deneysel sonuçlar sunulmuştur. Dördüncü bölümde ise çalışmadan elde edilen sonuçlar tartışılmıştır.

2. MATERYAL ve METOD

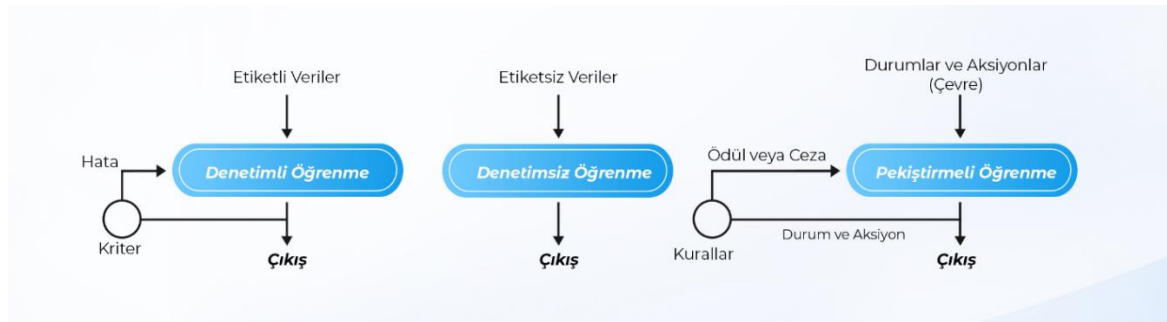
2.1. Derin Öğrenme

Derin öğrenme, ardışık katmanlarda işlenen verilerin git gide daha iyi sonuçlar verdiği makine öğrenmesinin bir alt alanıdır. *Derin* kelimesi ile ağ yapısındaki birbirini takip eden katmanlar kastedilmektedir. Derin öğrenme, güncel gelişen bir konu olduğu için tanımı da henüz tam olarak netleşmemiştir. Goodfellow [44] ve Skansi'nin [45] kitaplarında bazı tanımlar mevcuttur. Goodfellow derin öğrenmeyi bilgisayarların deneyimlerden öğrenmesi ve dünyayı kavramlar hiyerarşisi açısından anlamasını sağlayan bir araç olarak tanımlar. Skansi ise makine öğrenmesinin bir alt alanı olarak derin öğrenmeyi ifade eder. Bu bölümde, makine öğrenmesi, sinir ağlarının eğitimi, poz tahmini algoritmalarının temeli olan evrişimli ağlar ve hareket tanıma probleminin temel çözümü olan tekrarlayan sinir ağları incelenmiştir.

2.1.1. Makine öğrenmesi

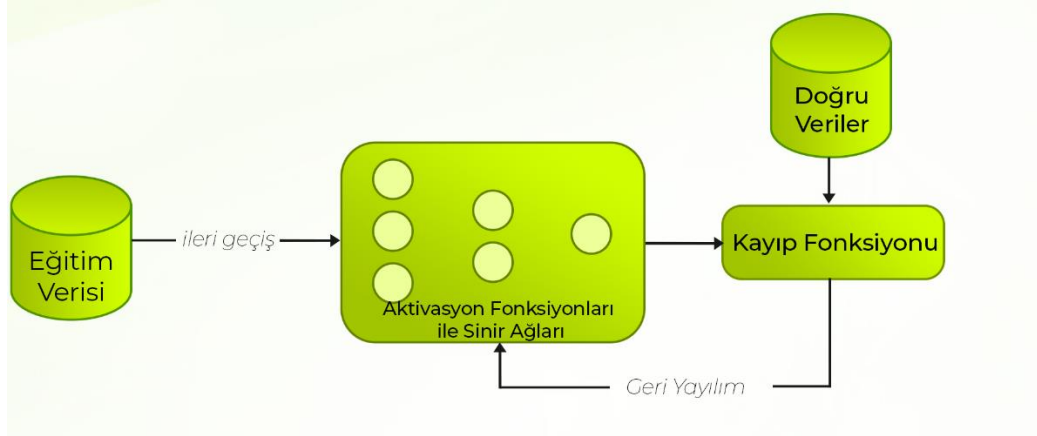
Yapay zekânın (YZ) bir alt alanı olan makine öğrenmesi modelleri, örnek verilerle karmaşık problemleri tahmin etmek için istatistiksel hesaplamalar kullanır. Makineler, probleme bağlı olarak girdileri nasıl işleyeceklerini gösteren bir algoritmaya ihtiyaç duyar. Her farklı giriş değeri için ayrı bir algoritma yazmak mümkün olmadığından makine öğrenmesi istatistiksel modeller kullanarak veri kümesi hakkında çıktı üretebilmektedir.

Makine öğrenmesi, genellikle Şekil 2.1’de gösterildiği gibi üç ana başlıkta incelenmektedir. Öğrenme modelleri giriş verilerinin yapısına göre farklılık göstermektedir.



Şekil 2.1. Makine öğrenmesinin üç farklı alt başlığı

Denetimli öğrenmede, etiketli veriler kullanılarak makine öğrenmesi modeli eğitilir. Giriş verileri modele uygulanır ve çıktılar kayıp fonksiyonu (loss function) denilen, kullanıcı tarafından önceden belirlenmiş kriterler kullanılarak beklenen doğru çıktılarla karşılaştırılır. Denetimli öğrenme eğitim döngüsü Şekil 2.2’de görülmektedir. Burada, veri kümesini en az kayıpla doğru etiketlere eşleştirmek amaçlanır. Literatürdeki poz tahmini ve poz izleme çözümleri denetimli öğrenmeye örnektir.



Şekil 2.2. Denetimli öğrenmenin genel yapısı

En gelişmiş poz takibi yöntemlerinin çoğu denetimli öğrenmeyi kullanır. Bununla birlikte, [46, 47] gibi birkaç yeni çalışma iyi performansa sahip diğer iki öğrenme yaklaşımını kullanmıştır. Denetimli öğrenme çözdükleri problemlere göre iki farklı alt başlıkta incelenebilir.

1. Sınıflandırma: Girdi değişkenlerini çıktı değişkenleri ile eşleştiren fonksiyonu elde etmek gibi düşünülebilir. Çıktı değişkenlerine etiketler, kategoriler ya da sınıflar denir. Örneğin; farklı insan iskeleti poz tahmini yapılarına sahip kişilerden oluşan bir veri kümesinden rastgele giriş değerleri verilirse, hangi poz tahmini yapısının hangi hareket sınıfına karşılık geldiği sınıflandırma yaparak belirlenebilir. Genellikle ayrık zamanlı sistemlerde kullanılır.

2. Regresyon: Eşleştirme fonksiyonunu yaklaşık olarak en düşük hata sapması ile belirlemek olarak düşünülebilir. Sürekli zamanlı çıktılar genellikle miktarlar ve boyutlar gibi sonuçları temsil eder. Genellikle sürekli zamanlı sistemlerde kullanılır. Sayısal bağımlılığın ayırt edilmesi beklenmektedir.

Denetimsiz öğrenme, eğitim için etiketlenmemiş verileri kullanır. Verilen girdilerin içindeki yapıları bulmayı amaçlar. Örneğin; bir LIDAR algılayıcı ile toplanan nokta kümesi içindeki

farklı nesneleri farklı nokta kümelerine ayırabilmek için kümeleme yöntemleri kullanılır. Bundan farklı olarak boyut azaltma gibi örüntülerin bulunması için denetimsiz öğrenme kullanılmaktadır.

Pekiştirmeli öğrenme, takviyeli öğrenme olarak da bilinir. Model tarafından verilen bir dizi karara cevap olarak yeni veriler sağlayan bir geri bildirim döngüsünü kullanır. Modele doğru etiketler verilmez fakat ajan doğru seçimde ödüllendirilir ve yanlış seçimde cezalandırılır. Amaç, en iyi karar politikasını oluşturarak toplam ödülü en büyük yapabilmektir.

Pekiştirmeli öğrenmeyi anlayabilmek için öncelikle; çevre, ajan, durum, aksiyon ve ödül olmak üzere bu beş terimin tanımını bilmek gerekir. Çevre, ajanın etkileşimde bulunduğu ortamdır. Ödül değerini ve yeni durumu gerçekleştirdiğini durum ile öğrenir. Ortam olarak da isimlendirilebilir. Ajan, öğrenir ve karşılaştığı durumlara tepki verir. Çevre içinde hareket eder. Çevreye belirli bilgiler gönderip alır. Ajan ya da çalışan olarak da isimlendirilir. Ajanın öğrenmesi ve doğru aksiyonları alır hale gelmesi için gereken ortamın hazırlanması da modeli tasarlayan kişi tarafından gerçekleştirilir. Durum, ajanın çevre içinde nerede olduğu bilgisi tutar. Aksiyon, ajanın yapabileceği hareketi belirler. Örneğin, ileri git, sağa dön, sola dön ve geri git gibi farklı dört aksiyondan oluşan bir aksiyon seti olabilir. Ödül, ajanın amacı çevrede en uygun koşullar içinde ödül değerini en büyük değere ulaştırmaktır. Ödül negatif değere sahip olduğunda ceza anlamına gelir. Yani ödül, hem pozitif hem de negatif anlama gelebilmektedir.

2.1.2. Veri ön işleme

Derin öğrenme modellerinin öğrenme süreci, veri kümelerinin boyutundan ve kullanılan veri ön işleme tekniklerinden önemli derecede etkilenir [48]. Eğitim verileri tasarlanan ağa beslenmeden önce genellikle ön işlemlerden geçirilir. Ön işleme adımında ilk olarak veri kümesinden ortalamanın çıkarılması yaygın olarak kullanılır. Böylece veriler farklı boyutlarda da olsa orijin etrafında merkezlenir. Örneğin; RGB görüntüler için tüm örnekler alınarak her kanal için ortalama hesaplanır. Ardından o kanaldaki her pikselden ortalama değeri çıkarılır. Başka bir teknik ise örnekleri her boyutta yeniden ölçeklendirmektir. Bu ölçeklendirme işlemi probleme bağlı olarak her bir veri parçası için ya da genele göre ölçeklendirme gibi farklılıklar gösterebilir. Verilerin ön işlenmesi için kullanılan bazı yöntemler Eş. 1 – Eş. 3 ile gösterilir. Eş. 1 yeniden ölçeklendirmeyi, Eş. 2 normalleştirmeyi, Eş. 3 ise standartlaştırmayı gösterir.

$$x_i = \frac{x_i - \text{enk}(X_i)}{\text{enb}(X_i) - \text{enk}(X_i)} \quad (1)$$

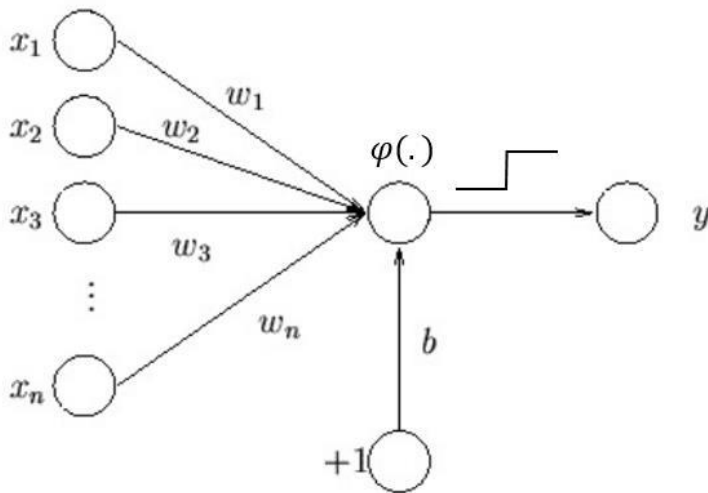
$$x_i = \frac{x_i}{\|x_i\|} \quad (2)$$

$$x_i = \frac{x_i - \mu_x}{\sigma_x} \quad (3)$$

x_i giriş örneğinin x kanalını temsil eder ve X_i tüm veri kümesinin i numaralı kanalıdır. μ , giriş örneğinin ortalama değeridir. σ , giriş örneğinin standart sapmasıdır. Dizideki en küçük değer $\text{enk}()$ ile en büyük değer ise $\text{enb}()$ ile bulunur.

2.1.3. Yapay sinir ağı

Bir yapay sinir ağı (YSA), insan beyninin çalışma şeklini taklit eder. İnsan beynindeki nöronlar, diğer sinir hücrelerine bilgi iletebilen hücrelerdir. Hücreler, bir sinyal iletim ağına bağlı olan diğer nöronlardan girdileri toplar. YSA'daki temel birimler, insan beynindeki nöronlarla taklit edilir. Şekil 2.3'teki gibi, bir nöronun n adet $\{x_1, x_2, \dots, x_n\}$ girdisi alabilir. Bu girdiler, ilişkili ağırlıklarla ($w_1, w_2, w_3, \dots, w_n$) çarpılır. Bu değer bir aktivasyon fonksiyonuna $\varphi(.)$ iletilir ve aktivasyonun çıktısı bir eşik değeri (genellikle sıfır) aşarsa, nöron etkinleştirilir ve bir çıktı gönderir. Çıktı değeri y , kullanılan aktivasyon fonksiyonuna bağlıdır. Bu basit örneğin matematiksel formülü Eş. 4'te verilmiştir. $\mathbf{w} \cdot \mathbf{x}$, ağırlıklar ve girdi vektörleri arasındaki nokta çarpımını temsil eder.

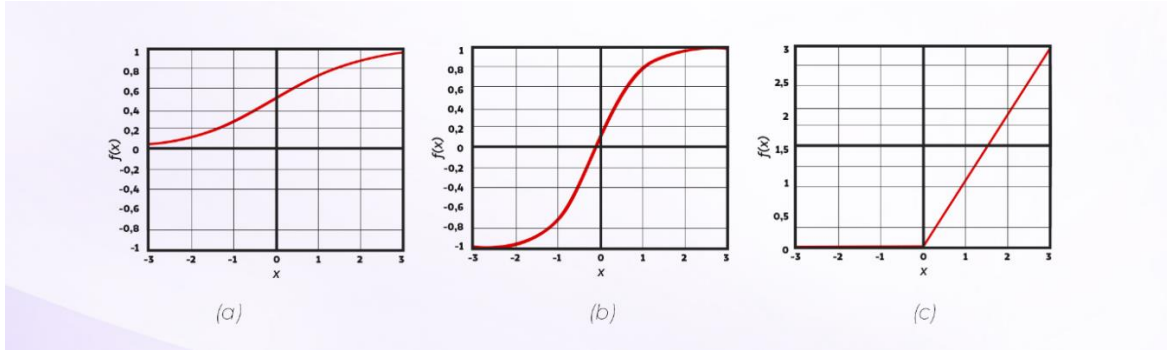


Şekil 2.3. n girişli bir nöron [44]

$$y(x) = \begin{cases} 1, & \varphi(\mathbf{w} \cdot \mathbf{x} + b) \geq \text{eşik değeri} \\ -1, & \text{diğer durumlar} \end{cases} \quad (4)$$

İki katmandan fazla olan bir sinir ağı, derin sinir ağı olarak adlandırılır. Modelin hataları, bir kayıp hesaplama fonksiyonu ile hesaplanır ve “geri yayılım süreci” adı verilen bir süreçle ağırlıkları günceller.

Aktivasyon fonksiyonları nöronların temel elemanlarından biridir. Nöronun etkinleştirilmesi gerekli gerekmediğini belirleyen değeri üretir. Genellikle sinir ağı modellerinde ortak sorun, bazı gradyanların bir yönde hızlı ve diğer yönlerde sıfır ya da yavaş olması nedeniyle kaybolmasıdır. Öğrenme performansının iyileştirilmesi için doğru aktivasyon fonksiyonu seçilmeli ve daha iyi özellikler çıkarılmalıdır. Aktivasyon fonksiyonunun kullanılmaması durumunda sadece doğrusal problemler öğrenilebilmektedir. Fakat gerçek hayat problemlerinde çok sayıda doğrusal olmayan durum vardır. Bu nedenle giriş özelliklerinden çıkış verilerine kadar doğrusal olmayan ilişkileri tahmin etmek için doğrusal olmayan aktivasyon fonksiyonları kullanılmaktadır. Doğrusal olmayan aktivasyon fonksiyonlarının bir avantajı da türevinin alınabilir olmasıdır. Bu fonksiyonların türevi alınamasa da geri yayılım esnasında çalışmazdı. Güncel ve yaygın kullanılan aktivasyon fonksiyonlarından sigmoid, ReLU ve tanjant hiperbolik fonksiyonlarının yapıları Şekil 2.4’te görülmektedir.



Şekil 2.4. (a) Sigmoid fonksiyonu, (b) Tanh fonksiyonu, (c) ReLU fonksiyonu

Sigmoid fonksiyonu, Şekil 2.4.(a)’daki gibi bir eğriye sahiptir. Gerçek değerli, türevlenebilir ve kesinlikle artan bir fonksiyondur. Gerçek giriş değerlerini ve (0, 1) aralığındaki çıkış değerine dönüştürür. Genellikle σ ile gösterilir. Giriş değeri x olan sigmoid fonksiyonunun matematiksel formülü Eş. 5’te gösterilmiştir. Sigmoid, bir nöronun belirli bir zaman aralığında başlatılma şansını tanımlayan bir olasılık fonksiyonu olarak yorumlanabilir. Bu nedenle genellikle sinir ağlarının sınıflandırma katmanında kullanılır. Birden çok sınıfın olduğu problemlerde kullanılan softmax katmanına benzer olarak uygulanabilir.

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (5)$$

ReLU, Şekil 2.4.(c)'deki gibi bir eğriye sahiptir. ReLU kullanarak derin sinir ağlarının eğitiminde iyileşmelerin olduğu gösterilmiştir. Giriş değeri x olan ReLU fonksiyonunu formülü Eş. 6'daki gibidir. Özetle giriş değeri 0'dan küçükse çıktı olarak 0 verir, giriş değeri 0'a eşit ya da büyükse doğrusal fonksiyon ile elde edilen bir değer çıkar. Sigmoid fonksiyonunda olduğu gibi kaybolan gradyan probleminden etkilenmediği için genellikle gizli katmanlarda tercih edilir.

$$f(x) = \max(0, x) \quad (6)$$

Hiperbolik tanjant (tanh) olarak da gösterilir. Şekil 2.4.(b)'deki gibi bir eğriye sahiptir. Sıfır merkezli aktivasyon fonksiyonudur. Eş. 7'de görüldüğü gibi giriş değeri x olan fonksiyonun denklemi verilmiştir. Çıkış değerleri $(-1, 1)$ aralığında değişmektedir.

$$h(x) = \frac{e^{2x} - 1}{e^{2x} + 1} \quad (7)$$

Kayıp fonksiyonlarının kullanılmasındaki amaç, kayıpları en aza indirmeye çalışmaktır. Mümkün olduğunca en düşük kayıp puanına sahip aday çözümü bulmak hedeflenir. Hareket tanıma görevi için yaygın olarak kullanılan bazı sınıflandırma kayıpları incelenmiştir.

ℓ_2 kayıp fonksiyonu, Öklid mesafesini kullanarak maliyet hesaplaması yapar. Bu mesafenin sahip olduğu özellikler dikkate alındığında, iskelet tabanlı hareket tanıma modelleri için ciddi bir sorun olan eklem noktalarının elde edildiği kişilerin görüntüye yakınlık ya da uzaklıklarından ve görüntü boyutlarından önemli ölçüde etkilenir. Ortalama kare hata (mean squared error) denklemi Eş. 8'de verilmiştir. Yönlerden bağımsız olarak sadece ortalama büyüklük hatası ile ilgilidir. n toplam örneklerin sayısını temsil ederken y_i , i numaralı örnek için sinir ağının çıktısıdır. $\hat{y}_i$ ise çıktı olarak beklenen etiket değeridir. Poz tahmini adımında, koordinat tahmini için yumuşatılmış ℓ_1 mesafesi kullanılır.

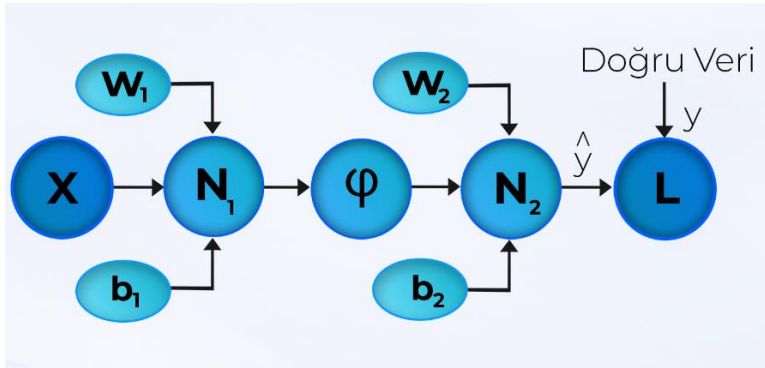
$$C(y, \hat{y}) = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (8)$$

Çapraz entropi kayıp fonksiyonu, entropiye dayalı iki olasılık dağılımı arasındaki farkı hesaplamak için kullanılır. ℓ_2 yerine çapraz entropi kullanmak, daha hızlı eğitimin yanı sıra daha iyi bir genelleme sağlar [49]. Genellikle olasılık değerini 0 ile 1 arasında veren modeller için kullanılır. UKSB kullanarak zaman dizilerinde sınıflandırma yapılırken sınıflandırma kaybı olarak çapraz entropi kullanılır. Eş. 9 çapraz entropi kaybının matematiksel formülünü gösterir.

$$C(p) = - \sum_{c=1}^N y_{(o,c)} \cdot \log(p_{(o,c)}) \quad (9)$$

N sınıfların toplam sayısını temsil eder. p , model tarafından üretilen N boyutlu bir vektördür ve girdinin hedeflenen (c) sınıflarından biri olma olasılıklarından oluşur. y ikilik bir göstergedir. Tahmin edilen sınıf c , gözlem o için doğru sınıflandırma ise sonuç 1'dir. $p(o,c)$ c sınıfının o gözlemi için tahmin olasılığıdır.

Geri yayılım, her bir girdi değeri için sinir ağı ağırlıklarına göre (zincir kuralı kullanılarak) kayıp fonksiyonlarının gradyanını hesaplamak için yaygın olarak kullanılan bir yöntemdir. Geri yayılım yöntemi bazı avantajlar sağlar. Bunlardan ilki; yaklaşık gradyan yerine gradyanın tam değerini ölçer. İkincisi, dinamik programlamaya uygundur. Üçüncüsü ise her bir parametre için tek seferde gradyan değerini hesaplar. Şekil 2.5'te basit bir iki katmanlı sinir ağı görülmektedir. Giriş değeri x , aktivasyon fonksiyonu φ ve L kayıp fonksiyonunu gösterir.



Şekil 2.5. İki katmanlı bir sinir ağı

Tam bağlantılı katmanlar, akıl yürütmeleri gerçekleştirmeyi amaçlar. Tam bağlantılı katmanlar, ESA mimarisinde genellikle havuzlama katmanından sonra kullanılır. ÖSA yapılarında da genellikle, UKSB ve bırakma katmanlarından sonra kullanılır. Önceki katmanlardan çıkarılan yüksek seviyeli özelliklerle tam bağlantılı bir sınıf puanı çıkarmaya veya aktivasyonlardan nesne koordinatlarını tahmin etmeye çalışacaktır. UKSB ile sınıflandırma yaparken sınıf puanı çıkarma görevini üstlenmektedir. Tam bağlantı katmanı çıkışları ile softmax katmanının kullanılması yaygın bir uygulamadır. Tam bağlantı katmanı, önceki katmandaki tüm nöronları alır ve genel anlamsal bilgi üretmek için bunları mevcut katmanın her bir nöronuna bağlar.

2.1.4. Özyinelemeli sinir ağı (ÖSA)

Doğal dil işleme, nesne takibi, müzik oluşturma ve benzer konular sıralı verilerle çalışmayı gerektirir. Sıralı verileri kullanarak girdiler ve çıktılar arasında bir bağlantı oluşturarak daha iyi tahminler elde edebilmeyi amaçlar. Özyinelemeli Sinir Ağları, mevcut zaman anına kadar hesaplananlarla ilgili bilgileri depolayabilen bir durum oluşturarak dizilerin dinamiklerini yakalayabilen modellerdir. Kaybolan gradyan (vanishing gradient) problemine Geçitli Tekrarlayan Birim (GRU) ve UKSB ile çözüm önerileri getirilmiştir. UKSB yapısındaki kapıların daha çok kontrol edilebilir olması yönüyle gelişmiş çözümler sunar.

Uzun-Kısa Süreli Bellek (UKSB)

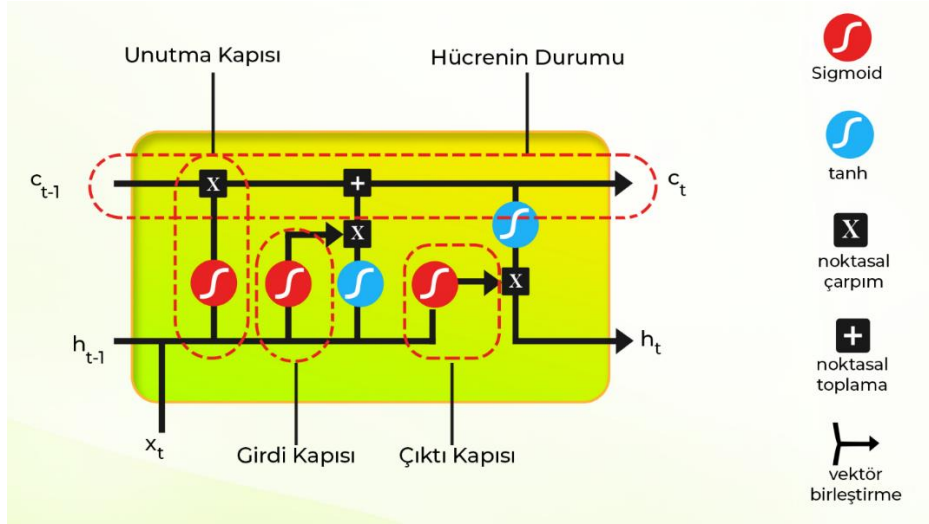
Önceki konular, Uzun-Kısa Süreli Bellek (UKSB) modelini tasarlayanın temel kaynağıdır. UKSB'ler, zaman dizilerindeki uzun vadeli bağımlılıkları öğrenebilen özel bir ÖSA türüdür. Hochreiter and Schmidhuber [50] tarafından 1997 yılında tanıtılmıştır. Standart bir UKSB hücresi Şekil 2.6'da gösterildiği gibi a, b, c, d maddelerinde açıklanan dört ana ögeye sahiptir:

a) Giriş Kapısı: C_{t-1} hücre durumunu günceller. En iyi tahmin sonucunu elde edebilmek üzere, hangi ağırlıkların güncellenmesi gerektiğine karar vermek için önceki gizli durum h_{t-1} ile t zaman adımında x_t girdisini alır.

b) Hücre Durum Bağlantısı: yeni hücre durumunu C_t çıkarmak için tüm kapıları birbirine bağlayan bir bağlantıdır.

c) Unutma Kapısı: Dahili durum ağırlıklarının sınırsız büyümesini engelleyerek gradyan genişleme problemini çözer. UKSB'ye ihtiyaç duyduğu sürece C_{t-1} hücre durumunu sıfırlama yeteneği sağlar. Başka bir deyişle, temel durum değerleri bire yakın bir sayı ile ağırlıklandırılır.

d) Çıkış kapısı: Önceki durumu h_{t-1} yeni hücre durumu C_t ile hesaplayarak günceller.



