Video Görüntülerinde Gerçek Zamanlı Yüz Tanıma ve Zaman İşaretleme için Yeni Bir Derin Öğrenme Modeli

Araştırma Makalesi/Research Article

Hüseyin GÖZE[1], Oktay YILDIZ[2]

[1]Gazi Üniversitesi, Bilişim Enstitüsü, Bilgisayar Bilimleri Ana, Kavaklıdere, Ankara, 06500, Türkiye
[2]Gazi Üniversitesi, Mühendislik Fakültesi, Bilgisayar Mühendisliği Bölümü, Maltepe, Ankara, 06570, Türkiye
hgoze01@gmail.com, oktyildiz@gmail.com

(Geliş/Received:31.12.2021; Kabul/Accepted:25.02.2022)
DOI: 10.17671/gazibtd.1051738

Özet— Video görüntülerinde gerçek zamanlı yüz tanıma ve görüntü akışı içerisinde etiketlenmesi birçok alanda yüksek öneme sahip bir konudur. Son yıllarda, video görüntülerinde gerçek zamanlı yüz tanıma problemlerinde derin sinir ağları başarılı bir şekilde kullanılmaktadır. Ancak video görüntülerinde yer alan küçük ölçekli yüzlerin tespiti ve aynı zamanda model yanıt süresinin düşürülmesi karşılaşılan önemli zorluklardır. Gerçekleştirilen bu çalışmada, video görüntülerinde gerçek zamanlı yüz tanıma ve zamanın tespiti için yeni bir derin öğrenme modeli önerilmiştir. Yapılan deneysel çalışmalarda önerilen Evrişimli Sinir Ağı tabanlı modelin MTCNN, OPENCV-CNN, HOG+SVM, SSD-CAFFEMODEL modellerine göre daha yüksek performansa ve daha yüksek doğruluk oranına sahip olduğu gösterilmiştir.

Anahtar Kelimeler— yüz tespiti, yüz tanıma, zaman işaretleme, derin öğrenme

A New Deep Learning Model for Real-Time Face Recognition and Time Marking in Video Footage

Abstract— Real-time face recognation and tagging in the video stream are high importance in many areas. In recent years, deep neural networks have been successfully used in real-time facial recognition problems in video images. However, detection of small-scale faces in video images and at the same time reducing response time of model are significant challenges. In this study, new deep learning model is suggested for real-time face recognition and time detection in video images. In experimental studies, the proposed CNN-based model has been shown to have higher performance and more accuracy rate than MTCNN, OPENCV-CNN, HOG+SVM, SSD-CAFFEMODEL models.

Keywords— face detection, face recognition, time marking, deep learning

1. GİRİŞ (INTRODUCTION)

Görüntü işleme uzun yıllardır bilgisayar biliminin ve araştırmacılarının üzerinde çalıştığı ve günümüzde popülerliği artarak devam eden bir problemdir. Görüntü içerisinde bulunan çeşitli nesne veya örüntülerin tespit edilmesi, tanımlanması, sınıflandırılması ve gruplandırılması görüntü işleme uygulamalarının temel araştırma konularıdır. Her türden görüntü arşivlerinde, internet ve sosyal medya platformlarında veya güvenlik amaçlı uygulamalarda, üretilen video görüntülerinin sayı ve boyutları her geçen gün artmaktadır. Artan video görüntüleri içerisinde bulunan kişi veya kişilerin tespit edilmesi, tanımlanması ve doğru bir şekilde etiketlenmesi günümüzde oldukça yüksek bir öneme sahiptir. Video içerisinde yüz tespiti, cinsiyet tespiti, duygu tespiti ve yaş tespiti gibi farklı alanlarda çalışmalar mevcuttur. Video görüntülerinde yüz tespiti ve tanımlanması, görüntüdeki kişilerin tümünün tespit edilmesi gereken durumlarda yüksek önem kazanmaktadır. Akan görüntüler içerisinde insan gözünün ayırt edemeyeceği kadar yüksek sayıdaki yüzün tespiti ve tanımlanması mümkün olamamaktadır. Ayrıca video görüntü taramalarında çok uzun zamanlar kaybedilmekte ve yüksek maliyetler ortaya çıkmaktadır. Bu tür durumlarda akan görüntüler için otomatik

tanımlama, birliktelik tespiti ve zamana göre işaretleme zorunlu bir hal almaktadır. Bu işlemlerin daha hızlı, daha az maliyetli ve daha doğru şekilde yapılabilmesi ihtiyacı her geçen gün artmaktadır. Bu problemlerin çözümü için zaman içerisinde çok çeşitli algoritma ve yöntemler geliştirilmiştir. Önerilen mimarilerden beklenen en temel sonuçlar yüksek başarı ve performanstır.

