**BURSA TEKNİK ÜNİVERSİTESİ ❖ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

# 3-BOYUTLU ARTIK AĞ EYLEM TANIMA MODELİ İLE SÜPERMARKET VİDEO GÖRÜNTÜLERİNDE HIRSIZLIK TESPİTİ

**YÜKSEK LİSANS TEZİ**

**Emirhan TOPRAK**

**Elektrik-Elektronik Mühendisliği Anabilim Dalı**

**Elektrik-Elektronik Programı**

**MAYIS 2022**

**BURSA TEKNİK ÜNİVERSİTESİ ❖ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

# 3-BOYUTLU ARTIK AĞ EYLEM TANIMA MODELİ İLE SÜPERMARKET VİDEO GÖRÜNTÜLERİNDE HIRSIZLIK TESPİTİ

**YÜKSEK LİSANS TEZİ**

**Emirhan TOPRAK**
**(19278034073)**

**Elektrik-Elektronik Mühendisliği Anabilim Dalı**

**Elektrik-Elektronik Programı**

**Tez Danışmanı: Dr. Öğr. Üyesi MUSTAFA ÖZDEN**

**MAYIS 2022**

BTÜ, Lisansüstü Eğitim Enstitüsü'nün 19278034073 numaralı Yüksek Lisans Öğrencisi Emirhan TOPRAK, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "3-BOYUTLU ARTIK AĞ EYLEM TANIMA MODELİ İLE SÜPERMARKET VİDEO GÖRÜNTÜLERİNDE HIRSIZLIK TESPİTİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

**Tez Danışmanı :** **Dr. Öğr. Üyesi MUSTAFA ÖZDEN** .....
Bursa Teknik Üniversitesi

**Jüri Üyeleri :** **Dr. Öğr. Üyesi Fatmatülzehra USLU** .....
Bursa Teknik Üniversitesi

**Dr. Öğr. Üyesi Savaş TAKAN** .....
Bursa Uludağ Üniversitesi

**Teslim Tarihi :**
**Savunma Tarihi : 13 Mayıs 2022**

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, Bursa Teknik Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Lisansüstü Eğitim Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

## İNTİHAL BEYANI

Bu tezde görsel, işitsel ve yazılı biçimde sunulan tüm bilgi ve sonuçların akademik ve etik kurallara uyularak tarafımdan elde edildiğini, tez içinde yer alan ancak bu çalışmaya özgü olmayan tüm sonuç ve bilgileri tezde kaynak göstererek belgelediğimi, aksinin ortaya çıkması durumunda her türlü yasal sonucu kabul ettiğimi beyan ederim.

Öğrencinin Adı Soyadı: Emirhan TOPRAK

İmzası :

***Aileme,***

## ÖNSÖZ

Derslerim ve tez çalışmam boyunca her konuda yol gösteren, yardım ve desteklerini benden esirgemeyen saygıdeğer hocam Sayın Dr. Öğr. Üyesi Mustafa ÖZDEN'e içten teşekkürlerimi sunarım.

Desteğini esirgemeyen ve her konuda yanımda olan aileme saygı ve sevgilerimi sunarım.

Mayıs 2022 Emirhan Toprak

**İÇİNDEKİLER**

**Sayfa**

## KISALTMALAR

**1-B** : 1-Boyutlu
**2-B** : 2-Boyutlu
**3-B** : 3-Boyutlu
**ABD** : Amerika Birleşik Devletleri
**BN** : Toplu Normalleştirme
**BoKP** : Bags of Key Points
**BoVW** : Bag of Visual Words
**BoW** : Bag of Words
**C3D** : Konvolüsyonel 3-B
**CNN** : Evrişimsel Sinir Ağı
**CSN** : Kanalla Ayrılmış Evrişimsel Ağlar
**DT** : Dense Trajectory
**ET** : Eylem Tanıma
**FC** : Tam Bağlantılı
**FV** : Fisher Vectors
**GMM** : Gaussian Mixture Model
**I3D** : Şişirilmiş İki-Akış 3-B
**IDT** : Improved Dense Trajectories
**P3D** : Yalancı-3B
**ReLu** : Rectified Linear Units
**RGB** : Red Green Blue
**ROI** : İlgi Bölgesi
**S3D** : Ayrılabilir 3-B
**SGD** : Stokastik Gradyan İnişi
**SVM** : Support Vector Machine
**TSN** : Geçici Segment Ağları
**VLAD** : Vector of Locally Aggregated Descriptors

**SEMBOLLER**

| | |
|---|---|
| **C** | **:** Kanal sayısı |
| ***c*** | **:** Toplam sınıf sayısı |
| **CE** | **:** Çapraz entropi kaybı |
| $f(s_i)$ | **:** Softmax aktivasyon fonksiyonu olasılık sınıf skoru |
| $F(x)$ | **:** Artık öğrenme |
| **H** | **:** Görüntü yüksekliği |
| $H(x)$ | **:** Gerçek çıktı |
| ***i*** | **:** Tek sınıf |
| **k x k x k** | **:** Çekirdek konvolüsyon filtresi |
| $s_i, t_i$ | **:** Tüm sınıftaki her bir sınıf için mimarinin oluşturduğu skor |
| **T** | **:** Çerçeve uzunluğu |
| **W** | **:** Görüntü genişliği |
| $W_1, W_2$ | **:** Ağırlık katmanı |
| $W_i$ | **:** Ağırlık katmanları |
| ***x*** | **:** Girdi |
| $\sigma$ | **:** ReLu aktivasyon fonksiyonu |

## ÇİZELGE LİSTESİ

**Sayfa**

## ŞEKİL LİSTESİ

**Sayfa**

# 3-BOYUTLU ARTIK AĞ EYLEM TANIMA MODELİ İLE SÜPERMARKET VİDEO GÖRÜNTÜLERİNDE HIRSIZLIK TESPİTİ

## ÖZET

Son zamanlarda, süpermarketlerde hırsızlık tespiti için yapay zekâ modellerine ilgi artmaktadır. Süpermarket hırsızlıkları, süpermarketleri finansal açıdan marketleri zarara sokmaktadır. Bu zararların önüne geçmek için insan hırsızlık eylemine yönelik modeller geliştirilmektedir.

Gündelik olarak gerçekleştirilen insan eylemlerini sınıflandırmak için 2-B CNN ve 3-B CNN eylem tanıma modelleri kullanılmaktadır. İnsan eylemi gerçekleştirilirken hem görsel hem de hareket bilgisi içermektedir. Görsel ve hareket bilgisi, uzam-zamansal bilgiyi ifade etmektedir. Eylem tanıma modelleri ile uzam-zamansal eylem bilgisi çıkarılmaktadır.

Bu tez çalışmasında, hırsızlık ve hırsızlık olmayan eylem video veri seti oluşturulmuştur. Oluşturulan hırsızlık ve hırsızlık olmayan eylem veri seti videoları 3'er eylemden oluşmaktadır. Hırsızlık eylemleri: eşyaları; cebe koymak, çantaya koymak ve el çantasına koymak şeklindedir. Hırsızlık olmayan eylemler: süpermarkette; yürümek, sabit durmak ve raftan eşya almak şeklindedir.

Eğitim veri seti Youtube'den toplanmış ve test veri seti ise bir süpermarket güvenlik kamerasından toplanmıştır. Eğitim veri seti, 161 hırsızlık olmayan eylem, 139 hırsızlık eylemi olarak 300 videodan oluşmaktadır. Test veri seti, 140 hırsızlık olmayan eylem, 130 hırsızlık eylemi olarak 270 videodan oluşmaktadır.

3-B CNN modellerini sıfırdan optimize etmek için büyük ölçekli veri setleri gerekmekte aksi halde ağın doğruluk oranı hızla düşmektedir. Oluşturulan hırsızlık eğitim veri seti küçük ölçekli olduğu için büyük ölçekli Kinetics-700 veri setinde önceden eğitilmiş olan 18 katmanlı 3-B Artık Ağ modeli transfer öğrenme ile kullanılmıştır. Temel alınan 3-B Artık Ağ modelinin FC katmanı dışındaki ağırlıkları kullanılmış ve model sadece FC katmanı ağırlıkları güncellenerek eğitilmiştir.

Hırsızlık eylemini daha detaylı incelemek ve sınıflandırmak için modele ait 12 versiyon oluşturulmuştur. Oluşturulan versiyonlar, parametre olarak birbirinden farklıdır. Versiyonlar, girdi görüntüsü boyutu, çerçeve uzunluğu ve parti büyüklüğü olarak farklılık göstermektedir. Versiyonlar; RGB girdi görüntüsü almakta ve 200 adımda eğitilmiştir. Elde edilen eğitim ve test sonuçları doğruluk oranları bakımından karşılaştırılmıştır.

Eğitim ve test sonuçları neticesinde Versiyon 1 sırasıyla, %88,0 ve %77,0 doğruluk oranları ile en iyi sonuca sahip modeldir. Versiyon 1: 224×224×3 RGB girdi görüntüsü, 32 çerçeve uzunluğu ve 12 parti büyüklüğüne sahiptir. Süpermarkette hırsızlık tespiti yapabilen 18 katmanlı 3-B Artık Ağ modeli geliştirilmiştir.

**Anahtar kelimeler:** Hırsızlık Tespiti, Yapay Zekâ, Evrişimsel Sinir Ağları, Eylem Tanıma.

## THEFT DETECTION IN SUPERMARKET VIDEOS WITH 3-DIMENSIONAL ACTION RECOGNITION RESIDUAL NETWORK MODEL

## SUMMARY

Recently, there has been increasing interest in artificial intelligence models for theft detection in supermarkets. Supermarket thefts are making to lose money for supermarkets financially. In order to prevent these losses, models for human theft action are being developed.

2-D CNN and 3-D CNN action recognition models are used to classify daily human actions. It contains both visual and movement information while performing human action. Visual and motion information refers to spatio-temporal information. Spatio-temporal action information is extracted with action recognition models.

In this thesis study, theft and non-theft action video dataset are generated. The generated theft and non-theft action dataset videos consist of 3 actions each. Acts of theft: belongings to; put in a pocket, put in a bag, and put in a handbag. Non-theft acts: in the supermarket; walking, standing still, and picking up items from the shelf.

The training dataset was collected from Youtube and the test dataset was collected from a supermarket security camera. The training dataset consists of 300 videos as 161 non-theft actions and 139 theft actions. The test dataset consists of 270 videos as 140 non-theft actions and 130 theft actions.

Optimizing 3-D CNN models from scratch requires large-scale datasets, otherwise, the accuracy of the network drops rapidly. Since the generated theft training dataset is small-scale, the 18-layer 3-D Residual Network model, which was pre-trained in the large-scale Kinetics-700 dataset, was used with transfer learning. The weights of the underlying 3-D Residual Network model except the FC layer are used and the model is trained by updating only the FC layer weights.

In order to examine and classify the act of theft in more detail, a deep model is trained 12 times with different parameters. The versions created are different from each other in terms of parameters. Versions differ in input image size, frame length, and batch size. Versions; takes an RGB input image and is trained in 200 epochs. Obtained training and test results were compared in terms of accuracy.

As a result of the training and test results, Version 1 is the model with the best results with 88.0% and 77.0% accuracy rates, respectively. Version 1 has, 224×224×3 RGB input image, 32 frame length and batch size of 12. As a result, an 18-layer 3-D Residual Network model has been developed, capable of detecting theft in the supermarket. A 18 layer 3-D Residual Network model has been developed that can successfully classify the theft action in supermarket.

**Keywords:** Theft Detection, Artificial Intelligence, Convolutional Neural Networks, Action Recognition.

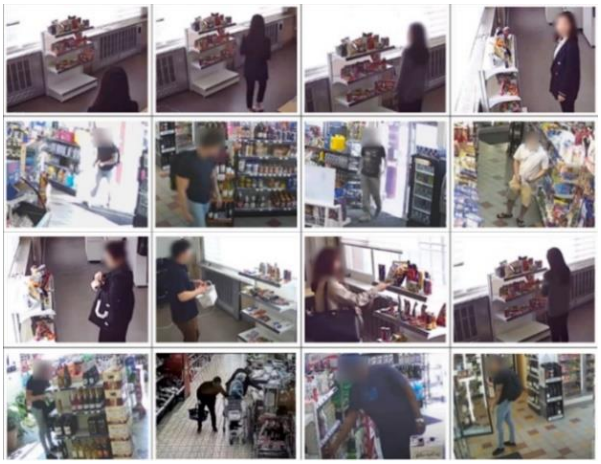
# 1. GİRİŞ

Süpermarketler insanların günlük ihtiyaçlarını karşılamak için bir gerekliliktir ancak bu ihtiyaç ile süpermarketlerde hırsızlıklar da görülmekte ve finansal açıdan marketleri zarara sokmaktadır. Amerika Birleşik Devletleri'nde (ABD) her yıl yaklaşık 27 milyon kişinin hırsızlık yaptığı, sonuç olarak her yıl en az 13 milyar ABD doları değerinde malın çalındığı ve dünya çapında da yılda 100 milyar dolar tutarından fazla malın çalındığı bildirilmiştir [1]. Ayrıca, bazı süpermarketler önlem olarak etiket alarm (paper alarm tag) kullanmaktadır. Ancak, bu etiketler kolayca sökülebileceği için hırsızlığın önüne tam anlamıyla geçememektedir. Çoğu süpermarket hırsızlığı önlemek için güvenlik kameraları kullanmakta ve kameralar görevliler tarafından izlenmektedir. Ancak kamera sayısının fazla olması ve bu durumda, insan görüşünün sınırlı olması ve çalışmadan doğan yorgunluk ile hırsızlık tespitinin önüne çoğu zaman geçilememektedir. Böylece, süpermarketlerdeki hırsızlık olayının anında tespiti çözülmesi gerekli bir problem haline gelmektedir. Yapay zekânın gelişmesiyle birlikte, süpermarketlerdeki hırsızlıkları önlemek için bu alandan faydalanılabilir. Bu yüzden, bu tez çalışmasında, yapay zekânın konusu olan insan eylem tanıma (ET) modeline dayanan hırsızlık tespiti konusuna odaklanılmıştır.

ET videodaki eylemi anlama ile ilgilidir ve onlarca yıldır araştırılmakta ve çalışılmaktadır. ET, kamera sistemleri, insan-bilgisayar etkileşimi ve robotik sistemler gibi birçok alanda uygulanmaktadır. ET insanlığın bir parçası olmuş ve genellikle günlük yaşantıdaki insan eylemlerini içeren veri setlerini [2-5] sınıflandırmak için kullanılmıştır. ET, önceden tanımlanmış veya etiketlenmiş kısa video kliplerindeki (bu çalışma için hırsızlık videoları için kısa video uzunluğu 1-2 saniye aralığındadır) etkinlikleri hem uzaysal hem de zamansal olarak inceleyerek sınıflandırmaktadır. Bir video klip, ET için iki önemli bilgi içermekte ve bunlar da uzay ve zaman bilgileridir. Uzaysal bilgi, bir video klibindeki tek bir çerçevedeki (frame) statik bilgiyi ifade ederken, zamansal bilgi ise farklı zamanlara ait sıralı görüntüleri temsil etmektedir. Başka bir ifadeyle uzaysal bilgi, algılanmak istenen nesnenin görsel bilgisini öğrenmek için işe yaramaktadır. Süpermarketteki insanın görsel bilgisinin ET modeli

ile işlenmesiyle, insanın tespiti uzaysal bilgiye örnek verilebilir. Zamansal bilgi ise, nesnenin zamandaki değişimini inceleyerek, sisteme öğretilmiş olan eylemler arasından en doğru tahmini yapmaya çalışmaktadır. Süpermarketteki insanların davranışları ele alınıldığında, insanın süpermarkette yürümesi, raftan eşya, mal alması veya hırsızlık yaparken ki hareketlerinin bir sıra içerisinde zamanda incelenmesi, zamansal bilgiye örnek verilebilir. Sonuç olarak ET, video klibindeki nesnenin dinamik değişliklerini izleyebilmektedir. Böylece, ET modelleri süpermarketlerdeki hırsızlık eylemlerini gözlemlemek ve belirlemek için kamera sistemlerinde kullanılabilir.

Bu çalışmada, süpermarketlerdeki hırsızlık eylemini sınıflandırmada yeni bir yaklaşım önerilmiştir. Önceden eğitilmiş (pre-trained) bir ET 3-Boyutlu (3-Dimensional) Evrişimsel Sinir Ağı (Convolutional Neural Network) modelini transfer öğrenim (transfer learning) yaparak kullanılmıştır. Eğitim veri seti için hırsızlık ve hırsızlık olmayan eylemler, YouTube'dan toplanmıştır. Ayrıca, test veri seti Bursa'daki bir süpermarketten, kişiler sanki hırsızlık yapıyormuş şeklinde davranarak, hırsızlık test veri seti oluşturulmuş ve hırsızlık olmayan eylemler için ise marketteki normal davranışlar toplanarak, hırsızlık olmayan test veri eylemleri oluşturulmuştur. Hırsızlık ve hırsızlık olmayan eylemlerde, modellerin parametre etkilerini incelemek için on iki farklı model eğitilmiştir. Hırsızlık eylemlerini sınıflandırmak için derin öğrenme (deep learning) literatüründe 2-Boyutlu (2-B) ve 3-Boyutlu (3-B) Evrişimsel Sinir Ağı (CNN) modelleri kullanılmıştır [6-8]. Ayrıca, çalışmada hazırlanan veri seti diğer hırsızlık veri setlerine benzer örnekler içermektedir. Örneğin; hırsızlık olmayan eylemler için süpermarkette: sabit olarak durmak, eşyaları (mal) raftan almak, yürümek. Hırsızlık olan eylemler için süpermarkette eşyaları: cebe sokmak, çantaya sokmak. Bahsedilen örnekler Şekil 1.1'de verilmiştir.



**Şekil 1.1 :** Süpermarket veri seti örnekleri [8].

### 1.1 Uzam-Zamansal Kavramı

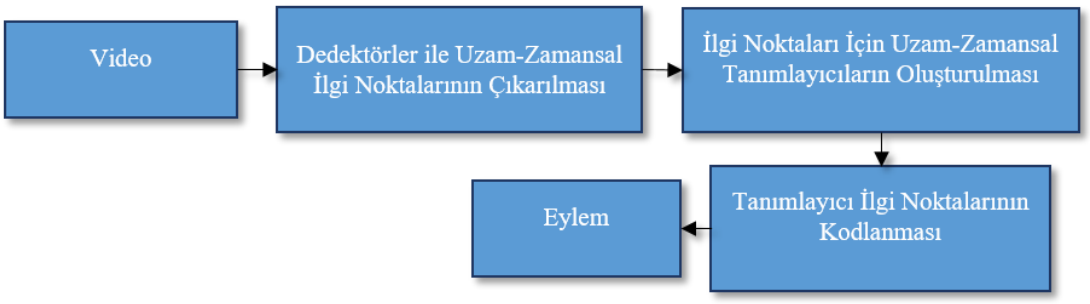
Uzam-zamansal kavramı, uzaysal veri serisinin zamandaki değişimidir. Bu kavram hem uzaya hem de zaman bağımlıdır. Uzam-zamansal 3-B'dir. 3-B olan bu kavramdaki uzay bilgisi 2-B ve zamansal bilgi ise 1-Boyutlu'dur (1-B). Buradaki uzay-zamansal veriye, hareket eden nesneler örnek gösterilebilir [9]. Bu tezde ele alınan uzam-zamansal bilgi, video görüntüsünde hareket eden nesnedir. Buna bağlı olarak, bir videodaki ardışık görüntünün herhangi bir görüntüsü 2-B uzayı temsil etmektedir. Bu 2-B görüntü ise nesnenin görsel bilgisini içermektedir. 1-B olan uzaysal bilgi ise sınıflandırılmak istenen eylemin zamana bağlı değişimini ifade etmektedir.

### 1.2 Derin Öğrenmeden Önce Eylem Tanıma Çalışmaları

Video kliplerindeki eylemleri tanımak için derin öğrenme öncesinde, matematiksel veya algısal modellere dayalı olarak geliştirilmiş formül ve algoritmalar aracılığıyla sınıflandırma çalışmaları yapılmıştır. Çalışmalar, 3-B uzam-zamansal dedektör ve tanımlayıcılar [10-15] ve Trajectory-Tabanlı dedektör ve tanımlayıcılar [16-21] olarak iki başlık altında incelenmiştir. Bu çalışmalarda, uzam-zamansal bilginin çıkarılıp işlenmesi ön plana çıkmaktadır.