Şekil 2.6. Standart bir UKSB hücresi

2.2. PoseTrack Veri Kümesi

Bu çalışmada, Pishchulin ve arkadaşları tarafından derlenen PoseTrack veri kümesi kullanılmıştır [51]. PoseTrack video tabanlı çok kişili insan pozu tahmini ve poz takibi için kullanılan büyük ölçekli bir veri kümesidir. 66 374 görüntü karesine sahip 550 adet videodan oluşur. Çizelge 2.1 PoseTrack veri kümesini oluşturmada kullanılan kaynakları, etiketli pozların sayısını, videoda birden çok kişinin bulunup bulunmadığını, video poz etiketlerinin kullanılabilirliğini ve videoların türünü göstermektedir [51].

Çizelge 2.1. PoseTrack veri kümesini oluşturmada kullanılan kaynaklar ve videoların özellikleri [51]

Dataset	# Poses	Multi- person	Video-labeled poses	Data type
LSP [25]	2,000			sports (8 act.)
LSP Extended [26]	10,000			sports (11 act.)
MPII Single Person [1]	26,429			diverse (491 act.)
FLIC [38]	5,003			feature movies
FashionPose [9]	7,305			fashion blogs
We are family [10]	3,131	✓		group photos
MPII Multi-Person [1]	14,993	✓		diverse (491 act.)
MS COCO Keypoints [28]	105,698	✓		diverse
Penn Action [51]	159,633		✓	sports (15 act.)
JHMDB [23]	31,838		✓	diverse (21 act.)
YouTube Pose [6]	5,000		✓	diverse
Video Pose 2.0 [39]	1,286		✓	TV series
Multi-Person PoseTrack [22]	16,219	✓	✓	diverse
Proposed	153,615	✓	✓	diverse

Toplamda 153615 adet açıklamalı poz ve 22000'den fazla etiketli görüntü bulunmaktadır. Veri kümesinde etiketleme işlemi için VATIC [52] aracı kullanılmıştır. Bu araç sayesinde ardışık tüm görüntü karelerini etiketlemeden boş bırakılan görüntü kareleri nesne takibi yapılarak birleştirilebilmektedir. Çizelge 2.1'den görülebileceği gibi veri kümesi spor, film ve blog gibi çok farklı özelliğe sahip çok kişili videolar içermektedir. Şekil 2.7, veri kümesinde bulunan farklı özelliklerdeki videolardan elde edilmiş örnek görüntüler içermektedir. Veri kümesinde çok kalabalık, insanların kameraya göre birbirinin önünü kapattığı ve az kişili görüntü örnekleri Şekil 2.7'de bir araya getirilmiştir. Bazı sahneler oldukça kalabalık ve karmaşık iken bazı sahneler ise az kişili ve sadedir.



Şekil 2.7. PoseTrack veri kümesi örnek görüntüleri ile elde edilen poz tahminleri (Üstteki iki satır OpenPose çıktıları, alttaki iki satır DCPose çıktıları)

2.2.1. Performans ölçütleri

Performans değerlendirmesi yapabilmek için poz tahmini ve poz takibi yönteminden PoseTrack veri kümesinde bulunan etiketler ile aynı formatta çıktılar oluşturmak gerekmektedir. Çok kişili poz takibi için bir görüntü karesindeki bir kişi ele alınırsa bu kişi için bulunması gereken veriler $x1$, $y1$, $x2$, $y2$, $score$, $trackID$ ve $annopoints$ 'tir. $x1$, baş bölgesini sınırlayan kutunun sol üst köşesinin piksel sütun değeri, $y1$, baş bölgesini sınırlayan kutunun sol üst köşesinin piksel satır değeri, $x2$, baş bölgesini sınırlayan kutunun

sağ alt köşesinin piksel sütun değeri, y_2 , baş bölgesini sınırlayan kutunun sağ alt köşesinin piksel satır değeri, $score$, kişinin güvenilirlik puanıdır. Kişilerin güvenilirlik puanı 0-1 arasında değişmektedir. Annopoints değerinin ise poz tahmini ile elde edilen tespit edilmiş eklem noktalarının eklem numarası, sütun piksel değeri, satır piksel değeri ve görünürlüğü bilgilerini içermesi gerekmektedir. Örneğin bir görüntüde bir kişinin sağ ayak bileği görünmemiş ve diğer tüm eklem noktaları görünmüş olabilir. Bu durumda sağ ayak bileğinin görünürlük değerine 0, diğer eklem noktalarının görünürlük değerine 1 verilir.

Çok kişili nesne takibi performansını değerlendirmek için MOTP (Multiple Object Tracking Precision) ve MOTA (Multiple Object Tracking Accuracy) ölçütleri kullanılmaktadır. PoseTrack veri kümesi de yöntemlerin performanslarını kıyaslarken bu ölçütleri kullanır. MOTP ve MOTA değerleri Eş. 10 ve Eş. 11'deki gibi hesaplanır [53].

$$MOTP = \frac{\sum_{i,t} d_t^i}{\sum_t c_t} \quad (10)$$

$$MOTA = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{\sum_t g_t} \quad (11)$$

Eş. 10 tüm çerçeveler üzerindeki eşleşen nesne-hipotez çiftleri için tahmini konumdaki toplam hatadır ve yapılan toplam eşleşme sayısı ile ortalaması alınır. Takipçinin nesne konfigürasyonlarını tanıma, tutarlı yörüngeler tutma vb. becerisinden bağımsız olarak, kesin nesne konumlarını tahmin etme yeteneğini gösterir. Burada c_t , t anı için bulunan eşleşme sayısıdır. Bu eşleşmelerin her biri için o_i nesnesi ve ona karşılık gelen hipotez arasındaki d_t^i mesafesi hesaplanır. Kalan tüm hipotezler yanlış pozitif olarak kabul edilir. Benzer şekilde, kalan tüm nesneler kayıp olarak kabul edilir. Eş. 11'te m_t , fp_t , ve mme_t t görüntüsü için sırasıyla yanlış pozitiflerin, kayıpların ve uyumsuzluk hatalarının sayısıdır. g_t , t anındaki mevcut nesnelerin sayısını ifade etmektedir [53].

PoseTrack veri kümesi üç farklı görevde yöntemlerin karşılaştırılmasına imkân sağlar;

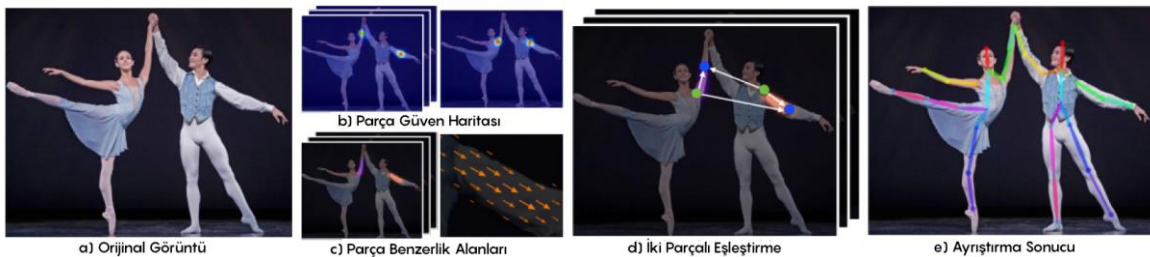
- i. Tek görüntü ve çok kişili poz tahmini
- ii. Videolarda çok kişili poz tahmini
- iii. Çok kişili poz takibi

2.3. Poz Tahmini

Poz tahmini, kamera ile çekilen görüntülerde insan iskeletindeki anahtar noktaların piksel konumlarının tespit edilmesi amacıyla ortaya çıkmıştır. Bir görüntüdeki birden fazla kişinin 2B eklem noktalarını yani pozunu algılamak için yüksek performanslı yaklaşımlar ortaya çıkmıştır. Poz tahmini yöntemlerinin çıktıları görüntüde tespit edilen tüm eklem noktalarının piksel değerlerini ilişkilendirdiği kişiye göre vermektedir. Videolarda kişilerin hareketlerini anlamlandırmak için kişilerin ardışık görüntü kareleri boyunca kimliklendirilmesi gerekir. Böylece kişilerin video boyunca ne zaman hangi hareketleri yaptığı tespit edilebilir. Bu amaçla çeşitli poz tahmini yöntemleri kullanılmaktadır. Poz tahmini yöntemleri içinde en yaygın olarak kullanılan OpenPose ve veri kümelerinde en yüksek doğruluk değerini elde eden DCPose yaklaşımları incelenmiştir.

2.3.1. OpenPose yöntemi

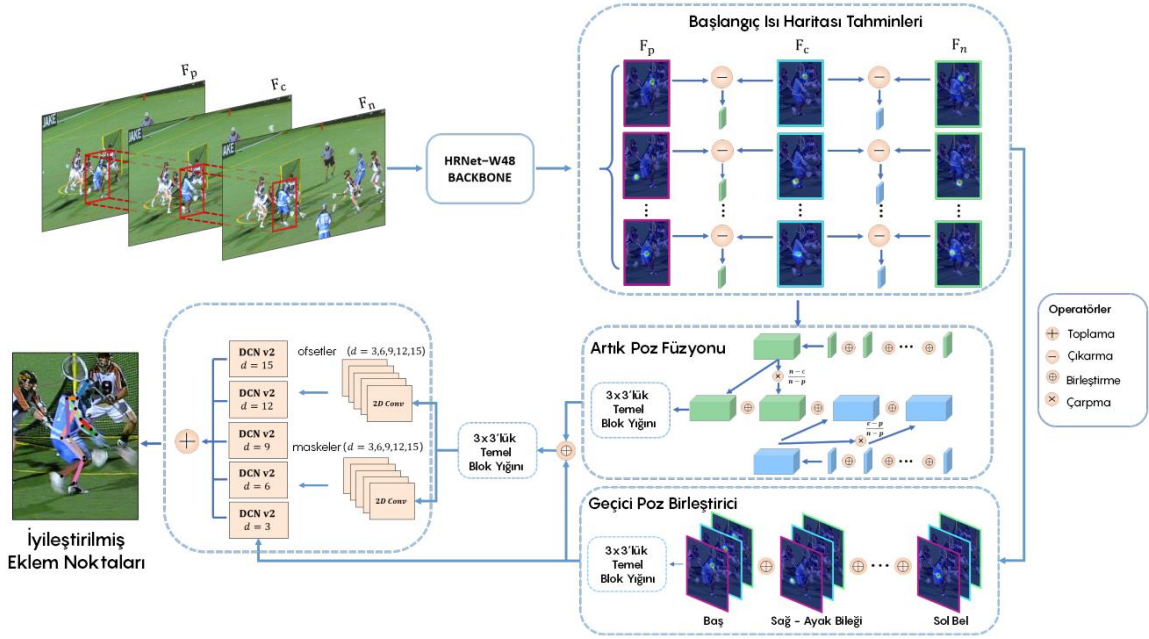
OpenPose gerçek zamanlı çok kişili poz tahmini yapan bir yöntemdir. Aşağıdan yukarı bir yaklaşıma dayanmaktadır. Bu yaklaşımda vücut parçaları model tarafından tespit edilir ve sonuçları elde etmek için son bir ayrıştırma işlemi yapılır. İnsan vücudunu, ellerini ve yüzdeki detay noktaları aynı anda gerçek zamanlı olarak tespit edebilir. Kişi başına bir görüntüde 130 anahtar nokta tespiti yapabilmektedir. Eklem noktalarının hesaplanma performansı, görüntüdeki kişi sayısından bağımsızdır. OpenPose Şekil 2.8'deki adımları izleyerek çalışmaktadır. Görüntünün tamamı Şekil 2.8 (a)'da, vücut parçası tespiti Şekil 2.8 (b)'de, parça ilişkilendirmesi Şekil 2.8 (c)'de görülmektedir. Kısmi parça alanları için ortaklaşa tahminde bulunmak üzere kısmi parçaların vektörel değerleri bir ESA'nın girdisi olarak alınır. Şekil 2.8 (d) ise ayrıştırma adımıdır. Vücut parçası adaylarını ilişkilendirmek için bir dizi ikili eşleştirme gerçekleştirir. Şekil 2.8 (e) görüntüdeki tüm insanlar için eklem noktaları arasında ikili eşleştirmelerin sonuçlarının tam vücut pozlarında birleştirilmiş gösterimidir.



Şekil 2.8. OpenPose'un genel işlem adımları [8]

2.3.2. DCPose yöntemi

Çok kişili poz tahmini çalışmaları arasında PoseTrack veri kümesi ile doğruluk değerleri en yüksek sonuçları elde eden çalışma DCPose olmuştur. Yukarıdan aşağı bir yaklaşımdır. Bu tür yaklaşımlarda öncelikle kişiler tespit edilir. Ardından bu kişilerin bulunduğu bilinen bölgede eklem noktaları tespit edilir. DCPose kişi başına 15 eklem noktası tespit etmektedir.



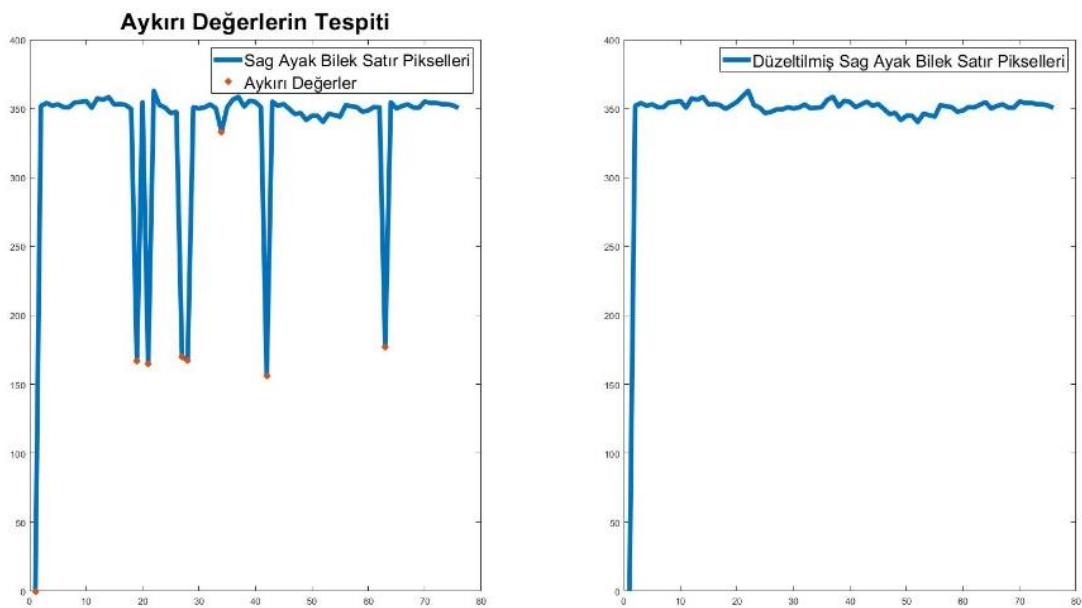
Şekil 2.9. DCPose yönteminin genel akış diyagramı [4]

Şekil 2.9'da gösterilen mevcut F_c görüntüsündeki kişilerin eklem noktalarının konumlarını bulmak amaçlanmaktadır. İlk olarak, c görüntüsündeki bir i . kişisi, ardışık görüntülerdeki (p, c, n) bölgeler birleştirilerek bir dizide tutulur. Bir HRNet [6] yapısı kullanarak, $h_i(p)$, $h_i(c)$, $h_i(n)$ başlangıç eklem noktalarının ısı haritalarını tahmin eder. Geçici poz birleştirici ağı etkili bir arama işlemi yapar. Ardından, artık poz füzyonu ağı, bölgelerin dışında kalan artık eklem noktalarını kişilerle birleştirmek için çalışır. Daha sonra ise poz düzeltme ağına girerek son çıktı olan 15 adet eklem nokta elde edilir.

2.4. Poz Takibi Ön İşleme Adımları

Bu çalışmada, video içindeki insan hareketlerinin ardışık görüntü karelerindeki değişimlerinin bir dizide tutulabilmesi amacıyla her bir kişinin takip edilmesi gerekmektedir. Bu nedenle her bir kişi için daha iyi takip performansı elde edilebilmesi amacı ile poz tahmini çıktılarına çeşitli ön işleme adımları uygulanmıştır. Ön işleme uygulaması beş ana

adımında gerçekleştirilmiştir. Birinci adım, eklem noktaları arasındaki aykırı değerlerin tespit edilerek düzeltilmesi işlemidir. Aykırı değerleri tespit etmek için bir videoda bulunan tüm eklem noktaları arasında inceleme yapılması gerekmektedir. Ardışık görüntüler arasında bir eklem noktasının değerlerinde büyük değişimler olmamalıdır. Eğer bu aykırı değerler poz tahmini algoritmasının çıktısı olarak elde edildiyse, bu değerler poz takibi adımına geçmeden önce düzeltilmelidir. Bu çalışmada, aykırı değerlerin tespit edilmesi için üç ölçekli mutlak medyan sapma kullanılmıştır. Bu yöntemde alt eşik, medyan ve üst eşik hesaplanmaktadır. Bu eşik değerlerinin dışında kalan değerler aykırı değer olarak tespit edilmektedir. Böylece piksel değerlerinde meydana gelebilecek aşırı değişimlerin tespit edilmektedir. Örneğin bir eklem noktasına ait ardışık 5 video karesinde tespit edilen satır piksel değerleri (50, 54, 59, 182, 53) şeklinde ise 4. görüntüdeki 182 değerinin poz tahmin yöntemi tarafından yanlış tahmin edildiği tespit edilerek bu aykırı değer için düzeltme işlemi yapılmaktadır. Aykırı değerın gerçek değerinin eğer önceki görüntü ile sonraki görüntü bir aykırı değere sahip değilse bu değerlerin arasında olması gerektiği ön görülmektedir. Bu nedenle aykırı değerın önceki ve sonraki değerleri aykırı değer değilse bu değerlerin aritmetik ortalaması alınarak düzeltme işlemi yapılmıştır. Şekil 2.10, tespit edilen aykırı değerler ve düzeltme sonucunda elde edilen değerleri göstermektedir. Grafiklerde x eksenini video kare numarasını y eksenini ise sağ ayak bileğinin bu karedeki satır pikselini göstermektedir. Sol tarafta aykırı değerlerin tespit edildiği grafik, sağ tarafta ise tespit edilen bu aykırı değerlerin düzeltildikten sonraki değerleri gösterilmektedir.



Şekil 2.10. Poz tahmini algoritmalarının çıktılarındaki aykırı değerlerin tespit edilmesi ve düzeltilmesi adımı

İkinci ön işleme adımı merkez nokta çıkarımıdır. Bu çalışmadaki poz takibi yöntemi her kişi için bir merkez nokta belirlenerek bu noktanın video boyunca takip edilmesini amaçlar. Bu yüzden merkez noktanın, anormal değişimlerden uzak bir nokta olması gerekir. İnsanların el, kol, bacak ve ayak noktaları gövde ve boyun noktasına göre daha hareketlidir. Bu nedenle merkez nokta olarak gövdeye bağlı bir nokta ya da boyun noktası merkez olarak kabul edilmelidir. Bu çalışmada merkez nokta olarak boyun noktası kabul edilmiştir ve boyun noktasının görünmediği durumlarda tespit edilen diğer eklem noktaları kullanılarak vektörel boyun yaklaşımı olarak isimlendirilen yöntem ile boyun noktasının konumu hesaplanabilmektedir. Vektörel boyun yaklaşımı gövde ve kafa arasında ya da omuzlar arasındaki Öklid mesafelerini kullanarak ve gövdenin eğimini hesaplayarak boyun noktasının değerlerini tahmin eder. Bu da poz takibi için eksik bilgi kullanılmamasını ve kararlı bir merkez noktası çıkarımının yapılmasını sağlar.

Üçüncü ön işleme adımı baş bölgesini sınırlayan alanın bulunmasıdır. Poz tahmini yöntemleri genellikle kulakları, gözleri, başın üst noktasını ve boynu tespit etmektedir fakat baş bölgesinin piksel konumlarının tespitini yapmamaktadır. Fakat PoseTrack veri kümesi, poz takibi başarımını ölçmek için bu bilgiyi istemektedir. Bu nedenle sonuçları değerlendirmeden önce bu ön işleme adımı ile analitik olarak baş bölgesini sınırlayan kutunun piksel değerleri hesaplanmıştır. DCPose baş bölgesinin sadece alt ve üst noktasını bulmaktadır. Kulakları tespit etmediğinden başın sağ ve sol tarafını bulamamaktadır. OpenPose ise yalnızca göz, kulak, burun ve boyun noktalarını bulmaktadır. Bu bilgiler kullanılarak baş bölgesini sınırlayan alan DCPose ve OpenPose için ayrı ayrı tespit edilmiştir. Şekil 2.11. örnek bir video karesi için poz tahmini algoritmalarından alınan sonuçları ve bu sonuçlara ek olarak hesaplanan baş bölgesinin konumunu göstermektedir.



Şekil 2.11. Poz tahmininden elde edilen eklem noktaları ile baş bölgesinin bulunması

Baş bölgesini sınırlayan alanı tespit etmek için eklem noktaları içinde sabit iki tanesi arasında referans bir uzunluk belirlenir. Bu referans uzunluğa göre sol göz-sol kulak ve sağ

göz-sağ kulak arasındaki Öklid mesafesi karşılaştırılır. Eğer sol göz ve sol kulak arasındaki mesafe sağ göz ve sağ kulak arasındaki mesafeden anormal şekilde büyükse referans uzunluk oranına göre kişinin kafasının bulunduğu görüntü koordinat sistemine göre başın dönük olduğu yönün çıkarımı yapılabilir. Kişinin yönü Şekil 2.12'deki son baş görüntüsünde olduğu gibi sol tarafa bakıyorsa baş bölgesini sınırlayıcı alan başın arka tarafını da çevreleyebilmek için sağ tarafı da kapsamalıdır. Doğru sınırlayıcı bölgeyi bulabilmek için bu çalışmada genişleme yönü, geçici baş merkezi ve genişleme miktarı olmak üzere 3 adımlı hesaplama önerilmektedir. Genişleme yönü başın görüntü koordinat sistemine göre mevcut yönü ve referans uzunluğa göre durumuna bakılarak hesaplanır. Geçici baş merkezi tespit edilen mevcut noktaların en büyük ve en küçük değerlerinin ortalaması alınarak hesaplanır. Genişleme miktarı ise referans uzunluğa göre göz-kulak arasındaki Öklid mesafesi arasındaki ilişkiye bağlı olarak hesaplanır. Önerilen yöntem başın 3 boyutta gerçekleştirebileceği dönüşlere dayanıklıdır.



Şekil 2.12. Baş pozisyonlarına göre sadece poz tahmini ile hesaplanan baş bölgeleri ve tespit edilmesi gereken baş bölgeleri

Dördüncü ön işleme adımında videonun başrol oyuncularları arasında olmayan ve yanlış pozitif olma ihtimali yüksek kişilerin filtrelenmesi işlemi yapılır. Videolarda kişilerin tespit edilme yoğunluklarını hesaplayarak belirli bir eşik değerinin altında videoda görünen kişiler silinir. Örneğin bu eşik değeri %10 olarak belirlenirse ve video 70 görüntüden oluşuyorsa videoda 7 görüntü karesinden daha az bulunan kişiler hiç tespit edilmemiş varsayılır. Geçici olarak başlatılan kimliklendirilmiş kişiler filtrelenerek iptal edilir. Bu ön işleme adımı videolarda bulunan başrol oyuncularının tespit edilmesi için de kullanılabilir.

Beşinci ön işleme adımı ise sadece DCPose poz tahmini yöntemi için uygulanmaktadır. Bu adım el ve ayak bilekleri için bulunan sonuçların iyileştirilmesine yöneliktir. DCPose'un el ve ayak bilekleri sonuçları, PoseTrack veri kümesinin gerçek değerlerine göre Öklid mesafesi olarak bakıldığında kısadır. Örneğin ayak bileği için bu durumu düzeltebilmek amacıyla öncelikle kaval kemiğinin görüntü koordinat sistemine göre yönü hesaplanmalıdır. Bu yönde eklem noktasının morfolojik kemik yapılarına göre uzatılması gerekmektedir.