Derin Öğrenme modelleri görüntü analizinde sıklıkla kullanılmaktadır [1-4]. Derin öğrenme ilk olarak 1943 yılında McCulloch ve Pitts [5] tarafından ortaya atılmış, insan beyninin yapısını ve işleyişini modelleyen bir hesaplama sistemidir. İlerleyen zaman içerisinde bu fikir destek görmüş ve çeşitli çalışmalarla geliştirilmiş ve farklı alanlar için yeniden yorumlanmıştır. 1997 yılında, bilgiyi tekrarlayan geri yayılımla uzun süre boyunca saklamayı öğrenme için Hochreiter ve Schmidhuber Uzun Kısa-Süreli Bellek (LTSM) modelini önermişlerdir [6]. Derin öğrenme en büyük atılımı 2012 yılında gerçekleştirilen ve Google ekipleri tarafından yüksek işlem gücüne sahip yapay desen tanıma algoritmalarının performansı ile yapmıştır. Bu çalışma ile Derin Öğrenmenin performansı insanların algı seviyesine ulaşmıştır [7].

Viola Jones, dijital görüntülerdeki nesnelerin etkin ve hızlı bir biçimde tespitini gerçekleştirebilen ilk algoritma olmuştur [8]. Hızlı bir şekilde gelişimini sürdüren donanımlar sayesinde daha efektif çalışan, özellikle Evrişimli Sinir Ağı (ESA) (Convolutional Neural Network – CNN) tabanlı yaklaşımlar yukarıda söz edilen problemin çözümünde yüksek doğruluk oranına sahip yöntemlerin geliştirilmesine olanak sağlamıştır. Bu yöntemlerin problemi ele alış şekilleri genellikle 4 ana adımdan oluşmaktadır. Bu adımlar sırasıyla görüntünün; ön işlemlerden geçirilip hazırlanması, ilgili nesnenin tespiti, tespit edilen nesnelerin sınıflandırılması ve nesne takibi şeklindedir [9]. Nesne tespit aşamasının başarı oranı sonraki aşamaların da başarı oranını doğrudan etkilemektedir.

Luo, X. ve arkadaşları, çalışmalarında derin öğrenme tabanlı görüntü işleme uygulaması ile araç tespiti ve akıllı trafik izleme sistemi geliştirmişlerdir [10]. Bu çalışmada dokuz katmandan az olmayan bir evrişimli sinir ağı önerilmiştir. Geliştirilen uygulama için birçok farklı açıdan toplanan ve içerisinde araçların bulunduğu görüntü dosyaları kullanılmıştır. Önerilen dokuz katmanlı yöntemde, evrişim katmanları, havuzlama katmanları ve tam bağlantılı katmanlar bulunmaktadır. Evrişim katmanları, özellik çıkarma ve gürültü gidermeden sorumludur. Havuzlama katmanları, parametrelerin azaltılmasında ve işlem hızının iyileştirilmesinde rol oynar. Ayrıca, sınıflandırma ve regresyondan tamamen bağlantılı katmanlar sorumludur. Kullanılan veri kümesi yaklaşık 90.000 görüntü içermektedir. Bu verilerin %80'i eğitim aşamasında, %20'si ise test aşamasında kullanılmıştır. Çalışmanın sonucunda önerilen dokuz katmanlı model ile %97,51 başarı oranı elde edilmiştir.