Derin öğrenme öncesi ET akış diyagramı Şekil 1.2'de gösterilmiştir. Sisteme, ilk olarak girdi videosu verilirmiştir. Daha sonra, el yapımı (hand-crafted) dedektörler (detectors) ile uzam-zamansal ilgi noktaları (interest points) çıkartılmıştır. İlgi noktaları ile uzam-zamansal tanımlayıcılar (spatio-temporal descriptors)

oluşturulurmuştur. Daha sonra, uzam-zamansal tanımlayıcılar kodlanmış ve eylem sınıflandırılmıştır.



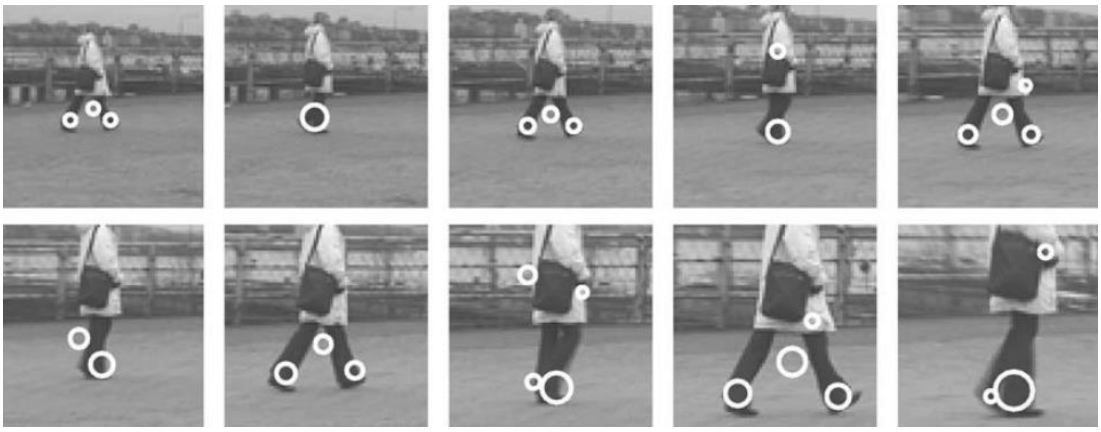

**Şekil 1.2 :** Derin öğrenme öncesi eylem tanıma akış diyagramı.

İlgi noktalarının 3-B dedektörler ile çıkarılması işlemi, uzaysal dedektörlerin zamansal boyutta genişletilmesiyle gerçekleşmektedir. Bu genişleme ile dedektörler, video görüntüsünde 3-B damlalar (blobs) ve köşeler (corners) bulmaktadır [10-15]. Trajectory-tabanlı dedektörler de ilgi noktalarının bulunması ise optik akış (optical flow), KLT ve SIFT gibi izleyici algoritmalar ile gerçekleştirilmiştir [16,17]. İlgi noktaları için uzam-zamansal tanımlayıcıların oluşturulması ise Bag of Visual Words (BoVW), Bags of Key Points (BoKP) [18], Fisher Vectors (FV) [19], Vector of Locally Aggregated Descriptors (VLAD) [20] gibi modeller ile kodlanarak gerçekleştirilmiştir. Bu modeller genel olarak lokal tanımlayıcıları bir kümeye kodlar. Buradaki kodlanan küme frekans histgoramı olabilir. BoKP'da kümeleme işlemi için, Naive Bayes ve Support Vector Machine (SVM) kullanılmış ancak SVM'nin daha iyi sonuç verdiği görülmüştür [18]. FV'de olasılıksal Gaussian Mixture Model (GMM) ile yerel tanımlayıcılar oluşturulmuştur. FV'nin BoVW'a göre avantajları; gradyan hesaplaması GMM'nin karışım ağırlığı parametrelerine bağlı olmasıyla ek gradyanlar ile doğruluk açısından daha üstün olduğu ve FV lineer sınıflandırıcılarla başarılı bir şekilde çalıştığı için geniş ölçekli sınıflandırma işlemlerinde daha başarılı olduğu gösterilmiştir [19]. VLAD, FV'nin basitleştirilmiş haline benzetilmiştir. SIFT tanımlayıcılar kullanılmış, tanımlayıcının kodlama işlemi ise komşu noktaların kestirimi ile yapılmış ve kestirim için Euclidean Locality-Sensitive Hashing kullanılmıştır [20].

### 1.2.1 3-Boyutlu uzam-zamansal dedektör ve tanımlayıcılar

2004'te Laptev, uzaysal boyutta tespit edilen ilgi noktalarını uzam-zamansal boyuta genişleterek yeni bir kavram ortaya çıkarmıştır [10]. Uzam-zamansal boyutta ilgi

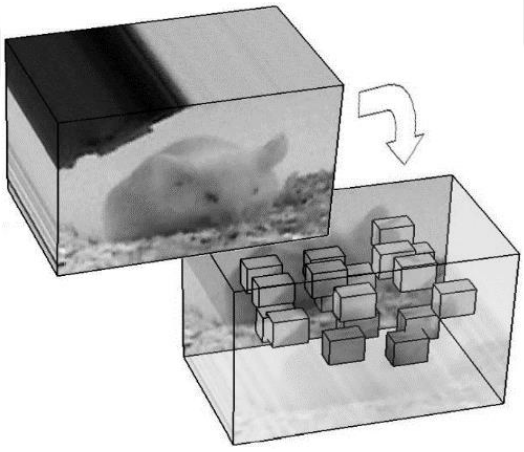
noktalarının tespiti için uzaysal ilgi noktası tespit eden Harris ve Förster operatörleri temel alınmıştır. Bu operatörler, görüntü ölçek uzayında, Gaussian penceresi üzerine entegre edilmiş ikinci moment matrisi kullanmaktadır. İkinci moment tanımlayıcısı, bir noktanın yerel komşuluğundaki 2-B görüntünün, yönelimlerinin ve dağılımının kovaryans matrisi olarak düşünülebilir. Kovaryans matrisi, yerel ölçekte birinci mertebeden türevleri hesaplanarak oluşturulmaktadır. Böylece, kovaryans matrisinde oluşan öz değerlerdeki farkların büyük olduğu değerleri inceleyerek kenar tespiti yapmaktadır. Sonuç olarak, görüntü üzerinde gezdirilen pencere ile görüntü öz değerleri elde edilir ve pencere üzerinde her iki yönde yüksek değişimlere uğrayan kenar etrafındaki noktaların tespit edilmesiyle ilgi noktaları çıkartılmaktadır. Yeni oluşturulan uzam-zamansal operatör ile zamansal boyutta, görüntü dizilerindeki hareketleri ve olayları tespit eden 3-B bir operatör geliştirilmiştir. Temel alınan operatördeki gibi uzam-zamansal hacimlerdeki görüntü değerlerinin hem uzaysal hem de zamansal yönlerdeki büyük varyasyonlara sahip noktaları hareket noktaları olarak kabul edilmektedir. Oluşan noktalar, sabit olmayan hareketli yerel uzam-zamansal komşuluklara karşılık gelen, zaman içinde farklı konumlara sahip uzamsal ilgi noktalarına karşılık gelmektedir. Böylece, Gaussian penceresine bağımsız zamansal varyans eklenmiştir. Sonuç olarak, öz değerler matrisi, maksimum Laplacian of Gaussian değerleri aranarak oluşturulmuştur. Şekil 1.3'te geliştirilmiş olan model ile görüntü dizisinde tespit edilen ilgi noktaları gösterilmiştir.



**Şekil 1.3 :** Tespit edilmiş uzam-zamansal ilgi noktaları [10].

2005'te yapılan bir çalışmada, 2-B ilgi noktası dedektörlerinin doğrudan 3-B karşılıklarının yetersiz olduğu gösterilmiş ve bir alternatif önerilmiştir [11]. İlgi noktaları temel alınarak, uzam-zamansal pencerelenmiş verilere dayalı bir algoritma geliştirilmiştir. İnsan yüz ifadesi ve bazı kemirgen davranışları gibi eylemler

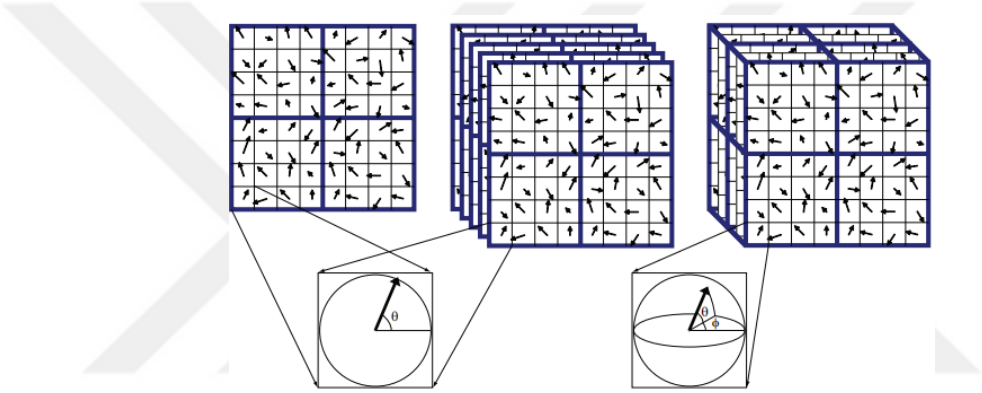
incelenmiştir. Bu eylemlerin tespiti için kullanılan nesne tanıma yaklaşımları ile sınıflandırılması sonucu ve [10]'da geliştirilmiş olan 3-B operatör ile ilgi noktaları çıkarıldığında tespit edilen ilgi noktalarının çok az olmasından dolayı bu veri seti temel alınmıştır. Bu eylemler ise ilgi noktalarının kullanılmasıyla uzam-zamansal boyutta dedektör küboidler (cuboid) oluşturulmuş ve geliştirilen tanımlayıcı ile karakterize edilmiştir. Uzam-zamansal noktaların çıkarımı için ise uzaysal boyutta 2-B Gaussian yumuşatma çekirdeği (kernel) ve 1-B Gabor filtresinin kareleme çifti (a quadrature pair) ile konvolüsyonu yapılarak uygulanmıştır. Böylece oluşturulan dedektör, lokal görüntü yoğunluklarındaki periyodik frekans bileşenlerinin değişikliklerinin nerede fazla olduğunu algılayacak hale getirilmiştir. Sonuç olarak, video görüntüsünden uzam-zamansal ilgi noktaları küboidler olarak çıkarılmıştır. Küboidler aracılığı ile görüntüden, normalize edilmiş piksel değerleri, parlaklık gradyanı ve pencereli optik değerleri çıkarılmıştır. Çıkarılan bu değerler k-means ile kümeleme işlemi yapılarak ayrılmış, vektöre çevrildikten sonra eylem sınıflandırılmıştır. Uzam-zamansal ilgi noktaları çıkarılmış örnek küboid görüntüsü Şekil 1.4'te gösterilmiştir.



**Şekil 1.4 :** Uzam-zamansal küboidler [11].

2007'de yapılan bir çalışmada, 3-B SIFT tanımlayıcı geliştirilmiştir [12]. 2-B Bag of Words (BoW) SIFT tanımlayıcıları görüntü sınıflandırma işleminde, görüntünün sınırlarını çok güçlü şekilde ifade ettiği için bu çalışma temel alınmıştır. 3-B SIFT modelinde; 2-B BoW modeli genişletilerek 3-B uzam-zamansal hale getirilmiş ve ilgi noktalarının bulunmasından ziyade 3-B tanımlayıcıların oluşturulmasına daha çok önem verilmiştir. Ayrıca, modelin hızlı çalışması için ilgi noktaları rastgele seçilmiş ve seçilen noktalar sınıflandırıldıktan sonra yönleri incelenmiştir. Tanımlayıcının ilk adımı, komşu noktaların genel yönelimini yani oryantasyonunu hesaplamaktadır.

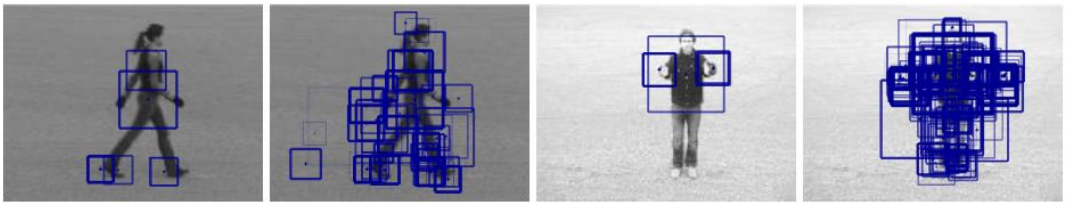
Oryantasyon hesabı, gradyan büyüklüğünün (magnitude) hesaplanmasıyla gerçekleşmektedir. Böylece, görüntüdeki her bir piksel iki değere sahip olmakta ve bu değerler de gradyanın 3-B bölgedeki yönünü temsil etmektedir. Sonraki adım da 3-B komşuluğu bulmak için ağırlıklı histogramlar oluşturulmaktadır. Son olarak, histogramlar arasındaki korelasyona göre eylemler sınıflandırılmaktadır. 3-B SIFT ve 2-B SIFT test sonuçları karşılaştırılmış ve 3-B SIFT'in 2-B'a göre çok daha başarılı olduğu ve uzam-zamansal boyutun önemi gösterilmiştir. Şekil 1.5'te ilk olarak 2-B SIFT tanımlayıcısı gösterilmiş, ortadaki şekil ise orijinal 2-B SIFT modelinin bir video görüntüsündeki ardışık kullanımını ifade etmekte ve sağdaki şekil ise uzam-zamansal boyuta genişletilmiş olan 3-B SIFT modelini göstermektedir. Bahsedildiği gibi 3-B SIFT tanımlayıcının bölünmüş hacimleri ile histogramlar oluşturulmuştur.



**Şekil 1.5 :** 2-B SIFT'in uzaysal boyuttaki ve 3-B SIFT'in uzam-zamansal boyuttaki oryantasyonu [12].

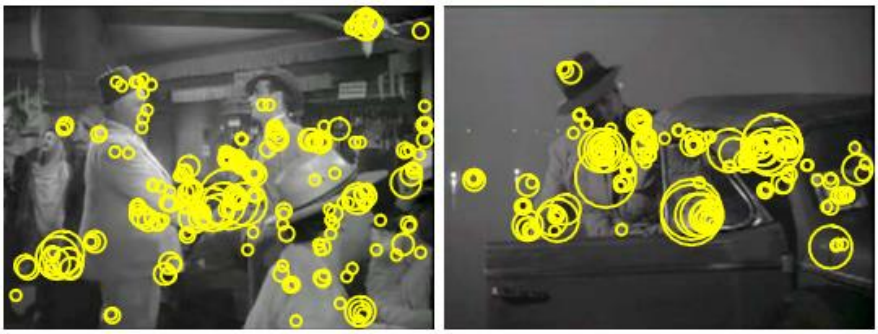
[13]'deki çalışma 2-B SURF [21] tanımlayıcının uzam-zamansal boyuta genişletilmiş halidir. Önceki çalışmalardan farklı olarak bu çalışmadaki dedektör, ilk kez uzam-zamansal ilgi noktalarının, video içeriğini yoğun bir şekilde kaplayan ve aynı zamanda ölçekte değişmez olan bir model oluşturulmuştur. Bunun için, ölçek-uzay teorisi, Hessian matris determinantı ile uygulanmış ve ölçekte normalize edilmiş değerlerin türevlerinin alınması ile gerçekleştirilmiştir. Buradaki matris, 3 x 3 uzam-zamansal ikinci moment matrisidir. Ayrıca, ölçekteki normalizasyon işlemi ile Gaussian damlasının (blob) en ideal olduğu yerler bulunmaktadır. Gaussian damlasının merkezinde determinant matrisi oluşturulmuştur. Merkezdeki matrisin ikinci mertebeden türevleri alınmış ve integral video yapısı kullanılarak karmaşık işlemlerin zorluğu azaltılmıştır. Türevlerin hesaplanması için kutu-filtresi (box-filter) kullanılmıştır. Kutu filtresinin ölçeği yükseltilerek farklı boyutlarda kutular oluşturulmuş ve verimli bir şekilde uygulanmıştır. Son olarak, oluşturulmuş

kutulardan tanımlayıcının çıkarımı için Haar-wavelets kullanılmıştır. Şekil 1.6'da bir eylemden 3-B SURF ile çıkarılmış olan ilgi noktaları gösterilmiştir. Ayrıca geliştirilen 3-B SURF metodu ile önceki metotlar değerlendirilmiştir [10,11]. [11]'deki çalışmanın ölçekte değişmez olduğu ve küboidlerin boyutunun kullanıcı tarafından belirlendiği gösterilmiştir. [10]'daki çalışmanın iteratif olduğu ancak geliştirilen bu metot ölçekte değişmez olduğu için iteratif bir yaklaşımın gerekli olmadığı gösterilmiştir.



**Şekil 1.6 :** 3-B SURF ile tanımlayıcıların oluşturulması [13].

[14]'deki video sınıflandırma çalışması, görüntü tanıma modellerinden olan uzaysal piramitleri (spatial pyramids) temel almaktadır. Piramitler uzam-zamansal boyuta genişletilerek 3-B hale getirilmiştir. İlgi noktalarının tespiti için 3-B Harris [10] dedektörü kullanılmıştır. Geliştirilen dedektör ile tespit edilen uzam-zamansal ilgi noktaları Şekil 1.7'de gösterilmiştir. Ölçek seçimi için çok ölçekli yaklaşım temel alınmış ve uzam-zamansal ölçeklerin çoklu seviyelerinde özellikler çıkarılmıştır. Böylece, iteratif yaklaşım kullanılmamış ve hesaplama zorluklarından kaçınılmıştır. Lokal özelliklerin, hareketini ve görünümünü karakterize etmek için tespit edilen komşu noktaların uzam-zamansal hacimlerindeki histogram tanımlayıcıları oluşturulmuştur. Her hacim tespit edilen noktanın ölçek büyüklüğüyle alakalıdır. Her hacim, bir küboid ızgarasına bölünmüştür. Her küboid için BoW ile yönlendirilmiş gradyan (Histogram of Oriented Gradient) ve optik akışın (Histogram of Oriented Flow) kaba histogramları hesaplanmıştır. Son olarak, eylemleri sınıflandırmak için SVM kullanılmıştır.



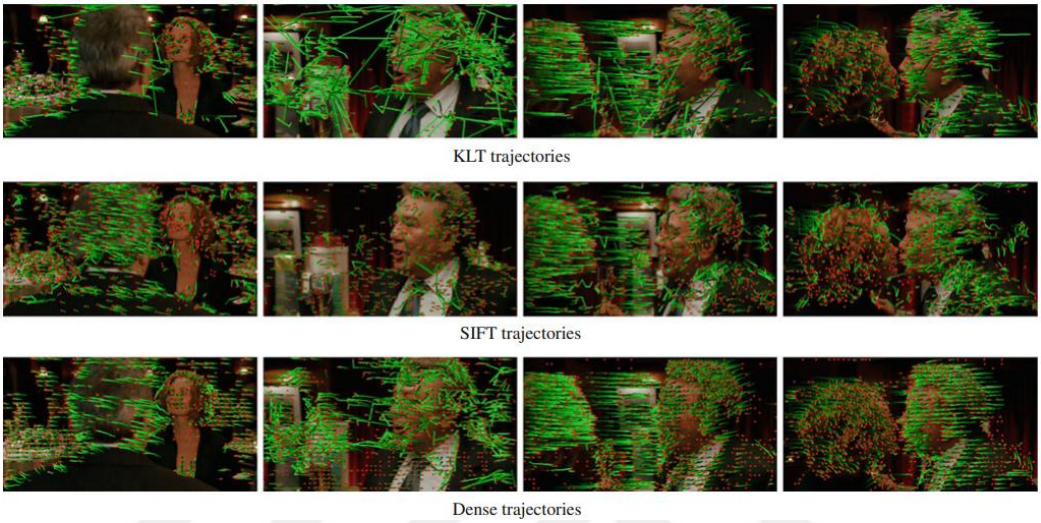
**Şekil 1.7 :** Uzam-zamansal ilgi noktalarının tespiti [14].

[15]'teki çalışma 3-B SIFT çalışmasını temel almaktadır. 3-B uzam-zamansal gradyanları hesaplamak için ölçekler, rastgele seçilmiştir. Geliştirilen algoritma, integral videolara dayalıdır. Ayrıca, yumuşatma operatörü olarak genel olarak Gaussian filtre kullanılmakta ancak bu çalışmada kutu filtresi kullanılmıştır. Çalışmada 3-B gradyanlar kullanıldığı için bellek açısından daha verimli bir modeldir. Düzenli çokyüzlülere (polyhedrons) dayanan bir 3-B oryantasyon kullanılmıştır. 3-B SIFT [12] modelinde oryantasyon niceleme işlemi için gradyanlar, kutupsal koordinatlarla temsil edilmiştir. Oluşan gradyanlar da meridyenler ve paraleller kullanılarak gradyan histogramlara dönüştürülmüştür. Ancak bu işlemde oluşan gradyan histogram kutular (bins) git gide küçüldüğü için kutuplardaki tekilliklere yol açarak problem oluşturduğu görülmüştür. Bu problemin çözümü için oryantasyon işleminde, çokyüzlüler kullanılarak çözülmüştür.