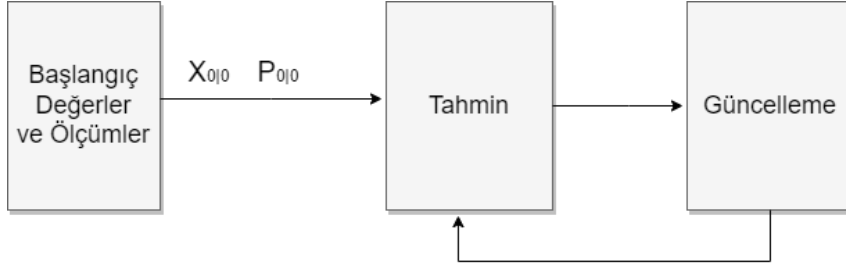
İnsanların iskelet yapısına göre uyluk kemiği ve kaval kemiği uzunlukları ile çalışmayı Hasegawa ve arkadaşları [54] yapmıştır. Buna göre kemikler arası oran kullanılarak ayak bileğinin uzatılması ile gereken konum tahmin edilmiştir. Deneysel çalışmalar bölümünde bu işlem adımının sonuçlar üzerindeki etkisi görülmektedir.

2.5. Poz Takibi

Poz takibi problemi videolarda bulunan çok sayıdaki kişinin pozlarını (duruşlarını) tahmin ederek video kareleri boyunca bu kişileri takip edebilmek için poz ile ilişkili her bir anahtar noktaya benzersiz kimlik numaraları atanmasıdır. Bu amaçla çeşitli algılayıcılardan elde edilen veriler kullanılarak bilinmeyen parametreleri tahmin etmek gerekir. Bu çalışmada Kalman filtresi ve KEYK kullanılarak gerçek zamanlı çalışabilen bir takip modülü geliştirilmiştir ve sonuçlar incelenmiştir. Geliştirilen ön işleme adımlarının poz tahmini yöntemlerinin kalitesini artırarak poz takibine etkisi de tespit edilmiştir. Bu amaçla PoseTrack veri kümesi üzerinde DCPose ve OpenPose poz tahmini yöntemlerinin başarımları değerlendirilmiştir. Ön işleme adımları ile sistemin başarımının her iki yöntem için de yükseldiği görülmüştür. Gerçek zamanlı çalışabilen, başarılı bir poz tahmini yöntemi olan ve düşük kaynak tüketimine sahip OpenPose yöntemi ile literatürde en iyi sonuçları veren DCPose yönteminin sonuçları incelendiğinde çok kişili poz takibi konusunda DCPose yönteminin daha iyi sonuçlar verdiği görülmüştür.

2.5.1. Kalman filtresi

Kalman filtresi doğrusal ve doğrusallaştırılmış sistemlerde sistem değişkenlerini kullanarak durum tahmini yapabilmek amacıyla kullanılır. Filtreleme işlemi genel olarak tahmin ve güncelleme olmak üzere iki adımda yapılır. Tahmin adımında hareket denklemleri ve ölçümleri kullanılarak bir sonraki zaman adımındaki durum ve kovaryans değeri tahmin edilir. Güncelleme adımında ise ölçüm değerleri kullanılarak durum ve kovaryans değeri düzeltilerek güncellenir. Kalman filtresinin çalışma prensibi Şekil 2.13'te akış diyagramı olarak verilmiştir.



Şekil 2.13. Kalman filtresi akış diyagramı

Şekil 2.13'teki döngü incelenirse $X_{0|0}$ ile gösterilen durum ve $P_{0|0}$ ile gösterilen durum kovaryansı için ilk değer ataması yapıldıktan sonra hareket modeline göre bir sonraki adımda kişilerin beklenen durumları ve beklenen kovaryans matrisi Eş. 12 ve Eş. 13 eşitlikleri ile hesaplanır. Eş. 14 eşitliği ile de güncellenen zamandaki ölçüm değeri tahmin edilir [55]. Tahmin adımında da Eş. 12-14 eşitlikleri kullanılır. Güncelleme adımında ise Eş. 15-19 eşitlikleri kullanılır [56, 57].

$$x_{k+1|k} = F_k x_{k|k} + G_u u_k \quad (12)$$

$$P_{k+1|k} = F_k P_{k|k} F_k^T + Q_k \quad (13)$$

$$z_{k+1|k} = H_{k+1} x_{k+1|k} \quad (14)$$

Eş. 12-14'teki $(k+1|k)$ gösterimi, k zaman anındaki değerlerle $k+1$ anındaki değerlerin tahmin edilmesidir. A priori olarak da bilinir. $(k+1|k+1)$ gösterimi ise a posteriori olarak bilinir ve güncelleme sonrası tahmin anlamına gelir. Burada k zaman anını, F durum geçiş matrisini, x durum vektörünü, G giriş geçiş matrisini, u giriş kontrol vektörünü, z ölçüm vektörünü ve H ölçüm matrisini ifade etmektedir. w değeri Q kovaryans matrisi ile sıfır ortalama sahip süreç gürültü vektörü, v ise R kovaryans matrisi ile sıfır ortalama sahip olan ölçüm gürültü vektörüdür. Güncellenen zaman anında $(k+1)$ durumu düzeltmek için $k+1$ ölçümü ile k zaman anındaki ölçüm kullanılarak tahmin edilen değer arasındaki fark kullanılır. Güncelleme yapabilmek için önce ölçüm tahmin kovaryansı S_{k+1} (Eş. 15) hesaplanarak ardından Kalman kazancının K_{k+1} (Eş. 16) hesaplanması gerekir [56, 57].

$$S_{k+1} = H_{k+1} P_{k+1|k} H_{k+1}^T + R_{k+1} \quad (15)$$

$$K_{k+1} = P_{k+1|k} H_{k+1}^T S_{k+1}^{-1} \quad (16)$$

Hareket modeline göre tahmin edilen değer ölçümle güncellenir. Yani boyun noktasının tespiti ile hareket modelini kullanarak bir sonraki görüntüde nerede olacağı tahmin edilir ve bu tahmin ile bir sonraki görüntüde gelen boyun noktası değeri arasındaki farka göre düzeltme yapılır. Bu noktada ölçüm ile tahmin arasında doğrusal bir ilişki olduğu varsayılır. Eş. 17'deki $z_{k+1} - z_{k+1|k}$ bölümü ölçüm yeniliği (measurement innovation) olarak bilinmektedir (Eş. 22). Bu değer in ortalaması sıfır olmalıdır. Ölçüm yeniliği $k + 1$ anındaki ölçüm değeri ile k anındaki ölçüm kullanılarak hesaplanan $k + 1$ 'deki tahmin arasındaki fark olarak da ifade edilebilir [56, 57].

$$X_{k+1|k+1} = X_{k+1|k} + K_{k+1}(Z_{k+1} - Z_{k+1|k}) \quad (17)$$

$$P_{k+1|k+1} = P_{k+1|k} - K_{k+1}S_{k+1}K_{k+1}^T \quad (18)$$

$$z_{k+1|k+1} = H_{k+1}Z_{k+1|k+1} \quad (19)$$

Deneylerde hareket modeli olarak sabit hızlı hareket modeli kullanılmıştır. Sabit hızlı hareket modeli doğrusal Kalman filtresi ile kullanılabilir. Bu modelde bir sonraki zaman adımında yapılan durum güncellemesi mevcut zamandaki durumun bir fonksiyonudur (Eş. 20)

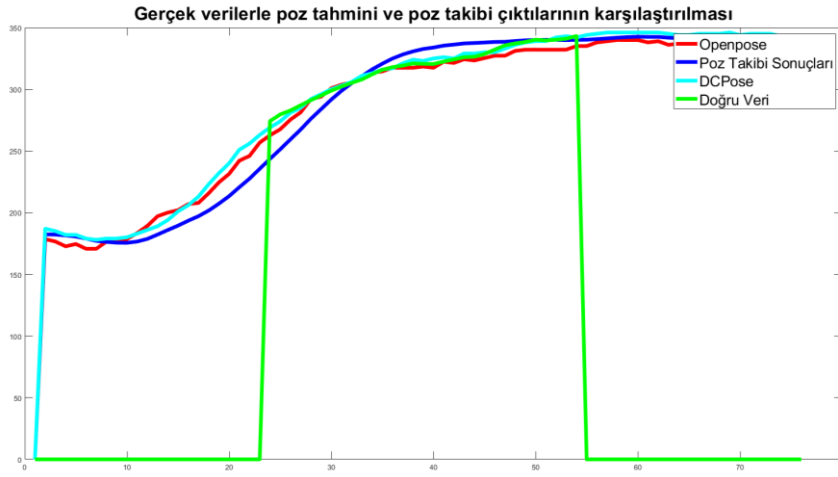
$$\begin{bmatrix} col_{k+1} \\ v_{col_{k+1}} \\ row_{k+1} \\ v_{row_{k+1}} \end{bmatrix} = \begin{bmatrix} 1 & T & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & T \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} col_k \\ v_{col_k} \\ row_k \\ v_{row_k} \end{bmatrix} \quad (20)$$

Burada col_k değeri poz tahmini çıktısındaki merkez noktanın sütun değeridir. col_{k+1} ise hareket modeli kullanılarak $k + 1$ anındaki durumun tahmin değeridir. Benzer şekilde row_k , row_{k+1} değerleri de merkez noktanın satır değerlerini ifade eder. v değerleri ise satır ve sütun pikselleri için hızları temsil eder.

Kalman filtresinde sistem durumları ile ölçümler arasındaki ilişkiyi tanımlamak için ölçüm modeli kullanılmaktadır. Ölçüm modeli durum vektörünün doğrusal bir fonksiyonudur. Poz tahmini çıktıları ölçümler olarak kullanıldığı için ölçüm modelinde merkez eklem noktasının satır ve sütun değerleri kullanılır (Eş. 21).

$$\begin{bmatrix} z_{col_k} \\ z_{row_k} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} col_k \\ v_{col_k} \\ row_k \\ v_{row_k} \end{bmatrix} \quad (21)$$

Poz tahmini çıktılarının gerçek değerlerden farkını belirleyebilmek için PoseTrack 2017 eğitim veri kümesindeki değerlerle DCPose ve OpenPose çıktıları kullanılmıştır. Bu farkların kovaryansı ölçüm gürültüsü olarak kullanılır. Veri kümesindeki bir videodan bir kişinin merkez noktası için poz tahmini değerleri Şekil 2.14'te gösterilmiştir. Bu tespitlerin gerçek değerden ne kadar saptığı incelenmiştir.



Şekil 2.14. Bir kişinin video boyunca merkez noktasının sütun değerlerinin gerçek verilerle poz tahmini ve poz takibi çıktıların karşılaştırılması

Kişilerin eklem noktalarının hareketleri ivmeli ya da hızlı hareketlerle tam olarak tanımlanamaz. Çünkü videoların içinde insanların hareketleri tamamen değişebilir ya da bir anda sabit olabilir. Kollar aniden dönüş hareketi yaparken gövde sabit gibi görünebilir. Bunun gibi çeşitli durumların da etkisini göz ardı etmemek ve sabit ivmeli hareketleri de yakalayabilmek için süreç gürültüsü matrisi kullanılabilir.

Bir ölçümün yeni bir hedef olup olmadığından emin olmak için geçici bir iz oluşturulmaktadır. Bu geçici iz belirli bir süre boyunca yeterli sayıda ölçüm alırsa onaylanmış bir iz olarak kabul edilir aksi takdirde silinir. M/N kriteri olarak tanımlanan bu çalışma mantığında kişilerin poz tahminlerinin ardışık N görüntünün M tanesinde ilişkilendirilmesi durumunda poz takibine başlamak için onaylanması gerekir. Bu şartı sağlayan kişilerin poz tahminleri izlerle ilişkilendirilir.

2.5.2. Veri ilişkilendirme

Poz tahmini çıktıları olan eklem noktası ölçümlerini kimliklendirilmiş kişilere yani izlere atayabilmek için bir veri ilişkilendirme yöntemi kullanmak gerekir. Bu çalışmada bu amaçla düşük kaynak tüketimi nedeniyle KEYK algoritması kullanılmıştır. KEYK veri ilişkilendirmesi geçitleme (gating) ve veri atama (data assignment) olmak üzere iki adımdan oluşur. Geçitleme adımında ölçümlerle tahminler arasında mümkün olmayan eşleştirmelerin dikkate alınmaması sağlanır. Kovaryans değerine bağlı olarak bir tahminin konumu etrafında geçit (elipsoid) oluşturulur. Elipsoidin içinde olmayan ölçümler dikkate alınmaz. Ölçümlerle tahminler arasındaki fark Eş. 22'deki gibi hesaplanır [58].

$$\Delta\xi(t) = (z_{k+1} - z_{k+1|k}) \quad (22)$$

Burada $\Delta\xi(t)$ ölçüm ve ölçüm tahmini arasındaki vektörel farktır. Kovaryans matrisi Eş. 15'deki gibi ve izlerle ölçümler arası d mesafesi Eş. 23'deki gibi hesaplanır.

$$d = \Delta\xi_k^T S^{-1} \Delta\xi_k \quad (23)$$

Oluşturulan elipsoidin sınırları D eşik değeri olarak kabul edilirse $d \leq D$ koşulunun karşılanması doğru bir ölçüm-iz ilişkilendirilmesi yapılabileceğini gösterir. Bu koşul ölçümün tahmin edilen izin hata elipsoidinin içinde olduğunu gösterir. Tek bir ölçüm, tek bir geçidin içine girdiyse doğrudan bir atama yapılır. Fakat bir geçidin içine birden çok ölçüm girebiliyorsa o zaman karmaşık durumlar ortaya çıkar. KEYK algoritması, bu karmaşık durumlarda da en uygun ilişkilendirmeyi yapar. Veri atama adımında Munkres, Jonker-Volgenant, Auction gibi maliyet hesapları yapan yöntemler kullanılmaktadır. Daha net sonuçlar verdiği için bu çalışmada maliyet hesabı için Munkres yöntemi kullanılmıştır. Munkres yöntemi algılamalar ve izler arasında bir maliyet matrisi hesaplar. Bir iz için en düşük maliyete sahip algılama arasında atama işlemi yapılır [58, 59].

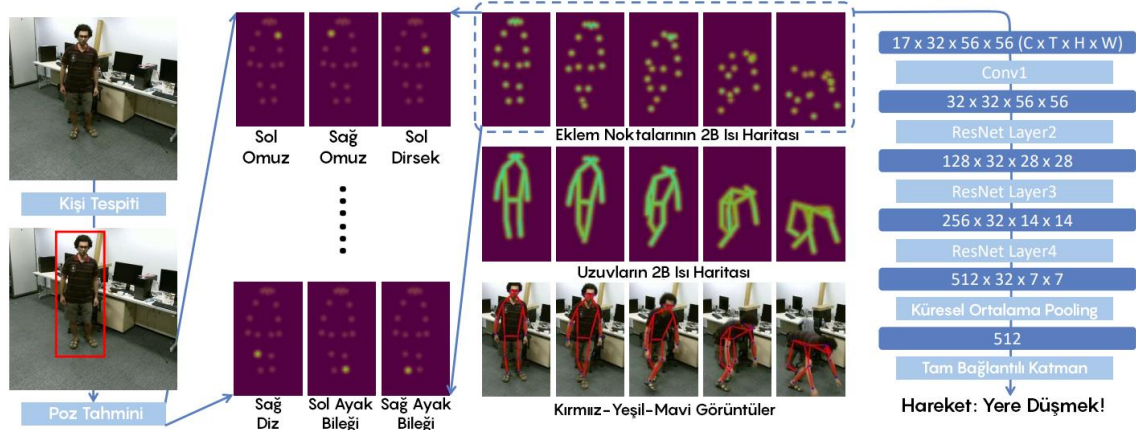
Mevcut izlerle ilişkilendirilmeyen ölçümler için yeni izler oluşturulur. Ayrıca izlerin onaylanması ya da silinmesi adımı ardışık M adet görüntünün N tanesinde izler ilişkilendirilirse onaylanma yapılacağından bu değerlerin belirlenmesi gerekmektedir. Belirtilen şartı sağlayan algılamalar izlerle ilişkilendirilir.

2.6. Hareket Tanıma

Video tanıma sistemlerinin aktif olarak gelişmesiyle birlikte video içeriklerinin anlaşılabilmesi için video içindeki insan hareketlerini tanıma önemli bir araştırma konusu haline gelmiştir. İnsan eylemlerini tanımayı amaçlayan çalışmalarda genellikle üç ana bilgi kullanılmıştır: derinlik [15, 16], RGB [17, 18] ve eklem noktaları [2, 19, 20, 60]. RGB, optik akış ve derinlik bilgisini kullanarak yapılan hareket tanıma algoritmaları hareketli kameralarla kaydedilen videolara karşı dayanıklı değildir. Sadece eklem noktaları ve eklem noktalarının dizisini kullanarak yapılan hareket tanıma; görüntülerdeki arka plan değişimi, parlaklık değişimi gibi görüntü özellikleri kullanılarak yapılanlara göre daha dayanıklıdır ve iyi sonuçlar vermektedir [61]. Bölüm 2.3'te anlatıldığı gibi insanların eklem noktaları tespit edilerek ve Bölüm 2.5'teki gibi poz takibi yapılarak video içinde eklem noktalarının dizileri biriktirilir. Poz tahmini, poz takibi ve hareket tanıma modüllerinin birbirinden bağımsız olarak incelenmesi herhangi bir modüldeki iyileştirmenin videodan insan hareketlerini tanımadaki etkisini ne kadar değiştirdiğini gösterebilmektedir. İnsan hareketlerinin anlamlandırılabilmesi için çeşitli yaklaşımlar mevcuttur.

İskelet tabanlı hareket tanıma yöntemleri arasında grafik evrişimli ağlar (GEA) yaygın olarak kullanılan başarılı yöntemlerden bir tanesidir. GEA, insanın her bir eklem noktasını uzamsal ve zamansal boyutlar olarak ardışık görüntüler arasında birleştirilir. Ardından grafik evrişim katmanları, eylem kalıplarını keşfetmek için oluşturulan grafiklere uygulanır. GEA ile geliştirilen son teknoloji yöntemler, diğer yöntemlere göre daha iyi sonuçlar vermektedir. GEA tabanlı yöntemlerin de bazı noktalarda başarısızlıkları mevcuttur. Örneğin, GEA insan eklemlerinin koordinatlarını doğrudan kullanırken farklı poz tahmini algoritması kullanıldığında koordinatların dağılımlarında kaymalar meydana gelebilmektedir. Koordinatlardaki bozulma ise farklı tahminlere yol açmaktadır [62]. İskelet temsiliinin grafik formu; RGB ve optik akış gibi ek bilgilerin kullanılabilmesi açısından sınırlıdır. Dolayısıyla GEA'nın bu yapısı farklı kombinasyonlarla çalışabilme etkinliği bir dezavantaj sayılabilmektedir. İskelet tabanlı eylem tanıma konusunda literatürdeki en başarılı yöntemlerden bir tanesi yapısı Şekil 2.15'te gösterilen PoseC3D [60] olarak bilinen eylem tanıma yöntemidir. PoseC3D yöntemi GEA'nın sınırlı ve dezavantajlı olduğu durumlara yönelik geliştirilen çözümleriyle FineGYM [63], NTURGB-D [64], Kinetics-400 [65] ve Voleybol [66] dahil olmak üzere çeşitli veri kümelerinde literatürdeki en başarılı yöntemdir. PoseC3D, videodaki her görüntü için önce 2B poz tahmini ve buna bağlı olarak zaman dizisi

boyunca eklemlerin ve uzuvların ısı haritalarını biriktirerek 3B ısı haritası hacimleri oluşturur. 3B ısı haritası hacimlerini sınıflandırmak için de 3B-ESA kullanmaktadır.

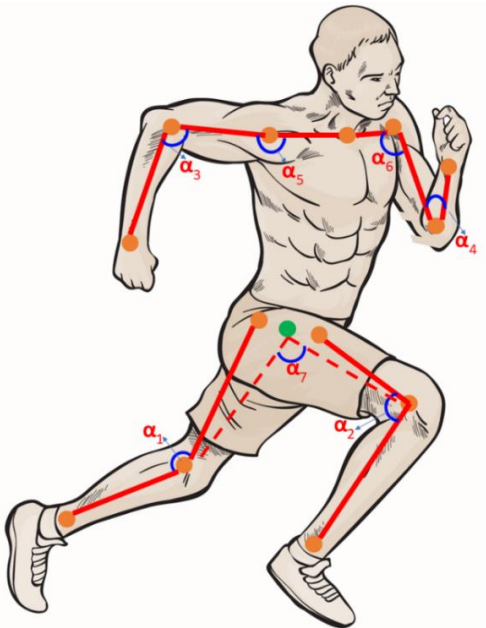


Şekil 2.15. PoseC3D’nin hareket tanıma için kullandığı model ve genel yapısı [60]

İnsan eklem noktalarının zamansal dizilerle işlem yapılabilmesini sağlayan yaklaşımlardan biri özyinelemeli sinir ağları (ÖSA) olarak bilinir. ÖSA’ların başlıca uygulama alanları; farklı diller arası makine çevirisi, ses tanıma ile cümlelerin yazıya dönüştürülmesi, nota dizileri meydana getirerek müzik oluşturma, bir cümlenin olumlu ya da olumsuz anlamda olduğunun tespiti ve video içeriğindeki hareketlerin anlamlandırılması gibi birçok konuda başarıyla kullanılmaktadır. ÖSA’lar, etkisi kaybolan gradyanlar (vanishing gradient) için geliştirilen UKSB yapılarına evrilerek daha başarılı sonuçlar vermektedir. Tüm makine öğrenmesi modellerinde olduğu gibi UKSB için veri kalitesinin yüksek olması daha iyi sonuçların alınmasını sağlar. Yao ve diğerleri [21], iskelet tabanlı özelliklerin görünüm tabanlı özelliklerden daha iyi performans gösterdiğini aynı veri kümesini ve sınıflandırıcıyı kullanarak kanıtlamıştır. ÖSA ile yapılan güncel araştırmalar görüntü dizilerinin işlenmesi üzerinde önemli etki oluşturmuştur [25-27]. İnsan eklem noktalarına dayalı çok katmanlı ÖSA yapıları ile hareket tanıma konusunda başarılı sonuçlar elde edilmiştir [19, 20, 23, 24]. Mevcut çalışmalarda girdiler eklem noktalarının görüntüdeki piksel konumları ile sınırlıdır. Du ve diğerleri [19] kolların bacakların ve gövdenin ilişkilerini ÖSA ile modellemektedir ve ardışık eklemler arasındaki ilişkileri dikkate almamaktadır. Zhu ve diğerleri [20] ardışık olmayan uzuvlar arasındaki ilişkilerden tam bağlantılı UKSB’ye karma mesafelerden oluşan bir düzenleme terimi ekler. Bu yöntemler vücut bölümleri arasındaki ilişkileri başarılı bir şekilde inceler ancak vücudu parçalara bölmek her zaman iyi sonuçlar vermeyebilir. Mevcut çalışmalar eklem noktaları arasındaki geometrik ilişkileri kullanarak belirli hareketleri tanımayla yönelik başarılı çalışmalardır. Bu çalışmaları ilham alarak ardışık eklemler

arasındaki açıları kullanarak eklemlerin açılıp kapanmasını ve birbirlerine uzaklıklarını temel alan farklı bir yöntem geliştirilmiştir. Bu çalışmada doğrudan hareketleri sınıflandırmak yerine hareket ettiğinden emin olunan uzuvları sınıflandırarak bu hareketlerin meydana getirebileceği insan hareket sınıfı eşleştirmesi ile hareketleri tanıyabilecek ara bir iyileştirme adımı önerilmektedir. Örneğin; diz çöken bir kişi çeşitli hareketler yapıyor olabilir. Bu diz çökme hareketini, evlilik teklifi anı ile ilişkilendirmek videonun kurgusu ile ilgilidir. Bu nedenle bu çalışmada, mevcut insan hareketlerinin tamamının sınıflandırılabilmesi için yeni bir muhtemel yapay hareket alt sınıfları meydana getirilmektedir.

Bu tez çalışmasında, diğer çalışmalardan farklı olarak poz tahmini çıktıları üzerinden tanınmak istenen hareketlerin doğrudan o hareket sınıfı için eğitilmesi yerine hareketleri gerçekleştiren eklemler hakkında bilgiler üreten bir derin öğrenme yapısı tasarlanmıştır. Böylece fiziksel olarak birbirine benzeyen ama gerçekte farklı hareketler olan dans etmek ve sarılmak gibi hareketlerin de ayırt edilebileceği bir yöntem önerilmektedir. Eklem noktalarının piksel konumları ile birbirleri arasındaki geometrik ilişkileri hesaplayarak UKSB'nin kaliteli verilerle beslenmesi sağlanmaktadır. Çünkü insanlar hareket ederken fizyolojik olarak kaslarını hareket ettirir ve çeşitli eklem noktaları arasındaki açılar ve eklem noktalarının birbirine göre Öklid mesafeleri değişmektedir. Buna dayanarak; belirli eklem noktaları arasındaki geometrik ilişkiler tanımlanmaktadır. Bu ilişkiler Şekil 2.16'da gösterilmektedir.