Pranav, K. ve arkadaşları, çalışmalarında yüksek çözünürlüklü görüntü üreten cihazlar için gerçek zamanlı yüz tanımlama yapılabilmesi için ESA tabanlı bir mimari önermişlerdir [11]. Tasarlanan yöntem, AT&T yüz veri kümesi üzerinde denenmiştir. Önerilen ESA mimarisi, en basit haliyle [INPUT – CONV − RELU − POOL − FC] olan bir dizi katmandan oluşur. INPUT katmanı, görüntülerin ham piksel değerlerini tutar, CONV katmanı, özellikleri çıkarmak için pencereli görüntü üzerinde evrişim işlemini gerçekleştirmek için bir pencere tarzında kayan sabit boyutlu bir çekirdek veya filtreden oluşur. Filtre boyutuyla eşit olmayan eşlemenin üstesinden gelmek için giriş görüntüsünün boyutuna dolgu uygulanır. RELU, gizli birimlere sıfır değeri atayan bir etkinleştirme işlevi olan düzeltilmiş doğrusal birimler anlamına gelir. POOL, verileri işlemek için gereken hesaplama gücünü azaltan aşağı örneklemeden ve boyutsallığın azaltılmasından sorumlu olan havuzlama katmanını ifade eder. Havuzlama katmanı ayrıca, dönme ve konumsal değişmeyen baskın özellikleri çıkarmak için girdiye bir pencere gibi kayan bir çekirdeğe veya işleve sahiptir. Maksimum havuzlama ve ortalama havuzlama, kullanılan iki yaygın işlevdir. FC, girdideki her nöronun çıktıdaki her bir nörona bağlandığı katmandır, ayrıca bu katman, belirli bir sınıfın puanını hesaplamaktan sorumlu olduğu, n'nin sınıflandırılacak (sınıf / kategori) sayısını gösterdiğin çıktı ile sonuçlanan tam bağlantılı katmandır. Maksimum puana sahip sınıf, ESA mimarisinin öngörülen sınıfı olarak belirlenir. FC katmanına DENSE katmanı da denir. ESA mimarisinin, sistemin tasarım gereksinimleri ve performansına bağlı olarak değiştirilebileceği ifade edilmiştir. ESA mimarisinde kullanılan diğer katmanlardan bazıları DROPOUT ve FLATTEN içerir. DROPOUT katmanı, ESA'nın aşırı uyumunu önlemek için bir düzenleme tekniğidir, burada girişlerin bir kısmı (bırakma oranı olarak adlandırılır), eğitim sırasında her güncellemede değerleri 0'a ayarlanarak çıkarılır. Tutulan girdilerin değerleri ölçeklendirilir, böylece toplamları eğitim sırasında değişmez. FLATTEN katmanlar, iki boyutlu unsurları tek boyuta dönüştürmek için FC katmanından önce eklenir. Bu çalışmada kameradan alınan canlı yayın akışında yüz algılama için "Viola Jones" algoritması kullanılmıştır. Algılanan yüz kırpıldıktan sonra 120x120 piksel ölçülerinde yeniden boyutlandırılır. Daha sonra bu görüntü ESA (ilk 32 katman) katmanlarına iletilir. Bu katmandan elde edilen 3x3 piksel boyutundaki resimler bir sonraki katmana iletilir ve daha sonra tasarlanan pooling ve dropout katmanlarından geçirilir. Belirlenen başarım parametrelerine göre tanımlama gerçekleştirilmiştir. AT&T veri kümesi üzerinde çalıştırılan bu model ile %98,75'e varan maksimum doğru tanımlama oranına ulaşılmıştır.

Mahmood, Z. ve arkadaşları, yapmış oldukları çalışmada Yüz Tanıma (FR) algoritmalarına genel bir bakış sunmuştur. Her yaklaşımın tanıma oranı ile ilgili olarak görüntü veri tabanlarının koşulları vurgulanmaktadır. Bu çalışmanın, yüz tanıma ile ilgili güncel metodolojilere ve genel kurallara hızlı bir bakış için ve algoritma seçimleri açısından yararlı olduğu düşünülmüştür [12]. Deneyler sırasında, insan yüzü görüntüleri içeren FERET veri tabanından eğitim için rastgele 300 resim seçilmiştir. Seçilen resimlerin 80 tanesi eğitim aşamasında

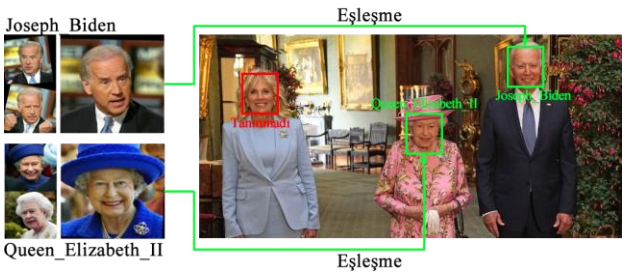
kullanılmıştır. Eğitim kümesindeki 80 görüntünün tamamı doğru şekilde etiketlenmiş ve diğer 220 görüntüden 207'si doğru bir şekilde belirlenmiştir. Böylece ortalama %95,67'lik bir tanıma oranı elde edilmiştir. FERET veri tabanından sonra CAS-PEALR1, LFW ve HFB veri tabanları üzerinde gerçekleştirilen deneyler, sırasıyla %86,8, %84,02 ve %92,2'lik ortalama doğruluk oranları elde edilmiştir.

Gerçekleştirilen çalışmada, video görüntülerinde geçen tanınan veya tanınmayan kişileri düşük donanım gereksinimi ile gerçek zamanlı olarak tespit edebilen bir mimari geliştirilmiştir. Gerçek zamanlı tanımlanan kişiler için etiketleme işlemi yapılmakta, tanımlanamayan kişiler için de ayrı bir havuz oluşturulmakta ve uygulamanın daha sonra bu kişileri de tanıyıp sisteme dahil etmesi sağlanmaktadır. Bu sayede yüksek işlem gücü gerektirmeden, düşük zaman ve iş gücü maliyeti ile sürekli bir şekilde öğrenen, tanıyan ve etiketleyen bir mimari oluşturulmuştur. Oluşturulan bu mimari gerçek zamanlı tespit gerektiren birçok alanda kullanılabilecektir.

