KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

UZAY-ZAMAN POZ ÇANTASI İLE 3B İNSAN HAREKETLERİNDEN EYLEM TANIMA

DOKTORA TEZİ

Saeid AGAHIAN

NİSAN 2018
TRABZON

KARADENİZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

UZAY-ZAMAN POZ ÇANTASI İLE 3B İNSAN HAREKETLERİNDEN EYLEM TANIMA

Saeid AGAHIAN

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsünce
"DOKTOR (BİLGİSAYAR MÜHENDİSLİĞİ)"
Unvanı Verilmesi İçin Kabul Edilen Tezdir.

Tezin Enstitüye Verildiği Tarih **10.04.2018**
Tezin Savunma Tarihi **30.04.2018**

Tez Danışmanı : Prof. Dr. Cemal KÖSE

Trabzon 2018

Karadeniz Teknik Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalında

Saeid AGAHIAN Tarafından hazırlanan

UZAY-ZAMAN POZ ÇANTASI İLE 3B İNSAN HAREKETLERİNDEN EYLEM TANIMA

başlıklı bu çalışma, Enstitü Yönetim Kurulunun 10/04/2018 gün ve 1748 sayılı kararıyla oluşturulan jüri tarafından yapılan sınavda

DOKTORA TEZİ

olarak kabul edilmiştir.

Jüri Üyeleri

Başkan : Prof. Dr. İbrahim TÜRKOĞLU …….

Üye : Prof. Dr. Nurhan KARABOĞA ……....

Üye : Prof. Dr. Cemal KÖSE ……

Üye : Dr. Öğr. Üyesi. Eyüp GEDİKLİ ……

Üye : Dr. Öğr. Üyesi. Tolga BERBER …...……

Prof. Dr. Sadettin KORKMAZ

Enstitü Müdürü

ÖNSÖZ

Görüntü tabanlı insan hareket tanıma, kolay ve etkili uygulanabildiğinden yaygın olarak kullanılmaktadır. Yapılmış çalışmaların çoğunda geleneksel kameralardan elde edilen iki boyutlu RGB verileri kullanılmıştır. Hâlbuki 3B verileri kullanınca daha başarılı sistemler elde edilebilir. Ancak 3B veriyi güvenilir bir şekilde elde etmek maliyet ve zaman açısından dezavantajdır. Son yıllarda hem Kinect algılayıcısının ortaya çıkmasıyla ve hem derin öğrenme tekniklerindeki gelişmelerle 3B verilerin daha ucuz, hızlı ve güvenilir bir şekilde elde etmesi sağlanabilmektedir. Bir insan eyleminde 3B iskelet dizilerin sağlanabilmesi, birçok araştırmacıyı Kinect algılayıcısını kullanmaya teşvik etmektedir. Bu tez çalışmasında eylem gerçekleştiren bir insanın 3B iskelet dizilerini ele alınarak, bu iskeletlerden çıkarılan uzay-zaman öznitelikleri kullanan bir poz çantası yöntemi ile eylemlerin tanınmasına çalışılmıştır.

Doktora çalışmamda danışmanlığımı üstlenerek bana çalışmalarımı yürütmemde her zaman sınırsız destek veren hocam Prof. Dr. Cemal KÖSE'ye, çalışma sürecinde değerli görüş ve katkılarını esirgemeyen sayın Dr. Öğr. Üyesi. Eyüp GEDİKLİ'ye ve Dr. Öğr. Üyesi. Tolga Berber'e teşekkürü bir borç bilirim. Doktora tezimde yer alan arkadaşlarım Farhood NEGIN, Mohammed MILANI ve Aref YELGHI'e ve ayrıca ilgileri ve desteklerini üzerimden eksik etmeyen sevgili babam, anneme ve 2016'dan beri hayat yoldaşım olan sevgili eşimden çok teşekkür ederim.

Saeid AGAHIAN

Trabzon, 2018

TEZ ETİK BEYANNAMESİ

Doktora Tezi olarak sunduğum "Uzay-Zaman Poz Çantası İle 3B İnsan Hareketlerinden Eylem Tanıma" başlıklı bu çalışmayı baştan sona kadar danışmanım Prof. Dr. Cemal Köse'nin sorumluluğunda tamamladığımı, verileri/örnekleri kendim topladığımı, deneyleri/analizleri ilgili laboratuvarlarda yaptığımı/yaptırdığımı, başka kaynaklardan aldığım bilgileri metinde ve kaynakçada eksiksiz olarak gösterdiğimi, çalışma sürecinde bilimsel araştırma ve etik kurallara uygun olarak davrandığımı ve aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim. 30/04/2018

Saeid AGAHIAN

İÇİNDEKİLER

Sayfa No

Doktora Tezi

ÖZET

UZAY-ZAMAN POZ ÇANTASI İLE 3B İNSAN HAREKETLERİNDEN EYLEM TANIMA

Saeid AGAHIAN

Karadeniz Teknik Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı
Danışman: Prof. Dr. Cemal KÖSE
2018, 101 Sayfa

Video işleme çalışmaları 1980'li yıllardan beri çalışılmaktadır. O zamandan beri bilgisayar görü alanında, insan aktivite tanıma en zorlu işlemlerden biri haline gelmiştir. Konuyla ilgili gerçekleştirilmiş çalışmalara rağmen, eylem tanıma ile ilgili birçok sorun halen daha çözümlenememiştir. Son yıllarda Kinect sensörün ortaya çıkması ve derin öğrenme tekniklerindeki gelişmeler güvenilebilir ve maliyeti ucuz olarak 3B insan iskeleti çıkarılabilmektedir. Yapılan tez çalışmasında insan eylem tanınması için 3B iskelet verilerini kullanan bir poz çantası yöntemi önerilmiştir. Çalışmada her bir eylem, önceden tanımlanmış uzay-zamansal anahtar pozlarla temsil edilmektedir. 3B pozları temsil eden uzay-zamansal tanımlayıcıların tanımlanması, alana en çok katkıda yapılan kısımdır. Poz tanımlayıcılar üç parçanın birleştirmesinden oluşturulmuştur. Birincisi 3B iskelet dizisinde ele alınan pozun normalleştirilmiş eklem konumları, ikincisi önceden belirlenmiş bir zaman ofset üzerinden aynı eklemlerden elde edilen yer değişim ve üçüncüsü ele alınan iskeletle bir önceki iskeletin eklemlerinden elde edilen yer değişim vektörleridir. Eğitim pozları üzerinde k-means kümeleme yöntemi uygulayarak anahtar pozlar elde edilmiştir. Daha sonra her eylem anahtar pozlar dizisine dönüştürülmüş ve anahtar poz histogramların elde edilmiştir. Son aşamada ELM sınıflandırıcı olarak kullanılmıştır. Önerilen yöntemin testi için 3B iskelet verilerine sahip ve yaygın kullanılan 5'tane eylem veri tabanı kullanılmıştır. Bunların üçünde, bilinen en başarılı sonuçlar elde edilmiş ve diğerlerinde en iyilerle karşılaştırılabilen sonuçlar elde edilmiştir.

Anahtar Kelimeler: 3B İskelet tabanlı eylem tanıma, Kelime çantası, Anahtar pozlar, Aşırı öğrenme makinesi, RGB-D

PhD Thesis

SUMMARY

ACTION RECOGNITION FROM 3D HUMAN MOVEMENTS WITH SPATIO-TEMPORAL BAG OF POSES

Saeid AGAHIAN

Karadeniz Technical University
The Graduate School of Natural and Applied Sciences
Computer Engineering Graduate Program
Supervisor: Prof. Dr. Cemal KÖSE
2018, 101 Pages

Video processing work has been worked since 1980's. Since that time, human activity recognition has become one of the most challenging tasks in the field of computer vision. Despite the studies that belong to the subject, many problems related to action recognition have not been solved yet. In recent years, with the emergence of the Microsoft Kinect sensor and the resurgence of deep learning methods, is provided cost-efficient and reliable 3D human skeleton. In this thesis, a bag-of-pose method which uses 3D skeletal data for the human action recognition have been proposed. In this study each action is represented by a set of predefined Spatio-temporal key poses. The definition of temporal-spatial descriptors to represent 3D poses is the main contribute of the study. The pose descriptors are consist of three parts concatenation. The first part is the normalized positions of 3D skeleton joints. The second is the displacement of the same joints of the poses over a predetermined time offset and the third part is the displacement vectors that obtained from the joints of the current and the previous skeleton. The Key poses has obtained by applying k-means clustering method on all of training poses. Later every action has been converted to a sequence of key poses and key poses histograms has obtained. In the last stage ELM was used as classifier. For the evaluation of the proposed method, have been used five popular action data sets that have 3D skeleton. Achieved state-of-the-art results on three of the datasets and competitive results on the other two datasets compared to the other methods.

Key Words: Skeleton-based 3D action recognition, Bag-of-words, Key poses, Extreme learning machine, RGB-D

ŞEKİLLER DİZİNİ

Sayfa No

TABLOLAR DİZİNİ

Sayfa No

SEMBOLLER DİZİNİ

DTW	Dynamic Time Warping
HMMs	Hidden Markov Models
DBNs	Dynamic Bayesian Networks
SVM	Support Vector Machines
KNN	K-Nearest Neighbors
ANN	Artificial Neural Network
GMM	Gaussian Mixture Model
CRF	Conditional Random Fields
CNNs	Convolutional Neural Networks
RNN	Recurrent Neural Networks
LSTM	Long Short Term Memory
ELM	Extreme Learning Machine
MEI	Motion Energy Image
MHI	Motion History Image
SIFT	Scale-Invariant Feature Transform
SURF	Speeded Up Robust Features
BoW	Bag of Words
BoF	Bag of Features
BoP	Bag of Poses
HoF	Histogram of Flow
HoG	Histogram of Gradients
LOSubO	Leave-one-subject-out cross validation
LOSeqO	Leave-one-sequence-out cross validation
CS	Cross-Subject test
MoCap	Motion Capture systems
PCA	Principal Component Analysis
LDA	Linear Discriminant Analysis
MLE	Maximum Likelihood Estimation
MAP	Maximum a posteriori probability
SOS	Skeleton Optical Spectra
RF	Random Forest Classifier
STIPs	Spatio-Temporal Interest Points
RGB	Red-Green-Blue
RGB-D	Red-Green-Blue and depth

1. GENEL BİLGİLER

1.1. Giriş

Video tabanlı insan eylem tanıma, etkin uygulanabilirlik açısından birçok araştırmacı tarafından ayrıntılı olarak incelenmektedir. İnsan aktivite tanıma sistemindeki amaç, bir videoda (bir görüntü kare dizisi) devam eden faaliyetleri otomatik olarak analiz etmektir. Sistemin amacı, bir videonun yalnızca bir insan faaliyetinin yürütülmesini içerecek şekilde bölümlendirildiği basit bir durumda videoyu eylem kategorisine göre doğru bir şekilde sınıflandırmaktır. Daha genel durumlarda ise, insan aktivitelerinin sürekli bir şekilde tanınması bir giriş videosundan ortaya çıkan tüm aktivitelerin başlangıç ve bitiş zamanlarının tespit edilerek gerçekleştirilmesidir. Aktivite tanıma, akıllı ev sistemleri (smart home systems), insan bilgisayar etkileşimi (human computer interaction), robotik görme (robot vision), sağlık hizmetleri (healthcare), davranışsal biyometri (behavioral biometrics), animasyon (animation), artırılmış gerçeklik (augmented reality), video özetlemek ve indeksleme (video summarization and indexing) gibi çeşitli alanlarda kullanılmaktır [1-3]. Konuyla ilgili yapılmış çalışmalara rağmen, eylem tanıma ile ilgili birçok sorun halen daha çözümlenememiştir. Bu nedenle, eylem tanıma günümüzde halen aktif bir araştırma alanı olarak görülmektedir.

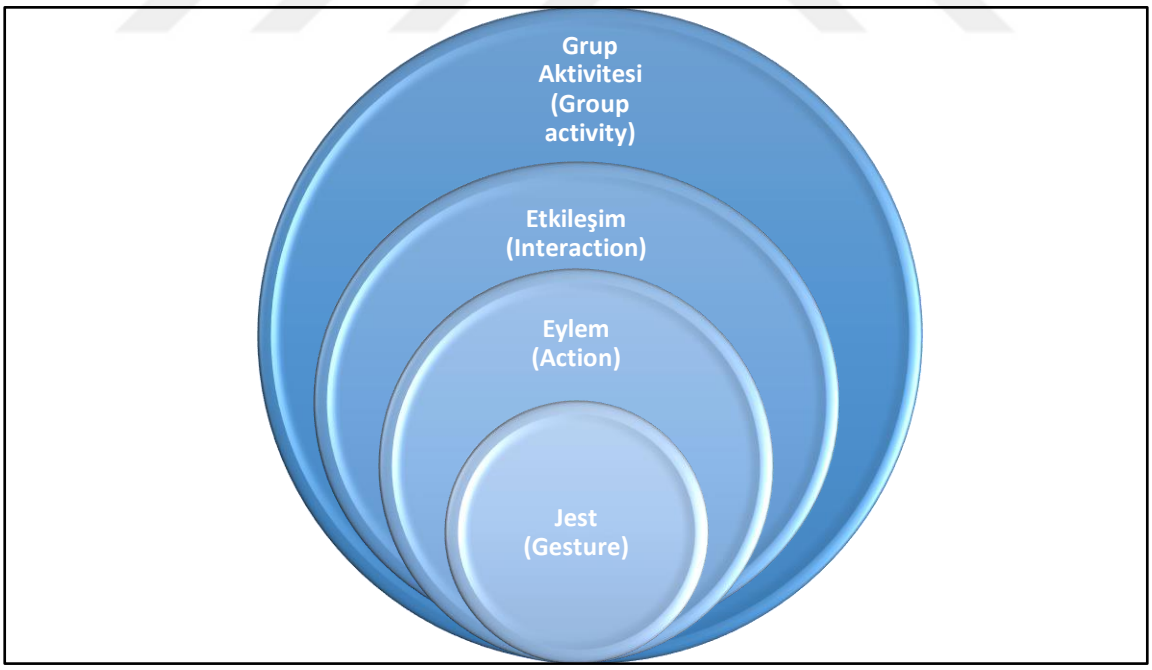
Görüş açısı, hız, ivme ve gövde boyutunda oluşabilecek değişkenlikler, eylemlerin sınıf içi farklılıkları ve sınıflar arası benzerlikler bu çalışma alanında karşılaşılan zorlukların en önemlileridir. Bununla birlikte, videoda gözlemlenen bir eylemin zamansal ve uzamsal bölütlenmesi, eylemlerin anlamsal olarak ayrıştırılması ve yeterli eğitim verisinin temin edilememesi konuyla ilgili karşılaşılabilecek diğer zorluklardır. Genel bir eylem tanıma sisteminin oluşturulabilmesi için bu sorunlara karşı oluşturulabilecek olası çözümler ele alınmalıdır [4].

Birçok araştırmacı tarafından kabul edilen Aggarwal vd. [1]'in incelemesine göre insan aktiviteleri karmaşıklık açısından Şekil 1'de gösterildiği gibi 4 düzeyde incelenebilir:

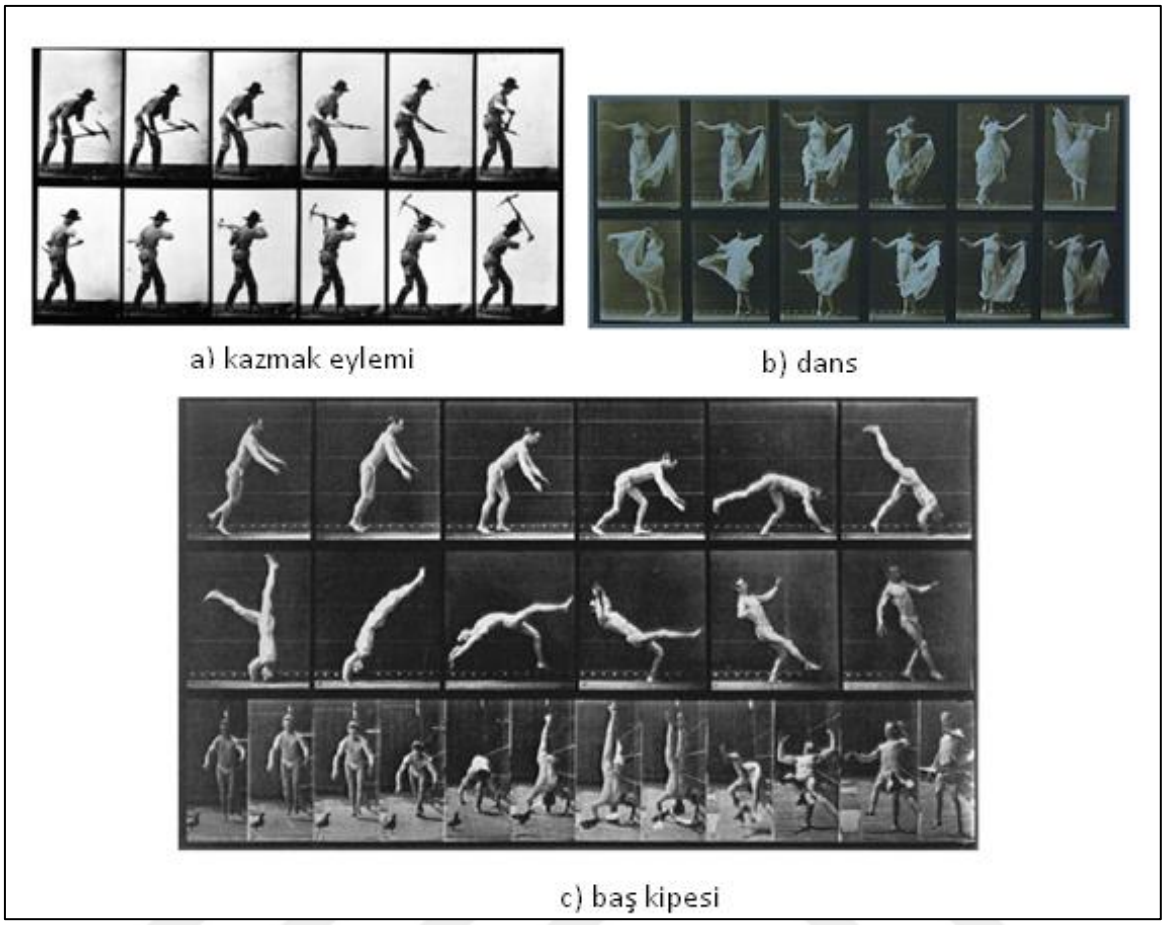
- Jest (Gesture): Bir kişinin vücut parçalarının temel hareketleri olup kişinin anlamlı bir hareketini anlatan atomik bileşenleridir. Örneğin "kolun gerdirilmesi, esnetilmesi", "bacağın kaldırılması" vb.
- Eylemler (Action): Zamansal olarak düzenli birçok jestten oluşan tek kişilik aktivitelerdir. Örneğin: "yürüyüş", "el sallamak", "yumruk atmak" vb.

- Etkileşimler (Interaction): İki veya daha fazla kişi/nesneyi içeren insan aktiviteleridir. Örneğin "tokalaşma" aktivitesi iki kişi arasındaki bir etkileşimdir. "bavul taşımak" aktivitesi ise iki kişi ve bir nesneyi içeren bir insan-nesne etkileşimidir.
- Grup aktiviteleri (Group activity): Bu aktiviteler birçok kişi ve/veya nesneyi içeren anlamsal gruplar tarafından yapılan aktivitelerdir. Örneğin "bir grup halinde yürüyen kişiler", "futbol oynamak" vb.

Turaga vd. [5] insan aktivitelerini karmaşıklık açısından "Action" ve "Activity" olmak üzere iki ana kategoriye ayırmıştır "Action" genelde bir kişi tarafından gerçekleştirilip basit örüntü içeren hareketleri belirtir ve genellikle çok kısa süreli hareketlerdir. Örneğin yürüme, yüzme ve esneme hareketleri. Diğer taraftan "Activity" birçok kişinin katılımı ve tanımlı kuralların etkileşimiyle yapılan karmaşık bir dizi "Action" hareketlerini içerir. Bu hareketler genellikle uzun sürelidir. Örneğin iki kişinin el sallaması, bir futbol takımının gol atması ve birden fazla soyguncunun katılımıyla gerçekleştirilen banka saldırısı bu hareketlere örnek verilebilir.



Şekil 1. İnsan aktivitelerinin karmaşıklık düzeyi



Şekil 2. İnsan hareketleri üzerinde ilk fotoğrafçılık çalışmaları [6].

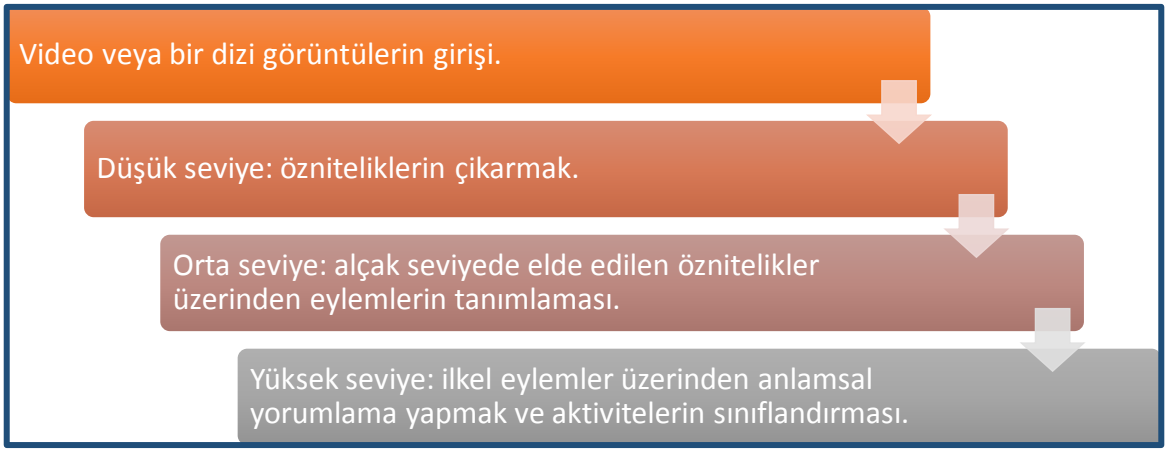
Pope [2], eylem ilkelerini uzuv seviyesinde tanımlanabilen bir atomik hareket olarak tanımlar. Buna göre, eylem terimi basit ve ilkel olanlardan döngüsel beden hareketlerine doğru çeşitli hareketleri tanımlar. Aktivite terimi, karmaşık bir hareketi temsil eden bir dizi müteakip eylemi tanımlamak için kullanılır. Örneğin, sol bacak ileriye doğru koşma bir ilkeli eylemdir. Engelleri atlama ise başlama, koşma ve atlama eylemleriyle gerçekleştirilen bir aktivitedir [7].

Günümüzde birçok araştırmacı insan hareketlerinin ayırt edilmesi ve eylemlerinin tanınmasıyla ilgilenmektedir. Eadweard Muybridge [8]'in1887'de başlayan (Şekil 2) ve sonrasında devam eden hareket-eylem analizi ve sentezi ile ilgili çalışmaları oldukça ilgi çekmiştir ve hareket resimlerinin gelişmesine neden olmuştur.

Birçok araştırmacı tarafından, insan aktivitesini tanımak için aşağıdan yukarıya (bottom-up) bir yapı takip edilir. Bu sistemlerin ana aşamaları şu şekildedir: öznitelik çıkarma, aktivite öğrenme ve sınıflandırma, aktivite tanıma ve segmentasyon [9]. En basit yapıdaki görüntü tabanlı insan hareket tanıma sistemleri öznitelik çıkarma ve özniteliklerin

sınıflandırılması işlemlerinin birleşimi olarak tanımlanabilir [2]. Turaga vd. [5] çalışmasında bir aktivite tanıma sistemini bir dizi görüntülerden başlayan ve yüksek düzeyli yorumlamayla biten çok aşamalı bir işlem olarak tanımlanmıştır. Bu sistemin ana aşamaları Şekil 3'de gösterilmiştir.

- Düşük seviye: Videolar, uzay-zamansal boyutta büyük miktarda piksel biçiminde ham bilgiler içerir. Ancak bu bilgilerin çoğu anlama görevi ve videoda meydana gelen aktivitenin belirlenmesiyle doğrudan alakalı değildir. Bu aşama arka ve ön planda bölütleme işlemini yapan, nesnelerin tanınma ve takip işlemini gerçekleştiren modüllerden oluşmaktadır.
- Orta seviyede: Eylem modelleme yapan yaklaşımlar üç ana sınıfa şu şekilde ayrılabilir: parametrik olmayan (non-parametric), hacimsel (volumetric), ve parametrik zaman serileri (Parametric time-series), Parametrik olmayan yaklaşımlar genellikle videonun her karesinden (frame) birtakım özellikler çıkarırlar. Özellikler daha sonra saklanmış bir şablonla eşleştirilir. Diğer taraftan, hacimsel yaklaşımlar kare bazında özellikleri çıkarmaz. Bunun yerine, bir videoyu 3B boyutlu bir piksel şiddetleri olarak düşünürler ve ölçek-uzay ekstrema (scale-space extrema), uzamsal filtre tepkileri (spatial filter responses) vb. gibi standart görüntü özniteliklerini 3B duruma genişletirler. Parametrik zaman serileri yaklaşımları, özellikle hareketin zamansal dinamikleri üzerine bir model oluştururlar. Daha sonra, sınıf eylemleri için modelde olan özel parametreler, eğitim verilerinden tahmin edilirler. Parametrik yaklaşımlara örnek olarak gizli Markov Modelleri (HMM'ler), Lineer Dinamik Sistemler (LDSs) vb. verilebilir.
- Yüksek seviyede: Gözetleme ve içerik tabanlı indeksleme uygulamalarının ilgi alanındaki aktiviteler birden fazla kişi içerir. Bunlar sadece birbiriyle değil, aynı zamanda içeriksel şeyler ile de etkileşime girerler. Şimdiye kadar tartışılan yaklaşımların çoğu, tek bir kişi eylemlerinin modellemeyi ve tanımı ile ilgilidir. Karmaşık bir sahneyi, karmaşık aktivitelerin içyapısı ve semantiğin modellemesi için daha üst düzey temsil ve mantık yürütme yöntemleri gerekir. Yapılan yaklaşımlar üç ana sınıfa; grafiksel modeller, sözdizimsel yaklaşımlar (syntactic approaches) ve Bilgi ve Mantık Tabanlı Yaklaşımlar (Knowledge and Logic-based Approaches) ayrılabilir.



Şekil 3. Aktivite tanıma sisteminin aşamaları

Bir eylem, düşük, orta ve yüksek adlandırılan üç öznitelik düzeyde ele alınabilir [10, 11]. İlk çalışmaların çoğunda, duruş (posture) yüksek seviyeli insan poz tanımlayıcısı ve bunların eklem yörüngelerinin tümünün birleştirmesi eylem tanıması için kullanılmıştır. Fakat insan vücut parçası algılamasında, güvenilir poz düzeltmesi ve yüksek hesaplama maliyetinde olan sıkıntılar nedeniyle araştırmacılar başka bir alternatif yol bulmaya zorlanmışladır [11, 12].

İlk çalışmalarda, araştırmacılar RGB videolarından seyrek veya yoğun olarak çıkartılmış düşük düzeyli özellikleri tanıtmaya çalışılmıştır. İnsan siluetleri, arka plan çıkarma işleminin uygulanabilir olduğu ortamlarda, insan eylemi tanıma için yaygın olarak kullanılan bu tür özelliklerin ilk örneklerinden biridir [13].

İlk kez Laptev ve Lindeberg [14], Harris kenar algılayıcısını 3B uzayda çalışacak bir şekilde geliştirmiş ve eylem tanıma için uzay-zamansal ilgi noktaları (Spatio-Temporal Interest Points)(STIPs) olarak adlandırılan seyrek özellik üretmiştiler. Bu özelliğin tanıtımı, önceden metin işlemede kullanılan kelimeler-çantası yönteminin video üzerinde eylemin tanınması için uydurup kullanmasında büyük başarıya yol açmıştır [15]. Optik akış, Efros vd. [22] tarafından insan eylemini temsili ve tanıması için kullanılan bir başka düşük düzey özniteliktir. Aynı anda, görüntü sınıflandırmasında bazı görüntü temelli özelliklerin başarısı, araştırmacının video sınıflandırma için onları kullanmasını teşvik etmiştir. Bunlar arasında eylem tanıma için HOG3D [16] ve SIFT3D [17] kullanılmıştır. Fakat düşük düzeyli özniteliklerin çıkarılması sadece RGB verileriyle sınırlı değildir. Örneğin, [18]'te sağlanan derinlik görüntüleri, parlaklık görüntüleri olarak kabul edilir ve düşük düzeyli görünüm (appearance) tabanlı öznitelik çıkarma yöntemleri kullanılır.

Düşük düzeyli özniteliklerle birlikte kelimeler-çantası yöntemlerini kullanmak için bazı sınırlamalar vardır. Esas dezavantaj, öznitelikler arasındaki uzay ve zamansal ilişkileri temsil etme kısıtlamasıdır [19]. Bu sınırlamanın üstesinden gelmek için, bazı araştırmacılar, düşük seviyeli özellikler arasındaki zamansal ve mekânsal bağımlılığı modellemek için orta düzey özellikler önermişlerdir. Örneğin, [20, 21]'de eylem de olan yerel hareketin temsili için ilgili nokta yerine bir semantik yapıya sahip hareket yörüngeleri incelenmiştir.

Bunlara rağmen düşük ve orta düzeyli öznitelikleri kullanan yöntemlerin ortak dezavantajı bunların anlamsal bilgi sunma sınırlamalarından dolayı, karmaşık hareketleri ifade etme yetersizliğidir [22]. Olası bir çözüm yüksek düzeyli anlamsal öznitelik tanımlanmasıyla sağlanmıştır [11], eylem tanımlayıcısı, bir dizi vücut pozunun uzay-zamansal bilgilerini içeren anlamsal sözlüklerden oluşmaktadır.

1.2. Tezin Kapsamı ve Amacı

Son zamanlarda maliyeti düşük Microsoft Kinect algılayıcı uzun süreli zor bir sorun olan yüksek düzeyli poz çıkarma işlemini gerçek zamanda işaretleyici kullanmadan tek bir derinlik görüntüsünden sağlamıştır [23, 24]. Etiket kullanmadan doğru şekilde poz bulma konusunda gerçekleştirilmiş gelişmeler bu konuda birçok avantaj sağlamıştır. Örneğin düşük ve orta özniteliklerle kıyasladığında görüş açısı, kişinin dış görünüşü ve ölçeğinde olan değişikliklere karşı, dirençli eylem tanımlayıcı sağlamıştır. Bu avantajlar birçok araştırmacının ilgisini çekmiş ve bu tür veri çok sayıda çalışmada öznitelik çıkarma işlemi için girdi olarak kullanılmıştır [11]. Eylem tanıma için bu tür bilgilerin kullanılmasındaki esas zorluk, anlamsal benzerliği olan eylemlerin bazen sayısal temsillerinin çok farklı olabileceğidir.

Pozların zamansal bilgilerinin saklaması için birçok araştırma yapılmıştır. Hareket örüntülerinde meydana gelen süre değişkenlikle baş edebilmek için en popüler şekilde kullanılan yöntem sıralı piramitlerdir [25]. Bu yöntemde ilk eylemler dizisi zaman ekseninde parçalara bölünüp ve daha sonra her bir parça için ayrı histogram elde edilir. Bazı araştırmacılar [11, 26] tarafından poz tanımlayıcısında zaman bilgiyi korumak amacıyla hız gibi zamansal özniteliklerin eklemesi önerilmiştir.

Geleneksel üretken yöntemlerde zamansal yapı, eylemlerin dinamiğinin anlaşılması için gerekli gösterilse bile bu derin ağlar (örneğin ConvNet ana yapısı) için kritik bir husus olarak sayılmaz. Bu yöntemlerin ilk odak noktası videolarda uzun dönemli zamansal yapılar

içeren eylemler yerine görünüş ve kısa dönem hareketleridir (kısıtlı sayıda görüntü kareleri). Buna rağmen zamansal sıra birçok eylem için önemli karakterlerden birisidir.

Metin ve ses işleme tanımasında Recurrent Neural Network (RNN) ve Long Short-Term Memory (LSTM) ağları dizide olan zamansal bağımlılıkların modellenmesinde oldukça başarılı sonuçlar elde etmiştir [27]. Bu alanlarda RNN ve LSTM'le elde edilen başarılar, araştırmacılara RNN [28, 29] ve LSTM [30-32]'in iskelet bilgiler ile kullanması için ilham kaynağı olmuştur. Bu ağların hesapsal karmaşıklığının fazla olması nedeniyle bunlar gerçek zamanlı ve online işlemler için uygun değildir [33, 34]. Buna rağmen son zamanlarda LSTM kullanarak eylem dinamiğinin modellemesinde HMMs ve sıralı piramitlere göre daha iyi sonuçlar elde edilmiştir [31].

Öngörüyle zamansal sıralamayı içeren bir video gösteriminin daha iyi ayrımcı özelliklere sahip olması beklenmektedir, Oysa her şeyi kapsayan bir gösterimin elde edilmesi halen daha önemli bir zorluk olarak görülmektedir.

Yukarıda adı geçen araştırmalar incelendiğinde, önerilen yöntemlerin (özellikle Kelime çantası tabanlı yöntemler) pozlar arasındaki zaman kavramını ve ilişkiyi tamamen modellemede halen daha başarısız olduğu görülmektedir [11, 25, 35]. Bu çalışmada, bu sorunu çözmek için poz tabanlı eylem tanıma yöntem önerilmiştir. Basitlik, yorumlanabilirlik ve eylem tanıma işlemindeki yüksek işlem performansı önerilen yöntemin en önemli avantajlarındandır. Ana fikir, bir eylemi bir dizi önceden tanımlanmış pozlarla tarif etmek ve bu işlemin ardından bunu pozların histogramı ile kodlamaktır [36]. Önerilen yöntemin akış diyagramı Şekil 32'de gösterilmektedir.

Dizide bulunan pozları basit ve etkili öznitelik tanımlanarak tarif ediyoruz. Farklı düzende, aynı pozlardan oluşan iki eylem, bu öznitelikler kullanınca ayrıt edilebilir.

Yapılan çalışmada eğitim pozlarından elde edilen tanımlayıcılar üzerinden anahtar pozlar oluşturulmuştur. Anahtar pozlardaki gömülü zamansal bilgilerin, eylem kodlaması için oluşturulan anahtar poz histogramında kullanılması kelime çantası (BoW) yöntemlerinin sınırlamalarını aşmamızda yardımcı olmuştur. Eylemleri temsil eden öznitelik vektörünün uzunluğu sabittir ve görüntü çerçeve sayısından bağımsızdır. Son aşamada, eylemleri sınıflandırmak için sınıflandırıcı olarak Aşırı Öğrenme Makinesi [37] kullanılmıştır.

Önerilen yöntem, 3B iskelet verileri kapsayan, halka açık araştırma amaçlı beş veri seti (benchmark) üzerinde test edilmiştir. Deneyler, çalışmada önerilen yöntemin, üç veri

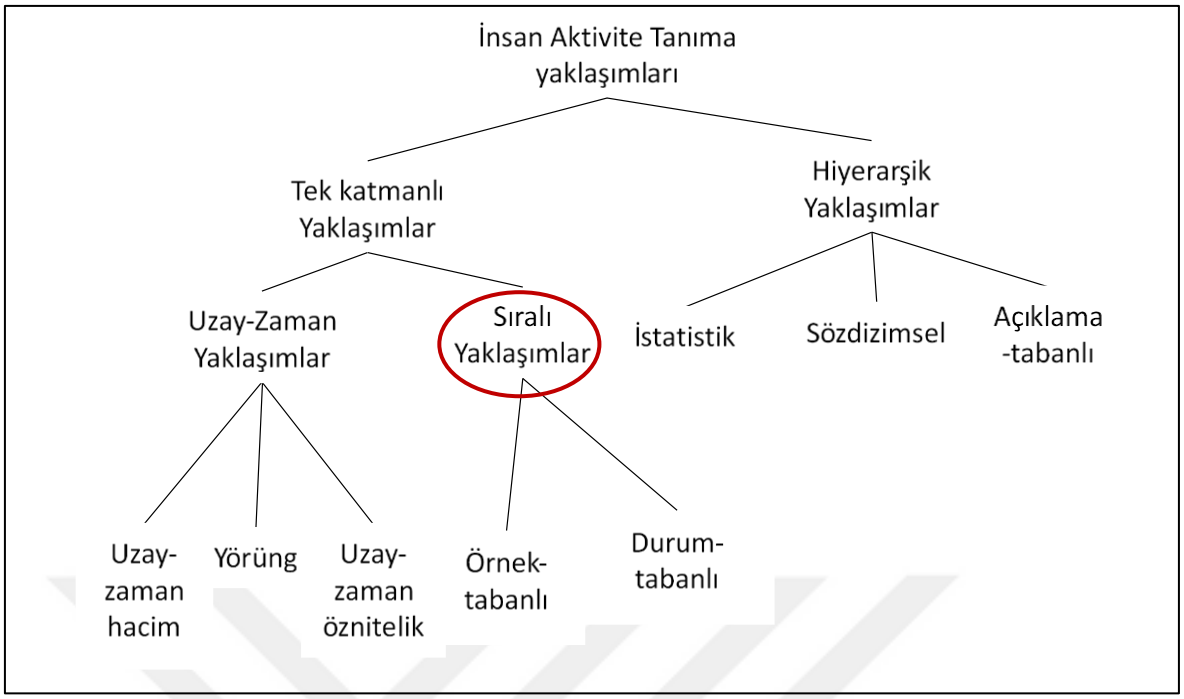
setinde şu ana kadar bilinen en iyi sonucu, dördüncü ve beşinci veri setinde iskelet tabanlı yöntemler arasında kıyaslanabilir sonuçların üretebildiğini göstermiştir.