Şekil 2.16. Eklem noktaları arasında hesaplanan açı ilişkileri

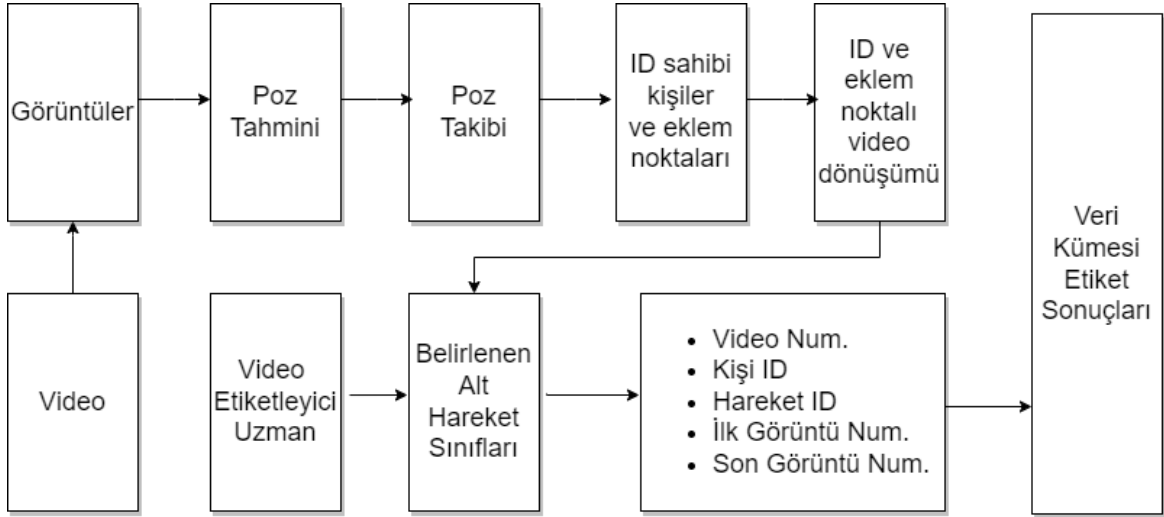
Önceki hareket tanıma çalışmaları genellikle zıplamak ve yürümek ya da yere düşmek ve koşmak gibi birbirinden kolaylıkla ayırt edilebilen hareketleri tanımak üzerine yoğunlaşmıştır [60, 64, 67]. Örneğin iki başrol oyuncusunun dans etmesi ile sarılması arasındaki farkı ayırt edecek kadar iyi tanıma yapabilecek ve bunun gibi karmaşık problemler ile ilgili çalışmalar mevcut değildir. Literatürdeki çalışmalar genellikle yere düşmek ve yürümek gibi birbirinden çok farklı hareket sınıfları arasında tanıma işlemi gerçekleştirir. Bu nedenle bu çalışmada Bölüm 2.9’da bahsedilen karmaşık ve gerçek hayat problemlerini yansıtan veri kümesi *PoseDrone* ortaya çıkarılmıştır.

Kol ve bacakların yapabileceği alt hareketler görüntü koordinat sistemine göre 4 farklı açı bilgisi hesaplanarak modellenmiştir. Açı değerlerindeki değişime bağlı olarak hareketlerin tanınması işleminde UKSB modelinin yapısındaki gradyan işlemleri sebebiyle başarı sağlayacağı sezgisel olarak keşfedilmiştir. Geliştirilen yöntemin avantajı, basit geometrik özelliklerin eklem noktalarının piksel konumlarından daha iyi performans göstermesidir. Bu özelliklerin anlaşılması ve analizi kolaydır. Sonuç olarak, iskelet tabanlı eylem tanıma için geometrik özellikleri hesaplayan ve UKSB modelini kullanan basit ve başarılı bir hareket tanıma algoritması geliştirilmiştir.

Bölümün devamında hareket tanıma için gerçekleştirilen veri etiketleme adımından bahsedilecektir. Ardından veri hazırlama işlemleri, derin öğrenme yapısı, yapay hareket alt sınıfları ve yapay hareket alt sınıflarının hedef hareketlere dönüştürülmesi işlemleri açıklanacaktır.

2.6.1. Veri etiketleme

Bölüm 2.9’da bahsedilen *PoseDrone* veri kümesindeki hareketlerin tanınabilmesi için geliştirilen poz takibi yöntemi ile veri etiketleme adımı hızlandırılmıştır. Veri etiketleme için geliştirilen yöntemin blok diyagramı Şekil 2.17’de görülmektedir.



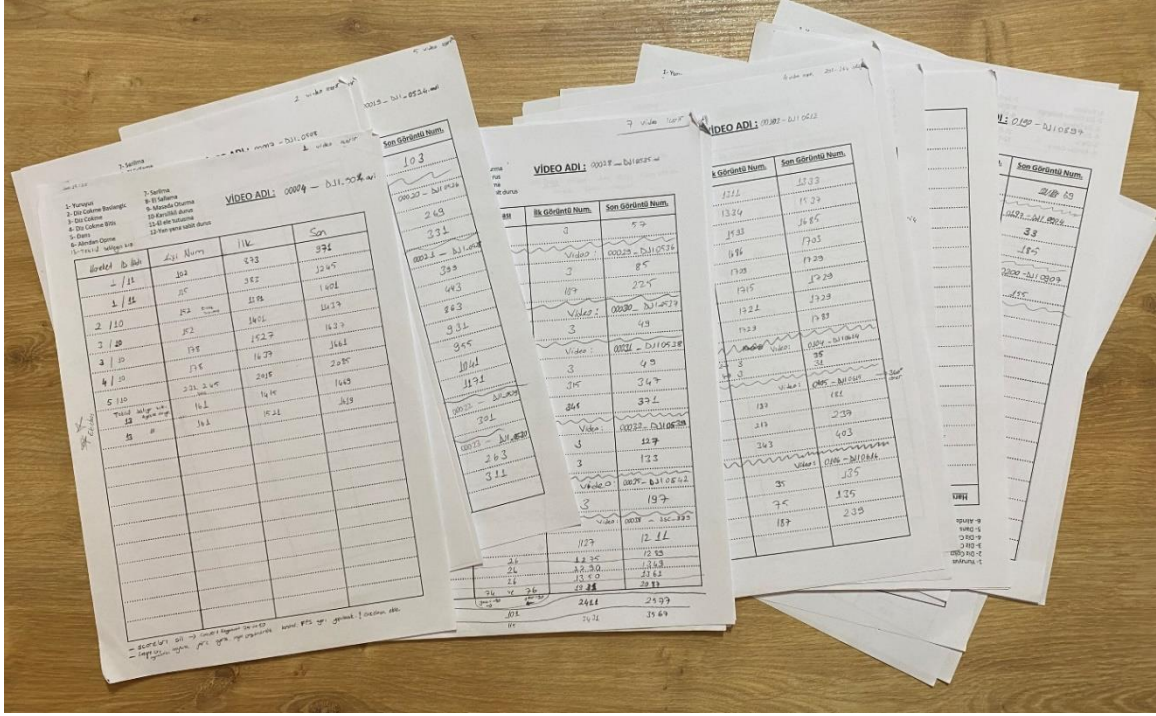
Şekil 2.17. Veri etiketleme adımının blok diyagramı

İlk olarak veri kümesi için kaydedilmiş videolar, görüntülere dönüştürülür. Bu adımda videonun saniyedeki görüntü sayısına göre görüntülere geçiş aşaması seyrekleştirilebilir. Seyrekleştirme işlemi veri çeşitliliğini artıracığından tercih edilebilir. Burada en önemli nokta tüm videoların saniyedeki görüntü karelerinin eşit olmasıdır. Böylece etiketli verilerin dağılımlarındaki ağırlıklar eşit olacaktır. Görüntüler elde edildikten sonra poz tahmini yöntemlerinden en düşük işlem maliyetine sahip olan OpenPose kullanılarak görüntülerdeki kişilerin eklem noktaları tespit edilir. Her görüntü için .json uzantılı dosyalar elde edilir. Json dosyaları okunarak kişilerin eklem noktaları elde edilir. Görüntüler, videoların içinden elde edildiği için buradan bulunan veriler de zamana bağlı olarak kullanılabilir. Böylece Kalman Filtresi ve KEYK yöntemi kullanılarak geliştirilen poz takibi yöntemi ile kişiler video boyunca takip edilir ve tüm hareketleri bir yapıcı dizi içinde tutularak saklanır. Ardından poz tahmini çıktısı olarak elde edilen görüntüler tüm hareketleri tutan yapıcı dizi içindeki verileri de kullanarak yeniden videolara dönüştürülür. Böylece veri etiketleyici uzman kişiler videoları istedikleri gibi durdurabilir, başa sarabilir ve hareketlerin detaylarını rahatlıkla inceleyebilir. Şekil 2.18’de uzman kişinin etiketlediği sahnelerden diz çökme hareketini yani evlilik teklifi anını içeren bir görüntü verilmiştir.



Şekil 2.18. *PoseDrone* veri kümesinden diz çökme hareketinin etiketlendiği sahneden bir görüntü

Bir diğer kolaylık ise tüm görüntüleri tek tek etiketlenmek zorunda değildir. Bu veri etiketleme açısından çok kıymetli bir kolaylık sağlamaktadır. Örneğin; bir kişi $t=1$ zamanında dans ediyorsa ve $t=150$ zamanına kadar dans etmeye devam ediyorsa aralardaki eklem noktaları da etiketleyici uzman tarafından onaylandıysa her görüntü karesi için etiketleme yapılmasına gerek kalmamaktadır. 2 adet veri etiketleyerek 150 adet veri etiketlemiş olur. Bu sırada, veri etiketleyici uzman; etiketlediği videonun numarasını, poz takibi sonucu olan kişi ID'lerini, hareket ID'lerini ve hareketin başladığı görüntü numarasını ve hareketin bittiği görüntü numarasını kaydeder. 33 binden fazla etiket verisi Şekil 2.19'deki gibi etiketlenmiştir.



Şekil 2.19. 33 bin adet veri etiketleme işlemi yapılırken doldurulan dosya kağıtları

En sonunda bu dosya kağıtları Excel formatına aktarılarak buradan okuma işlemi yapılmıştır. Böylece hareketlere ait veri etiketlerinin tüm detayları hakkında bilgi elde edilmiştir. Etiketleme işlemine başlarken 12 adet farklı hareket için etiketleme yapılmıştır. Fakat bazı hareketlerin etiket sayısının çok az olduğu veri hazırlama işlemleri esnasında tespit edildikten sonra etiketlenen bazı hareketler eğitimde kullanılmamıştır.

2.6.2. Veri hazırlama işlemleri

Derin öğrenme modelinin eğitiminden önce verilerin hazırlanması adımı büyük önem taşımaktadır. Bu bölümle ilgili gerçekleştirilen deneysel çalışmalara ait sonuçlar Bölüm 3'te verilmiştir. Veri etiketleme adımının sonuçları olarak kişilerin eklem noktaları ve hedef hareket sınıfları elde edilmiştir. Hareket numaraları, bu hareketlerin isimleri ve her bir sınıfa ait etiket sayıları Çizelge 2.2'de verilmiştir. Verilerin hazırlanması adımı ilk olarak etiketleme adımıyla doldurulan Excel dosyasından veriler okunmaktadır. Excel dosyasındaki bir hedef hareket sınıfına ait aksiyonu gerçekleştiren kişinin numarası ve harekete başladığı görüntünün ilk numarası ve hareketin bittiği görüntünün numarası bulunmaktadır. Her bir kişinin gerçekleştirdiği hareket için eklem noktaları ve hareket sınıfı bir dizide tutulur. Bu etiketleme işleminin sonrasında insanların hareketlerinin tanınması işleminin poz tahmini yönteminin doğru çıktılar üretip üretmemesine dayanaklı olabilmesi

için iki farklı filtreleme adımı uygulanmaktadır. Birincisi eklemler arası açılara dayalı filtreleme, ikincisi ise eklem uzunluklarına göre filtrelemedir. Filtreleme işlemleri hakkında detaylar Bölüm 3’te verilmiştir.

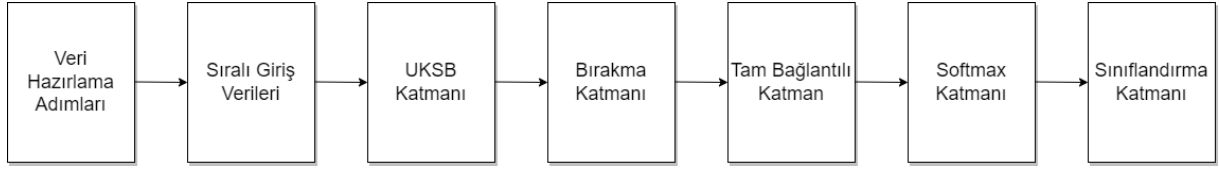
Çizelge 2.2. Her bir hareket sınıfı için elde edilen veri etiketi sayıları

Hareket Numaraları	Hareket Adları (Etiketler)	Etiket Adedi
1	Yürüyüş	4349
2	Diz Çökme Başlangıç	278
3	Diz Çökme	1644
4	Diz Çökme Bitiş	232
5	Dans	10443
6	Alından Öpme	0
7	Sarılma	1414
8	El Sallama	77
9	Masada Oturma	3247
10	Karşılıklı Duruş	5206
11	El Ele Tutuşma	5136
12	Yan Yana Sabit Duruş	1003

2.6.3. Derin öğrenme modeli

Derin öğrenme modelinde ilk olarak veri hazırlama adımları gerçekleştirilerek verilerin tüm veri kümesi için aynı formata getirilmesi gerekmektedir. Şekil 2.16’da gösterilen $\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \alpha_7$ açıları ile veri listesi hazırlanmıştır. Bu listeden veriler UKSB yapısına uygun olacak şekilde, her bir kişi için videoda devam eden hareket boyunca farklı zaman dizilerinde tutulur. Bu dizilerin satır sayıları tüm veri kümesinde yedi açı bilgisinden oluştuğu için yedi satırdır. Sütun sayıları hareketin yapıldığı süreye göre değişiklik göstermektedir. UKSB katmanında sigmoid ve tanh aktivasyon fonksiyonları kullanılarak olasılıklar hesaplanarak ağırlıklar elde edilir. Yinelemeli olarak değerler güncellenir. Bırakma katmanında belirlenen olasılık değerinin altında kalan değerler sıfırlanır. Tam bağlantılı katmanda ise ağırlıkların öğrenme faktörü, kayıp fonksiyonunun ağırlık faktörü, bias öğrenme oranını ve kayıp fonksiyonunun bias faktörü değerleri ayarlanarak çıktı olarak kaç farklı sınıf için sınıflandırma yapılıyorsa bunun sayısı belirlenir. Hemen ardından

softmax katmanı kullanılarak sonuç sınıfları elde edilir. Doğru verilerle karşılaştırılarak kayıp fonksiyonu ile sonuç hesaplanarak ağırlıklar özyinelemeli olarak güncellenir.



Şekil 2.20. Tek bir UKSB katmanı ile sınıflandırma

UKSB, ÖSA'nın kaybolan gradyan probleminin [68] üstesinden gelen uzun vadeli bağımlılıklarını modelleyebilen gelişmiş bir yapıdır. ÖSA'dan farklı olarak UKSB nöronu; giriş kapısı, çıkış kapısı, bir hücre ve bilginin nörona nasıl girip çıkacağını belirleyen bir unutma kapısından oluşmaktadır. Giriş ve aktivasyon değerleri ilk değer atamaları yapılır. Ardından zaman dizisindeki her bir zaman adımı için aynı hesaplama adımları gerçekleştirilir. Giriş değeri ile aktivasyon değeri sigmoid fonksiyonu ile çarpılarak hücreden gelen değeri yani bellekteki değeri ne kadar unutması gerektiği Eş. 26 ile hesaplanır. Benzer şekilde giriş ve aktivasyon değeri ile sigmoid fonksiyonu çarpılarak güncelleme olasılığı Eş. 25. ile çıkış olasılığı da Eş. 27 hesaplanır. Diğer taraftan Eş. 24'teki gibi tanh aktivasyon fonksiyonu ile bir sonraki zaman anı için aday değer hesaplanır. Eş. 28'de ise bu aday değer ne kadarının bir sonraki adımda kullanılacağı güncelleme olasılığı ile hesaplanır ve önceki adımdaki bellek değerinin ne kadarının unutulacağı Eş. 26'da hesaplanan olasılık değeri ile çarpılarak toplanır. Çıkış kapısında ise hücre birimi tanh ile aktivasyona girerek ne kadarı dışarı verilmeli diye hesaplamaktadır. Bu aktivasyon değeri Eş. 29'da hesaplanmaktadır ve aynı zamanda hücrenin çıkış değerini de etkilemektedir. Bu eşitliklerde W_c hücrenin ağırlık değerini, W_u güncelleme kapısına bağlı ağırlık değerini, W_f unutma kapısına bağlı ağırlık değerini, W_o çıkış kapısına bağlı ağırlık değerini, Γ_u güncelleme olasılığını, Γ_f unutma olasılığını, Γ_o çıkışa verilme olasılığını, $\tilde{c}^{<t>}$ hücrenin aday değerini, $a^{<t-1>}$ önceki adımdan gelen aktivasyon değerini, $x^{<t>}$ şimdiki zamandaki giriş değerini, b_c hücrenin aday değeri hesaplanırken eklenen bias değerini, b_u güncelleme adımındaki bias değerini, b_f unutma adımındaki bias değerini, b_o çıkış adımındaki bias değerini, $c^{<t>}$ hücrenin değerini, $c^{<t-1>}$ bir önceki zaman anından gelen hücrenin değerini göstermektedir.

$$\tilde{c}^{<t>} = \tanh\left(W_c \left[a^{<t-1>}, x^{<t>} \right] + b_c\right) \quad (24)$$

$$\Gamma_u = \sigma \left(W_u \left[a^{<t-1>}, x^{<t>} \right] + b_u \right) \quad (25)$$

$$\Gamma_f = \sigma \left(W_f \left[a^{<t-1>}, x^{<t>} \right] + b_f \right) \quad (26)$$

$$\Gamma_o = \sigma \left(W_o \left[a^{<t-1>}, x^{<t>} \right] + b_o \right) \quad (27)$$

$$c^{<t>} = \Gamma_u * \tilde{c}^{<t>} + \Gamma_f * c^{<t-1>} \quad (28)$$

$$a^{<t>} = \Gamma_o * c^{<t>} \quad (29)$$

2.6.4. Yapay hareket alt sınıfları

Yapay hareket alt sınıfları kişilerin görüntü koordinat sistemine göre düz ya da yan durmasına dayanıklı olacak şekilde tasarlanmıştır. Alt sınıfların sağladığı sonuçlar aynı anda birçok hareketi tanımaya da olanak sağlayarak kişilerin yürürken el ele tutuşmasını algılamak gibi yetenekleri de beraberinde getirir. Çizelge 2.2'deki hedef hareketlerin her biri yapay hareket alt sınıfı ile ifade edilebilir. Bir kişinin evlilik teklifi sahnesini gerçekleştirmesi için diz çökme hareketi yapması gerekir ya da sarılma hareketini yapabilmesi için el bileklerini omuz hizasına kaldırarak yaklaştırması gerekir. Her bir hareket sınıfı en az bir hareket alt sınıfı ile bağlantılıdır. Buna bağlı olarak yapay hareket alt sınıfları ile hedeflenen hareketler arasında bir veri ilişkilendirme yapılması gerekmektedir.

Çizelge 2.3. Yapay hareket alt sınıfları

Yapay Hareket Alt Sınıfları	Hedef Hareket Sınıfları
Dizlerin bükülmesi	Evlilik Teklifi (Diz Çökme)
Dizlerin düz durması	Sabit Duruş, Karşılıklı Duruş
Bacakların birbirine göre açılıp kapanması	Yürüyüş
Bacakların sabit durması	Sabit Duruş, Karşılıklı Duruş
Kollar omuzdan aşağıda	Sabit Duruş, Karşılıklı Duruş
Kollar omuz hizasında	Dans, Sarılma
İki farklı kişinin ellerinin yaklaşması	El Ele Tutuşma

Veri ilişkilendirme için her bir alt hareket diğer tüm hedef hareket sınıfları ile kıyaslanır. Kovaryans değerlerine bağlı olarak hareketler verileri arasında birbirlerine en yakın olanlar ilişkilendirilir. Veri ilişkilendirme için Bölüm 2.5.2'de verilen KEYK kullanılmıştır.

İlişkilendirme yapılmayan hareketler için hedef hareket sınıfı belirsiz olarak isimlendirilmektedir. İlişkilendirme yapılan hareket sınıfları ise ilişkilendirildikleri hedef sınıf ya da sınıflar ile anılmaktadır.

2.6.5. Hedef hareket sınıfları

Hedef hareket sınıfları, hareket alt sınıfları ile ilişkilendirilen bir ya da birden çok sınıftan oluşabilir. El ele yürüyen başrol oyuncularını, hem el ele tutuşma hareketini hem de yürüme hareketini icra etmektedir. Etiketlenmiş hedef hareket sınıfları ile hareket alt sınıfları arasındaki yakınlıklar açı ve uzunluk değerleri ile ayrı ayrı hesaplanmaktadır. Her iki değerde de ilişkilendirme aynı hedef sınıf ya da sınıflarla gerçekleştirildiyse hedef sınıf ataması yapılır. Gerçekleştirilmediyse sınıf bilinmiyor olarak atanır. En yakın değerlerle ilişkilendirme için Bölüm 2.5.2’de verilen KEYK veri ilişkilendirme yöntemi kullanılmaktadır. Alt sınıflardaki detaylara bağlı olarak hedeflenen hareket sınıflarının daha detaylı sonuçlarının elde edilebileceği öngörülmektedir. Hedef hareket sınıfları, veri etiketleme adımında etiketleyici uzman tarafından belirlenen ve her görüntüde analiz edilen sınıflardır. Bu sınıflar Çizelge 2.2’de de etiket sayıları ile birlikte belirtilmiştir. Başrol oyuncularının aynı anda iki hareket sınıfı ile birden ilişkilendirildiği el ele tutuşma ve yürüyüş sınıflarından örnek bir sahne Şekil 2.21’de görülmektedir. Şekil 2.22’de evlilik teklif anını gösteren diz çökme sahnesi bulunmaktadır. Şekil 2.23’de dans etme sahnesi gösterilirken Şekil 2.24’te ise başrollerin sabit duruşu sahnesi görülmektedir.



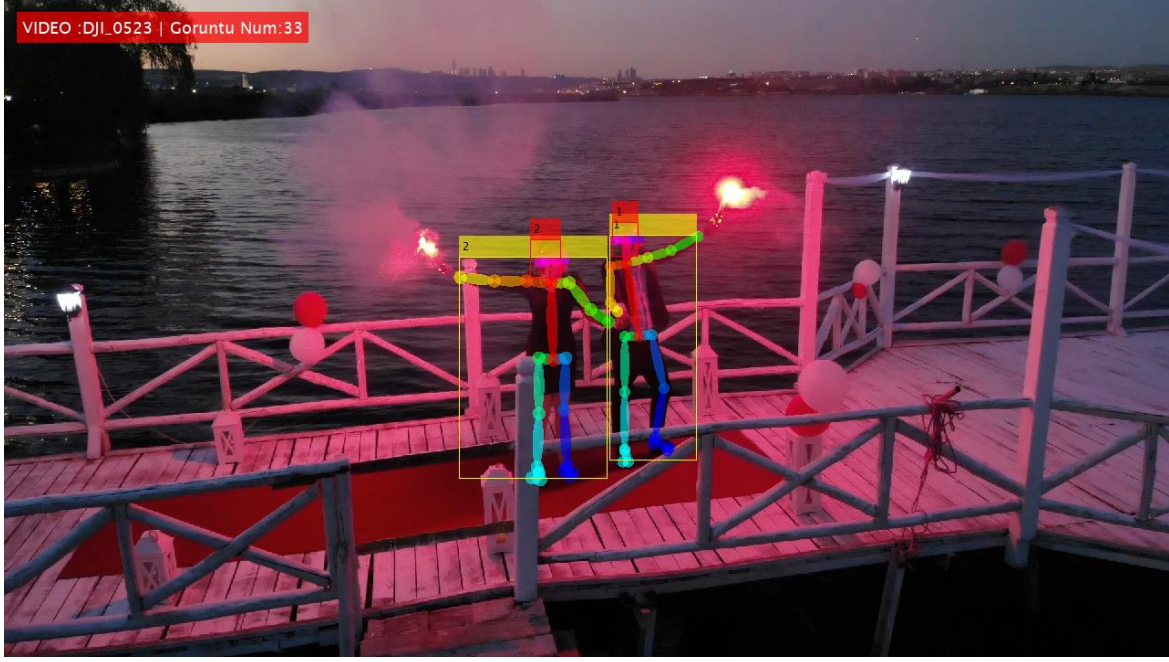
Şekil 2.21. El ele tutuşma ve yürüyüş sınıflarının örnek gösterimi



Şekil 2.22. Diz çökme hareket sınıfı örnek gösterimi



Şekil 2.23. Dans hareket sınıfı örnek gösterimi



Şekil 2.24. Sabit duruş hareket sınıfı örnek gösterimi

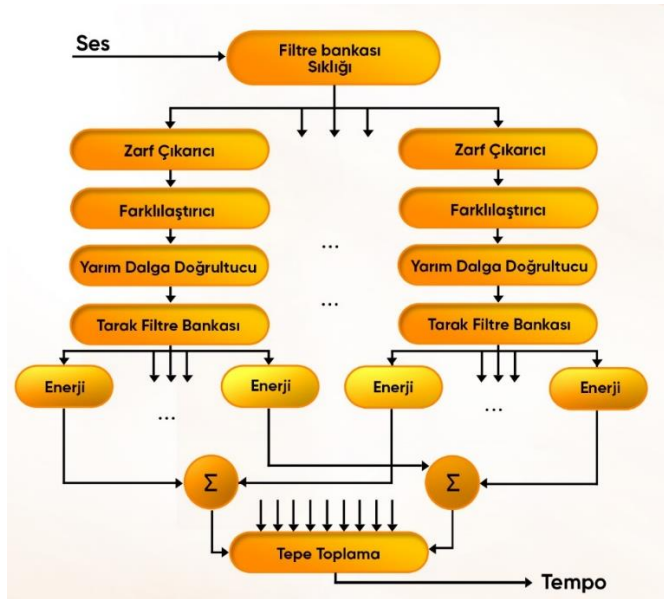
2.6.6. Hareket Sınıflarının Birleştirilmesi

Derin öğrenme modelinde girdi olarak açı bilgileri ve uzunluk bilgileri kullanılmıştır. Bu bilgiler birbirine benzeyen ancak veri giriş sayısı, gizli katman sayısı gibi farklılıklara sahip UKSB kullanılan modellerde eğitilmiştir. Herhangi bir test verisi ile ağ yapısı kullanılarak çıkış değeri tahmin edildiğinde hem açı bilgileri için hem de uzunluk bilgileri için hareket sınıfı bilgileri elde edilir. M/N kriteri kullanılarak ardışık M görüntünün N tanesinde hareket aynı sınıfa ait çıkış değerine sahipse hareketin ardışık M görüntü boyunca aynı hareket sınıfına ait olması gerektiği kabul edilmiştir. Çünkü poz tahmini işleminden kaynaklı olarak bu değerlerde hata olabilmektedir. Bu aşamadan sonra her iki tahmin sonucu birbiri ile aynı ise bu sonuç doğru kabul edilmiştir. Farklı ise hareket tahminlerinin beklenen değerden sapmasına bakılmıştır. Açı değerleri o hareket sınıfının ortalama açı değerinden ne kadar saptığı hesaplanmıştır. Aynı şekilde uzunluk değeri için de bu hesap yapılmıştır. Standart sapması düşük olan hareket sınıfı, çıkış hareket sınıfı olarak kabul edilmiştir.

