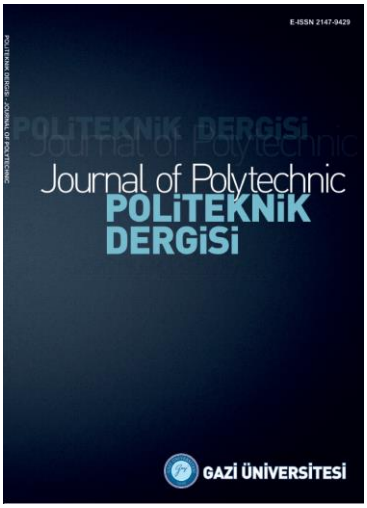


POLİTEKNİK DERGİSİ

JOURNAL of POLYTECHNIC

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik

Derin öğrenme tabanlı video üzerinde olay sınıflandırma

Deep learning based video event classification

***Yazarlar (Authors)**: Serim GENÇASLAN[1], Anıl UTKU[2], M. Ali AKCAYOL[3]*

ORCID[1]: 0000-0001-8404-3099

ORCID[2]: 0000-0002-7240-8713

ORCID[3]: 0000-0002-6615-1237

To cite to this article: Gençaslan S., Utku A. ve Akcayol M.A., "Derin Öğrenme Tabanlı Video Üzerinde Olay Sınıflandırma", ***Journal of Polytechnic***, 26(3): 1155-1165, (2023).

Bu makaleye şu şekilde atıfta bulunabilirsiniz: Gençaslan S., Utku A. ve Akcayol M.A., "Derin Öğrenme Tabanlı Video Üzerinde Olay Sınıflandırma", ***Politeknik Dergisi***, 26(3): 1155-1165, (2023).

Erişim linki (To link to this article): http://dergipark.org.tr/politeknik/archive

DOI: **10.2339/politeknik.775185**

Derin Öğrenme Tabanlı Video Üzerinde Olay Sınıflandırma

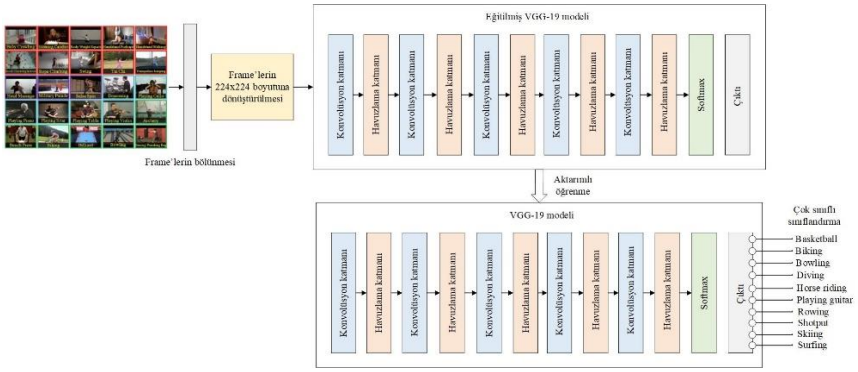
Deep Learning Based Video Event Classification

Önemli noktalar (Highlights)

- *Video üzerinde olay sınıflandırma/ Video event classification*
- *CNN ve VGG-19 modellerinin karşılaştırmalı analizi/ Comparative analysis of CNN and VGG-19 models*
- *Spor dallarının sınıflandırılması/ Classification of sports branches*

Grafik Özet (Graphical Abstract)

CNN kullanılarak geliştirilen modelde, VGG-19 ile öğrenme aktarımı uygulanmıştır./ In the model developed using CNN, learning transfer has been applied with VGG-19.



Şekil. Geliştirilen VGG-19 modeli /**Figure.** Developed VGG-19 model

Amaç (Aim)

VGG-19 modeli ile öğrenme aktarımı uygulanarak video üzerinden olay sınıflandırma problemine yönelik daha yüksek başarı oranı elde edilmesi./ A higher success rate for the event classification problem over video by applying the learning transfer with the VGG-19 model.

Tasarım ve Yöntem (Design & Methodology)

VGG-19 modeli ile birlikte öğrenme aktarımı uygulanmıştır. / Transfer learning has been implemented with the VGG-19 model.

Özgünlük (Originality)

Öğrenme aktarımı uygulanarak geleneksel derin öğrenme modellerinden daha yüksek başarı oranı elde edilmiştir./ By applying transfer learning, a higher success rate has been achieved than traditional deep learning models.

Bulgular (Findings)

VGG-19 modeli kullanılarak %83 doğruluk oranı elde edilmiştir./ By using the VGG-19 model, 83% accuracy rate has been obtained.

Sonuç (Conclusion)

Geleneksel derin öğrenme yöntemlerine kıyasla öğrenme aktarımının önemli ölçüde doğruluk artışı sağladığı gözlemlenmiştir. / Compared to traditional deep learning methods, it has been observed that the learning transfer provides a significant increase in accuracy.

Etik Standartların Beyanı (*Declaration of Ethical Standards*)

Bu makalenin yazar(lar)ı çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler. / The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.

Derin Öğrenme Tabanlı Video Üzerinde Olay Sınıflandırma

Araştırma Makalesi / Research Article

Serim GENÇASLAN[1], Anıl UTKU[2*], M. Ali AKCAYOL[3]

[1,3] Mühendislik Fakültesi, Bilgisayar Müh. Bölümü, Gazi Üniversitesi, Türkiye
[2]Mühendislik Fakültesi, Bilgisayar Müh. Bölümü, Munzur Üniversitesi, Türkiye

(Geliş/Received : 28.07.2020 ; Kabul/Accepted : 18.07.2022 ; Erken Görünüm/Early View : 24.08.2022)

ÖZ

Son yıllarda, dijital kütüphanelerin ve video veritabanlarının büyümesi nedeniyle, videolardan aktivitelerin otomatik olarak tespit edilmesi ve büyük veri kümelerinden örüntülerin elde edilmesi ön plana çıkmaktadır. Görüntüden nesne algılama, çeşitli uygulamalar için bir araç olarak kullanılır ve video sınıflandırmanın temelidir. Videolardaki bilgilerin zaman sürekliliği kısıtlaması olduğundan, videolardaki nesneleri tanımlamak tek görüntüye göre daha zordur. Bilgisayarlı görme alanındaki gelişmelerin ardından, makine öğrenmesi ve derin öğrenme için açık kaynaklı yazılım paketlerinin kullanımı ve donanım teknolojilerinde yaşanan gelişmeler, yeni yaklaşımların geliştirilmesine imkân sağlamıştır. Bu çalışmada, video üzerinde spor dallarının sınıflandırılmasına yönelik derin öğrenme tabanlı bir sınıflandırma modeli geliştirilmiştir. CNN kullanılarak geliştirilen modelde, VGG-19 ile öğrenme aktarımı uygulanmıştır. 32827 adet frame üzerinde, CNN ve VGG-19 modelleri kullanılarak yapılan deneysel çalışmalar, VGG-19'un %83 doğruluk oranı ile CNN'den daha başarılı bir sınıflandırma performansına sahip olduğunu göstermiştir.

Anahtar Kelimeler: Video sınıflandırma, derin öğrenme, CNN, VGG-19.