2. MATERYAL VE METODOLOJİ (MATERIALS AND METHODOLOGY)

2.1. Veri Kümesi (Dataset)

Gerçekleştirilen deneysel çalışmalarda, 5749 farklı kişiye ait 13.000'den fazla yüz görüntüsü içeren Labeled Faces in the Wild Home [13] veri kümesi kullanılmıştır. Her yüz görüntüsü kişinin adıyla etiketlenmiş durumdadır. Veri kümesi içerişindeki kişilerin 1680'inin 2 veya daha fazla görüntüsü bulunmaktadır. Bu yüzlerdeki tek kısıtlama Viola-Jones yüz dedektörü tarafından tespit edilmiş olmalarıdır. Veri kümesinde bulunan resimler .jpg uzantılı ve 250x250 piksel boyutlara sahiptir. Veri kümesinin dosya yapısı Şekil-1'de gösterilmiştir.

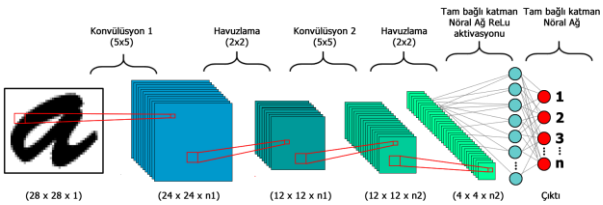


Şekil 1. Labeled Faces in the Wild Home veri kümesi yapısı
(Labeled Faces in the Wild Home dataset structure)

Deneysel çalışmada kullanılan video görüntüleri, Labeled Faces in the Wild Home veri kümesinde bulunan kişilerin içerisinde bulunduğu, internet ortamında herkese açık bir şekilde paylaşılan videolardan oluşturulmuştur. Kullanılan bu video görüntüleri farklı çözünürlüklere sahip ayrı dosyalara dönüştürülmüştür.

2.2. Metodoloji (Methodology)

Evrişimli Sinir Ağı derin öğrenme yöntemi benimsenerek tasarlanmıştır. Evrişimsel katman, lineer olmayan katman, havuzlama katmanı ve tam bağlı katman olmak üzere beş adet katmandan oluşmaktadır. ESA modelleri görüntü işleme ve analizinde yaygın olarak kullanılmaktadır [2]. Temel çalışma prensibi Şekil-2'de gösterilmiştir.



Şekil 2. Temel bir ESA mimarisi
(A basic CNN architecture)

Bir ESA, bir girdi görüntüsünü alabilen, görüntüdeki yön ve nesnelere öğrenilebilir ağırlıklar ve önyargılarla önem atayabilen ve bu şekilde görüntüleri birbirlerinden ayırabilen bir Derin Öğrenme algoritmasıdır. ESA modeli için gerekli olan ön işleme algoritması, diğer sınıflandırma algoritmalarına kıyaslandığında çok daha düşük kalmaktadır. Filtreler, ilkel yöntemlerde yeterli eğitimle elle üretilirken, ESA bu filtre ve özellikleri öğrenme yeteneğine sahiptir. Bir ESA'nın mimarisi, biyolojik insan beyninden esinlenmiştir ve insan beynindeki nöronların bağlantı modeline benzerdir. Bireysel nöronlar uyaranlara yalnızca Alıcı Alan olarak bilinen görme alanının sınırlı bir bölgesinde yanıt verir. Buna benzer alanlardan oluşan bir koleksiyon, görsel alanın tamamını kapsayacak biçimde örtüşür.

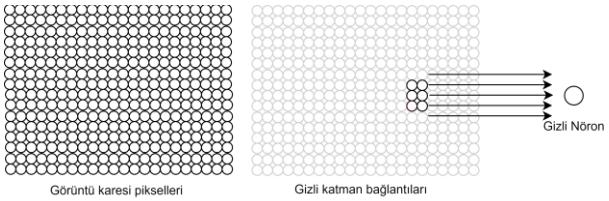
MTCNN (Multi-task Cascaded Convolutional Networks), hem yüz algılama hem de yüz hizalama için geliştirilmiş bir modeldir. MTCNN, yüzleri; göz, burun ve ağız gibi dönüm noktası konumunu tanıyabilen farklı ağ aşamalarından oluşmaktadır. MTCNN modeli üç evrişimli ağa (P-Net, R-Net ve O-Net) sahiptir. Model, gerçek zamanlı performansı sağlarken diğer yüz algılama modellerinden daha iyi performans sonuçları vermektedir. Bu model görüntü içinde yer alan yüzleri tespit etmek için farklı boyutlarda alt resimler üreterek bir piramit yapısı oluşturur. Oluşturulan her kopyada yüz taraması yapılır ve yüz bulunması durumunda bir sonraki aşamaya geçiş sağlanır. MTCNN ilk aşamada yukarıdaki üç aşamalı ağın girdisi olan bir görüntü piramidi oluşturmak için görüntüyü alıp farklı ölçeklerde yeniden boyutlandırmaktır [3].