2.7. Müzik İşleme ve Vuruş Tespit Algoritması

Video özetleme yöntemi için parçalanacak videoların uzunlukları arka planda çalan müziğin ritminin periyoduna göre belirlenmiştir. Bunun için kullanılması planlanan önceden belirlenmiş müzik sinyalinin vuruşlarının yani ritminin algılanması için [69]'de verilen algoritma mevcut probleme uyarlanmıştır. Çalışmanın genel blok diyagramı Şekil 2.25'te

görülmektedir. Temel olarak algoritma belirlenen müzik sinyalindeki darbeleri vurgulamak ve bu sinyaldeki darbelerin periyodunu bulmaya yöneliktir. Müzik sinyali Fourier dönüşümü ile frekans düzlemine geçirilir. Ardından frekans bandında sinyallerin zarfı çıkarılarak sesteki ani değişiklikleri vurgulamak için sinyaller bir filtreden geçirilir. Bu filtre; sütun vektöründe tutulan bir zaman sinyalini alır ve her sütun farklı bir bandı temsil edecek şekilde frekans düzleminde sinyalin bir vektörünü verir. Bant sınır değeri, her elemanın bir bandın frekans sınırlarını temsil ettiği bir sıranın vektörüdür. Son bant, bant sınır değeri ve en büyük frekansın son ögesiyle sınırlıdır. Müziğin temposunun belirlenmesi, en yüksek enerjiye sahip bandın seçilmesi ile gerçekleştirilir. Müzik sinyalinin içinde tempo değişebilir. Müzik yavaş ilerliyorken bir anda hızlanabilir. Bu yöntem müzik içindeki farklı ritimlerin olma olasılığını hesaba katmaz. Bu yönüyle başarılı değildir. Fakat bu tez çalışması, müzik içindeki vuruşlar arası sürelerin iyi bir şekilde hesaplandığını varsayarak iskelet tabanlı hareket tanıma ile video özetleme yöntemine odaklanır.



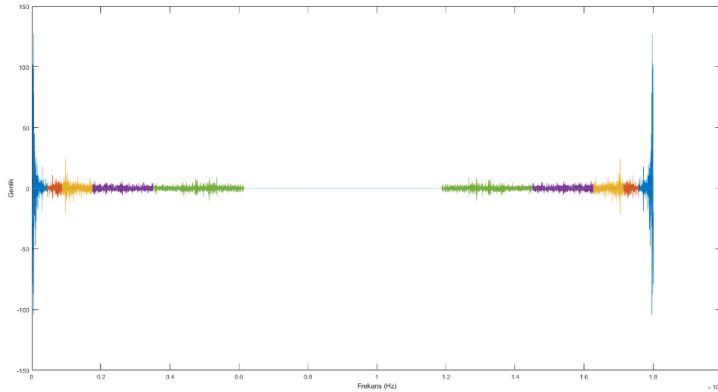
Şekil 2.25. Müzik sinyalindeki vuruşların tespit edilmesi için kullanılan algoritmanın blok diyagramı [69]

Müzik sinyalinden vuruşların algılanması adımları tek tek incelenecek olursa; ilk olarak müziğin ilk ve son 20 saniyeleri haricinde herhangi bir bölümünden bir örnek alınır. Bu örnek kısa müzik parçasının süresi ve müziğin alabileceği var sayılan en küçük dakikadaki vuruş sayısı (DVS) değerine göre Eş. 30 ile hesaplanır.

$$t_{\text{örnek}} = \frac{enk(DVS)}{60} \times 2 + 0.1 \quad (30)$$

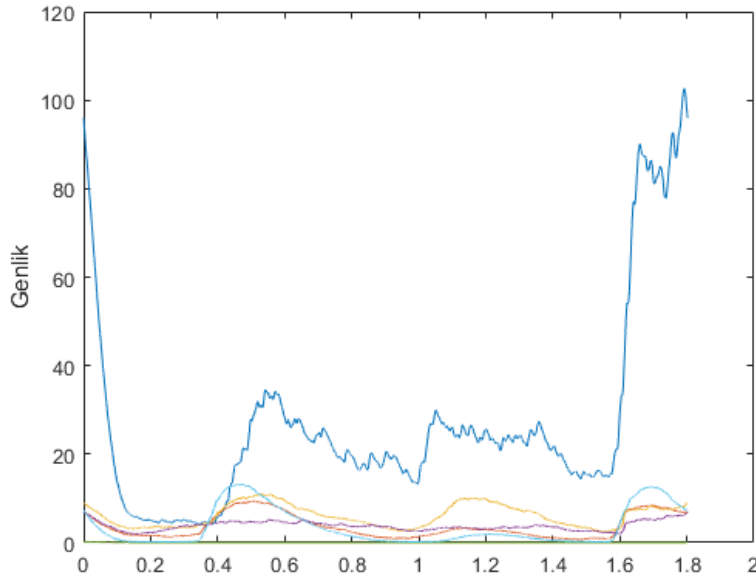
Eş.30'de $enk(DVS)$ değeri, müziğin alabileceği var sayılan dakikadaki en küçük vuruş sayısıdır. Örneğin $enk(DVS)=60$ kabul edilirse belirlenen müziğin ortasından $t_{\text{örnek}}$ saniyelik bir müzik parçası çıkarılır. Bunun üzerinde tempo analizi yapılır ve bulunan temponun uzun süreli müzik sinyali ile ilişkili olduğu kabul edilir. 2 ile çarpılmasının nedeni işlemlerin yapılabilmesi için en az 2 adet vuruş bilgisine ihtiyaç vardır. Alınan bu müzik parçasının en az 2 adet vuruş bilgisini içerebilmesi için de 0.1 saniye gibi küçük bir zaman dilimi eklenmiştir.

Tempo çıkarma işlemi genel olarak 4 adımda incelenmektedir: Birincisi, orijinal sinyalin frekans düzlemindeki hali birbirine eşit 6 ayrı sinyale bölünür. Bu farklı enstrüman gruplarından ayırma ve bunların ayrı ayrı incelenmesine imkan sağlar. Orijinal sinyalin tempo analizi farklı enstrümanların alt vuruşları ile karmaşıklığa yol açabileceği için bu yol izlenir. 6 ayrı sinyale ayırma işlemi öncesinde açılımını yaz (FFT) alınır ve belirlenen frekans bantlarına atamalar gerçekleştirilir. Scheirer'in yöntemi kullanılarak örnekleme frekansına ve sinyal uzunluğuna bağlı olarak sinyal 0-200Hz, 200-400Hz, 400-800Hz, 800-1600Hz, 1600-3200Hz ve 3200Hz-Fs bantlarına bölünmüştür [69]. Fs, örnekleme frekansını ifade etmektedir. Örnek bir müzik sinyali için Şekil 2.26'daki gibi grafik elde edilir. Böylece 6 farklı sinyalin ters Fourier dönüşümü yapılarak zaman düzlemine geçilir.



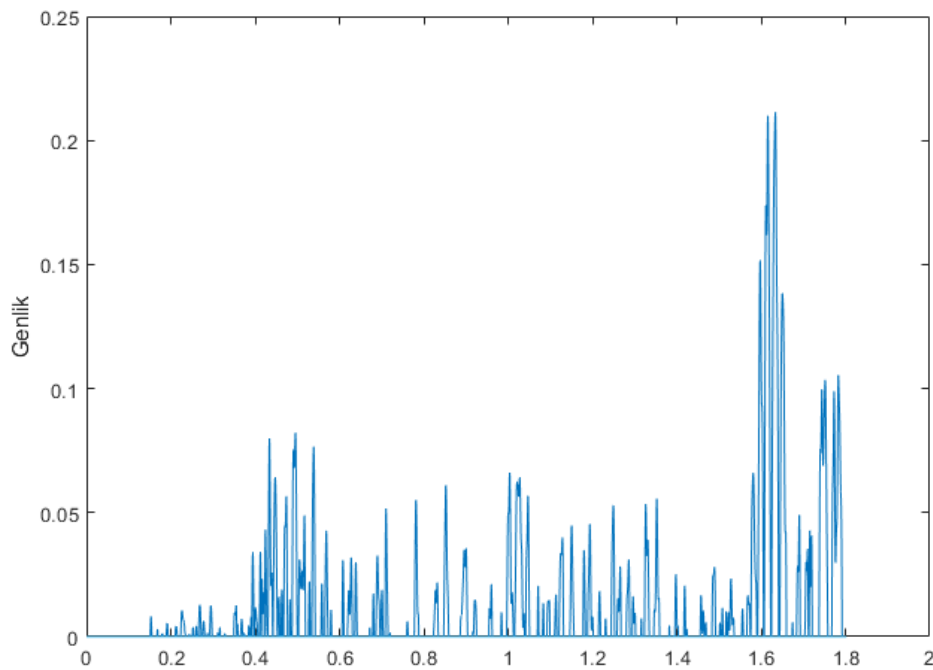
Şekil 2.26. Farklı frekans bantlarına bölünmüş örnek bir müzik sinyali

İkinci adımda, sinyalin ritmi arandığı için sesteki ani değişimlerin incelenebileceği bir forma indirgenmesi gerekir. Bu işlem, sinyalin taşıdığı frekansların hesaba katılmaması için yapılır. 6 farklı frekans bandındaki sinyallerin her biri için sinyalin içindeki küçük değişimler düşük geçişli filtre kullanılarak filtrelenir. Örnek bir müzik sinyali ile elde edilen sonuç Şekil 2.27'de gösterilmiştir. Yatay eksen zaman eksenidir.



Şekil 2.27. Örnek müzik sinyalinin FFT'si ile Yarım Hanning Penceresinin FFT'sinin çarpılması sonucu elde edilen sinyal

Üçüncü adımda, genel olarak fark düzeltme işlemi yapılır. Sinyallerin zarfı alındığı için genliği değiştiğinde tespit etmek de daha kolay duruma gelir. Vuruşlar, müzikteki periyodik bir vurgu olduğundan en büyük değişiklikler vuruşların olduğu anlamını taşır. 6 farklı sinyal yarım dalga ile doğrultularak sinyallerdeki sadece artışların görülmesi sağlanır. Ritmin analiz edilmesi için engeller kaldırılmıştır. Bu adımın çıktısı ise Şekil 2.28'de gösterilmektedir. Yatay eksen zaman eksenidir.



Şekil 2.28. Farkı alınan ve yarım dalga doğrultucu ile doğrultulan örnek bir müzik sinyali

Dördüncü adımda ise, tarak filtresi ile hangi sinyalin en yüksek enerjiyi verdiğini belirlemek amaçlanır. Belirlenen tempoda periyodik olarak meydana gelen bir darbe dizisidir. Tarak filtresi ve sinyalin temposu eşleşirse daha yüksek enerjili bir çıktı meydana gelecektir. Frekans bandındaki her sinyal, tarak filtresinin Fourier dönüşümü yapılmış hali ile çarpılarak enerjileri analiz edilmiştir. Bu çıktının en büyük değeri seçilen müziğin periyodik temposu olarak belirlenmektedir.

2.8. Video Özetleme

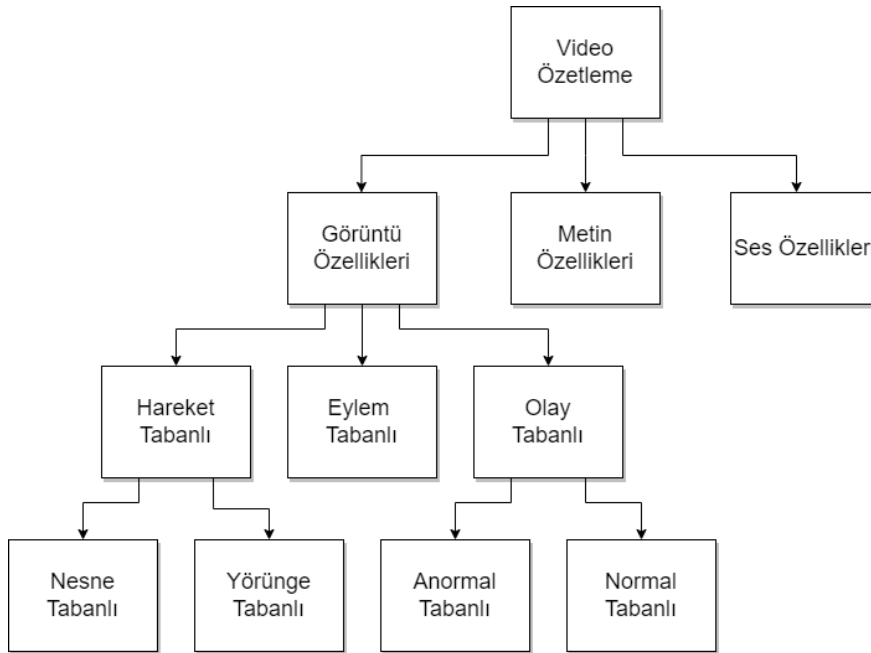
Video paylaşım kanallarının artması ile dünya genelinde kaydedilen video verilerinin miktarı uzun süredir artış göstermektedir [70]. Sosyal medya platformlarının büyümesi ile bu platformlara yüklenen video miktarları gün geçtikçe hızlı bir şekilde artmaktadır. YouTube'un icra kurulu başkanı Susan Wojcicki, 2015 yılında VidCon konferansında sadece YouTube'a her dakika 400 saatten fazla video yüklendiğini bildirmiştir [71]. Bu da, çevrimiçi yayın platformlarında video verilerinin büyük miktarda olduğunu ve sürekli arttığını gösterir. Kaydedilen videoların artması; video içi arama, akıllı depolama, video düzenleme gibi teknolojik ihtiyaçların artmasına neden olmuştur. Bu teknolojik ihtiyaçların geliştirilebilmesi için video verilerinin etkili bir biçimde anlaşılması ve özetlenmesi gerekmektedir. Aynı zamanda video paylaşım platformlarında, videolar yayınlanmadan önce bir düzenleme adımından geçerek yayına hazır hale getirilir. Videoların düzenlenmesi adımının gerçekleştirilebilmesi için videolarda düzenleme yapan kişi tarafından video içeriklerinin anlaşılması gerekir. Video özetlerinin, zaman sırası korunarak birleştirilen bir dizi görüntüden oluşan uzun videoların kısa hali olduğunu belirten özetleme tanımları, video düzenleme üzerine çalışmayan kişilerin yaptığı eksik bir tanımdır. Özet videoların içerisinde zaman sıralarının korunması gerekli değildir. Benzer içerikli video parçaları da kronolojik sıra aranmaksızın ardışık dizilerek birleştirilebilir. Bu çalışmaya göre video özetleme; video içeriklerinin en bilgilendirici, en kaliteli, en ilgili çekici sahnelerini çeşitli algoritmalarla seçerek parçalara ayıran ve bu parçaları bir uyum içerisinde tekrar birleştiren yöntem olarak tanımlanmaktadır.

Videoların düzenlenmesi işlemi için görüntü öznitelik bilgileri, görüntüdeki yazının öznitelik bilgileri ve videodaki seslerin öznitelik bilgilerini anlamlandırmaya ihtiyaç vardır. Literatürde bu bilgileri kullanan çeşitli yaklaşımlar bulunmaktadır. Bazı yaklaşımlar bu bilgilerin kombinasyonlarını da kullanabilmektedir [72]. Şekil 2.29'de gösterilen video

özetleme yaklaşımlarını daha da geliştirilerek büyük resmin incelenebilmesi bu çalışmada Şekil 2.30’da verilen video özetleme yaklaşımı uygulanmıştır.



Şekil 2.29. İçerik tabanlı video özetleme yaklaşımlarının alt dalları [69]



Şekil 2.30. Video özetleme yaklaşımlarının genel yapısı

Görüntü özellikleri, videoda bulunan görüntülerdeki nesnelerin renk, şekil ve hareketindeki değişiklikleri hakkında bilgi verir. Görüntünün özellikleri ile videodaki değişimler ya da görüntünün parlaklığı/solukluğu gibi bilgiler edinilebilir. Ekin ve arkadaşları [73], bir futbol maçındaki değerli sahneleri kaydeden kameraların futbol sahasına mesafesine göre özetleme tekniği geliştirmiştir. Görüntüler hakkındaki ön bilgiler, video içindeki nesnelerin ve olayların tanımlanmasında etkili olabilmektedir. Örneğin; bir haber yayınında sunucu kişinin tespit edilmesi ve haber yazılarının tespit edilebilmesi olay örgüsü hakkında bilgi edinmeyi

ve buna göre en değerli görüntü karelerinin belirlenmesine katkı sağlayabilmektedir. Görüntü bilgileriyle video özeti çıkarmayı amaçlayan yöntemler genellikle belirli ve özel bir konudaki videoları özetleme üzerine gerçekleştirilmiştir [74, 75].

Videoda bulunan ses özellikleri kullanılarak spor içerikli videolarda önemli kolaylıklar sağlanmıştır. Spor yorumcularının konuşmaları, topa vuruş sesleri ya da seyirci seslerindeki heyecanlar video sahnelerinin ne kadar değerli olduğu ve içeriklerin anlamları hakkında bilgiler vermektedir. Video içindeki konuşmalar analiz edilerek ve yazıya dönüştürülerek de önemli bilgiler edinilmektedir [72].

Görüntü üzerine yerleştirilen metinler çeşitli yazı algılama algoritmaları ile tespit edilebilmektedir. Bununla birlikte yazıların yazılış şekillerine göre alt yazı olup olmadığı ya da haber yazısı olup olmadığı gibi farklı tespitler de yapılmaktadır. Sözcüklerin sıralı dizilimi kullanılarak her farklı konu için bir özet çıkarılır [73]. Edinilen bu bilgiler doğrultusunda bu üç farklı bilgiyi bir arada kullanarak birleştirmek daha doğru sonuçlar verecektir. Ancak bu çalışma, üç farklı bilgi arasında özetleme için en zorlayıcı problem olan videoların görüntü bilgisine göre anlamlandırılarak özetlenmesi ile ilgilenmektedir.

Görüntü özelliklerine dayalı video özetlemenin genel amacı, önemli video alt dizilerini belirlemek ve seçilen orijinal videonun en değerli görüntüleriyle temsil edilen daha kısa bir versiyonunu oluşturmaktır. Videonun görüntü bilgisine göre çeşitli özetleme yaklaşımları sahne tabanlı ve içerik tabanlı olmak üzere iki ana başlıkta incelenmektedir. Sahne tabanlı yaklaşımlar da kendi içinde statik ve dinamik olmak üzere iki alt başlıkta incelenir. İçerik tabanlı yaklaşımlar ise önce hareket, eylem ve olay tabanlı olmak üzere üç alt başlıkta incelenirken hareket tabanlı yaklaşımlar nesne ve yörünge tabanlı, olay tabanlı yaklaşımlar ise anomali ve normal tabanlı olmak üzere iki farklı alt başlıkta incelenmektedir. İçerik tabanlı yaklaşımlar Şekil 2.30'da alt başlıkları ile birlikte gösterilmiştir.

Sahne tabanlı yöntemlerin alt başlıklarından statik tabanlı yöntemler, video içindeki ana karelerin seçilmesi ve bunların birleştirilmesi yaklaşımı ile gerçekleştirilir. Dinamik tabanlı yöntemler ise ana kare yerine kısa video kliplerden oluşur. Sahne tabanlı yöntemler, genellikle değişen sahnelerin olduğu videolarda kullanılır. Hareketli kameraların kullanıldığı videolara elverişlidir. Önceden elde edilen görüntü öznetelik bilgileri kullanılarak kümeleme yöntemleri ile ana kare seçim işlemi yapılabilmektedir. Nihayetinde anlamsız görüntülerin kaldırılması için çeşitli yöntemler uygulanabilmektedir [74, 75].

İçerik tabanlı yöntemler, video görüntülerinden anlam çıkararak özetleme işlemini yapar. Bu yöntemlerde özetleme adımından önce bir ön işleme yapmak gereklidir. İçerik tabanlı yöntemleri, sahne tabanlı yöntemlerin içinde kullanmak da mümkündür. Video özetleme problemindeki en önemli zorluk, özetlenecek videonun anlamlandırılmasına, içeriğindeki video parçalarının sınıflandırılmasına ve video içindeki nesnelerin türlerine hatta bu nesnelerin eylemlerinin türlerine doğrudan bağlı olmasıdır. Sahnelerin karmaşıklığı genel olarak bilinen yöntemlerin kullanılmasında çeşitli zorluklar meydana getirir [76-78].

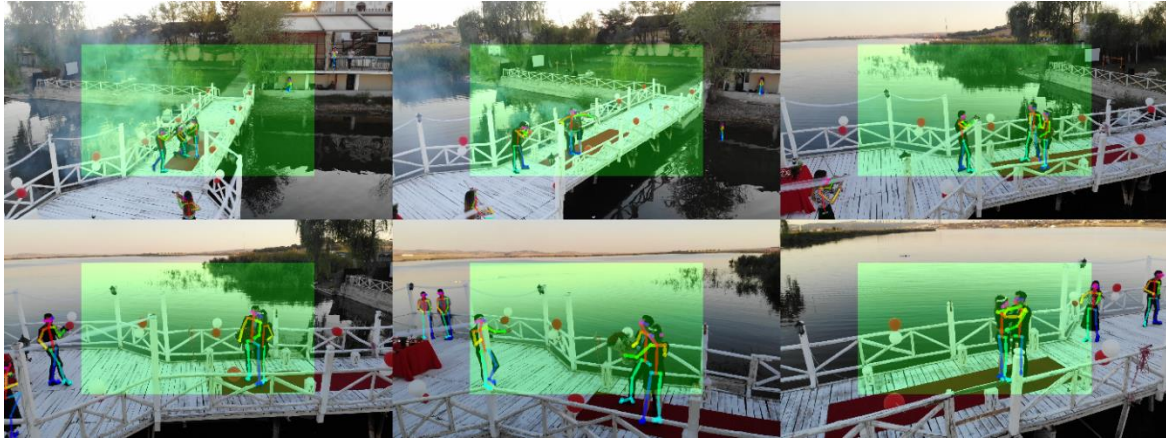
Bu çalışmada, insan içerikli videoların özetlenmesi için iskelet tabanlı hareket tanıma yöntemlerine dayalı olarak arka fon müziği destekli yeni bir yöntem geliştirilmiştir. İlk olarak görüntülerdeki kişilerin poz tahmini algoritması kullanılarak eklem noktaları belirlenir. Ardından bu eklem noktalarının poz takibi öncesinde iyileştirme adımlarından geçirilmesi sağlanır. Böylece video içindeki kişilerin video boyunca eklem noktalarındaki değişimler analiz edilebilir. Hareket tanıma aşamasında zaman dizilerinde başarılı sonuçlar elde eden UKSB katmanları kullanılarak derin öğrenme ağı tasarlanmıştır. Bu çalışma için kaydedilen ve yayınlanan *PoseDrone* veri kümesi üzerinde etiketlenen veriler kullanılarak hareket tanıma yapılmıştır. Video boyunca kişiler arası ilişkileri incelemek için hareketleri hakkında çeşitli hesaplamalar yapılmıştır. Bu hesaplamalar sonucunda Bellman denklemi (Eş. 31) ile video içeriğinde keşif yapabilen bir algoritma tasarlanmıştır. Bunun sonucunda en değerli video anları belirlenebilmektedir. Bu değerli video anlarının başlangıç ve bitiş zamanlarının belirlenebilmesi gerekmektedir. Arka planda çalması planlanan müziğin dakikadaki vuruş sayısı sinyal işleme yöntemleri kullanarak bulunmuş ve video parçalama işlemi gerçekleştirilir.

2.8.1. Video özetleme için ön işlem adımları

Bu bölüm hareket tanıma verilerinin çıktılarına ek olarak videolardan elde edilen tüm bilgilerin derlenmesi için gerçekleştirilmiştir. Buradaki veriler kullanılarak pekiştirmeli öğrenmenin temeli olan Bellman denklemi ile videonun keşfedilmesi ve kurguya göre en değerli sahnelerin belirlenebilmesi için bu tez çalışması kapsamında geliştirilen ön işleme adımları uygulanmıştır. Videonun özetlenebilmesi için kurguya göre en değerli görüntü parçalarının keşfedilmesi gereklidir. Bunun için kişilerin başrol olma olasılığı, görüntü merkez bölgesinde bulunma durumu, görüntünün merkezine uzaklığı, hedef hareket sınıfı, tespit edilen hareket sınıfına güven değeri, uzuvların uzunlukları, eklem noktalarına göre açılar, eklem noktalarının tespitinin güven değeri, muhtemel başrol oyuncularının birbirine

yakınlıkları, başrol oyuncularının birbirinin önünü kapatıp kapatmaması ve videodaki kişilerin video boyunca kaç görüntüde bulunduğu verileri kullanılarak görüntünün ne kadar değerli olduğu Bellman denklemi ile hesaplanmaktadır. Bu ön işleme adımları, pekiştirmeli öğrenmedeki çevrenin tasarlanması adımıyla benzetilebilir. Ön işleme adımları ile videolardaki görüntülerin ne kadar değerli olduğu hakkında detaylı hesaplamalar yapılmış ve hangi bilgilerin sonucu ne kadar etkileyeceği kaydedilmiştir. Bu doğrultuda ödül ve ceza puanları dikkate alınarak en değerli görüntü kareleri belirlenmektedir. Yapılan çalışmaların sonuçları Bölüm 3’te verilmektedir.