Deep Learning Based Video Event Classification

ABSTRACT

In recent years, due to the growth of digital libraries and video databases, automatic detection of activities from videos and obtaining patterns from large datasets have come to the fore. Object detection from images is used as a tool for various applications and is the basis of video classification. Objects in videos are more difficult to identify than in single images, as the information in videos has a time-continuity constraint. Following the developments in the field of computer vision, the use of open source software packages for machine learning and deep learning and the developments in hardware technologies have enabled the development of new approaches. In this study, a deep learning-based classification model has been developed for the classification of sports branches in video. In the model developed using CNN, transfer learning has been applied with VGG-19. Experimental studies on 32827 frames using CNN and VGG-19 models showed that VGG-19 has a more successful classification performance than CNN with an accuracy rate of 83%.

Keywords: Video classification, deep learning, CNN, VGG-19.

1. GİRİŞ (INTRODUCTION)

1981 yılından itibaren aktif olarak kullanılmaya başlayan İnternet teknolojileri, her geçen gün hızla büyümeye devam etmektedir. İnternet kullanımındaki bu artış, anlık olarak üretilen dijital içeriklerin sayısında da artış yaşanmasına neden olmuştur. International Data Corporation'ın raporuna göre, yapılandırılmamış veriler 2020 yılı itibariyle 40 Zettabayttan fazla ve yapısal verilere göre 50 kat daha büyük olacaktır [1].

Videolardan gerçek zamanlı olay tespiti, sınıflandırma gibi konular her geçen gün dikkat çekmekte ve gelişmeye devam etmektedir. Buna ek olarak günümüzde video paylaşım sitelerinin çoğalmasıyla milyarları aşan sayıda video bulunmaktadır. Bu gibi platformların haricinde marketlerde, bankalarda ve spor merkezi gibi alanlarda oldukça fazla sayıda kamera bulunmaktadır. Bu videolardan veya kameralardan alınan hareketli görüntüler, suçlu yakalama, araç hareketleri, olağan dışı insan hareketleri, trafik ışığı ihlalleri ve insan/hayvan/araç gibi sınıfların ayırt edilmesi gibi konular açısından önemlidir.

Video sınıflandırma, multimedya ve bilgisayarli görü konularında hayati bir araştırma konusudur. Başarılı sınıflandırma sistemleri, çıkarılan video özelliklerine büyük ölçüde güvenmektedir. Video sınıflandırma çalışmaları büyük ölçüde videolardaki olayları algılama, sınıflandırma ve nesneleri algılama ile ilgilidir [2-8]. Ma tarafından, 2016 yılında yapılan çalışmada zamansal derin öğrenmenin eğitilmesinin geliştirilmesi ve videodan olay tespitine yönelik Uzun Kısa Vadeli Hafıza (Long Short Term Memory-LSTM) tabanlı bir sistem geliştirilmiştir [2]. Çalışmada LSTM'e ek olarak LSTM-s ve LSTM-m modelleri kullanılmıştır. Canotiho tarafından 2005 yılında yapılan çalışmada, yürüme, koşma, kavga etme gibi insan davranışlarının tespitine yönelik bir sistem geliştirilmiştir [3,4]. Kim tarafından

*Sorumlu Yazar (Corresponding Author)
e-posta : anilutku@munzur.edu.tr

2009 yılında yapılan çalışmada, temel eylemleri tanımak için Gizli Markov Modeli ve Koşullu Rastgele Alan modeli kullanılmıştır [5].

Anguita [6] tarafından 2013 yılında yapılan çalışmada, akıllı telefon sensörlerinden alınan verilerle Destek Vektör Makinelerini (Support Vector Machines-SVM) kullanarak yürüme, merdiven çıkma, merdiven inme, oturma, ayakta durma gibi 30 farklı temel insan hareketini tanıyan bir sistem geliştirilmiştir. Literatürdeki çalışmalarda Kategori Özellik Vektörleri de (Category Feature Vectors-CFV) yaygın olarak kullanılmıştır. Örneğin Lin tarafından 2020 yılında yapılan çalışmada, bu yöntem geliştirilmiştir. Her bir etkinlik Gauss Karışım Modelleri'nin (Gaussian Mixture Models-GMM) bir kombinasyonu olarak belirlenmiştir [7]. Sonuç olarak, model olağandışı eylemler için daha esnek bir hale getirilmiştir. Dai [8] tarafından 2017 yılında yapılan çalışmada, yapısal olarak Faster R-CNN'e benzeyen Geçici Bağlam Ağı (Temporal Context Network-TCN) yapısı insanların yer tespiti için kullanılmış ve bu tespitlerin ardından ActivityNet ve THUMOS14 veri setleri üzerinde LSTM kullanılarak sınıflandırma işlemi yapılmıştır.

Bu çalışmada, belirlenen 10 farklı sınıf için video sınıflandırma yapılmış ve öncelikle Evrişimli Sinir Ağları (Convolutional Neural Network-CNN) modelleri denenmiştir. Bu modellerden yeterli düzeyde başarı oranı sağlanamadığından VGG-19 ile öğrenme aktarımı uygulanarak modelin başarısının artırılması sağlanmıştır.

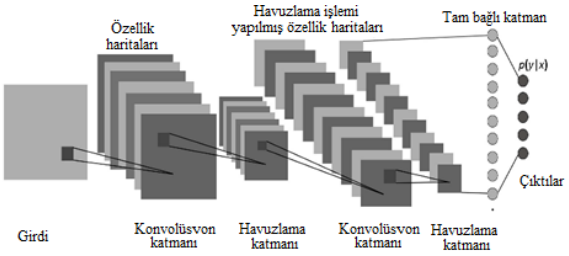
2. VİDEO SINIFLANDIRMA YÖNTEMLERİ (VIDEO CLASSIFICATION METHODS)

CNN, bilgisayarlı görü alanında yoğun olarak kullanılmakta olan bir derin öğrenme modelidir. Derin öğrenmede, evrişimli bir sinir ağı çoğunlukla görüntüleri analiz etmek için uygulanan bir derin sinir ağ sınıfıdır [21]. Görüntü ve video tanıma, öneri sistemleri [22], görüntü sınıflandırma, tıbbi görüntü analizi, doğal dil işleme [23] ve finansal zaman serilerinde [24] kullanılmaktadır.

CNN çok katmanlı algılayıcıların düzenli bir versiyonudur. Çok katmanlı algılayıcılar genellikle tamamen bağlı ağlar anlamına gelmektedir, yani her bir katmandaki bir nöron, bir sonraki katmandaki tüm nöronlara bağlanmaktadır. Bu ağların tamamen bağlılığı, verileri aşırı sığmaya eğilimli hale getirmektedir. CNN normalleşmeye farklı bir yaklaşım getirmektedir. Verilerdeki hiyerarşik kalıptan yararlanmakta ve daha küçük ve daha basit kalıplar kullanarak daha karmaşık kalıplar oluşturmaktadır. Bu nedenle, bağlılık ve karmaşıklık ölçeğinde CNN alt uçtadır.