1.3. Görüntü Tabanlı İnsan Eylem Tanıma Yaklaşımlarının Sınıflandırılması

İnsan aktivite tanıması için çok sayıda yaklaşımlar önerilmiştir. Literatürde yaklaşımlar farklı kriterlere göre sınıflandırılmıştır [9]. Poppe [2] görüntü üzerinden insan aktivite tanımasını ve aktivitelerin sınıflandırmasını ayrı ayrı ele almıştır. Turaga vd. [5] tanıma problemini karmaşıklık açısından eylem ve aktivite (Activity) olarak ikiye ayırmış ve tanıma yaklaşımlarını aktivitelerin karmaşıklık derecesi üzerinden ele alınma yeteneğine göre sınıflandırmıştır. Aggarwal ve Ryoo [1] çalışması bu konudaki en kapsamlı çalışmalardan birisidir. Çalışmada, bu alanda önemli olan gelişmeler karşılaştırılmış ve Şekil 4'de gösterildiği gibi yöntemler kategorize edilmiştir. Daha sonra yapılan çalışmalarda bu sınıflandırma sıkça kullanılan bir taksonomi haline dönüşmüştür.

Aktivitelerin giriş görüntülerinden doğrudan algılanıp algılanmamasına bağlı olarak, aktivite tanıma yöntemleri iki önemli kategoriye ayrılmıştır: Tek-katmanlı yaklaşımlar ve Hiyerarşik yaklaşımlar.

Doğası nedeniyle tek-katmanlı yaklaşımlar jest ve ardışık bir karaktere sahip eylemlerin tanınması için daha uygundurlar. Diğer yandan, hiyerarşik yaklaşımlar yüksek seviyeli insan aktivitelerini, alt olay (subevents) olarak adlandırılan diğer basit eylemler aracılığıyla tanımlamaktadırlar. Birden çok katmandan oluşan tanıma sistemleri inşa edilmiştir ve dolaysıyla karmaşık aktivitelerin analizi için uygundurlar [1].

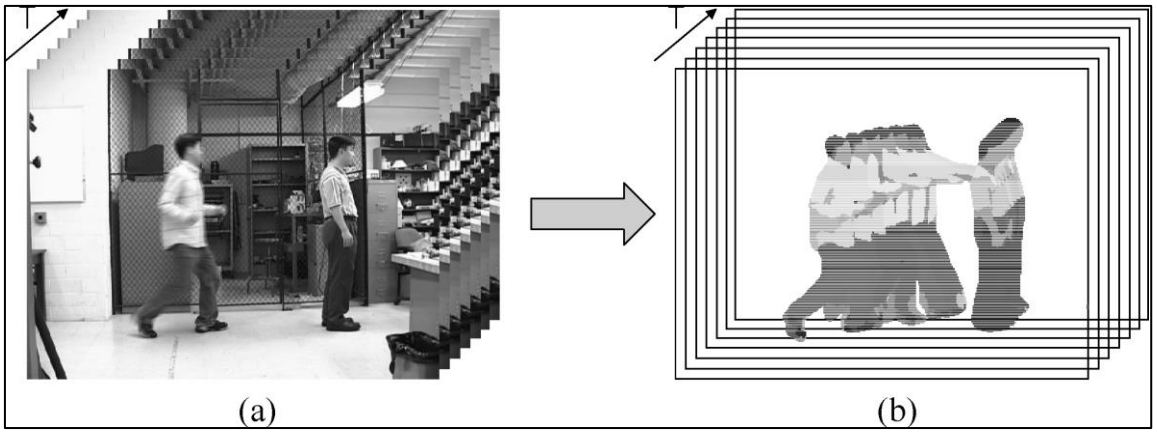


Şekil 4. Hiyerarşik yaklaşıma dayalı taksonomi [1].

1.3.1. Tek-Katmanlı Yaklaşımlar

İnsan aktiviteleri nasıl modellendiğine bağlı olarak iki sınıfa ayrılırlar; uzay-zaman yaklaşımları (space-time approaches), ardışık yaklaşımlar (sequential approaches). Aralarındaki en önemli fark, zamana ait (3B XYZ uzayında üçüncü boyut) boyuttur. Uzay-zamansal yaklaşımları, zamanı normal bir boyut olarak ele alırlar ve 3B hacimden öznitelik kümesini oluştururlar. Ardışık yaklaşımlar ise, bir insan aktivitesini zaman ekseni boyunca sıralı gözlemler olarak dikkate alırlar. Daha spesifik olarak, bir insan aktivitesini bir dizi görüntüden elde edilen öznitelik vektörleri ile temsil ederler. Ardışık ilişkileri dikkate aldıkları için, ardışık yaklaşımlar genellikle uzay-zaman yaklaşımlarından başarılıdırlar [9].

Çoğu eylem tanıma sistemi, giriş videolarından elde edilir. Burada tartışılan tüm videolar 2-B uzay (XY) görüntülerin zamansal (T) dizisinden oluşur. Bu nedenle, bir video uzamsal-zamansal hacim Şekil 5'de verildiği gibi temsil edilebilir. Bu hacim insan ve bilgisayar için hacimdeki eylemleri ve aktiviteleri tanımak için gerekli bilgileri içerir.



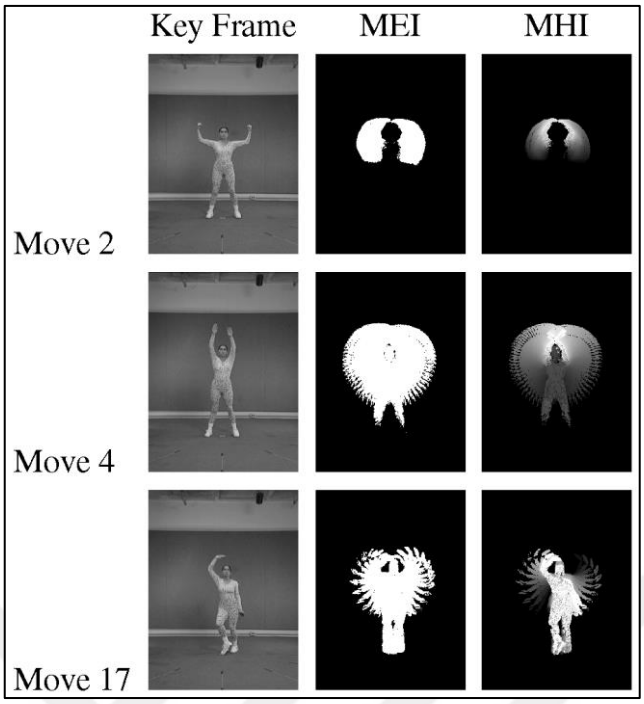
Şekil 5. Yumruk atma eylemi için oluşturulan örnek XYT hacim a) bütün görüntü kareler b) çıkarılmış ön plan [1].

1.3.1.1. Uzay-Zaman Yaklaşımları

Şekil 4'de verildiği gibi 3B uzay-zaman hacim üzerinden kullanılan özeliklere bağlı olarak uzay-zaman yaklaşımları 3'e ayrılır;

Uzay- Zaman hacim: En sezgisel uzay-zaman hacim yaklaşımı, 3 boyutlu hacmin tamamını öznitelik veya şablon olarak kullanır ve bilinmeyen eylem videolarını mevcut olanlarla eşleştirerek sınıflandırmayı elde eder. Doğru benzerlikleri hesaplamak için, çeşitli uzay-zaman hacim temsilleri ve tanıma metodolojileri geliştirilmiştir. Gürültü ve anlamsız arka plan bilgileri yöntemi negatif olarak etkiler, Dolaysıyla hareket modellenmesi için sadece kişinin bulundu yerin ön planı (silhouettes) üzerinden bazı girişimler yapılmıştır.

Şekil 6'de verildiği gibi Bobick ve Davis [38] şablon eşleştirmeyi kullanarak gerçek zamanlı bir hareket tanıma sistemi oluşturmuşturlar. Her bir eylemin 3 boyutlu uzay-zaman hacmini hesaplamak yerine, her eylemi 2 boyutlu görüntüden oluşan bir şablonla temsil etmişler; iki boyutlu binary değer Hareket Enerji Görüntüsü (motion-energy image) ve skaler değerli Hareket Geçmişi Görüntüsü (motion history image). Uzay-zaman hacim yaklaşımlarının en büyük dezavantajı, birden fazla kişiyi içeren aktivitelerin tanımasında yaşanan zorluktur. Yaklaşımların çoğu, bu sorunu çözmek için kayan pencere algoritmasını (sliding window) kullanırlar. Ancak, bu algoritmada eylemlerin doğru yerini belirlemek için çok sayıda hesap gerekir.

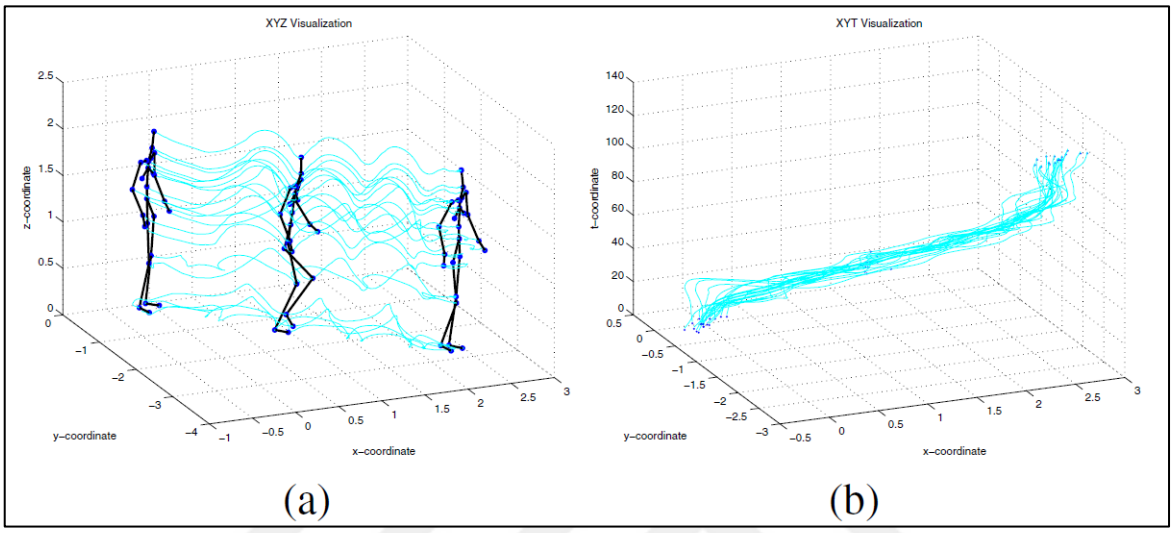


Şekil 6. Hareket enerji görüntüsü ve hareket geçmiş görüntüsü [38].

Yörünge ye dayalı yaklaşımlar (Trajectories): Yörünge temelli yaklaşımlar, bir aktiviteyi bir dizi uzay-zaman yörüngesi olarak yorumlayan tanıma yaklaşımlarıdır. Yörünge temelli yaklaşımlarda bir kişi, genel olarak eklem pozisyonlarına karşılık gelen 2 boyutlu **(**XY) veya 3 boyutlu (XYZ) noktalarla temsil edilir. İnsan vücudu parçaların tahmin yöntemleri, özellikle çubuk figür (stick figure) modellemesi, her bir görüntü karesindeki bir kişinin eklem pozisyonlarını çıkarmak için yaygın olarak kullanılmıştır. Bir insan bir eylem gerçekleştirdiğinde, eklem pozisyonundaki değişiklikler, uzay-zamanı yörüngeleri olarak kaydedilir, inşa edilmiş 3B XYT ya da 4B XYZT eylemi temsil eder. Şekil 7'de örnek yörüngeleri gösterilmiştir. Johansson [39] tarafından yapılan öncü çalışmada, eklem pozisyonlarının izlenmesinin, insan eylemlerinin ayırt edebilmeleri için yeterli olduğu öne sürülmüştür.

Bu yaklaşımların en büyük avantajı, insan hareketlerinin ayrıntılarını analiz edebilme yetenekleridir. Ayrıca, bu yöntemlerin çoğu, kamera bakış açısından bağımsızdır. Ancak, bunu yapmak için, bu tür yöntemlerde bir sahnede görünen kişilerin 3-B XYZ eklem yerlerinin doğru bir şekilde tahmin edilebilmesi için genellikle düşük seviyeli bir bileşen gerektirir.

Uzay-Zaman Yerel Öznitelikleri Kullanan Yaklaşımlar: Bu bölümde ele alınan yaklaşımlar, aktiviteleri temsil etmek ve tanımak için 3B uzay-zaman hacimlerinden çıkarılan yerel öznitelikleri kullanırlar. Bu yaklaşımların ardındaki motivasyon, 3 boyutlu bir uzay-zaman hacmi aslında katı bir 3B objesi olmasıdır.



Şekil 7. Yürüme eyleminde iskelet eklemlerinin yörüngesi a) 4B XYZ uzayı, b) 3B XYT uzayı [40].

Bir sistem, her bir eylemin 3B hacminin niteliklerini açıklayan uygun öznitelikleri çıkarabiliyorsa, eylem tanıma, nesne eşleştirme problemi çözünmesi anlamına gelir.

Nesne tanıma sürecine benzer şekilde, sistem ilk olarak bir kişinin 3-D uzay-zaman hacminden yerel hareket bilgisini yakalamak için tasarlanmış belirli yerel öznitelikleri (local-feature or interest points or local descriptors) çıkarır. Bu öznitelikler daha sonra uzay-zamansal ilişkiler dikkate alınırken veya ilişkiler görmezden gelinerek aktivitelerin temsil edilmesi için birleştirilir. Son olarak, aktiviteleri sınıflandırmak için tanıma algoritmaları uygulanır. Literatürde bulunan eylem tanıması için yerel öznitelikleri kullanan yaklaşımların daha detaylı bir şekilde incelenmesi için [12] derlemesi önerilir.

Yerel tanımlayıcıları çıkaran uzay-zaman yaklaşımları çeşitli avantajlara sahiptir. Doğası gereği, arka plan çıkarma veya diğer düşük düzey bileşenlere genellikle ihtiyaç yoktur ve yerel öznitelikler çoğu durumda ölçek, rotasyon ve ötelemeden bağımsızdır. Periyodik eylemler tekrar tekrar öznitelik desenleri ürettiğinden dolayı yürüyüş ve sallama gibi özellikle basit periyodik eylemleri tanımak için uygundurlar. Uzay-zaman öznitelik-temelli yaklaşımların temel sınırlaması, daha karmaşık aktivitelerin modellenmesi için

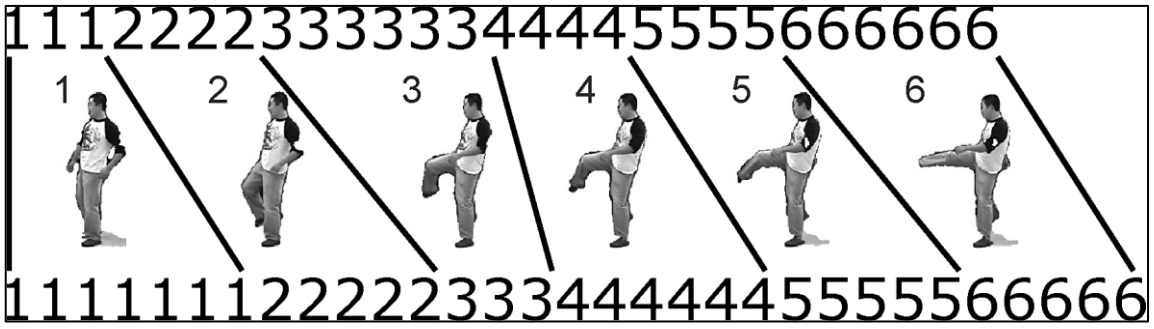
uygun olmamalarıdır. Belirli bir süre alan periyodik olmayan aktivite için öznitelikler arasındaki ilişkiler önemlidir.

1.3.1.2. Ardışık Yaklaşımları

Ardışık yaklaşımlar, özniteliklerin dizilerini analiz ederek insan faaliyetlerini tanıyan tek katmanlı yaklaşımlardır. Bu tür yaklaşımlar bir girdi videosunu bir görüntü dizisi olarak (özellik vektörleri) dikkatte alırlar ve aktiviteyi karakterize eden belirli bir dizi gözlemleyebildikleri takdirde videoda bir aktivite meydana geldiği sonucuna varırlar.

Ardışık yaklaşımlar önce bir dizi görüntüyü bir dizi özellik vektörüne dönüştürür, bu işlem her bir görüntü karesi üzerinde bulunan şahsın konumunu açıklayan öznitelikleri (örneğin, eklem açıların derecesi) çıkararak yapılabilir. Özellik vektörleri çıkarıldıktan sonra, sıralı yaklaşımlar, özellik vektörlerinin aktiviteyi gerçekleştiren kişi tarafından üretilme olasılığını ölçmek için diziyi analiz eder. Dizi ile aktivite sınıfı (veya dizi aktivite sınıfına ait posterior olasılığı) arasındaki olasılık yeterince yüksekse, sistem aktivitenin gerçekleştiğine karar verir. Ardışık yaklaşımlar, metodolojiye dayalı bir taksonomi kullanarak iki kategoriye ayırılmıştır;

Örnek Tabanlı Yaklaşımları (Exemplar-based): Örnek-tabanlı ardışık yaklaşımlar, doğrudan eğitim örneklerini kullanarak insani eylem sınıflarını tanımlar. Her sınıf için temsilci dizi veya bir set eğitim dizileri sağlanır ve bunlarla eşleşme yapılarak gelen yeni dizi aktivite sınıfı tanımlanır. Her bir sınıf temsilci veya örnekleriyle eşleşme yapılarak benzerlikleri yeterince yüksekse, sistem verilen girdinin aktiviteyi gerçekleştirdiğini belirleyebilir. İnsanlar farklı tarzlarda ve farklı tempoda aynı aktiviteyi gerçekleştirebilirler ve bu çeşitlilikler dikkate alınarak benzerlik ölçülmelidir. Başlangıçta ses işleme için geliştirilmiş dinamik zaman bükmesi (DTW) algoritması, iki farklı boyutlu dizilerin eşleştirilmesi için yaygın olarak kullanılmıştır. DTW algoritması, iki dizi arasında optimal bir doğrusal olmayan eşleşmeyi polinom zamanda bulur. Şekil 8'de farklı icra tempolarına sahip iki sekans (yani diziler) arasında kavramsal bir eşleşmeyi gösterilmiştir.



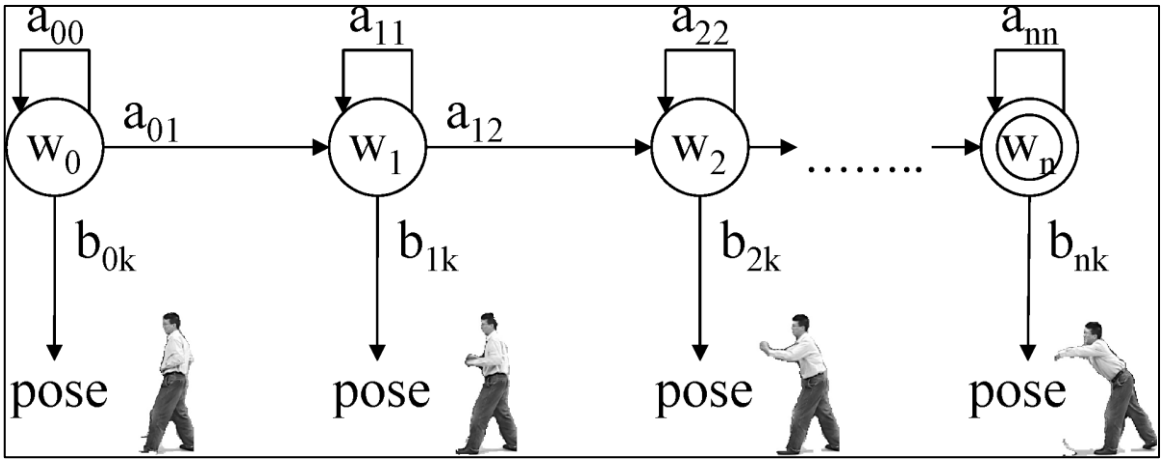
Şekil 8. Farklı tempoda iki "bacak germe" dizileri arasındaki eşleştirme [1].

Durum Tabanlı Yaklaşımlar (State-based): bir aktiviteyi Bir set durum ile oluşan modelle temsil eden ardışık yaklaşımlardır. Model istatistiksel olarak eğitilmiştir, böylece aktivite sınıfına ait özellik vektörlerinin dizilerine karşılık gelir. Daha spesifik olarak, istatistiksel model belirli bir olasılıkla bir dizi oluşturmak için tasarlanmıştır. Genel olarak, her aktivite için bir istatistiksel model oluşturulur. Her model için, modelin gözlemlenen bir dizi özellik vektörünü oluşturma olasılığı, eylem modeli ile girdi görüntüsü dizisi arasındaki olasılığın ölçülmesi için hesaplanır. Faaliyetleri tanımak için, en yüksek olasılık tahmini (MLE) veya maksimum bir posteriori olasılık (MAP) sınıflandırıcısı oluşturulur.

Durum tabanlı yaklaşımlar için gizli Markov modelleri (HMM'ler) ve dinamik Bayes ağları (DBN'ler) yaygın olarak kullanılmıştır. Her iki husus da bir aktivite bir set gizli durumlarla temsil edilir. Bir şahısın her zaman diliminde sadece bir durumda olduğu varsayılır ve her durum bir gözlem (yani, bir özellik vektörü) üretir. Bir sonraki görüntü karesinde, durumlar arasındaki geçiş olasılığını dikkate alarak sistem başka bir duruma geçmektedir. Modellerin geçiş (transition) ve gözlem (observation) olasılıkları eğitildikten sonra, genelde değerlendirme problemi çözülerek eylemler teşhis edilir.

Değerlendirme problemi, verilmiş bir dizinin (yani yeni girdi) olasılığının belirli bir durum tabanlı modeli tarafından üretildiğinde, hesaplanmasıdır. Hesaplanan olasılık yeterince yüksekse, durum tabanlı yaklaşımlar, söz konusu girdide kendilerine karşılık gelen aktivitenin gerçekleştiğine karar verebilir. Şekil 9'da sıralı bir HMM'in bir örneğini göstermektedir.

Temel HMM'in esas sınırlaması, iki veya daha fazla kişiyi hareketlerinden oluşan aktiviteleri temsil edememesidir. Bir HMM sıralı bir modeldir ve bir seferde sadece bir durum aktive edilir ve bu birden çok kişi içeren aktivitelerini modellemesini engeller.



Şekil 9. Kol germe eylemi için örnek bir gizli markov modeli [1].

Ardışık yaklaşımlar, uzay-zaman yaklaşımlarla kıyasladığında daha karmaşık aktiviteleri modelleyebilir. Örnek-tabanlı yaklaşımların eğitilmesi durum tabanlı yaklaşımlara göre daha az örnekle yapılabilir. Durum tabanlı yaklaşımların sınırlamalarından birisi, tanımak istedikleri aktivite daha karmaşık hale geldikçe çok sayıda eğitim videosuna ihtiyaç olmasıdır.

1.3.2. Hiyerarşik Yaklaşımlar

Hiyerarşik yaklaşımların ardındaki esas fikir, diğer benzer faaliyetlerin tanınma sonuçlarına dayalı olarak üst düzey etkinliklerin tanınmasını sağlamaktır. Diğer benzer aktivitelerin tanıma sonuçlarına dayalı olarak üst düzey aktivitelerin tanınmasını sağlar. Motivasyon, ilk önce tanınması daha basit olan alt aktivitelerin tanımasına izin vermek ve daha sonra üst düzeyli aktivitelerin tanımasında bunları kullanmaktır. Örneğin, "kavga" gibi üst düzey bir etkileşim, bir dizi yumruklama ve tekme etkileşimleri tespit edilerek tanınabilir.

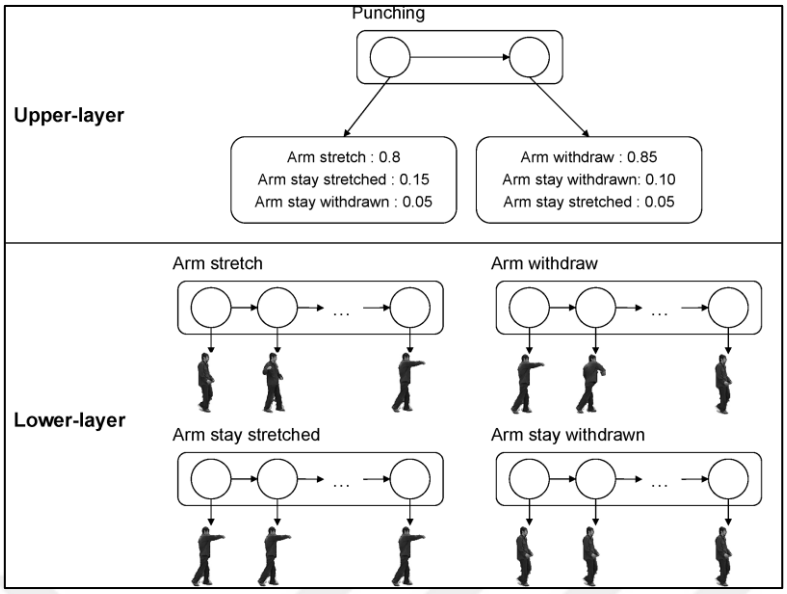
Çoğu hiyerarşik yaklaşımda, bu atomik eylemler, önceki bölümde sunduğumuz tek katmanlı tanıma yöntemlerini kullanarak tanınır. Hiyerarşik yaklaşımların hiyerarşik olmayan yaklaşımlarla kıyasladığında en büyük avantajı (yani, tek katmanlı yaklaşımlar) daha karmaşık yapılara sahip üst düzey aktiviteleri tanıma yetenekleridir. Hiyerarşik yaklaşımlar özellikle, insanlar veya nesneler arasındaki etkileşimlerin yanısıra karmaşık grup aktivitelerin semantik düzey analizi için uygundur. Bu avantaj, hiyerarşik yaklaşımların

iki kabiliyetinin bir sonucudur; daha az eğitim verisi ile başa çıkma yeteneği ve ön bilgiyi temsile dâhil etme yeteneği. Hiyerarşik yaklaşımlar kendi içerisinde üç kategoriye ayrılır;

1.3.2.1.Statiksel Yaklaşımlar

İstatistiksel (statistical) yaklaşımlar, aktiviteleri tanımak için istatistiksel durum tabanlı modelleri kullanır. Hiyerarşik istatistiksel yaklaşımlar söz konusu olduğunda, HMM'ler ve DBM'ler gibi çok sayıda durum tabanlı model (genellikle iki katman) katmanları, sıralı yapılarla aktiviteleri tanımak için kullanılır. Alt katmanda, atomik eylemler, tek katmanlı ardışık yaklaşımlarda olduğu gibi, özellik vektörlerinin dizilerinden tanınır. Sonuç olarak, bir dizi özellik vektörü bir dizi atomik eylemlere dönüştürülür. İkinci katmanda ki modeller atomik eylemlerin dizisinin gözlemler olarak ele alır. Her model için, bir gözlem (yani, atomik seviyeli eylemler) dizisinden üretilen olasılık değeri hesaplanır. Hesaplanan değer aktivite ve giriş görüntüsü dizisi arasındaki benzerliği ölçmesi için kullanılır. Sonuç olarak, en yüksek olasılık tahmini (maximum likelihood estimation, MLE) veya maksimum bir posteriori olasılık (maximum a posteriori probability, MAP) sınıflandırıcısı oluşturulur.

Şekil 10'de bir kişini yumruk atma eyleminin tanıması için tasarlanmış bir istatistiksel hiyerarşik model gösterilmektedir. Önerilen model iki katmandan oluşmuştur. Alt katmanda HMMs'leri kullanılarak germe ve geri alma gibi çeşitli atomik eylemler tanımlanıyor ve üst katmanda bulunan HMM, alt katmandaki HMM'lerin tanıma sonuçlarını bir girdi olarak kabul eder. Bu da yumruk vurma eyleminin bir dizide gerilme ve geri çekilmenin gerçekleştiğinden kaynaklanır.



Şekil 10. Yumruk atma eyleminin tanınması için tasarlanmış bir istatistiksel hiyerarşik model [1].

Ayrıca, karmaşık aktivitelerin tanınması için DBN'leri kullanan hiyerarşik yaklaşımlar üzerinde çalışılmıştır. DBN'ler, birden fazla gizli durum katman içerebilir ve bu da, hiyerarşik insan aktivitelerini temsil edecek şekilde formüle edilebileceğini ileri sürer. Özellikle ardışık aktivitelerin tanınması için istatistiksel yaklaşımlar uygundur. Yeterli eğitim verisiyle, istatistiksel modeller gürültülü girişler durumunda bile ilgili aktiviteleri güvenilir bir şekilde tanıyabilir. İstatistiksel yaklaşımların esas sınırlamaları eşzamanlı alt öğelerden oluşan bir aktivite gibi, karmaşık zamansal yapıya sahip aktiviteleri için doğasında olan tanıma yetersizlikleridir. Sonuç olarak bu yaklaşımlarda kullanılan HMM ve DBN'ler eşzamanlı ilişkiler değil, ardışık ilişkileri modellemek için uygundurlar.

1.3.2.2.Sözdizimsel Yaklaşımlar

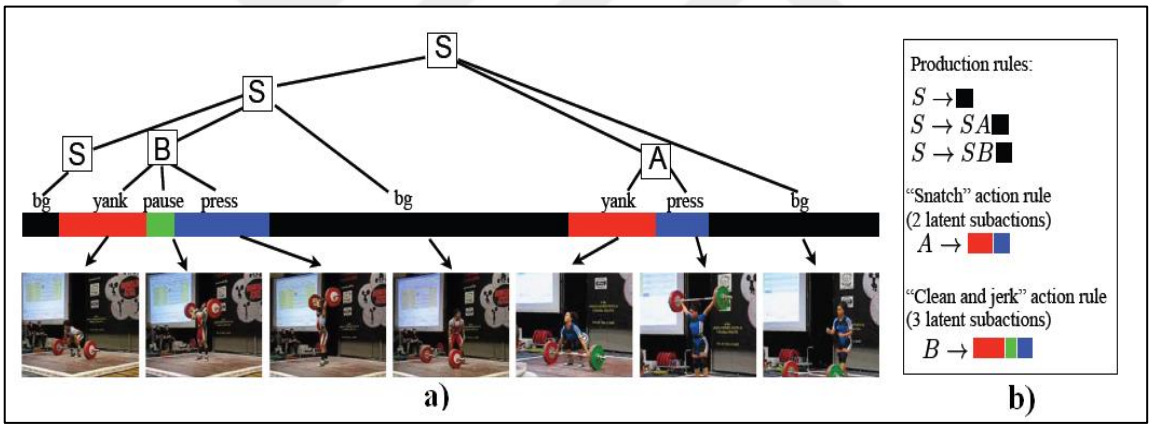
Sözdizimsel (syntactic) yaklaşımlar, insan aktivitelerini, semboller dizisi olarak modellemektedir. Burada her sembol atomik düzeydeki bir eyleme karşılık gelmektedir. Hiyerarşik istatistiksel yaklaşımlara benzer olarak, sözdizimsel yaklaşımlarda, daha önceki tekniklerden herhangi birini kullanarak, ilk önce atomik düzeydeki eylemlerin tanınmasını gerektirir. İnsan aktiviteleri, bir dizi atomik eylemi üreten bir set üretim kuralı olarak temsil edilir ve programlama dilleri alanından ayrıştırma (parser) tekniklerini benimsemeyi kullanıp bu aktiviteleri tanırlar.

İçerikten bağımsız gramer (Context-free grammars, CFG'ler) ve stokastik metinden bağımsız gramerler (stochastic context-free grammars, SCFG'ler) önceki araştırmacılar tarafından üst düzey aktiviteleri tanımak için kullanılmıştır. CFG'lerin üretim kuralları doğası nedeniyle aktivitelerin hiyerarşik bir temsile ve tanınmasına yol açmaktadır.

Şekil 11'de kavga etkileşiminin basitleştirilmiş bir tarifi ve tanıması için SCFG üretim kuraları gösterilmektedir. Şekil 12'de video üzerinde olan halter kaldırma eyleminin ayrıştırılması için CFG grameri verilmektedir.

Fighting	*→ Punching*	*:0.3*	*Punching*	*→ stretch withdraw*	*:0.8*
	| Punching Fighting	*:0.7*		*| stretch stay_withdraw*	*:0.1*
				| stay_stretch withdraw	*:0.1*

Şekil 11. Kavga etkileşimini beyan ve tanıma için örnek SCFG [1].



Şekil 12. a) Bir videonun hiyerarşik ayrıştırması ile eylemlere bölütlemesi b) bunu gerçekleştiren CFG gramer [41].

Sözdizimsel yaklaşımların kısıtlamalarından biri, eş zamanlı aktivitelerin tanınmasıdır. Sözdizimsel yaklaşımlar, olasılıksal olarak sıralı alt maddelerden oluşan hiyerarşik aktiviteleri tanımlayabilir, ancak eşzamanlı alt gruplardan oluşan aktiviteler için doğası gereği sınırlıdır. Sözdizimsel yaklaşımlar üst düzey bir aktiviteyi bunu oluşturan bir dizi atomik düzeydeki eylemlerle modellediğinde, atomik düzey etkinliklerin zamansal sıralaması kesinlikle sıralı olmalıdır.

Ayrıca, sözdizimsel yaklaşımlar tüm gözlemlerin üretim kurallarını uygulayarak ayrıştırıldığını varsaymaktadır. Bu sistemlerde, kullanıcı tüm olası olaylar için bir set üretim

kuralları sağlamalıdır. Bu nedenle, bilinmeyen bir gözlem (örneğin bir yaya) sisteme müdahale ettiğinde zorluğa sebep olur. Böyle bir sınırlamanın üstesinden gelmek için otomatik olarak gözlemlerden dilbilgisi kurallarını öğrenmek için çalışmalar yapılmıştır.

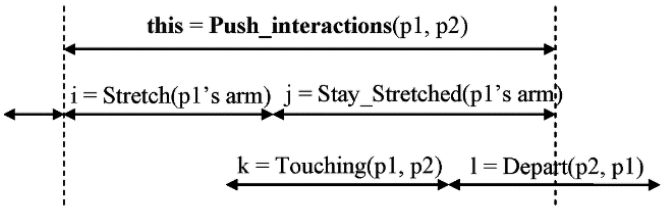
1.3.2.3.Açıklama Tabanlı Yaklaşımlar

Açıklama tabanlı (description-based) bir yaklaşım, insan aktivitelerinin uzay-zamansal yapılarını açık bir şekilde koruyan bir tanıma yaklaşımıdır. Aktiviteyi oluşturan daha basit eylemler (diğer bir deyişle, alt gruplar) ve onlarda olan zamansal, uzay ve mantıksal ilişkiler üzerinden yüksek düzeyde bir insan aktivitesini temsil eder. Yani, açıklamaya dayalı yaklaşımlar, bir insan faaliyetini, belirli ilişkileri tatmin eden (kendi alt bölümlerinden oluşmuş) alt eylemlerin bir oluşumu olarak modellemektedir.

Bu nedenle, aktivitenin tanınması, temsilinde belirtilen ilişkileri tatmin eden alt eylemleri araştırarak gerçekleştirilir. Tüm açıklama tabanlı yaklaşımlar, doğası gereği hiyerarşiktir (insan aktivitelerini temsil etmek için alt eylemleri kullanmaları nedeniyle) ve bunlar eşzamanlı yapıya sahip aktiviteleri ele alabilirler.

Açıklama tabanlı yaklaşımlarda, bir zaman aralığı (time interval), genellikle alt eylemler arasında gerekli zamansal ilişkileri belirtmek için onlarla birleşmiştir. Allen'in zamansal belirtmeler (Allen's temporal predicates), zaman aralıkları arasındaki ilişkileri belirlemek için bu yaklaşımlarda yaygın olarak kullanılmıştır. Allen'in tanımlanmış yedi temel belirtecileri şunlardır; before, meets, overlaps, during, starts, finishes ve equals. Before ve meet belirticileri sıralı ilişkileri temsili ederken diğer belirticiler eşzamanlı ilişkileri belirtmek için kullanılmıştır. Şekil 13 (a)'de, insan-insan itme etkileşiminin kavramsal zamansal yapısı, zaman aralıklar cinsinden temsil edilir.

Açıklama tabanlı yaklaşımların sınırlamalarından biri, çoğunlukla deterministik olmaları ve düşük seviyeli bileşenlerinin gürültülü olmasından dolayı kırılgan olmalarıdır.



(a)

```
Push_interaction(p1, p2) = {
  list( def('i', Stretch(p1's arm)),
      list( def('j', Stay_stretched(p2's arm)),
          list( def('k', Touching(p1, p2)),
              def('l', Depart(p2, p1)) ))),
  and( meets('i', 'j'),
      and( during('k', 'this'),
          and( meets('k', 'l'),
              and( starts('i', 'this'),
                  finishes('j', 'this') ))))
};
```

(b)

Şekil 13. a) İtme etkileşim ve alt eylemreninde zaman aralığı b) biçimsel gösterim [42].

1.4. 3B Verilerini Kullanarak İnsan Aktivitesi Tanıma

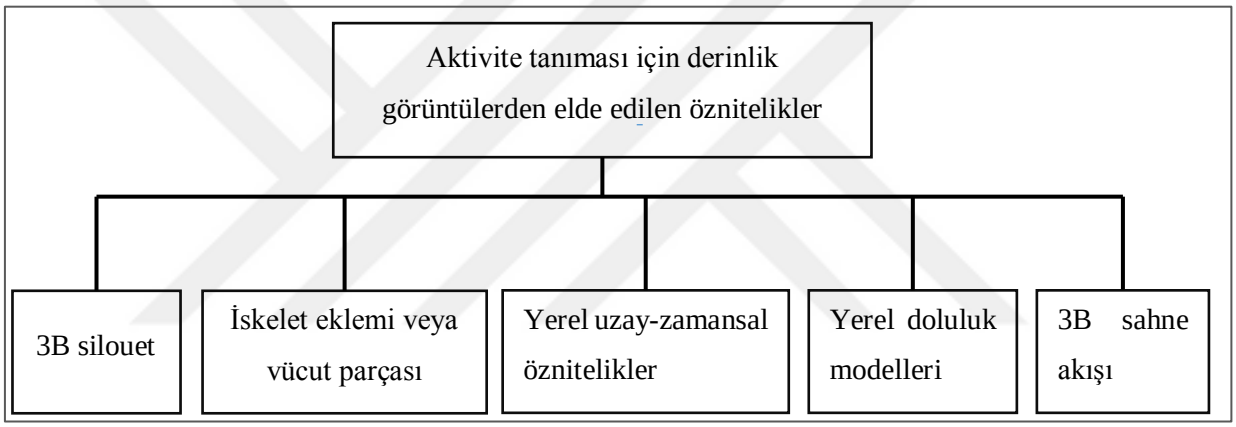
Derinlik sensörlerinden insan aktivitesinin tanınması 1980'lerin başlarında başlamıştır [43]. Aktivite öğrenme ve tanımada üzerinde geçmişte yapılan araştırmalar çoğunlukla görünür ışıklı kameralar tarafından çekilen video sekanslarının üzerine odaklanmıştır. Görünür ışıklı videolar ile ilgili esas sorun, monoküler video sensörlerinden eklemli insan hareketini yakalamak, önemli ölçüde bilgi kaybına yol açmasıdır. Bu, video tabanlı insan aktivite tanımasının performansını sınırlandırır [43]. Son on yıldaki çabalara rağmen, videolardan insan faaliyetlerini tanımak hala zor bir iştir.