HOG (Histogram Of Oriented Gradients), nesnelerin tespiti için görüntülerde kullanılan özellik tabanlı bir tanımlayıcıdır ve genelde 3D fotoğraflarda kullanılmaktadır. Bundan kaynaklı olarak yavaşlığı konusunda genel kanı bulunmaktadır. HOG, basit ve güçlü bir özellik tanımlayıcısıdır. Sadece yüz tanıma için değil, aynı zamanda arabalar, evcil hayvanlar ve meyveler gibi nesne tespiti için de yaygın olarak kullanılmaktadır. HOG, nesne algılama için sağlamdır çünkü nesne şekli, yerel yoğunluk gradyan dağılımı ve kenar yönü kullanılarak

karakterize edilir. Gradyan değerlerinin hesaplanması aşamasında görüntü üzerinde hem yatay hem de dikey yönde bir filtre uygulanmaktadır. Bu sayede görüntünün renk ve doygunluk verileri filtrelenebilmektedir. İkinci aşama olan yönlendirme gruplaması ile hesaplanan gradyan değerlerine göre her piksel için bir ağırlıklandırma işlemi yapılır. Son aşamada ise kontrast ve aydınlık bölgelerdeki gradyan değerinin güçlü yönleri normalleştirilir [1].

SSD (Single Shot Detector), nesne algılama için ESA tabanlı sınıflandırıcılar olan bir temel derin öğrenme ağını kullanarak özellik haritasını çıkarır ve nesneleri algılamak için evrişim filtreleri uygular. Caffe framewok ise Berkeley AI Research ve topluluğa katkıda bulunanlar tarafından geliştirilmiş bir derin öğrenme çerçevesidir. Caffe modeli diğer modellere kıyasla daha hızlı ve daha verimli çalışmaktadır [4].

ESA, görüntülerden özellik çıkarımı için üç temel yöntem kullanır. Bunlar Yerel Alıcı Alanlar, paylaşılan ağırlıklar ve havuzlamadır [14]. Yerel Alıcı Alanlar: Evrişimsel sinir ağlarında, gizli nöronların katmanı küçük bağlantılara sahiptir ve diğer sinir ağlarından farkı giriş piksellerinin lokal bölgesi, her giriş pikselinin birebir bağlantısına sahip olmasıdır. Diğer bir ifadeyle; ilk gizli katmandaki her nöron, giriş nöronlarının küçük bir bölgesi ile bağlantılı olarak çalışmaktadır. Belirli gizli nöronlar için bağlantı aşağıdaki Şekil-3'teki gibi olacaktır [15].



Şekil 3. ESA mimarisinde bulunan gizli nöronların yapısı
(Structure of hidden neurons in CNN architecture)

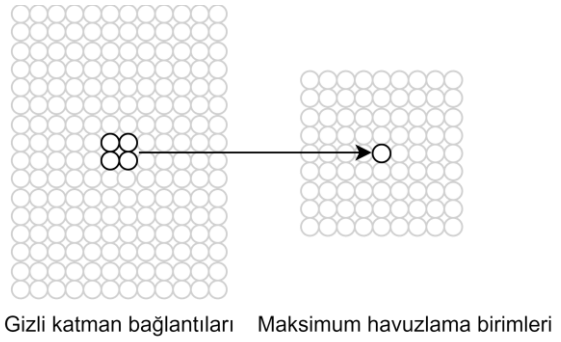
Giriş görüntüsündeki bu bölge, giriş pikselleri üzerinde küçük bir pencere olan gizli nöron için yerel alıcı alan olarak adlandırılır. Alıcı alan, her gizli katman nöronlarıyla bağlantı kurmak için kayarak çalışır.

Paylaşılan Ağırlıklar: Aynı uzaklığa sahip, gizli katmana bağlı yerel alıcı alanlar, gizli katman boyunca aynı ağırlığı ve sapmayı paylaşacaktır. j^{th} için, k^{th} gizli nöron çıktısı Eşitlik-1'de gösterildiği gibi hesaplanabilir.

$$o(j,k) = \sigma\left(b + \sum_{l=0}^{4}\sum_{m=0}^{4} W(l,m)\,\alpha_j^{+l,k+m}\right) \quad (1)$$