CNN, nöronlar arasındaki bağlantı paterninin hayvan görsel korteksinin organizasyonuna benzemesi nedeniyle biyolojik süreçlerden ilham almıştır [25-27]. Bireysel kortikal nöronlar uyaranlara sadece alıcı alan olarak bilinen görme alanının sınırlı bir bölgesinde tepki vermektedir. Farklı nöronların alıcı alanları, tüm görme alanını kaplayacak şekilde kısmen çakışmaktadır.

CNN, adını oluşturduğu gizli katmanların türünden almaktadır. Genel olarak bu gizli katmanlar evrişimli katmanlar, havuzlama katmanları, tam bağlı katmanlar, doğrusal olmayan katmanlar, düzleştirme katmanları ve normalizasyon katmanlarından oluşmaktadır. CNN'ler diğer görüntü sınıflandırma algoritmalarına kıyasla nispeten daha az ön işleme kullanmaktadır. Bu, ağın geleneksel algoritmalarda el ile tasarlanan filtreleri öğrendiği anlamına gelmektedir. Şekil 1'de, temel CNN yapısı görülmektedir.



Şekil 1. Temel CNN yapısı (Basic CNN architecture)

Matematiksel olarak *f* ve *g* şeklindeki iki fonksiyonunun evrişimi Eşitlik 1'de görüldüğü gibi tanımlanmaktadır:

$$(f * g)(i) = \sum_{j=1}^{m} g(j) f(i - j + \frac{m}{2}) \quad (1)$$

3. DERİN ÖĞRENME TABANLI VİDEO ÜZERİNDE OLAY SINIFLANDIRMA MODELİ (DEEP LEARNING BASED EVENT CLASSIFICATION MODEL ON VIDEO)

Mevcut olay tanıma veri kümelerinin çoğu iki temel dezavantaja sahiptir. Birincisi, sınıfların sayısının gerçekte insanlar tarafından gerçekleştirilen olayların zenginliğine kıyasla tipik olarak çok düşük olmasıdır. Örneğin, KTH [15], Weizmann [16], UCF Sports [17], IXMAS [18] veri setleri sırasıyla 6, 9, 9, 11 sınıf içermektedir. İkincisi ise videoların gerçekçi olmayan şekilde kontrol edilen ortamlara kaydedilmesidir. Örneğin KTH, Weizmann, IXMAS aktörler tarafından sahnelenmekte; HOHA [19] ve UCF Sports, profesyonel çekim ekibi tarafından çekilen film kliplerinden oluşmaktadır. Özellikle Kinetics veri setinin [9] 2017'de piyasaya sürülmesinin ardından çeşitli veri setlerinin performanslarında oldukça önemli iyileşmeler yapılmıştır. Bu veri setlerinden bazıları: UCF-101 [10], HMDB-51 [11], Charades [12], AVA [13], Thumos [14]. Bu çalışma kapsamında UCF-101 veri seti kullanılmıştır.

3.1. Veri Seti (Dataset)

UCF-101 veri seti, beş farklı türde 101 kategoriye ait 13.320 videodan oluşmaktadır. Bu türler insan-nesne etkileşimi, sadece vücut hareketleri, insan- insan etkileşimi, müzik enstrümanları çalma ve spordur. Şekil 9'da UCF-101 veri setinde bulunan her olay sınıfı için kaç tane video klip olduğu gösterilmiş ve klip sürelerinin dağılımları da renklerle ifade edilmiştir.

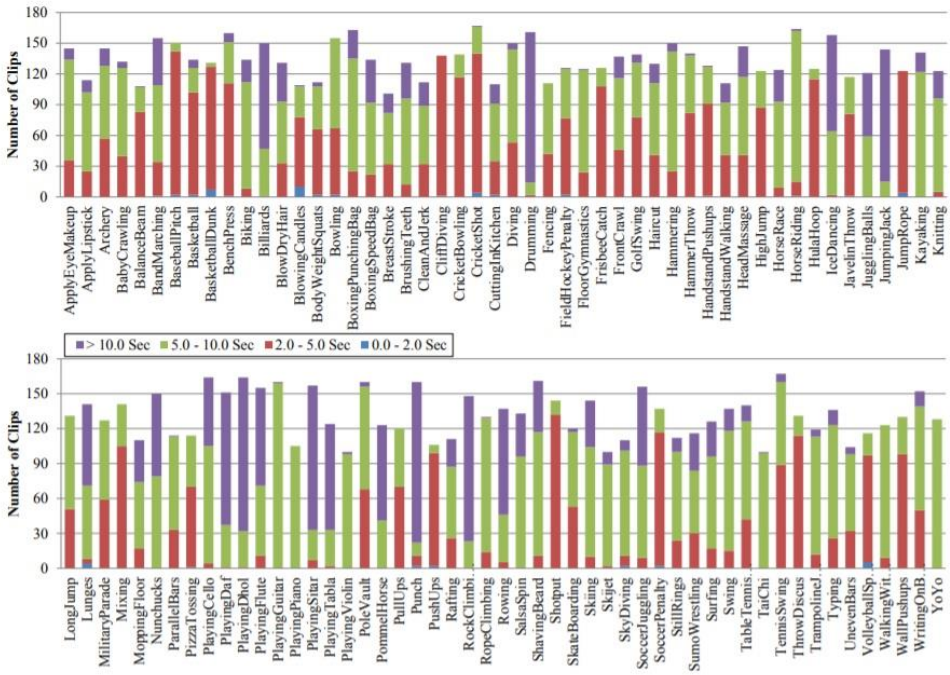
Bu çalışmada UCF-101 veri seti içinden belirlenen 10 farklı spor dalı kullanılmıştır. Seçilen sınıflar Şekil 2'de görülmektedir.

Şekil 2. Seçilen sınıflar (Selected classes)

Bu çalışma için seçilen 10 farklı spor dalına ait 0-2 saniye arası, 2-5 saniye arası, 5-10 saniye arası ve 10 saniyeden büyük olan video kliplerin sayısı Çizelge 1'de gösterilmektedir. Şekil 3'te her bir olay sınıfı için sürelere göre video klip sayıları görülmektedir.

Çizelge 1. Video kliplerin uzunluklarına göre dağılımları (Distribution of video clips by length)

Sınıflar	0-2 saniye	2-5 saniye	5-10 saniye	>10 saniye
Basketball	2	100	24	8
Biking	0	8	104	22
Bowling	0	61	94	0
Diving	0	53	91	6
HorseRiding	1	14	147	2
PlayingGuitar	0	0	4	156
Rowing	0	5	39	93
Shotput	0	128	16	0
Skiing	0	10	88	37
Surfing	0	13	81	32



Şekil 3. Her olay sınıfı için video klip sayısı (Number of video clips for each action class)

Olay sınıflarının klipleri, her biri 4-7 video klip içeren 25 gruba ayrılmaktadır. Bir gruptaki klipler arka plan veya aktörler gibi bazı ortak özellikleri paylaşmaktadır.

Veri setindeki videolar YouTube'dan indirilmiş videolar olup, alakasız olan videolar manuel olarak kaldırılmıştır. Tüm kliplerin saniyedeki kare sayısı ve çözünürlüğü sabit olup sırasıyla 25 FPS ve 320x240'tır. Videolar, k-lite paketinde bulunan DivX codec bileşeni kullanılarak sıkıştırılmış .avi dosyalarına kaydedilmektedir [20]. Çizelge 2'de UCF-101 veri setinin karakteristik özellikleri görülmektedir.