Yakın zamanda uygun maliyetli derinlik sensörlerinin piyasaya sürülmesiyle, 3B veriler üzerinde başka bir araştırma büyümesi görüyoruz. Aggarwal J. ve Xia Lu [43] son 20 yıl içinde 3B veriler üzerinde araştırmalar artmıştır. Birincisi MoCap gibi işaretçi kullananlar, ikincisi 3B verileri oluşturmak için iki ve ya fazla sayıda kameralar kullanan stereo yöntemleri ve üçüncüsü de derinlik sensörlerini kullanan yöntemlerdir.

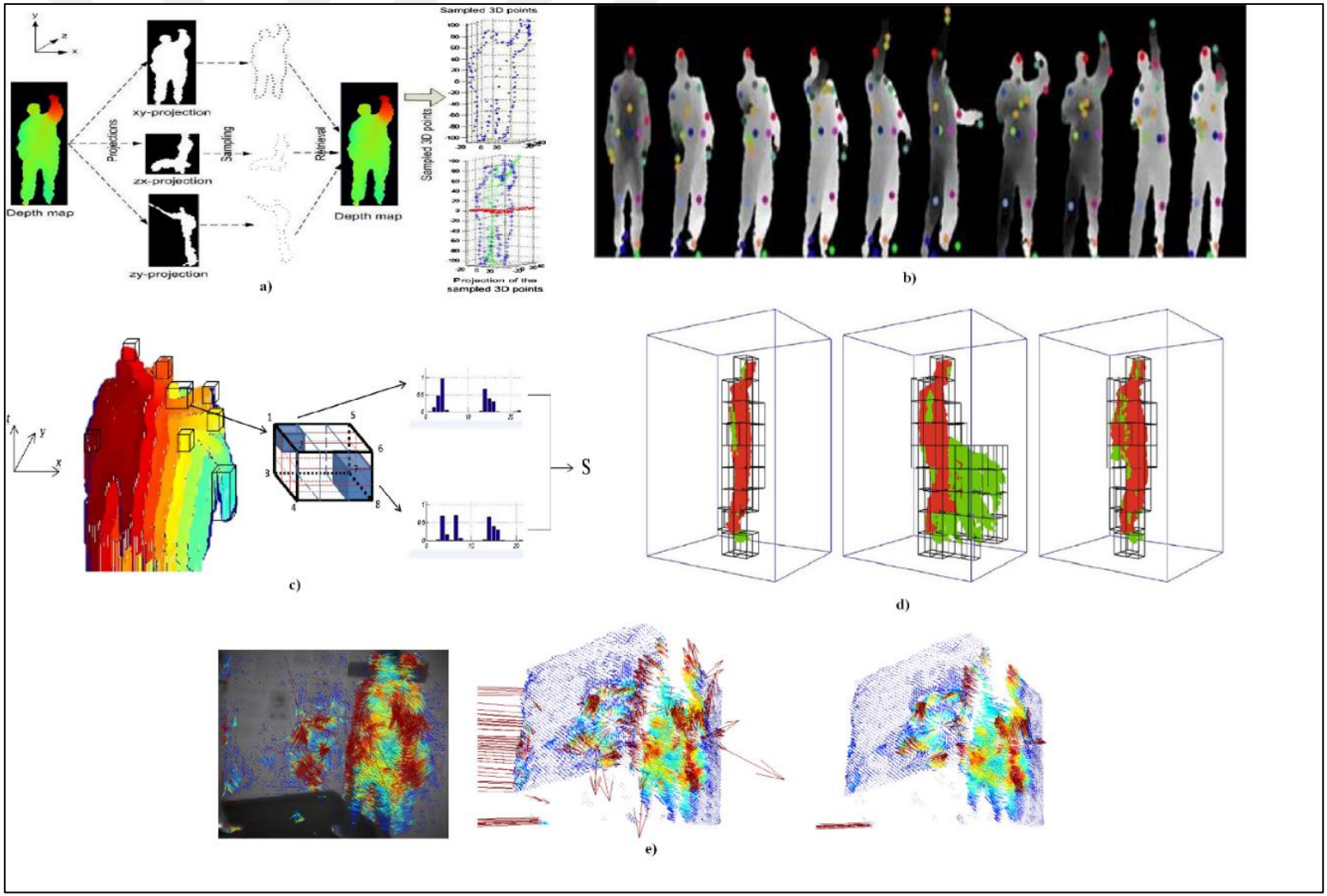
Görüntü temelli insan eylemi tanıması için dört ana sorun bulunmaktadır. Birincisi düşük seviyeli zorluklardır. Örtüşme, dağınık arka plan, gölgeler ve değişen aydınlatma koşulları, hareket segmentasyonu için zorluklar üretebilir ve eylemlerin algılanma şeklini değiştirebilir. Bu, RGB videolar üzerinden insan aktivite tanımaya çalışan yöntemlerin esas zorluğudur. 3D verilerinin tanıtımı, sahnenin yapı bilgisini sağlayarak düşük seviyedeki zorlukları büyük ölçüde hafifletir. İkinci zorluk görüş açısı değişiklikleridir. Aynı aktivite farklı görüş açılarından farklı gözükebilir. Bu sorunun çözümü geleneksel RGB kamera kullandığında sadece senkronize edilmiş birden fazla kamerayla mümkün olabilir. Üçüncü zorluk ölçek değişikliğidir ve bu problem kişilerin kamerayla aralarında farklı mesafelerin

bulunması veya kişilerine farklı ebatlarda olmasından kaynaklanır. Derinlik videolarında ise kişilerin gerçek ebat bilindiğinden dolayı kolaylıkla ayarlanabilir.

Bir derinlik görüntüsünde, her bir pikselin değeri, gerçek dünya noktası ile sensörler arasındaki mesafeye karşılık gelir ve bu sahnenin 3B yapısal bilgilerin sağlar. Son birkaç yılda, derinlikli görüntülerden insan aktivite tanıma problemini ele alan çeşitli yöntemler önerilmiştir. Şekil 14'de gösterildiği gibi Aggarwal J. ve Xia Lu [43] 3B verileri kullanan yöntemleri, kullandıkları özniteliklere göre, beş kategoriye ayırmıştırlar: 3B siluetlerin öznitelikleri, iskelet eklemi veya vücut parçası konumlarından öznitelikler, yerel uzay-zamansal öznitelikler, yerel doluluk modelleri ve 3B sahne akışı öznitelikleri. Şekil 15'da her bir kategoride olan öznitelik için görsel olarak bir örnek verilmektedir.



Şekil 14. Derinlik verilerinden insan aktivite tanımada kullanılan özelliklerin taksonomisi [43].



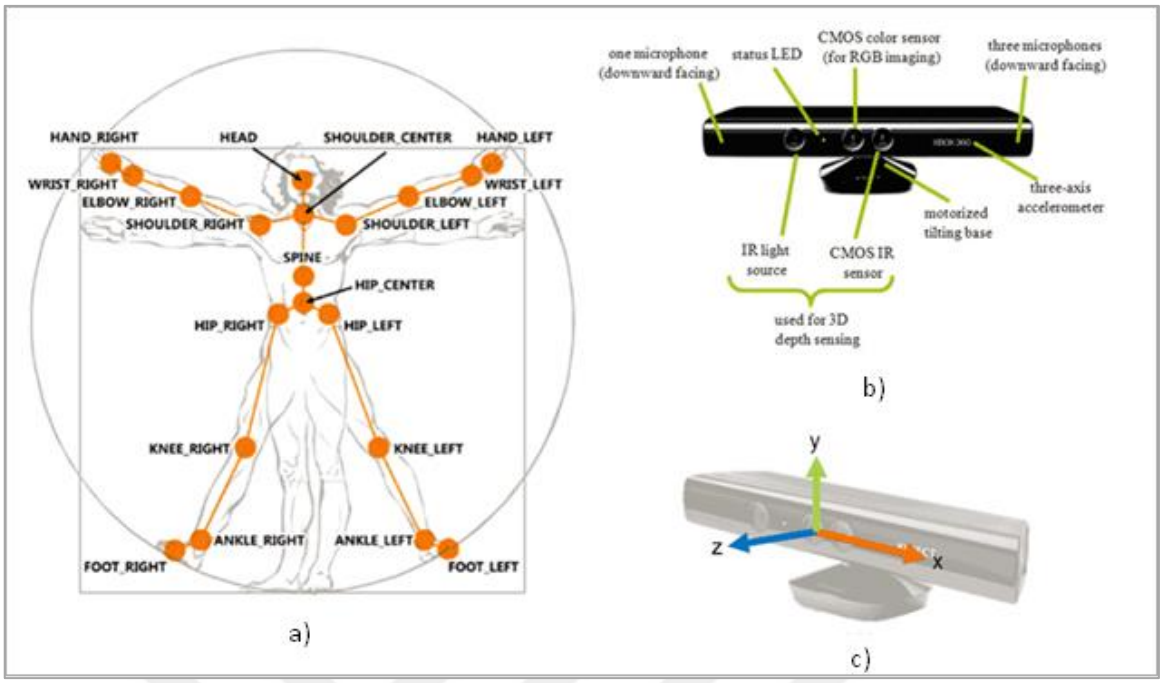
Şekil 15. a) 3B silüetlerin b) iskelet eklemi c) yerel uzay-zamansal d) yerel doluluk e) 3B sahne akışı öznitelikleri [43].

2. 3B POZ ÇIKARMA TEKNİKLERİ

Literatürde tek kamera kullanarak RGB görüntüden 3B insan pozun çıkarması için çalışmalar [44] olsa bile, 3B iskelet elde etmenin geleneksel yöntemleri; MoCap veya stereo kullanmasıdır Önceden bahsettiğimiz gibi MoCap kullanmak hem maliyet açısından ve hem hesaplama karmaşıklığı açısından çok pahalıdır ve bu nedenle günlük aktivite tanıması için kullanılmamıştır. Son yıllarda bu verileri elde etmesi daha basitleşip ve daha az maliyetle sağlanır. Bunun birinci nedeni az maliyetli Kinect kameranın ortaya çıkması ve diğeri derin öğrenme tekniklerinde olan gelişmelerdir.

2.1.1. Kinect Sensör ile Çıkartılan 3B Poz

İlk nesil Microsoft Kinect (Xbox 360) tabanlı sensörler, derinlik tahmini için yapılandırılmış ışık (structured light) kullanır. Sistem renkli kamera, kızılötesi kamera ve kızılötesi lazer içerir (Şekil 16 (B)). Kızıl ötesi lazer, kırınım ızgarası mekanizmasından geçen birden çok ışığa bölündüğü bir ışık demeti yayar. Bu bir projektörün çalışmasını taklit eder. Yansıtılan model kızılötesi kamera tarafından yakalanır ve bilinen referans koduna göre karşılaştırılır. Yansıtılan kod ile gözlenen kod arasındaki fark, sahnedeki derinlik bilgisini verir [45]. Hem RGB video hem de 3D derinlik verilerinin sağlanması, bilgisayar görme araştırmasının birçok klasik probleminde önemli bir gelişmedir. Kinect [46] algoritmayı kullanarak derinlik verilerinden gerçek zamanlı olarak 20 eklemden oluşan 3B iskeleti elde eder.

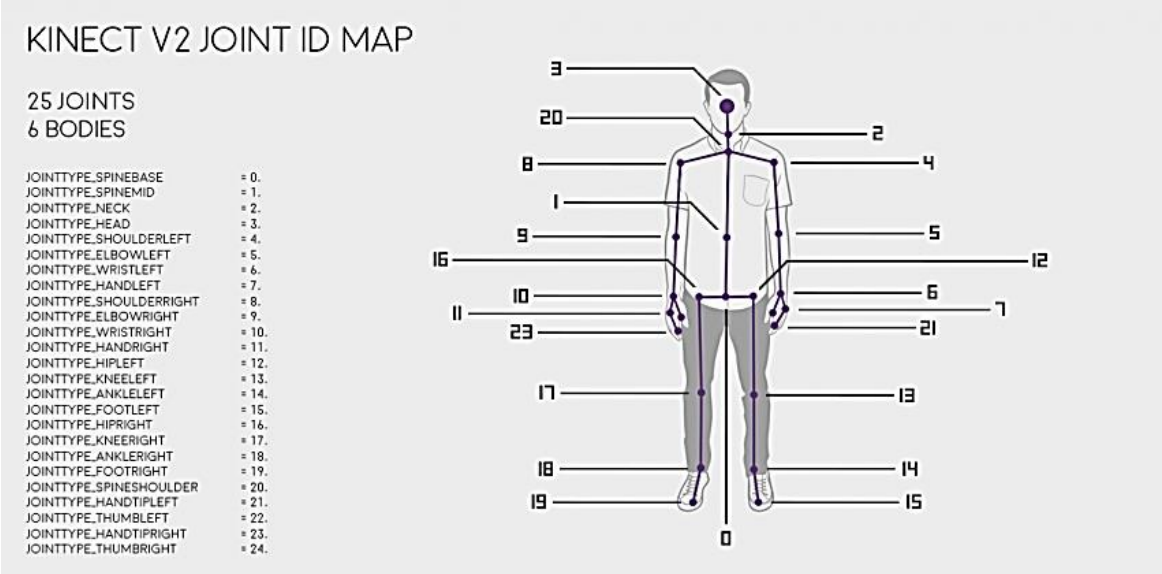


Şekil 16. Xbox 360 için a) tüm takip edilen eklemler b) cihazda olan aygıtlar c) koordinat eksenleri [47].

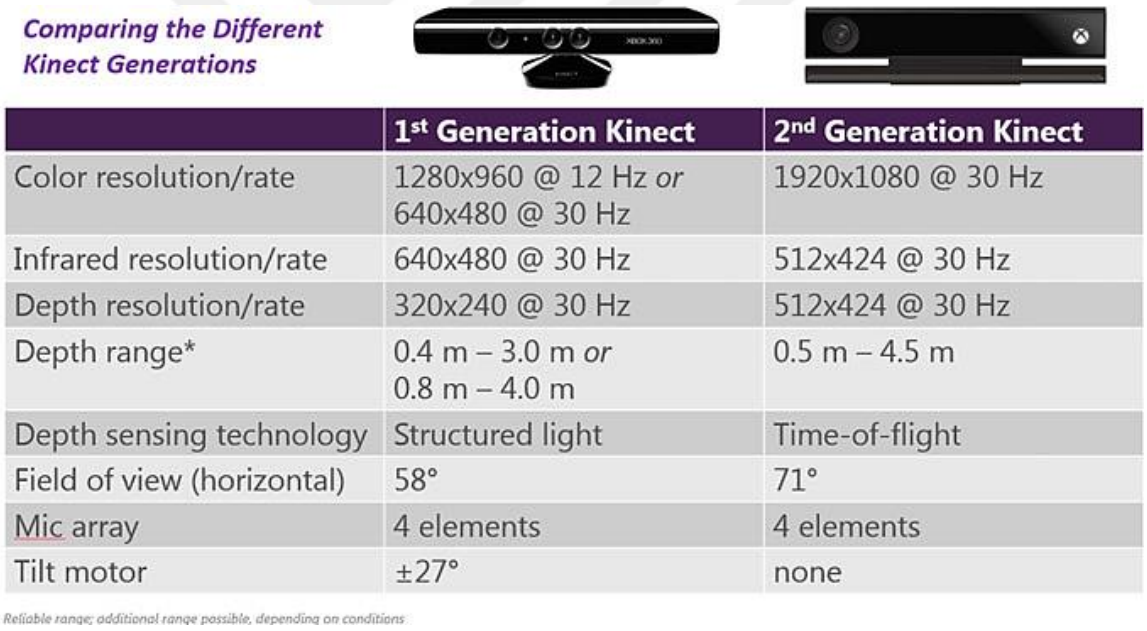
Kinect de kullanılan algoritmaların özelliklerine bağlı olarak, iskelet takibi kullanıcı Kinect'e baktığı zaman optimize edilmiştir. Kinect'e görünmeyen kullanıcı bölümleri ile yan pozlar, iskelet takibini zorlaştırır. Clark vd. [47] Microsoft SDK'dan elde edilen iskeletin performansı, gürültü, doğruluk ve çözünürlüğü açısından araştırılmıştır.

Birinci nesil Kinect, yapılandırılmış bir ışık (structured light) görme sistemi, ikinci nesil Kinect ise uçuş zamanı (time-of-flight) sistemidir. İkinci nesil Kinect birincisiyle kıyasladığında Şekil 17'de gösterdiği gibi bir insan üzerinde daha fazla eklemi güvenilir bir şekilde takip edebilir, ayrıca karşısında çok kişi olduğu zaman, birinci nesil iki kişinin tüm eklemlerini takip ederken yeni nesil altısının tüm eklemlerin takip edebilir.

Kinect nesillerinde olan özellikler Şekil 18'de karşılaştırarak verilmiş. Wasenmüller vd. [48] iki nesil den üretilen derinlik görüntüleri üzerinde titiz bir değerlendirme ve karşılaştırma yapmıştır.



Şekil 17. İkinci nesil Kinect ile takip edilen eklemler.



Şekil 18. Birinci (Xbox 360) ve ikinci (Xbox One) nesil Kinect'lerin özellikleri [45].

2.1.2. Derin Öğrenme Teknikleriyle Çıkartılan 2B/3B Poz

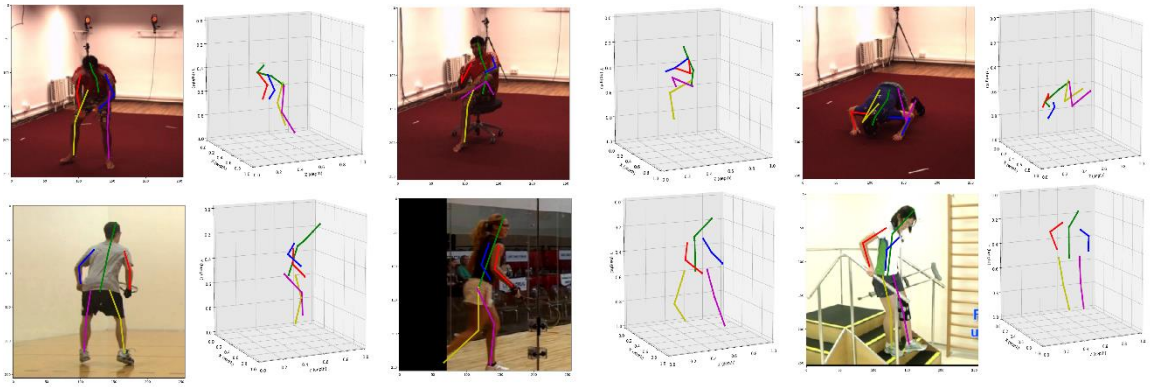
Son zamanlarda derin öğrenme yöntemlerin güçlenmesi ile birlikte poz çıkarma sadece derinlik algılayıcılara sınırlı kalmayarak RGB görüntüler üzerinden bile güvenilir ve hatasız şekilde elde edilir [49-53]. Literatürde, RGB video veya görüntüsü üzerinden poz tahmini gerçekleştiren yaklaşımlar genellikle Algılama-Tabanlı (Detection based) ve Regresyon-

Tabanlı (Regression based) adlandırılan iki ana sınıfta gruplanabilir [51]. Diğer bir sınıflandırma çıkarılan pozların 2B veya 3B üzerinden yapılabilir. Algılama-Tabanlı yaklaşımlar poz keşfini, ısı haritası (heat map) tahmin problemi olarak ele alırlar. Bir ısı haritasındaki her piksel, ilgili bir eklemin tespit puanını temsil eder. Ancak, algılama tabanlı yaklaşımlar doğrudan eklem koordinatlarını vermez. Pozun (x, y) koordinatlarını çıkarmak için işlem sonrası genellikle argmax fonksiyonu kullanılır. Diğer taraftan, regresyon temelli yaklaşımlar, girişi doğrudan istenen çıktıya (eklem koordinatlar olabilir) eşleyen, doğrusal olmayan bir işlev kullanır. Regresyon yöntemlerinin sınırlanması, regresyon fonksiyonunun sıklıkla sub-optimal olmasıdır.

Guilhem vd. [49] renkli video görüntü karelerinde bulan pozları temsil için P-CNN adlandırılan öznitelikler tanımlamış ve bu öznitelikler üzerinden vücut eklemlerinin (kafa, dirsek, bilek ve…) iki boyutlu uzayda olan koordinatlarını belirleyerek 2B pozların elde etmesi için önerilen sistemi FLIC veri küme üzerinde eğitmiştir. Daha sonra vücut parçalarının koordinatlarını elde ettikten sonra bunlardan yararlanarak hem RGB görüntüleri hem de işlem karmaşıklığı pahalı olan optik akış haritalarını birleştirmek için iki akışlı ağlar kullanmışlardır. Bu yaklaşımla çıkartılan 2B pozlar Şekil 19'da gösterilmektedir.

Şekil 19. JHMDB veri setinde (soldaki iki resim) ve MPII pişirme faaliyetlerinde (sağdaki iki resim) başarılı örnekler ve hata durumları [49].

Luvizon vd. [51] 2B ve 3B poz tahmini ve eylem tanıma işlemin birleştirilmiş şekilde gerçekleştiren bir yaklaşım önermişler. Poz tahmini için kullanılan yöntem regresyon temelli yaklaşımlar sınıfına giriyor. Çalışmada 2B ısı haritalarını hacimsel temsillere genişleterek 2B poz regresyonunu 3B senaryolara genişletirmiş. Bu yaklaşımı kullanarak tahmin edilen 3B pozlar Şekil 20'de gösterilmiştir.



Şekil 20. Human3.6M (üst sıra) ve MPII (alt sıra) veri tabanlarından tahmin edilen 3B pozlar [51].

2.2. Literatürde Bulunan İskelet Tabanlı Yaklaşımlar

Bu bölümde, hareket tanıma için 3B iskelet eklemlerin verilerini kullanan poz temelli yöntemler kısaca açıklandı. Ana odak noktamızın tek bir kişi tarafından gerçekleşen günlük yaşam eylemleri olduğuna dikkat edilmelidir (etkileşimli olmayan).

3B iskelet veriler eklemler arasındaki ilişkileri ve insan pozlarının tam konfigürasyonlarını temsil eder. Bu bilgi hareket yakalama sistemleri (MoCap), stereo ve derinlik algılayıcı gibi çok farklı modalitelerden çıkarılabilir [34, 43].

İnsan eylemi tanıma konusundaki öncü bir araştırmacı olan Johansson [39], eklem konum dizisinin kullanılabilirliğinin insan eylemlerini tanımak için yeterli olduğunu göstermiştir. Yao vd. [54] kapalı ortam eylem tanıma senaryolarında, poz temelli özniteliklerin kullanılması, görünüme dayalı özniteliklere kıyasla daha iyi bir tanıma performansıyla sonuçlandığını göstermiştir.

Genel olarak tüm poz tabanlı eylem tanıma yaklaşımları iki önemli adımdan oluşur; birincisi, her görüntü karesindeki insan pozları, ham 3B iskelet verilerinden çıkarılan öznitelikler ile tanımlanır ve ikinci adımda sınıflandırma veya mantıksal karar verme için kullanılacak nihai öznitelik vektörü bütün eylem dizisi üzerinden hesaplanır. Han vd. [34] sırasıyla bu iki adımı "bilgi modalitesi" (information modality) ve "temsil kodlaması" (representation encoding) olarak adlandırılmıştır.

2.2.1. Bilgi Modalitesi

Bilgi Modalitesi (information modality) taksonomiye göre, 3B iskelet üzerinden insan tarif etmek için ham 3B iskelet verilerinden üretilen çeşitli öznitelikler, eklem değişimi (joint displacement), oryantasyonu (orientation) [55], ham eklem pozisyonları (raw joint positions) [28, 30, 56, 57] ve çoklu model (multi-modal) [11, 58-61] baz alınarak dört gruba ayrılmıştır.

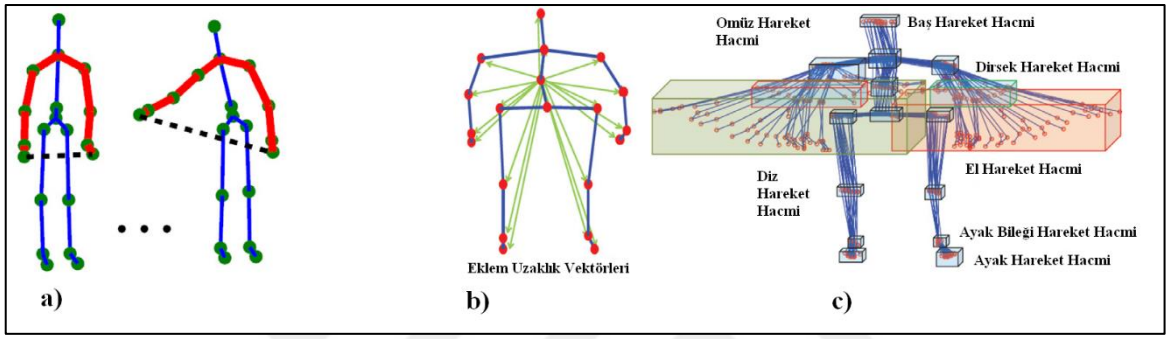
2.2.1.1. Yer Değişim Tabanlı Temsil

Yer değişim tabanlı temsil (displacement-based representation), iskelet eklemlerin değişiminden çıkarılan özellikler, basit yapı ve kolay uygulama nedeniyle birçok iskelet tabanlı temsilde yaygın olarak uygulanır. Bunlar iskelet eklemlerinin yer değişim bilgisini kullanırlar, aynı görüntü karesinde bulunan iskeletin farklı eklemler arasında (uzay) veya aynı eklemin farklı zaman aralıklarında (farklı görüntü karelerinde) olan yer değişim (zamansal) vektörü olabilir. Birinci türe uzamsal denilir ve bir görüntü karesinde olan farklı eklemlerin 3B koordinatlarından hesaplanan yer değişimidir.

İnsanın 3B iskelet üzerinden seçilen her iki eklem konumların arasından hesaplan yer değişim vektörleri, insan temsili için en yaygın olarak incelenen yer değişim özniteliğidir [62-64]. Bir görüntü karesinden elde edilen iskelet modelinde, 3B uzayda her bir eklem $p = (x, y, z)$ olursa, i'inci eklem ve j'inci eklem konumlarının arasındaki fark vektörü $p_{ij} = p_i - p_j$, $i \neq j$ tarafından hesaplanır. Eklem konumları, p genellikle normalleştirilir, bu nedenle öznitelik vücudun bulunduğu pozisyonun, başlangıç vücut oryantasyonuna ve vücut büyüklüğüne göre değişmezdir. Yapılan çalışmalarda normalleşmiş değerler elde etmek için orijinal formülün biraz değişik hali $\| p_i - p_j \| / \sum_{i \neq j} \| p_i - p_j \|$, $i \neq j$ kullanılmıştır (Şekil 21-a).

İskelet tabanlı temsil inşası için bir başka eklem yer değişim öznitelik grubu, aynı görüntü karesinde referans eklemi üzerinden çıkarılan uzaklığa dayanmaktadır (Şekil 21-b). Bu öznitelikler için yer değişim vektörleri, tüm eklemlerin koordinat farkının genellikle elle seçilen bir referans eklemine göre hesaplanmasıyla elde edilir. Dünya koordinat sisteminde verilen bir eklemin konumu (x, y, z) ve verilen bir referans ekleminin konumu (x_c, y_c, z_c) düşünüldüğünde uzamsal eklem uzaklığı $(\Delta x, \Delta y, \Delta z) = (x, y, z) - (x_c, y_c, z_c)$ olarak hesaplanır [65].

Yaygın olarak kullanılan bir zamansal yer değişim özniteliği, eklem koordinatlarının farklı zaman adımlarında karşılaştırılmasıyla gerçekleştirilir. Yang vd. [66] çalışmalarında EigenJoints olarak adlandırılan ve eklemlerin pozisyon farkına dayanan yeni bir özellik ortaya koydular ve statik duruş, hareket ve ofset öznitelikleri de dâhil olmak üzere üç öznitelik kategorisini birleştirmişler. Özellikle, mevcut görüntü karesinde bulunan poz eklemlerin yer değişimi bir önceki karede buluna pozun eklemleri ile veya başlangıç karesinde bulunan poza göre hesaplanır.



Şekil 21. a) İki eklem yer değişim b) bağıntılı eklem yer değişim c) hareket hacim öznitelikleri [34].

Eklem hareket hacmi, insan temsili için bir başka öznitelik yapım yaklaşımıdır. Ayrıca, öznitelik çıkarma için eklem yer değişim bilgileri kullanır ve özellikle bir eklem büyük bir hareket sergilediğinde kullanılır [65]. Belirli bir eklem için, tam eklem hareketi sırasında uç nokta pozisyonlar $x, y\ ve\ z$ eksenleri boyunca hesaplanır. Her bir boyuttaki her bir eklemin maksimum hareket aralığı daha sonra $L_a = \max(a_j) - \min(a_i)$ ile hesaplanır, burada $a = x, y\ ve\ z$ ve eklem hacmi $V_j = L_x L_y L_z$ olarak Şekil 21-c'de gösterilmiştir.

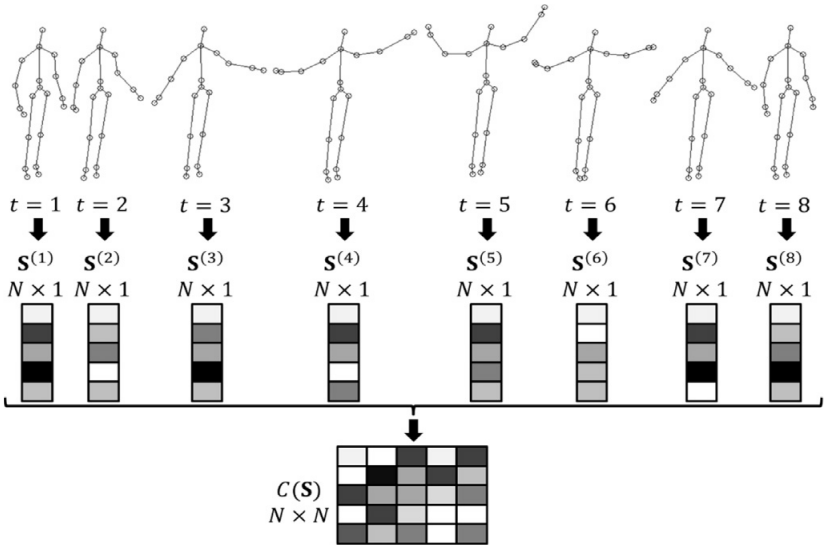
2.2.1.2.Oryantasyon Tabanlı Temsil

İnsan temsili sağlamak için yaygın olarak kullanılan bir başka bilgi mod de eklem oryantasyonlarına dayanmaktadır. Genel olarak oryantasyon tabanlı özellikler, insan konumuna, vücut büyüklüğüne ve kameranın görüş açısından bağımsızdır. Yer değiştirme özniteliklerin benzer, uzay ve zamansal olarak üretilebilirler. İkili eklemlerin uzamsal oryantasyona dayanan yaklaşımlar, aynı görüntü karesinde elde edilen insan iskelet ekleminden seçilen iki eklemin yer değişim vektörlerinin oryantasyonun hesaplar. Popüler

oryantasyona dayalı bir insan temsili, her bir eklemin 3 boyutlu uzayda insanı koordinat merkez nokta kabul ederek oryantasyonun hesaplar. Zamansal eklem oryantasyonlarına dayanan insan temsilleri, genellikle aynı eklemin bir görüntü kareler dizisi boyunca oryantasyonu arasındaki farkı hesaplar.

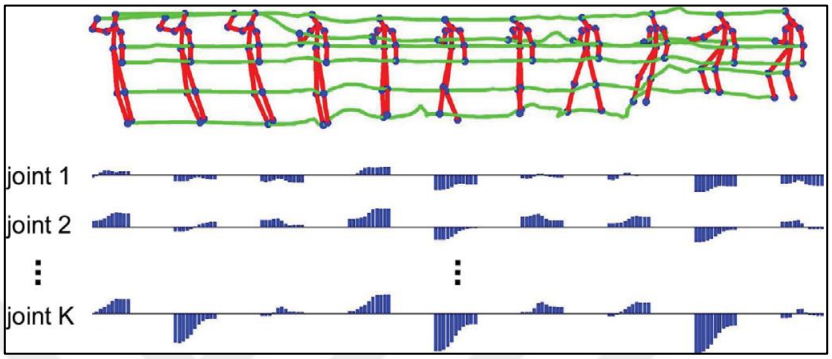
2.2.1.3.Ham Eklem Konum Tabanlı Temsil

Eklem yer değişimleri ve oryantasyonlarının yanı sıra, sensörlerden doğrudan elde edilen ham eklem pozisyonları da, uzay-zaman 3D insan temsili oluşturmak için birçok yöntem tarafından kullanılmaktadır. Bir yaklaşım kategorisi aynı görüntü karesinden elde edilen eklem pozisyonlarını birleştirerek, bir sütun vektörüne düzleştirir. Bir dizi iskelet kareleri verildiğinde, diziyi saf olarak kodlamak için bir matris oluşturulabilir. Bu matrisin her sütunu belirli bir zaman noktasında elde edilen düzleştirilmiş eklem koordinatlarını içerir. Şekil 22'de gösterildiği gibi bu doğrultuda Hussein vd. [67] öznitelik olarak 3B eklemlerin istatistiksel kovaryansını (Cov3DJ) hesaplamıştır. Verilen K eklemli bir iskeletin her bir eklem konumu $p_k = (x_k, y_k, z_k)\,, k = 1, 2, \dots, K$ varsayılınca, t zamanında bulunan iskeletin kodlaması için $s^{(t)} = \left[x^{(t)}{}_1, \dots, x^{(t)}{}_K, y^{(t)}{}_1, \dots, y^{(t)}{}_K, z^{(t)}{}_1, \dots, z^{(t)}{}_K\right]^T$ özellik vektörü oluşturulmuş. İskelet dizisinden elde edilen bu öznitelik dizisini kovaryans işlemine tabi tutarak Cov3DJ özniteliği elde edilmiştir.



Şekil 22. Cov3DJ tanımlayıcı dayalı 3B insan modeli [67].

Bir diğer temsil sağlama teknik grubu, ham eklem pozisyon bilgilerini kullanarak bir yörünge oluşturur ve daha sonra bu yörünge üzerinden öznitelikleri çıkarır. Bu çoğunlukla yörüngeye dayalı temsil olarak adlandırılır. Örneğin Wei vd. [68] bir dizi 3B insan iskelet eklemini kullanarak eklem yörüngeleri oluşturmuş ve Şekil 23'de görüldüğü gibi her zamansal eklem dizisini özelliklere kodlamak için dalgacıklar uygulamıştır.



Şekil 23. Dalgacık öznitelikleri kullanan yörünge tabanlı modeli [68].

Görüntüler üzerinden derin öğrenme teknikleriyle öznitelik çıkartan yöntemlerde genellikle ham pikseller girdi olarak kullanılır. Benzer şekilde derin öğrenme yöntemleriyle oluşturulan iskelet tabanlı insan temsilleri genellikle ham eklem pozisyon bilgilerine dayanır.

2.2.1.4.Çoklu Modlu Temsil

Çoklu bilgi modaliteleri mevcut olduğu için, bir insan temsilinin tanımlayıcı gücünü geliştirmek için sezgisel bir yol, çoklu bilgi kaynaklarını entegre etmek ve insanları 3B uzayda kodlamak için çok modlu bir temsil oluşturmaktır. Örneğin, uzamsal eklem uzaklığı ve oryantasyon insan temsili oluşturmak için birlikte entegre edilebilir.

Uzamsal ve zamansal bilgileri entegre edebilen ve 3 boyutlu alanda insan hareketlerini temsil edebilen çok modlu uzay-zaman insan gösterimleri de aktif olarak incelenmiştir. Yu vd. [69], ikili eklem yer değişimları, eklem koordinatları ve eklem yerlerinin zamansal değişimlerini içeren uzay-zamansal bir temsili oluşturmak için üç tip özelliği entegre etmiştir. Zanfir vd. [60] insan eklem hareketlerinin hız ve hız değiştirme (acceleration) davranışlarını kuadratik dereceli fonksiyonlarla doğru bir şekilde tanımlanabileceği

beklentisine dayanarak, ham 3B eklem pozisyonlarını ve eklem yörüngelerinin birinci ve ikinci türevlerini birleştirerek hareketli poz (moving pose) olarak adlandırılan bir özniteliğini tanıttı.

2.2.2. Temsil Kodlama

Öznitelik kodlamak (Feature encoding), temsil oluşturulmasında gerekli ve önemli bir bileşendir, amacı tüm çıkarılan öznitelikleri nihai bir öznitelik vektöründe entegre edip sınıflandırıcıya veya mantık yürüten sistemlerine girdi olarak kullanılabilmektir. 3D iskelet tabanlı temsilin oluşturulması senaryosunda, kodlama yöntemleri genel olarak üç sınıfa ayrılabilir:

2.2.2.1.Birleştirme Tabanlı Yaklaşımlar

Birleştirme tabanlı yaklaşımlar (concatenation-based approach) yaklaşımlar, elde edilen özniteliklerin birleştirilmesi ile tek boyutlu nihai öznitelik vektörü oluşturur ve diğer kodlamalarla karşılaştırdığında aralarında en basit ve popüler olanı sayılmaktadır. Pek çok yöntem, 3B insan eklemlerinin yer değişimleri ve oryantasyonları gibi çıkarılmış iskelet tabanlı öznitelikleri doğrudan kullanarak ve bunları 1B öznitelik vektöründe birleştirerek insan temsili oluşturmuşlardır [64, 70].