### 1.2.2 Trajectory dedektör ve tanımlayıcılar

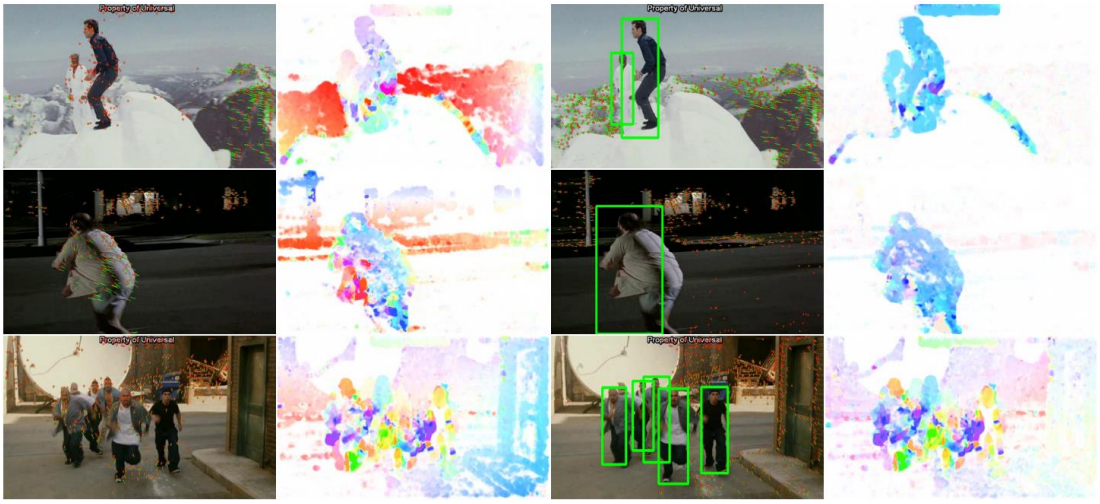
İlgi noktalarının bulunması için uzam-zamansal boyutta inceleme yapılırken yörünge (trajectory) tabanlı çalışmalarda ilgi noktaları uzaysal boyuttan çıkarılmakta ve bu noktaları zamansal boyutta takip etmektedir. Yörüngelerin oluşturulması için KLT ve SIFT tanımlayıcıları kullanılmıştır [16]. 2012'de yapılan bu çalışmada [16] Dense Trajectory (DT) ve hareket sınırı tanımlayıcıları tanıtılmıştır. Yörüngeler, video klibinden yerel hareket bilgilerini yakalamaktadır. Optik akış algoritmaları yoğun yörüngelerin çıkarılmasını sağlamaktadır. Bu sebeple yoğun yörüngelerin kalitesi, optik akış algoritmasının kalitesiyle sınırlanmaktadır. İlgi noktalarının oluşturulması için yoğun (dense) örnekleme yapılmıştır. İlgi noktalarının takip edilmesi, yani şekil verilmesi için gradyan histogramı, optik akış histgoramı ve hareket sınır histogramları kullanılmıştır. Yeni bir tanımlayıcı olarak hareket sınırı histogramları oluşturulmuştur. Hareket sınırı tanımlayıcısı, optik akışı vektöre dönüştürür ve her bir noktanın uzaysal

olarak birinci mertebeden türevini hesaplamaktadır. Oryantasyon bilgileri ise histogramlara nicelenmiştir. Ayrıca, DT ölçekte değişmezdir. Bu çalışmada KLT, SIFT ve DT tanımlayıcıları kıyaslanmış ve en iyi sonucu DT'nin verdiği tespit edilmiştir. Bu tanımlayıcılarla oluşturulan ilgi noktaları ve yörüngeleri Şekil 1.8'de gösterilmiştir. Şekil 1.8'de görüldüğü üzere, DT ile oluşturulmuş ilgi noktaları ve noktaların takibi, SIFT ve KLT yörüngelere göre daha pürüzsüz ve temizdir.




**Şekil 1.8 :** KTL, SIFT ve DT tanımlayıcılarla oluşturulan ilgi noktalarının takibi [16].

Improved Dense Trajectories (IDT), DT'nin geliştirilmiş halidir [17]. Arka plan yörüngelerini ortadan kaldırarak ve optik akışı kamera hareketine yaklaşan bir şekilde tahmin edilen, homografi ile çarpıtarak performansın önemli ölçüde iyileştirilebileceği gösterilmiştir. Kamera hareketini tahmin etmek için, SURF tanımlayıcıları ve yoğun optik akışı kullanarak görüntüler arasındaki özellik noktaları eşleştirilmiştir. Bu eşleşmeler, RANSAC ile homografiyi tahmin etmek için kullanılmıştır. İnsan hareketi tutarsız eşleşmeler oluşturabilmektedir. Bunun önüne geçmek için, başarılı insan dedektörleri ile potansiyel olarak tutarsız eşleşmeler, kamera hareket tahmini sırasında kaldırılmıştır. Özellikleri kodlamak için, BoW ve FV kullanılmıştır. IDT ile yapılan insan hareket tespit örnekleri Şekil 1.9'da gösterilmiştir. Ayrıca Şekil 1.9'da insan dedektörünün olduğu ve olmadığı örnek görüntüler gösterilmiştir. İnsan dedektörü ile oluşturulan yaklaşımın optik akış görüntüleri karşılaştırıldığında daha başarılı olduğu gösterilmiştir.



**Şekil 1.9 :** IDT ile insan dedektörü yokken ve varken oluşturulan optik akış görüntüleri [17].

### 1.3 Derin Öğrenme ile Eylem Tanıma Çalışmaları

Derin öğrenme öncesi ET için geliştirilmiş modeller, modele bağlı algoritmalar olduğu için üzerinde çalışılan veri setine özel geliştirilmiş ve geliştirildiği veri setine bağımlı kalmıştır. Bu sebeple, geliştirilen algoritmalar uygulandığı veri setinde yüksek performans göstermesine rağmen, farklı veri setlerine uygulandığında aynı performansı gösterememiştir.

Lecun vd. 1988'de derin öğrenmenin gelişmesiyle ve artan bilgi birikimiyle birlikte, birden çok özellik katmanındaki bilgileri öğrenebilen, karmaşık ham girdilere karşı en az ön işleme gerektiren gradyan tabanlı öğrenme algoritması geliştirmiş ve 2-B CNN'nin temelini atmıştır [22]. 2-B CNN mimarilerinin [22,23] herhangi bir ön işleme adımı olmadan doğrudan görüntü piksellerinden, görsel örüntüleri öğrenebildiği ve özellikleri otomatik olarak çıkarıldığı gösterilmiştir. Böylece, derin öğrenme modellerinin 2-B görüntü işleme veri setlerinde gösterdiği başarılardan dolayı, ET video veri setlerini sınıflandırmak için uygulanabileceği belirtilmiştir [24].

ET için derin öğrenme mimarileri: 2-B [25-28] CNN ve 3-B [29-35] CNN modelleri olarak ikiye ayrılmaktadır. Literatürdeki 2-B CNN modelleri: İki-Akış Ağları (Two-Stream Networks) [25-27] ve Geçici Segment Ağları (Temporal Segment Networks) [28] olarak ikiye ayrılmaktadır. 3-B CNN modelleri: Konvolüsyonel 3-B (C3D) [29], Şişirilmiş İki-Akış 3-B (I3D) [30], Yalancı-3B (P3D) [31], 3-B Artık Ağlar (3-D ResNets) [32], R(2+1)D [33], Ayrılabilir 3-B (S3D) [34] ve Kanalla Ayrılmış Evrişimsel Ağlar (CSN) [35].

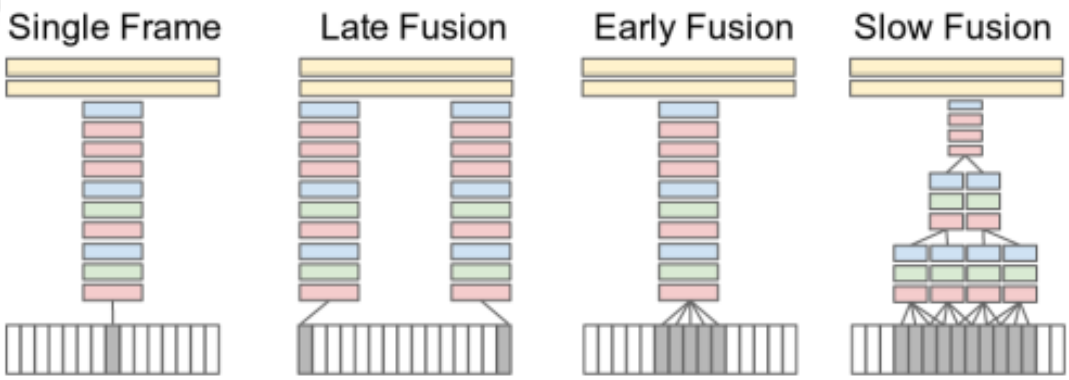
### 1.3.1 2-B CNN modelleri

#### 1.3.1.1 İki-akış ağları ve mimarileri

İki-akış ağları [25-27], hem uzaysal hem de zamansal boyutu inceleyen iki akış ağına sahiptir. Uzaysal ağ, giriş olarak tek bir Red Green Blue (RGB) görüntü alır ve bu görüntüyü işler. Zamansal ağ, giriş olarak birden çok ardışık görüntü alır ve bu görüntü serisi arasında optik akış yer değiştirme alanlarının istiflenmesiyle oluşturulmaktadır. Uzaysal akış ağ bölümünde işlenen görüntüde, sınıflandırılacak olan nesnenin görsel bilgisinin çıkarılması amaçlanırken, zamansal akış ağ bölümünde işlenen ardışık görüntü serisinde ise sınıflandırılmak istenen eylemin hareket bilgisinin çıkarılması amaçlanmaktadır. Uzaysal ve zamansal akış ağ mimarileri genel olarak: konvolüsyonel katmanlardan (convolutional layer), tam bağlantılı (FC) (fully-connected layer) katmanlardan ve softmax katmanından oluşmaktadır. İki-akış ağlarında uzaysal ve zamansal bilginin birleştirilmesi için kaynaşma (fusion) işlemi yapılmaktadır. Kaynaşma işleminde [25]'de geç kaynaşma, [26]'de ise erken, geç ve çoklu katman kaynaşma fonksiyonları kullanılmıştır.

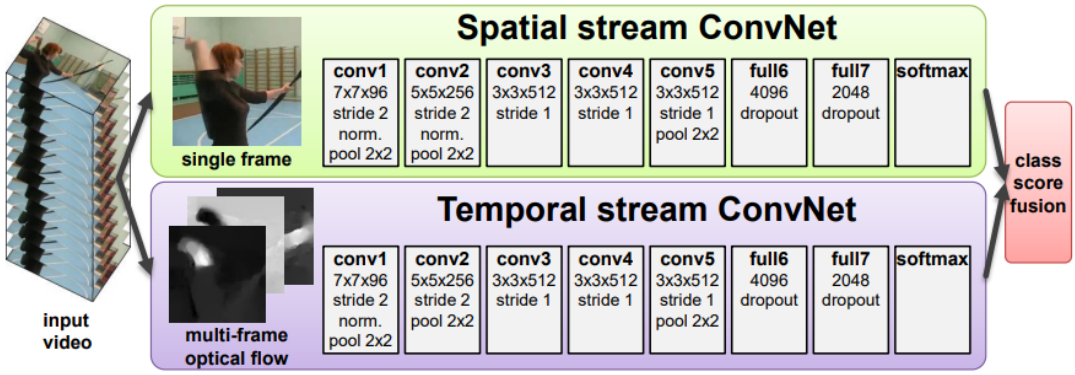
Karpathy vd. tarafından kaynaşma işlemi, videoda bulunan yerel hareket bilgisinden yararlanmak için bir CNN mimarisindeki zamansal bağlantı modeli araştırılarak, ek hareket bilgisinin bir CNN'nin tahminini nasıl etkilediği ve genel olarak performansının ne kadar değiştiği araştırılarak oluşturulmuştur [3]. Oluşturulacak modelin uzam-zamansal boyuta uzatıldığı düşünüldüğünde, klasik CNN modellerinin bir girdi görüntüsü alarak eğitildiği modellerin eğitiminin uzun sürmesinden dolayı uzam-zamansal modelin girdi olarak birçok görüntü alacağı için eğitim aşamasında milyonlarca parametreden oluşacağı öngörülerek, zamandan kazanmak için oluşturulan mimari iki akışa ayrılarak uzaysal ve zamansal ağlar oluşturulmuştur [3]. Mimaride, bağlam akışı ve fovea akışı vardır. Bağlam akışı, çerçeveyi daha düşük çözünürlüklü çerçeveye dönüştürür ve çerçevelerdeki (frame) sınıflandırılmak istenen nesnenin görsel bilgisini öğrenmektedir. Fovea akışı, yalnızca çerçevenin orta yüksek çözünürlüklü kısmında işlem yapmakta ve çerçevenin hareket bilgisini öğrenmeyi amaçlamaktadır. Statik görünümün sınıflandırma katkısını anlamak için tek çerçeveli temel bir mimari oluşturulmuştur. Temel alınan mimaride girdi görüntüsü boyutu 224×224×3 şeklinde kullanılmıştır. Kaynaşma işlemi, kovolüsyon filtre katmanının zamanda uzatılmasıyla gerçekleştirilmiştir. Üç çeşit kaynaşma modeli oluşturulmuş

olup bunlar; erken, geç ve yavaş kaynaşmadır. Erken kaynaşma uzantısı, zaman penceresinde 10 tane ardışık girdi görüntüsünün, giriş konvolüsyon katmanında kaynaşmasıyla gerçekleştirilmektedir. Erken kaynaşma ile ağın, video eyleminin lokal hareket yönünü tespit ettiği belirtilmiştir. Geç kaynaşma modeli, iki ayrık tek çerçeve ağı ile oluşturulmuştur. Geç kaynaşma, 15 ardışık görüntüden iki görüntü alır ve son katmandaki konvolüsyon katmanında kaynaşmayı gerçekleştirir. Yavaş kaynaşma modeli, erken ve geç kaynaşma modelinin dengeli halidir. Yavaş kaynaşma, zamansal bilgiyi yavaşça işlemektedir. Yavaş kaynaşma 10 girdi görüntüsü alır ve kovolüsyon katmanlarıyla girdileri yavaş yavaş işleyerek kaynaşmayı gerçekleştirmektedir. Şekil 1.10'da kaynaşma çeşitleri ve mimarileri gösterilmiştir [3]. En başarılı kaynaşma modelinin yavaş kaynaşma olduğu gösterilmiştir.




**Şekil 1.10 :** Kaynaşma çeşitleri [3].

Çalışma([25])'da oluşturulan iki-akış mimarisi Şekil 1.11'de gösterilmiştir. Mimarinin uzaysal bölümünde bir girdi çerçevesi ve zamansal bölümünde ise çoklu optik akış girdileri alınmış ve geç kaynaşma metodu kullanılmıştır. Girdiler işlendikten sonra softmax katmanından çıkan skorlar ile geç kaynaşma uygulanmış ve iki kaynaşma metodu kullanılmıştır. Bu metotlar: ortalama (averaging) ve SVM'dir. SVM'nin daha başarılı olduğu gösterilmiştir [25].



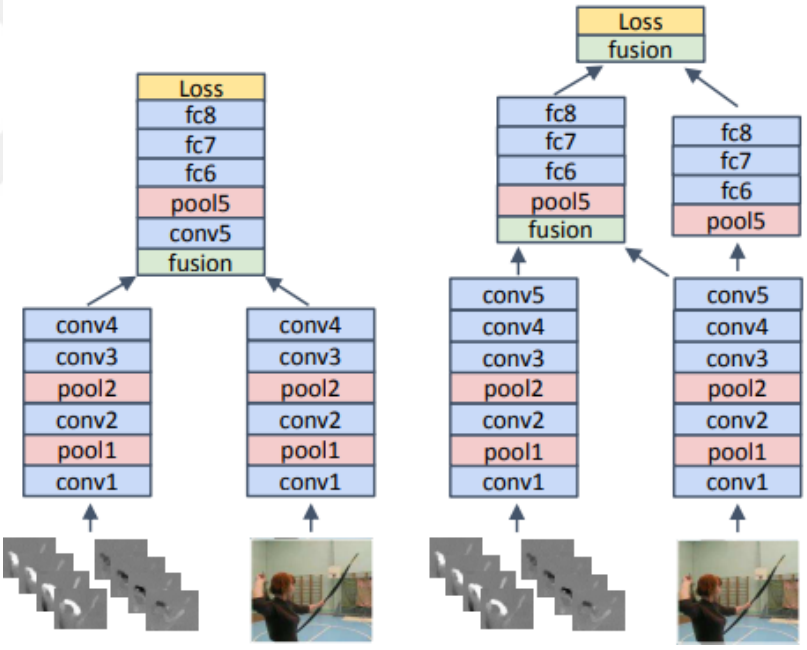

**Şekil 1.11 :** İki-akış mimarisi [25].

Uzaysal akış ve zamansal akış ağlarının performans etkileri değerlendirilmiştir [25]. Uzaysal ağlarda, RGB çerçeveler üç şekilde değerlendirilmiştir. İlk olarak, UCF-101 [5] veri seti üzerinde sıfırdan eğitim gerçekleştirilmiştir. İkinci olarak, ağ ilk önce ImageNet ILSVRC-2012 [36] veri setinde eğitilmiş ve eğitilen ağ UCF-101 veri setinde üzerinde ince ayar (fine-tuning) yapılarak eğitilmiştir. Üçüncü olarak, önceden eğitilmiş veri setindeki ağ yapısı korunarak, sadece son katmandaki sınıflandırıcı da ince ayar yapılarak eğitilmiştir. Bırakma (dropout) oranı 0,5 ve 0,9 olarak iki eğitim doğruluğu oluşturulmuştur. Sonuç olarak, sıfırdan eğitim yapıldığında her iki bırakma oranı için ağın ezberlemeye (overfitting) yöneldiği gösterilmiştir. Bırakma oranı 0.5 olan ince ayar eğitimi ve bırakma oranı 0,9 olan son katman ince ayar eğitim doğruluklarının en iyi sonuçları verdiği gösterilmiştir.

Zamansal ağlarda, optik akışların farklı çeşitlerinin etkileri incelenmiştir. Bunun için, tek çerçeve optik akış, optik akışlar 5 ve 10 çerçeve ile yığılmış, yörünge (trajectory) de yığılmış 10 optik akış çerçeve ve yığılmış 10 çerçeve iki-yönlü optik akış kullanılmıştır. Optik akışlarda, ortalama akış çıkarma (mean flow subtraction) uygulanmıştır. Bu işlem genel olarak, girdinin merkezine odaklanarak, doğrusal olmayan düzeltmelerden daha iyi yararlanması sağlamaktadır. Böylece, çerçeveler arasındaki küresel hareketin etkisini azaltmaktadır [25]. Mimariler UCF-101 veri setinde sıfırdan eğitilmiş ve bırakma oranı 0.9 seçilmiştir. Sonuç olarak, 10 çerçeve ile yığılmış ve yığılmış 10 çerçeve iki-yönlü optik akış modellerinin en iyi sonuçları verdiği gösterilmiştir.

[26]'de kaynaşma işleminden en iyi uzam-zamansal bilgileri toplamak için zamansal ve uzaysal akış mimarileri oluşturulmuş ve Şekil 1.12'de mimariler gösterilmiştir. Eğitim veri seti olarak, HMDB-51 [4] ve UCF-101 [5] seçilmiştir. Bu çalışmada, üç

bulgu araştırılmıştır. İlk olarak, [25]'de olduğu gibi softmax sınıflandırma katmanında kaynaşma yapmak yerine bu işlem performans kaybı olmaksızın konvolüsyon katmanında gerçekleştirilmiştir. Bunun sebebi, [25]'deki mimarinin uzaysal ve zamansal özellikler arasındaki piksel bazındaki haberleşmeyi öğrenememesi olmuştur [26]. İkinci olarak, özellikle uzaysal ağlarda, son katmanlarda kaynaşma yapmanın sınıflandırma skorunun değerini arttırabileceği ve sınıflandırmada olumsuz etki yapabileceği söylenmiş ve bu sebeple erken katmanlarda kaynaşma yapmanın daha iyi olacağı belirtilmiştir. Üçüncü olarak, soyut konvolüsyonel özelliklerin uzam-zamansal komşular üzerinde havuzlama (pooling) işlemiyle performansının artabileceği söylenmiştir [26]. Kaynaşma işleminin araştırılmasının asıl amacı, farklı sınıf ancak aynı hareket örüntüsüne sahip eylemlerin, mimari tarafından zamansal ağ ile hareketin ne olduğu ve uzaysal ağ ile hareketin konumunun tespit edilmesiyle birlikte bu olayların birleşimi ile eylemin daha doğru ayrıştırılması amaçlanmıştır.