Çizelge 2. UCF-101 veri setinin karakteristik özellikleri (Characteristics of UCF-101 dataset)

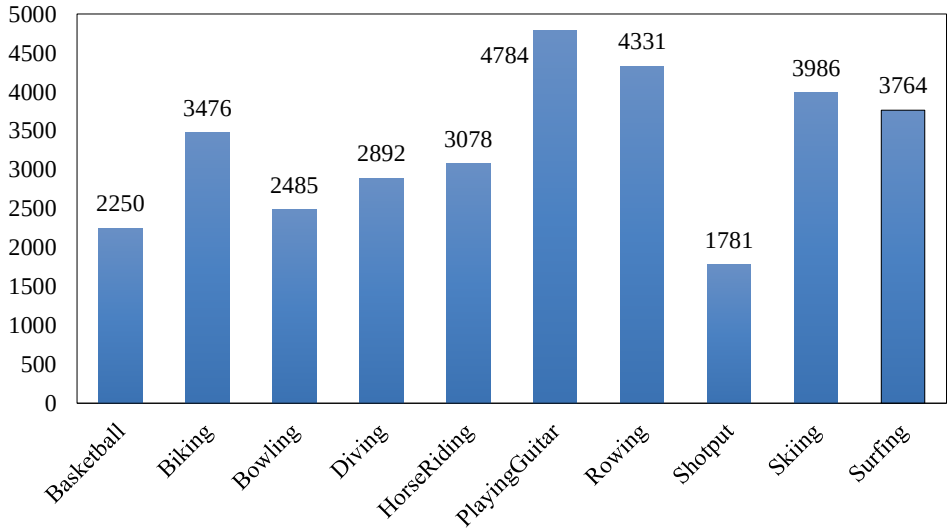
Olay Sınıfları	101
Video sayısı	13320
Her olay için grup sayısı	25
Her grup için video sayısı	4-7
Ortalama video uzunluğu	7.21 saniye
Toplam uzunluk	1600 dakika
En az video uzunluğu	1.06 saniye
En fazla video uzunluğu	71.04 saniye
Saniyedeki kare sayısı	25 FPS
Çözünürlük	320x240
Ses	Var (Yeni 51 sınıf için)

3.2. Video Önişleme (Video Pre-processing)

Bu çalışmada kullanılan UCF-101 veri setinden 10 farklı spor dalı (Basketball, Biking, Bowling, Diving, HorseRiding, PlayingGuitar, Rowing, Shotput, Skiing, Surfing) seçilmiş ve her olay sınıfından 120 video alınarak, toplam 10 sınıfa ait 1200 video elde edilmiştir.

Elde edilen bu 1200 video 4 FPS değeri ile karelere bölünerek 10 farklı sınıfa ait toplam 32827 adet frame elde edilmiş ve bu framelerin boyutları 224x224 olacak şekilde ayarlanmıştır. Toplam elde edilen 32827 adet frame, model oluşturulurken %70'i eğitim verisi (22983 frame) ve %30'u test verisi (9844 frame) olacak şekilde bölünmüştür.

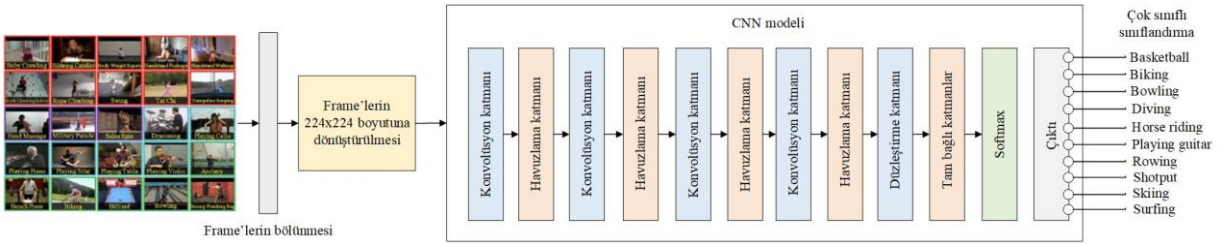
Elde edilen 32827 adet görüntünün sınıflara göre dağılımı Şekil 4'te görülmektedir.



Şekil 4. Her sınıf için frame sayıları (Frame counts for each class)

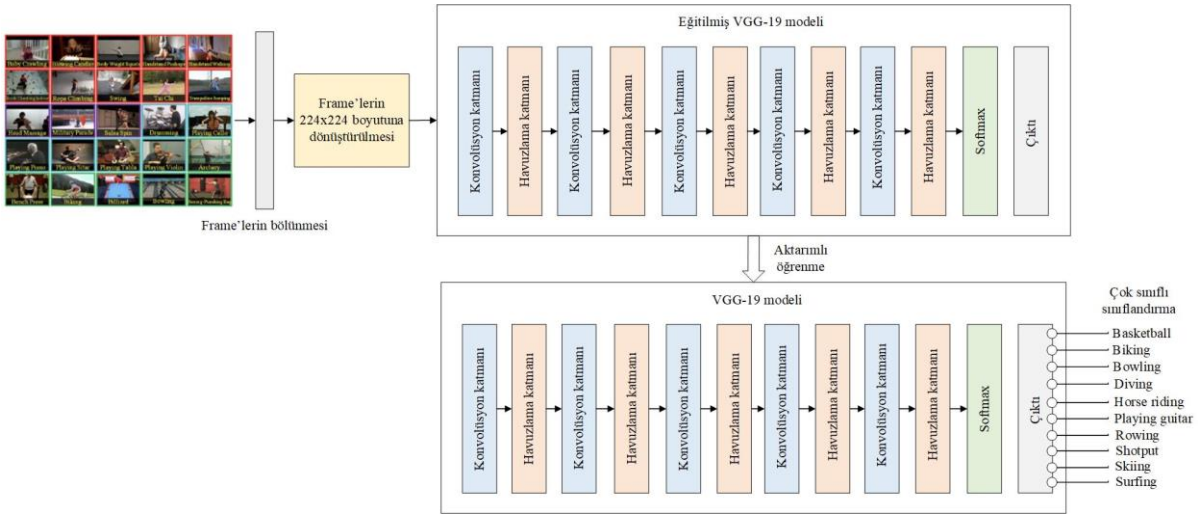
3.3.Geliştirilen Derin Öğrenme Tabanlı Sınıflandırma Modeli (Developed Deep Learning Based Classification Model)

Oluşturulan CNN modelinin 64, 128, 256 ve 512 düğümlü 4 tane konvolüsyon (convolution) katmanı, 4 tane seyreltme (dropout) katmanı, 4 tane havuzlama (pooling) katmanı, 1 tane düzleştirme (flattening) katmanı ve 2 adet 2048 nöronlu tam bağlı (fully-connected) katman bulunmaktadır. Konvolüsyon katmanlarının kernel_size değerleri 3, padding değerleri "same" olarak ayarlanmış, aktivasyon fonksiyonu olarak Doğrultulmuş Lineer Birim (Rectified Linear Unit-ReLU) kullanılmıştır ve ilk konvolüsyon katmanının input_shape değeri (224,224,3) olarak verilmiştir. Optimizer olarak ise Adaptif Moment Tahmini (Adaptive Moment Estimation-Adam) kullanılmıştır. Oluşturulan modelin yapısı Şekil 5'te görülmektedir.