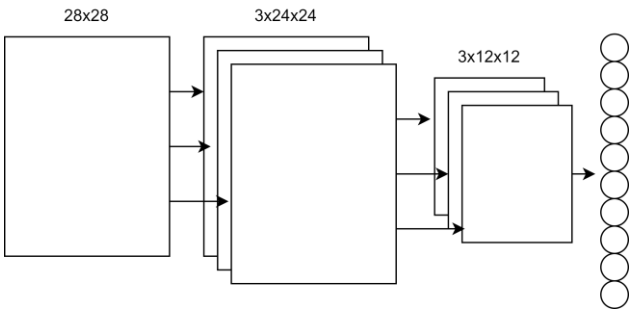
σ nöral aktivasyon fonksiyonu veya sigmoid fonksiyon olduğunda, b paylaşılan sapma, W(l,m) paylaşılan ağırlıklardır ve $\alpha_{x,y}$ x, y konumunda girdi aktivasyonunu beslemektedir. İlk gizli katmandaki nöronlar, girişte bulunan özellikleri farklı konumlarda algılamaktadır. Bu nedenle, girişin gizli katmanından gelen harita, özellik haritası için kastedilen sapma değeri ortak sapma olarak adlandırılmaktadır. Paylaşılan ağırlıklar ve sapmalar, çekirdek veya filtre olarak adlandırılmaktadır. Yüz tanıma birçok özellik eşler, ancak çekirdek parametreleri etkili bir şekilde azaltır [15].

Havuzlama Katmanı: Havuzlama katmanı, konvolüsyon katmanının özellik haritasından, çıktıdaki bilgileri yoğunlaştırılmış özellik haritasına basitleştirmek için kullanılan konvolüsyon katmanı tarafından takip edilir. Şekil-4'te görüldüğü gibi maksimum havuzlama, 2x2 giriş bölgesindeki maksimum etkinleştirmenin çıktısını döndürür [15].



Şekil 4. ESA mimarisinde kullanılan havuzlama katmanı
(The pooling layer used in the CNN architecture)

Yüz tanıma aşamasında kullanılan model, 28 dönüşüm katmanına sahip bir ResNet ağıdır. Zhang ve diğerleri tarafından yayınlanan "Görüntü Tanıma için Derin Kalıntı Öğrenme" [16] isimli çalışmada önerilen ResNet-34 ağının birkaç katmanı kaldırılmış ve katman başına filtre sayısı yarı yarıya azaltılmış, Şekil-5'te temel yapısı gösterilen bir versiyonudur. ResNet-34 özellikle görüntü sınıflandırma modeli olarak kullanılabilen 34 katmanlı bir evrişimsel sinir ağıdır. Bu ağ 200 farklı sınıfta 100.000 görüntüye sahip bir veri kümesi olan ImageNet veri kümesinde önceden eğitilmiş bir modeldir.



Şekil 5. 28 katmanlı Evrişimli Sinir Ağları modeli
(28-layer Convolutional Neural Networks model)

2.3. Değerlendirme Metrikleri (Evaluation Metrics)

Bu makalede yüz tespitinin ve tanımlanmasının doğru kabul edilebilmesi için Eşitlik-2'de gösterildiği gibi bir eşleşme eşik değeri kullanılmıştır. Literatürdeki [17, 18] yüz tespiti için farklı eşik değerleri kullanılmış olsa da

gerçekleştirilen deneysel çalışmalarda en uygun eşik değeri 0,6 olarak belirlenmiştir. Makalede kullanılan diğer değerlendirme ölçütleri sınıflandırma doğruluğu (Accuracy), Kesinlik (Precision), Duyarlılık (Recall) ve F1 Skoru aşağıda sırasıyla Eşitlik-3, Eşitlik-4, Eşitlik-5 ve Eşitlik-6'da gösterilmiştir. F1 Skor değerleri ile bir saniyede işlenen toplam görüntü karesi (FPS) dikkate alınmıştır. Eşitliklerde yer alan TP, FP, TN ve FN sırasıyla Doğru Pozitif, Yanlış Pozitif, Doğru Negatif ve Yanlış Negatiftir.

$$e=\begin{cases} \geq 0.6 \text{ Doğru} \\ <0.6 \text{ Yanlış} \end{cases} \quad (2)$$

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \quad (3)$$

$$\text{Precision}=\frac{TP}{TP+FP} \quad (4)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (5)$$