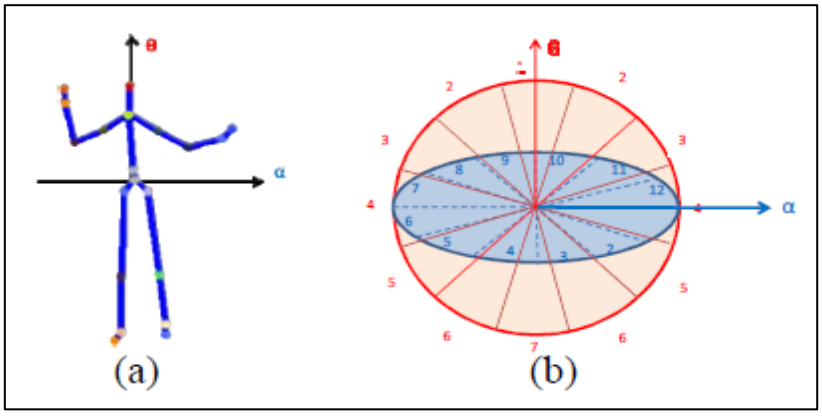
Örneğin, Fothergill vd. [70], her bir görüntü karesi için 35 iskelet eklem açısını, 35 eklem açısal hızını ve 60 eklem hızını bir 130 boyutlu vektöre birleştirerek öznitelik vektörünü kodlamıştır. Daha sonra, bir dizi görüntü karesinden elde edilen özellik vektörleri, büyük bir nihai özellik vektöründe birleştirilir ve mantıksal yorum yapması için bir sınıflandırıcıya girdi olarak hazırlanır.

Bu kodlama yönteminin avantajlarından olan basitlik ve düşük hesaplama maliyeti nedeniyle gerçek zamanlı uygulamalarda yaygın şekilde kullanılmıştır. Bunun yanı sıra oluşturulan öznitelik vektörünün çok uzun olması ve dolayısıyla uygulama açısından sınıflandırıcı için yüksek boyutlu uzayı idare etmesinin zor olması dezavantajlarından sayılabilir.

2.2.2.2.İstatistiksel Kodlama

İstatistiksel kodlama (statistics-based encoding) herhangi bir öznitelik kuantalama işlemi kullanmadan, çıkarılan öznitelik vektörlerinin entegresi için istatistiksel analitik uygulayan yaygın ve etkili bir yöntemdir [67]. Bu kodlama metodolojisi, basit istatistikler kullanarak öznitelikleri işler ve düzenler. Örneğin, Şekil 22'te gösterildiği gibi Cov3DJ temsili [67], iskelet karelerinin bir dizisi boyunca toplanan bir dizi 3B eklem konum vektörünün kovaryansını hesaplar. Bu çalışmada elde edilen kovaryans matrisi simetrik olduğundan dolayı nihai özniteliği oluşturmak için sadece üst üçgen değerleri kullanılır. Bu istatistik tabanlı kodlama yaklaşımının bir avantajı, nihai öznitelik vektörünün boyutunun görüntü karelerinin sayısından bağımsız olmasıdır. En çok kullanılan istatistik tabanlı kodlama metodolojisi, çıkarılmış iskelet tabanlı özelliklerin dağılımını tahmin etmek için 1B histogram kullanan histogram kodlamasıdır.

Örneğin, Xia vd. [55], 3B uzayı değiştirilmiş küresel koordinat sistemi (Şekil 24-a) kullanarak birçok dilime ayırmış (Şekil 24-b) ve her bir dilime düşen eklemlerin sayısını sayarak bir boyutlu histogram oluşturup, ve bunu3B eklem pozisyonlar histogramı (HOJ3D) olarak tanımlamıştır.



Şekil 24. a) HOJ3D için referans koordinatları b) eklem konum dilimi için değiştirilmiş küresel koordinat sistemi [55].

Benzer histogram kodlama yöntemlerini kullanan çok sayıda iskelet tabanlı insan temsili yöntemleri de tanıtılmıştır. Eklem Pozisyon Farkı Histogram (Histogram of Joint Position Differences (HJPD)) [65], Yönlendirilmiş Hız Vektörlerinin Histogramı (Histogram of Oriented Velocity Vectors (HOVV))[71] ve Yönlendirilmiş Deplasmanların

Histogramı (Histogram of Oriented Displacements (HOD))[72] vb. Multimodal iskelet tabanlı öznitelikler kullanıldığında, birden çok histogramı nihai bir öznitelik vektörüne entegre etmek için genellikle birleştirme tabanlı kodlama kullanılır [73].

Öznitelik öğelerinde düzen(sıra) eksikliği ve zamansal ilişki yokluğu bu yöntemlerin en önemli dezavantajları olarak düşünülebilir. Kuantalama kullanılmadığı için bu yöntemler gürültüye karşı daha dayanaklıdır.

2.2.2.3.Kelime Çantası Kodlama

Kelime çantası kodlama (bag-of-words encoding) birleştirme ve istatistik tabanlı kodlama yöntemlerinden farklı olarak, kelimeler çantası kodlama, tüm kodları içeren eğitilmiş bir kod çizelgesi(codebook) (veya sözlük(dictionary)) kullanarak her yüksek boyutlu özellik vektörünü olası tek bir kod (veya sözcük) haline dönüştürmek için bir kodlama operatörü uygular. Bu prosedüre ayrıca öznitelik kuantalama da denir. Kelimeler çantası kodlama, çok sayıda iskelet tabanlı insan temsili tarafından yaygın olarak kullanılmaktadır [35, 57, 60, 74]. Sözlük öğrenimi açısından, kodlama yöntemleri genellikle iki ana kategoriye ayrılır; kümeleme ve seyrek kodlama (sparse coding) tabanlı yöntemler [34].

Popüler gözetimli olmayan öğrenme yöntemlerinden, K-means algoritması, yaygın olarak sözlük oluşturmak için kullanılır. Wang vd. [75] insan eklemlerini beş vücut bölümüne ayırmış ve eğitim verilerini kümelemek için k-means algoritmasını kullanmıştır. Küme merkezlerinin indisleri bir sözlük oluşturmak için kod olarak kullanılır. Test sırasında, vücut parçası pozları eğitilmiş sözlük kullanılarak kuantalanmıştır. Kapsoura vd. (2014) iskelet tabanlı öznitelikleri olan eklem oryantasyon ve çoklu zamansal ölçeklerde oryantasyon farklılıkları üzerinde k-means kümeleme yöntemini kullanmış ve temsili kalıpları seçerek bir sözlük oluşturmuşlardır.

2.2.3. Yapı ve Topolojik Dönüşüm

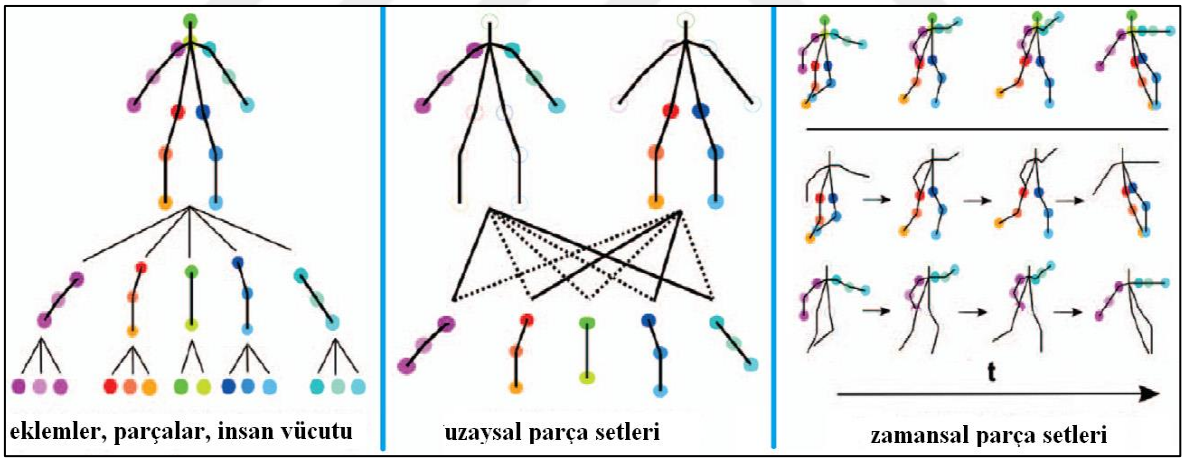
Çoğu iskelet temelli 3B insan temsili, 3B öklid (Euclidean) uzayındaki iskelet verisinden çıkarılan saf düşük düzeyli özniteliklere dayanırken, bazı çalışmalar orta düzey öznitelikler üzerinde çalışmış veya diğer topolojik uzaylara geçiş özelliklerine sahiptir. Han

vd. [34] literatürde bulunan çalışmaları yapı ve geçiş perspektifinden diğer bir deyişle temsil için kullandığı özniteliklerin çıkarma tarzına (uzayı) göre üç ana sınıfa ayırmışlar;

2.2.3.1.Vücut Parça Modellerle Temsil

Vücut parçası modellerine (Representations based on body part models) dayanan orta düzey özellikler, iskelet tabanlı insan temsillerini oluşturmak için de kullanılır. Bu orta düzey özellikler insan vücudunun fiziksel yapısını kısmen dikkate almasından dolayı daha güçlü ve ayrıt edici insan poz temsilin sağlanmasına neden olurlar [73, 76].

Wang vd. [35] kinematik bir insan vücudu modelini, sol / sağ kollar / bacaklar ve gövde dahil beş parçaya ayrıştırdı. Daha sonra, bir uzay-zamansal insan temsili elde etmesi için veri madenciliği tekniğini kullanarak bir görüntü karesinde vücut parçalarının uzamsal konfigürasyonlarını (uzamsal-parça setleri ile) ve aynı zamanda bir dizi kare boyunca vücut parçası hareketlerini (zamansal-parça setleri ile) yakalamışlardır.



Şekil 25. Vücut parçalarından çıkarılmış orta düzeyli özniteliklerle insan modeli [35].

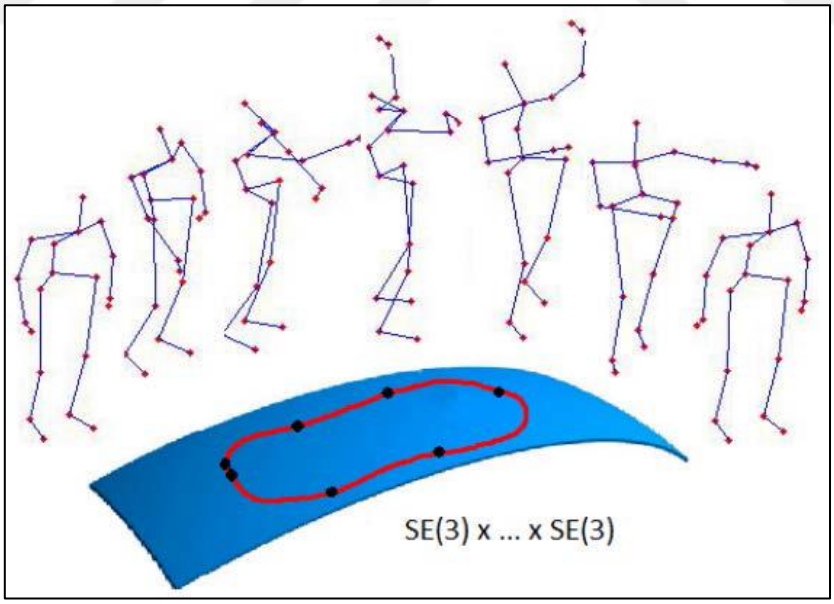
Nie vd. [77], insanları pozlar, uzay-zamansal parçalar ve parçalar da dâhil olmak üzere üç seviyede temsil etmek için uzay-zamansal bir And-Or graf modelini uygulamıştır. Du vd. [29] bir vücut parçası modeli oluşturmak ve vücut parçalarının korelasyonunu araştırmak için derin bir sinir ağını tanıtmışlardır. Vücut kinematiğine veya insan anatomisine dayanarak iskelet tabanlı temsili sağlaması için vücuttan esinlenmiş (Bio-inspired) adlandırılan orta düzey özellikler vücut parçası yöntemleriyle elde edilmiştir [78].

2.2.3.2.Manifold Tabanlı Temsil

Literatürdeki bir dizi yöntem, 3 boyutlu Öklid (Euclidean) uzayındaki iskelet verilerini başka bir topolojik uzaya (yani, manifold) aktarmıştır, amaç iskelet yörüngelerini yeni uzayda eğriler olarak işlemektir.

Vemulapalli vd. [58] $SE(3) \times \ldots \times SE(3)$ Lie grubundan oluşmuş bir iskelet temsili tanımladı ve bu 3B vücut parça hareketlerinin üye olduğu uzayda gözlemlerden üretilmiş manifold eğrisidir. Bu temsili kullanarak, eklem yörüngeleri Şekil 26'da gösterilen Lie grubunda eğriler olarak modellenebilir. Bu manifold tabanlı gösterim, 3B alanda döndürme ve ötelemeleri kullanarak eklemler arasındaki 3B geometrik ilişkileri modelleyebilir. Lie grubundaki eğrileri analiz etmek kolay olmadığına göre yaklaşım, Lie grubundan eğrileri bir vektör uzay olan Lie cebirine eşleştirir.

Amor vd. [56], insan iskelet şekillerinin, Kendall'in şekil manifoldları üzerindeki yörüngeleri olarak modellemeyi önermişler ve parametrelere değişmez ölçü kullanmışlardır.



Şekil 26. Lie Gruplar uzayında iskelet dizisinden oluşan eğri [58].

2.2.4. Eylem Tanıma Yaklaşımlarında Kullanılan Sınıflandırma Türleri

Önceki bölümlerde tanıtıldığı gibi, belirli ilgi noktalarından veya eylemin bütün şeklinden çeşitli öznitelikler çıkarılabilir. Bu nedenle, bir eylem dizisini temsil etmenin birkaç yolu vardır ve bunun uygun sınıflandırma yöntemlerinin seçimi üzerinde etkisi vardır [79]. Eylem tanımada yaygın olarak kullanılan sınıflandırma yöntemleri iki ana çeşitte bölünmüştür.

2.2.4.1.Değişken Boyutlu Öznitelik Vektör

Hareket içeren tüm görüntü karelerinden çıkarılan öznitelik vektörleri bir matriste yığılabilirler. Bir eylemi temsil etmek için bu matristeki tüm satır vektörleri birleştirilerek zaman serisi şeklinde bir vektör olarak ifade edilebilir.

Genel olarak, eylemler farklı sayıda görüntü karelerinden oluşur ve bu yüzden bu zaman serisi şeklindeki diziler farklı uzunlukta olabilirler. Bu tür eylem temsilleri için eylem tanıma sistemlerinde yaygın olarak kullanılan sınıflandırıcılar, Saklı Markov Modelleri (HMMs) [80-83], en uzun ortak alt dizi (Longest Common Subsequence) (LCSS) [84], oyun grafları (action graph) [85, 86], RNN[28, 29], LSTM[30-32] ve dinamik zaman bükmesinin (DTW) [58, 87-89] çeşitli varyasyonlarıdır. Ön işlem olarak, doğrudan çıkarılan özellikler genellikle anahtar kelimeler halinde kümelenir ve ardından kuantalanır. Kümeleme için yaygın olarak kullanılan yöntemler arasında K-means [36, 55, 78] ve Gauss Karışım Modelleri (Gaussian Mixture Model) [90] bulunmaktadır. Kümelerin merkezleri anahtar kelimeler veya kod sözcükleri olarak kullanılır. Bu nedenle, her bir eylem, değişen uzunluktaki bir anahtar kelime dizisi olarak temsil edilir.

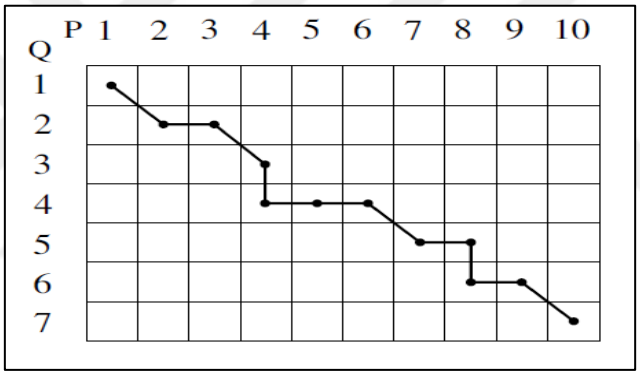
Dinamik Zaman Bükmesi (Dynamic Time Warping, DTW): Dinamik zaman bükmesi (DTW), sinyallerin benzerliğini ölçmek için en popüler algoritmadır [91, 92]. Örneğin konuşma tanımada ve iskelet verilerini kullanan jest tanımada yaygın olarak kullanılmıştır [93-95]. Eğer $P = p_1, p_2, \dots, p_N$ ve $Q = q_1, q_2, \dots, q_M$ iki zaman serisi olarak tanımlanırsa (burada p_n ve q_m sırasıyla n'inci ve m'inci zaman indeksindeki sinyal değerleridir). Bir yer değişim fonksiyonu, her bir p_n ve q_m elemanın arasındaki benzerliğin hesaplanması için kullanılır. Yaygın olarak kullanılan fonksiyonlar arasında öklid ve kosinüs mesafeleri bulunur.

Daha sonra dizilerdeki her bir eleman çifti arasındaki benzerlik (farklılık), mesafe hesaplanmasıyla bir maliyet matrisi oluşturulabilir. DTW algoritması, P ve Q uyuşmuş eleman çiftleri arasındaki mesafelerin toplamını minimize ederek diğer bir deyişle minimum

toplam maliyeti bulur. Minimum maliyetli uyuşma, bükme yolu (warping path) olarak belirtilir, bu yol dizilerdeki eşleşen (benzetilmiş) elemanların indislerini gösterir. Temel DTW için, bükme yolu iki koşulu yerine getirmek zorundadır:

- P ve Q'nun ilk ve son elemanları birbiriyle uyuşmak zorundadır. Yani $p_1 \leftrightarrow q_1$ ve $p_N \leftrightarrow q_M$
- Eğer $p_N \leftrightarrow q_M$ olursa, o zaman $p_N \leftrightarrow q_{M+1}$ ya $p_{N+1} \leftrightarrow q_M$ ya $p_{N+1} \leftrightarrow q_{M+1}$ olmaktadır. $\leftrightarrow$ sembolü, sol sinyal elemanları sağ da olan sinyal elemanlar ile uyuşma ya da eşleşme anlamına gelmektedir.

Hesaplama verimliliği ve belirtilen kısıtlamaları göz önüne alarak bükme yolunun bulunması için, genellikle dinamik programlama yöntemi kullanılır. Şekil 27 ile bu işlemi gösteren bir örnek verilmiştir [96].



Şekil 27. Dinamik programla kullanarak DW yöntemde eşleşen en iyi yolun bulması [96].

2.2.4.2.Sabit Boyutlu Öznitelik Vektör

Zaman serileri sinyalleri olarak temsil edilen jestlerin aksine, bazı sistemlerde, her bir hareket, bir bütün olarak, sabit bir boyuta sahip tek bir özellik vektörü ile temsil edilir. Bu sabit boyutlu öznitelikleri elde etmek için bir strateji, tüm hareketleri sabit sayıda görüntü kareye örneklemektir. Bu görüntü karelerden çıkarılan özellikler daha sonra bir özellik vektörü oluşturmak için birleştirilebilir ve genellikle yüksek boyutlu hale gelir. Başka bir strateji, hareketi bir histogram olarak temsil etmektir. Öznitelikler tüm eylemler boyunca çıkarılır ve ardından kod sözcüklerine kümelenir. Eylem daha sonra kod kelimelerinin bir histogramı olarak temsil edilir. Hareketler, sabit bir boyuta sahip özellikler ile temsil

edildiğinde, KNN [97], naive Bayes [98] ve SVM [99] gibi birçok örüntü tanıma yöntemi uygulanabilir.

2.3. Fisher Vektör Kodlama

Fisher çekirdeği (kernel) [100] kavramı yeni olmamasına rağmen üzün süre geçtikten sonra ilk olarak görsel kelimeleri kodlamak için Fisher vektör (FV) olarak Perronnin ve Dance [101] tarafından kullanmış ve görüntü temsili gerçekleştirilmiş. Sadece birinci mertebe istatistiklerinden sağlanan BOF'le (öznitelikler çantası) farkı, FV'nin kelime sözlüğün oluşturmak için k-means kümeleme yerine Gaussian karışım modellerine (GMM'ler) ihtiyacı var, böylece hem birinci hem de ikinci mertebe istatistiklerini kodlamasıdır. FV'yi hesaplamak için, ilk olarak, verilen eğitim tanımlayıcıların dağılımına uyan K tane GMM dağılımının parametreleri, $\lambda = \{\omega_k, \mu_k, \Sigma_k\}_{k=1}^{K}$ tespit edilir [102].

$\{x_1, x_2, \dots, x_N\}$, N tane tanımlayıcıyla temsil edilen bir görüntü verilirse, FV, ortalama (u_k) ve sapmaların (v_k) normalize edilmiş kısmi türevlerinin birleşiminden oluşur:

$$u_k = \frac{1}{N\sqrt{\omega_k}} \sum_{i=1}^{N} \gamma_{ki} \left(\frac{x_i - \mu_k}{\Sigma_k}\right), \quad v_k = \frac{1}{N\sqrt{2\omega_k}} \sum_{i=1}^{N} \gamma_{ki} \left[\left(\frac{x_i - \mu_k}{\Sigma_k}\right)^2 - 1\right] \qquad (1)$$

Burada γ_{ki}, her x_i vektörün GMM'de olan k inci bileşen ile ilişkilendiren ardıl olasılığına işaret eder. FV vektörünün nihai boyutu $2DK$ olur, burada D tanımlayıcıların boyutudur.

BoW hem görüntü ve hem video tanımada en yaygın kullanılan yöntemlerinden biridir ve bu alanlarda diğer yöntemlere karşılaştığında çoklu başarılar elde etmişler [103, 104]. Bu yöntemlerde anahtar sözcüklerinden oluşan sözlüğü genelde kümeleme yöntemleri kullanarak sağladıktan sonra girdi bir veri üzerinden elde edilen öznitelik, sözlükte üzerinde sadece bir anahtar sözcüğe dönüştürülüyor. FV yöntemleri BoW yöntemlerinden farklı olarak bu eşleştirmeni sadece bir sözcükle değil belki GMM de elde edilen bileşenlerin farklı katsayıda olan karışımlarıyla gerçekleştiriyor.

Oneata vd. [105] FV kodlama yöntemin insan hareket tanıma için geniş ve kapsamlı değerlendirme yaparak kullanan ilk çalışmalardan birisidir. Bu çalışmada düşük seviyeli özniteliklerin çıkarılması ve daha sonra bunlar üzerinde FV kodlama yöntemin kullanarak eylemi temsil eden tanımlayıcının sağlanması üzerinde odaklanmışlar. Elde edilen

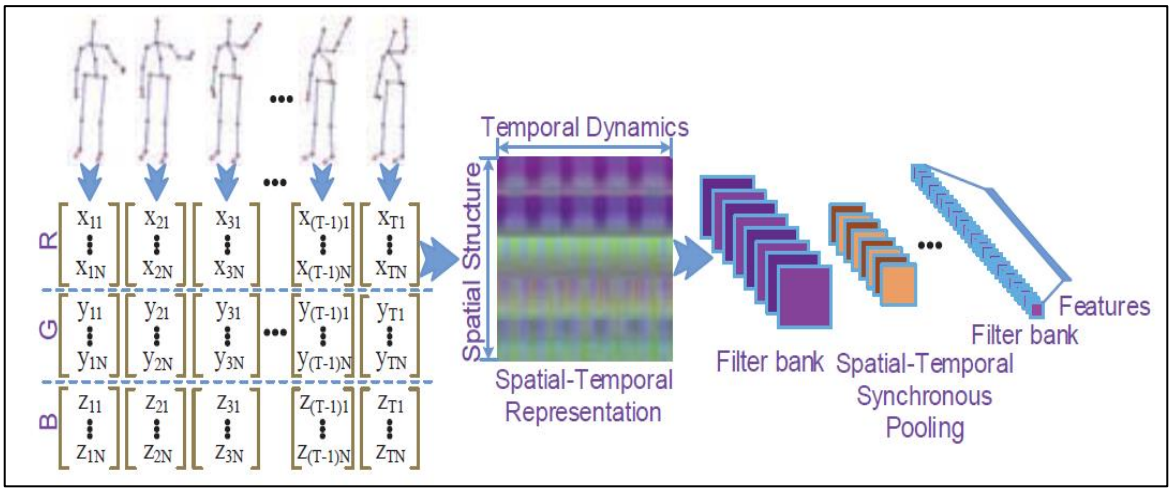
tanımlayıcıların sınıflandırması için linear sınıflandırıcı kullanılmış. Varol ve Salah [102] yaptıkları çalışmada iyileşmiş yoğun yörünge (improved dense trajectory) öznitelikleri ile birlikte orta düzeyli renk öznitelikleri kullanarak, bunlarını entegre etmesi için FV kullanmışlar. Eylemler için elde edilen nihai vektörlerin sınıflandırması ELM gerçekleştirmişler. Önerilen yöntemi THUMOS 2014 benchmark veri seti üzerinde değerlendirme yaparak iyi sonuçlar elde etmişler.

Evangelidis vd. [106] 3B iskelet üzerinden FV kullanarak, görüş açıdan değişmez düşük seviyeli "skeletal quad" adlandırılan öznitelik tanımlamışlar. Elde edilen tanımlayıcıların sınıflandırmasın SVM ile gerçekleştirmişler.

2.4. İskelet Dizisini RGB Görüntüye Dönüştürmek

Günümüzde, Evrişimsel Sinir Ağları (Convolutional Neural Networks)(CNN) gibi görüntü işleme ve derin öğrenme tabanlı sınıflandırma yöntemlerinde kaydedilen gelişmeler nedeniyle, araştırmacılar bu yöntemleri iskelet tabanlı eylem tanıma için kullanmaya teşvik edilmektedir.

Ancak, çözülmesi gereken birçok zorluk halen daha mevcuttur. Bu yöntemler, görüntüleri girdi olarak kabul edecek şekilde tasarlanmıştır ve iskelet dizilerindeki dinamik bilgileri yakalayamazlar. Bu nedenle, bir dizinin uzay-zamansal bilgisini içeren ve bunları iki boyutlu uzaylı bir görüntüye aktaran bir kodlama yöntemine ihtiyaç vardır.

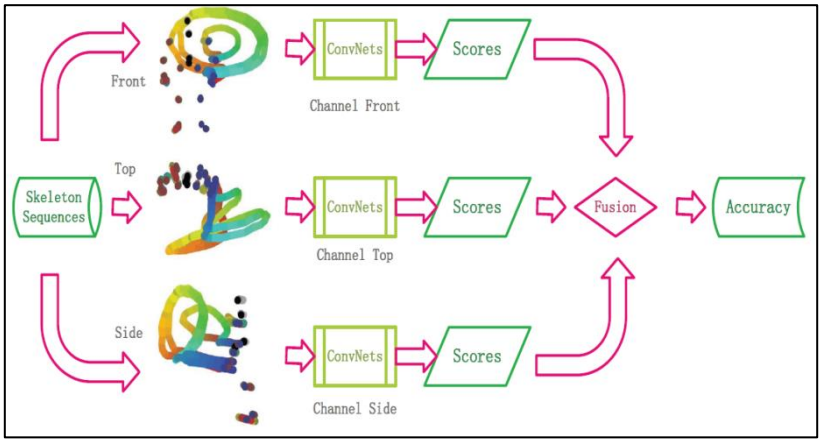


Şekil 28. İskelet dizisin görüntüye dönüştürme ve sonra CNN ile sınıflandırması [107].

Literatürdeki bazı çalışmalar, iskelet poz dizilerinin dinamik bilgileri içeren bir görüntüye dönüştürülmesini önerir ve sonra ağdan (CNN) sentezlenmiş görüntüleri sınıflandırmasını istenilir.

Bu yaklaşımdaki ana adım, iskelet dizilerinin görüntülere dönüştürülmesidir, bu işlemi gerçekleştirmek için dizideki uzay-zamansal bilgiler renk ve doku gibi görüntü özellikler ile tek görüntüye yansıtılır. Du vd. [107], bir iskelet dizisini bir matris olarak temsil ettiler. Bu matristin oluşturması için tüm örnek koordinatları birleştirerek ve vektör gösterimlerini kronolojik sırayla düzenleyerek yapılmıştır (Şekil 28). Matris daha sonra bir görüntüye kuantalanmıştır ve dizilerin değişken uzunluklarından kaynaklanan sorununla baş etmek için normalleştirilmiştir. Nihai görüntü, özellik çıkarma ve tanıma için bir CNN modeline giriş olarak verilmiştir.

Wang ve diğ [108] iskelet dizisinde yer alan uzay-zamansal bilgileri çoklu doku görüntülerine, yani eklem yol haritalarına (Joint Trajectory Maps) (JTM) kodlamayı önermiştirler, bu kodlama eklemlerin yörüngeleri HSV (ton, doygunluk, değer) uzayına eşleştirilmesiyle gerçekleştirilmiştir. Imagenet üzerinde önceden eğitilmiş CNN modelleri, özellikleri çıkarmak ve eylemleri tanımak için JTM'ler üzerinde üzere kullanılmıştır (Şekil 29).

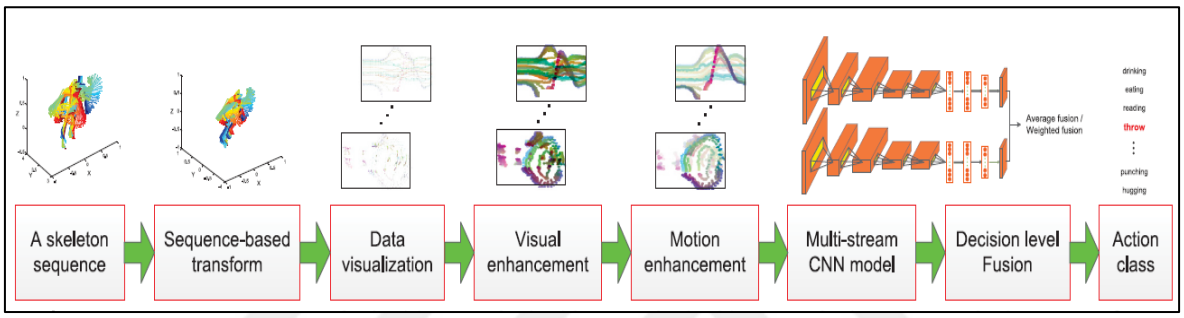


Şekil 29. JTM yöntemiyle CNN kullanarak eylem tanıma [108].

Liu ve diğ [109], bir iskelet dizisini temsil etmek için bir dizi görsel ve hareket arttırılmış renk görüntüsünden oluşan geliştirilmiş iskelet görselleştirme (enhanced skeleton visualization) yöntemini ortaya koymuştur. Bu görüntüler sayesinde eylemlerde olan yerli örüntüler (görsel, uzamsal, hız ve yön) daha iyi bir şekilde temsil edilebilir. Görüş açısı

değişim problemiyle başa çıkmak için dizi-tabanlı görüş açısından bağımsız bir dönüşümü önermişler ve tanıma yürütmek için çoklu akış CNN füzyon yöntemi kullanılmıştır.

Hou vd. [110] iskelet dizilerini doku görüntülerine dönüştüren "Skeleton Optical Spectra"(SOS) adlı yeni bir kodlama yöntemi önerdi. Önerilen yöntemin görüş açısından bağımsız hale gelmesi için iskelet verilerin üç farklı yönden ele alıp ve her yön için farklı doku görüntüler elde edilmiştir. Oluşturulan doku görüntülerinden ayrılabilir özellikleri çıkarmak için bir CNN ağına girdi olarak verilip ve sınıflandırma işlemi CNN ağının ortalama çıktısını kullanarak gerçekleşti.



Şekil 30. Geliştirilmiş iskelet görselleştirme yönteminin iş diyagramı [109].

2.5. Kelime Çantası Yaklaşımları

Han vd. [34] tarafından yürütülen çalışmada, iskelet verilerinden farklı öznitelikler çıkartılarak üzerlerinde birleştirme, istatistiksel ve kelimeler çantası kodlama yöntemlerin uyguladıktan sonra incelediği dört benchmark veri küme üzerinde ki sonuçlar, kelimeler çantası kodlama yönteminin, diğer yöntemlerle kıyasladığında daha iyi bir performans verdiğini gösterdi.

Öznitelikler arasında zamansal bilginin kaybolması kelimeler çantası yöntemlerin büyük bir dezavantajıdır. Bunun üstesinden gelmesi ve kodlama yöntemlerin güvenilirliğini iyileştirmesi için literatürde yapılan çalışmalar bulunmaktadır [35, 60, 74]. Her eylem sınıfında pozların uzaysal / zamansal yapısını çıkarmak için [35, 60]'de veri madenciliği teknikleri kullanılmıştır. Eğitim verileri üzerinde iskelet eklemleri, k-means kümelemesi ile gruplandırıldı ve küme merkezleri, eylemin uzamsal bilgisini kodlayan kod kelimeleri olarak kullandı. Her bir eylem sınıfının zamansal yapısını kodlamak için, her grubun dizileri arasında sıkça oluşan alt dizileri çıkarmak için veri madenciliği teknikleri (Contrast Mining [35]'gibi) kullanılmıştır. Bu yöntem, görüntülerden poz algılamayı iyileştirmeye yardımcı

olan bir poz kurtarma tekniğinden yararlanmaktadır. Bunla birlikte her iki kodlama adımlarında veri madencilik tekniklerinin kullanması yüksek hesaplama maliyetine yol açmaktadır.

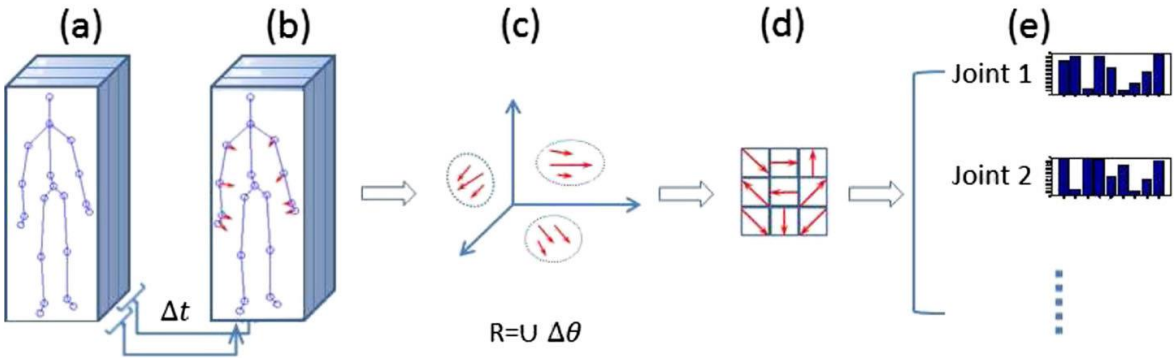
Pozlar üzerinde hesaplama maliyeti yüksek olan madencilik süreçleri yerine yöntemimiz her eylemin poz dizisini oluşturmak için sınıflandırmayı kullanarak onu daha verimli hale getirir.

Zamansal piramit yöntemi, kelime kümesi yönteminde zamansal bilgiyi temsil etmenin alternatiflerinden biridir [11, 74]. Pozların zamansal lokalizasyonu için bu yöntemi kullanan çalışmaların çoğu eylem dinamiğini göz ardı eder (Örneğin dizinin eşit parçalara bölünmesi).

Sonuç itibariyle, bu tür zamansal piramit yöntemleri farklı hızlarla gerçekleştirilen aynı eylemi tarif etmede yetersiz kaldı. Uygun olmayan segmentasyonun etkisini azaltmak için, Liu vd. [111] dizileri kırpmak için hareket enerjisi kullanan bir tanımlayıcı önermiştir.

Zamanı kodlama için iskeletlerin enerji değerlerini kullanmak yerine, önerilen tanımlayıcımız, geometrik özelliklerin zaman bağımlılığı modellemesi için her bir poz kelimesin önceki pozlarla rasgele seçilmiş bir kaydırma ofseti ile ilişkilendirir. Her bir poz kelimesi için aynı zamanda dizide olan bir önceki pozla ilgili kaydırma bilgisinin saklar. Başarılı çalışmalar eğitim aşamasında daha temsilci öznitelikler veri madenciliği veya diğer özellik seçme mekanizmaları gibi pahalı hesaplama yöntemleri ile seçiliyor [11]. Kelime kümesi yöntemleri için bu mekanizmaları kullanarak uzamsal / zamansal bilginin sağlanması daha karmaşık bir seviye beraberinde gelir.

Önerilen yöntem için literatürde bulunan kelime çantası yaklaşımını kullanan en benzer çalışma, Lu vd. [112] tarafından önerilmiştir (Şekil 31). Bu çalışmada, t'inci görüntü karesinde bulunan poz tanımlayıcının inşası için her bir eklem için zamansal kaydırma hesaplamış, öylesine ki ele alınan pozun i $(i \in 1,2,\cdots,m)$ eklemi için önceden rasgele olarak seçilmiş gecikme süresin Δt kullanarak, diferansiyel zaman farkı $\Delta\theta^i(\mathrm{t}) = \left(x^i{}_t - x^i{}_{t-\Delta t}, y^i{}_t - y^i{}_{t-\Delta t}, z^i{}_t - z^i{}_{t-\Delta t}\right)$ elde edilmiş. Ele alınan t'inci pozda tüm eklemler için hesaplanan $\Delta\theta^i(\mathrm{t})$ birleştirerek $\Delta\emptyset(\mathrm{t}) = \left\{\Delta\theta^1(\mathrm{t}), \Delta\theta^2(\mathrm{t}), \cdots, \Delta\theta^m(\mathrm{t})\right\}$ poz tanımlayıcısı elde etmişlerdir. Eylem dizisin temsilinin oluşturması için kelimeler çantası kodlama yöntemini kullanmıştır. Eğitim verilerinden elde edilen tüm $\Delta\theta^i(\mathrm{t})$'lerin üzerinde öklid mesafeli k-means kümeleme yaptıktan sonra sözlükte olan K tane kelimeni elde etmişler (Şekil 31-c).



Şekil 31. Eylem tanıma için yerel zaman offset ile kelime çantası kullanımı [112].