Şekil 5. CNN modeli (CNN model)

CNN kullanılarak yeterli başarı oranına ulaşılamadığı için VGG-19 modeli ile birlikte öğrenme aktarımı uygulanarak iyileştirme yapılması amaçlanmıştır. VGG-19 ile öğrenme aktarımı uygulanarak oluşturulan modelde 5 tane havuzlama katmanı, 3 tane tam bağlı katman, 16 konvolüsyon katmanının ardından 2 tane seyreltme katmanı ve 2 tane tam bağlı katman katman eklenmiştir. Bunlara ek olarak aktivasyon fonksiyonu olarak ReLU, optimizer olarak ise learning rate değeri 0.0001, momentum değeri 0.9 ve decay değeri 0.01 olan Olasılıksal Dereceli Azalma (Stochastic Gradient Descent-SGD) kullanılmıştır. Şekil 6'da geliştirilen VGG-19 modeli görülmektedir.



Şekil 6. Geliştirilen VGG-19 modeli (Developed VGG-19 model)

Kullanılan sınıflandırma algoritmalarının performansını değerlendirmek için karışıklık matrisinden elde edilen doğruluk, kesinlik, duyarlılık ve F-1 puanı metrikleri kullanılmıştır. Karışıklık matrisi, Çizelge 3'te görüldüğü gibi gerçek değerlerin bilindiği bir dizi test verisi üzerindeki bir sınıflandırma modelinin performansını tanımlamak için kullanılan bir tablodur. Karışıklık matrisinde doğru ve yanlış tahminlerin sayısı, sayı değerleriyle özetlenir ve her sınıfa göre ayrılır.

Çizelge 3. Karışıklık matrisi (Confusion matrix)

		Gerçek değerler	
		Pozitif	Negatif
Tahmin değerleri	Pozitif	TP	FN
	Negatif	FP	TN

Doğru pozitif (True Positive-TP), gerçek değeri pozitif olan ve sınıflandırıcı tarafından da pozitif olarak tahmin edilen değerleri ifade etmektedir. Yanlış negatif (False

Negative-FN), gerçek değeri negatif olan ancak sınıflandırıcı tarafından pozitif olarak tahmin edilen değerleri ifade etmektedir. Yanlış pozitif (False Positive-FP), Gerçek değeri pozitif olan ancak sınıflandırıcı tarafından negatif olarak tahmin edilen değerleri ifade etmektedir. Doğru negatif (True Negative-TN) ise gerçek değeri negatif olan ve sınıflandırıcı tarafından da negatif olarak tahmin edilen değerleri ifade etmektedir.

Sınıflandırma performansını ölçmek için doğruluk (Accuracy), kesinlik (Precision), duyarlılık (Recall) ve F-1 puanı (F1-Score) değerleri hesaplanmıştır. Doğruluk metriği en sezgisel performans ölçüsüdür ve sadece doğru tahmin edilen gözlemin toplam gözlemlere oranıdır. Doğruluk metriği Eşitlik 2 kullanılarak hesaplanmaktadır.

$$\text{Doğruluk}=\frac{TP+TN}{TP+FP+FN+TN} \quad (2)$$

Kesinlik, doğru tahmin edilen pozitif gözlemlerin toplam tahmin edilen pozitif gözlemlere oranıdır. Kesinlik Eşitlik 3 kullanılarak hesaplanmaktadır.

$$\text{Kesinlik}=\frac{TP}{TP+FP} \quad (3)$$

Duyarlılık, doğru tahmin edilen pozitif gözlemlerin gerçek sınıftaki tüm gözlemlere oranıdır. Duyarlılık Eşitlik 4 kullanılarak hesaplanmaktadır.

$$\text{Duyarlılık}=\frac{TP}{TP+FN} \quad (4)$$

F-1 puanı, hassasiyet ve duyarlılığın ağırlıklı ortalamasıdır. Bu nedenle, bu puan hem yanlış pozitifleri hem de yanlış negatifleri hesaba katar. F-1 puanı Eşitlik 5 kullanılarak hesaplanmaktadır.

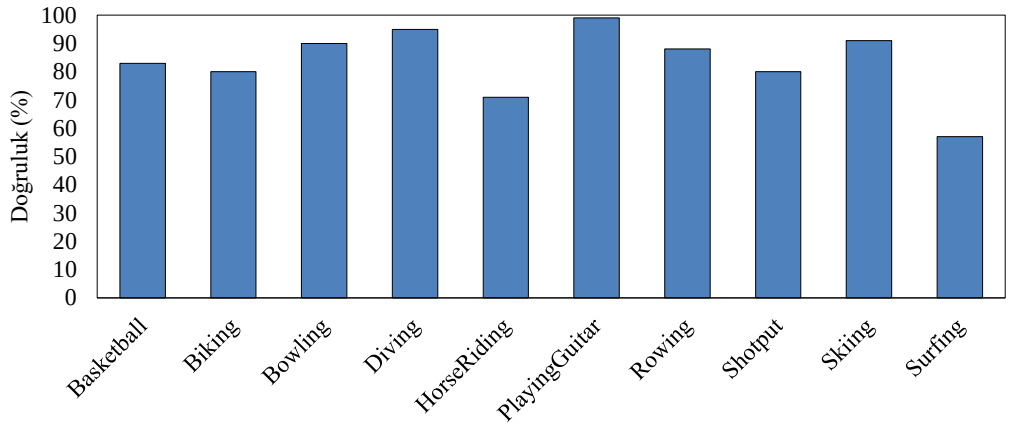
$$\text{F-1 puanı}=\frac{2*\text{Hassasiyet}*\text{Duyarlılık}}{\text{Hassasiyet}+\text{Duyarlılık}} \quad (5)$$

Çizelge 4'te her bir sınıf için sınıflandırma sonuçlarına göre elde edilen karışıklık matrisi görülmektedir.

Çizelge 4. Sınıflandırma sonuçları için karışıklık matrisi (Confusion matrix for classification results