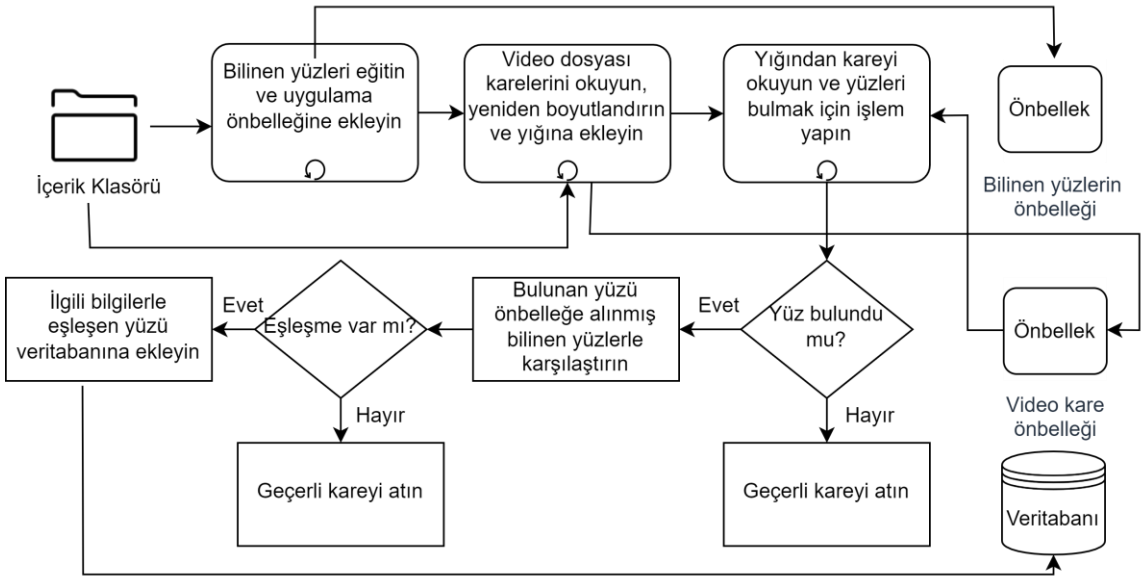
$$F1=2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (6)$$

3. DENEYSEL ÇALIŞMA (EXPERIMENTAL STUDY)

Deneysel çalışmalar 3.80 GHz CPU, 32 GB RAM, 8 GB Nvidia Cuda ekran kartına sahip bir bilgisayar üzerinde gerçekleştirilmiş, Şekil-6'da gösterilen adımlar izlenmiştir. İlk aşamada veri kümesinde bulunan kişilere ait yüz görüntüleri işlenmiş ve bu kişilere ait yüzlere karşılık gelen eğitim modelleri oluşturulmuştur. Her bir kişi için en az bir en fazla beş görüntü dosyası modele dahil edilmiştir. Sonraki aşamada, uygulamanın daha hızlı işlem ve tanımlama yapabilmesi amacıyla, eğitim modelleri kurulan Cache (Uygulama Ön Belleği) yapısına dahil edilmiştir. Bu sayede disk okuma/yazma işlemlerinden kaynaklanan gecikmeler en aza indirgenmiştir. Havuzdan alınan ilgili karede bir veya birden fazla yüz tespit edilmesi durumunda, bulunan yüzler sırası ile ön bellekte bulunan tanınmış kişilerin yüzleri ile karşılaştırılmakta ve eşleşme bulunması durumunda veri tabanına kayıt edilmektedir. Mimarinin Sözde kodu (Pseudocode) Şekil-7'de gösterilmiştir.

Görüntü içinde yer alan nesne veya örüntülerin tespit ve tanımlanması aşamalarında, başarı oranı ve çalışma performansı öne çıkan iki temel problemdir. Görüntü içindeki nesneleri doğru bir şekilde tespit edilmesi, tanımlanması, etiketlenmesi ve bu işlemlerin mümkün olan en hızlı ve en az maliyetli olacak şekilde yapılması beklenmektedir. Bu çerçevede gerçekleştirilen çalışmada CNN, MTCNN, OPENCV-CNN, HOG+SVM, SSD-CAFFEMODEL ve önerdiğimiz ESA tabanlı mimari ile geliştirilen farklı uygulamalarda Labeled Faces in the Wild Home veri kümesi ortak kullanılmıştır. Veri kümesinde etiketlenmiş kişilere ait farklı çözünürlükteki birbirinden farklı video dosyaları kullanılmıştır. Video dosyalarında bulunan kişilerin, minimum gerçek zamanlı olacak şekilde tespit edilmesi, tanımlanması ve video içinde geçtiği zamana göre etiketlenmesi amaçlanmıştır.

Önerdiğimiz ESA tabanlı mimari ile görüntü içerisinde bulunan Şekil-8'de gösterilen örüntüye sahip desenler yüz olarak kabul edilmektedir. Çalışmamızda 68 noktadan oluşan yüz tanımlayıcı model kullanılmıştır. Görüntü karelerinde bulunan yüzlerin tespit edilmesi için ESA tabanlı modeller kullanılmıştır. Yüz örüntüsü tespit edildikten sonra görüntü içerisinde bulunan yüz dikey hizalama işlemi uygulanır. Son aşamada ise yüz görüntüsü belirlenen çözünürlükte kırpıldıktan sonra tanıma işlemine hazır hale gelir.