Video özetleme için gerçekleştirilen ön işleme adımlarından birincisi, Şekil 2.31’de gösterilmektedir. Görüntünün içindeki yeşil bölgede videodaki önemli olayların gerçekleştiği ve başrol oyuncularının genellikle bu yeşil bölgede bulunduğu daha kıymetli sahnelerin gerçekleştiği öne sürülmektedir. Böylece en değerli görüntü kareleri hesaplanırken bu bölgede bulunan kişilerin gerçekleştirdiği hareketlerin olasılıklarının etkisi daha büyük olacaktır. İnsansız hava araçları ile kaydedilen görüntülerde başrol oyuncuları genellikle bu bölgenin dışında bulunmamaktadır. Bulunuyorsa o görüntü video özetlemeye dahil edilecek kadar değerli değildir.



Şekil 2.31. Görüntünün merkez bölgesi ve başrol oyuncularının ilişkisi

İkinci ön işleme adımı video boyunca kişilerin takip edilme oranının hesaplanmasıdır. Örneğin bir video 100 görüntüden oluşuyorsa ve bir kişi 78 görüntü boyunca kimlik numarasını koruyabildiyse bu kişinin takip edilme oranı %78’dir. Başrol oyuncuları çok sayıda ardışık görüntü boyunca takip edilirken başrol olmayan oyuncular kadrajdan çıkıp tekrar girebilir. Böylece uzun süre takip edilememektedir. Başrol oyuncularında da poz takibinden kaynaklı olarak kimlik numarası değişimleri meydana gelebilmektedir.

Geliştirilen ön işleme algoritması buna dayanıklıdır. Bir kişinin bir görüntü anında kimlik numarası değiştiyse bu kişinin sınırlayıcı kutusu ile bir önceki görüntüdeki tüm kişilerin sınırlayıcı kutularının kesişim bölgeleri arasındaki değerleri tek tek hesaplar. En yüksek değere sahip olan kişinin bilgileri de özetleme için kullanılacak videonun tüm bilgilerini içeren bir dizide tutulur.

Ön işleme adımlarından üçüncüsü ise, muhtemel başrol oyuncularının birbirlerine olan uzaklıklarıdır. Etkinlik videolarında başrol oyuncularının birlikte hareket ettiği sahnelerin en değerli sahneler olduğu kabul edilir. Bu doğrultuda, görüntüdeki her bir kişinin diğer tüm kişiler arasındaki mesafeleri hesaplanır. Bu adım ikinci adımdaki en çok takip edilen kişilerle birbirine en yakın iki kişinin kesişim kümesi alınır. Nihai olarak hesaplanan tüm değerler 0-1 aralığına normalleştirilir. Böylece Bellman denklemini kullanarak çalışan ile video içinde keşif yapılarak en değerli video parçalarının belirlenmesi gerçekleştirilebilecektir.

2.8.2. Bellman Denklemi ile Videonun Keşfedilmesi

Pekiştirmeli öğrenmenin temelinde yatan Bellman denklemi ile video keşfedilerek ve hareket tanıma verilerinin işlenmiş çıktıları kullanılarak en değerli görüntülerin keşfedilmesi amaçlanmaktadır. İlk olarak Bellman denklemi Eş. 31’de incelenmektedir.

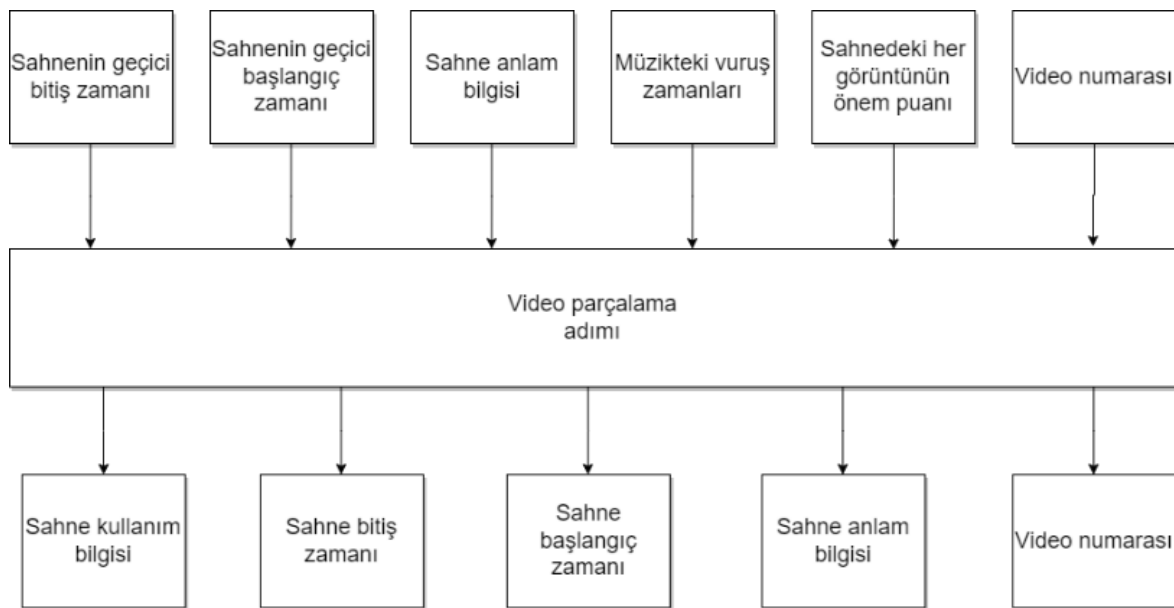
$$V(s) = \max_a (R(a, s) + \delta * V(s')) \quad (31)$$

Burada $V(s)$, çalışanın mevcut bulunduğu durumu gösterir. Video boyunca ödülü en fazla yapabileceği görüntüler arasında keşif yaparken her görüntü için ne kadar değerli olduğu hakkında hesap yapmaktadır. a , ileri ve geri giderek görüntüler arasında gezinebileceği için +1 ve -1 olarak tanımlanmıştır. s ise çalışanın mevcut durumunu gösterir. İndirim oranı δ ile gösterilmektedir. 0 ile 1 aralığında değer alabilmektedir. Sonraki görüntü adımlarındaki ödüllerin ihmal edilmemesi için bu değer mümkün olduğunca 1’e yakın seçilmektedir. Bu nedenle bu çalışmada $\delta = 0.99$ olarak alınmıştır. Bellman denklemi ile video için elde edilen sonuçlar Bölüm 3’te verilmiştir.

2.8.3. Videoların Parçalanması

Çalışmada videoların en önemli sahnelerinin kurguya göre birleştirilebilmesi amacıyla videolar parçalara ayrılmıştır. Videoların parçalanabilmesi için gerekli olan video numarası, değerli sahnelerin başlangıç ile bitiş zamanları, sahne anlam bilgisi ve sahnedeki her

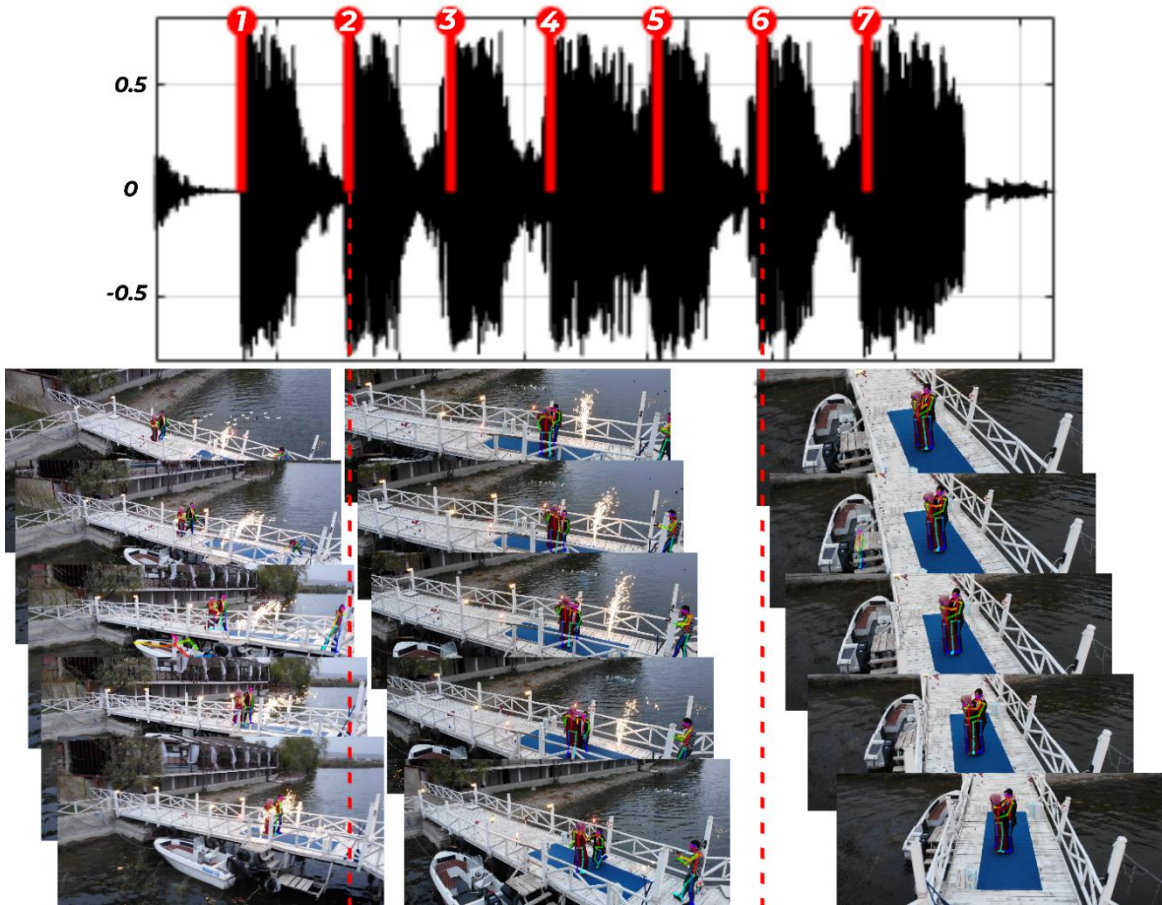
görüntünün değer puanı bir dizide tutulmalıdır. Her bir video numarası için çalışanın video keşfi sonucunda birbirinden farklı ve önemli sahneler belirlenmişti. Başrol hareketlerine bağlı olarak sahne anlam bilgisinin değiştiği an, geçici sahne bitiş zamanı olarak kabul edilmektedir. Aynı zamanda bu diğer sahne için geçici sahne başlangıç zamanıdır. Belirlenen müzik dosyasının vuruşlarının zamanları Bölüm 2.7’de tespit edilerek bir dizide tutulmaktadır. Her vuruşta bir sahne geçişi olmak zorunda değildir. En değerli görüntünün %70’inden az değerli görüntüler olduğu an sahne geçiş kararı verilir. Bu eşik değer deneysel olarak elde edilmiştir. Geçici sahne başlangıç zamanı ve geçici sahne bitiş zamanı, bu sahne içindeki en değerli görüntü çevresine ve görüntü önem puanına bağlı olarak müzik vuruş zaman aralıklarına göre sahne başlangıç ve bitiş zamanları hesaplanır. Sahne anlam bilgisi, kurgu sırasına göre kullanılacak bilgiyi içermektedir. Müzikteki iki ya da dört vuruşluk zaman aralığında video sahneleri parçalanmadan analiz edilmektedir. Üç farklı sahne kararı için de eşik değer kontrolleri yapılır ve kriterler sağlanıyorsa dört vuruşluk süre için sahne parçalanır, eğer kriterlere uymayan görüntüler varsa iki vuruşluk süre için uygunluğu kontrol edilir. Kriterler sağlanmıyorsa tek vuruş süresi için sahne parçalanabilmektedir. Buna bağlı olarak Şekil 2.32’deki bilgiler kullanılarak video parçalama adımı gerçekleştirilir. Parçalama adım sonunda her sahnenin ait olduğu video numarası, sahne anlam bilgisi, sahne başlangıç zamanı, sahne bitiş zamanı ve kullanılma bilgisi üretilir. Parçalanmış tüm videoların kullanılma bilgisi 0’dır. Video birleştirme adımında kullanılan sahneler 1 değeri atanarak tekrar kullanılmamaları sağlanır.



Şekil 2.32. Video parçalama adımının girdi ve çıktıları

2.8.4. Videoların Birleştirilmesi

Parçalanmış videolar müzik ritmine ve kurguya göre belirlenen sıraya uyumlu olacak şekilde birleştirilir. Belirlenen kurguya göre videoların sıralanarak birleştirme işleminin gerçekleştirilmesi önerilmektedir. Örneğin *PoseDrone* veri kümesindeki evlilik teklifi konulu videolar için kurgunun; ilk olarak çiftlerin el ele yürümesi, ardından evlilik teklif anı, ardından sarılma ile dans sahneleri ve tekrar yürüyüş sahnesi ile video klibin sonlandırılması şeklinde olması planlanmaktadır. Tüm bu sahneler müzik ritmi ile zamansal olarak uyumlu olacak şekilde sahneler arası geçişler sağlanarak birleştirilir. Şekil 2.33'te örnek bir müzik sinyalinin 7 tane vuruşu gösterilmiştir. 2. ve 6. vuruşlarda sahneler arası geçiş gösterilmiştir. Müzik sinyalindeki 2. Vuruş anına kadar el ele yürüyüş hareketi gösterimdeyken, 6. vuruşa kadar sahne değişikliği olmadan diz çökme sahnesi video özetine dahil edilir ve 6. Vuruştaki sahne değişikliği ile dans sahnesine geçiş yapıldığının gösterimi mevcuttur.



Şekil 2.33. Örnek bir müzik sinyalindeki vuruş zamanlarına göre farklı sahneler arası geçişlerin gerçekleştirilme örneği

2.9. PoseDrone Veri Kümesi

Tekil kamera kullanarak kaydedilen videolar derinlik bilgisini kaydeden kameralardan daha yaygın kullanıldığı için ve derinlik gibi bir bilgiye sahip olmadan hareketlerin analizinin yapılması daha zor bir problemdir. Sabit kameralarla kaydedilen videolardan oluşan eylem tanıma etkinlik videoları basit bir problemin çözümü ile ilgilenmektedir. Literatürde rastgele ve hızlı hareketler yapan insansız hava aracı ile kaydedilen eylem tanıma veri kümesi bulunmamaktadır. Bu çalışmada insansız hava aracı ile yapılan bu kayıtlar evlilik teklifi etkinliklerinden oluşmaktadır. Video paylaşım platformlarının ve sosyal medyanın kullanımının artması ile evlilik teklifi gibi etkinliklerden video klip oluşturma ihtiyacı artmıştır. Bu veri kümesinin aynı zamanda video özetleme alanında da kullanılabilmesi amaçlanmaktadır. Mevcut çalışmalar sadece eylem tanıma ya da sadece video özetleme için ayrı ayrı veri kümeleri oluşturmak üzerinedir. Literatürde ilk kez bir veri kümesi hem eylem tanıma hem de video özetleme için kullanılmak üzere hazırlanmıştır. Bu çalışmada, 494 adet video örneği ve 396 037 adet görüntüden oluşan insan eylemi tanıma ve video özetleme için özel bir veri kümesi sunulmaktadır.

Veri kümesinde her bir kişiye ait eklem noktalarının tespit edilmesi için OpenPose kullanılmıştır. OpenPose ile tespit edilen eklem noktaları derlenerek etiketleme işlemi yapılmıştır. Bu eklem noktalarına dayalı olarak 33 083 adet kişi için hareket sınıflarına etiketlemeler yapılmıştır. Eylem sınıflandırması için özyinelemeli sinir ağları yaygın olarak kullanılmaktadır ve en başarılı sonuçlar zaman dizilerinin kullanılması ile gerçekleşmektedir. Bu nedenle veri kümesi, her bir kişinin hareketini zaman dizileri haline getirilmiş halini de sunmaktadır. Deneysel sonuçlar, verilerin bir ön işleme adımından geçirilerek eklemler arası açılara ve uzuvların Öklid uzunluklarına göre filtreleme yapılmasının faydalı olduğu göstermektedir.

Veri kümesi kaydedilirken Şekil 2.34'deki ekipmanlardan yararlanılmıştır. İnsansız hava araçları olarak; DJI Mavic Air, DJI Mini 2 ve DJI Mavic Air 2 kullanılmıştır. Yerden çekimlerde ise Nikon D90 ve DJI Osmo Pocket kameraları ile videolar kaydedilmiştir. Bu kameraların avantajı, titreşimi engelleyici mekanizmalara sahip olmalarıdır. Bu sayede görüntüde anlık kaymalar engellenerek anlık titreşimlerden yoksun kaliteli görüntüler elde edilmiştir. Profesyonel tanıtım ve klip çekimlerinde de bu tip kameralar kullanılmaktadır. Bazı etkinlik videoları aynı zaman anını farklı açılardan kaydeden sabit kameralarla da

desteklenmektedir. Böylece hem hareket tanıma konusunda hem de video özetleme alanında farklı çalışmaların da önünü açabilmesi amaçlanmaktadır.



Şekil 2.34. *PoseDrone* veri kümesini kaydederken kullanılan ekipmanlar

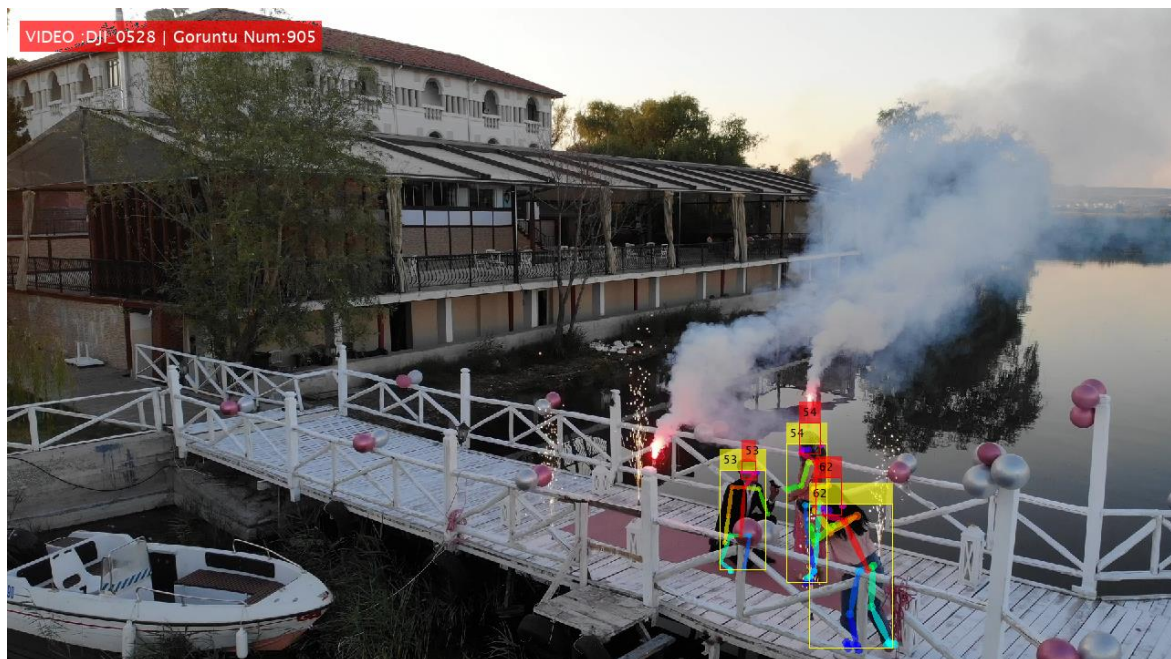
Her bir görüntünün çözünürlüğü 1920 x 1080 olarak kaydedilmiştir. Videolar RGB olarak kaydedilmiştir. Eylem sınıfları; evlilik teklif etmek, ayakta sabit durmak, yürümek, dans etmek, sarılmak ve el ele tutuşmak üzere 6 ana sınıftan oluşmaktadır. Her bir kişi için 25 eklem noktasının piksel konumları ve poz tahmininin güven değeri bulunmaktadır. Her kişinin her bir videoda benzersiz bir kimlik numarası bulunmaktadır. Kişilerin fiziksel boyutları genel olarak benzerdir. Videolar, 20 ile 35 yaş aralığındaki kişilerden oluşmaktadır. Bazı hareketlerde insansız hava aracı 360 derece dönüş sağlayarak kişilerin hareketlerini mümkün olduğunca çeşitlendirmektedir. Böylece karşılaşılabilecek senaryoları artırarak veri kümesi veri çeşitliliği açısından zengindir. Kameraların yükseklikleri anlık olarak değişebilmektedir. İnsansız hava araçları bölgede izin verilen azami 120 metreye kadar çıkabilmektedir. Kaydedilen videolarda 50 metrenin yukarısına çıkmadan veri kaydı yapılmıştır. Görüntülerin kaydedildiği insansız hava araçlarının yükseklikleri ağırlıklı olarak 3-15 metre aralığında değişmektedir.

Eş zamanlı olarak kaydedilen videolarda görüntü üzerinden zaman senkronizasyonu çalışmaları yapılacağı için hareket tanıma ve anormal hareket tespiti üzerine de çalışmaların geliştirilebilmesine fırsat tanınmaktadır. Şekil 2.35, Şekil 2.36 ve Şekil 2.37’de verilen görüntüler aynı zaman anına ait 3 farklı açıdan farklı kameralarla kaydedilmiş görüntülere aittir. Yer kamerası ile kaydedilen Şekil 2.35’teki görüntünün sağ üst köşesine doğru bakıldığında insansız hava aracı küçük birkaç siyah piksel olarak görülmektedir. Şekil 2.36’da ise insansız hava aracı ile kaydedilen görüntü bulunurken Şekil 2.37’de aynı sahnenin yer kamerası ile çekilen görüntüsü bulunmaktadır.

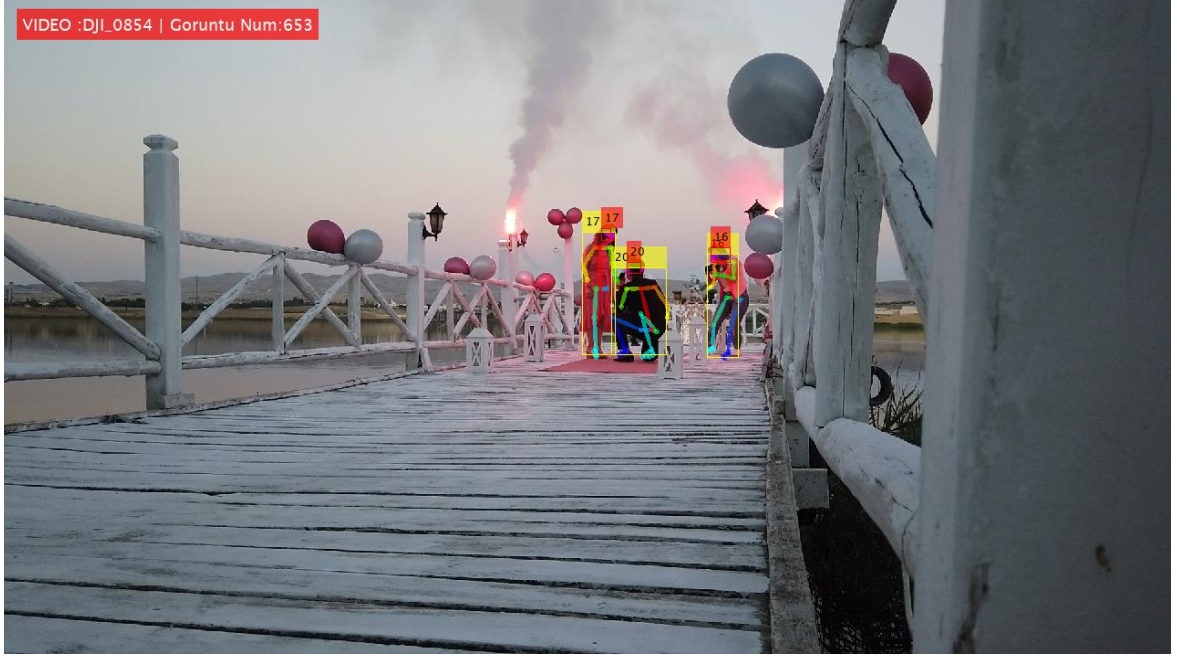
Veri kümesi ile ilgili en güncel çalışmalara çalışmanın resmi internet sayfası olan posedrone.com üzerinden ulaşılabilecektir. Şu anda veri kümesi hakkında verilen bilgiler ve değerlendirme metrikleri internet sayfasında paylaşılarak güncellenebilecektir.



Şekil 2.35. Etkinlik videosunda eş zamanlı kaydedilen yerdeki sol kamera görüntüsü



Şekil 2.36. Etkinlik videosunda eş zamanlı kaydedilen insansız hava aracının kamera görüntüsü



Şekil 2.37. Etkinlik videosunda eş zamanlı kaydedilen yerdeki sağ kamera görüntüsü

3. DENEYSEL SONUÇLAR

3.1. Poz tahmini ve takibi ile ilgili çalışmalar

Poz takibi ile ilgili yapılan çalışmaların başarısını ölçebilmek için gerçek hayat problemlerini yansıtan videoların olduğu bir veri kümesine ihtiyaç vardır. PoseTrack [15] veri kümesi bu ihtiyacı karşılamaktadır ve en güncel çalışmaların sonuçları herkese açık şekilde yayınlanmaktadır. Poz tahmini aşamasında yaygın olarak kullanılan ve gerçek zamanlı çalışabilen OpenPose ile son verilere göre poz tahmini için PoseTrack veri kümesi üzerinde en yüksek başarıyı elde eden DCPose yöntemleri kullanılarak PoseTrack veri kümesi üzerinde çok kişili poz takibi konusunda deneysel çalışmalar yapılmıştır.