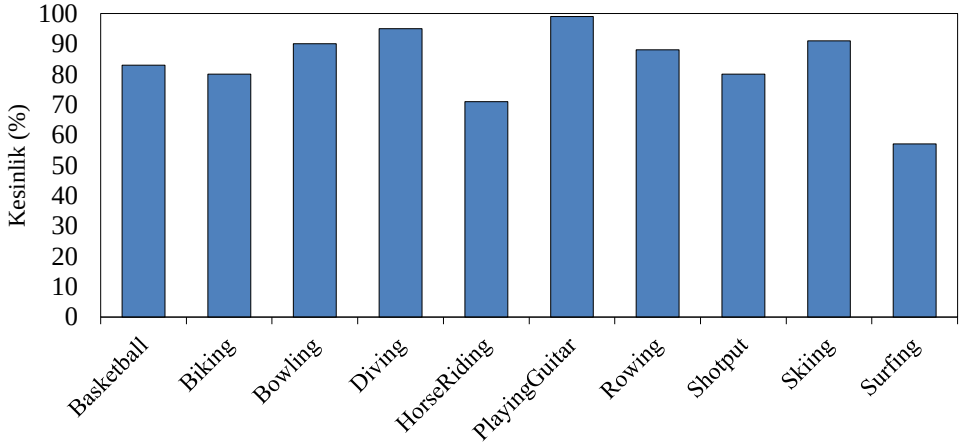
	Basketball	Biking	Bowling	Diving	HorseRiding	PlayingGuitar	Rowing	Shotput	Skiing	Surfing
Basketball	472	44	0	23	61	0	9	63	1	2
Biking	28	858	0	5	121	9	12	0	7	2
Bowling	1	0	731	0	0	0	0	3	10	0
Diving	20	8	10	698	16	1	93	0	2	19
HorseRiding	8	29	0	9	747	1	25	0	37	67
PlayingGuitar	0	44	63	0	0	1328	0	0	0	0
Rowing	0	0	0	1	4	0	1050	0	0	244
Shotput	31	40	5	0	2	0	4	416	15	21
Skiing	6	43	4	0	100	3	5	40	745	249
Surfing	0	1	0	0	2	0	2	0	4	794

Oluşturulan karışıklık matrisinin ardından her bir olay sınıfı için doğruluk, kesinlik, duyarlılık ve F1-puanı hesaplanmıştır. Her bir sınıf için hesaplanan doğruluk değerleri Şekil 7'de görülmektedir.



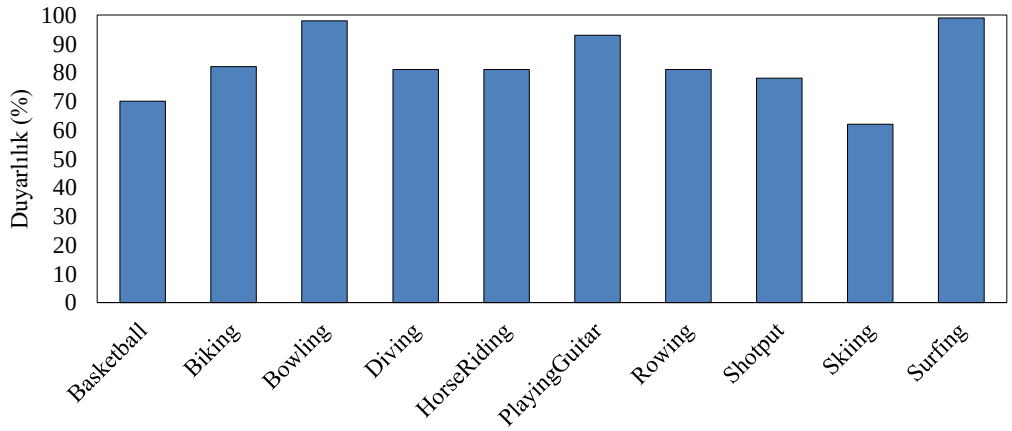
Şekil 7. Her bir sınıf için doğruluk değerleri (Accuracy values for each class)

Şekil 7'de PlayingGuitar, Diving, Skiing ve Bowling sınıfları için elde edilen doğruluk değerlerinin %90'ın üzerinde olduğu görülmektedir. Her bir sınıf için hesaplanan kesinlik değerleri Şekil 8'de görülmektedir.



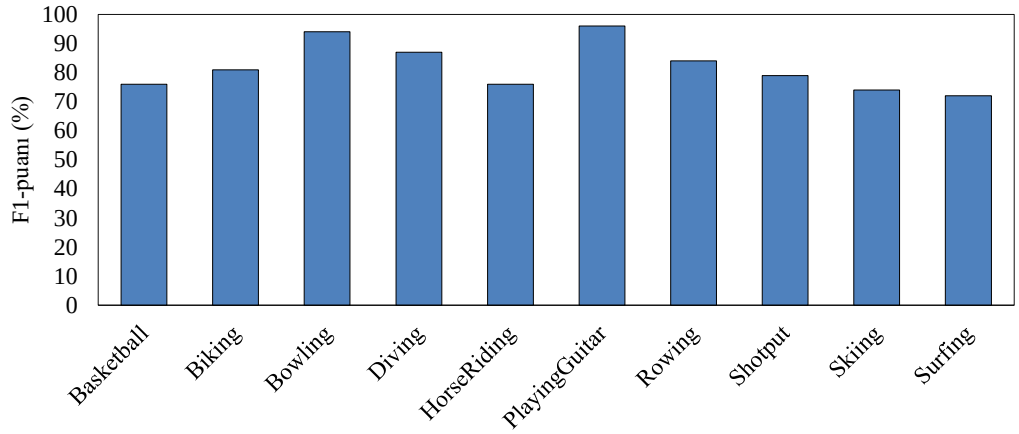
Şekil 8. Her bir sınıf için kesinlik değerleri (Precision values for each class)

Şekil 8'de PlayingGuitar, Diving, Skiing ve Bowling sınıfları için elde edilen kesinlik değerlerinin %90'ın üzerinde olduğu görülmektedir.Her bir sınıf için hesaplanan duyarlılık değerleri Şekil 9'da görülmektedir.



Şekil 9. Her bir sınıf için duyarlılık değerleri (Recall values for each class)

Şekil 9'da PlayingGuitar, Surfing ve Bowling sınıfları için elde edilen duyarlılık değerlerinin %90'ın üzerinde olduğu görülmektedir.Her bir sınıf için hesaplanan F1-puanı değerleri Şekil 10'da görülmektedir.

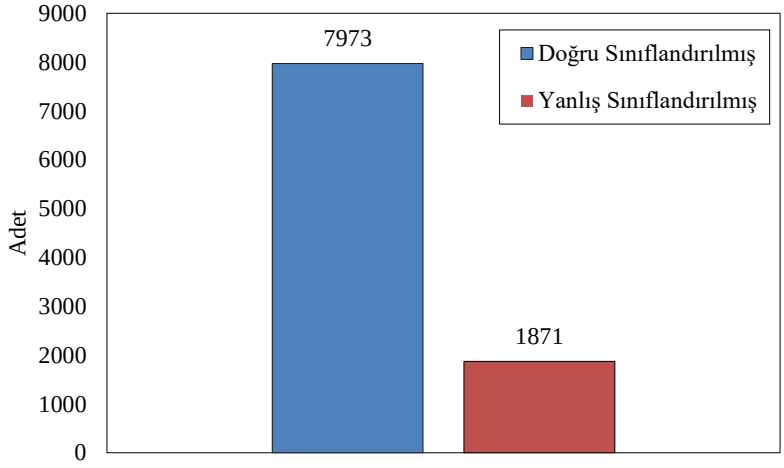


Şekil 10. Her bir sınıf için F1-puanı değerleri (F1-score values for each class)

Şekil 10'da PlayingGuitar ve Bowling sınıfları için elde edilen F1-puanı değerlerinin %90'ın üzerinde olduğu görülmektedir.Oluşturulan karışıklık matrisinin ardından hesaplanan ortalama kesinlik, duyarlılık ve F1-puanı değerleri Çizelge 5'te verilmiştir.