**Şekil 1.12 :** İki-akış mimarilerinde kaynaşma işleminin gerçekleştirildiği katmanlar [26].

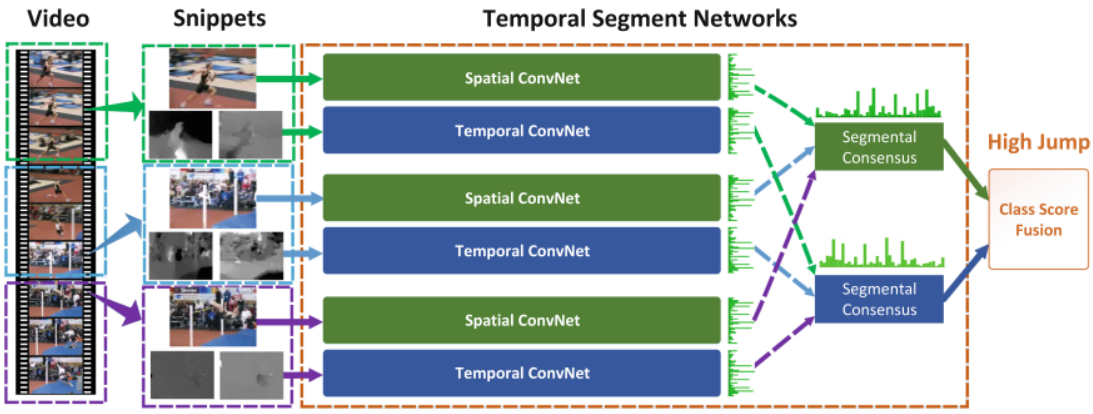
Kaynaşma işlemi için toplama (sum fusion), maksimum (max fusion), birleştirme (concatenation fusion), konvolüsyonel (conv fusion) ve bilinear kaynaşma aynı uzaysal konumlara sahip iki özellik haritalarında (feature maps) uygulanmıştır. En iyi sonucu konvolüsyonel kaynaşmanın verdiği gösterilmiştir [26]. Ayrıca, zamansal kaynaşma kısmında, zamansal boyutta 3-B konvolüsyon kaynaşma kullanılmış, 2-B havuzlama yerine 3-B maksimum havuzlamanın (max-pooling) daha başarılı olduğu

gösterilmiştir. Önceki çalışmalardan farklı olarak 3-B parametreler kullanılmasına rağmen yeni mimaride parametre sayısı çok fazla artmamıştır [26]. Mevcut veri setlerinin 3-B parametreleri eğitmek için çok küçük ölçekli (small-scale) veya çok kirli olduğu söylenmiştir.

#### 1.3.1.2 Geçici segment ağları (TSN) ve mimarisi

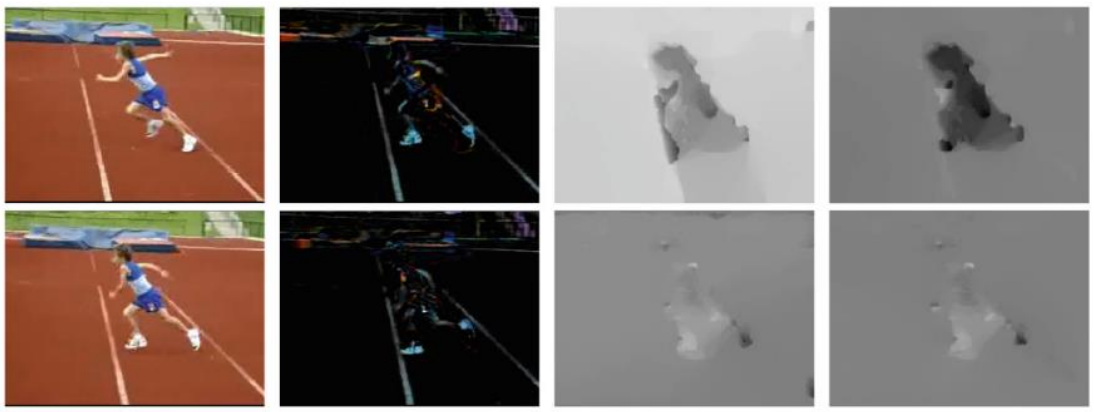
TSN, uzun menzilli zamansal yapı modelleme fikrine dayanmaktadır [28]. TSN, eylemin etkili sınıflandırılması için eylem videosunun tamamını kullanarak seyrek zamansal örnekleme stratejisini kullanmaktadır. Mevcut olan uzun menzilli zamansal modellerin, genellikle kısa süreli hareketlere ve nesnenin görüntüsüne odaklanmasından dolayı uzun menzilli zamansal yapıyı birleştirme kapasitesinden yoksun olduğu belirtilmiştir [28]. Dahası, kullanılan bu metotlar önceden tanımlanmış bir örnekleme aralığı ile yoğun zamansal örneklenmeye dayanmaktadır. Ancak bu yaklaşım uzun video dizilerine uygulandığında, aşırı hesaplama maliyetine sebep olmakta ve pratikte uygulanmasını sınırlamaktadır. Ayrıca, önceki çalışmalar değerlendirildiğinde ağları optimize etmek için büyük ölçekli video veri setlerinin gerektiği ancak küçük ölçekli mevcut veri setleriyle [4,5] sınırlı kalındığı bildirilmiştir. Mevcut küçük ölçekli veri setleri ile eğitim yapılacağı için bu probleme yönelik, gelişmiş veri büyütme (enhanced data augmentation), ön eğitim (pre-training) ve düzenleme (regularization) yöntemleri kullanılmıştır [28].

Zamansal yapı modelinde, ardışık çerçevelerin benzer bilgiler içermesinden dolayı gereksiz olduğu söylenmiştir. Seyrek uzaysal örnekleme seçiminin daha uygun olacağı belirtilmiş ve Şekil 1.13'te seyrek örnekleme ile TSN ağ mimarisi gösterilmiştir. Şekil 1.13'te gösterildiği gibi, eylem videosunu K tane eşit bölümlere (segments) böldükten sonra kısa parçacıklar (snippets) her bir bölümden rastgele seçilmekte ve seyrek bir örnekleme ile uzun video klibi küçük parçalara ayrılmaktadır. Böylece, hesaplama maliyeti düşürülmektedir. TSN mimarisi farklı küçük parçacıkları, RGB görüntü ve çoklu optik akış görüntüleri olarak uzaysal ve zamansal ağlarda işlendikten sonra Consensus fonksiyonundan çıkan bilgiler ile sınıf skorunda kaynaştırmaktadır. Consensus fonksiyonunda, maksimum havuzlama, ortalama havuzlama (average pooling) ve ağırlıklı ortalama fonksiyonları kullanılmıştır.




**Şekil 1.13 :** Geçici segment ağları [28].

TSN modelini güçlendirmek için dört farklı girdi modeli değerlendirilmiş ve yeni olarak iki girdi modeli tanıtılmıştır [28]. Şekil 1.14'te girdi örneklerine ait görüntüler gösterilmiştir. RGB görüntünün bir önceki ve sonraki kareler hakkında bağlamsal bilgiden yoksun olduğu düşünülerek RGB fark görüntüsü oluşturulmuştur. RGB fark görüntüsü, iki ardışık kare arasındaki görüntü değişimini ifade etmekte ve hareket bölgesine karşılık gelmektedir. Optik akış alanlarında, x veya y koordinat değerlerine filtrelerle işlem yapılarak, yatay (x) veya dikeydeki (y) pikseller daha belirgin hale getirilerek eylem hareketine ilişkin bilgi oluşturulmaktadır. Gerçek uygulamalarda, kamera hareketi yüzünden optik akış alanlarının insan eylemine odaklanamayabileceği düşünülerek çarpık optik akış alanları geliştirilmiştir. Çarpık optik akış alanları, IDT'den [17] esinlenerek oluşturulmuştur. Çarpık optik akış alanlarını oluşturmak için önce homografi matrisi tahmin edilmekte ve sonra kamera hareketi dengelenmektedir. Sonuç olarak, eğitim sonuçları ile çarpık optik akışın etkisi gösterilmiştir. Consensus fonksiyonunda, ortalama havuzlama yöntemi öne çıkmıştır.



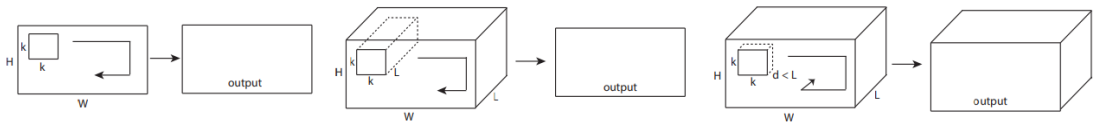
**Şekil 1.14 :** Dört tip girdi örneği: 1. RGB görüntü, 2. RGB fark görüntü, 3. optik akış alanları (x, y yönleri) ve 4. çarpık optik akış alanları (x, y yönleri) [28].

### 1.3.2 3-B CNN modelleri

#### 1.3.2.1 Konvolüsyonel 3-B (C3D)

ET literatüründeki ilk 3-B CNN mimarisi C3D'dir (Convolutional 3D) [29]. C3D girdi olarak tüm video çerçevelerini almakta ve herhangi bir ön işlem gerektirmediği için büyük veri setlerinde kolayca ölçeklenmektedir. Bu çalışma, iki-akış ağından farklı olarak [25] 2-B konvolüsyon ve 2-B havuzlama yerine bu fonksiyonları bütün ağa (network) yayarak 3-B olarak gerçekleştirmektedir.

3-B ağda konvolüsyon ve havuzlama işlemleri uzam-zamansal boyutta gerçekleşmektedir. Ancak, 2-B ağda bu işlemler sadece uzaysal boyutta gerçekleşmektedir. 2-B ve 3-B konvolüsyon işlemlerinin karşılaştırılması Şekil 1.15'te gösterilmiştir. Şekil 1.15 incelendiğinde, ilk olarak, 2-B görüntüye 2-B konvolüsyon işlemi uygulanmış ve 2-B görüntü elde edilmiştir. İkinci olarak, video görüntüsüne (üst üste birikmiş çoklu çerçeve görüntüsü) 2-B konvolüsyon işlemi uygulanmış ve yine sonuç olarak 2-B görüntü elde edilmiştir. Üçüncü olarak, video görüntüsüne 3-B konvolüsyon işlemi uygulanmış ve 3-B bir sonuç elde edilerek giriş sinyalindeki zamansal bilgi korunmuştur [29]. 2-B ağlar, 2-B her bir konvolüsyon işleminde giriş sinyalindeki zamansal bilgiyi kaybederken 3-B ağlar bu bilgiyi korumaktadır. Sonuç olarak 3-B ağ, 3-B konvolüsyon ve 3-B havuzlama işlemleriyle zamansal bilgileri daha iyi modellemektedir [29].



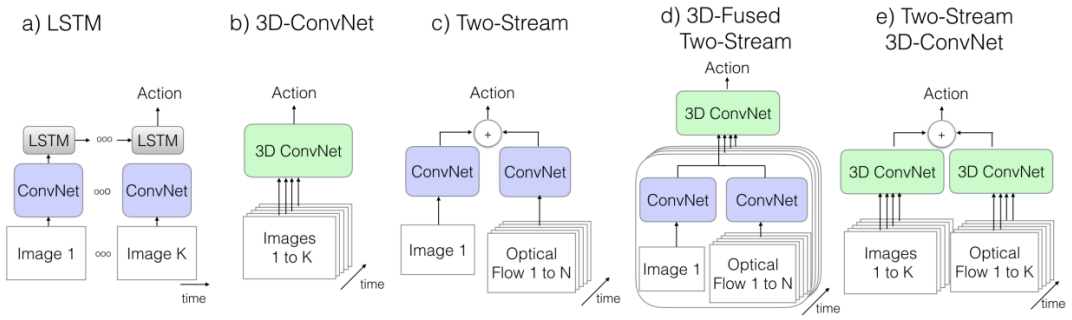

**Şekil 1.15 :** 2-B ve 3-B konvolüsyon işlemlerinin uygulanması [29].

C3D mimarisi için en uygun çekirdek derinliğinin (kernel depth) değerleri UCF-101 [5] veri setinde araştırılmıştır. İlk olarak, çekirdek derinliği; 1, 3, 5, 7 seçilmiş ve tüm konvolüsyon katmanları boyunca aynı kalarak araştırılmıştır. Ayrıca, çekirdek derinliğinin 1 olması Şekil 1.15'teki ikinci şekli ifade etmektedir. Çekirdek derinliği; artan 3-3-5-5-7 ve azalan 7-5-5-3-3 şeklinde iki mimari oluşturularak araştırılmıştır. Bu üç deney değerlendirildiğinde en iyi sonucu değişmeyen 3×3×3 çekirdek derinliği vermiş ve derinlik 3 seçilmiştir [29]. C3D mimarisi: 8 konvolüsyon katmanı, 5 havuzlama katmanı, 2 FC katman ve softmax çıkış katmanı şeklinde oluşturulmuştur.

Öncelikle C3D, video tanımlayıcı olarak çok sınıflı doğrusal SVM sınıflandırıcı ile bir arada kullanılmış ve büyük ölçekli Sports-1M [3] veri setinde uzam-zamansal özellikleri öğrenmesi hedeflenerek eğitilmiştir. Sonuç olarak, C3D'nin video klibindeki özellikleri öğrenmede, görünümü ve hareketi aynı anda modelleyen öğrenme makineleri olduğu deneysel olarak gösterilmiştir. Son olarak, video tanımlayıcı C3D modeli, ET modeli olarak eğitilmiştir. Video tanımlayıcı C3D modeli; I380K, Sports-1M [3] ve I380K'de ön eğitim (pre-training) yapılarak Sports-1M'de ince ayar yapılarak eğitilmiştir. En iyi sonucu ince ayar yapılan ağ vermiştir. Eğitin bu modeller UCF-101 [5] veri setinde değerlendirilmiş ve bu veri setinde son teknoloji (state-of-the-art) mimarilerinden daha iyi sonuç verdiği gösterilmiştir.

#### 1.3.2.2 Şişirilmiş iki-akış 3-B (I3D)

I3D [30] mimarisi, hem iki-akış mimarisinden [25,26] hem de Inception-V1 [37] 2-B CNN modelinden esinlenerek oluşturulmuştur. I3D ağı, Şekil 1.16'da Two-Stream 3D-ConvNet ismiyle gösterilmiştir. I3D mimarisinin girdi bölümleri iki-akış mimarisi ile aynıdır. İki-akış mimarisinin uzaysal akış ağı, girdi olarak tek RGB görüntüsü alıyorken, I3D uzaysal akış ağı, girdi olarak 64 RGB görüntüsü almaktadır. I3D zamansal akış ağıda girdi olarak 64 çerçeve almakta ve iki-akış ağında olduğu gibi çoklu optik akış görüntüsü almaktadır. Şekil 1.16'da gösterilen I3D modelinin 3D ConvNet katmanı: 2-B görüntü sınıflandırma Inception-V1 [37] modelinin 3-B uzaya uzatılmasıyla oluşturulmuştur. Ayrıca, I3D akışları ayrı ayrı eğitilmiş ve ağların tahmin sonuçlarının ortalaması alınarak sınıflandırma skoru oluşturulmuştur.




**Şekil 1.16 :** Birbirleriyle karşılaştırılan farklı video mimarileri [30].

I3D modeli ilk olarak, Inception-V1'in ImageNet'te [36] önceden eğitilmiş (pre-trained) olduğu ağırlık değerleri kullanılarak Şekil 1.16'da gösterilen a, b, c, d mimarileriyle, HMDB-51 [4], UCF-101 [5] ve miniKinetics test veri setlerinde sınıflandırılmıştır. Ancak, 3D-ConvNet modeli ayrı tutularak sıfırdan eğitilmiş ve test

edilmiştir. Sonuç olarak, I3D modeli tüm veri setlerinde en yüksek doğruluk oranlarına sahip olmuştur. Daha sonra, Şekil 1.16'daki modeller miniKinetics veri setinde ön eğitim yapılarak UCF-101 ve HMDB-51 test veri setlerini sınıflandırmış ve I3D modeli tekrardan en yüksek doğruluk oranlarına sahip olmuştur. MiniKinetics veri setinde ön eğitimin yapılma amacı, Kinetics-400 [38] veri setine göre küçük ölçekli olması ve bu sayede eğitim, test işlemlerini hızlandırmış olmasıdır. Son olarak, I3D modeli Kinetics-400 [38] ve miniKinetics veri setlerinde ön eğitim yapılarak, HMDB-51 ve UCF-101 veri setlerinde, son teknoloji (state-of-the-art) mimarileri (IDT [17], iki-akış ağı [25], C3D [29]) ile karşılaştırılmış ve en yüksek doğruluğa I3D Kinetics-400 ön eğitim modelinin sahip olduğu gösterilmiştir [30].

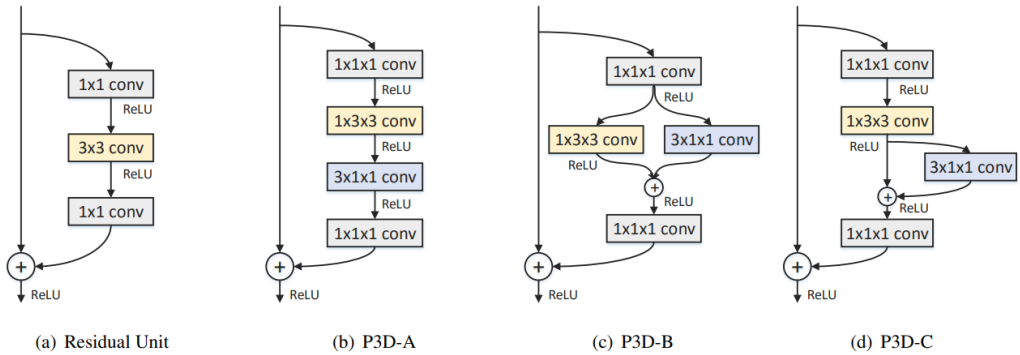
Sonuç olarak, 3-B CNN [29,30] modellerinin çok fazla parametreye sahip olmasından dolayı, hareket özelliklerini doğru öğrenebilmeleri için büyük miktarda doğru veriye ihtiyaç duydukları kanıtlanmıştır [30]. I3D modeli C3D modeline göre daha küçük veri setiyle eğitilmiş olmasına rağmen daha iyi sonuç vermiştir. Sebep olarak, Sports-1M veri setinin kirli ve doğru eylem başlıklarına sahip olmadığı ve mimari olarak I3D modelinin daha iyi olduğu rapor edilmiştir [30].

#### 1.3.2.3 Yalancı-3B (P3D)

P3D (Pseudo-3D) [31], 2-B Artık Ağlar'ı [39] temel almaktadır. 11 katmanlı C3D [29] ve 152 katmanlı 2-B Artık Ağ [39] modellerinin hafıza (memory) olarak çok yer kapladığı için çözüm olarak P3D modeli geliştirilmiştir. Sıfırdan çok derin 3-B CNN modelinin geliştirildiği düşünüldüğünde, çok sayıda parametre olacağından dolayı pahalı hesaplama maliyeti ve bellek talebi ile sonuçlanacağı belirtilmiştir [31]. Bunun önüne geçmek için P3D modelinde, 2-B uzaysal bölgede $3 \times 3 \times 3$ konvolüsyonları $1 \times 3 \times 3$ konvolüsyon filtrelerle değiştirilerek oluşturulmuştur. 1-B zamansal boyutu oluşturmak için 1-B CNN özelliği ile özellik haritalarında (feature maps) zamansal bağlantılar $3 \times 1 \times 1$ konvolüsyonları kullanılarak oluşturulmuştur. Yalancı 199 katmanlı bu model, 11 katmanlı C3D modelinden hafıza olarak daha az yer kaplamaktadır. Ayrıca, 2-B CNN görüntü sınıflandırma veri setlerinde ön eğitim sağlayarak sahne ve nesne bilgisi ile daha güçlü bir model oluşturulabilir [31].