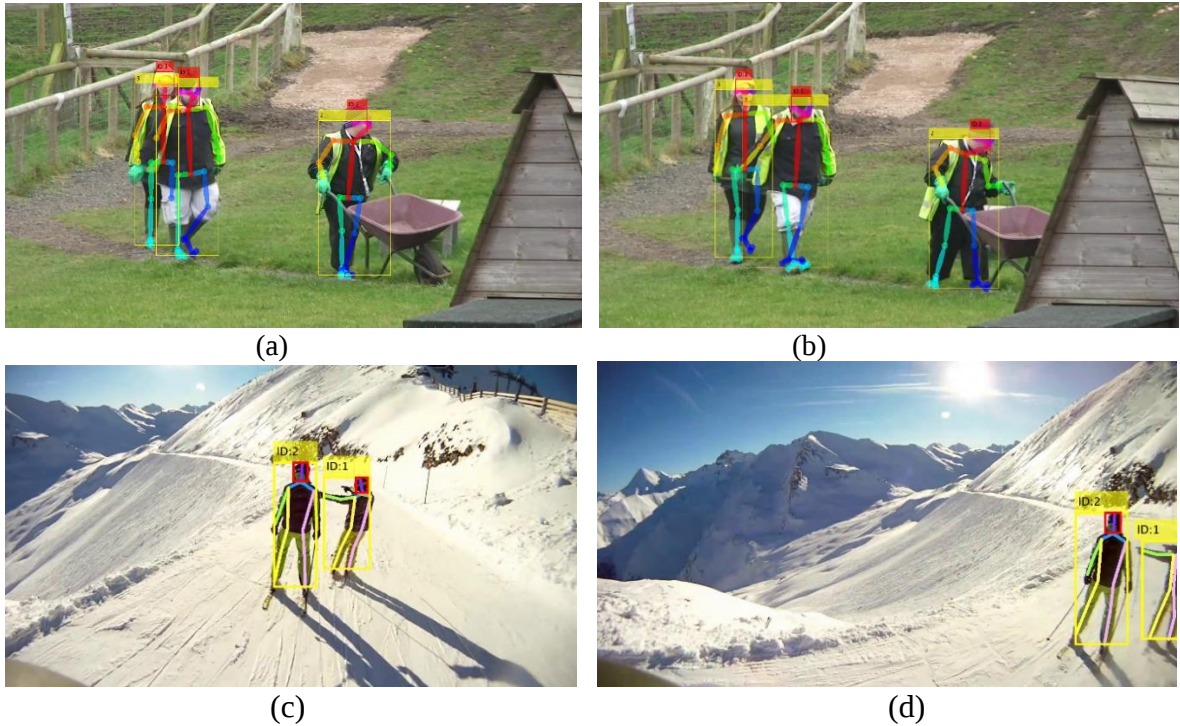
Çizelge 3.1. Ön işleme adımlarının poz takibi performansına etkisi

M/N Kriteri	Algoritma	MOTA Baş	MOTA Omuz	MOTA Dirsek	MOTA El Bilek	MOTA Bel	MOTA Diz	MOTA Ayak Bilek	MOTA Toplam	MOTP Toplam	Prec. Toplam	Rec. Toplam
2/3	Openpose	-14.6	33.3	13.9	11.3	21.1	16.8	13.0	11.6	56.3	58.9	69.6
	Openpose ön işlemli	33.5	46.2	32.6	27.6	38.3	32.6	28.9	34.2	62.5	75.5	52.8
	DCPose	46.7	49.0	21.3	7.7	36.8	19.3	-4.9	26.6	67.6	63.5	79.5
	DCPose ön işlemli	55.5	56.3	42.6	39.0	52.7	49.0	35.9	47.8	67.2	77.9	72.1
2/5	Openpose	-15.3	32.4	13.2	10.7	20.3	15.9	12.3	10.9	56.2	58.6	69.7
	Openpose ön işlemli	33.4	46.0	32.6	27.6	38.1	32.5	28.7	34.1	62.6	75.7	52.1
	DCPose	46.6	49.0	21.2	7.6	36.7	19.2	-5.0	26.5	67.7	63.4	79.6
	DCPose ön işlemli	55.5	56.3	42.6	39.0	52.7	49.0	35.9	47.8	67.2	77.9	72.1

Çizelge 3.1 PoseTrack 2017 veri kümesi kullanılarak OpenPose ve DCPose poz tahmin algoritmalarından elde edilen verilere poz takibi işlemi öncesinde uygulanan ön işleme adımlarının etkisini farklı M/N değerleri için göstermektedir. Çizelgeden de görüleceği gibi izlerin onaylanması için kullanılan M/N kriterinin sonuçlar üzerinde büyük bir etkisi olmamıştır. Fakat toplam MOTA değeri olarak 2/3 M/N kriteri ile daha iyi sonuçlar alınmıştır. Ön işleme adımları uygulanarak toplam çok kişili poz takibi MOTA değerinde

OpenPose için %22.6, DCPose için %21.2 performans artışı elde edilmiştir. DCPose yönteminin genel olarak OpenPose yöntemine göre daha başarılı olduğu görülmektedir. Fakat DCPose poz tahmini el/ayak bilekleri, dirsek ve diz noktalarında başarısız sonuçlar vermektedir. Görüntüde bu eklem noktaları olmasa bile DCPose bu eklem noktaları için bir değer tahmin etme eğilimindedir. Bu nedenle net olarak belirlenemeyen bilek noktaları için beşinci ön işleme adımı ile tahminde iyileştirme yapılmıştır. Çizelge 3.1’de bulunan MOTA Baş değerleri incelendiğinde ön işleme adımının bu bölgenin poz takibi performansını DCPose için %8.8, OpenPose için %48.1 arttırdığı görülmektedir. Benzer şekilde diğer eklem noktalarındaki iyileşmeler de toplam MOTA değerindeki artış ile birlikte incelendiğinde birbirini kanıtlar niteliktedir.

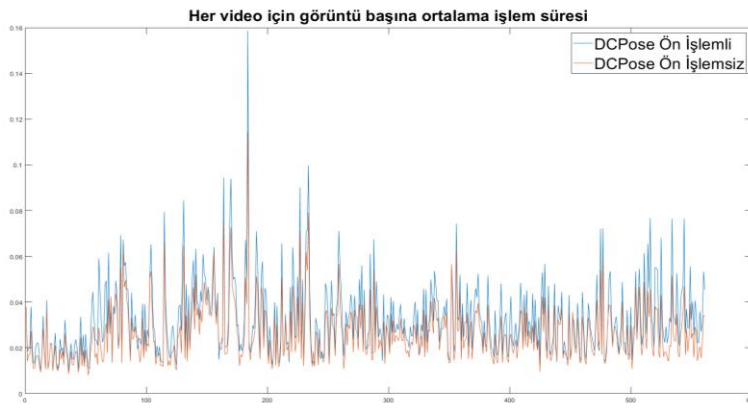
Şekil 3.1 peş peşe video karelerinde kişilerin takip edildiğini gösteren bir örnektir. Kişilere birer numara verilmiş ve takip eden karelerde bu numaralar korunarak kişilerin takibi yapılmıştır.



Şekil 3.1. Örnek birer videoda OpenPose ile videonun (a) 22. ve (b) 40. karelerinde, DCPose ile videonun (c) 53. ve (d) 77. karelerinde bulunan kişilerin takip sonuçları

Şekil 3.2’de ön işleme adımlarının DCPose için işlem süresine etkisi görülmektedir. Mavi değerler ön işleme adımının ortalama sürelerini gösterir. Her video için poz takibi süresi

hesaplanmıştır. Bu süre videodaki görüntü sayısına bölünerek bir görüntü karesi için ortalama takip işlem süresi bulunmuştur. Şekil 3.2’de ön işleme adımının poz takibinde ne kadar işlem gecikmesi getirdiği işlem süreleriyle birlikte verilmiştir. Tüm görüntülerin poz takibi için birim ortalama işlem süreleri hesaplandığında ön işleme adımı olmadan 26.1 milisaniye, ön işleme adımıyla birlikte 32.4 milisaniye gerekmektedir. Ön işlemler çalışma süresini 5.7 milisaniye yavaşlatmaktadır. Performansta %21.2 artış sağladığı için çalışma süresinden 5.7 milisaniyelik ödün vererek ön işlem adımlarının kullanılabileceği düşünülmektedir.



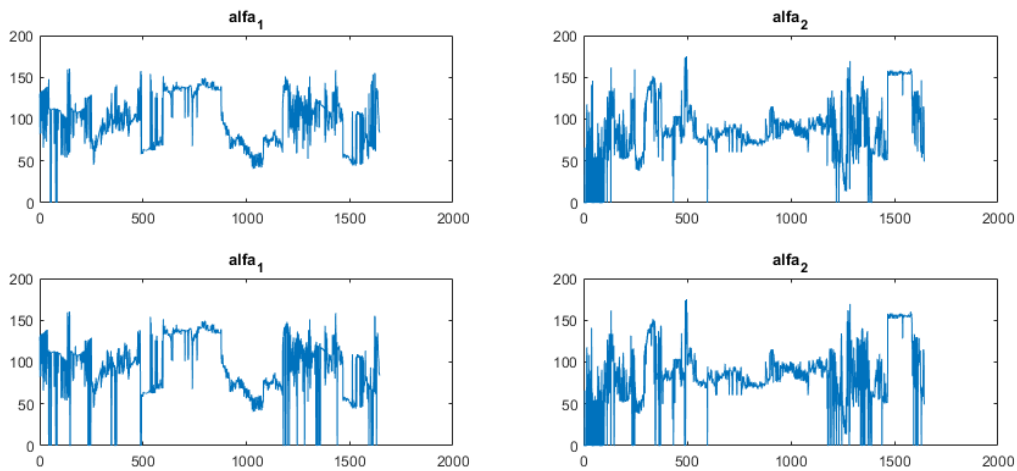
Şekil 3.2. Ön işleme adımlarının her bir videonun poz takip süresinin videodaki görüntü sayısına bölünmesi ile elde edilen ortalama işlem sürelerine etkisi

Poz tahmini ve poz takibi çalışmalarından sonra hareket tanıma probleminin çözülebilmesi için çeşitli deneysel çalışmalar yapılmıştır. Hareket tanıma için eklem noktaları tespit edildikten sonra derin öğrenme ağını eğitmeden önce verilerin ön işleme adımları gerçekleştirilmiştir.

3.2. Hareket tanıma ile ilgili çalışmalar

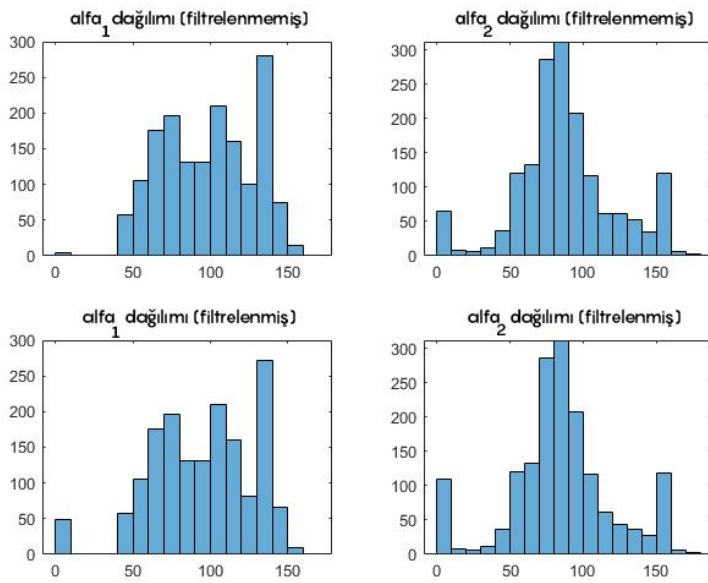
Hareketli kameralarla kaydedilen videolarda insan hareketlerinin tanınması için iskelet tabanlı yaklaşımların daha iyi performans gösterdiği en güncel çalışmalarla gösterilmiştir [67]. Bu nedenle bu çalışmada iskelet tabanlı insan hareketlerinin tanınması üzerine çeşitli deneysel çalışmalar gerçekleştirilmiştir. Bölüm 2.6’da da bahsedildiği gibi derin öğrenme modelinde eğitim yapılmadan önce verilerin kalitesinin artırılması üzerine çalışmalar yapılmıştır. Çünkü poz tahmin algoritmaları eklem noktaları hakkında detaylı tespit yapabilme konusunda henüz yeterince kararlı sonuçlar üretmemektedir. Özellikle insansız

hava aracı ile kaydedilen ve gerçek hayat problemlerini doğrudan yansıtan videoların ardışık görüntüleri arasında poz tahminlerine bağlı olarak eklem noktalarında çeşitli bozulmalar tespit edilmektedir. Bu bozulmalar eklemler arası açı değerlerini ve Öklid mesafelerini doğrudan etkilemektedir. Dolayısıyla takip eden adımda bu verilerle yapılacak eğitim başarısızlıkla sonuçlanacaktır. Bunun önüne geçebilmek için bir hareket sınıfına ait eklem noktalarının tüm açı değerlerinin dağılımı incelenmiştir. Şekil 2.16’da gösterilen 7 tane açı için her bir alt hareket sınıfı da göz önünde bulundurularak yeni analizler yapılmıştır. Açı ve mesafe değerlerinin dağılımı incelenerek poz tahmininden kaynaklı olarak tespit edilen hatalı eklem noktası değerlerini temizlemek amacıyla çeşitli filtreler eklenmiştir. Örneğin; diz çökme hareketi için etiketlenen veriler incelendiğinde sağ ve sol dizdeki açılar Şekil 2.16’da gösterilen α_1 ve α_2 açı değerleridir. Bu değerler, yapılan gözlemlere göre diz çökme hareketi icra edildiyse α_1 ve α_2 açı değerleri aynı anda 135 dereceden büyük olmamalıdır. Dar açıya sahip her değer doğru olabileceği için filtrelenmez. 90 ve 135 derece arasında kalan değerler ortalama değerden standart sapmanın 2 katından büyükse bu değerler de filtrelenir. Açı değerlerinin dağılımının Gauss olduğu kabul edilmektedir. Standart sapmanın 2 katı değeri deneysel olarak yapılan incelemeler sonucunda belirlenmiştir. Benzer olarak diğer hareket sınıfları için de açı ve uzunluk filtrelemeleri yapılmıştır. Açılardan α_1 ve α_2 için Şekil 3.4’te filtrelenmiş ve filtrelenmemiş grafikler görülmektedir. Dikey eksen 0-180 derece arasında değişen açı değerlerini, yatay eksen incelenen veri sayısını göstermektedir. Şekil 3.4’te üstteki grafikler filtrelenmemiş, alttaki grafikler ise filtrelenmiş grafikleri göstermektedir. Filtrelenmiş verilerde sıfır değerlerinin arttığı görülmektedir.



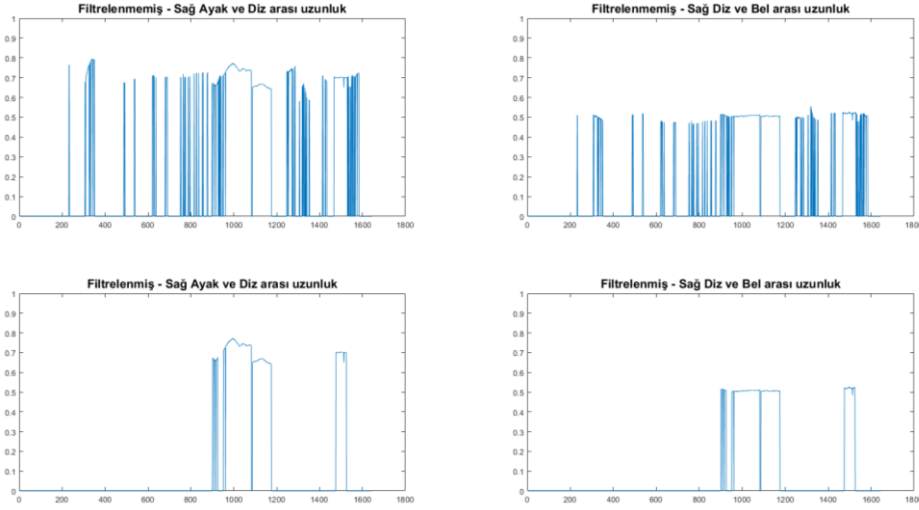
Şekil 3.3. Diz çökme hareketi için etiketlenmiş tüm kişilerin sağ ve sol dizinde meydana gelen açı değerlerinin filtrelenmiş ve filtrelenmemiş değerleri

Yukarıda bahsedilen iki açı değerinin dağılımları da filtrelenmiş ve filtrelenmemiş olarak Şekil 3.4'te incelenmektedir. Filtreleme işleminden önce ortalama değer 98.92 iken filtrelendikten sonra 95.3'e düşmüştür. Böylece poz tahmininden kaynaklı bazı değerlerde düzeltme yapılmıştır. Şekil 3.4'te üst bölümdeki iki grafik filtrelenmemiş α_1 ve α_2 verilerini, alttakiler ise filtrelenmiş α_1 ve α_2 açı dağılımlarını gösterir. Yatay eksen 0-180 derece arasında değişen açı değerlerinden oluşmaktadır. Dikey eksen kaç adet olduğunu göstermektedir. Sıfır değerindeki artış filtreleme işleminin etkisini göstermektedir. Eğitimde sıfır değerine sahip olan veriler kullanılmamaktadır.



Şekil 3.4. Sağ ve sol dizdeki açı değerlerinin filtrelenmemiş ve filtrelenmiş dağılımı

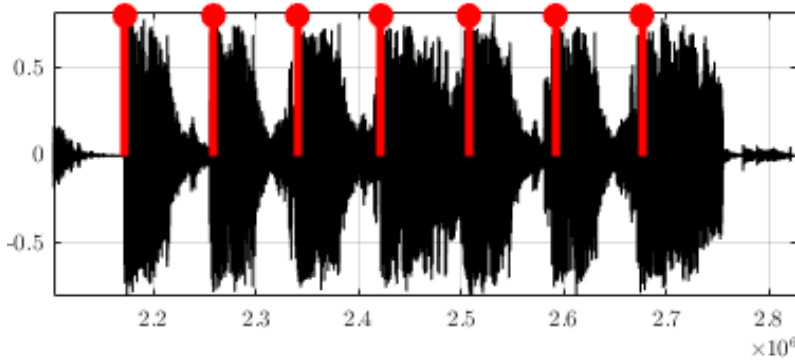
Uzuvların uzunlukları için ise Şekil 3.5'te filtrelenmiş ve filtrelenmemiş grafikler görülmektedir. Yatay eksen veri sayısını dikey eksen ise uzuv uzunluğunun kişiyi sınırlayan dikdörtgenin bir köşegen uzunluğuna oranını göstermektedir.



Şekil 3.5. Sağ ayak ve sağ diz ile sağ diz ve sağ bel arasındaki uzunlukların filtrenmemiş ve filtrenmiş gösterimleri

3.3. Müzik sinyalinde vuruş tespit çalışmaları

Video özetleme için bir müzik sinyalinin belirlenmesi gerekmektedir. Örnek bir müzik sinyali olarak “Oshins featuring Hael - Darkside” belirlenmiştir. Bu müziğin 47. ve 64. saniyelerinin arasındaki müzik sinyalinin vuruşlarına ait tespitler Şekil 3.6’da kırmızı ile gösterilmektedir. Yatay eksen sinyalin zamana bağlı değerlerini gösterirken dikey eksen sinyalin genliğini göstermektedir.



Şekil 3.6. Bir müzik sinyalindeki vuruş zamanlarının tespitinin gösterimi

Bölüm 2.7’de belirlenen bir müzik sinyalinin vuruşlarının tespiti üzerine gerçekleştirilen çalışmanın detaylarından bahsedilmiştir. Çalışmanın asıl amacı müzik sinyallerinin tespiti olmamakla birlikte, fon müziğiyle birlikte video özetleme olarak yeni bir yaklaşım ortaya konulmuştur.

3.4. Video özetleme ile ilgili çalışmalar

Video özetleme için ön işlem adımlarının sonunda elde edilen bilgilerin bir bölümü Çizelge 3.2’de paylaşılmaktadır. 2088 görüntüden oluşan bir videonun diz çökme hareketinin gerçekleştirildiği bir sahnenin 1603. ile 1611. görüntüleri arasındaki özetleme ön verileri incelenmektedir. Başrol tespitleri görüntüdeki tüm kişilerin ikili kombinasyonlarıyla yapılan işlemler sonucunda elde edilmiştir. Çizelge 3.2’de 62 ve 88 kimlik numaralı kişiler başrol oyuncularındır. Tespit edilen başrol oyuncularının birbirlerine göre her görüntüdeki piksel uzaklıkları hesaplanmaktadır. Başrol oyuncularının sahip olduğu sınırlayıcı kutuların birbirleri ile kesişim alanları arasındaki ilişki incelenmiştir. Sınırlayıcı Kutu Kesişim-1 birinci başrol oyuncusunun ikinci başrol oyuncusunun sınırlayıcı kutusu ile kesişimini gösterirken Sınırlayıcı Kutu Kesişim-2 değeri de tam tersini göstermektedir. Merkez Nokta Mesafesi değeri ise görüntünün merkez noktasından birinci başrol oyuncusunun uzaklığını göstermektedir. Takip Yüzdesi değeri, video boyunca takip edilen her kişinin en çok takip edilen kişinin oranına bölünmesi ile elde edilmektedir. Merkez bölge kesişim değeri, Bölüm 2.8.1’deki görüntünün önemli bölgesi ile birinci başrol oyuncusunun kesişim yüzdesini göstermektedir. Bu değerler Bellman denklemindeki bir zaman anındaki durumun ve ödül puanının belirlenmesini sağlamaktadır.

Çizelge 3.2. Her bir video için görüntü önem değerlerinin hesaplanması için oluşturulan tablodan 8 görüntülük bir örnek

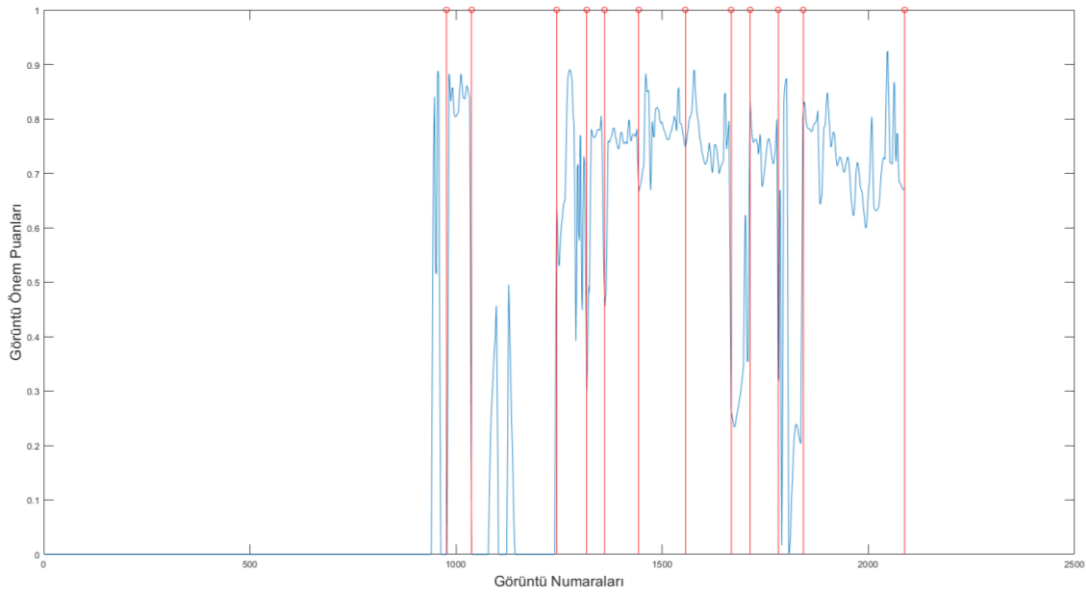
Görüntü Numarası	Birinci Başrol Kimlik	İkinci Başrol Kimlik	Başrol Mesafe (Piksel)	Sınırlayıcı Kutu Kesişim-1	Sınırlayıcı Kutu Kesişim-2	Takip Yüzdesi	Merkez Bölge Kesişim	Merkez Nokta Mesafesi (Piksel)	Hedef Hareket Sınıfı
1603	62	88	60.19	0	0	93.4	100	83.73	3
1604	62	88	58.86	0	0	93.4	100	84.68	3
1605	62	88	59.16	0	0	93.4	100	86.32	3
1606	62	88	58.92	0	0	93.4	100	87.97	3
1607	62	88	58.04	75.1	20.7	93.4	100	89.56	3
1608	62	88	57.33	53.7	19.4	93.4	100	91.50	3
1609	62	88	57.02	0.1	0.1	93.4	100	93.01	3
1610	62	88	57.09	93.3	28.3	93.4	100	94.42	3

Şekil 3.7.’de ise Çizelge 3.2’deki değerlerin elde edildiği görüntüler paylaşılmaktadır. Sol üst köşede 1603 numaralı görüntü olmak üzere sırayla soldan sağa artan şekilde gösterilmektedir.



Şekil 3.7. Çizelge 3.2’de değerleri paylaşılan bir video sahnesinin görüntüleri

Video sahnelerinin önem puanlarına göre sahnelerin parçalanması Bölüm 2.8.3’te detaylandırılmıştır. Video değerlerinin önem puanlarının sonuçları Şekil 3.8’de verilmektedir. Kırmızı ile görülen çizgiler, video parçalama adımına göre belirlenen sahneler arası geçişleri göstermektedir.