Eylemler dizisinin kodlamada aşamasında her bir eklemin dizi buyunca elde edilen özelikleri eğitilmiş sözlüğü kullanarak bir kelimeye üretmiştiler, daha sonra bu kelimeler üzerinden her bir eklem için ayrı histogramlar üretilmiştir (Şekil 31-e). Eylem sınıflandırması için her bir eklemden elde edilen histogramların birleştirip eylem için bir öznitelik yerine histogramları ayrı olarak ele almıştırlar, böylece pozda olan uzamsal bilgileri korumuşturlar. Sınıflandırma yöntemi olarak Naive Bayes sınıflandırma yöntemin geliştirerek sınıflandırmada olan sınıf uzaklığı yerine, histogramların tümünden elde edilen yer değişim bilgisini kullanmışlardır. Sınıflandırma için oy tabanlı (voting-based) ikinci bir diğer yöntem önermişler, öylesine ki her bir histogramı ayrı ve diğerlerinden bağımsız bir şekilde sınıflandırmışlar ve nihai kararı bunlarda elde edilen sonuçları oylama uygulayarak gerçekleştirmişler. Daha sonra benchmark veri kümeleri kullanarak her iki yöntemden elde edilen sonuçları rapor etmişler.

Yöntemleri ile tezde önerilen yöntemin arasında olan önemli fark poz kodlama fazında ortaya çıkar, bunlarda sırasıyla düşük düzey ve yüksek düzey poz kodlama gerçekleştirilir. Tezde önerilen yöntemde her kelime gerçek bir pozu açıklarken, [112]'de bir kelime, her bir yerel kısmı tanımlayan bir yön vektörüdür. Tez de önerilen tanımlayıcı, iskelet elemanlarının (eklemlerinin) sayısından bağımsızdır ve sadece anahtar poz sayısına bağlıdır ve düşük boyutlu öznitelik vektörlerini ürettiği için daha etkilidir. [112]'de global uzamsal bilgiyi yok sayarken, tezde önerilen yöntemde uzamsal poz bilgisiyle birlikte zamansal bilgileri kullanılır. Önerilen yaklaşım, kodlama için kelime kümesi (poz kümesi) yöntemini kullanan poz tabanlı bir yöntemdir ve iskelet verilerinin ham eklem pozisyonlarından çıkarılan basit öznitelikleri kullanması nedeniyle daha yüksek hesaplama verimliliğine ulaşarak diğer mevcut yöntemlerden ayırt edilmiştir. Bu özellikler Lie Grupları [58, 113] gibi başka bir

uzaya dönüştürülmeden doğrudan ham eklem pozisyonlarından çıkarılır. Zamansal bilgi, karmaşık veri madencilik yöntemlerini kullanmadan kelime kümesi sözlüğünde gömülmüştür [114]. Bu sözlük kelimesi olarak uzamsal / zamansal pozlar üretilmesi ile gerçekleştirilir. Bu nedenle, üretilen histogramlar doğal olarak zamansal bilgileri içerir ve zaman bilgisini ele almak için çoklu histogramların kullanmasına gerek kalmaz.

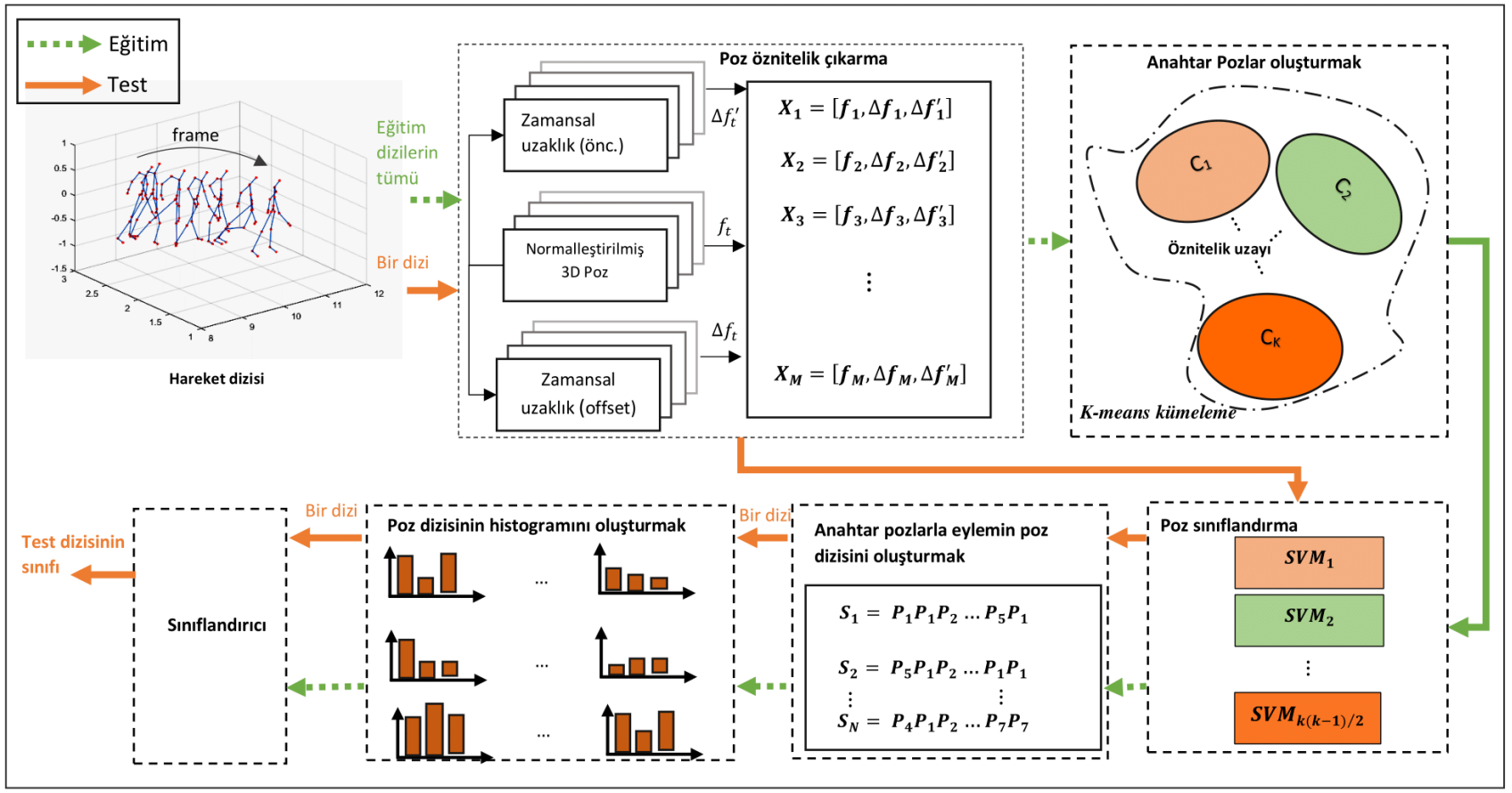
3. YAPILAN ÇALIŞMALAR

3.1. Giriş

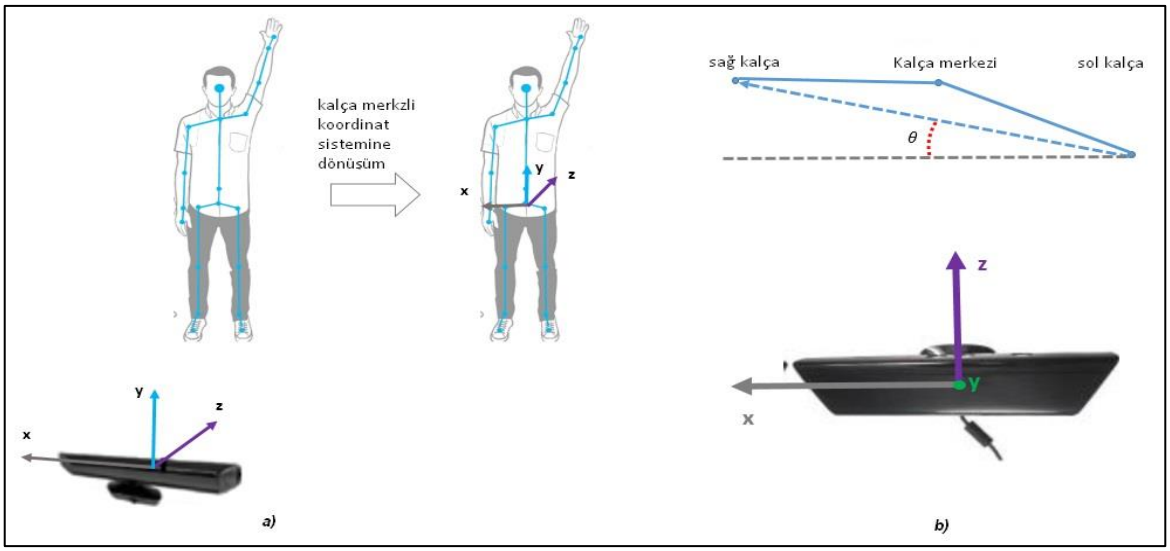
Video işleme çalışmaları 1980'li yıllardan beri çalışılmaktadır. O zamandan beri bilgisayar görü alanında, insan aktivite tanıma en zorlu işlemlerden biri haline gelmiştir. Konuyla ilgili gerçekleştirilmiş çalışmalara rağmen, eylem tanıma ile ilgili birçok sorun halen daha çözümlenememiştir. Son yıllarda Kinect sensörün ortaya çıkması ve derin öğrenme tekniklerindeki gelişmeler güvenilebilir ve maliyeti ucuz olarak 3B insan iskeleti çıkarılabilmektedir. Yapılan tez çalışmasında insan eylem tanınması için 3B iskelet verilerini kullanan bir poz çantası yöntemi önerilmiştir. Çalışmada her bir eylem, önceden tanımlanmış uzay-zamansal anahtar pozlarla temsil edilmektedir. Önerdiğimiz yöntemin akış diyagramının tümü Şekil 32'de gösterilmektedir.

Önerilen yöntem, girdi olarak T görüntü kare sayısına sahip her bir eylem için $S = \{P_t | \forall\, t \in (1, \cdots, T)\}$ iskelet eklemlerinden oluşan yüksek boyutlu bir vektör dizisini kabul etmektedir. Burada, $P_t = \left\{p_t^i \middle| \forall\, i \in (1, \cdots, J)\right\}$ t'inci görüntü karesinde olan iskeletin eklemler kümesini, $p_t^i = (x_t^i, y_t^i, z_t^i)$ t'inci görüntü karesinde olan p iskeletinin i'inci eklem pozisyonu ve J ise her bir iskelet üzerinde olan eklem sayısını temsil etmektedir. Şekil 33-a'da gösterildiği gibi, sabit kamera (x, y, z) koordinat sisteminin orjinine yerleştirilir.

Geleneksel kelime çantası yöntemlerinden esinlenerek önerilen yöntem, bir eylemi bir dizi poz küme (anahtar poz) olarak tanımlar. Eylem tanıma problemi üzerinde uygulanan kelime çantası yöntemlerinin kapsamlı derleme makalesi için [104]'a başvurabilir.



Şekil 32. Önerilen yöntemin akış diyagramı [36].



Şekil 33. a) Kinect kamerasının koordinat sisteminin orijinine konumlandırılması, b) döndürmek için hesaplanan açı [36].

3.2. Önişlem ve Öznitelik Çıkarma

Eklem koordinatları koordinat sistemine bağlı olarak farklılık gösterir. Aynı koordinat sisteminde bile, aynı pozun örneklerinin öteleme ve rotasyon nedeniyle farklı koordinatlara sahip olması muhtemeldir. Bunun yanında farklı vücut ölçülerine sahip iki kişi tarafından yapılan aynı pozlar için bile farklı koordinatlara sahip olabilir. Eklemlerin koordinatlarının birbiriyle karşılaştırılabilmesi için ilk önce verileri ortak bir koordinat sistemine taşımamız gerekir.

Önişlem aşamasında, her veri setinde bulunan iskelet verileri aşağıda belirtilen işlemlere tabi tutulur. Yapılan işlemler sayesinde, eylem tanıma sistemlerinde bulunan kişinin vücut yapısı, kişi konumu ve kamera görüş açısı gibi sorunlar çözülür. Bu girdi verileri üzerinden elde edilen öznitelikler daha doğru bir eylem temsilinin oluşmasını sağlar.

3.2.1. Ötelemeden Bağımsızlık

Eylemleri kaydeden kamera farklı pozisyonlarda olabilir veya bir kişi bir hareketi farklı konumlardan başlayıp gerçekleştirebilir. Farklı kamera pozisyonlarının etkisini azaltmanın bir çözümü iskelet koordinat sistem merkezini kamera merkezinden daha önce belirlenmiş bir ekleme ötelemektir. Önerilen yöntemde her görüntü karesi için, Şekil 33-a'da

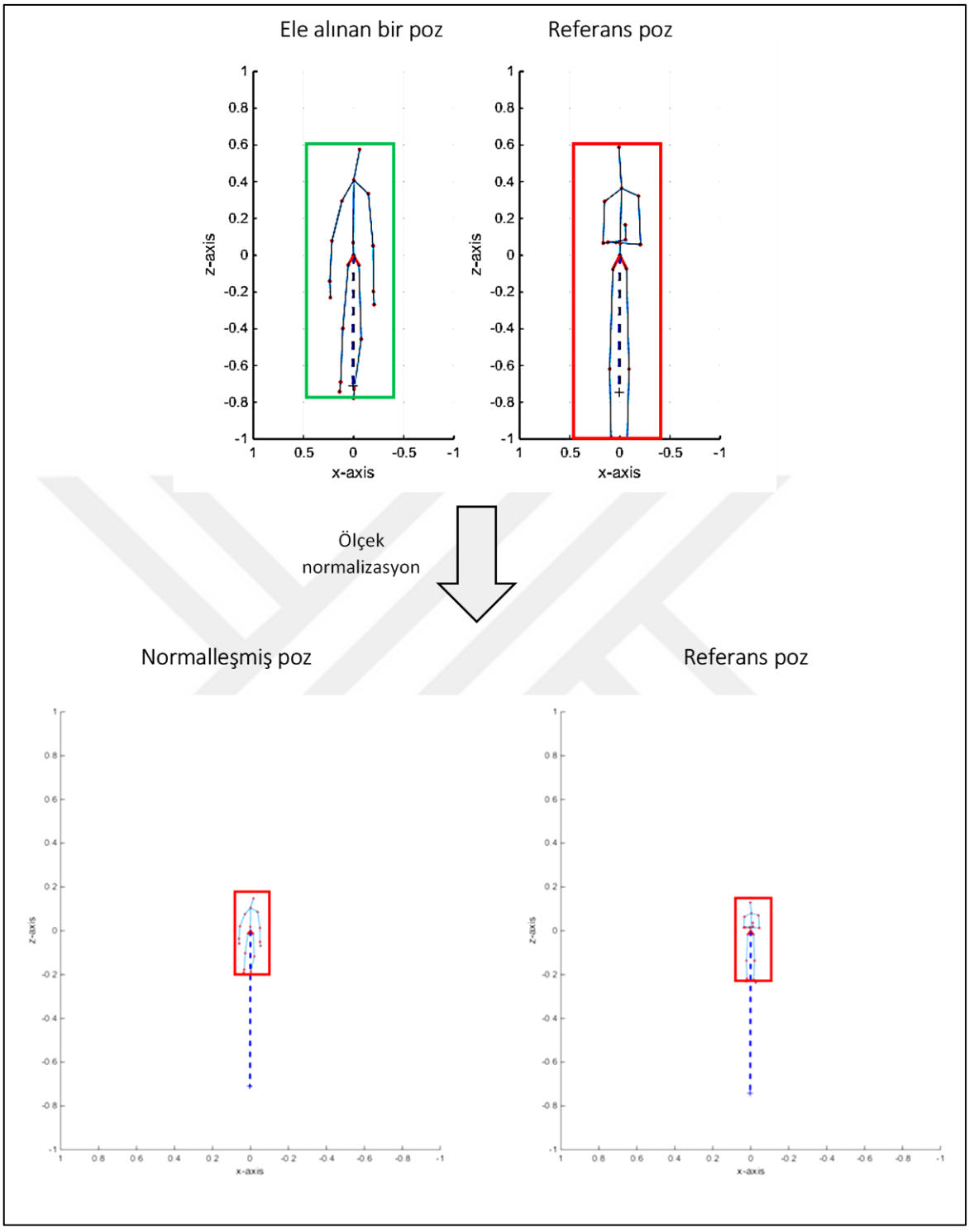
gösterildiği gibi koordinat sisteminin merkezi kamera koordinatlarından kişinin merkez kalça eklemine ötelenmiştir. Bu dönüşüm iskelet eklemlerinin pozisyonlarını kişinin konumundan değişmez kılar. Dönüşüm aşağıda gösterilmiştir. Burada *j* eklem indeksidir:

$$\left(x'_j, y'_j, z'_j\right) = \left(x_j - x_{kalça_{merkezi}}, y_j - y_{kalça_{merkezi}}, z_j - z_{kalça_{merkezi}}\right) \qquad (2)$$

3.2.2. Ölçekten Bağımsızlık

Genelde, eylemi yapan kişiler çeşitli vücut ebat boyutlarına sahiptirler. Güçlü eylem modellerine sahip olmak için, farklı vücut ebatlarına sahip kişilerden oluşturulan eylem öznitelikler eylem temsilleri arasındaki tutarlılığı korumalıdır. Ölçeğin değişmezliğini sağlamak için literatürde farklı yöntemler önerilmiştir.

Örneğin iskeletten elde edilen bir değer kullanılarak sabit bir aralığa ölçeklendirilme işlemi yapılmaktır. Bu değer, iskeletin uzunluğu, omuz genişliği veya toplam parça uzunlukların bire eşitlenmesi olabilir. Bunlar göz önüne alarak çalışmamızda [113]'de kullanılan benzer yöntemi kullanılmıştır. Şekil 34'de gösterildiği gibi, önce her veri seti için herhangi bir poz referans poz olarak seçilir, daha sonra iskelet dizilerinde bulunan diğer pozlardaki iskelet parçaları arasındaki orijinal açılar korunarak referans pozunda karşılık gelen iskelet parçalarının boyutları yeniden ölçeklendirilmiştir.



Şekil 34. Ölçek normalleştirme

3.2.3. Kamera Bakış Açısından Bağımsızlık

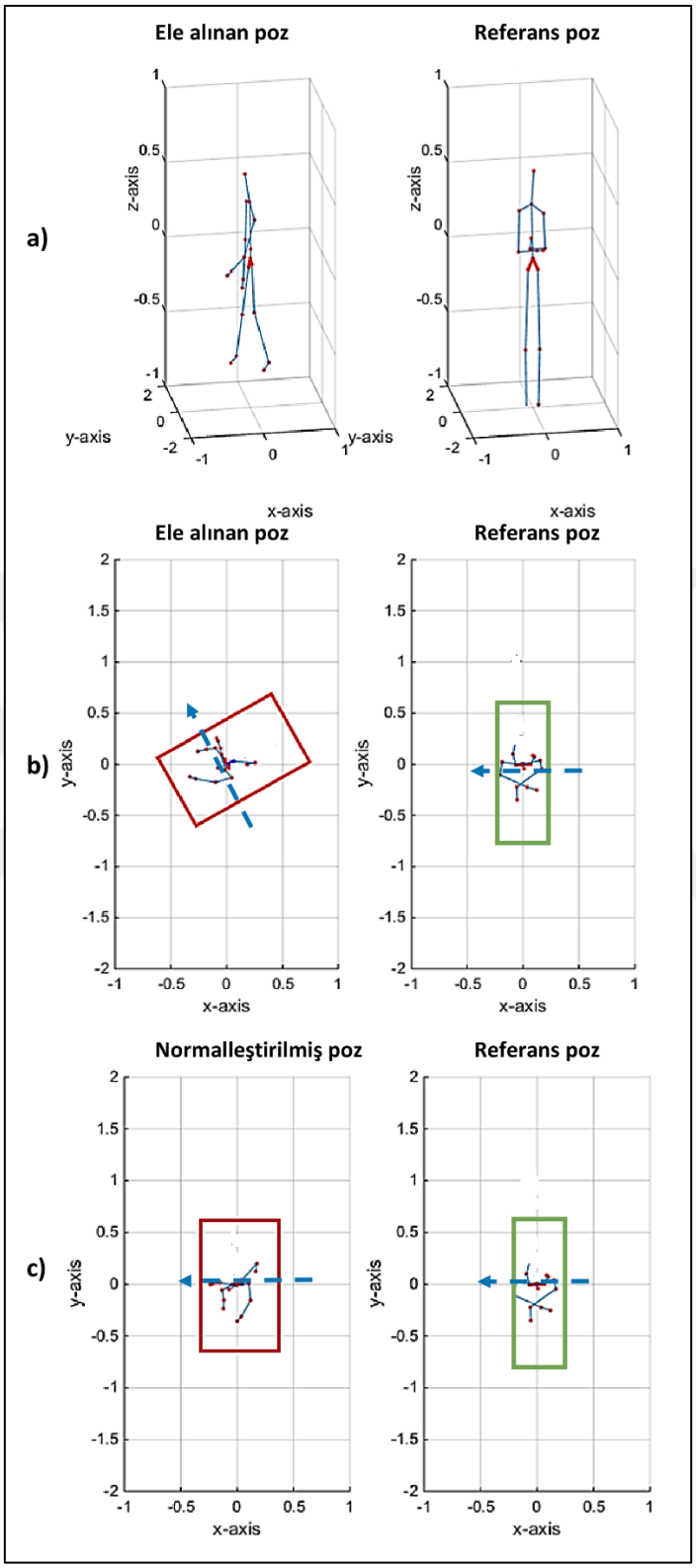
Bir eylem farklı görüş açılarından farklı gözükür. İskelet eklemlerini kamera görüş açısından bağımsız yapmak için kameranın belirtilen bakış açısına göre belirli bir döndürme işlemi gerçekleştirilir. Şekil 33-b'de gösterildiği gibi bu dönüşüm, sol kalçadan sağ kalçaya geçen vektörün, zemin düzlemindeki projeksiyonunun (izdüşümü), gerçek dünya koordinatlarında (kamera merkez noktası olan) x ekseni ile paralel kalmasını sağlar. Burada döndürme açısı Denklem (3) ile hesaplanır:

$$\theta = tan^{-1}\left(\frac{z_{sağ\_kalça} - z_{sol\_kalça}}{x_{sağ\_kalça} - x_{sol\_kalça}}\right) \quad (3)$$

Sapma açısını elde ettikten sonra, görüntü karedeki her bir iskelet eklemi için y ekseni etrafında saat yönünün tersine döndürme işlemi için Denklem (4) kullanılır.

$$\begin{bmatrix} x'_i \\ y'_i \\ z'_i \\ 1 \end{bmatrix} = \begin{bmatrix} cos\,\theta & 0 & sin\,\theta & 0 \\ 0 & 1 & 0 & 0 \\ -sin\,\theta & 0 & cos\,\theta & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ z_i \\ 1 \end{bmatrix} \quad (4)$$

İskelet dizisinde olan her iskelet için döndürme açısı hesaplandıktan sonra Denklem (4) kullanarak normalleştirilmiş poz Şekil 35-c'de gösterildiği gibi elde edilir. Üst bölümde açıklandığı gibi normalleştirilmiş pozda, Kinect kameraya doğru ve sol kalçadan sağ kalçaya geçen vektörün x eksenine paralel şekilde olması sağlanır.



Şekil 35. a) 3B uzayda ele alınan ve referans poz b) pozların 2B XY sayfasına izdüşümü c) normalleştirilmiş pozun XY sayfada izdüşümü

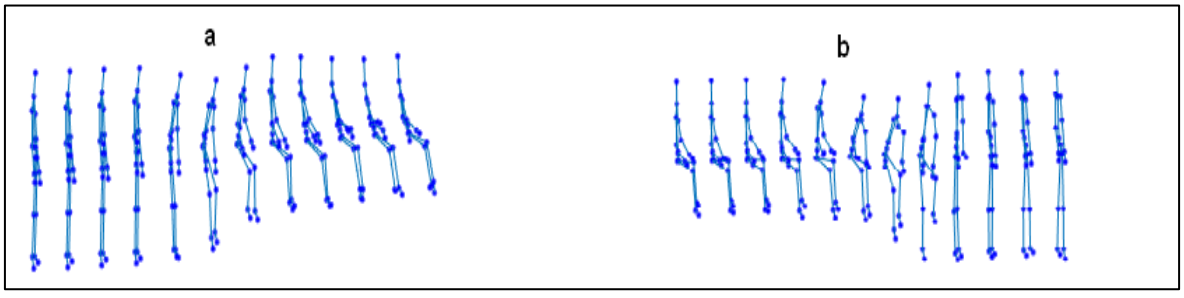
3.3. Öznitelik Çıkarma

Normalleştirilmiş bir poz verildiğinde bir sonraki adımda bir poz tanımlayıcı üretilir. Lillo vd. [59] çalışmasında poz tanımlayıcı özniteliklerini iki kategoride sınıflandırmıştır:

- Geometrik tanımlayıcı: Bu tanımlayıcılar, her görüntü karedeki iskelet eklemlerinin uzamsal konfigürasyonunu temsil eder. İskelet parçalarının vektörleri arasındaki açı veya farklı metrikleri kullanarak eklemler arasındaki mesafeyi hesaplarlar.
- Hareket tanımlayıcı: Bu Geometrik tanımlayıcılar iskelet eklemlerinin uzamsal konfigürasyonlarını tanımlayabilse bile, pozların dinamik bilgilerini kodlayamazlar. Poz hareket tanımlayıcılarının temsilinde hareket dinamiklerini kodlamak için, hesaplamalarda hız, ivme ve optik akış gibi bilgiler kullanılır. Şekil 36'da gösterildiği gibi, hareket tanımlayıcıları aynı zamanda benzer uzamsal konfigürasyonlarla farklı eylem karakteristiklerini temsil eden iki poz arasındaki belirsizliği önler.

Önerilen tanımlayıcı içsel olarak geometrik bilgiyi içeriyor olsa da, ardışık görüntü kareler arasındaki zamansal bağımlılığı göz önüne alarak hareket dinamiklerinin izini sürmeye çalışmaktadır. Nihai poz temsilcisi çeşitli tamamlayıcıların birleştirilmesi ile oluşmuştur.

Popüler bir birleştirme stratejisi, çıkarılan tüm özniteliklerin peş peşe birleştirilmesidir. Tanımlayıcının boyutlarındaki artışla orantılı olarak sınıflandırmanın hesaplama maliyeti de yükselir. Sınıflandırma aşamasının hesaplama maliyetini düşürmek için, sıklıkla PCA veya LDA gibi boyut azaltma prosedürü kullanılmıştır. Boyut azalması, tanımlayıcıların işlenmesinde verimlilik sağlamasına rağmen, hesaplama açısından pahalıdır ve bazı durumlarda doğruluğu yükseltmez [115].



Şekil 36. a) Örnek bir koltukta oturma eylemi b) örnek bir koltuktan kalkmak eylemi [36].

Özniteliklerin kör birleştirmeleri yerine, öznitelik mühendisliği adı verilen alternatif bir strateji öznitelik kümesindeki en iyi temsilcileri seçmeye çalışır. Öznitelik mühendisliği genellikle manuel (elle tasarlanmış) veya otomatiktir (öğrenme tabanlı). Örneğin, gözetimli seyrek sözlük öğrenimi (supervised sparse dictionary learning), yapay sinir ağları, genetik programlama, CNN ve rasgele karar ormanları (random decision forests) otomatik öznitelik belirleme yöntemleridir [116].

Öznitelik seçme mekanizmaları hesaplama açısından pahalı olduğundan, gerçek zamanlı bir uygulama için uygun bir seçenek olamazlar [34]. Öznitelik seçimi tabanlı yöntemlerin aksine, Özniteliklerimiz [79]'deki gibidir ve etkin bir poz tanımlayıcısı sağlar.

Şekil 32'de gösterildiği gibi, her görüntü karesindeki iskeletin uzamsal konfigürasyonunun tarifi için, t'inci görüntü karesi için öznitelik vektörünü $f_t = \left[x_t^1, y_t^1, z_t^1, x_t^2, y_t^2, z_t^2, \ldots, x_t^J, y_t^J, z_t^J\right]$ tanımlanmış, bu iskelet eklemlerinin normalleştirilmiş koordinatlarının birleştirmesidir (J iskelet üzerinde eklemler sayısıdır).

Önceden belirtildiği üzere, farklı görüntü karelerde olan pozlar arasındaki zamansal bağımlılığı modellemek ve çeşitli zamansal bağımlılıklar ile benzer eylem konfigürasyonlarının bilgisini içeren tanımlayıcıların sağlaması için Denklem (5) ile ifade edilen başka bir $\Delta \boldsymbol{f_t}$ vektörü üretiyoruz. Bu vektör rasgele seçilen bir görüntü kare ofsetini (t') dikkate alarak zamansal bağımlılığı aşağıdaki şekilde modeller:

$$\Delta \boldsymbol{f_t} = \begin{cases} f_t & 1 \leq t < t' \\ \frac{f_t - f_{t-t'+1}}{\|f_t - f_{t-t'+1}\|} & t' \leq t \leq T \end{cases} \qquad (5)$$

Mevcut poz ofsetten önce bulunursa, hesaplanan vektör normal eklem özniteliklerini içerir. Aksi halde (eğer olmazsa), mevcut poz ile zamansal ofset bağlı olarak üretilen diğer bir poz arasındaki kayma vektör hesaplanır. T değeri bir eylem dizisinde olan görüntü kare sayısı ve t' zamansal sabit bir ofsettir.

Ayrıca Denklem (6) ile ifade edildiği şekilde, dizide mevcut pozla bir önceki pozun arasındaki $\Delta f_t'$ kaydırma vektörünü hesaplayarak başka bir öznitelik vektörü oluşturulur.

$$\Delta f_t' = \begin{cases} f_t & t = 1 \\ f_t - f_{t-1} & 2 \leq t \leq T \end{cases} \qquad (6)$$

$X_t = [f_t, \Delta f_t, \Delta f_t^{'}] \in \mathbb{R}^D$, t'inci poz için nihai öznitelik vektör uzay-zamansal özniteliklerin peş peşe birleştirmesi ile oluşur ve doğrusal olarak iskelet eklemlerinin sayısına bağlı bir $D = 3 * J * 3$ boyuta sahiptir.

3.4. Anahtar Poz Üretmek

Kelime kümesi yöntemlerine benzer şekilde, önerilen yöntem, bir eylem dizisini önceden öğrenilen anahtar pozlar (sözlükteki kelimeler) ile temsil eder. Bu nedenle anahtar pozların sözlüğü öğrenilmelidir (Key poses selection) ve daha sonra, yüksek boyutlu poz öznitelikleri tek bir sözcükte kodlanır. Geleneksel olarak, sözlükleri öğrenmenin iki yolu vardır.

- Birinci yöntem, öznitelik uzayını alt bölgelere bölmek ve daha sonra her bölgeyi temsilcisi (kod sözcüğü) ile ifade etmektir. K-means algoritması bu amaçla yaygın şekilde kullanılmıştır [35, 74, 115].
- İkinci yöntem, özniteliklerin dağılımını üreten bir model kullanarak belirlemektir. Gaussian Mixture Model (GMM) bu konuda kullanılan en popüler yöntemdir.

Öznitelikler üzerinden sözcüklerin üretimi için K-means algoritması kesin atamalara (yani en yakın merkezi bulmak için öklid uzaklığını kullanır) dayalı iken GMM bunun yerine esnek atama gerçekleştirir (yani kod kelimeleri ataması için ortalama değer yerine özelliklerin olasılık dağılımını kullanır) [104].

Sınıflandırma doğruluğu eğitilmiş sözlük ve öznitelik kodlama kalitesi ile doğrudan orantılıdır. K-means algoritmasında, özellik vektörlerinin boyutu arttıkça öklid uzaklığı düşük performans gösterir ve güvenilmez kodlamalar üretmeye başlar.

Dolayısıyla, sözlük öğrenimi ve kodlamayı iyileştirmek için bunu iki adımda gerçekleştiriyoruz (Şekil 32) [115]. Poz kelimeleri (Anahtar pozlar) üretmek için, K-means algoritması Denklem (7)'de belirtilen tüm eğitim görüntü karelerinde olan pozlar için elde edilmiş öznitelik vektörlerine uygulanır.

$$Poses = \{\cup_m \cup_t X_t(m) \,|\forall m \in (1,2,\cdots,M) \;\; and \; \forall t \; \in (1,\cdots,T)\} \qquad (7)$$

Denklem (7)'de, M eğitim kümesinde bulunan tüm denemelerin (trials) sayısı ve T ele alınan denemenin görüntü kare sayısıdır. Sonuç olarak (neticede), öznitelik uzayı K kümeye bölünür ve bunlar küme merkezlerine karşılık gelir. Elde edilen küme merkezleri anahtar pozlar olarak kabul edilir ve bir sonraki adıma devredilir. Bu aşama önerilen yöntemin aşamaları arasında hesaplama karmaşığı açısından en yüksektir. Dolaysıyla en fazla zaman harcanan bölüm olarak tanımlanabilir. Bu sorunun eğitim sayısı az olan veri kümeleri çok fazla etkilememiş ama büyük veri kümelerde etkisi daha baskın olmuştur. Bu negatif etkiyi azaltmak için hızlandırılmış k-means kümeleme olan VLFeat kütüphanesi kullanılmıştır. k-means yönteminde yer alan yer değişim ölçeği için *L2* ve hızlandırılmış Elkan algoritması kullanılmıştır [117]. Diğer bir çözüm kümeleme işlemin eğitim pozlar üzerinden alınan örnekler üzerinde gerçekleştirmektir. Lu vd. [112] yaptıkları çalışmada, örnekleme işleminin eşit olasılıkla rastgele olarak yapılmasını önermişlerdir.

3.5. Poz Kodlama ve Sınıflandırma

Kodlama aşamasında öklid mesafesi kullanımından kaynaklanan sorunun gidermesi için sözlükte bulunan anahtar pozları kullanılarak bir dizi SVM sınıflandırıcısı eğitilir. SVM için LIBSVM [118] one-against-one tekniği ile kullanılmıştır. Bu adımda kullanılan SVM için farklı çekirdekler (kernel) denenmiş ve en başarılı sonuçlar polinom çekirdeğiyle elde edilmiştir. Sınıflandırmanın eğitmesi sadece bir örnekle gerçekleşir ve bundan dolayı literatürde benzer durumlarda SVM den başka sınıflandırıcı kullanılmamıştır [115].

K türlü pozun sınıflandırması için $G = K(K-1)/2$ tane ikili SVM eğitilmiştir. Öznitelik vektörlerinin anahtar pozuna atanması için, eğitilmiş ikili SVM'leri maksimum kazanım ("max wins") oylama stratejisiyle kullanılmıştır.

3.6. Anahtar Pozlar Histogramı ile Eylem Temsili

Bu adımda eğitilmiş SVM sınıflandırıcı kullanılarak her eylemin öznitelik vektörü anahtar poz dizilimine dönüştürülür. Üretilen pozların dizisi videolardaki görüntü kare sayısının çeşitliliği nedeniyle değişken uzunluktadır. Değişken uzunluk dizilerinin sınıflandırılması için Gizli Markov Modeli, Bayes Ağı ve Dinamik Zaman Bükmesi (Dynamic Time Warping) gibi yöntemler kullanılır [2].

Eylemlerin sınıflandırılması için, SVM, KNN ve ANN gibi ayrımcı sınıflandırıcıları kullanabilir. Öznitelik vektörlerinin uzunluğunu sabit bir uzunluğa normalleştirmek genellikle iki şekilde yapılı; video görüntü kare sayısın istenen boyuta örneklemek ve ardından öznitelik vektörlerini çıkarmak.

Diğer yöntem, öznitelik vektörlerinin değerlerini kuantalar ve eylemin tümünün temsili için kuantalanmış değerlerin histogramını kullanır [4]. Her eylemi sabit uzunlukta bir öznitelik vektörü ile tanımlanır. Daha sonra, oluşturulan anahtar pozları içeren dizinin histogramı hesaplanır. Bu hesaplamalardan önce, histogramların uzunluğu çıkarılan anahtar poz sayısıyla belirlenir.

3.7. Eylem Sınıflandırma

Sabit uzunlukta özellik vektörlerinin sınıflandırılması için KNN, SVM, ANN ve rastgele orman gibi birkaç popüler sınıflandırıcı vardır. Bu çalışmada, eylemleri sınıflandırmak için Aşırı Öğrenme Makinesi (ELM) sınıflandırıcısı ("hardlim" aktivasyon fonksiyonlu) kullanılmıştır [37]. ELM, çeşitli uygulamalarda başarılı bir şekilde uygulanan ve yüksek öğrenme hızı ve tutarlı doğruluk sergileyen (kanıtlayan) tek katmanlı ileri beslemeli sinir ağı sınıflandırıcısıdır.

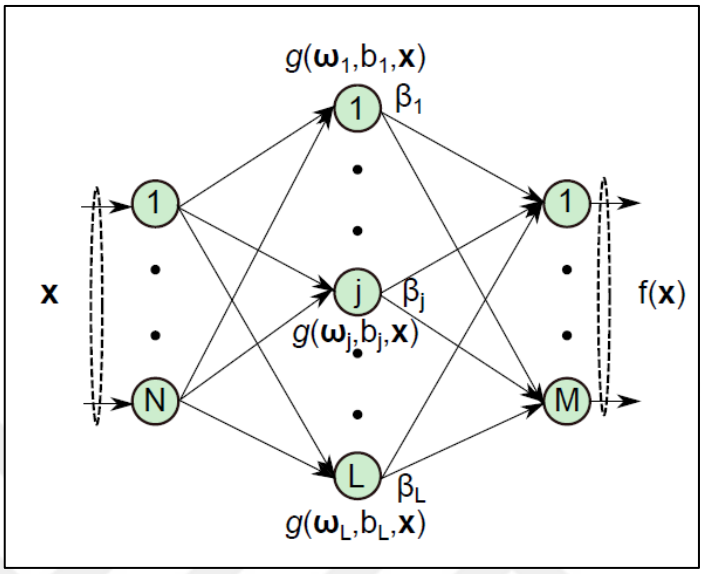
Aşırı öğrenme makinesi(Extreme learning machine*):* Aşırı öğrenme makinesi, sınıflandırma ve regresyonda kullanılan tek gizli tabaka ileri beslemeli sinir ağları (SLFN) için bir öğrenme algoritmasıdır [37].