Şekil 1.17'de temel alınan Artık darboğaz bloğu (bottleneck) [39] ve 3 çeşit P3D darboğaz blokları gösterilmiştir [31]. Artık ağın [39] ana fikri, kısa devre bağlantısıyla girdi bilgisinin çıkışa direkt aktarılması ve konvolüsyon katmanlarıyla işlenen girdinin

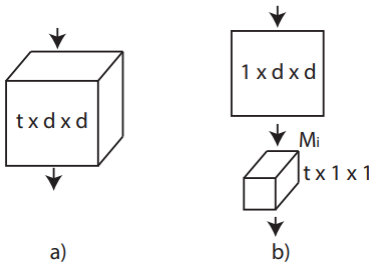
çıkış katmanında toplanmasıyla, girdide kaybolabilecek önemli bilgileri korumasıdır. P3D-A bloğu: 1-B ve 2-B konvolüsyon filtreleri birbirine kademeli şekilde bağlanmasıyla oluşmaktadır. Böylece, iki tür filtre birbirini doğrudan etkilemektedir. P3D-B bloğu: İki tür filtre arasında doğrudan bir bağlantı yoktur ve birbirini dolaylı olarak etkilemektedir. P3D-C bloğu: Filtreler birbirini hem doğrudan hem de dolaylı olarak etkilemektedir. Bloklar, UCF-101 [5] veri setinde test edilmiş ve en iyi sonucu P3D-A bloğu vermiştir. Son olarak, oluşturulan düşük maliyetli P3D modeli, UCF-101 veri setinde son teknoloji mimarilerinden (IDT [17], iki-akış ağı [25], TSN [28], C3D [29]) daha iyi sonuç vermiştir [31].




**Şekil 1.17 :** Artık Birim Bloğu (Residual Unit) ve P3D Blokları [31].

### 1.3.2.4 R(2+1)D

R(2+1)D [33] modelinde, artık öğrenme [39] çerçevesinde 3-B CNN'lerin 2-B CNN'lere göre doğruluk avantajları deneysel olarak gösterilmiştir. Oluşturulan konvolüsyon bloğu Şekil 1.18'de gösterilmiştir [33]. Artık ağlar [39], 3-B ve (2+1)D konvolüsyon bloklarıyla 18 ve 34 katmanlı olarak, Kinetics-400 [38] veri setinde 16 girdi alarak sıfırdan eğitilmiş ve test edilmiştir. R(2+1)D modeli [33], 3-B artık ağ modeline göre daha iyi sonuç vermiştir. Ayrıca, R(2+1)D modeli, Sports-1M [3] veri setinde de sıfırdan eğitilerek test edilmiştir. Ancak, Kinetics-400 veri seti, Sports-1M veri setine göre daha iyi sonuç vermiştir [33]. R(2+1)D mimarisi, Sports-1M [3] ve Kinetics-400 [38] veri setlerinde ön eğitim yapıldıktan sonra HMDB-51 [4] ve UCF-101 [5] veri setlerinde test edilmiş ve son teknoloji mimarilerine yakın veya daha yüksek doğruluğa sahip sonuçlar elde etmiştir. Sonuç olarak, 2-B uzaysal ve 1-B zamansal ayrılarak oluşturulan ve 3-B artık ağ ile aynı parametre boyutlarına sahip (2+1)D bloğunun performansta artış sağladığı gösterilmiştir.

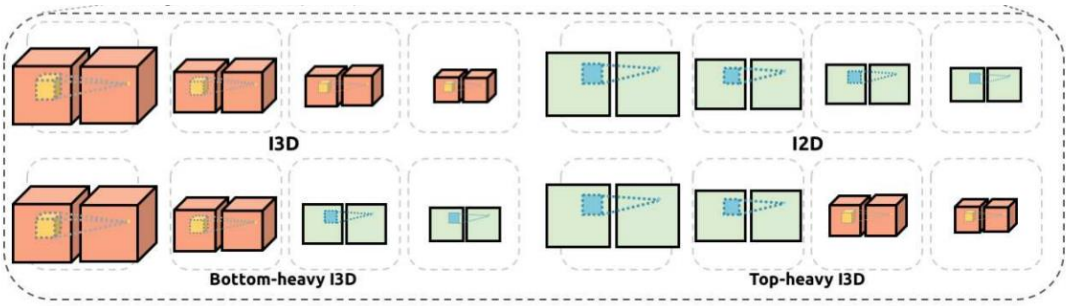



**Şekil 1.18 :** a) 3-B konvolüsyon bloğu b) Oluşturulan 2-B uzaysal ve 1-B zamansal konvolüsyon blok [33].

### 1.3.2.5 Ayrılabilir 3-B (S3D)

S3D [34] modeli I3D [30] modelini temel almakta, 2-B ve 3-B konvolüsyon filtrelerden oluşmaktadır. S3D modelinde kullanılan 3-B konvolüsyon filtreler, P3D [31] ve R(2+1)D [33] modellerinde olduğu gibi 2-B uzaysal ve 1-B zamansal boyuta ayrılarak oluşturulmuştur. Böylece, I3D modelindeki bazı 3-B filtreler, hafıza olarak az yer kaplayan 2-B filtrelere dönüştürülmüştür.

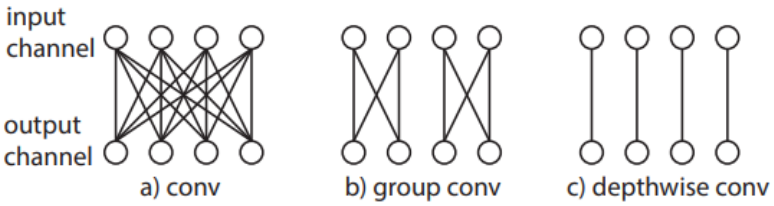
Çalışmada, dört farklı model varyantı incelenmiş ve bu modeller: I3D [30], I2D, Bottom-heavy, Top-heavy olarak Şekil 1.19'da gösterilmiştir. I3D [30] modelinin orijinali kullanılmıştır. I2D modeli, I3D modelinin 2-B modele dönüştürülmüş hali olarak ve çoklu çerçeve girdisi alarak oluşturulmuştur. Bottom-heavy I3D modeli, düşük katmanlarda 3-B ve yüksek katmanlarda 2-B konvolüsyon filtreler kullanmaktadır. Top-heavy I3D ise Bottom-heavy modelinin tam tersidir, yüksek katmanlarda 3-B ve düşük katmanlarda 2-B konvolüsyon filtreler kullanmaktadır. I3D ve oluşturulan I2D modeli, Kinetics-400 [38] veri setinde karşılaştırılmıştır. I3D modeli daha başarılı olmuştur. Top-heavy ve Bottom-heavy modelleri miniKinetics veri setinde karşılaştırıldığında, Top-heavy modeli daha başarılı gelmiştir. Bunun sebebi ise girdiden gelen önemsiz yoğun bilgilerin 2-B filtreler ile es geçilmesiyle, yüksek katmanda daha zengin bilgi içermesidir. Böylece, 3-B konvolüsyon filtrelerin, yüksek katmanda uzam-zamansal örüntüleri daha başarılı çıkardığı gösterilmiştir [34]. Son olarak S3D modeli, ImageNet ve Kinetics-400 [38] ile ön eğitim yapıldıktan sonra RGB girdi ile HMDB-51 [4], UCF-101 [5] veri setlerinde, C3D [29], I3D [30], P3D [31] ve R(2+1)D [33] modellerinden daha yüksek doğruluk oranına sahip olduğu gösterilmiştir [34].




**Şekil 1.19 :** İncelenen dört farklı model: I3D [30], I2D, Bottom-heavy I3D ve Top-heavy I3D [34].

### 1.3.2.6 Kanalla ayrılmış evrişimsel ağlar (CSN)

CSN [35] modelleri 101 ve 152 katmanlı olarak oluşturulmuştur. Bu çalışma, video sınıflandırması için 3-B CNN ağlarda farklı tasarım seçeneklerinin etkilerini incelemektedir. 3-B konvolüsyonları, uzaysal 2-B ve zamansal 1-B olarak ayrılmıştır. Farklı kanal bağlantıları kullanılmıştır. Kanallar: Conv, group conv ve depthwise conv olarak Şekil 1.20'de gösterilmiştir. Conv bağlantı: Yoğun bağlantı içerir ve her konvolüsyon filtresi, önceki katmanından girdi almaktadır. Group conv bağlantı: Konvolüsyon filtreleri alt küme halinde gruplayarak dağıtılmıştır. Bir alt kümedeki filtreler, yalnızca kendi grubundaki kanallardan sinyal almaktadır. Depthwise conv: Grup sayısının giriş ve çıkış kanallarının sayısına eşit olmasıdır. Group conv ve depthwise conv'un amacı daha iyi doğruluk ve daha düşük hesaplama maliyeti sağlamasıdır [35]. CSN konvolüsyonu, 3-B konvolüsyonlara kıyasla daha düşük eğitim doğruluğuna sahiptir. Ancak, daha yüksek test doğruluğu sağlamıştır [35]. CSN modeli, Sports-1M [3] ve Kinetics-400 [38] veri setlerinde, son teknoloji mimarilerinden daha iyi sonuç verdiği gösterilmiştir [35]. Ayrıca, CSN modeli mevcut ağlardan birkaç kat daha hızlıdır.




**Şekil 1.20 :** Giriş, çıkış kanal bağlantıları [35].

### 1.4 Eylem Tanıma Veri Setleri

Öne çıkan ET veri setleri: Youtube-8M [2], Sports-1M [3], HMDB-51 [4], UCF-101 [5] ve Kinetics'tir [38,40,41]. ET veri setlerine ait örnek video sayıları ve toplam sınıfları Çizelge 1.1'de gösterilmiştir.

**Çizelge 1.1 :** ET veri setleri.

| Veri Seti Adı | Örnek Video Sayısı | Sınıf Sayısı |
|---|---|---|
| Youtube-8M | 8264650 | 4800 |
| Sports-1M | 1000000 | 487 |
| HMDB-51 | 6766 | 51 |
| UCF-101 | 13320 | 101 |
| Kinetics-400 | 306245 | 400 |
| Kinetics-600 | 495547 | 600 |
| Kinetics-700 | 650317 | 700 |

#### 1.4.1 Youtube-8M

Youtube-8M [2] veri seti, 4800 sınıf içermekte ve yaklaşık 8 milyondan fazla video klipten oluşmaktadır. Her sınıf, alt küme sınıflarına ayrılmakta ve alt küme sınıfı en az 200 video içermektedir. Veri kümesindeki her etiket, yalnızca görsel bilgiler kullanılarak ayırt edilebilmektedir. Video klipleri ortalama 2,5 saniyeden oluşmaktadır. Veri alt kümesindeki sınıflar, faaliyetleri (spor, oyunlar, hobiler), nesneleri (otomobiller, yiyecekler, ürünler), sahneleri (seyahat) ve olayları kapsamaktadır [2].

#### 1.4.2 Sports-1M

Sports-1M [3] veri seti, 487 sınıf içermekte ve açıklamalı YouTube'den alınmış 1 milyon video klibinden oluşmaktadır. Sınıflar: Su sporları, takım sporları, kış sporları, top sporları, dövüş sporları, hayvanlarla sporlar gibi sınıflar içermekte ve manuel olarak seçilmiş bir sınıflandırmada düzenlenmiştir [3]. Sınıf başına yaklaşık 1000-3000 arası videoya sahiptir. Videolar yarım saniyelik 100 çerçeveden (frame) oluşmaktadır.

#### 1.4.3 HMDB-51

HMDB-51 [4] veri seti, 51 sınıf içermekte ve yaklaşık 6 binden fazla video klibinden oluşmaktadır. Sınıflar beş kategoriye ayrılmıştır. Kategoriler: Genel yüz hareketleri (gülmek, çiğnemek, konuşmak), nesne yüz eylemleri (sigara içmek, yemek, içmek),

genel vücut hareketleri (el çırpma, tırmanış, dalış), nesne etkileşimli vücut hareketleri (bisiklete binme, at binme, golf) ve insan etkileşimi için vücut hareketlerinden (yumruk atmak, el sıkışmak, sarılmak) oluşmaktadır [4]. Sınıflar en az 101 video klibi içermekte ve ortalama en az 1 saniye uzunluğundadır.

### 1.4.4 UCF-101

UCF-101 [5] veri seti, 101 sınıf içermekte ve yaklaşık 13 binden fazla video klibinden oluşmaktadır. Sınıflar beş kategoriye ayrılmıştır. Kategoriler: İnsan nesne etkileşimi (göz makyajı yapmak, ruj sürmek, diş fırçalamak), yalnızca vücut hareketi (şınav çekmek, barfiks çekmek, ip tırmanma), insanla insan etkileşimi (saç kesme, kafa masajı, bando yürüyüşü), müzik aletleri çalma (gitar çalmak, piyano çalmak, keman çalmak), spordan (bisiklete binme, basketbol oynamak, at binme) oluşmaktadır [5]. Video klipleri ortama 7 saniyedir.

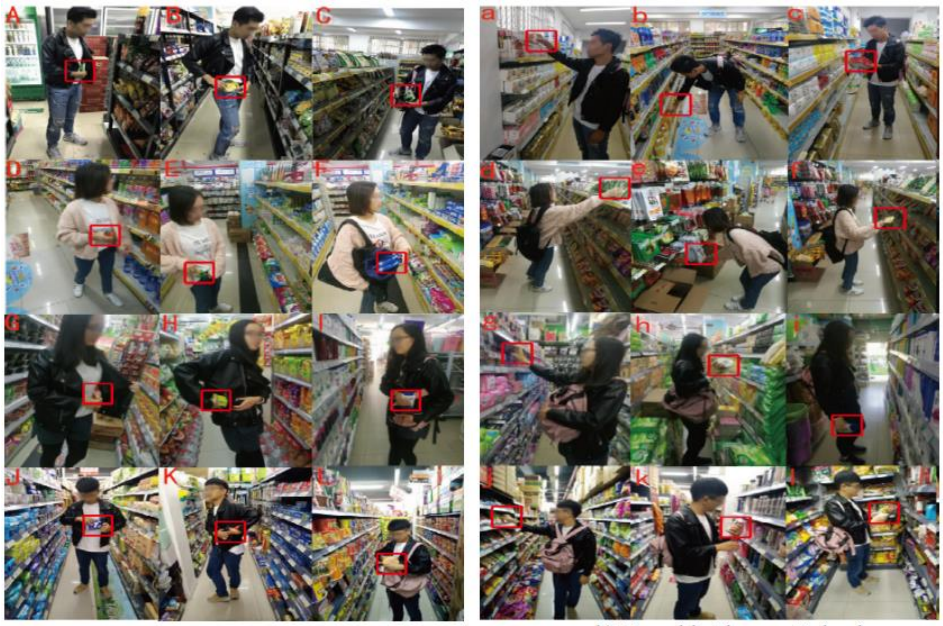
### 1.4.5 Kinetics

Kinetics veri setleri, 400, 600 ve 700 sınıflı olmak üzere üç tanedir [38,40,41]. Kinetics-400 veri seti [38], yaklaşık 300 binden fazla video klibinden oluşmaktadır. Kinetics-400 insan eyleminden oluşan 400 sınıflı bir veri setidir. Sınıf başına 400-1150 arası videoya sahiptir. Video uzunluğu ortalama 10 saniyedir. Sınıflar genel olarak üç kategoriye ayrılmaktadır. Kategoriler: Tekil kişi eylemleri (robot dansı, bacak germe), insanla insan eylemleri (el sıkışmak, gıdıklamak), kişi nesne eylemlerinden (bisiklete binme) oluşmaktadır [38]. Kinetics-600 veri seti [40], yaklaşık 500 binden az video klibinden oluşmaktadır. Kinetics-700 veri seti [41], yaklaşık 650 binden fazla video klibinden oluşmaktadır. Kinetics-400 veri seti eylemleri daha da detaylandırılarak Kinetics-600 [40] ve Kinetics-700 [41] veri setleri oluşturulmuştur.

## 1.5 Hırsızlık Tespiti Çalışmaları

Zhang ve ark. çalışmalarında [6], yüksek çözünürlüklü bir kamera ile hırsızlık ve hırsızlık dışı eylemler sınıflandırılmış ve 2-B görüntü sınıflandırma CNN modeli ile birleştirilmiştir. Sistemin girişi 100 x 100 RGB görüntüdür ve sistemin 2-B CNN yapısı LeNet-5 [22] modeline dayanmaktadır. Araştırmacılar veri toplamada, eğitimde kişisel faktörlerin etkisini ortadan kaldırmak için dört farklı kişi (uzun boylu, kısa) faktörü seçmişlerdir. Örnek olarak, her kişi üç farklı kombin giymiştir. Ayrıca,

hırsızlık davranışını daha iyi tanımlamak için sırt çantası üç farklı konumda (arkada, önde ve yanda) kullanılmıştır. Her kişi, raftaki ürünü rastgele seçerek ürün faktörünün düşürülmesi hedeflemektedir. Hırsızlık davranışı üç durumdan oluşur: Eşyaları; omuz altına sokmak, cebe sokmak ve sırt çantasına koymak. Hırsızlık olmayan davranış da üç durumdan oluşur: Raftan eşya almak, rafta eşya aramak ve eşya taşımak. Eğitim aşamasında, görüntülerin yarısı hırsızlık, diğer yarısı ise hırsızlık dışı olmak üzere toplam 10800 görüntü kullanılmıştır. Hırsızlık eylemleri ve hırsızlık olmayan eylemler Şekil 1.21'de gösterilmiştir [6]. Deney aşaması beş aşamada tamamlanmıştır. Aşama 1-3, dört tür eğitim modeline sahiptir. Bu nedenle, cinsiyet faktörlerinden kaçınmak için dört kişi birbirinden ayrı olarak değerlendirilmiştir. Aşama 1-3, üç farklı kombin ve sırt çantası kullanarak kıyafet faktörünü incelemektedir. Aşama 4, aynı cinsiyetten farklı boyda ve aynı boyda farklı cinsiyette olmak üzere üç tip eğitim modelinde kişilerin boy ve cinsiyet faktörünü değerlendirmiştir. Aşama 5, tüm veri setini içermektedir. Deneysel sonuçlar, her aşama için doğruluğun arttığını, yani eleme faktörlerinin işe yaradığını göstermektedir. Aşama 5, en iyi model ve %83 doğruluk oranına sahiptir. Araştırmacılar, doğruluk oranını artırmak için modelin daha fazla deneysel veriye ihtiyaç duyduğunu, CNN yapı modelinin daha derin olabileceğini ve daha yüksek çözünürlüklü kameraların kullanılabileceğini belirtmiştir [6].



a) Hırsızlık Eylemleri b) Hırsızlık Olmayan Eylemler

**Şekil 1.21 :** a) Hırsızlık eylemleri b) Hırsızlık olmayan eylemler [6].

Diğer bir çalışmada [7], süpermarket kameralarından otomatik olarak hırsızlık davranışlarını tespit eden bir sistem oluşturulmuştur. Tüm çerçeveden özellikler

çıkarmak yerine, gerekli özellikleri daha doğru bir şekilde vurgulamak için İlgi Bölgesi (ROI) optik akış kaynaşma ağı kullanılmıştır. Çıkarılan tüm görüntülerin optik akışı ve ROI'nin optik akışı Şekil 1.22'de gösterilmiştir [7]. Sistem, anormal davranışları tespit etmek için kullanıcı hareketlerinin optik akışını dikkate almaktadır. Önce, Mask-R-CNN kullanılarak bir kişi nesnesini ROI olarak çıkarmakta ve sonra optik akışa çevrilmektedir. Optik akış görüntüsünde, ROI kişi nesnesindeki değişiklik miktarı kullanılarak hırsızlık belirlenmektedir. Model girdi olarak 3 ROI çerçeve almaktadır. Hırsızlık eylemleri: Eşyaları; çantaya, omuz altına ve cebe koymak. Hırsızlık olmayan eylemler ise hırsızlık eylemleri dışındaki tüm eylemler olarak tanımlanmıştır. Veri seti örnek sayısı ve çalışma sonuçları hakkında bir bilgi verilmemiştir.