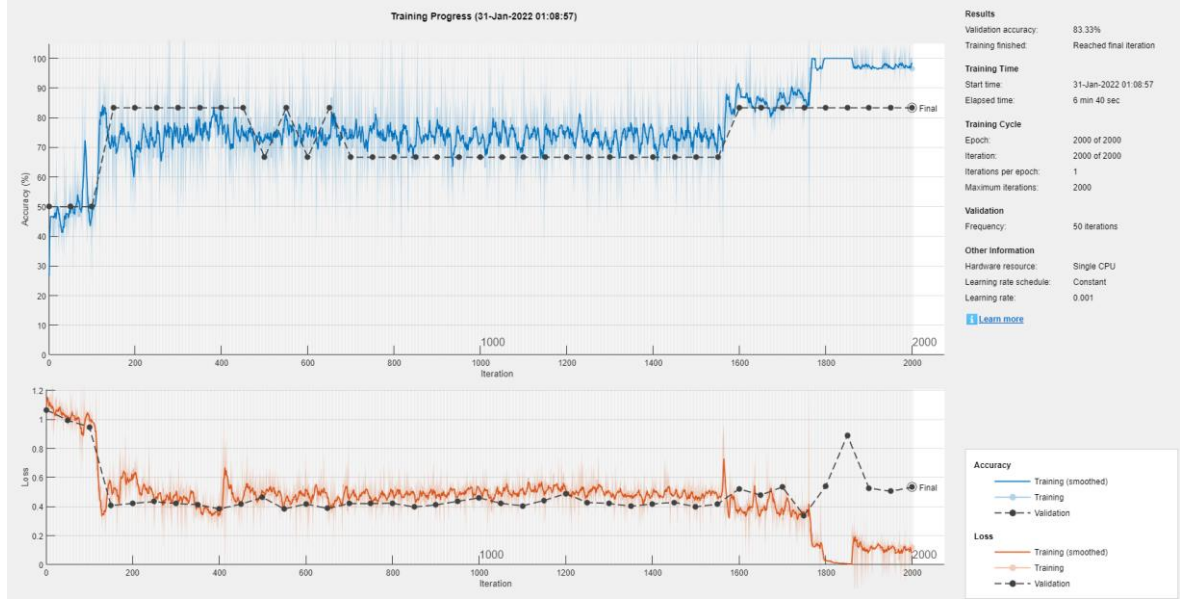
Şekil 3.8. Bir etkinlikteki 2088 adet görüntüden oluşan en uzun videonun görüntülerinin önem puanları ve en değerli görüntüler

Şekil 3.8’de kırmızı her iki kırmızı çizginin arasında kalan bölümler birleştirilmek üzere kullanılmaya hazır aday sahnelerden oluşmaktadır. Bu bölümlerin içindeki her görüntü kullanılmak zorunda değildir. Bölümün içindeki en değerli görüntünün ardışık olarak öncesinde ve sonrasında gelen görüntüler, müzik vuruşları arasında geçen süreye ve saniyedeki görüntü sayısı bilgisine göre kısaltılabilir ya da uzatılabilir. Şekil 3.9 bu bölümlerin arasında kalan en değerli görüntülerden en değerli dört örneği içermektedir.



Şekil 3.9. Bir etkinlikteki bazı sahnelerin en değerli örnek görüntüleri

Şekil 3.10’da derin öğrenme modelinin doğruluk ve kayıp grafikleri gösterilmektedir. UKSB katmanında kullanılan ara katman sayısı 70’dir. Verilerin eğitimi, MATLAB 2021A versiyonu kullanılarak gerçekleştirilmiştir.



Şekil 3.10. Derin öğrenme modelinin eğitim sürecindeki doğrulama ve kayıp grafikleri

Literatürde UKSB ile gerçekleştirilen eylem tanıma öğrenme katsayısı gibi parametreler ayarlanarak gerçekleştirilen eğitimler bu veri kümesi için iyi sonuçlar vermediği için deneysel olarak bazı parametrelerin en iyi değerleri bulunmuştur. Giriş ağırlıklarının öğrenme oran faktörü 0.23 alınırken giriş ağırlıklarının L2 faktörü de 0.46 alınmıştır. Kayıp fonksiyonu olarak çapraz entropi kullanılmıştır. Doğrulama frekansı 50 iterasyon olarak belirlenirken 2000 iterasyon yapılmıştır ve öğrenme oranı 0.001 alınmıştır. Eğitimde posedrone veri kümesinden %14.6’sı doğrulama için, %14.36’sı test için %70.04’ü ise eğitim için kullanılmıştır.

4. SONUÇ VE ÖNERİLER

Bu çalışmada, etkinliklerde kaydedilen videoların içerik bilgisine ve belirlenen arka plan müziğine göre özetlenmesi amaçlanmıştır. Video içeriklerinin anlamlandırılabilmesi için meydana geldiği görüntülerde poz tahmini algoritması kullanılarak görüntülerdeki kişilerin eklem noktaları tespit edilir. Ardından bu eklem noktaları kullanılarak video boyunca kişilerin takibi yapılır. Meydana getirilen veri kümesinde etiketlenen verilerle kişilerin video içinde zaman dizisi olarak hangi hareketleri yaptığı bilgisine sahip olunmaktadır. Aynı zamanda yine poz takibi sayesinde başrol oyuncularının tespiti ile görüntülerin önem puanlarının hesaplanabilmesi amacıyla çeşitli olasılık değerleri hesaplanmaktadır. Bu olasılık değerleri ödül puanlarına dönüştürülmektedir. Ödül puanları kullanılarak Bellman denklemi ile videodaki görüntülerin önem puanları hesaplanmaktadır. Tüm etkinlik boyunca kaydedilen videolar içinde önem puanlarına ve sahne bilgisine göre sıralama yapılmaktadır. Farklı sahnelere ait en değerli görüntüler kurgu bilgisine göre sıralanmaktadır. Görüntülerin özet içinde kullanılmak üzere parçalanmış videolarda kullanım bilgisi tutulmaktadır. Belirlenen müzik vuruşları arasında geçen süre ile oluşturulacak video özetinde saniyedeki görüntü sayısı çarpılarak her iki vuruş arasında kaç görüntü kullanılması gerektiği hesaplanmaktadır. Tüm bu veriler doğrultusunda videolar birleştirilmektedir. Saatlerce kaydedilen etkinlik videoları belirlenen bir müziğin süresi kadar kısaltılarak özetlenmektedir.

Bu çalışmada çok kişili gerçek zamanda çalışabilen bir poz takibi yöntemi kullanılarak farklı poz tahmini yöntemlerinin başarımını arttıracak ön işlemler geliştirilmiştir. Çalışma sonuçları 550 farklı videodan oluşan PoseTrack 2017 veri kümesi ve bu veri kümesinin herkese açık olarak paylaştığı sonuç hesaplama yöntemi ile gerçek değeri bilinen 250 video için değerlendirilmiştir. OpenPose ve DCPose yöntemleri kullanılarak bu veri kümesi üzerinde kişilerin eklem noktalarının tespitleri elde edilmiştir. Bu eklem noktaları Kalman filtresi ve KEYK yöntemi ile geliştirilen poz takibi modülünde kullanılmıştır. Poz takibi öncesinde yapılan ön işlemlerin sonuçlara etkisi incelenmiştir ve işlem süresi analizi yapılmıştır. Bu analizlerin sonuçlarına göre DCPose ile yapılan poz takibi sonuçları PoseTrack veri kümesine göre daha başarılıdır. Ön işleme adımları çok kişili poz takibi toplam MOTA sonuçlarını OpenPose poz tahmini yöntemi için %22.6 artırırken DCPose için %21.2 arttırmıştır. Poz takibinden önce yapılan işlemlerin her bir videoyu işlem süresi bakımından ne kadar etkilediği hesaplanmıştır. Elde edilen bu değerlerle ortalama bir video

için ön işlemler ve ön işlemsiz poz takibi işlem süreleri karşılaştırıldığında tüm videolarda görüntü başına ortalama 5.7 milisaniye ilave işlem süresi gözlemlenmiştir. Ön işleme adımlarına ek olarak farklı poz takibi sonuçlarını birleştirerek ileri işleme adımları ile daha başarılı sonuçlar da elde edilebileceğini ön görülmektedir. Gelecek çalışmalarda paralel poz takipçilerini birleştirerek yüksek doğruluğa sahip düşük kaynak tüketimi olan bir poz takibi yöntemi geliştirilmesi hedeflenmektedir.

Tez kapsamında hızlı yer değiştirebilen insansız hava aracı üzerindeki kamera ve yer kameraları ile kaydedilen ve 396037 adet görüntüden oluşan PoseDrone veri kümesi oluşturulmuştur. Bu görüntüler üzerinde 7 farklı sınıf için 33 bin adetten fazla veri etiketleme yapılmıştır. Hareket tanıma için çeşitli filtreleme adımları ile poz tahminindeki hatalı tespitler giderilerek derin öğrenme ağı %83.89 doğrulukla eğitilmiştir.

Farklı sahneler, başrol oyuncularının farklı hareketleri gerçekleştirilmesi olarak nitelendirilir. Sahneler arası geçişler için müzikteki her iki vuruş arasındaki geçen süreye göre saniyedeki görüntü sayısı bilgisi kullanılarak sahnede kullanılacak görüntü sayısı belirlenmiştir. Sahne içindeki önem puanı en yüksek görüntülerin öncesinde ve sonrasında gelen ardışık görüntüler kullanılmıştır. Video özetleme işlemi yeni bir veri kümesi için başarılı olarak gerçekleştirilmiştir. PoseDrone veri kümesi üzerinde kişilerin eklem noktalarının tespit edilmesi için OpenPose yöntemi kullanılmıştır. Her bir sınıf için veri adedini artırarak ve doğruluk oranı yüksek poz tahmini yöntemleri ile hareket tanıma sonuçlarının daha başarılı olabileceği ön görülmektedir. Gelecek çalışmalarda, bu veri kümesi için değerlendirme kriterleri meydana getirilerek dünya genelinde bir poz takibi yarışma başlatılması planlanmaktadır.

Bu tez çalışması kapsamında pek çok yeni yaklaşım önerilmektedir. Bunlardan birincisi poz takibi öncesinde geliştirilen ön işleme adımlarından baş bölgesinin bulunmasıdır. Geliştirilen ön işleme adımlarının poz takibine olumlu etkileri bulunmaktadır. İkincisi, hareket tanıma konusunda alt hareket sınıflarından hedef hareket sınıflarına geçiş yöntemidir. Alt hareket sınıfları ile hangi eklem noktasında nasıl değişimin olduğu tespit edilir ve etiketlenmiş hedef hareket sınıfı ile ilişkilendirilir. Aynı zamanda uzunluk bilgileri ve açı bilgileri için sınıflandırma yapılarak iki farklı sonuç için birleştirme yapılmaktadır. Üçüncüsü, video özetleme alanında arka plan müziğindeki vuruş zamanlarının tespit edilerek bu zamanlara göre video sahnelerinin değiştirilmesi yaklaşımı literatürde bir ilktir. Bu

yaklaşım gerçek hayatta video düzenleme problemine getirilen kıymetli bir çözümdür. Dördüncüsü ise literatürde hem video özetleme hem de hareket tanıma için kullanılabilecek PoseDrone isimli bir veri kümesi hazırlanmıştır. Veri kümesi, insansız hava araçları üzerindeki ani hareketler sergileyebilen kameralar ile karmaşık problemlerin çözülebilmesinin önünü açmak üzere çeşitli etkinliklerde kaydedilen videolardan oluşmaktadır.

KAYNAKLAR

1. Cote, M., Jean, F., Albu, A. B., and Capson, D., (2016). *Video summarization for remote invigilation of online exams*. IEEE Winter Conference on Applications Of Computer Vision (WACV), New York, 1-9.
2. Yan, S., Xiong, Y., and Lin, D. (2018). *Spatial temporal graph convolutional networks for skeleton-based action recognition*. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 3482-3489.
3. Snower, M., Kadav, A., Lai, F., and Graf, H. P. (2020). *15 keypoints is all you need*. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Washington, 6738-6748.
4. Liu, Z., Chen, H., Feng, R., Wu, S., Ji, S., Yang, B., and Wang, X. (2021). *Deep dual consecutive network for human pose estimation*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Tennessee, 525-534.
5. Wang, M., Tighe, J., and Modolo, D. (2020). *Combining detection and tracking for human pose estimation in videos*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, 11088-11096.
6. Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). *Deep high-resolution representation learning for human pose estimation*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, California, 5693-5703.
7. Bulat, A., and Tzimiropoulos, G. (2016). *Human pose estimation via convolutional part heatmap regression*. European Conference on Computer Vision, Springer, 717-732.
8. Cao, Z., Simon, T., Wei, S. E., and Sheikh, Y. (2017). *Realtime multi-person 2d pose estimation using part affinity fields*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Hawaii, 7291-7299.
9. Kuhn, H. W. (1955). The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2), 83-97.
10. Ladicky, L., Torr, P. H., and Zisserman, A. (2013). *Human pose estimation using a joint pixel-wise and part-wise formulation*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Oregon, 3578-3585.
11. Jin, S., Liu, W., Ouyang, W., and Qian, C. (2019). *Multi-person articulated tracking with spatial and temporal embeddings*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, California, 5664-5673.
12. Raaj, Y., Idrees, H., Hidalgo, G., and Sheikh, Y. (2019). *Efficient online multi-person 2d pose tracking with recurrent spatio-temporal affinity fields*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, California, 4620-4628.

13. Girdhar, R., Gkioxari, G., Torresani, L., Paluri, M., and Tran, D. (2018). *Detect-and-track: Efficient pose estimation in videos*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Utah, 350-359.
14. Xiao, B., Wu, H., and Wei, Y. (2018). *Simple baselines for human pose estimation and tracking*. European Conference on Computer Vision (ECCV), Munich, 466-481.
15. Yang, X., and Tian, Y. (2014). *Super normal vector for activity recognition using depth sequences*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Ohio, 804-811.
16. Wang, P., Li, W., Gao, Z., Zhang, J., Tang, C., and Ogunbona, P. O. (2015). Action recognition from depth maps using deep convolutional neural networks. *IEEE Transactions on Human-Machine Systems*, 46(4), 498-509.
17. Donahue, J., Anne Hendricks, L., Guadarrama, S., Rohrbach, M., Venugopalan, S., Saenko, K., and Darrell, T. (2015). *Long-term recurrent convolutional networks for visual recognition and description*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Massachusetts, 2625-2634.
18. Zhang, Y., Liu, X., Chang, M. C., Ge, W., and Chen, T. (2012). *Spatio-temporal phrases for activity recognition*. European Conference on Computer Vision, Berlin, 707-721.
19. Du, Y., Wang, W., and Wang, L. (2015). *Hierarchical recurrent neural network for skeleton based action recognition*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Massachusetts, 1110-1118.
20. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L., and Xie, X. (2016). *Co-occurrence feature learning for skeleton based action recognition using regularized deep LSTM networks*. AAAI Conference on Artificial Intelligence, Arizona, 3697-3703.
21. Yao, A., Gall, J., Fanelli, G., and Van Gool, L. (2011). *Does human action recognition benefit from pose estimation?*. British Machine Vision Conference, Scotland.
22. Tenorth, M., Bandouch, J., and Beetz, M. (2009). *The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition*. IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops, Kyoto, 1089-1096.
23. Shahroudy, A., Liu, J., Ng, T. T., and Wang, G. (2016). *Spatio-temporal LSTM with trust rgb+ d: A large scale Spatio-temporal LSTM*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nevada, 1010-1019.
24. Liu, J., Shahroudy, A., Xu, D., and Wang, G. (2016). *Spatio-temporal lstm with trust gates for 3d human action recognition*. European Conference on Computer Vision, Amsterdam, 816-833.
25. Zhao, Z., Yang, Q., Lu, H., Weninger, T., Cai, D., He, X., and Zhuang, Y. (2017). Social-aware movie recommendation via multimodal network learning. *IEEE Transactions on Multimedia*, 20(2), 430-440.

26. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., and Zhuang, Y. (2017). *Video question answering via gradually refined attention over appearance and motion*. 25th ACM International Conference on Multimedia, California, 1645-1653.
27. Zhang, H., Wang, M., Hong, R., and Chua, T. S. (2016). *Play and rewind: Optimizing binary representations of videos by self-supervised temporal hashing*. 24th ACM International Conference on Multimedia, Amsterdam, 781-790.
28. Elharrouss, O., Almaadeed, N., Al-Maadeed, S., Bouridane, A., and Beghdadi, A. (2021). A combined multiple action recognition and summarization for surveillance video sequences. *Applied Intelligence*, 51(2), 690-712.
29. Mehmood, I., Sajjad, M., Rho, S., and Baik, S. W. (2016). Divide-and-conquer based summarization framework for extracting affective video content. *Neurocomputing*, 174, 393-403.
30. Sujatha, C., Chivate, A. R., Tabib, R. A., and Mudenagudi, U. (2014). *Multilevel framework for summarization of surveillance videos*. Fifth International Conference on Signal and Image Processing, Bangalore, 265-270.
31. Elharrouss, O., Al-Maadeed, N., and Al-Maadeed, S. (2019). *Video summarization based on motion detection for surveillance systems*. 15th International Wireless Communications and Mobile Computing Conference (IWCMC), Morocco, 366-371.
32. Zhang, S., Zhu, Y., and Roy-Chowdhury, A. K. (2016). Context-aware surveillance video summarization. *IEEE Transactions on Image Processing*, 25(11), 5469-5478.
33. Chen, F., De Vleeschouwer, C., and Cavallaro, A. (2013). Resource allocation for personalized video summarization. *IEEE Transactions on Multimedia*, 16(2), 455-469.
34. Bagheri, S., Zheng, J. Y., and Sinha, S. (2016). Temporal mapping of surveillance video for indexing and summarization. *Computer Vision and Image Understanding*, 144, 237-257.
35. Song, X., Sun, L., Lei, J., Tao, D., Yuan, G., and Song, M. (2016). Event-based large scale surveillance video summarization. *Neurocomputing*, 187, 66-74.
36. Xu, X., Hospedales, T. M., and Gong, S. (2016). Discovery of shared semantic spaces for multiscene video query and summarization. *IEEE Transactions on Circuits and Systems for Video Technology*, 27(6), 1353-1367.
37. Kalaivani, P., and Roomi, S. M. M. (2017). *Towards comprehensive understanding of event detection and video summarization approaches*. Second International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM), Tindivanam, 61-66.
38. Yun, S., Yun, K., Kim, S. W., Yoo, Y., and Jeong, J. (2014). *Visual surveillance briefing system: Event-based video retrieval and summarization*. 2014 11th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Seoul, 204-209.

39. Lin, W., Zhang, Y., Lu, J., Zhou, B., Wang, J., and Zhou, Y. (2015). Summarizing surveillance videos with local-patch-learning-based abnormality detection, blob sequence optimization, and type-based synopsis. *Neurocomputing*, 155, 84-98.
40. Tejero-de-Pablos, A., Nakashima, Y., Sato, T., and Yokoya, N. (2016). *Human action recognition-based video summarization for RGB-D personal sports video*. IEEE International Conference on Multimedia and Expo (ICME), Washington, 1-6.
41. Tejero-de-Pablos, A., Nakashima, Y., Sato, T., Yokoya, N., Linna, M., and Rahtu, E. (2018). Summarization of user-generated sports video by using deep action recognition features. *IEEE Transactions on Multimedia*, 20(8)
42. Truong, B. T., and Venkatesh, S. (2007). Video abstraction: A systematic review and classification. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 3(1)
43. Choi, J., Jeon, W. J., and Lee, S. C. (2008). *Spatio-temporal pyramid matching for sports videos*. 1st ACM International Conference On Multimedia Information Retrieval, British Columbia, 291-297.
44. İnternet: Goodfellow, I., Bengio, Y., and Courville, A. (2016). Deep learning. Book in preparation for MIT Press, <http://www.webcitation.org/query?url=https://mitpress.mit.edu/books/deep-learning> Son Erişim Tarihi: 03.02.2021.
45. İnternet: Skansi, S. (2018). Introduction to Deep Learning: From Logical Calculus to Artificial Intelligence. Springer, <http://www.webcitation.org/query?url=https://link.springer.com/book/10.1007/978-3-319-73004-2> Son Erişim Tarihi: 14.09.2021.
46. Uz Kent, B., Yeh, C., and Ermon, S. (2020). *Efficient object detection in large images using deep reinforcement learning*. IEEE/CVF Winter Conference on Applications of Computer Vision, Colorado, 1824-1833.
47. Crawford, E., and Pineau, J. (2019). *Spatially invariant unsupervised object detection with convolutional neural networks*. AAAI Conference on Artificial Intelligence, Hawaii, 3412-3420.
48. Nawi, N. M., Atomi, W. H., and Rehman, M. Z. (2013). The effect of data pre-processing on optimized training of artificial neural networks. *Procedia Technology*, 11, 32-39.
49. İnternet: Christopher, Bishop. (2006). Pattern recognition and Machine Learning, Springer, <http://www.webcitation.org/query?url=http://users.isr.ist.utl.pt/~wurmd/Livros/school/Bishop%20-%20Pattern%20Recognition%20And%20Machine%20Learning%20-%20Springer%20%202006.pdf> Son Erişim Tarihi: 10.01.2021
50. Hochreiter, S., and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, 9(8), 1735-1780.

51. Andriluka, M., Iqbal, U., Insafutdinov, E., Pishchulin, L., Milan, A., Gall, J., and Schiele, B. (2018). *Posetrack: A benchmark for human pose estimation and tracking*. IEEE Conference on Computer Vision and Pattern Recognition, Utah, 5167-5176.
52. Vondrick, C., Patterson, D., and Ramanan, D. (2013). Efficiently scaling up crowdsourced video annotation. *International Journal of Computer Vision*, 101(1), 184-204.
53. Bernardin, K., and Stiefelhausen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 1-10.
54. Hasegawa, I., Uenishi, K., Fukunaga, T., Kimura, R., and Osawa, M. (2009). Stature estimation formulae from radiographically determined limb bone length in a modern Japanese population. *Legal Medicine*, 11(6), 260-266.
55. Moon, S., Park, Y., Ko, D. W., and Suh, I. H. (2016). Multiple kinect sensor fusion for human skeleton tracking using Kalman filtering. *International Journal of Advanced Robotic Systems*, 13(2), 65.
56. Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1), 35-45.
57. İnternet: Brown, R. G. (1983). Introduction to Random Signal Analysis and Kalman Filtering, <http://www.webcitation.org/query?url=https://www.wiley.com/en-us/Introduction+to+Random+Signals+and+Applied+Kalman+Filtering+with+Matlab+Exercises%2C+4th+Edition-p-9780470609699> Son Erişim Tarihi: 11.06.2021
58. Wang, H., and Zhang, X. (2021). Real-time vehicle detection and tracking using 3D lidar. *Asian Journal of Control*, 2021, 1-11.
59. Munkres, J. (1957). Algorithms for the assignment and transportation problems. *Journal of The Society For Industrial And Applied Mathematics*, 5(1), 32-38.
60. İnternet: Duan, H., Zhao, Y., Chen, K., Shao, D., Lin, D., and Dai, B. (2021). Revisiting Skeleton-based Action Recognition. Yayınlanmamış metin. <http://www.webcitation.org/query?url=https://arxiv.org/pdf/2104.13586.pdf> Son Erişim Tarihi: 02.01.2022.
61. Han, F., Reily, B., Hoff, W., and Zhang, H. (2017). Space-time representation of people based on 3D skeletal data: A review. *Computer Vision and Image Understanding*, 158, 85-105.
62. Zhu, D., Zhang, Z., Cui, P., and Zhu, W. (2019). *Robust graph convolutional networks against adversarial attacks*. 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Anchorage AK, 1399-1407.
63. Shao, D., Zhao, Y., Dai, B., and Lin, D. (2020). *Finegym: A hierarchical video dataset for fine-grained action understanding*. IEEE/CVF Conference on Computer Vision and Pattern Recognition, Washington, 2616-2625.

64. Liu, J., Shahroudy, A., Perez, M., Wang, G., Duan, L. Y., and Kot, A. C. (2019). Ntu rgb+ d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(10), 2684-2701.
65. Carreira, J., and Zisserman, A. (2017). *Quo vadis, action recognition? a new model and the kinetics dataset*. IEEE Conference on Computer Vision and Pattern Recognition, Hawaii, 6299-6308.
66. Ibrahim, M. S., Muralidharan, S., Deng, Z., Vahdat, A., and Mori, G. (2016). *A hierarchical deep temporal model for group activity recognition*. IEEE Conference on Computer Vision and Pattern Recognition, Nevada, 1971-1980.
67. Zhang, S., Yang, Y., Xiao, J., Liu, X., Yang, Y., Xie, D., and Zhuang, Y. (2018). *Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks*. IEEE Transactions on Multimedia, 20(9), 2330-2343.
68. Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2), 157-166.
69. İnternet: Beat This - A Beat Synchronisation Project. URL: https://www.clear.rice.edu/elec301/Projects01/beat_sync/, Son Erişim Tarihi: 31.01.2022.
70. Masumitsu, K., and Echigo, T. (2000). *Video summarization using reinforcement learning in eigenspace*. International Conference on Image Processing, British Columbia, 267-270.
71. İnternet: Vidcon conference for and about online video, URL: <https://www.vidcon.com/>, Son Erişim Tarihi: 2021-07-15.
72. Brezeale, D., and Cook, D. J. (2008). Automatic video classification: A survey of the literature. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(3), 416-430.
73. Ekin, A., Tekalp, A. M., and Mehrotra, R. (2003). Automatic soccer video analysis and summarization. *IEEE Transactions on Image Processing*, 12(7), 796-807.
74. Ciocca, G., and Schettini, R. (2006). Dynamic storyboards for video content summarization. *8th ACM International Workshop on Multimedia Information Retrieval*, California, 259-268.
75. Jung, B., Kwak, T., Song, J., and Lee, Y. (2004). *Narrative abstraction model for story-oriented video*. 12th Annual ACM International Conference on Multimedia, New York, 828-835.
76. Xu, M., Maddage, N. C., Xu, C., Kankanhalli, M., and Tian, Q. (2003). *Creating audio keywords for event detection in soccer video*. International Conference on Multimedia and Expo, Maryland, II-281.
77. Money, A. G., and Agius, H. (2008). Video summarisation: A conceptual framework and survey of the state of the art. *Journal of Visual Communication and Image Representation*, 19(2), 121-143.

78. Qazi, H. A., Jahangir, U., Yousuf, B. M., and Noor, A. (2017). *Human action recognition using SIFT and HOG method*. International Conference on Information and Communication Technologies (ICICT), Karachi, 6-10.



GAZİ GELECEKTİR..