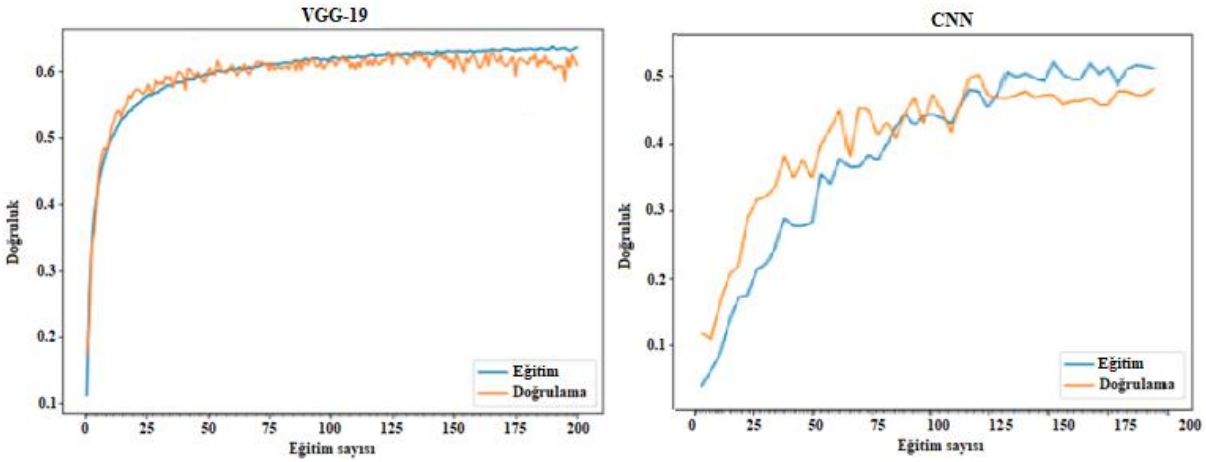
Çizelge 5. Kesinlik, duyarlılık ve F1-puanı sonuçlarının ortalamaları (Average results of precision, recall and F1-score)

	Kesinlik	Duyarlılık	F1-puanı
Ortalama	%85	%82	%83



Şekil 11. Doğru ve yanlış olarak sınıflandırılan örnek sayıları (Number of samples classified as true and false)

Toplam 9844 adet test verisi için VGG-19 tarafından yapılan doğru ve ve yanlış tahminlerin sayısı Şekil 11'de görülmektedir. Şekil 11'de görüldüğü gibi VGG-19 tarafından doğru sınıflandırılan örnek sayısı 7973, yanlış sınıflandırılan örnek sayısı ise 1871'dir. Doğru sınıflandırılmış örnek sayısının sonucu, doğru sınıflandırılmış pozitif örnek ve doğru sınıflandırılmış negatif örnek sayıları toplanarak hesaplanmaktadır. Benzer şekilde yanlış sınıflandırılmış örnek sayısının sonucu, yanlış sınıflandırılmış pozitif örnek ve yanlış sınıflandırılmış negatif örnek sayılarının toplanmasıyla hesaplanmaktadır. Karşılaştırılan modellerin eğitimi sırasında doğruluk değerlerinin eğitim sayısına göre değişimini gösteren grafikler Şekil 12'de görülmektedir.

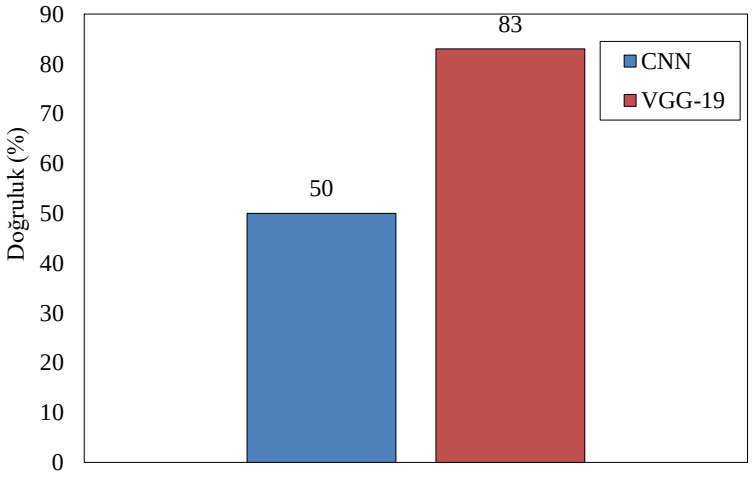


Şekil 12. VGG-19 ve CNN için doğruluk grafikleri (Accuracy graphics for VGG-16 and CNN)

CNN ve VGG-19 için karşılaştırmalı deneysel sonuçlar Şekil 13 ve Çizelge 6'da görülmektedir. Şekil 13 ve Çizelge 6'da görüldüğü gibi VGG-19 %83 doğruluk oranı ile CNN'e göre video üzerinden olay sınıflandırma probleminde daha başarılı olmuştur.

Çizelge 6. Kullanılan modellerin performansları (Performance of the models used)

Model	Doğruluk oranı
CNN	%83
VGG-19	%50



Şekil 13. Kullanılan modellerin başarı oranları (Accuracy rates of the models used)

6. SONUÇ (CONCLUSION)

Video sınıflandırma, insan-bilgisayar etkileşimi ve otonom sürüşteki potansiyel uygulamaları nedeniyle bilgisayarla görü alanında aktif bir araştırma konusu olmuştur. Bununla birlikte, sınıflar arası ve sınıf içi varyasyonlar, video dizilerinin uzamsal ve zamansal yönlerinden türetilen çok modlu karmaşık bilgilerin işlenmesi gibi sınırlamalar nedeniyle video sınıflandırma zorlu bir görevdir. Geleneksel video işleme problemlerine kıyasla videoda olay tanıma, olaylar arasındaki zamansal bağımlılıklara bağlı olduğundan karmaşık bir problemdir. Her adım, zamansal olarak diğerlerine bağlıdır.

Bu çalışmada, derin öğrenme yöntemlerinden biri olan CNN, UCF101 veri seti üzerinde video sınıflandırma için kullanılmış olup çok sayıda model üzerinde deneysel çalışmalar yapılmıştır. Uygulanan modellerde çeşitli parametreler değiştirilerek parametre optimizasyonu yapılmıştır. Ancak %50 doğruluk oranının üzerine çıkılamadığı görüldüğü için VGG-19 ile öğrenme aktarımı uygulanmıştır. VGG-19 modeli ile nesne özellikleri çıkarılarak orijinal video özellikleri ve nesnelerle karşılaştırılmıştır. UCF101 veri setinde elde edilen toplam 32827 görüntüden 22983'ü kullanılarak gerçekleştirilen deneysel çalışmalar sonucunda 9844 görüntüden 7973'ünün başarılı bir şekilde sınıflandırıldığı görülmüştür. Deneysel sonuçlar, VGG-

19 modelinin %83 doğruluk oranı ile video üzerinden olay sınıflandırma problemine uygulanabileceğini göstermiştir. Deneysel sonuçlar, geliştirilen modelin literatürdeki çalışmalara kıyasla daha başarılı sonuçlara sahip olduğunu göstermiştir.

ETİK STANDARTLARIN BEYANI (DECLARATION OF ETHICAL STANDARDS)

Bu makalenin yazarları çalışmalarında kullandıkları materyal ve yöntemlerin etik kurul izni ve/veya yasal-özel bir izin gerektirmediğini beyan ederler.