**Şekil 1.22 :** Tüm çerçeve optik akışı ve ROI'nin optik akışı [7].

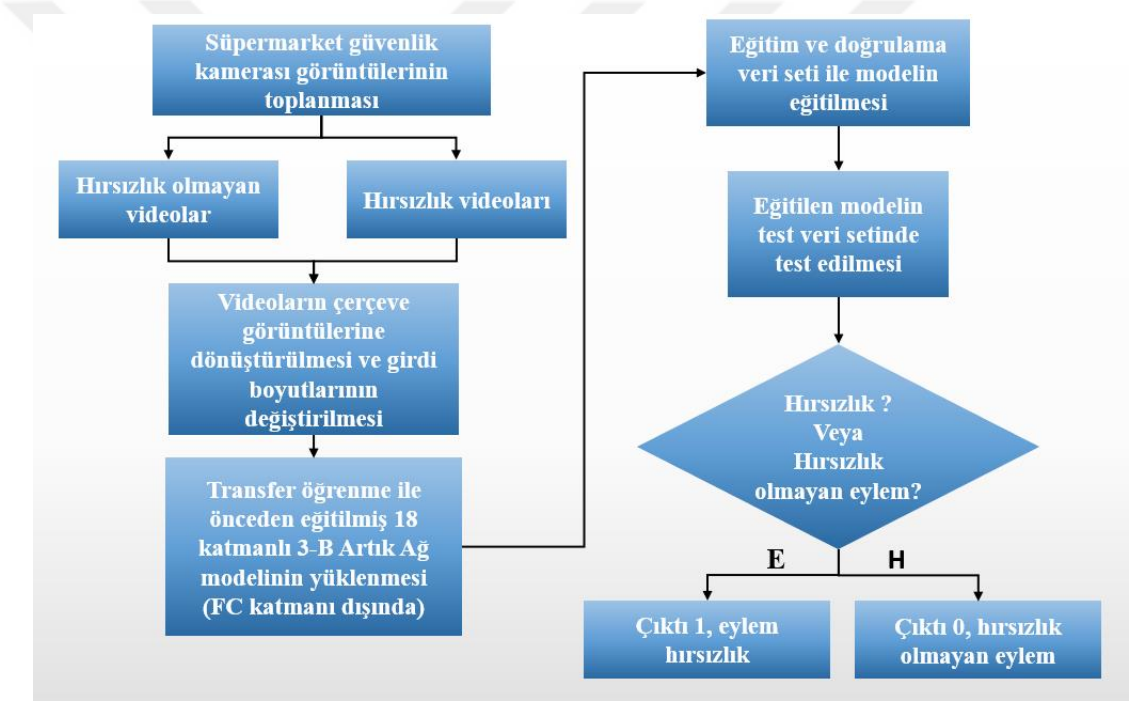
Süpermarket kameralarından otomatik olarak hırsızlık davranışlarını tespit eden başka bir sistem 3-B iki akış modeli kullanan bir çalışma [8] ile oluşturulmuştur. Burada 11 sınıflı bir veri seti oluşturulmuştur. Sınıfların 6 tanesi normal müşteri olarak seçilmiştir. Hırsızlık yapan kişi normal hareketleri de sergileyeceği için 11 sınıfın tüm hepsi hırsızlık eylemine dahil edilmiştir. Sınıflar: Markete girmek, yürümek, marketi gözetmek, kameraya bakmak, raftan eşya almak, eşyayı cebe sokmak, eşyayı çantaya sokmak, eşyayı market arabasına koymak, eşyayı geri rafa koymak ve eşyayı omuz altına sokmak olarak belirlenmiştir. Her bir eylem sınıfı için veri sayısı 90-95'tir ve 11 eylem için toplam veri sayısı yaklaşık 1000'dir. Veri ön işleme adımında, video akışlarındaki kişilerin gerçek zamanlı tespiti için YOLO v3 algoritması kullanılmıştır. Ardından, tespit edilen kişi tarafından gerçekleştirilen eylemleri tespit etmek için 3-B

iki-akış CNN modeli kullanılmıştır. Girdi olarak, çok çerçeveli yoğun optik akış kullanılmıştır. 11 eylem sınıfı için ortalama doğruluk oranı %85'tir.

**Şekil 1.23 :** Veri seti örnekleri [8].

## 2. MATERYAL VE YÖNTEM

3-B artık ağlar [32], 2-B artık ağlardan [39] esinlenerek oluşturulmuştur. 3-B artık ağ modeli 3-B konvolüsyon katmanlarını kullanarak eylemlerin uzam-zamansal bilgilerini öğrenmektedir. Bu tez çalışmasında, süpermarket hırsızlık eylemini tespit etmek için 3-B artık ağ modeli temel alınmıştır. Materyal ve yöntem için izlenen yol Şekil 2.1'de gösterilmektedir.

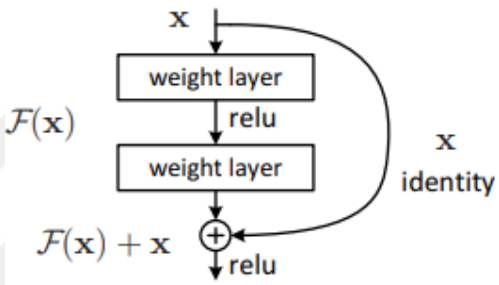



**Şekil 2.1 :** Sistem akış diyagramı.

Şekil 2.1'de görüldüğü üzere: ilk olarak veri seti için süpermarket güvenlik kamerası videoları toplanmıştır. Toplanan videolar, hırsızlık ve hırsızlık olmayan eylemler olarak iki sınıfa ayrılmıştır. Sınıf videoları, çerçeve görüntülerine dönüştürülerek görüntü boyutları model girdilerine uygun hale dönüştürülmüştür. Önceden eğitilmiş 3-B artık ağ modelinde FC katmanı hariç olarak, transfer öğrenme yapılarak yeni 3-B artık ağ modeli oluşturulmuştur. Model, eğitim ve doğrulama veri seti ile eğitilmiştir. Daha sonra, eğitilen model test veri seti ile test edilmiştir. Eğer eylem hırsızlık ise sistem çıktı olarak 1, eylem hırsızlık değilse sistem çıktı olarak 0 vermektedir.

## 2.1 2-B Artık Ağlar

Derin ağlar, düşük, orta, yüksek seviye özellikleri ve sınıflandırıcıları uçtan uca çok katmanlı bir tarzda entegre etmekte ve özelliklerin seviyeleri katmanların sayısı (derinlik) ile zenginleştirilebilmektedir [39]. Böylece, artan katman sayısı ile doğruluk oranı da artmaktadır. Ancak, derin ağlar yakınsamaya (converging) başladıkça bir bozulma sorunu ortaya çıkmaktadır. Ağ derinliği arttıkça doğruluk oranı doyuma ulaşmakta ve doğruluk oranı hızlıca düşmektedir. Böylece, yüksek katman sayısı daha yüksek eğitim hatasına yol açmaktadır. Çözüm olarak, Şekil 2.2'deki derin artık öğrenme bloğu geliştirilmiş ve bu blok ile kimlik (identity) eşleşmesi kısa devre bağlantıları ile oluşturulmuştur [39].




**Şekil 2.2 :** Artık öğrenme bloğu [39].

Artık öğrenme, çıktının ve girdinin farkının alınmasıyla oluşmakta ve bu işlem denklem 2.1'de ifade edilmektedir.

$$F(x) := H(x) - x \tag{2.1}$$

Denklem 2.1'de görüleceği üzere, *F(x)* artık öğrenme, *H(x)* artık öğrenme bloğunun bulmayı hedeflediği gerçek çıktı, *x* ise girdiyi ifade eder. Denklem 2.1 tekrar düzenlendiğinde, artık öğrenme Denklem 2.2'deki gibi ifade edilebilir.

$$H(x) = F(x) + x \tag{2.2}$$

Denklem 2.2'de görüleceği üzere, yığılmış katmanların *H(x)* değerini yakınsaması yerine, *H(x)* değeri artık öğrenme bloğu ile girdi katmanlarından işlemden geçen *F(x)* ve girdiden çıkışa kimlik kısa devre bağlantısı ile aktarılan *x*'in toplanmasıyla, ResNet ağları *H(x)* değerini daha kolay bir şekilde yakınsamaktadır.

Şekil 2.2'deki artık öğrenme bloğunun tüm mimarideki artık öğrenme bloklarını temsil ettiği formül, Denklem 2.3'te ifade edilmektedir.

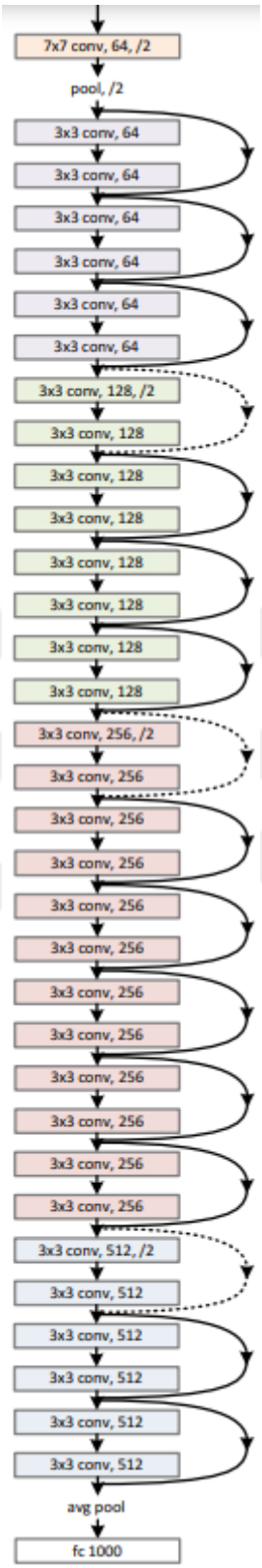
$$H(x) = F(x, \{W_i\}) + x \tag{2.3}$$

Denklem 2.3'te görüleceği üzere, $F(x,\{W_i\})$ fonksiyonu öğrenilecek artık haritalamayı temsil etmekte ve $x$ katmanın girdisini temsil etmektedir.

Ayrıca, Denklem 2.3'ün tek bir artık öğrenme bloğu olarak ifadesi Şekil 2.2'de gösterilmektedir. Denklem 2.3'ün tek bir blok olarak ifade edildiği formül, Denklem 2.4'te ifade edilmektedir.

$$H(x) = F(x, \{W_i\}) + x = (W_1 x) W_2 \sigma + x \tag{2.4}$$

Denklem 2.4'te görüleceği üzere, $W_1$ ve $W_2$ ağırlık katmanlarını, $\sigma$ ReLu (Rectified Linear Units) aktivasyon fonksiyonunu temsil etmektedir. Ayrıca, 2-B artık ağ blokları, ekstra parametre ve hesaplama karmaşıklığı eklememektedir. Tüm ağ, geri yayılım ile uçtan uca eğitilebilmektedir. Sonuç olarak, başarılı artık ağ modeli oluşturulmuş ve 34 katmanlı mimari Şekil 2.3'te gösterilmiştir.



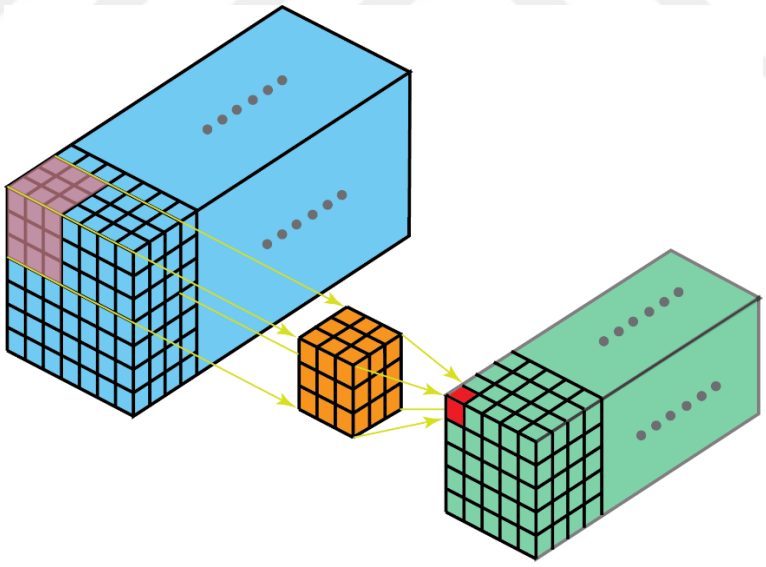

**Şekil 2.3 :** 34 katmanlı artık ağ mimarisi [39].

## 2.2 3-B Konvolüsyon İşlemi

3-B konvolüsyonda filtre derinliği giriş katmanı derinliğinden daha küçüktür (çekirdek boyutu < kanal boyutu) [42]. Böylece, 3-B filtre, görüntünün; yüksekliğinde, genişliğinde, kanalında ve zamansal boyutunda, konvolüsyon işlemini gerçekleştirmektedir. 3-B uzam-zamansal görüntü hacmine uygulanan 3-B filtre her seferinde 1 değer sağlamaktadır. Bu işlem Şekil 2.4'te gösterilmiştir [42]. 3-B konvolüsyon filtre işlemi sonucu, 3-B uzam-zamansal kodlanmış veri oluşmaktadır. 3-B konvolüsyon filtre işlemi Denklem 2.5'te ifade edilmektedir.

$$\text{3 B uzam zamansal kodlanmış veri} = \mathrm{C} \times \mathrm{T} \times \mathrm{H} \times \mathrm{W} * \mathrm{k} \times \mathrm{k} \times \mathrm{k} \tag{2.5}$$

Denklem 2.5'te görüleceği üzere, C kanal boyutu (RGB görüntü için 3 kanal), T çerçeve uzunluğu (temporal length), H görüntünün yüksekliği (image height), W görüntünün genişliği (image width), $*$ konvolüsyon operatörü ve $\mathrm{k} \times \mathrm{k} \times \mathrm{k}$'de çekirdek (kernel) konvolüsyon filtresini ifade etmektedir. Ayrıca, Denklem 2.5'teki $\mathrm{C} \times \mathrm{T} \times \mathrm{H} \times \mathrm{W}$, ağa girdi olarak verilen 4 boyutlu tensörü ifade etmektedir.



**Şekil 2.4 :** 3-B görüntü hacmine uygulanan 3-B çekirdek filtre [42].

## 2.3 3-B Artık Ağlar

Hara vd. tarafından 3-B artık ağ modeli, 2-B artık ağ modelini temel alarak oluşturulmuştur [32]. 3-B artık ağ [32] modeli, 2-B artık ağ [39] modelinin tüm özelliklerini korumaktadır. 3-B artık ağ, 2-B artık ağın sadece zamansal boyuta

uzatılmış halidir. 3-B CNN'ler ET veri setlerinden direkt olarak uzam-zamansal bilgiyi çıkardıkları için görüntü sınıflandırma 2-B CNN modellerine göre daha kullanışlıdır [32]. 3-B CNN'lerin çok sayıda parametre içermesinden dolayı, ağı optimize etmek için büyük ölçekli veri setlerine ihtiyaç duymaktadır [32].

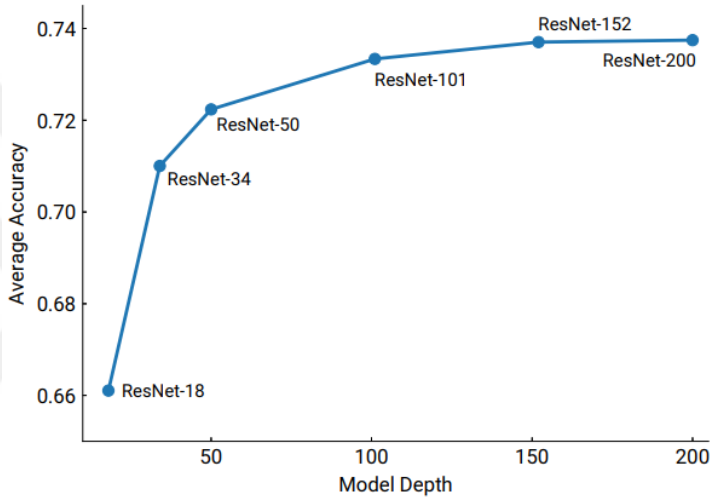
Oluşturulan 18 ve 34 katmanlı 3-B artık ağ modelleri ayrıntılı olarak Şekil 2.5'te gösterilmiştir [32]. 3-B artık ağ, girdi olarak 16 çerçeve RGB görüntü almaktadır. Şekil 2.5 incelendiğinde, girdi ilk olarak conv1 katmanında uzaysal atlama (stride) ile alt örnekleme (down-sampling) yapılarak görüntü çözünürlüğü yarıya düşürülmektedir. Ancak, ilk katmanda zamansal olarak alt örnekleme yapılmamaktadır. Girdilere alt örnekleme işlemi, özellik haritalarının 128, 256 ve 512 olarak arttığı, conv3_x, conv4_x ve conv5_x katmanlarında hem uzaysal hem de zamansal boyutta gerçekleşmektedir. Conv5_x katmanı çıktısı, ortalama havuzlama filtresinden geçtikten sonra FC katmana verilmektedir. FC katmanındaki bilgi softmax fonksiyonu ile olasılığı oluşturulduktan sonra girdinin sınıflandırılması yapılmaktadır.

| Katman Adı | Mimari | |
|---|---|---|
| | 18-katman | 34-katman |
| conv1 | $7 \times 7 \times 7, 64$, atlama 1 (T), 2 (XY) | |
| conv2_x | $3 \times 3 \times 3$ maksimum havuzlama, atlama 2<br>$\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 2$ | $\begin{bmatrix} 3 \times 3 \times 3, 64 \\ 3 \times 3 \times 3, 64 \end{bmatrix} \times 3$ |
| conv3_x | $\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 2$ | $\begin{bmatrix} 3 \times 3 \times 3, 128 \\ 3 \times 3 \times 3, 128 \end{bmatrix} \times 4$ |
| conv4_x | $\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 2$ | $\begin{bmatrix} 3 \times 3 \times 3, 256 \\ 3 \times 3 \times 3, 256 \end{bmatrix} \times 6$ |
| conv5_x | $\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 2$ | $\begin{bmatrix} 3 \times 3 \times 3, 512 \\ 3 \times 3 \times 3, 512 \end{bmatrix} \times 3$ |
| Ortalama havuzlama , 400-d fc, softmax | | |

**Şekil 2.5 :** 18 ve 34 katmanlı 3-B artık ağ modelleri [32].

18, 34, 50, 101 ve 152 katmanlı artık ağ modelleri oluşturulmuş ve bu ağlar HMDB-51 [4], UCF-101 [5], ActivityNet [43] ve Kinetics-400 [38] veri setlerinde sıfırdan eğitilmiştir. HMDB-51, UCF-101 ve ActivityNet veri setlerinde sıfırdan eğitilen ağların, küçük ölçekli veri seti olmasından dolayı doğruluk oranın hızla düşerek ezberlemeye yöneldiği (overfitting) gösterilmiştir. ActivityNet veri seti, Kinetics-400 veri setine göre daha az doğruluk oranına sahip olmuş ve Kinetics-400 veri setinin 3-

B CNN'leri eğitmek için daha iyi bir veri seti olduğu gösterilmiştir. Böylece, Kinetics-400 veri setinin parametreleri optimize etmek için yeterli veri büyüklüğüne sahip olduğu, sıfırdan eğitilen ağların sonuçlarıyla kanıtlanmıştır. 18, 34, 50, 101 ve 152 katmanlı artık ağların, Kinetics-400 veri seti doğrulama (validation) grafiği Şekil 2.6'te gösterilmiştir. Şekil 2.6'te görüleceği üzere, artık ağ katman sayısı arttıkça doğruluk oranı da artmaktadır. 152 ve 200 katmanlı artık ağlarda doğruluk oranı doyum noktasına gelmiştir. Ayrıca, Kinetics-400'de ön eğitilmiş model, transfer öğrenme yapılarak HMDB-51 ve UCF-101 veri setlerinde test edilmiştir. Sonuç olarak, kompleks 2-B mimarilerinden daha iyi performans göstermiştir.




**Şekil 2.6 :** 18, 34, 50, 101 ve 152 katmanlı artık ağların kinetics-400 doğrulama veri seti sonuçları [32].

## 2.4 Hırsızlık Tespiti Yapan 3-B Artık Ağ Model Versiyonları

Bu tezde oluşturulan veri seti videoları, hırsızlık ve hırsızlık olmayan eylemler için 10-32 arasında RGB görüntü çerçevesinden oluşmaktadır. Videoların farklı çerçeve aralığına sahip olmasından dolayı, süpermarketteki hırsızlık ve hırsızlık olmayan eylemleri incelemek için dört farklı 3-B artık ağ mimarisi oluşturulmuştur. Çerçeve etkisini incelemek için girdi olarak 16 ve 32 çerçeve girdisine sahip versiyonlar oluşturulmuştur. Ayrıca, çözünürlüğün etkisini incelemek için 112×112 ve 224×224 çözünürlüklerine sahip versiyonlar oluşturulmuştur. Oluşturulan 18 katmanlı 3-B artık ağ mimarisi Çizelge 2.1'de detaylı olarak gösterilmektedir.