Tek gizli katmanlı ileri beslemeli sinir ağı eğitiminde kullanılan ELM, gizli katman nöron sayısını ayarlayabilir ve girdi ağırlıkları ile gizli katman eşik değeri (biases) rasgele atayabilir, en küçük kareler yöntemi ile çıktı katmanı ağırlıklarını elde edilmiş, bütün öğrenme süreci tekrarlama olmadan sadece bir matematiksel dönüşümle tamamlandı.

Gardiyan iniş dayalı geleneksel geri yayılım (back propagation-BP) algoritması ile karşılaştırıldığında eğitim hızı önemli ölçüde iyileştirilmiştir (genellikle 10 kat veya daha fazla) [119]. Şekil 37'de L sayıda gizli nöron tek gizli katmanlı ileri beslemeli sinir ağı gösterilmektedir. Pratik uygulamalarda, öncelikle ELM eğitimi alınıp ardından onunla tahmin yapılır. Eğitim veri seti esas olarak belirli konularla birleştirilir. Veri setleri gerçek sonuçları ve ilgili faktörleri içermektedir.

Eğitim sırasında etki faktörleri ve bunlara karşılık gelen sonuçlar, eğitim için ELM'ye bir döngü sayesinde eğitim süreci bitene kadar verilir. Daha sonra, eğitimli ELM ile tahmin

etmek için, sadece girdi verilerinin girilmesi gerekir ve eğitim veri seti etkileyen faktörlere benzemektedir.



Şekil 37. Tek gizli katmanlı ileri beslemeli yapay sinir ağları [79].

Aşırı öğrenme makinesi kullanımı kolaydır ve tek gizli katman ileri besleme yapay sinir ağı için etkili bir algoritmadır. Geleneksel sinir ağı öğrenme algoritması (ör. BP algoritması), çok sayıda yapay ağ eğitim parametrelerini ayarlaması gerekir ve yerel optimuma takılabilir.

ELM algoritmasının sadece gizli katman nöron sayısını ayarlaması gerekir. Algoritma uygulama sürecinde ağ girdi ağırlıklarının ve gizli eşik değerlerinin ayarlanmasına gerek yoktur(rastgele bir şekilde atanır) ve hızlı öğrenme ve genelleme performansı avantajları ile benzersiz bir optimal çözüm üretir [119].

$\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{P}$ verilmiş P tane örnek ise ve burada $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iN}]^T \in \mathbb{R}^N$ ve $\mathbf{y}_i = [y_{i1}, y_{i2}, \dots, y_{iM}]^T \in \mathbb{R}^M$ olduğunda L tane gizli nörona sahip bir standart SLFNs çıkış fonksiyonu (8) ile tanımlanabilir;

$$\boldsymbol{y}_i = f(\boldsymbol{x}_i) = \sum_{j=1}^{L} \beta_j g(\omega_j . \boldsymbol{x}_i + b_j) \qquad (8)$$

Burada $\boldsymbol{\omega}_j = [\omega_{j1}, \omega_{j2}, \dots, \omega_{jN}] \in \mathbb{R}^N$ girdi katmanını j'inci gizli nöronuna bağlayan giriş ağırlık vektörüdür, b_j ise j'inci gizli nöronun eşik değeri, $g(.)$ doğrusal olmayan parçalı sürekli (piecewise) örneğin sigmoid veya Gaussian gibi bir fonksiyon ve

$\beta_j = [\beta_{j1}, \beta_{j2}, \ldots, \beta_{jM}]^T \in \mathbb{R}^M$ j'inci gizli nöron ve çıktı nöronlar bağlayan çıkış ağırlığı vektörüdür. *(8)* denklemi kısaca şu şekilde yazılabilir:

$$\boldsymbol{Y} = \boldsymbol{H}\beta \tag{9}$$

Burada **H** gizli katman çıkış matrisidir;

$$\boldsymbol{H} = \begin{bmatrix} g(\omega_1.\boldsymbol{x}_1 + b_1) \cdots g(\omega_L.\boldsymbol{x}_1 + b_L) \\ \vdots \quad \ddots \quad \vdots \\ g(\omega_1.\boldsymbol{x}_P + b_1) \cdots g(\omega_L.\boldsymbol{x}_P + b_L) \end{bmatrix} \tag{10}$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \ldots \\ \beta_1^T \end{bmatrix}_{L\times M} \quad and \quad \boldsymbol{Y} = \begin{bmatrix} y_1^T \\ y_2^T \\ \ldots \\ y_p^T \end{bmatrix}_{P\times M} \tag{11}$$

Geleneksel olarak, her gizli nöron için giriş ağırlık vektörü ω_j, eşik değer b_j ve çıkış ağırlık vektörü β_j değerleri, döngülü geri yayılım işlemle öğrenilir [15], ileri beslemeli sinir ağlarında kullanılan en popüler öğrenme algoritmasıdır. Geleneksel sinir ağlarından farklı olarak aşırı öğrenme makinesinde ω_j ve b_j değerleri öğrenilmez, bunun yerine sabit kalmaya devam eden rasgele değerler atanır.

ELM'yi eğitmek, Eşitlik *(8)*'e $\hat{\beta}$ en küçük kareler çözüm bulma eşdeğeridir. Gizli nöron sayısı L, eğitim numuneleri P sayısına eşitse, H matrisi karedir ve terslenebilir. Ancak, çoğu durumda $L \ll P$ ve doğrusal sistemin $\mathbf{H}\beta = \mathbf{Y}$ 'in en küçük kareler çözümünün en küçük normu *(11)* ile elde edilebilir;

$$\hat{\beta} = \boldsymbol{H}^{\dagger}\boldsymbol{Y} \tag{12}$$

Burada $\boldsymbol{H}^{\dagger}$, matris **H**'nin Moore-Penrose yalancı tersidir. $\boldsymbol{H}^{\dagger}$'yi hesaplamak için ortogonal projeksiyon yöntemi ve tekil değer ayrışması (singular value decomposition) da dahil olmak üzere çeşitli yöntemler mevcuttur ve $\hat{\beta}$ in **ç**özümü, doğrusal sistem için minimum eğitim hatası verir, ağırlıkların en küçük normunu sağlar ve benzersizdir.

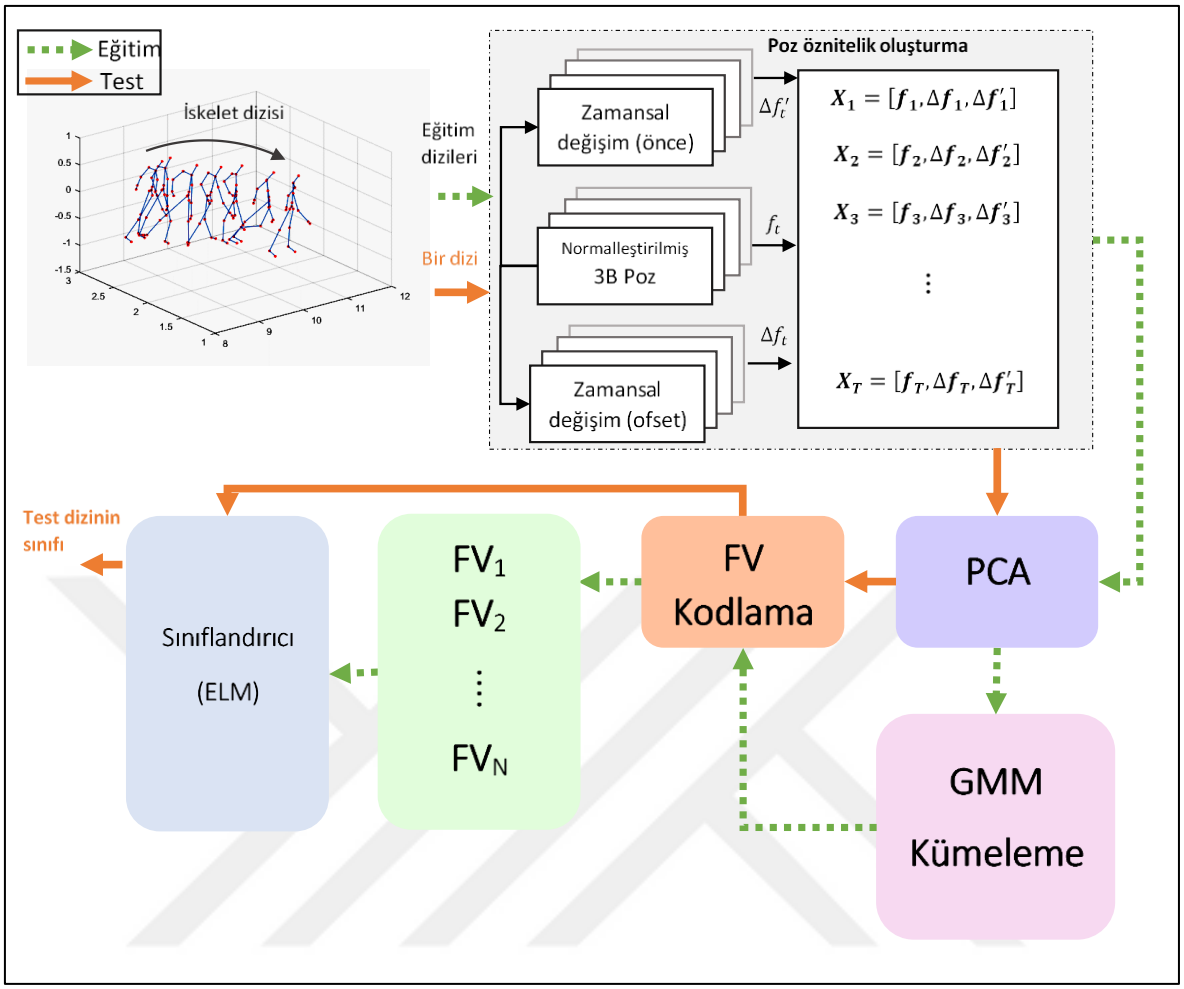
İlk kez, Minhas vd. [120] bu sınıflandırıcıyı, insan eylemlerinin belirlemesi için hareket esaslı özniteliklerde kullandılar ve umut verici sonuçlar elde ettiler. Üstelik bu yöntem düşük sınıf sayısı ve küçük ölçek sınıflamaları ile sınırlı değildir ve büyük ölçekli gerçekçi görevlerde kullanılabilir. Varol ve Salah [102] ELM'i gerçek video kliplerde eylem

tanıma için kullandılar ve derin sinir ağı yöntemlerinin ağır hesaplama maliyetini göz önünde bulundurarak onlarla kıyaslandığında kabul edilebilir sonuçlar elde etmiştirler [79, 121].

3.8. Fisher Vektör Kodlamaya Dayalı Eylem Tanıma

FV, görsel sözcüklerinden (özniteliklerden) oluşan görüntü veya video temsilin oluşturması için BoW ile birlikte en yaygın kullanılan yöntemlerdendir. Tez çalışmasına 3B iskelet verilerinden elde edilen uzay-zamansal poz tanımlayıcı kullanarak eylemin temsilin oluşturması için FV yöntemin kullanarak değerlendirmeler gerçekleşmiş. Yapılan işlemin tümü Şekil 38'de gösterilmiş.

Bölüm 2.3'de açıklandığı gibi eğer çıkarılan özniteliklerin uzunluğu D ve GMM de bileşenlerin sayısı K olursa o zaman FV boyutu $2DK$ olur. Oneata vd. [105], Varol ve Salah [102] aktivite tanıma için önerdikleri yöntemlerde FV kodlama aşamasından önce elde edilen öznitelikler üzerinde PCA yöntemin uygulayarak boyut küçültme işlemin gerçekleştirmişler. Benzer şekilde önerdiğimiz yöntemde de FV üretmeden önce PCA yöntemi boyut azaltmak için kullanılmıştır.



Şekil 38. Fisher vektör kodlamanı kullanan yöntemin iş akış diyagramı
Bu işlem bütün eğitim pozlar üzerinde gerçekleşmiş ve elde e

dilen küçülmüş öznitelikler GMM kullanarak belirlen K bileşenin parametreleri elde ediliyor. Şekil 38'de görüldüğü gibi sonraki aşamada FV kodlama elde edilen bu parametreleri kullanarak her bir eylemi temsil eden bir FV üretiliyor. Eğitim verilerinden üretilen FV'leri kullanarak ELM sınıflandırıcı eğitiliyor.

Sistem eğitildikten sonra gelen bir test dizisi öznitelik çıkarma ve ardından PCA işlemine dahil ediliyor. Bu işlem eylem dizisinde olan her poz için gerçekleşiyor ve ardından PCA çıktıları FV kodlama işlem vasıtasıyla bir FV dönüşüp ve sınıflandırması eğitilmiş ELM le gerçekleşiyor. Tez çalışmasında Matlab ortamda GMM ve FV kullanmak için VLFeat [117] kütüphanesi kullanılmıştır.

4. BULGULAR VE İRDELEME

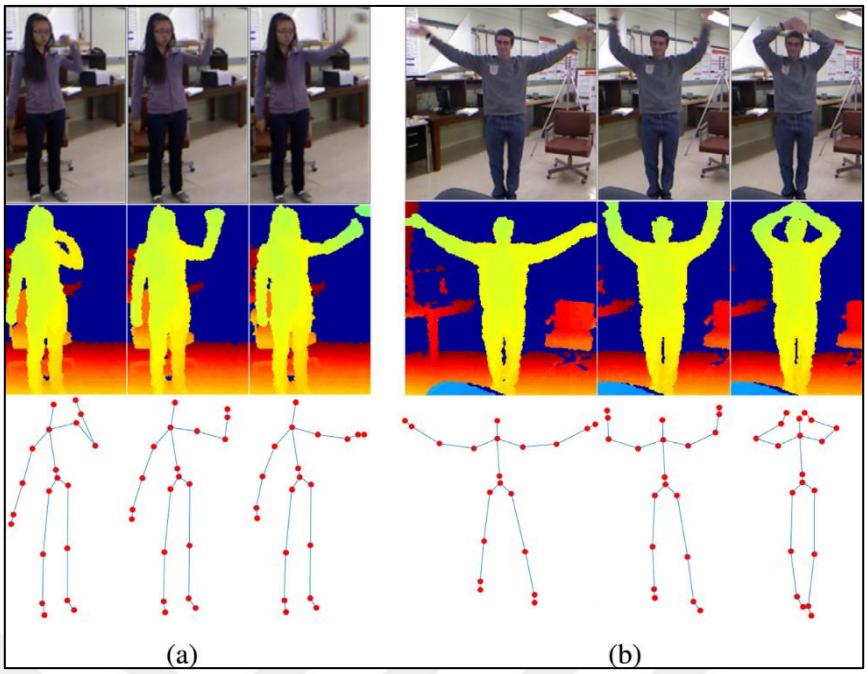
4.1. Halka Açık (Benchmark) Veri Setleri

Önerilen yöntem beş zorlu benchmark veri seti üzerinde değerlendirdik. Tahsis edilen eylemlerin gerçekleştirenleri tek bir kişi olduğu varsayıldı. Yani etkileşimli olmayan tek kişilik eylemleri ele alındı. Bu durum, etkileşimli eylemler değerlendirildiğinde neden bir performans düşüşü gözlemlediğimizi açıklamaktadır.

4.1.1. UTKinect-Action Veri Seti

Bu veri seti 2012 yılında Austin Üniversitesi'nde Xia vd.[55] tarafından toplanmıştır. Veriler Kinect v1 ile 30 fps hızla çekilmiş 10 eylem içermektedir. Her eylem, 10 kişi (9 erkek ve 1 kadın) tarafından 2 kez gerçekleştirilmiştir. Veri setinde toplam 200 dizi vardır. Veri seti RGB, derinlik ve iskelet veri dizilerin içerir, bu diziler manuel olarak kırpılmıştır (yani her dizi bir eylemin başlanma anından tam bitiş anına kadar şekliyle video üzerinden elle kırpılmış).

Şekil 39. UTKinect-Action veri setinde olan 10 eylemden örnek görüntüler [55].



Şekil 40. RGB görüntüleri, derinlik haritaları ve ilgili iskelet eklemlerinin (a) atmak (b) el salama eylemleri.

Benzer şekilde, her görüntü karesindeki iskelet veriler, 20 eklemin kamera merkez uzaylı Öklid konumu ile temsil edilmiştir. Eylemler de kişilerin kameraya karşı pozisyon ve yön değişkenliği, farklı kişilerde olan performans çeşitliği ve fark edilebilir hız farkı ve süresi bu veri setinin esas zorluklarıdır. İnsan-nesnelerin birbirini örtmesi ve görüş alanı gövde parçalarının dışına çıkması nedeniyle sensör tüm vücut parçalarını çıkaramıyor. Bunlar bu veri setinde karşılaşılan zorluklara eklenebilirler. Şekil 39'da RGB görüntüleri ile karşılıklı derinlik görüntüler gösterilmiştir. Burada eylemler soldan sağa ve yukardan aşağıya sırasıyla: yürüme (walk), ayağa kalkma (stand up), oturma (sit down), kaldırma (pick up), taşıma (carry), atma (throw), itme (push), çekme (pull), el sallama (wave hands) ve el çırpma (clap hands) [55] şeklindedir. Benzer şekilde Şekil 40'da RGB görüntülerin, derinlik haritalarının ve ilgili iskelet eklemlerinin atmak (a) ve el salama (b) eylemlerine karşılık gelen örnekleri verilmiştir.

4.1.2. CAD-60 Veri Seti

Kontrollü laboratuvar ortamında günlük faaliyetler nadiren meydana gelir. Bu, Cornell Üniversitesi'ndeki araştırmacıları, gerçek ortamlarda meydana gelen eylemler için CAD-60

veri seti [122] oluşturmaya motive etmiştir. Toplam 12 eylemi dört kişi tarafından 5 farklı ortamda gerçekleşmiş. Aşağıda her bir ortamda gerçekleşen eylemler verilmiştir.

- Banyo: ağzı durulamak, diş fırçalamak, göze lens takmak
- Yatak odası: telefonda konuşmak, su içmek, hap kutusunu açmak
- Mutfak: pişirmek (doğramak), pişirmek (karıştırmak), su içmek, hap kutusunu açmak
- Oturma odası: telefonda konuşmak, su içmek, koltukta konuşmak, koltukta dinlenmek
- Ofis: telefonda konuşmak, tahtada yazmak, su içmek, bilgisayarda çalışmak

Her bir örneğin derinlik, RGB ve iskelet verileri Kinect v1 kullanarak 30 fps hızıyla çekilmiştir. CAD-60 veri setinden bazı eylemlerin örnek resimleri, iskelet eklemleri ve derinlikleri Şekil 41'de gösterilmiştir. Burada, eylemler soldan sağa "diş fırçalamak", "telefonda konuşmak", "pişirmek (karıştırmak)", "koltukta dinlenmek" ve "göze lens takmak" şeklindedir.

Şekil 41. CAD-60 veri setinden bazı eylemlerin örnekleri

Her eylem en az bir kez her bir kişi tarafından gerçekleştirilmiş. Toplamda veri setinde, bir eylem için ortalama 45 saniyelik 60 diziyi içermektedir. Her görüntü kare için iskelet verisi, sensör koordinatın referans noktası alarak 15 eklemin Öklid konumu ile sunulmuştur. Bu veri setinin esas zorlukları, yetersiz eğitim verileri ve değişken arka plandır. Kişilerin biri solak olduğundan dolayı eylemler farklı ellerle yapılır. Farklı el etkisini telafi etmek amacıyla, literatürde önerilen yöntemlerden bazıları [122-124] çalışmalarında bu örneklerin

yansıtılmış bir halini eğitim verilerine ekleyerek kişilerin kullanılan el tercihine karşı bir değişmezlik elde edilmiştir.

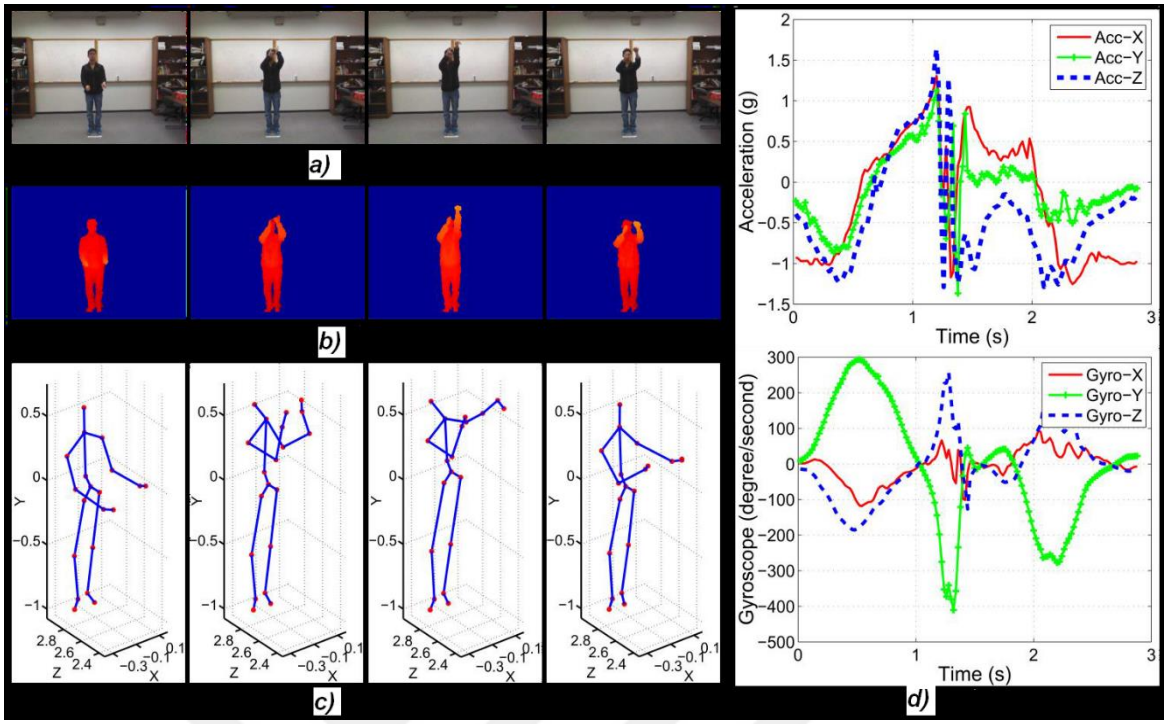
4.1.3. UTD-MHAD Veri Seti

UTD-MHAD [125] birçok modlu bir veri seti olarak Texas Üniversitesi tarafından aktivite tanıması için yayınlanmıştır. Veriler Kinect v2 tarafından (30 fps hızla) ve bir giyilebilir inertial sensörle çekilmiş. RGB, derinlik, iskelet ve inertial sinyali de dahil olmak üzere 4 veri modalitesi, bu sensörleri kullanarak zamansal senkronize modda kaydedilmiştir. Basketball-shoot eylemine karşılık gelen multimodalite verilerinin bir örneği Şekil 42'de verilmiştir. Burada sırasıyla renkli görüntüler, derinlik görüntüleri (her görüntü karesinde de arka plan çıkarılmış), iskelet eklemler ve inertial sensör verisi (acceleration ve gyro-scope sinyalleri) verilmiştir.

Veri seti 27 eylem içermektedir: (1) kolu sol kaydır, (2) kolu sağa kaydır, (3) el sallama, (4) el çırpma, (5) fırlatmak, (6) kol çapraz, (7) Basketbolda şut, (8) X çizmek, (9) saat yönünde daire çiz, (10) saatin ters yönünde daire dire çiz, (11) üçgen çiz, (12) bowling, (13) yumruk atmak, (14) beyzbol savurmak, (15) teniste savurmak, (16) kol kıvırmak, (17) teniste servis atmak, (18) itme, (19) kapıya vurmak, (20) el tutmak, (21) tutup atmak, (22) Koşmak, (23) Yürümek, (24) Kalkmak, (25) Oturmak, (26) Hamle, (27) Çömelmek.

Bu eylemler, sabit bir arka plana sahip bir ortamda 8 kişi (4 erkek ve 4 kadın) tarafından gerçekleştirildi. Her kişi her eylemi 4 kez icra etti. Her görüntü karesinde olan iskelet verileri, sensör merkezli koordinatlara göre 20 eklemin Öklid konumu ile sunulmuştur.

Başka bir taksonomide, bu veri seti, eylemleri dört alt kategoride sınıflandırmıştır: spor eylemleri (örneğin bowling, teniste servis atmak ve basketbolda şut), el hareketleri (örneğin X, üçgen ve daire çizmek), günlük aktiviteler (örneğin kapıyı dövmek, kalkma ve oturma) ve eğitim egzersizleri (örneğin kol kıvırmak, hamle ve çömelmek).

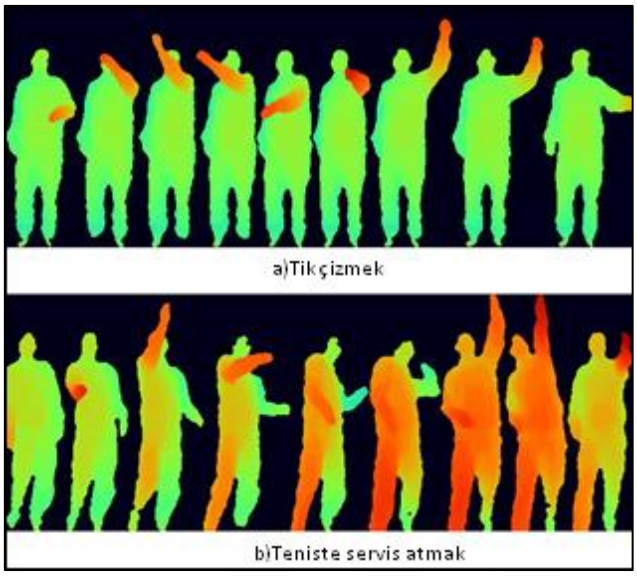


Şekil 42. Basketbolda şut eylemi: a) RGB görüntüler, b) derinlik görüntüleri c) iskelet eklemler ve d) inertial sensör verisi [125].

4.1.4. MSR Action 3D Veri Seti

MSR Action 3D veri seti [126], ilk halka açık (araştırma amaçlı kullanılabilen) RGB-D eylem veri setidir ve Microsoft Research Redmond tarafından oluşturulmuştur. Veri seti Kinect v1 tarafından 15 fps'de hızla kaydedilmiş ve farklı vücut parçalarını içeren 20 eylemi; Yukarı Kol Sallamak, Yana Kol Sallamak, Çakmak, El Tutmak, İleri Yumruk Atmak, Yükseğe Fırlatmak, X Çizmek, Tik Çizmek, Daire Çizmek, El Çırpmak, İki El Sallamak, Yandan Yumruk Atmak, Bel Çevirmek, İleri Tekme Atmak, Yana Tekme Atmak, Koşmak, Teniste Savurmak, Teniste Servis Atmak, Golf Oynamak, Tutup Atmak içerir. Her eylem 10 kişi tarafından 2-3 kez gerçekleştirilmiştir. Toplamda, veri setinde uzunlukları 13 ile 67 görüntü kare arasında değişen 567 dizi vardır. Her dizi Şekil 43'te gösterildiği gibi manuel olarak bölünmüştür ve sadece bir eylemi içermektedir. Veri seti ayrıca her bir eylemin derinlik ve iskelet verilerini de içermektedir.

Her görüntü karesindeki iskelet, sensör merkez koordinata göre 20 eklemin Öklid konumu ile temsil edilmiştir. Örneklerin tümünde, kişiler kameraya doğru bakan sabit bir konumda ve kontrollü bir arka planda eylemler gerçekleştirilmiştir.



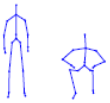
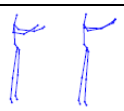
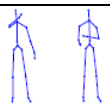
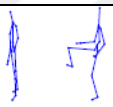
Şekil 43. a) Tik çizmek ve b) Teniste servis atmak için örnek derinlik harita dizisi [126].

4.1.5. MSRC-12 Veri Seti

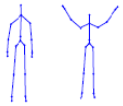
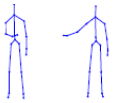
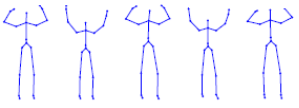
MSRC-12 veri seti [70] önceki veri setleriyle karşılaştırıldığında daha çok veri içermektedir. Dolayısıyla bu veri seti yaklaşımımızın ölçeklenebilirliğinin daha uygun değerlendirmesini sağlamaktadır. Bu veri seti sadece iskelet bilgilerini içermektedir ve farklı talimatların eylem tanınmasına olan etkilerini değerlendirmek için Microsoft Research tarafından toplanmıştır. Saniyede 30 görüntü kareye kaydedildi ve her görüntü karesinde olan iskeletin 20 tane ekleminin pozisyonları kayıt edilmiştir. Her biri 6 eylemden oluşan ve iki gruba ayrılmış toplam 12 eylem 30 kişi tarafından gerçekleşmiştir. Veri setinde yer alan iki grup jest Tablo 2 ve Tablo 2'de verilmektedir.

Toplamda, veri setinde 594 ardışık eylem dizisi vardır, her birinde, bir özne tarafından ardışık şekilde 5 farklı talimatla gerçekleştirilen bir eylem bulunur (toplamda, veri setinde 6244 eylem örneği vardır). Dizilerin bölütlemesi için, [67]'de sunulan etiketleme bilgileri kullanılır.

Tablo 1. Iconic jestler [70, 79].

Jest açıklama	Statik görüntüler	Önemli pozlar
Eğil, saklan		
Silah ateşle		
Bir nesne fırlat		
Silah değiştir		
Tekme		
Gözlük giy		

Tablo 2. Metaphoric jestler [70, 79].

Jest açıklama	Statik görüntüler	Önemli pozlar
Müziğin sesini aç		
Menüler arasında dolaş		
Müziği aç		
Alkışlamak		
Müziğe İtiraz		
Tempoyu yükselt		

4.2. Önerilen Yaklaşımın Deneylerinde Kullanılan Ortak Ayarlar

Önerilen yöntemin değerlendirmesi için deneylerimizde kullanılan beş veri setinin genel özellikleri Tablo 3'de özetlenerek gösterilmiştir. Deneyimlerimizde, iskelet eklemlerinin 3B koordinatları dünya koordinatlarından kişi koordinatlarına dönüştürülmüştür. Her bir görüntü karesi için koordinat sisteminin merkezi kalça merkezi eklemine ötelenmiştir. Her bir veri seti üzerinde elde edilen sonuçlar tanıma işlemi için yalnızca iskelet verileri kullanan yöntemlerle karşılaştırılmıştır.

Tablo 3. Veri setlerin özetleri

Veri seti	Eylem sayısı	Kişi sayısı	Örnek sayısı	Eklem sayısı	Yayınlanma yıllı
UTKinect-Action [55]	10	10	199	20	2012
CAD-60 [122]	12	4	60	15	2011
UTD-MHAD [125]	27	8	861	20	2015
MSRAction3D [126]	20	10	557	20	2010
MSRC-12 [70]	12	30	594$_{(6244\ \text{deneme})}$	20	2012

4.3. Önerilen Yaklaşımların Uygulaması

4.3.1. Poz Çantası Kodlama Uygulaması

Bu yöntem için gereken üç girdi parametresi her veri seti için ayrı ayrı ayarlanmıştır. Birinci parametre, öznitelik vektöründe zamansal farkın (Δf_t) oluşturulmasında kullanılan t' zamansal kayma ofsetidir. İkinci parametre, k-means kümeleme yöntemindeki kümelerin sayısıdır ve tüm eğitim pozlarından anahtar pozların oluşumu için kullanılmıştır. Başka bir deyişle, anahtar poz sayısını temsil eder. Ayarlanması gereken son parametre ise ELM'nin gizli katmanındaki nöronların sayısıdır. Tüm bu parametrelerin en optimum değerlerinin bulunmasında öncelikle geniş bir parametre aralığı büyük adımlarla taranmıştır. Sonrasında ise belirlenen en uygun aralık üzerinde aralıklar daraltılarak en uygun parametre değerleri belirlenmeye çalışılır. Tablo 4'de gösterildiği gibi önerilen tanıma sisteminde en iyi toplam

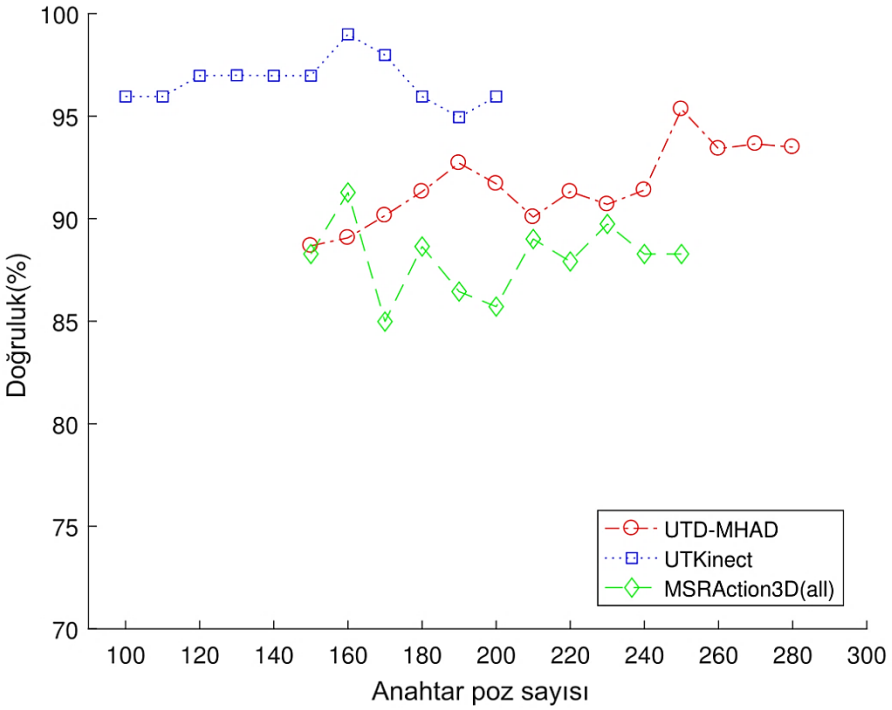
performansı sağlamak için deneysel olarak optimum adım aralıkları ve en uygun adım boyutu belirlenmiştir.

K-means yöntemi anahtar pozların hesaplamasında kullandığında, kümelenme merkezlerinin rasgele başlatılması gerçekleştirilir. Önerilen yöntem bu nedenle her veri setinde *20* kez tekrarlanır ve en iyi sonuç raporlanır ve şu ana kadar bilinen en iyi olan yaklaşımlarla karşılaştırılır.

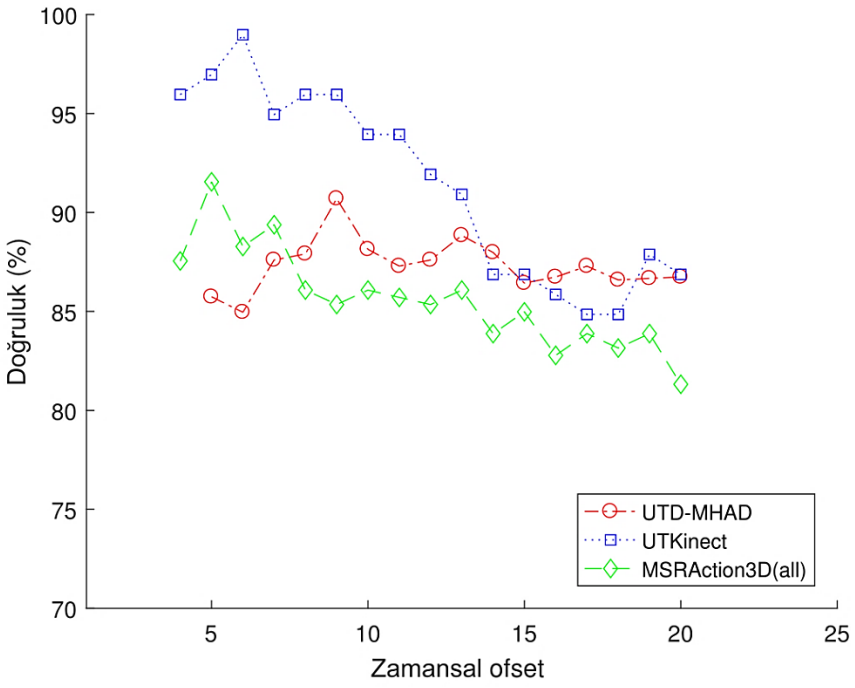
Tablo 4. Poz çantası yönteminde en optimum değerleri elde etmek için kullanılan parametre aralıkları ve adımları.

Veri seti	İncelenen aralık ve adımlar		
	Zamansal ofset	Anahtar poz sayısı	Nöron sayısı
UTKinect-Action [55]	4:1:20	100:10:200	500:100:3500
CAD-60 [122]	10:10:150	100:10:250	500:100:3500
UTD-MHAD [125]	4:1:20	150:10:280	500:100:3600
MSRAction3D [126]	4:1:20	150:10:250	500:100:3500
MSRC-12[70]	4:1:11	100:10:200	500:100:3100

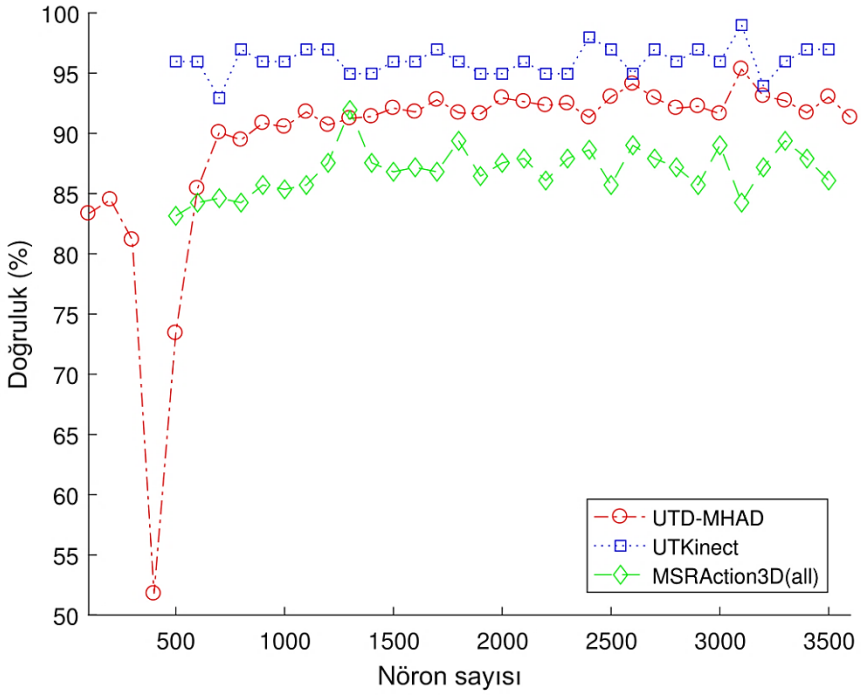
Ayrıca, parametre incelemelerinin ayrıntılı işlemi Şekil 44, Şekil 45 ve Şekil 46'da açıklanmıştır. Şekillerde her bir veri setinin farklı protokoller nedeniyle değerlendirilen veri kümelerinden üçünün ayar süreçleri gösterilmiştir. Veri setindeki pozları temsil etmede daha az sayıda anahtar pozun başarısız olduğu görülürken, daha fazla sayıda poz da gürültüyü arttırır ve doğruluğu düşürmektedir. Her veri setinde en iyi performansı veren anahtar poz sayısı için bir optimum nokta bulunmaktadır. Şekil 45'de, ofset değeri arttıkça doğruluğun önemli ölçüde düştüğü açıkça görülebilmektedir. Az sayıda nöronun UTD-MHAD veri setinin doğruluğu üzerinde olumsuz etkisi bulunmaktadır. Nöron sayısı artıkça ise sonuçlar doğrusal olarak iyileşmektedir. Diğer veri setleri için doğrulukta önemli dalgalanmalar yoktur.