YAZARLARIN KATKILARI (AUTHORS' CONTRIBUTIONS)

Serim GENÇASLAN: Deneyleri yapmış, sonuçlarını analiz etmiş, makalenin yazım işlemini gerçekleştirmiştir./ Performed the experiments, analyse the results, wrote the manuscript.

Anıl UTKU: Deneyleri yapmış, sonuçlarını analiz etmiş, makalenin yazım işlemini gerçekleştirmiştir./ Performed the experiments, analyse the results, wrote the manuscript.

M. Ali AKCAYOL: Deney sonuçlarını analiz etmiş, makalenin yazım işlemini gerçekleştirmiştir./ Analyse the experimental results, wrote the manuscript.

ÇIKAR ÇATIŞMASI (CONFLICT OF INTEREST)

Bu çalışmada herhangi bir çıkar çatışması yoktur. / There is no conflict of interest in this study.

KAYNAKLAR (REFERENCES)

[1] Çiğdem A.C.I. and Çırak A., "Türkçe Haber Metinlerinin Konvolüsyonel Sinir Ağları ve Word2Vec Kullanılarak Sınıflandırılması", ***Bilişim Teknolojileri Dergisi***, 12(3): 219-228, (2019).

[2] Ma S., Sigal L. and Sclaroff S., "Learning activity progression in lstms for activity detection and early detection", ***Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition***, Las Vegas, NV, USA, 1942-1950, (2016).

[3] Ribeiro P.C., Santos-Victor J. and Lisboa P., "Human activity recognition from video: modeling, feature selection and classification architecture", ***Proceedings of International Workshop on Human Activity Recognition and Modelling***, 61-78, (2005).

[4] Ribeiro P.C., Santos-Victor J. and Lisboa P., "Human activity recognition from video: modeling, feature selection and classification architecture", ***Proceedings of International Workshop on Human Activity Recognition and Modelling***, 61-78, (2005).

[5] Kim E., Helal S. and Cook D., "Human activity recognition and pattern discovery", ***IEEE pervasive computing***, 9(1): 48-53, (2009).

[6] Anguita D., Ghio A., Oneto L., Parra X. and Reyes-Ortiz J.L., "A public domain dataset for human activity recognition using smartphones", ***In Proceedings of the 21th international European symposium on artificial neural networks, computational intelligence and machine learning***, Belgium, 437-442, (2013).

[7] Lin W., Sun M.T., Poovandran R. and Zhang Z., "Human activity recognition for video surveillance", ***2008 IEEE International Symposium on Circuits and Systems***, Washington, USA, 2737-2740, (2008).

[8] Dai X., Singh B., Zhang G., Davis L.S. and Qiu Chen Y., "Temporal context network for activity localization in videos", ***Proceedings of the IEEE International Conference on Computer Vision***, Cambridge, MA, USA, 5793-5802, (2017).

[9] Kay W., Carreira J., Simonyan K., Zhang B., Hillier C., Vijayanarasimhan S. and Suleyman M., "The kinetics human action video dataset", ***arXiv preprint arXiv:1705.06950***, (2017).

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

[11] Kuehne H., Jhuang H., Garrote E., Poggio T. and Serre T., "HMDB: a large video database for human motion recognition", ***2011 International Conference on Computer Vision***, Barcelona, Spain, 2556-2563, (2011).

[12] Sigurdsson G.A., Varol G., Wang X., Farhadi A., Laptev I. and Gupta A., "Hollywood in homes: Crowdsourcing data collection for activity understanding", ***European Conference on Computer Vision***, Amsterdam, Netherlands, 510-526, (2016).

[13] Gu C., Sun C., Ross D.A., Vondrick C., Pantofaru C., Li Y. and Schmid C., "Ava: A video dataset of spatio-temporally localized atomic visual actions", ***Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition***, San Juan, Puerto Rico, USA, 6047-6056, (2018).

[14] Idrees H., Zamir A.R., Jiang Y.G., Gorban A., Laptev I., Sukthankar R. and Shah M., "The THUMOS challenge on action recognition for videos in the wild", ***Computer Vision and Image Understanding***, 155: 1-23, (2017).

[15] Schuldt C., Laptev I. and Caputo B., "Recognizing human actions: a local SVM approach", ***Proceedings of the 17th International Conference on Pattern Recognition***, Cambridge, UK, 32-36, (2004).

[16] Blank M., Gorelick L., Shechtman E., Irani M. and Basri R., "Actions as space-time shapes", ***Tenth IEEE International Conference on Computer Vision (ICCV'05)***, Beijing, China, 1395-1402, (2005).

[17] Rodriguez M.D., Ahmed J. and Shah M., "Action mach a spatio-temporal maximum average correlation height filter for action recognition", ***2008 IEEE conference on computer vision and pattern recognition***, Anchorage, Alaska, 1-8, (2008).

[18] Weinland D., Boyer E. and Ronfard R., "Action recognition from arbitrary views using 3d exemplars", ***2007 IEEE 11th International Conference on Computer Vision***, Rio De Janeiro, Brazil, 1-7, (2007).

[19] Marszalek M., Laptev I. and Schmid C., "Actions in context", ***2009 IEEE Conference on Computer Vision and Pattern Recognition***, Miami Beach, FL, 2929-2936, (2009).

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

[21] Valueva M.V., Nagornov N.N., Lyakhov P.A., Valuev G.V. and Chervyakov N.I., "Application of the residue number system to reduce hardware costs of the convolutional neural network implementation", ***Mathematics and Computers in Simulation***, (2020).

[22] Van den Oord A., Dieleman S. and Schrauwen B., "Deep content-based music recommendation", ***Advances in neural information processing system***s, 2643-2651, (2013).

[23] Collobert R. and Weston J., "A unified architecture for natural language processing: Deep neural networks with multitask learning", ***Proceedings of the 25th international conference on Machine learning***, Helsinki, Finland, 160-167, (2008).

[24] Tsantekidis A., Passalis N., Tefas A., Kanniainen J., Gabbouj M. and Iosifidis A., "Forecasting stock prices from the limit order book using convolutional neural networks", ***2017 IEEE 19th Conference on Business Informatics (CBI)***, Thessaloniki, Greece, 7-12, 2017.

[25] Fukushima K., "Neocognitron". ***Scholarpedia***, 2(1): 1717, (2007).

[26] Hubel D.H. and Wiesel T.N., "Receptive fields and functional architecture of monkey striate cortex", ***The Journal of physiology***, 195(1): 215-243, (1968).

[27] Fukushima K. and Miyake S., "Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition", ***Competition and cooperation in neural nets***, 267-285, (1982).

[28] Li S., Li W., Cook C., Zhu C. and Gao Y., "Independently recurrent neural network (indrnn): Building a longer and deeper RNN", ***Proceedings of the IEEE conference on computer vision and pattern recognition***, Salt Lake City, Utah, ABD, 5457-5466, (2018).

[29] Sundermeyer M., Ney H. and Schlüter R., "From feedforward to recurrent LSTM neural networks for language modeling", ***IEEE/ACM Transactions on Audio, Speech, and Language Processing***, 23(3): 517-529, (2015).