### 2.4.1 Farklı girdi çerçevesine ve çözünürlüğe sahip 3-B artık ağ mimarileri

Hırsızlık tespiti için oluşturulan 3-B model, 18 katmanlı 3-B artık ağ [32] mimarisinin aynısıdır. Modele ait farklı parametrelere sahip 12 versiyon oluşturulmuştur. Oluşturulan 12 versiyon, 18 katmanlı 3-B artık ağ modelinden oluşmaktadır. Versiyonlar sadece; girdi boyutu, çerçeve uzunluğu, parti büyüklüğü olarak değişiklik göstermekte yani parametre olarak farklılık göstermektedir.

Çizelge 2.1 modeli girdisi: C × T × H × W şeklindedir. C kanal sayısı (number of channels), T çerçeve uzunluğu (temporal length), H görüntünün yüksekliği, W görüntünün genişliğini temsil etmektedir.

Versiyonlar ilk olarak, 3 kanallı (C) RGB video girdi serisini, 16 veya 32 çerçeve (T) olarak ve 112×112 (H × W) veya 224×224 (H × W) girdi boyutunda almaktadır. Çizelge 2.1 girdisi, kanal sayısı 3, çerçeve uzunluğu 32, girdi görüntü boyutu 224 × 224 olduğu durum için geçirli olarak katman girdi ve çıktıları oluşturulmuştur. Conv1 katmanı konvolüsyon çekirdek filtre boyutu 7×7×7'dir. Conv1 katmanı dışındaki katmanların konvolüsyon çekirdek (kernel) filtre boyutu 3×3×3'tür. Her konvolüsyon katmanından sonra, toplu normalleştirme [44] (BN) (batch normalization) ve ReLu katmanları gelmektedir. BN, katman girdilerini yeniden ölçeklendirerek normalize etmektedir. Böylece ağın eğitimini hızlandırmakta ve sabit tutmaktadır. Ayrıca BN ile eğitim yakınsamasından ödün vermeden daha geniş bir öğrenme aralığı (learning rate) kullanabilmektedir. ReLu, doğrusal bir aktivasyon fonksiyonudur. ReLu fonksiyonu girdi değerleri pozitif ise çıktı olarak girdi değerleri kendi değerini korumaktadır. Ancak, girdi değerleri negatif ise çıkış değerleri sıfır olmaktadır. Artık ağ kısa devre bağlantısı olarak shortcut type B [39] kullanılmıştır. Shortcut type B: Girdi ve çıktı katmanlarında oluşabilen herhangi bir uyuşmazlıkta boyutları uygun hale getiren kısa devre bağlantı tipidir [39]. Conv1 katmanı, uzam-zamansal görüntü seri girdisine sadece uzaysal boyutta 1×2×2 atlama ile alt örnekleme yaparak girdi boyutunu zamansal olarak yarıya düşürmektedir. Girdi Pooling katmanına verilir ve uzam-zamansal olarak alt örnekleme yapılır. Özellik haritalarının (feature maps) 128, 256 ve 512 olarak arttığı Conv3, Conv4 ve Conv5 katmanlarında, 2×2×2 atlama ile uzam-zamansal girdiyi hem uzaysal hem de zamansal olarak alt örnekleme uygulanarak girdi boyutu uzam-zamansal olarak yarıya düşmektedir. Conv5 katmanı çıktısı FC katmanına verilmektedir. Son olarak, FC katmanında 2 sınıflı veri setimiz, hırsızlık

veya hırsızlık olmayan eylem olarak softmax aktivasyon fonksiyonu ile bir olasılık değeri üreterek eylemi sınıflandırmaktadır.

**Çizelge 2.1 :** 32 çerçeve girdisi alan 224×224 çözünürlüğe sahip 18 katmanlı 3-B artık ağ mimarisi.

| Katman adı | Girdi | Katman detayları | Çıktı |
|---|---|---|---|
| Conv1 | C × T × H × W | 7×7×7, atlama 1×2×2 | 64×32×128×128 |
| Pooling | 64×32×128×128 | 3×3×3 maksimum havuzlama ile atlama 2 | 64×16×64×64 |
| Conv2 | 64×16×64×64 | $\begin{pmatrix}3\times3\times3\\3\times3\times3\end{pmatrix}\times2$ | 64×16×64×64 |
| Conv3 | 64×16×64×64 | $\begin{pmatrix}3\times3\times3\\3\times3\times3\end{pmatrix}\times2$ | 128×8×32×32 |
| Conv4 | 128×8×32×32 | $\begin{pmatrix}3\times3\times3\\3\times3\times3\end{pmatrix}\times2$ | 256×4×16×16 |
| Conv5 | 256×4×16×16 | $\begin{pmatrix}3\times3\times3\\3\times3\times3\end{pmatrix}\times2$ | 512×2×8×8 |
| FC katmanı | 512×2×8×8 | Ortalama havuzlama 2 - d FC softmax | 2×1 |

### 2.4.2 Uygulama

Uygulama ayarları genel olarak [45]'ten alınmıştır. Bu tezde oluşturulan hırsızlık veri seti, 3-B CNN ağını sıfırdan eğitmek için küçük ölçekli olmasından dolayı sıfırdan eğitime uygun değildir. Bu sebeple, büyük ölçekli Kinetics-700 [41] veri setinde ön eğitim yapılmış 18 katmanlı 3-B artık ağ [45] transfer öğrenme ile kullanılmıştır. Transfer öğrenme ile FC katmanı dışındaki diğer katmanlar, ön eğitim yapılmış modelin ağırlıklarını kullanmaktadır. Böylece, oluşturulan yeni hırsızlık ağ mimarisi eğitiminde sadece FC katman ağırlıkları güncellenmiştir. Oluşturulan modelin 12 versiyonu için eğitim, doğruluk ve test uygulamaları PyCharm Community Edition 2020.3.3 programı ve Pytorch kütüphanesi kullanılarak uygulanmıştır. Uygulamada kullanılan ekran kartı NVIDIA GeForce GTX 1070'tir. Eğitilen versiyonlar: farklı girdi çözünürlüğü, farklı çerçeve uzunluğu (temporal length) ve farklı parti büyüklüğüne (batch size) sahiptir. Girdi çözünürlüğü 112×112 ve 224×224'tür. Çerçeve uzunluğu 16 ve 32'dir. Parti büyüklüğü 8, 10 ve 12'dir. Modele girdi olarak verilen videonun çerçeve uzunluğu eğer uygulanan çerçeve uzunluğundan küçükse, video klibi çerçeveleri iteratif bir şekilde yinelenerek model girdi çerçeve uzunluğuna ayarlanmaktadır. Oluşturulan modelin 12 versiyonu Çizelge 2.2'te gösterilmiştir.

**Çizelge 2.2 :** Hırsızlık tespiti için oluşturulan modelin 12 versiyonu.

| Versiyon | Girdi Çözünürlüğü (WxH) | Çerçeve Uzunluğu (Temporal Length) | Parti Büyüklüğü (Batch Size) |
|---|---|---|---|
| 1 | 224×224 | 32 | 12 |
| 2 | 224×224 | 32 | 10 |
| 3 | 224×224 | 32 | 8 |
| 4 | 224×224 | 16 | 12 |
| 5 | 224×224 | 16 | 10 |
| 6 | 224×224 | 16 | 8 |
| 7 | 112×112 | 32 | 12 |
| 8 | 112×112 | 32 | 10 |
| 9 | 112×112 | 32 | 8 |
| 10 | 112×112 | 16 | 12 |
| 11 | 112×112 | 16 | 10 |
| 12 | 112×112 | 16 | 8 |

Her veri örneğinin zamansal konumları, her adımda (epoch) giriş video klibinden eşit (uniformly) olarak seçilmektedir. Girdi görüntüleri [0-1] aralığında normalize edilerek baskınlık oluşturabilecek piksel değerleri baskılanmaktadır. Kinetik-700 ortalama çıkarma (mean subtraction) değerleri ile her örnek için ortalama çıkarma yapılmaktadır. Uzaysal kırpma (spatial crop) olarak uygulanan 10 eğitim görüntüsü örnekleri için veri büyütme ayrıntıları (data augmentation) dört köşe ve merkez ile uygulanır. Dört köşe ve merkez ölçeği $\{1, \frac{1}{2^{1/4}}, \frac{1}{\sqrt{2}}, \frac{1}{2^{3/4}}, \frac{1}{2}\}$ arasından seçilir. Ölçek 1, örnek genişliği ve yüksekliğinin çerçevenin kısa kenar uzunluğu ile aynı olduğu ve 0,5 ölçeği, örneğin kısa kenar uzunluğunun yarısı kadar olduğu anlamına gelmektedir. Örnek en boy oranı 1'dir ve örneğin seçilen konumlarında, ölçekte ve en boy oranında uzam-zamansal olarak kırpılmaktadır. Ardından kırpılan görüntünün genişliği ve yüksekliği mimarinin giriş çözünürlüğüne ayarlanmaktadır. Oluşturulan 12 mimari 200 adımda (epoch) eğitilmiştir. Veri büyütme ile oluşturulan örnek, %50 olasılıkla yatay olarak çevrilmektedir.

Optimize edici olarak SGD ve kayıp fonksiyonu olarak çapraz entropi kaybı (cross-entropy loss) kullanılmıştır. SGD, yüksek boyutlu optimizasyon problemlerinde hesaplama yükünü azaltır ve daha düşük bir yakınsama oranı ile daha hızlı yineleme sağlamaktadır. SGD, kayıp fonksiyonunu en aza indirmek için gradyanları ve öğrenme oranını kullanarak model parametrelerini güncellemektedir. Model parametreleri: Ağırlık azalması (weight decay), öğrenme hızı (learning rate) ve momentum sırasıyla 0,001, 0,1 ve 0,9 olarak ayarlanmıştır. Doğrulama (validation) kaybı doyuma ulaştığında öğrenme oranı 10'a bölünmektedir. Çapraz entropi kaybı, çıktısı 0 ile 1

arasında bir olasılık değeri olan bir sınıflandırma modelinin performansını ölçerek oluşan hatayı hesaplamaktadır. Çapraz entropi kaybının ifadesi Denklem 2.6'da gösterilmektedir.

$$\mathrm{CE} = -\sum_{i}^{C} t_i \, log\big(f(s_i)\big) \tag{2.6}$$

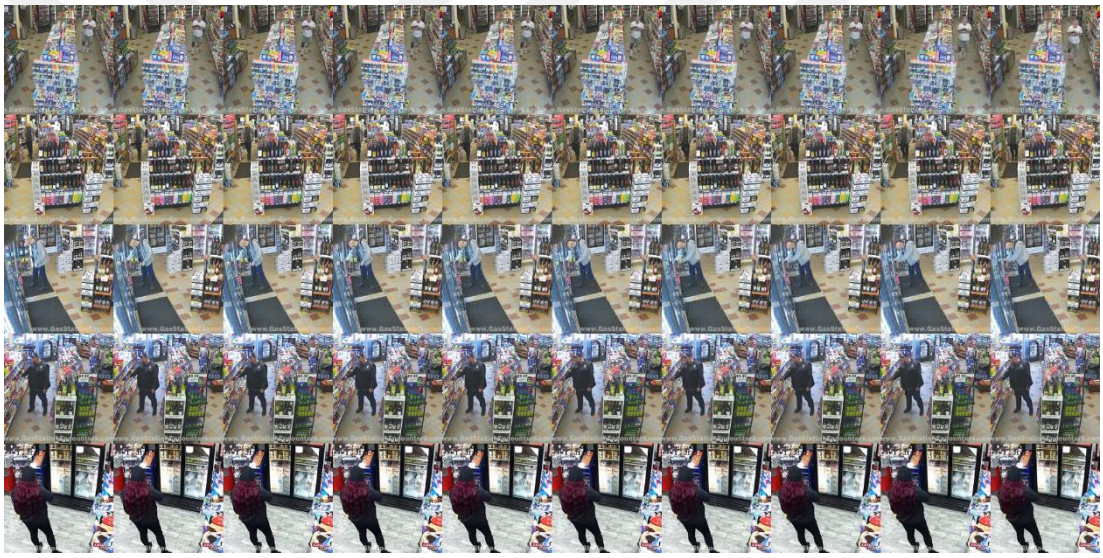
Denklem 2.6'da görüleceği üzere, CE çapraz entropi kaybı, *c* sınıf sayısı, $t_i$ ve $s_i$ temel gerçek (ground truth) ve *c*'deki her *i* sınıfı için mimarinin oluşturduğu skor, $f(s_i)$ softmax katmanı aktivasyon fonksiyonu olasılık sınıf skorunu ifade etmektedir.

Hırsızlık veya hırsızlık dışı eylemi sınıflandırmak için test veri kümesinde eğitilmiş model kullanılırken: Giriş klipleri oluşturmak için kayan pencere yöntemini kullanılır [45] ve her klip maksimum ölçekte bir merkez konumu etrafında kırpılır. Ardından, her videonun sınıf olasılıklarını tahmin etmekte ve test veri setinde genel doğruluk için bunların ortalaması alınmaktadır.

## 2.5 Süpermarket Hırsızlık ve Hırsızlık Olmayan Eylem Veri Seti

Bu tez çalışmasında, hırsızlık ve hırsızlık olmayan eylem olarak iki sınıflı bir veri seti oluşturulmuştur. Hırsızlık eylemi sınıfı: eşyaları; cebe koymak, çantaya koymak ve el çantasına koymak şeklinde üç farklı hırsızlık eylemini içeren bir sınıftır. Hırsızlık eylemlerine ait veri seti örnekleri Şekil 2.7'de gösterilmektedir. Hırsızlık olmayan eylem sınıfı: süpermarkette; yürümek, sabit durmak ve raftan eşya almak şeklinde üç farklı hırsızlık olmayan eylemi içeren bir sınıftır. Hırsızlık olmayan eylemlere ait veri seti örnekleri Şekil 2.8'de gösterilmektedir. Eğitim veri seti eylemleri YouTube'den toplanmıştır. Test veri seti eylemleri Bursa'daki bir süpermarketten toplanmıştır. Ancak, eğitim veri seti hırsızlık eylemlerinin aksine, süpermarketten toplanan test veri seti eylemlerinde gerçekleştirilen hırsızlık eylemi hırsızlık yapılıyormuş gibi davranılarak oluşturulmuştur. Eğitim ve test veri seti hırsızlık sınıfında, eşyaları cebe koymak eylemi en fazla örneğe sahip eylemdir. Eğitim ve test veri seti hırsızlık olmayan sınıfta ise üç eylem de neredeyse aynı örnek sayısına sahiptir. Eğitim ve test veri setlerinde, hemen hemen tüm kişiler hem hırsızlık hem de hırsızlık dışı eylemler için kullanılmaktadır. Toplanan videolar eylem sınıflarına uygun olarak kırpılmıştır.

**Şekil 2.7 :** Süpermarket hırsızlık eylem çerçeve örnekleri.

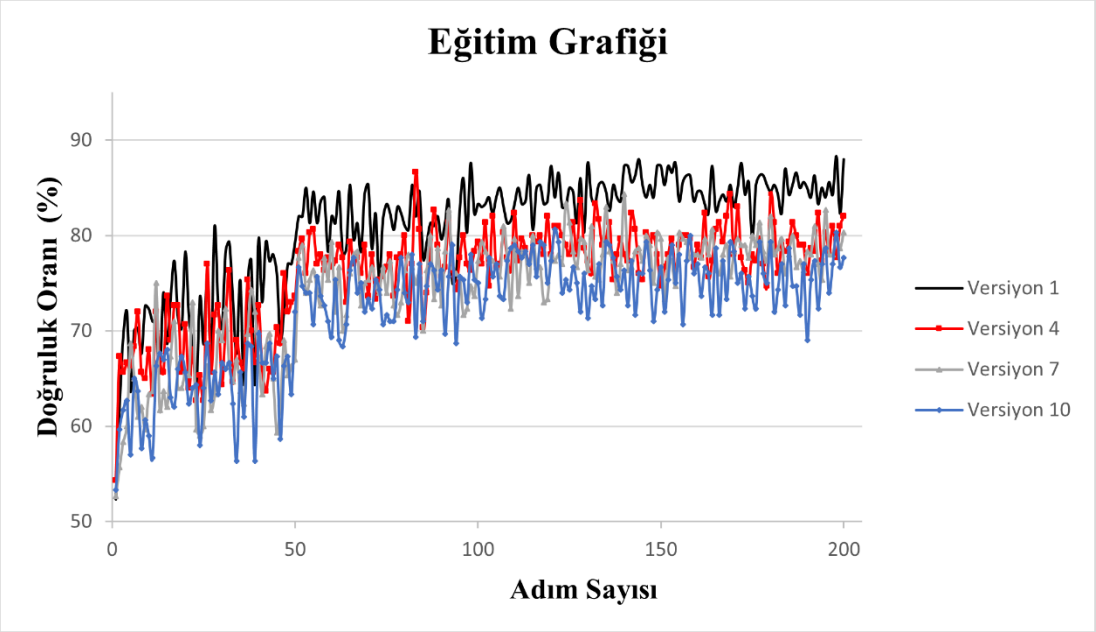


**Şekil 2.8 :** Süpermarket hırsızlık olmayan eylem çerçeve örnekleri.

Eğitim veri seti, 139 hırsızlık eylemi ve 161 hırsızlık dışı eylem videosu olmak üzere 300 videodan oluşmaktadır. Doğrulama veri seti, 7 hırsızlık eylemi ve 7 hırsızlık dışı eylem videosu olmak üzere 14 videodan oluşmaktadır. Test veri seti, 130 hırsızlık ve 140 hırsızlık olmayan eylem videosu olmak üzere 270 videodan oluşmaktadır. Test veri seti, eğitim veri seti kadar büyük olmasına rağmen eğitim veri setinde kişi sayısı çeşitliyken, test veri setinde tüm videolar için sadece bir kadın ve bir erkek olmak üzere 2 kişi ile sınırlıdır. Videoların ortalama süresi yaklaşık 1,5 saniye ve bu da 32 RGB görüntü çerçevesine eşittir. Her 32 çerçeveli girdi versiyonu, eğitim ve test için sırasıyla 10048 ve 8640 görüntüye sahiptir. Her 16 çerçeveli girdi versiyonu, eğitim ve test için sırasıyla 5024 ve 4320 görüntüye sahiptir.

# 3. BULGULAR VE TARTIŞMA

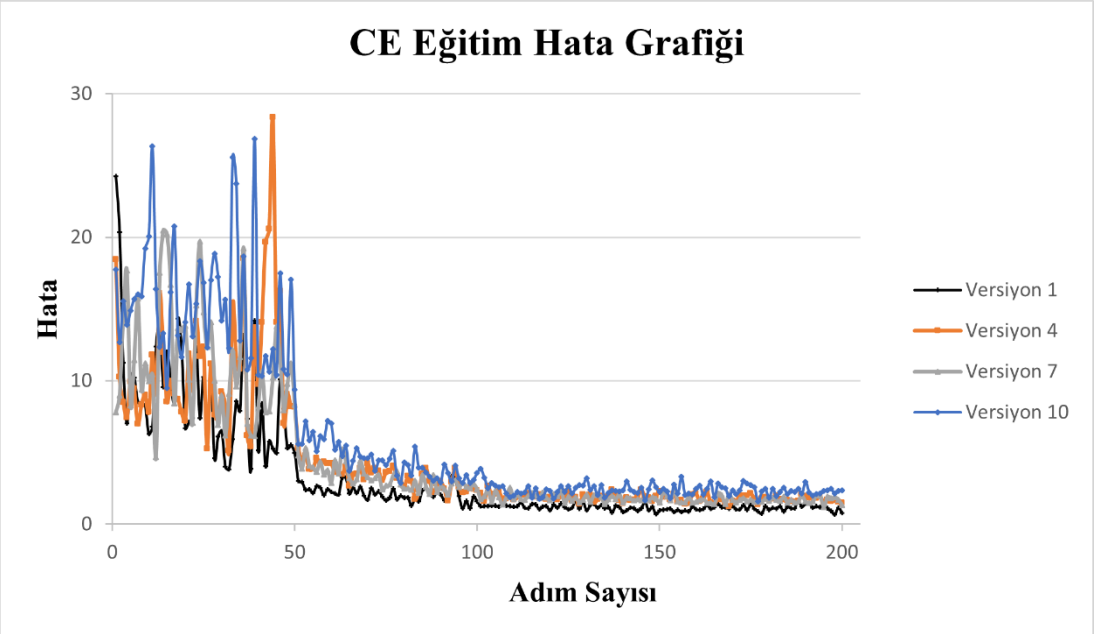
Eğitim ve test çalışmasında, Çizelge 2.2'te gösterilen versiyonların, farklı; girdi görüntülerinin, çerçeve uzunluklarının ve parti büyüklüklerinin etkilerinin karşılaştırılması amaçlanmıştır. Bu tez çalışmasında oluşturulan hırsızlık ve hırsızlık olmayan veri setinde, versiyonlar eğitilip test edilmiştir. Versiyonların performansını analiz etmek için tüm versiyonlar 200 adımda eğitilmiş ve son adımda oluşturulmuş olan 200. adım versiyonu tarafından test edilmiştir. Eğitim grafiği Şekil 3.1'de gösterilmektedir.