Şekil 44. Anahtar poz sayısının değerlendirilmesi [36].



Şekil 45. Zamansal ofset parametresinin değerlendirilmesi [36].



Şekil 46. Nöron sayısının değerlendirmesi [36].

4.3.2. FV Kodlama Uygulaması

FV kodlama yöntemin kullandığımızda poz dizileri üzerinde öznitelik çıkarma işlemi önerilen poz çantası yöntemindeki gibi gerçekleştirilmektedir. Böylece birinci girdi parametre zamansal kayma ofseti (t')'tır ve optimum değeri her veri seti için farklı olarak belirlenir. Bölüm 3.8'de açıkladığımız gibi işlemin devamında her bir poz için oluşturulan öznitelik vektörün üzerine PCA uygulanarak boyut azaltma işlemi gerçekleşiyor. Dolaysıyla ikinci girdi parametre çıkış boyutudur ve her veri seti için optimum değeri farkı olarak belirleniyor. İşlemin devamında eğitim pozlar tümünün üzerinde GMM uygulanarak istenilen K bileşen üretiliyor. K değerin belirlemek bu yöntemin üçüncü girdi parametresi oluyor. Her eylem için üretilen FV temsil sabit boyuta sahiptir ve aynen poz çantası yöntemlerde elde edilen histogramların sınıflandırması gibi bunlarda da ELM kullanılıyor, böylece sonuncu girdi parametre ELM'nin gizli katmanında kullanılan nöron sayısı oluyor.

Tablo 5'de gösterildiği gibi tanıma sisteminde en iyi toplam performansı sağlamak için deneysel olarak optimum adım aralıklarını ve en uygun adım boyutları belirlenmiştir. Önerilen yöntemin değerlendirmesi UTKinect, MSRAction 3D ve UTD-MHAD veri

setlerinde gerçekleşmiştir. Bu veri setlerinden yeterinde iyi sonuçlar elde edilmeyince diğer veri setlerinde deneme yapılmayıp diğer bir kodlama yöntemi (poz çantası) denenmiştir.

Tablo 5. FV yönteminde en optimum değerleri elde etmek için kullanılan parametre aralıkları ve adımları

Veri seti	İncelenen aralık ve adımlar			
	Zamansal offset	İndirgenen boyut	GMM'de bileşen sayısı	Nöron sayısı
UTKinect-Action [55]	4:1:20	50:20:150	4:1:15	500:100:3500
UTD-MHAD [125]	4:1:20	50:20:150	4:1:20	500:100:3600
MSRAction3D [126]	4:1:20	50:20:150	4:1:20	500:100:3500

Ele alınan veri kümelerinde FV kullanınca optimum parametreleri bulmak için Tablo 5 kullanılmıştır. Her veri seti için en iyi performans deneysel olarak Tablo 6'da gösterilen parametre değerlerle elde edilmiştir.

Tablo 6. FV kodlama yöntemi için belirlenen en optimum parametre değerleri

Parametreler	Veri setler		
	UTKinect	MSR Action 3D (All)	UTD-MHAD
offset	7	4	9
GMM sayısı	8	15	12
PCA katsayısı	100	60	40
Nöron sayısı	3100	2100	3100

4.4. Önerilen Yaklaşımların Literatür Karşılaştırmaları

4.4.1. UTKinect Veri Setinde Literatür Sonuçları

UTKinect-Action veri setine dayalı özgün [55] çalışmasındaki değerlendirmelerde test edilen dizini dışarda bırakma (leave-one-sequence-out) (LOSeqO) protokolü kullanmıştır. Bu protokolde, her bir seferde tüm veri setinden test örneği olarak rasgele bir dizi seçilmiştir ve kalan diğer dizileri eğitim verisi olarak kullanılmıştır.

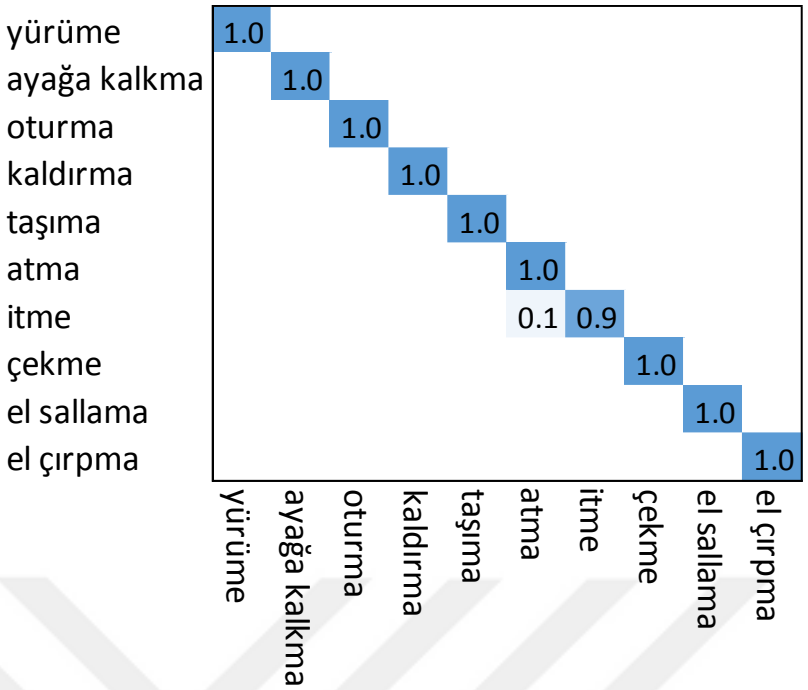
Bu süreç belirli defa tekrarlandı ve elde edilen sonuçların ortalaması son performans olarak kullanılmıştır [127]. Denemelerimizde, [58]'de kullanılan Çapraz kişi (Cross-subject) protokolünü takip edilmiştir. Eğitim için 1, 3, 5, 7 ve 9 kişiler tarafından yapılan eylemler ve 2, 4, 6, 8 ve 10 kişiler tarafından yapılanlar test için seçilmiştir.

Test kişilerinin eylemleri eğitim setinin dışında tutulduğundan bu değerlendirme protokolü daha gerçekçidir. UTKinect-Action veri setinde BoP yöntemi kullanılınca optimum parametreleri bulmak için Tablo 4'de verilen değerler kullanılmıştır. En iyi performans sonuçları zamansal ofset değeri 6, anahtar poz sayısı 160 ve nöron sayısını 3100 olduğunda elde edilmiştir. Aynı veri setinde FV yöntemi kullanılınca optimum parametreleri bulmak için Tablo 5'de verilen değerler kullanılmıştır. FV kodlama yönteminde en iyi performans sonuçları Tablo 6'da verilen parametre değerleri ile elde edilmiştir. Önerilen her iki yöntemle elde edilen sonuçlar ve önceki çalışmalarda bilinen en iyi yöntemlerle karşılaştırmalar Tablo 7'da verilmiştir. Bildiğimiz kadarıyla, Tablo 7'da gösterildiği gibi UTKinect eylem veri seti kullanan tüm iskelet tabanlı yaklaşımlar arasında elde edilen en iyi performans başarısı BoP kodlama kullanan yöntemimizle elde edilmiştir.

Tablo 7. UTKinect veri setinden elde edilen sonuçların literatürle karşılaştırılması

Öznitelik Müh.	Yöntemler	Doğruluk (%)
Hand-crafted	HOJ3D [55] (LOSeqO)	90.9
	Lie Group [58]	97.0
	Spatiotemporal SHs [121]	93.0
	PAIRWISE JOINTS [111]	94.4
	Önerilen Yöntem (FV)	97.0
	Önerilen Yöntem (BoP)	99.0
Learned representations	RDF-based [128]	92.0
	Max-Margin Multitask [129] (LOOCV)	98.8
	LMNN [33] (LOOCV)	98.0
RNN-LSTM	Multilayer LSTM [130]	95.9
	ST-LSTM [32]	95.0
	TS-LSTM [31]	97.0

Şekil 47'deki karışıklık matrisinden (confusion matrix) görüleceği üzere "itme" eylemi yapılan test örneklerinin %10'nu "atma" eylemi olarak yanlış sınıflandırılmıştır. İki eylemde bulunan pozlar arasındaki benzerlik ve elde edilen iskelet eklem pozisyonlarında yer alan gürültü tanıma başarısızlığının ana nedenleridir. Elde edilen sonuçlar incelendiğinde önerilen yöntemin on eylemden dokuzunu yüzde yüz doğrulukla tanımaktadır.



Şekil 47. UTKinect veri setinin karışıklık matrisi

4.4.2. CAD-60 Veri Setinde Literatür Sonuçları

Sung vd. [122] CAD-60 veri setinin değerlendirmesi için "yeni kişi" (new person) ve "görüldü" (have seen) olmak üzere iki tip protokol sundu. "yeni kişi" değerlendirme protokolünde bir kişinin gerçekleştiği eylemler test için ve diğer kişilerin tarafından gerçekleşen eylemler eğitim için kullanılmaktadır. "görüldü" protokolünde ise tüm kişilerin tarafından gerçekleşen eylemlerin yarısı eğitim için ve kalan diğer yarısı test aşamasında kullanılmaktadır. Önerilen yöntemlerini değerlendirmek için Kesinlik\Hassasiyet ölçekleri kullandılar.

Denememizde, değerlendirmeler için "yeni kişi" protokolünü kullanıldı. Bu protokol, bir kişi dışarıda bırakma doğrulama yöntemi olarak tanımlandı. Bu nedenle test için bir kişi, diğer üç kişi ise eğitim için tutulmuş. CAD-60 veri setinde, dört kişiden biri solaktır (3. Numara kişi). Eylemlerin sağ elle yapılanlara benzemek ve lateralliğini dönüştürmek için özellik vektörünü oluşturmadan önce yansıtma işlemleri kullanılır. CAD-60 veri setinde [131] bilinen en iyi sonuç veren yöntemdir. Bu yaklaşımda, test için 2 numara, eğitim için diğer 3 kişi (1, 3 ve 4) düşünülmüştür.

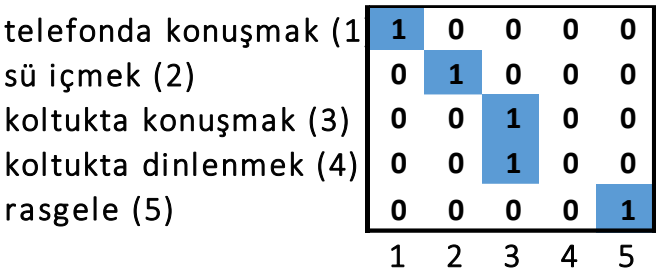
Denemelerimizde aynı ayarı kullanılmaktadır. Bu veri setindeki eylemlerin uzunluğu bir öncekinden daha uzundu. Tablo 4 kullanarak farklı parametre aralıklarını ve adım

boyutlarını denedik. Bu aralıklarla parametreler için olası senaryoları inceleyerek, zamansal ofset için 50, anahtar poz numarası için 210 ve nöron sayısı için 3100 ile CAD-60 veri setinde en iyi performansı elde ettik.

Önerilen yöntemin performansı ve literatürde CAD-60 kullanan başarılı yaklaşımlarla karşılaştırmalar Tablo 8'de gösterilmektedir. Bildiğimiz kadarıyla, Tablo 8'de gösterildiği gibi CAD-60 eylem veri seti kullanan tüm iskelet tabanlı yaklaşımlar arasında elde edilen en iyi performans, yöntemimizle elde edilmiştir.

Tablo 8. CAD-60 veri setinden elde edilen sonuçların literatürle karşılaştırılması

Öznitelik Müh.	Yöntemler	%	
		Kesinlik	Hassasiyet
Hand-crafted	MEMM [122]	67.9	55.5
	3D posture [115]	77.3	76.7
	Pose Kinetic Energy [123]	93.8	94.5
	Decision-level Fusion (Majority Voting) [131]*	96.4	84.6
	Önerilen Yöntem (BoP)	98.5	99.0
Learned representations	M-L key pose [132]	97.4	95.8
	Self-organizing neural int [124]	91.9	90.2
	RF-Key pose [133] (Random + Still)	81.8	80.0



Şekil 48. CAD-60 veri setinin oturma odası eylemlerinin karışıklık matrisi (3 uncu kişi)

Üçüncü kişi hariç, farklı ortamlardaki tüm eylemler yüzde yüz başarı ile tanınmıştır. Karışıklık matrisinden (Şekil 48) anlaşıldığı üzere, "koltukta dinlenmek" yerine "koltukta

konuşmak" hareketini tanımak, üçüncü kişinin hareketlerinde meydana gelen tek başarısızlıktır ve bu başarısızlığın temel sebebi yeterli sayıda eğitim örneklerinin bulunmamasıdır.

Üçüncü kişiye ilişkin "koltukta dinlenmek" eylemi için yalnızca bir test örneği olduğundan, hesaplanan kesinlik değeri $0/0$ gibi tanımsız bir değer üretilir. "oturma odası" ortamında eylemlerin ortalama kesinliğini hesaplamak için bu değeri sıfır olarak düşünmüştür.

4.4.3. UTD-MHAD Veri Setinde Literatür Sonuçları

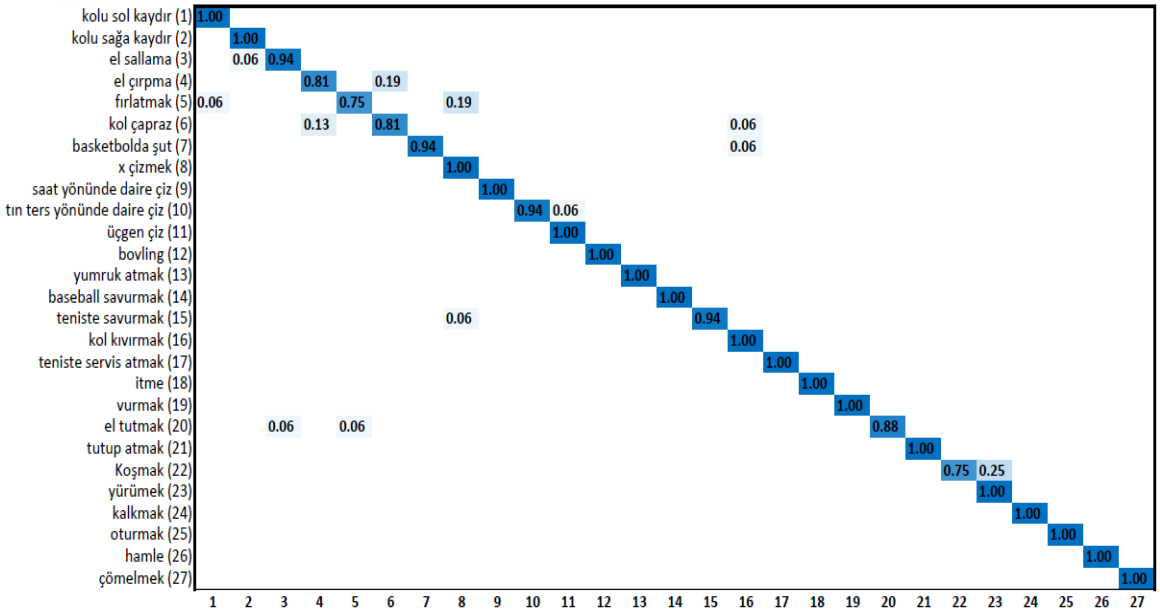
UTD-MHAD veri setindeki [125] çalışmasında sağlayıcıları tarafından önerilen çapraz kişi (cross-subject) değerlendirme protokolü kullanılmıştır. Bu protokolde kişilerin yarısı (1, 3, 5 ve 7) eğitim için ve kalan diğer yarısı ise (2, 4, 6 ve 8) test için alınmıştır.

Denemelerimizde, BoP yöntemini kullanarak önerilen yöntemin değerlendirilmesi için aynı ayarlar kullanılmıştır. Önceki veri setlerine benzer şekilde, Tablo 4'de belirtilen değerler üzerinden optimum parametreler araştırılmıştır. UTD-MHAD veri setinde zamansal ofset için 9, anahtar poz numarası için 250 ve nöron sayısı için 3100 ile en iyi performans elde edilmiştir (bu parametrelerin değerlendirilmesi Şekil 44, Şekil 45 ve Şekil 46'da gösterilmektedir). Aynı veri setinde FV yöntemi kullanınca optimum parametreleri bulmak için Tablo 5 kullanılmıştır. En iyi performans değerleri Tablo 6'da verilen parametre değerleri ile elde edilmiştir. Bildiğimiz kadarıyla, UTD-MHAD veri setindeki tüm iskelet tabanlı yaklaşımlar arasındaki en iyi performans değerleri BoP kullanan yöntemimiz tarafından Tablo 9'da gösterildiği gibi elde edilmiştir.

Bu veri seti üzerinde BP yöntemimizin karışıklık matrisinin (Şekil 49) analizi, ortak pozları paylaşan eylemlerin hatalı tanımaya neden olduğunu göstermiştir. Örneğin "fırlatmak" eylemi %75 doğrulukla sınıflandırılırken, örneklerin yüzde yirmisini "X çizmek" olarak yanlış sınıflandırılır. Benzer bir durumda, "koşmak" eyleminde örneklerin %25'inde "yürümek" eylemi olarak yanlış sınıflandırılmıştır. Buna rağmen, 27 eylemden 18'i yüzde yüz doğrulukla tanınmıştır. Ayırıcı pozların bulunması, önerilen yöntemin bu eylemleri mükemmel bir doğrulukla anlamasına yol açmıştır.

Tablo 9. UTD-MHAD veri setinden elde edilen sonuçların literatürle karşılaştırılması

Öznitelik Müh.	Yöntemler	Doğruluk (%)
Hand-crafted	Kinect & Inertial [125]*	79.1
	Kinect & Inertial fusion [134]*	91.5
	ELC-KSVD [135]	76.1
	Cov3DJ [67]	85.5
	Önerilen Yöntem (FV)	92.6
	Önerilen Yöntem (BoP)	95.3
CNN	SOS based CNN [110]	86.9
	JTM_CNN [108]	85.8



Şekil 49.UTD-MHAD veri setinin karışıklık matrisi

4.4.4. MSR Action 3D Veri Setinde Literatür Sonuçları

MSR-Action3D [126] veri setini değerlendirmek için önceki çalışmalarda kullanılan iki ayar vardır. Birincisi, MSR-Action3D veri setinin özgün (seminal) makalesinde [126] önerilmiştir ve Tablo 10'da gösterdiği gibi tüm eylemleri üç alt kategoriye (AS1, AS2 ve AS3) ayrılmıştır. Alt kategorilerin her biri sekiz sınıf eylemi içermekte ve her kategori üzerinde eğitim ve testler bağımsız olarak gerçekleştirilmektedir. AS1 ve AS2 alt kategorilerinde, benzer hareketleri olan eylemler gruplanmıştır. Bu kategoriler algoritmaların benzer yapıya sahip eylemlerin tanınmasında olan ayırt edici yeteneğinin

değerlendirilmesi için kullanılmıştır. AS3 alt kategorisi, karmaşık vücut dinamiklerinden oluşan eylemler içermektedir ve bir yöntemin çeşitliliğini değerlendirmede (çok farklı tür hareketlerle tanıma kabiliyeti) için kullanılmaktadır. Bir sistemin bu veri seti üzerinde genel performansı, alt kategorilerin performans ortalaması alınarak elde edilir.

Tablo 10. MSR-Action3D veri setinden elde edilen sonuçların literatürle karşılaştırılması.

AS1	AS2	AS3
Yana kol sallamak	Yukarı kol sallamak	Yükseğe Fırlatmak
Çakmak	El tutmak	İleri tekme atmak
İleri yumruk atmak	X çizmek	Yana tekme atmak
Yükseğe fırlatmak	Tik çizmek	Koşmak
El çırpmak	Daire çizmek	Teniste savurmak
Bel çevirmek	İki el sallamak	Teniste servis atmak
Teniste servis atmak	İleri tekme atmak	Golf oynamak
Tutup atmak	Yandan yumruk atmak	Tutup atmak

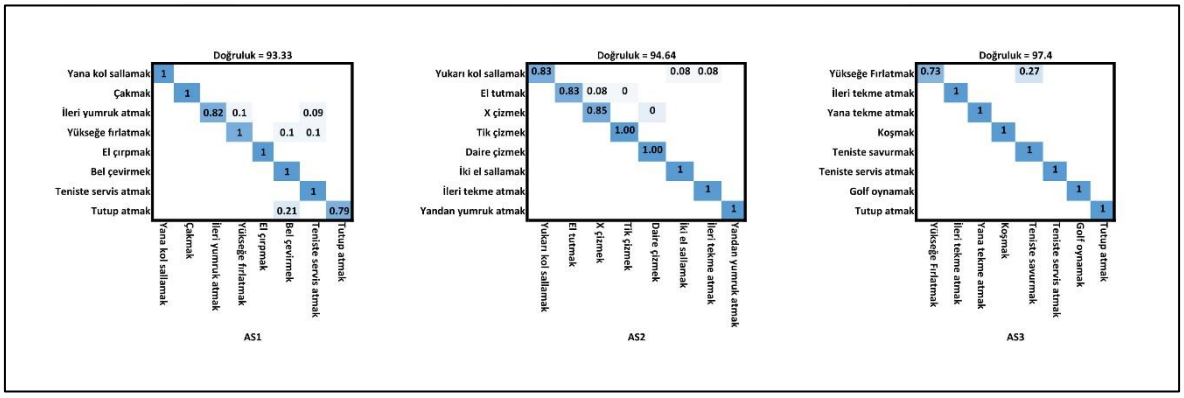
İkinci deney protokolü ise [63] çalışmasında önerilmiştir. Bu deneyde 20 eylemin tümü veri setini bölmeden eğitim ve test için tek bir kümede tutulmuştur. Bu deney protokolü ilk ayara kıyasla sınıflandırmayı daha da zorlaştırır. Denemelerimizde, her iki ayarda kullanılmıştır. İlk protokol için [58]'ye benzer şekilde çapraz kişi kullanılmıştır. Kişilerin yarısı (1, 3, 5, 7 ve 9) eğitim için diğer yarısı ise (2, 4, 6, 8 ve 10) test için kullanılmıştır.

BoP kodlaması kullanıldığında Tablo 4'de belirtilen parametrelerin tüm olası senaryolarını inceleyerek, MSR Action 3D veri seti üzerinde zamansal ofseti 4'e, anahtar poz sayısını 100'e ve nöron sayısını 3100'e ayarlayarak en iyi performans değerleri elde edilmiştir. Bu parametrelerin değerlendirmeleri sırasıyla Şekil 44, Şekil 45 ve Şekil 46'da gösterilmektedir. MSR Action 3D veri seti üzerinde yöntemimizin iki değerlendirme protokolüne göre performansı ve bilinen en iyi iskelet tabanlı yöntemlerle karşılaştırılması Tablo 11'de gösterilmektedir (ikinci protokolün sonuçları "All" ile belirtilen sütunun altındadır). Öznitelik türüne bağlı olarak, yöntemler elle tasarlanmış (handcrafted) veya otomatik türlere kategorize edilmiştir.

Önerilen yöntem, öznitelikleri yalnızca öklid uzayında, [58, 121]'deki gibi başka bir uzaya dönüştürülmeden hesaplanırken, elle tasarlanmış öznitelik yöntemleri arasında kabul

edilebilir bir performans elde etmiştir. Ayırt edici öznitelikleri seçmek için veri madenciliği tekniklerini kullanan [33] gibi yaklaşımlar üstün sonuçlar elde etmiştir. Bu yöntemlerle de olan eylem tanıma performans iyileşmesi, özellikle eğitim aşamasında olan hesaplama maliyetinin artışıyla örtüşmektedir. Tablo 11'de gösterildiği gibi yöntemimiz daha karmaşık yapılan eylemleri içeren AS3 üzerinde [29, 76, 121] ile kıyasladığında nispeten daha iyi sonuçlar üretmektedir.

Birinci protokolde, diğer iki alt kategoriyle karşılaştırıldığında, AS1 alt kategorisindeki eylemler önerdiğimiz yöntem için daha zorludur ve eylemlerin karmaşıklığına bağlı olarak daha düşük bir performans doğruluğu ile sonuçlanmıştır.



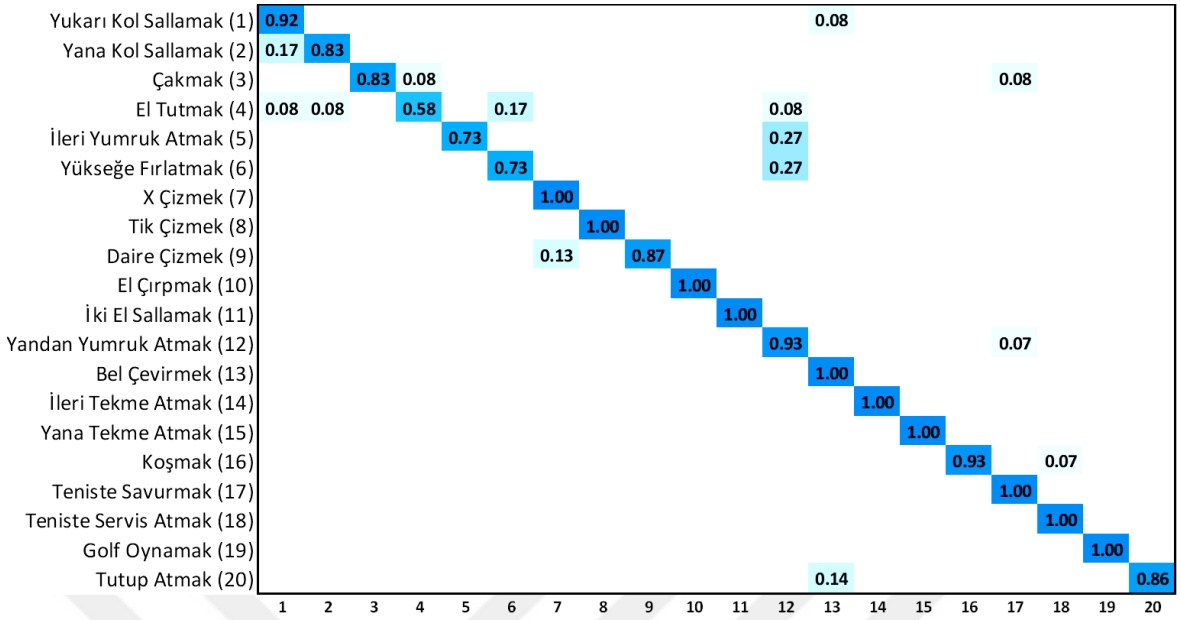
Şekil 50. Birinci protokole göre MSR-Action3D veri setinin karışıklık matrisi

Tablo 11. MSR-Action3D veri setinden elde edilen sonuçların literatürle karşılaştırılması

Öznitelik Müh.	Yöntemler	Doğruluk %				
		AS1	AS2	AS3	Ortalama	Tümü
Hand-crafted	Pose-based [35]	-	-	-	90.2	-
	HOJ3D [55]	-	-	-	78.9	-
	Lie Group [58]	95.3	83.9	98.2	92.5	89.4
	Spatiotemporal SHs [121]	89.7	91.7	92.5	90.9	-
	PAIRWISE JOINTS [111]	-	-	-	93.8	-
	RRV [136]	-	-	-	-	93.4
	Trajectory let [137]	96.4	97.5	100	97.9	-
	Önerilen Yöntem (FV)	-	-	-	-	81.3
	Önerilen Yöntem (BoP)	94.3	94.6	97.2	95.4	91.9
Learned representations	LMNN [33]	-	-	-	97.1	-
	Moving Pose lets [76]	89.8	93.5	97.0	93.5	93.6
	Max-Margin Multitask [130]	-	-	-	95.6	90.5
RNN/LSTM	HBRNN-L [29]	93.3	94.6	95.5	94.5	-
	ST-LSTM [32]	-	-	-	94.8	-
	TS-LSTM [31]	95.2	96.4	100	97.2	-

Şekil 50'de verilen karışıklık matrisinden ("Tutup atmak" eyleminin test örneklerinin %79'unda doğru olarak sınıflandırıldığı açıkça görülmektedir. Ancak, bu eylem örneklerinin %21'inde "Bel çevirmek" olarak yanlış sınıflandırılmıştır. AS2 alt kategorisinde en yüksek yanlış sınıflandırma oranı "El tutmak" eyleminde meydana gelmiştir ve burada örneklerin %8'inde "X çizme" ve 8%'de "Tik çizme" olarak yanlış sınıflandırılmıştır. Son alt kategori AS3'te ise, en yüksek yanlış sınıflandırma oranı "Yükseğe fırlatmak" eylemine aittir ve burada örneklerin %27'sinde "Teniste savurmak" olarak yanlış sınıflandırılmıştır.

Tanıma başarısızlığının esas nedeni, her bir eylem sınıfı için ayrıt edici anahtar pozların oluşturma eksikliğidir. Yaklaşımımız farklı zamanlarda ancak aynı anahtar pozu üretir. Oluşturulan pozlar zaman bilgisi içeriyor olsalar bile, yöntem karmaşık eylem kodlama prosedürü sırasında eylemlerin bazıları için dizideki pozların zaman sırasını kaybeder.



Şekil 51. İkinci protokole göre MSR-Action3D veri setinin karışıklık matrisi

Şekil 51'de gösterilen ikinci protokolün karışıklık matrisi, en başarısız sınıflandırma doğruluğunu "El tutmak" eylemi sırasında %58 doğru tanıma oranıyla elde edilmiştir. Diğer taraftan bu "El tutmak" eyleminin %17'si "Yükseğe fırlatmak", geri kalanı her biri yaklaşık %8 oranında "Yana kol sallamak", "Yukarı kol sallamak" ve "Yandan yumruk atmak" olarak yanlış sınıflandırılmıştır. Bu hata eylem dizisinde bulunan benzer pozlara dayanabilir. Bu eylemler dışında "İleri yumruk atmak" ve "Yükseğe fırlatmak" eylemleri %73 oranında doğru sınıflandırmayla en başarısız sonuçlar üreten diğer iki eylemdir. Bu eylemlerde elde edilen başarısız sonuçlara rağmen, bu protokol üzerinde yöntemimiz 20 eylemden 10'unu %100 doğru olarak tanımıştır.

4.4.5. MSRC-12 Veri Setinde Literatür Sonuçları

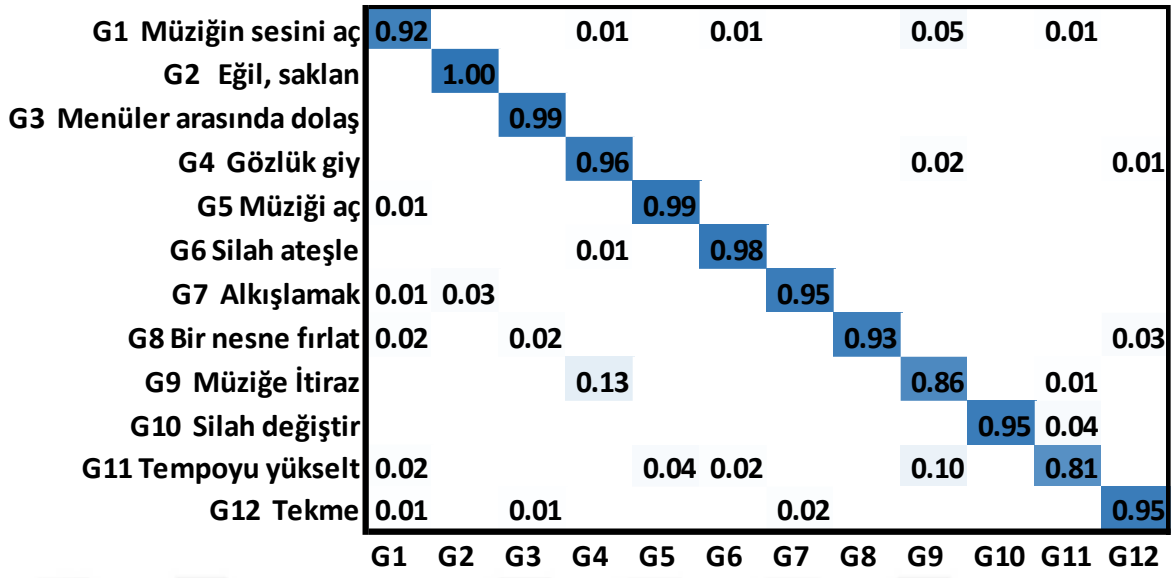
Literatürdeki incelendiğinde MSRC-12 veri setini kullanan çalışmaların değerlendirme aşamasında sıklıkla iki yaygın çapraz-kişi (cross-subject) protokolü kullanmıştır. [34] ise çalışmasında birini dışarda bırakma (leave-one-out) protokolü kullanılmıştır. Bu protokolde eğitim için 29 kişi ve kalan 1 kişi ise test için kullanılmıştır. Bu birini dışarda bırakma protokolünde veri setindeki her bir kişi test verisi olacak şekilde tekrarlanmıştır ve nihai sonuç elde edilen tüm performansların doğruluklarının ortalaması alınarak elde edilmiştir. İkinci protokol [29, 113] örneklemin yarısını (tek sayılı kişilerin

örneklerini) eğitim diğer yarısını ise test için kullanmıştır. Değerlendirmelerimizde ikinci protokol kullanılmış, optimal parametreler Tablo 4'e göre tespit edilerek ayarlanmıştır.

Yaptığımız değerlendirmelerin sonuçları ve bilinen en iyi iskelet temelli yöntemlerle yapılan karşılaştırmalar Tablo 12'de verilmiştir. Tablo 12'de gösterildiği gibi, yöntemimiz öznitelikleri elle tasarlanmış yöntemlere göre kıyaslanabilir bir performans sağlamaktadır ve örneklerin sayısı arttıkça daha iyi performans göstermesi beklen öğrenim tabanlı yöntemlerden üstün olmaktadır. Yapılan değerlendirmeler, yöntemimizin ölçeklenebilir (scalable) olduğuna ve örneklerin sayısı arttıkça güvenilir bir şekilde gerçekleştirildiğine işaret etmektedir. (MSRC-12 veri setinin boyutu UTD-MHAD veri setinin yaklaşık 7 katıdır).

Tablo 12. MSRC-12 veri setinden elde edilen sonuçların literatürle karşılaştırılması

Öznitelik Müh.	Yöntemler	Doğruluk %	
		LoSubO	Cross-Subject
Hand-crafted	Cov3DJ [67]	91.7	93.6
	RRV [136]	93.8	94.7
	Hierarchical model [138]	-	94.6
	ASM [139]	-	97.6
	ELC-KSVD [135]	90.2	-
	Position offset + NBNN [112]	-	90.2
	Trajectory let [137]	94.9	95.1
	Önerilen Yöntem (BoP)	94.2	-
Learned representations	DF-selected features [61]	-	94.0(5-fold)
RNN/LSTM	ConvNets [107]	84.4	-
	JTM_CNN [108]	93.1	-
	SOS_based CNN [110]	94.2	-
	Enhanced skeleton visualization [109]	96.6	-



Şekil 52. MSRC-12 veri setinin karışıklık matris

Bu veri seti üzerinde önerilen yöntemi değerlendirmek için ikinci protokol dikkate alınmıştır. Tablo 4 göz önüne alınarak ve bu veri seti üzerindeki girdilerin tüm olası değerlerini inceleyerek, 9 zamansal ofset ile 150 anahtar poz sayısı ve 600 nöron sayısı ile en iyi performans sağlanmıştır. Şekil 52'de görüleceği üzere eylemlerde benzer pozların varlığına bağlı olarak önerilen yöntem en başarısız tanıma oranlarını sırasıyla %81 ve %86 doğrulukla "Tempoyu yükselt" ve "Müziğe İtiraz" eylemlerinde elde etmiştir.