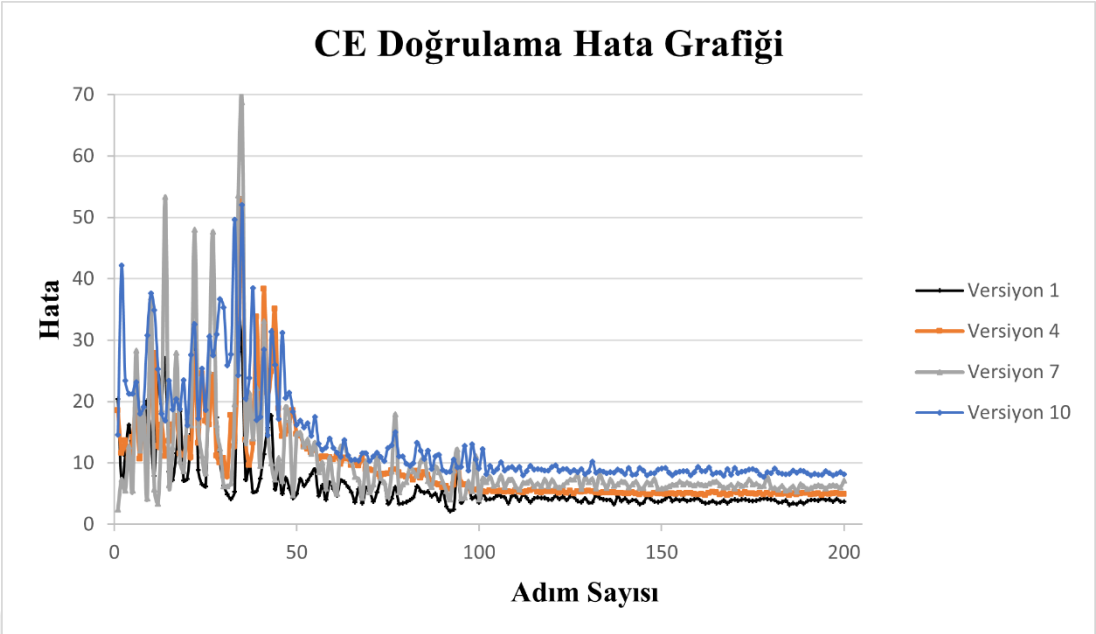

**Şekil 3.1 :** Versiyon 1, 4, 7 ve 10'un eğitim grafiği.

Versiyonlar için eğitim grafiği oluşturulurken parti büyüklüğü 12 olan versiyonlar baz alınmış ve versiyonlar, girdi görüntüsü, çerçeve uzunluğu olarak birbirinden farklıdır. Versiyon 1, 224×224 girdi görüntüsü ve 32 çerçeve uzunluğu ile %88,0 doğruluk oranına sahiptir. Versiyon 4, 224×224 girdi görüntüsü ve 16 çerçeve uzunluğu ile %82,0 doğruluk oranına sahiptir. Versiyon 7, 112×112 girdi görüntüsü ve 32 çerçeve uzunluğu ile %80,3 doğruluk oranına sahiptir. Versiyon 10, 112×112 girdi görüntüsü ve 16 çerçeve uzunluğu ile %77,6 doğruluk oranına sahiptir. Şekil 3.1'de görüldüğü üzere en yüksek doğruluk oranına Versiyon 1 sahiptir. Farklı girdi görüntüsü boyutuna

ve aynı çerçeve uzunluğuna sahip olan Versiyon 1 ve Versiyon 7, Versiyon 4 ve Versiyon 10 kendi aralarında karşılaştırıldığında, girdi görüntüsü boyutunun önemi görülmektedir. Aynı girdi görüntüsü boyutuna ve farklı çerçeve uzunluğuna sahip olan Versiyon 1 ve Versiyon 4, Versiyon 7 ve Versiyon 10 kendi aralarında karşılaştırıldığında çerçeve uzunluğunun önemi görülmektedir. Ayrıca, Versiyon 1, 4, 7 ve 10'nun Şekil 3.2'de CE eğitim hata eğrisi ve Şekil 3.3'te CE doğruluk (validation) hata eğrisi gösterilmektedir. Şekil 3.2 ve Şekil 3.3'te Versiyon 1 hem eğitim hem de doğrulama hata grafiği için en düşük hata değerlerine sahiptir.




**Şekil 3.2 :** Versiyon 1, 4, 7 ve 10'un CE ile eğitim hata eğrisi.




**Şekil 3.3 :** Versiyon 1, 4, 7 ve 10'un CE ile doğrulama hata eğrisi.

Tüm versiyonlara ait eğitim ve test doğrulukları Çizelge 3.1'de gösterilmektedir. Versiyon 1, sırasıyla %88,0 ve %77,0 ile en yüksek eğitim ve test doğruluğuna sahiptir. Versiyon 1, test veri setinde 131 hırsızlık eylemi videosundan 104 video eylemini doğru olarak etiketlerken 26 videoyu yanlış etiketlemiştir. Böylece, hırsızlık eylemi veri seti doğruluk sonucu %79,4'tür. Versiyon 1, test veri setinde 139 hırsızlık olmayan eylem videosundan 104 video eylemini doğru etiketlerken 36 videoyu yanlış etiketlemiştir. Hırsızlık olmayan eylem veri seti doğruluk sonucu %74,8'dir. Sonuç olarak Versiyon 1 270 test videosunu %77,0 doğruluk oranı ile sınıflandırmıştır. Test veri setinde Versiyon 1'in; duyarlılık (recall) oranı %80, kesinlik (precision) oranı %74,28 ve F1 skoru %77,03'tür.

**Çizelge 3.1 :** Modelin 12 versiyonuna ait eğitim ve test doğrulukları.

| Versiyon | Girdi Çözünürlüğü (WxH) | Çerçeve Uzunluğu (Temporal Length) | Parti Büyüklüğü (Batch Size) | Eğitim Doğruluğu (%) | Test Doğruluğu (%) |
|---|---|---|---|---|---|
| 1 | 224×224 | 32 | 12 | 88.0 | 77.0 |
| 2 | 224×224 | 32 | 10 | 84.3 | 72.5 |
| 3 | 224×224 | 32 | 8 | 85.0 | 68.8 |
| 4 | 224×224 | 16 | 12 | 82.0 | 55.1 |
| 5 | 224×224 | 16 | 10 | 79.6 | 55.1 |
| 6 | 224×224 | 16 | 8 | 82.3 | 56.6 |
| 7 | 112×112 | 32 | 12 | 80.3 | 55.1 |
| 8 | 112×112 | 32 | 10 | 78.0 | 57.0 |
| 9 | 112×112 | 32 | 8 | 77.6 | 55.5 |
| 10 | 112×112 | 16 | 12 | 77.6 | 44.8 |
| 11 | 112×112 | 16 | 10 | 77.0 | 53.3 |
| 12 | 112×112 | 16 | 8 | 75.0 | 53.7 |

Çizelge 3.1 incelendiğinde, aynı girdi görüntüsü 224×224 ve aynı parti boyutlarına sahip olan versiyonlar farklı çerçeve uzunluğunda, Versiyon 1 ve Versiyon 4, Versiyon 2 ve Versiyon 5, Versiyon 3 ve Versiyon 6 kendi aralarında karşılaştırıldığında Versiyon 4, Versiyon 5 ve Versiyon 6 için eşleştirilen versiyonlar ile eğitim doğruluk oranları yakın olsa da test doğruluk oranları sırasıyla %21,9, %17,4 ve %12,2 olarak büyük ölçüde düşmüştür. Aynı şekilde, aynı girdi görüntüsü 112×112 ve aynı parti boyutlarına sahip olan versiyonlar farklı çerçeve uzunluğunda, Versiyon 7 ve Versiyon 10, Versiyon 8 ve Versiyon 11, Versiyon 9 ve Versiyon 12 kendi aralarında karşılaştırıldığında eğitim doğruluk oranlarını birbirine daha da yakın ve test doğruluk oranları da Versiyon 7 ve Versiyon 10 çifti dışında birbirine yakındır. Versiyon 7 ve Versiyon 10 arasındaki test doğruluk düşüşü %10,3'tür. Sonuç olarak test doğruluk oranları ile hırsızlık eylemini sınıflandırmak için 32 çerçeve uzunluğunun 16 çerçeve uzunluğundan daha üstün olduğu gösterilmiştir.

Çizelge 3.1’de görüldüğü üzere, farklı görüntü girdi versiyonları için hem eğitim hem de test doğruluk oranları düşmektedir. Ancak, test doğruluk oranı eğitim doğruluk oranına göre daha fazla düşüş gösterdiği için girdi görüntüsü karşılaştırılmasında test doğruluk oranına göre karşılaştırma yapılmıştır. Versiyon 1 ve Versiyon 7, Versiyon 2 ve Versiyon 8, Versiyon 3 ve Versiyon 9, Versiyon 4 ve Versiyon 10, Versiyon 5 ve Versiyon 11, Versiyon 6 ve Versiyon 12 kendi aralarında karşılaştırıldığında test doğruluk oranı düşüşü sırasıyla %21,9, %15,5, %13,3, %10,3, %1,8 ve %2,9'dur. Böylece, hırsızlık eyleminin tanınması için çok önemli olan girdi görüntüsünün etkisi doğrulanmıştır. Buna ek olarak, parti boyutunun etkisi bilinmektedir ve iyi bir model eğitmek için büyük bir parti büyüklüğü önemlidir [44].

# 4. SONUÇ VE ÖNERİLER

Hırsızlık eylemi, hareket eylemi ile gerçekleşmektedir. Bu sebeple, hırsızlık eylemi uzam-zamansal bilgi içermektedir. Uzam-zamansal bilgiyi doğru bir şekilde çıkarmak ve sınıflandırmak için 2-B ve 3-B CNN olmak üzere insan eylemi tanıma modelleri kullanılmaktadır. Bu tez çalışmasında, 18 katmanlı 3-B artık ağ modeli [32] temel alınarak hırsızlık tespiti yapan başarılı bir model geliştirilmiştir.

Bu tez çalışmasında, süpermarketlerdeki hırsızlık ve hırsızlık olmayan eylemleri içeren eğitim video veri seti Youtube'den ve test veri seti bir süpermarketten toplanarak oluşturulmuştur. Hırsızlık veri seti sınıfı: eşyaları; cebe koymak, çantaya koymak ve el çantasına koymak şeklinde 3 eylemden oluşmaktadır. Hırsızlık olmayan veri seti sınıfı: süpermarkette; yürümek, raftan eşya almak ve sabit durmak şeklinde 3 eylemden oluşmaktadır. Eğitim veri seti, 161 hırsızlık olmayan eylem ve 139 hırsızlık eylemi videosu ile toplamda 300 videodan oluşmaktadır. Test veri seti, 140 hırsızlık olmayan eylem ve 130 hırsızlık eylemi videosu ile toplamda 270 videodan oluşmaktadır. Eğitim veri setinde kişi sayısı çeşitliyken, test veri seti için sadece bir kadın ve bir erkek olmak üzere 2 kişi ile sınırlıdır.

Hırsızlık tespitini gerçekleştirmek için tasarlanan CNN ağı, 12 versiyon şeklinde farklı parametreli olarak oluşturulmuş ve 200 adımda eğitilmiştir. Oluşturulan versiyonlar, girdi görüntüsü, çerçeve uzunluğu ve parti büyüklüğü olarak birbirinden farklıdır. Versiyonlar RGB girdi almaktadır. Eğitim ve test sonuçları Çizelge 3.1'de gösterilmekte ve Versiyon 1, 224×224 girdi görüntüsü, 32 çerçeve uzunluğu ve 12 parti büyüklüğü ile eğitim veri seti %88,0, test veri seti %77,0 ile en iyi sonuca sahiptir. Test sonuçları ile 224×224 girdi görüntüsünün 112×112 girdi görüntüsüne göre daha üstün olduğu net bir şekilde görülmektedir. Eğitim ve test videoları ortalama olarak 32 girdi çerçevesine sahip ve 32 çerçeve uzunluğuna sahip versiyonların 16 çerçeve uzunluğuna sahip olan versiyonlara göre doğruluk oranının daha yüksek olduğu gösterilmiştir. Parti büyüklüğünün etkisi BN [44] çalışması ile zaten bilinmekte ve parti büyüklüğü arttıkça doğruluk oranı da genel olarak artmaktadır. Sonuç olarak, süpermarket hırsızlık eylemi tespiti için başarılı bir model oluşturulmuştur. Ayrıca,

yapılan deneylerde, çerçeve uzunluğu ve parti büyüklüğü parametreleri donanım yetersizliği sebebiyle daha büyük değerlerde test edilememiştir.

İlerde yapılacak çalışmalar için çerçeve uzunluğu, parti büyüklüğü ve veri seti örnekleri arttırılabilir.

## KAYNAKLAR


**[1] Lasky, N. V., Fisher, B. S., & Jacques, S.** (2017). 'Thinking thief'in the crime prevention arms race: lessons learned from shoplifters. Security Journal, 30(3), 772-792.

**[2] Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., & Vijayanarasimhan, S.** (2016). Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675.

**[3] Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L.** (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 1725-1732).

**[4] Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T.** (2011, November). HMDB: a large video database for human motion recognition. In *2011 International conference on computer vision* (pp. 2556-2563). IEEE.

**[5] Soomro, K., Zamir, A. R., & Shah, M.** (2012). UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*.

**[6] Zhang, Y., Jin, S., Wu, Y., Zhao, T., Yan, Y., Li, Z., & Li, Y.** (2020). A NEW INTELLIGENT SUPERMARKET SECURITY SYSTEM. *Neural Network World*, *30*(2), 113-131.

**[7] Gim, U. J., Lee, J. J., Kim, J. H., Park, Y. H., & Nasridinov, A.** (2020, April). An Automatic Shoplifting Detection from Surveillance Videos (Student Abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 34, No. 10, pp. 13795-13796).

**[8] Kim, S., Hwang, S., & Hong, S. H.** (2021). Identifying shoplifting behaviors and inferring behavior intention based on human action detection and sequence analysis. *Advanced Engineering Informatics*, *50*, 101399.

**[9] Yin, Z., & Collins, R.** (2007, June). Belief propagation in a 3D spatio-temporal MRF for moving object detection. In *2007 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.

**[10] Laptev, I.** (2005). On space-time interest points. *International journal of computer vision*, *64*(2), 107-123.

**[11] Dollár, P., Rabaud, V., Cottrell, G., & Belongie, S.** (2005, October). Behavior recognition via sparse spatio-temporal features. In *2005 IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance* (pp. 65-72). IEEE.

**[12] Scovanner, P., Ali, S., & Shah, M.** (2007, September). A 3-dimensional sift descriptor and its application to action recognition. In *Proceedings of the 15th ACM international conference on Multimedia* (pp. 357-360).

**[13] Willems, G., Tuytelaars, T., & Van Gool, L.** (2008, October). An efficient dense and scale-invariant spatio-temporal interest point detector. In *European conference on computer vision* (pp. 650-663). Springer, Berlin, Heidelberg.

**[14] Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B.** (2008, June). Learning realistic human actions from movies. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1-8). IEEE.

**[15] Klaser, A., Marszałek, M., & Schmid, C.** (2008, September). A spatio-temporal descriptor based on 3d-gradients. In *BMVC 2008-19th British Machine Vision Conference* (pp. 275-1). British Machine Vision Association.

**[16] Wang, H., Kläser, A., Schmid, C., & Liu, C. L.** (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, *103*(1), 60-79.

**[17] Wang, H., & Schmid, C.** (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision* (pp. 3551-3558).

**[18] Csurka, G., Dance, C., Fan, L., Willamowski, J., & Bray, C.** (2004, May). Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV* (Vol. 1, No. 1-22, pp. 1-2).

**[19] Sánchez, J., Perronnin, F., Mensink, T., & Verbeek, J.** (2013). Image classification with the fisher vector: Theory and practice. *International journal of computer vision*, *105*(3), 222-245.

**[20] Jégou, H., Douze, M., Schmid, C., & Pérez, P.** (2010, June). Aggregating local descriptors into a compact image representation. In *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3304-3311). IEEE.

**[21] Bay, H., Tuytelaars, T., & Van Gool, L.** (2006, May). Surf: Speeded up robust features. In *European conference on computer vision* (pp. 404-417). Springer, Berlin, Heidelberg.

**[22] LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P.** (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-24.

**[23] LeCun, Y., Kavukcuoglu, K., & Farabet, C.** (2010, May). Convolutional networks and applications in vision. In *Proceedings of 2010 IEEE international symposium on circuits and systems* (pp. 253-256). IEEE.

**[24] Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A.** (2011, November). Sequential deep learning for human action recognition. In *International workshop on human behavior understanding* (pp. 29-39). Springer, Berlin, Heidelberg.

**[25] Simonyan, K., & Zisserman, A.** (2014). Two-stream convolutional networks for action recognition in videos. *arXiv preprint arXiv:1406.2199*.

**[26] Feichtenhofer, C., Pinz, A., & Zisserman, A.** (2016). Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1933-1941).

**[27] Wang, L., Xiong, Y., Wang, Z., & Qiao, Y.** (2015). Towards good practices for very deep two-stream convnets. *arXiv preprint arXiv:1507.02159*.

**[28] Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., & Van Gool, L.** (2016, October). Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision* (pp. 20-36). Springer, Cham.

**[29] Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M.** (2015). Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision* (pp. 4489-4497).

**[30] Carreira, J., & Zisserman, A.** (2017). Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6299-6308).

**[31] Qiu, Z., Yao, T., & Mei, T.** (2017). Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision* (pp. 5533-5541).

**[32] Hara, K., Kataoka, H., & Satoh, Y.** (2018). Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet?. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 6546-6555).

**[33] Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M.** (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* (pp. 6450-6459).

**[34] Xie, S., Sun, C., Huang, J., Tu, Z., & Murphy, K.** (2018). Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)* (pp. 305-321).

**[35] Tran, D., Wang, H., Torresani, L., & Feiszli, M.** (2019). Video classification with channel-separated convolutional networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5552-5561).

**[36] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... & Fei-Fei, L.** (2015). Imagenet large scale visual recognition challenge. *International journal of computer vision*, *115*(3), 211-252.

**[37] Ioffe, S., & Szegedy, C.** (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

**[38] Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., ... & Zisserman, A.** (2017). The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

**[39] He, K., Zhang, X., Ren, S., & Sun, J.** (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

**[40] Carreira, J., Noland, E., Banki-Horvath, A., Hillier, C., & Zisserman, A.** (2018). A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*.

**[41] Carreira, J., Noland, E., Hillier, C., & Zisserman, A.** (2019). A short note on the kinetics-700 human action dataset. *arXiv preprint arXiv:1907.06987*.

**[42] 3-B görüntü hacmine uygulanan 3-B çekirdek filtre** https://towardsdatascience.com/a-comprehensive-introduction-to-different-types-of-convolutions-in-deep-learning-669281e58215, **e**rişim tarihi 27.01.2022.

**[43] Caba Heilbron, F., Escorcia, V., Ghanem, B., & Carlos Niebles, J.** (2015). Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 961-970).

**[44] Ioffe, S., & Szegedy, C.** (2015, June). Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning* (pp. 448-456). PMLR.

**[45] Kataoka, H., Wakamiya, T., Hara, K., & Satoh, Y.** (2020). Would mega-scale datasets further enhance spatiotemporal 3D CNNs?. *arXiv preprint arXiv:2004.04968*.

## ÖZGEÇMİŞ

**Ad-Soyad** **: Emirhan TOPRAK**

**Doğum Tarihi ve Yeri** **:**

**E-posta** **:**

**ÖĞRENİM DURUMU:**

- **Lisans** **:** 2019, Kütahya Dumlupınar Üniversitesi, Simav Teknoloji Fakültesi, Elektrik-Elektronik Mühendisliği
- **Yüksek Lisans** **:** 2022, Bursa Teknik Üniversitesi, Lisansüstü Eğitim Enstitüsü, Elektrik-Elektronik Mühendisliği

**TEZDEN TÜRETİLEN ESERLER, SUNUMLAR VE PATENTLER:**

- **Emirhan TOPRAK, Mustafa ÖZDEN, Theft Detection in Video Surveillance Videos with 3D CNN Model, International Conference on Design, Research and Development, December 15-18, Kayseri, Türkiye, 2021**