Çalışmada yukarıda önerilen ve sonuçları verilen yöntemlerin dışında diğer birçok yöntemde denenmiştir. Önerilen yöntemin farklı aşamalarında denenen bazı yöntemler Tablo 13'de ki gibidir. Tablo da anahtar poz çıkarma, kodlama ve sınıflandırma aşamalarında denenen bazı yöntemler verilmektedir ve bu yöntemlerin hepsinde poz çantası yöntemi eylem kodlaması için kullanılmıştır. Yöntem 1'de anahtar poz çıkarma aşamasında k-means kümeleme algoritması "cityblock" (L1) uzaklık metriği, poz sınıflandırma SVM (polinom çekirdeği ile) ve anahtar poz histogramlarının sınıflandırılması için ELM yöntemi "hardlim" aktivasyon fonksiyonu ile kullanılmıştır. Yöntem 2'de anahtar poz çıkarma aşamasında k-means kümeleme algoritması "correlation" uzaklık metriği, poz sınıflandırma SVM (polinom çekirdeği ile) ve anahtar poz histogramlarının sınıflandırılması için ELM yöntemi "hardlim" aktivasyon fonksiyonu ile kullanılmıştır. Yöntem 3'de anahtar poz çıkarma aşamasında k-means kümeleme algoritması L2 uzaklık metriği, poz sınıflandırma SVM (polinom çekirdeği ile) ve anahtar poz histogramlarının sınıflandırılması için SVM (RBF çekirdeği ile) yöntemi ile kullanılmıştır. Yöntem 4'de anahtar poz çıkarma aşamasında

k-means kümeleme algoritması L2 uzaklık metriği, poz sınıflandırma ELM yöntemi (hardlim aktivasyon fonksiyonu) ve anahtar poz histogramlarının sınıflandırılması için ELM yöntemi hardlim aktivasyon fonksiyonu ile kullanılmıştır. Yöntem 5'de anahtar poz çıkarma aşamasında Fuzzy c-means kümeleme algoritması, poz sınıflandırma SVM (polinom çekirdeği ile) ve anahtar poz histogramlarının sınıflandırılması için ELM yöntemi "hardlim" aktivasyon fonksiyonu ile kullanılmıştır. Yöntem 6'de anahtar poz çıkarma aşamasında k-means kümeleme algoritması L2 uzaklık metriği, poz sınıflandırma öklid uzaklığı kullanan en yakın merkez yöntemiyle ve anahtar poz histogramlarının sınıflandırılması için ELM yöntemi hardlim aktivasyon fonksiyonu ile kullanılmıştır. Tablo 13'de verilen yöntemlerin UTKinect ve MSR Action 3D (All) veri setlerinde elde ettikleri sonuçlar ise Tablo 14'de ki gibidir. Bu sonuçlar hesaplanırken öncelikle önerilen BoP yöntemi (Önerilen BoP Yönt.) için her iki veri setinde en iyi parametre değerleri belirlenmiş, sonrasında aynı parametreler diğer tüm yöntemler (Yönt1-Yönt6) için uygulanmıştır.

Tablo 14'den görüleceği gibi en iyi sonuçların elde edildiği bu çalışmada önerilen yöntemde de anahtar poz çıkarma aşamasında k-means kümeleme algoritması L2 uzaklık metriği, poz sınıflandırma SVM (polinom çekirdeği ile) ve anahtar poz histogramlarının sınıflandırılması için ELM yöntemi "hardlim" aktivasyon fonksiyonu ile kullanılmıştır.

Tablo 13. Poz üretme ve sınıflandırma bölümlerinde uyguladığımız diğer yöntemler

Yöntem	Her adımda kullanılanlar			
	Kümeleme	Uzaklık ölçeği	Poz sınıflandırma	Hist. sınıflandırma
Yönt. 1	k-means	city block (L1)	SVM (polinom)	ELM (hardlim)
Yönt. 2	k-means	correlation	SVM (polinom)	ELM (hardlim)
Yönt. 3	k-means	L2	SVM (polinom)	SVM (rbf)
Yönt. 4	k-means	L2	ELM (hardlim)	ELM (hardlim)
Yönt. 5	Fuzzy c-means	Obj. F.	SVM (polinom)	ELM (hardlim)
Yönt. 6	k-means	L2	En yakın merkez	ELM (hardlim)
Önerilen BoP Yönt.	k-means	L2	SVM (polinom)	ELM (hardlim)

Tablo 14. Uyguladığımız diğer yöntemlerden elde edilen sonuçlar

Parametreler	Veri setler	
	UTKinect	MSR Action 3D(All)
Zaman offset	6	4
Anahtar poz sayısı	160	100
Nöron sayısı	3100	3100
Yöntem	Doğruluk	
Yönt. 1	91.3	84.0
Yönt. 2	90.9	83.1
Yönt. 3	92.9	87.1
Yönt. 4	87.8	83.2
Yönt. 5	53	48.0
Yönt. 6	93.9	88.3
Önerilen BoP Yönt.	99.0	91.9

5. TARTIŞMA VE SONUÇLAR

Bu çalışmada 3B eylem tanıma için, önceden tanımlanmış bir uzay-zamansal poz kümesine dayanan yeni bir poz çantası sistemi önerilmiştir. Literatürde poz dayalı eylem tanıma ile ilgili yapılan çalışmaların çoğunda durum tabanlı yöntemler veya poz çantası yöntemleri kullanılmıştır. Durum tabanlı yöntemlerin ana dezavantajlarından bazıları aşırı miktarda eğitim verisine ihtiyaç duyması ve parametrelerin genellikle elle ayarlanması zorluluğudur. Poz çantası yaklaşımlarının temel dezavantajı ise eylem kodlaması aşamasında pozlar arasındaki zaman kavramını dikkate almamasıdır.

Bu tez çalışmasında, zamansal bilgiler poz tanımlayıcılarına dâhil edilerek anahtar pozlar elde edilmiş ve bunların üzerinden yeni bir poz çantası yaklaşımını geliştirilmiştir. Önerilen tanımlayıcı, bir eylem dizisinde farklı zamansal sırada olan aynı iskelet konfigürasyonlarına sahip pozların ayırt edilmesini sağlamaktadır. Poz tanımlayıcısı ise iskelet eklem verileri başka bir uzaya taşınmadan öklid uzay koordinatlarında oluşturulmaktadır. Önerilen yöntemin değerlendirilmesi halka açık beş adet benchmark 3B eylem veri setinde gerçekleştirilmiştir. Üç veri setinden de elde edilen sonuçlar literatürde şu ana kadar bilinen en iyi sonucu üretirmiştir. Diğer iki veri setlerinden ise literatüre göre kıyaslanabilir sonuçlar elde edilmiştir.

Bu tez çalışmasında önerilen tüm yöntemlerin benchmark veri setlerinde elde ettiği sonuçlara dayanarak, küçük boyutlu kod sözlüklerinin (kelime veya anahtar poz) tüm eylemleri ayırt etmede yeterli sayıda farklı kod sözcüğü üretmediği görülmüştür. Diğer taraftan, kod sözcükleri büyük boyutlu olduğunda gürültüye maruz kalma eğiliminin daha fazla olduğu görülmüştür. Anahtar pozlara dayalı yöntemlerinin çoğu eylem tanımlama ve zamansal modelleme için sıklıkla HMM yöntemini kullanılmaktadır ve bu nedenle üretilen anahtar poz sayıları sınırlıdır. Önerdiğimiz yöntemde ise sözlük eğitimi aşamasında yeterince anahtar poz üretilebilmektedir. Ayrıca, anahtar pozlarının sayısının ayarlanması önerilen yöntemin tanıma sağlamlığının arttırılmasında önemli bir etkiye sahiptir ve her veri seti için ayrı bir şekilde ayarlanmıştır.

6. ÖNERİLER

Önerilen yöntemde iyileştirilmesi gereken en önemli unsur kişiler arasındaki etkileşimli eylemlerin tanınmasıdır. Bunun için yöntemin ortamdaki içerik ve nesnelerle etkileşim bilgilerinden faydalanması gerekmektedir. Bu nedenle gelecekte mevcut çalışmaların sonuçlarını iyileştirmek amacıyla hem derinlik hem de içerik bilgilerinin kullanılması gerekmektedir. Ayrıca, iskelet verilerinden faydalanarak önceden belirlenmiş eklem konumlarını göz önüne alınabilir. Bu belirlenmiş bölgele ait RGB görüntülerden bölgesel parçalar ayrılabilir ve Evrişimsel Sinir Ağları (CNN) ile derin öznitelikler çıkarılabilir. Böylelikle etkileşimli aktivite tanıma işleminde elde edilen derin özniteliklerle iskeletten çıkarılan öznitelikler birlikte kullanılabilir. Ek olarak, anahtar poz oluşturulması aşamasında Fuzzy kümeleme yöntemlerinin daha gelişmiş türleri kullanılarak daha ayrımcı pozlar elde edilebilir.

Gözetim sistemleri için gerçek zamanlı ve online tanıma yöntemleri gerekmektedir. Bu amaçla, sürekli bir video üzerinden tanınması gereken eylemlerin öncelikle eylemler içeren parçalara bölünmesi ve bu parçalar üzerinde sınıflandırma yapılaması geleceğe yönelik bir geliştirme olarak önerilebilir.

7. KAYNAKLAR

1. Aggarwal, J.K. ve Ryoo, M.S., Human activity analysis: A review, ACM Computing Surveys (CSUR), 43, 3 (2011) 1-43.

2. Poppe, R., A survey on vision-based human action recognition, Image and vision computing, 28, 6 (2010) 976-990.

3. Vishwakarma, S. ve Agrawal, A., A survey on activity recognition and behavior understanding in video surveillance, The Visual Computer, 29, 10 (2013) 983-1009.

4. Ramanathan, M., Yau, W.-Y. ve Teoh, E.K., Human action recognition with video data: research and evaluation challenges, IEEE Transactions on Human-Machine Systems, 44, 5 (2014) 650-663.

5. Turaga, P., Chellappa, R., Subrahmanian, V.S. ve Udrea, O., Machine recognition of human activities: A survey, Circuits and Systems for Video Technology, IEEE Transactions on, 18, 11 (2008) 1473-1488.

6. Ikizler, N., Understanding human motion: recognition and retrieval of human activities, Doktora Tezi, Bilkent Üniversitesi, Mühendislik ve Fen Bilimleri Enstitüsü , Ankara, 2008.

7. Herath, S., Harandi, M. ve Porikli, F., Going deeper into action recognition: A survey, Image and vision computing, 60 (2017) 4-21.

8. Muybridge, E., Animal locomotion, Da Capo Press, New York, 1887.

9. Cheng, G., Wan, Y., Saudagar, A.N., Namuduri, K. ve Buckles, B.P., Advances in human action recognition: A survey, Computer Vision and Pattern Recognition, 1 (2015) 1-30.

10. Tran, D. ve Torresani, L., EXMOVES: Mid-level Features for Efficient Action Recognition and Video Analysis, International Journal of Computer Vision, 119, 3 (2016) 239-253.

11. Eweiwi, A., Cheema, M.S., Bauckhage, C. ve Gall, J., Efficient pose-based action recognition, Asian Conference on Computer Vision 2014, Kasım 2014, Singapore, Bildiriler Kitabı: 428-443.

12. Dawn, D.D. ve Shaikh, S.H., A comprehensive survey of human action recognition with spatio-temporal interest point (STIP) detector, The Visual Computer, 32, 3 (2016) 289-306.

13. Bobick, A.F. ve Davis, J.W., The recognition of human movement using temporal templates, IEEE Transactions on pattern analysis and machine intelligence, 23, 3 (2001) 257-267.

14. Laptev, I. ve Lindeberg, T., Space-time interest points, Proceedings Ninth IEEE International Conference on Computer Vision, Ekim 2003, Nice, Bildiriler Kitabı: 432-439

15. Laptev, I., Marszalek, M., Schmid, C. ve Rozenfeld, B., Learning realistic human actions from movies, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Haziran 2008, Anchorage, Bildiriler Kitabı : 1-8.

16. Scovanner, P., Ali, S. ve Shah, M., A 3-dimensional sift descriptor and its application to action recognition, Proceedings of the 15th ACM international conference on Multimedia 2007, Eylül 2007, Augsburg, Bildiriler Kitabı: 357-360.

17. Klaser, A., Marszałek, M. ve Schmid, C., A spatio-temporal descriptor based on 3d-gradients, BMVC 2008-19th British Machine Vision Conference 2008, Eylül 2008, Leeds, Bildiriler Kitabı: 271-210.

18. Yang, X., Zhang, C. ve Tian, Y., Recognizing actions using depth motion maps-based histograms of oriented gradients, Proceedings of the 20th ACM international conference on Multimedia, Eylül 2012, Nara, Bildiriler Kitabı: 1057-1060.

19. Kovashka, A. ve Grauman, K., Learning a hierarchy of discriminative space-time neighborhood features for human action recognition, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Haziran 2010, San Francisco, Bildiriler Kitabı: 2046-2053.

20. Wang, H., Kläser, A., Schmid, C. ve Liu, C.-L., Action recognition by dense trajectories, 2011 IEEE Conference on Computer Vision and Pattern Recognition, Haziran 2011, Washington, Bildiriler Kitabı: 3169-3176.

21. Wang, H., Kläser, A., Schmid, C. ve Liu, C.-L., Dense trajectories and motion boundary descriptors for action recognition, International Journal of Computer Vision, 103, 1 (2013) 60-79.

22. Sadanand, S. ve Corso, J.J., Action bank: A high-level representation of activity in video, 2012 IEEE Conference on Computer Vision and Pattern Recognition, Haziran 2012, Providence, Bildiriler Kitabı:1234-1241.

23. Shotton, J., Sharp, T., Kipman, A., Fitzgibbon, A., Finocchio, M., Blake, A., Cook, M. ve Moore, R., Real-time human pose recognition in parts from single depth images, Communications of the ACM, 56, 1 (2013) 116-124.

24. Han, J., Shao, L., Xu, D. ve Shotton, J., Enhanced computer vision with microsoft kinect sensor: A review, IEEE transactions on cybernetics, 43, 5 (2013) 1318-1334.

25. Zelnik-Manor, L. ve Irani, M., Event-based analysis of video, Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on 2001, Aralık 2001, Rehovot, Bildiriler Kitabı II: 123-130.

26. Gong, W., Bagdanov, A.D., Roca, F.X. ve Gonzàlez, J., Automatic key pose selection for 3d human action recognition, International Conference on Articulated Motion and Deformable Objects 2010, Temmuz 2010, Heidelberg, Bildiriler Kitabı: 290-299.

27. LeCun, Y., Bengio, Y. ve Hinton, G., Deep learning, Nature, 521, 7553 (2015) 436-444.

28. Veeriah, V., Zhuang, N. ve Qi, G.-J., Differential recurrent neural networks for action recognition, Proceedings of the IEEE International Conference on Computer Vision 2015, Aralık 2015, Chile, Bildiriler Kitabı: 4041-4049.

29. Du, Y., Wang, W. ve Wang, L., Hierarchical recurrent neural network for skeleton based action recognition, Proceedings of the IEEE conference on computer vision and pattern recognition 2015, Aralık 2015, Chile, Bildiriler Kitabı: 1110-1118.

30. Zhu, W., Lan, C., Xing, J., Zeng, W., Li, Y., Shen, L. ve Xie, X., Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks, AAAI 2016, Şubat 2016, Arizona, Bildiriler Kitabı: 3697-3703.

31. Lee, I., Kim, D., Kang, S. ve Lee, S., Ensemble Deep Learning for Skeleton-Based Action Recognition Using Temporal Sliding LSTM Networks, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017, Ekim 2017, Venice, Bildiriler Kitabı: 1012-1020.

32. Liu, J., Shahroudy, A., Xu, D., Chichung, A.K. ve Wang, G., Skeleton-Based Action Recognition Using Spatio-Temporal LSTM Network with Trust Gates, IEEE Transactions on pattern analysis and machine intelligence, 99 (2017) 1-1.

33. Luvizon, D.C., Tabia, H. ve Picard, D., Learning features combination for human action recognition from skeleton sequences, Pattern Recognition Letters, 99 (2017) 13-20.

34. Han, F., Reily, B., Hoff, W. ve Zhang, H., Space-time representation of people based on 3D skeletal data: A review, Computer Vision and Image Understanding, 158 (2017) 85-105.

35. Wang, C., Wang, Y. ve Yuille, A.L., An approach to pose-based action recognition, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2013, Haziran 2013, Portland, Bildiriler Kitabı: 915-922.

36. Agahian, S., Negin, F. ve Köse, C., Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition, The Visual Computer, (2018) 1-17.

37. Huang, G.-B., Zhu, Q.-Y. ve Siew, C.-K., Extreme learning machine: theory and applications, Neurocomputing, 70, 1 (2006) 489-501.

38. Bobick, A.F. ve Davis, J.W., The recognition of human movement using temporal templates, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 23, 3 (2001) 257-267.

39. Johansson, G., Visual motion perception, Scientific American, 232, 6 (1975) 76-88.

40. Sheikh, Y., Sheikh, M. ve Shah, M., Exploring the space of a human action, ICCV '05 Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Ekim 2005, Washington, Bildiriler Kitabı: 144-149.

41. Pirsiavash, H. ve Ramanan, D., Parsing videos of actions with segmental grammars, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition , Haziran 2014, Columbus, Bildiriler Kitabı: 612-619.

42. Ryoo, M.S. ve Aggarwal, J.K., Semantic representation and recognition of continued and recursive human activities, International Journal of Computer Vision, 82, 1 (2009) 1-24.

43. Aggarwal, J. ve Xia, L., Human activity recognition from 3d data: A review, Pattern Recognition Letters, 48 (2014) 70-80.

44. Talu, M.F., İnsan Hareketlerinin Takibinde Karşılaşılan Problemlerin Çözümüne Yeni Yaklaşımlar, Firat Üniversites, Fen Bilimler Enstitüsü, Elazığ, 2010.

45. Shao, L., Han, J., Kohli, P. ve Zhang, Z., Computer vision and machine learning with RGB-D sensors, Springer, Heidelburg, 2014.

46. Girshick, R., Shotton, J., Kohli, P., Criminisi, A. ve Fitzgibbon, A., Efficient regression of general-activity human poses from depth images, 2011 IEEE International Conference on Computer Vision, Kasım 2011, Washington, Bildiriler Kitabı: 415-422.

47. Clark, R.A., Pua, Y.-H., Fortin, K., Ritchie, C., Webster, K.E., Denehy, L. ve Bryant, A.L., Validity of the Microsoft Kinect for assessment of postural control, Gait & posture, 36, 3 (2012) 372-377.

48. Wasenmüller, O. ve Stricker, D., Comparison of kinect v1 and v2 depth images in terms of accuracy and precision, Asian Conference on Computer Vision 2016: 34-45.

49. Chéron, G., Laptev, I. ve Schmid, C., P-cnn: Pose-based cnn features for action recognition, Proceedings of the IEEE International Conference on Computer Vision 2015, Aralık 2015, Santiago, Bildiriler Kitabı:3218-3226.

50. Cao, Z., Simon, T., Wei, S.-E. ve Sheikh, Y., Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields, Computer Vision and Pattern Recognition, 1, 2 (2016) 7291-7299.

51. Luvizon, D.C., Picard, D. ve Tabia, H., 2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning, Computer Vision and Pattern Recognition, 1 (2018) 1-12.

52. Mehta, D., Sridhar, S., Sotnychenko, O., Rhodin, H., Shafiei, M., Seidel, H.-P., Xu, W., Casas, D. ve Theobalt, C., Vnect: Real-time 3d human pose estimation with a single rgb camera, ACM Transactions on Graphics (TOG), 36, 4 (2017) 1-14.

53. Rogez, G. ve Schmid, C., Image-based synthesis for deep 3d human pose estimation, International Journal of Computer Vision, (2018) 1-16.

54. Yao, A., Gall, J., Fanelli, G. ve Van Gool, L., Does human action recognition benefit from pose estimation?, Proceedings of the 22nd British machine vision conference-BMVC 2011, Ağustos 2011, Dundee, 67.1- 67.

55. Xia, L., Chen, C.-C. ve Aggarwal, J., View invariant human action recognition using histograms of 3d joints, Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on 2012, Haziran 2012, Providence, Bildiriler Kitabı: 20-27.

56. Amor, B.B., Su, J. ve Srivastava, A., Action recognition using rate-invariant analysis of skeletal shape trajectories, IEEE Transactions on pattern analysis and machine intelligence, 38, 1 (2016) 1-13.

57. Chaaraoui, A.A., Padilla-López, J.R., Climent-Pérez, P. ve Flórez-Revuelta, F., Evolutionary joint selection to improve human action recognition with RGB-D devices, Expert systems with applications, 41, 3 (2014) 786-794.

58. Vemulapalli, R., Arrate, F. ve Chellappa, R., Human action recognition by representing 3d skeletons as points in a lie group, Proceedings of the IEEE conference on computer vision and pattern recognition 2014, Haziran 2014, Columbus, Bildiriler Kitabı: 588-595.

59. Lillo, I., Niebles, J.C. ve Soto, A., Sparse composition of body poses and atomic actions for human activity recognition in RGB-D videos, Image and vision computing, 59 (2017) 63-75.

60. Zanfir, M., Leordeanu, M. ve Sminchisescu, C., The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection, Proceedings of the IEEE International Conference on Computer Vision 2013, Aralık 2013, Washington, Bildiriler Kitabı: 2752-2759.

61. Negin, F., Özdemir, F., Akgül, C.B., Yüksel, K.A. ve Erçil, A., A decision forest based feature selection framework for action recognition from rgb-depth cameras, International Conference Image Analysis and Recognition 2013, Haziran 2013, Heidelberg, Bildiriler Kitabı: 648-657.

62. Chen, X. ve Koskela, M., Online RGB-D gesture recognition with extreme learning machines, Proceedings of the 15th ACM on International conference on multimodal interaction 2013, Aralık 2013, New York, Bildiriler Kitabı: 467-474.

63. Wang, J., Liu, Z., Wu, Y. ve Yuan, J., Mining actionlet ensemble for action recognition with depth cameras, Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on 2012, Ekim 2012, Providence, Bildiriler Kitabı:1290-1297.

64. Wang, J., Liu, Z. ve Wu, Y., Human Action Recognition with Depth Cameras, Springer, Cham, 2014.

65. Rahmani, H., Mahmood, A., Huynh, D.Q. ve Mian, A., Real time action recognition using histograms of depth gradients and random decision forests, 2014 IEEE Winter Conference on Applications of Computer Vision, Mart 2014, Steamboat Springs, Bildiriler Kitabı: 626-633.

66. Yang, X. ve Tian, Y., Effective 3d action recognition using eigenjoints, Journal of Visual Communication and Image Representation, 25, 1 (2014) 2-11.

67. Hussein, M.E., Torki, M., Gowayyed, M.A. ve El-Saban, M., Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations, Twenty-Third International Joint Conference on Artificial Intelligence 2013, Ağustos 2013, Beijing, Bildiriler Kitabı: 2466-2472.

68. Wei, P., Zheng, N., Zhao, Y. ve Zhu, S.-C., Concurrent action detection with structural prediction, 2013 IEEE International Conference on Computer Vision, Aralık 2013, Sydney, Bildiriler Kitabı: 3136-3143.

69. Yu, G., Liu, Z. ve Yuan, J., Discriminative orderlet mining for real-time recognition of human-object interaction, Asian Conference on Computer Vision 2014, Kasım 2014, Singapore, Bildiriler Kitabı: 50-65.

70. Fothergill, S., Mentis, H., Kohli, P. ve Nowozin, S., Instructing people for training gestural interactive systems, Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, Mayıs 2012, Austin, Bildiriler Kitabı: 1737-1746.

71. Boubou, S. ve Suzuki, E., Classifying actions based on histogram of oriented velocity vectors, Journal of Intelligent Information Systems, 44, 1 (2015) 49-65.

72. Gowayyed, M.A., Torki, M., Hussein, M.E. ve El-Saban, M., Histogram of Oriented Displacements (HOD): Describing Trajectories of Human Joints for Action Recognition, IJCAI 2013, Ağustos 2013, Beijing, Bildiriler Kitabı: 1351-1357.

73. Zhang, H. ve Parker, L.E., Bio-inspired predictive orientation decomposition of skeleton trajectories for real-time human activity prediction, 2015 IEEE International Conference on Robotics and Automation (ICRA), Mayıs 2015, Seattle, Bildiriler Kitabı: 3053-3060.

74. Kapsouras, I. ve Nikolaidis, N., Action recognition on motion capture data using a dynemes and forward differences representation, Journal of Visual Communication and Image Representation, 25, 6 (2014) 1432-1445.

75. Wang, C., Wang, Y. ve Yuille, A.L., An approach to pose-based action recognition, 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Haziran 2013, Portland, Bildiriler Kitabı: 915-922.

76. Tao, L. ve Vidal, R., Moving poselets: A discriminative and interpretable skeletal motion representation for action recognition, Proceedings of the IEEE International Conference on Computer Vision Workshops 2015, Aralık 2015, Chile, Bildiriler Kitabı: 61-69.

77. Nie, B.X., Xiong, C. ve Zhu, S.-C., Joint action recognition and pose estimation from video, 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Haziran 2015, Boston, Bildiriler Kitabı: 1293-1301.

78. Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R. ve Vidal, R., Bio-inspired dynamic 3d discriminative skeletal features for human action recognition, 2013 IEEE Conference on Computer Vision and Pattern Recognition, Haziran 2013, Portland, 471-478.

79. Chen, X. ve Koskela, M., Skeleton-based action recognition with extreme learning machines, Neurocomputing, 149 (2015) 387-396.

80. Yamato, J., Ohya, J. ve Ishii, K., Recognizing human action in time-sequential images using hidden markov model, Proceedings. 1992 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Haziran 1992, Champaign, Bildiriler Kitabı: 379-385.

81. Lehrmann, A.M., Gehler, P.V. ve Nowozin, S., Efficient nonlinear markov models for human motion, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014, Haziran 2014, Columbus, Bildiriler Kitabı: 1314-1321.

82. Lv, F. ve Nevatia, R., Recognition and segmentation of 3-d human action using hmm and multi-class adaboost, European conference on computer vision 2006, Mayıs 2006, Graz, Bildiriler Kitabı: 359-372.

83. Ding, W., Liu, K., Fu, X. ve Cheng, F., Profile HMMs for skeleton-based human action recognition, Signal Processing: Image Communication, 42 (2016) 109-119.

84. Pazhoumand-Dar, H., Lam, C.-P. ve Masek, M., Joint movement similarities for robust 3D action recognition using skeletal data, Journal of Visual Communication and Image Representation, 30 (2015) 10-21.

85. Jiang, Y.-G., Li, Z. ve Chang, S.-F., Modeling scene and object contexts for human action retrieval with few examples, IEEE Transactions on Circuits and Systems for Video Technology, 21, 5 (2011) 674-681.

86. Vieira, T., Faugeroux, R., Martínez, D. ve Lewiner, T., Online human moves recognition through discriminative key poses and speed-aware action graphs, Machine Vision and Applications, 28,1-2 (2017) 185-200.

87. Reyes, M., Dominguez, G. ve Escalera, S., Featureweighting in dynamic timewarping for gesture recognition in depth data, 2011 IEEE International Conference on Computer Vision, Kasım 2011, Barcelona, Bildiriler Kitabı: 1182-1188.

88. Sempena, S., Maulidevi, N.U. ve Aryan, P.R., Human action recognition using dynamic time warping, 2011 International Conference on Electrical Engineering and Informatics, Temmuz 2011, Bandung, Bildiriler Kitabı: 1-5.

89. Tang, J., Cheng, H., Zhao, Y. ve Guo, H., Structured Dynamic Time Warping for Continuous Hand Trajectory Gesture Recognition, Pattern Recognition, 80 (2018) 21-31.

90. Chung, H. ve Yang, H.-D., Conditional random field-based gesture recognition with depth information, Optical Engineering, 52, 1 (2013) 0172011-7.

91. Berndt, D.J. ve Clifford, J., Using dynamic time warping to find patterns in time series, KDD workshop 1994, Temmuz 1994 , Washington, Bildiriler Kitabı: 359-370.

92. Müller, M., Information retrieval for music and motion, 2, Springer, Verlag Berlin Heidelberg, 2007.

93. Adistambha, K., Ritz, C.H. ve Burnett, I.S., Motion classification using dynamic time warping, 2008 IEEE 10th Workshop on Multimedia Signal Processing, Ekim 2008, Cairns, Bildiriler Kitabı: 622-627.

94. Deng, L., Leung, H., Gu, N. ve Yang, Y., Automated recognition of sequential patterns in captured motion streams, International Conference on Web-Age Information Management 2010, Temmuz 2010, Jiuzhaigou, Bildiriler Kitabı: 250-261.

95. Raptis, M., Kirovski, D. ve Hoppe, H., Real-time classification of dance gestures from skeleton animation, Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation, Ağustos 2011,Vancouver, Bildiriler Kitabı: 147-156.

96. Chen, X. ve Koskela, M., Sequence alignment for RGB-D and motion capture skeletons, International Conference Image Analysis and Recognition 2013: Bildiriler Kitabı:630-639.

97. Lai, K., Konrad, J. ve Ishwar, P., A gesture-driven computer interface using Kinect, Image Analysis and Interpretation (SSIAI), Nisan 2012, Santa Fe, Bildiriler Kitabı: 185-188.

98. Yang, X. ve Tian, Y.L., Eigenjoints-based action recognition using naive-bayes-nearest-neighbor, 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2012), Haziran 2012, Providence, Bildiriler Kitabı II: 14-19.

99. Wang, J.-Y. ve Lee, H.-M., Recognition of human actions using motion capture data and support vector machine, 2009 WRI World Congress on Software Engineering, Mayıs 2009, Xiamen, Bildiriler Kitabı: 234-238.

100. Jaakkola, T. ve Haussler, D., Exploiting generative models in discriminative classifiers, Advances in neural information processing systems (1999): 487-493.

101. Perronnin, F. ve Dance, C., Fisher kernels on visual vocabularies for image categorization, Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on 2007, Bildiriler Kitabı: 1-8.

102. Varol, G. ve Salah, A.A., Efficient large-scale action recognition in videos using extreme learning machines, Expert systems with applications, 42, 21 (2015) 8274-8282.

103. Niebles, J.C., Wang, H. ve Fei-Fei, L., Unsupervised learning of human action categories using spatial-temporal words, International Journal of Computer Vision, 79,3 (2008) 299-318.

104. Peng, X., Wang, L., Wang, X. ve Qiao, Y., Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice, Computer Vision and Image Understanding, 150 (2016) 109-125.

105. Oneata, D., Verbeek, J. ve Schmid, C., Action and event recognition with fisher vectors on a compact feature set, Computer Vision (ICCV), 2013 IEEE International Conference on 2013, Bildiriler Kitabı:1817-1824.

106. Evangelidis, G., Singh, G. ve Horaud, R., Skeletal quads: Human action recognition using joint quadruples, Pattern Recognition (ICPR), 2014 22nd International Conference on 2014, Bildiriler Kitabı: 4513-4518.

107. Du, Y., Fu, Y. ve Wang, L., Skeleton based action recognition with convolutional neural network, The 3rd IAPR Asian Conference on Pattern Recognition,Kasım 2015, Kuala Lumpur, Bildiriler Kitabı: 579-583.

108. Wang, P., Li, Z., Hou, Y. ve Li, W., Action recognition based on joint trajectory maps using convolutional neural networks, Proceedings of the 2016 ACM on Multimedia Conference, Ekim 2016, Amsterdam, Bildiriler Kitabı: 102-106.

109. Liu, M., Liu, H. ve Chen, C., Enhanced skeleton visualization for view invariant human action recognition, Pattern Recognition, 68 (2017) 346-362.

110. Hou, Y., Li, Z., Wang, P. ve Li, W., Skeleton optical spectra based action recognition using convolutional neural networks, IEEE Transactions on Circuits and Systems for Video Technology, (2016).

111. Liu, M., Chen, C. ve Liu, H., Learning informative pairwise joints with energy-based temporal pyramid for 3D action recognition, 2017 IEEE International Conference on Multimedia and Expo, Temmuz 2017, Hong Kong, Bildiriler Kitabı: 901-906.

112. Lu, G., Zhou, Y., Li, X. ve Kudo, M., Efficient action recognition via local position offset of 3D skeletal body joints, Multimedia Tools and Applications, 75, 6 (2016) 3479-3494.

113. Vemulapalli, R., Arrate, F. ve Chellappa, R., R3DG features: Relative 3D geometry-based skeletal representations for human action recognition, Computer Vision and Image Understanding, 152 (2016) 155-166.

114. Wang, C., Wang, Y. ve Yuille, A.L., Mining 3D Key-Pose-Motifs for Action Recognition, 2016 IEEE Conference on Computer Vision and Pattern Recognition, Haziran 2016, Seattle, Bildiriler Kitabı: 2639-2647.

115. Gaglio, S., Re, G.L. ve Morana, M., Human activity recognition process using 3-D posture data, Human-Machine Systems, IEEE Transactions on, 45, 5 (2015) 586-597.

116. Zhu, F., Shao, L., Xie, J. ve Fang, Y., From handcrafted to learned representations for human action recognition: a survey, Image and vision computing, 55 (2016) 42-52.

117. Vedaldi, A. ve Fulkerson, B., VLFeat: An open and portable library of computer vision algorithms, Proceedings of the 18th ACM international conference on Multimedia 2010, Bildiriler Kitabı: 1469-1472.

118. Chang, C.-C. ve Lin, C.-J., LIBSVM: a library for support vector machines, ACM Transactions on Intelligent Systems and Technology (TIST), 2, 3 (2011) 1-27.

119. Ding, S., Zhao, H., Zhang, Y., Xu, X. ve Nie, R., Extreme learning machine: algorithm, theory and applications, Artificial Intelligence Review, 44, 1 (2015) 103-115.

120. Minhas, R., Baradarani, A., Seifzadeh, S. ve Wu, Q.J., Human action recognition using extreme learning machine based on visual vocabularies, Neurocomputing, 73, 10 (2010) 1906-1917.

121. Youssef, C., Spatiotemporal representation of 3D skeleton joints-based action recognition using modified spherical harmonics, Pattern Recognition Letters, 83 (2016) 32-41.

122. Sung, J., Ponce, C., Selman, B. ve Saxena, A., Unstructured human activity detection from rgbd images, 2012 IEEE International Conference on Robotics and Automation, Mayıs 2012, St Paul, Bildiriler Kitabı: 842-849.

123. Shan, J. ve Akella, S., 3D human action segmentation and recognition using pose kinetic energy, Advanced Robotics and its Social Impacts (ARSO), 2014 IEEE Workshop on 2014, Bildiriler Kitabı: 69-75.

124. Parisi, G.I., Weber, C. ve Wermter, S., Self-organizing neural integration of pose-motion features for human action recognition, Frontiers in neurorobotics, 9, 3 (2015) 1-14.

125. Chen, C., Jafari, R. ve Kehtarnavaz, N., Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor, 2015 IEEE International Conference on Image Processing, Eylül 2005, Quebec, Bildiriler Kitabı:168-172.

126. Li, W., Zhang, Z. ve Liu, Z., Action recognition based on a bag of 3d points, 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, Haziran 2010, Bildiriler Kitabı:9-14.

127. Zhang, J., Li, W., Ogunbona, P.O., Wang, P. ve Tang, C., RGB-D-based action recognition datasets: A survey, Pattern Recognition, 60 (2016) 86-105.

128. Negin, F., Akgül, C.B., Yüksel, K.A. ve Erçil, A., An RDF-based action recognition framework with feature selection capability, considering therapy exercises utilizing depth cameras, Journal of Theoretical and Applied Computer Science, 8, 3 (2014) 3-22.

129. Yang, Y., Deng, C., Tao, D., Zhang, S., Liu, W. ve Gao, X., Latent Max-Margin Multitask Learning With Skelets for 3-D Action Recognition, IEEE transactions on cybernetics, 47, 2 (2017) 439-448.

130. Zhang, S., Liu, X. ve Xiao, J., On geometric features for skeleton-based action recognition using multilayer LSTM networks, 2017 IEEE Winter Conference on Applications of Computer Vision, Mart 2017, Santa Rosa, Bildiriler Kitabı:148-157.

131. Zhu, Y., Chen, W. ve Guo, G., Fusing multiple features for depth-based action recognition, ACM Transactions on Intelligent Systems and Technology (TIST), 6, 2 (2015) 1-20.

132. Zhu, G., Zhang, L., Shen, P. ve Song, J., Human action recognition using multi-layer codebooks of key poses and atomic motions, Signal Processing: Image Communication, 42 (2016) 19-30.

133. Nunes, U.M., Faria, D.R. ve Peixoto, P., A human activity recognition framework using max-min features and key poses with differential evolution random forests classifier, Pattern Recognition Letters, 99, 1 (2017) 21-31.

134. Chen, C., Jafari, R. ve Kehtarnavaz, N., A real-time human action recognition system using depth and inertial sensor fusion, IEEE Sensors Journal, 16, 3 (2016) 773-781.

135. Zhou, L., Li, W., Zhang, Y., Ogunbona, P., Nguyen, D.T. ve Zhang, H., Discriminative key pose extraction using extended lc-ksvd for action recognition, Digital lmage Computing: Techniques and Applications (DlCTA), 2014 International Conference on 2014,Kasım 2014, Wollongong, Bildiriler Kitabı: 1-8.

136. Guo, Y., Li, Y. ve Shao, Z., RRV: A Spatiotemporal Descriptor for Rigid Body Motion Recognition, IEEE transactions on cybernetics, 99 (2017) 1-13.

137. Qiao, R., Liu, L., Shen, C. ve van den Hengel, A., Learning discriminative trajectorylet detector sets for accurate skeleton-based action recognition, Pattern Recognition, 66 (2017) 202-212.

138. Jiang, X., Zhong, F., Peng, Q. ve Qin, X., Online robust action recognition based on a hierarchical model, The Visual Computer, 30,9 (2014) 1021-1033.

139. Ibañez, R., Soria, Á., Teyseyre, A., Rodríguez, G. ve Campo, M., Approximate string matching: A lightweight approach to recognize gestures with Kinect, Pattern Recognition, 62 (2017) 73-86.

ÖZGEÇMİŞ

Saeid AGAHİAN 1983 yılında İran'nin Bonab (Binab, Binev Azeri Türkçesinde) Şehrinde doğdu. İlköğrenimini 1990 yılında Bonab'da tamamladıktan sonra 2002 yılında Bonab'ın Şeykh behayi lisesinden mezun oldu. 2002 yılında Peyam Noor Üniversitesi Bilgisayar Mühendisliği Bölümü'ne girmeye hak kazanmış ve öğrenimini burada 2006 yılında tamamlanmış ve 2010 yılında Karadeniz Teknik Üniversitesi Fen Bilimleri Bilgisayar Mühendisliği Anabilim Dalında Yüksek Lisans öğrenimine başlamıştır ve 2012 haziran ayında tamamlanmış. Daha sonra 2013 yılının güz döneminde Karadeniz Teknik Üniversitesi Fen Bilimleri Bilgisayar Mühendisliği Anabilim Dalında Doktora öğrenimine başlamıştır. Evli olan Saeid AGAHIAN iyi derecede Azerice, Türkçe, Farsça ve orta derecede İngilizce bilmektedir. Yayınları aşağıda verilmiştir;

SCI / SCI-expanded İndeksli Dergi Yayınları

S. Agahian, F. Negin, and C. Köse, "Improving bag-of-poses with semi-temporal pose descriptors for skeleton-based action recognition," *The Visual Computer,Springer,* February 21, 2018.page 1-17.

Uluslararası Konferans Yayınları

S. Agahian, F. Negin, and C. Köse, "An Efficient Human Action Recognition Framework with Pose-based Spatiotemporal Features" International Conference on Advanced Technologies, Computer Engineering and Science (ICATCES 2018), Safranbolu, Turkey, May 2018.