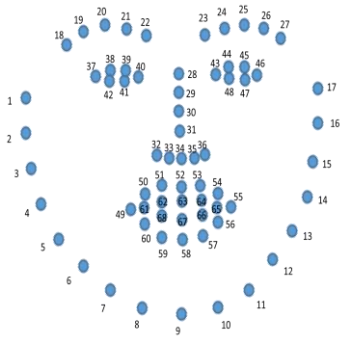
Şekil 6. Önerilen sistem mimarisi.
(Recommended system architecture)

Önerilen Mimarinin Sözde Kodu

Video dosyalarında gerçek zamanlı yüz tanımlama.

Giriş: Resim ve video görüntü kareleri f ∈ {1,...,N}, T toplam video kare sayısını ifade eder.
Çıkış: Tespit edilen ve tanınan yüzler, zaman damgaları, kare içindeki yüz koordinatları, kare numarası
Prosedür:
1: Tanınmış yüzleri eğit.
2: Her yüzün sayısal karşılığını ön belleğe al
3:Video karelerini oku, yeniden boyutlandır ve kare kuyruğuna ekle
4: for t = 1 de T'ye kadar do
5: Geçerli kareden yüzleri bul
6: if Geçerli çerçevede herhangi bir yüz bulundu ise
7: Bulunan yüzleri eğitilmiş yüzlerle karşılaştır
8: if Herhangi bir yüz eşleşmesi var ise
9: Veritabanına eşleşen yüzü ekle
10: end if
11: end if
12: end for

Şekil 7. Uygulanan algoritmanın sözde kodu
(Pseudo code of the algorithm)

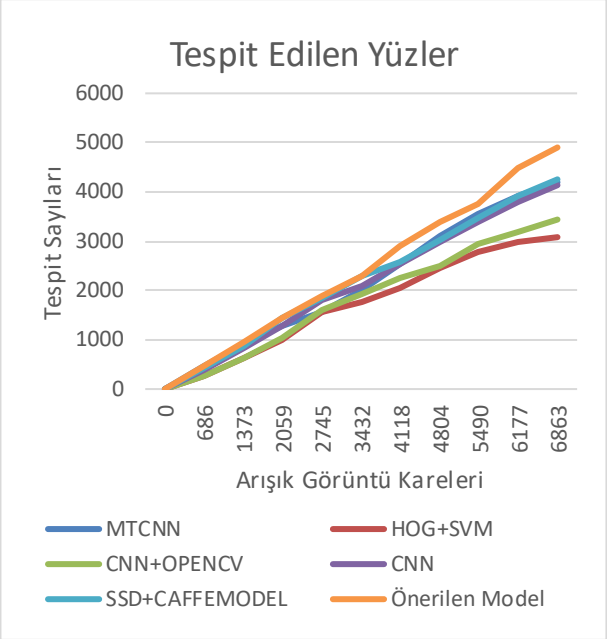


Şekil 8. Tespit edilen yüz örüntüsü
(Detected face pattern)

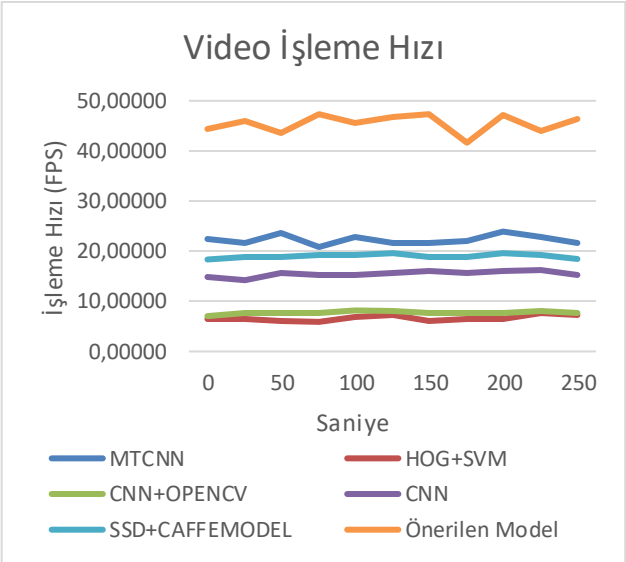
Önerilen model ile beraber diğer tüm modeller aynı şartlar altında, ortak veri kümesi ile test edilmiş, performans ve başarım oranları hesaplanmıştır. Şekil-9'da tespit edilen yüz sayıları görülmektedir. Kullanılan video görüntüsünde toplam 6863 görüntü karesi mevcuttur, bu görüntü karelerinin 4941 tanesinde yüz örüntüsü bulunmaktadır. Kullanılan mimarilerden HOG+SVM 3082, CNN+OPENCV 3440, CNN 4137, MTCNN 4196, SSD+CAFFEMODEL 4255 ve önerdiğimiz ESA tabanlı model 4902 adet yüz örüntüsünü başarılı bir şekilde tanımlamıştır. Kullanılan video görüntüsünün toplam işlenme süreleri Şekil-10'da gösterilmiştir.

4. DENEYSEL SONUÇLAR (RESULTS OF WORK)

Tasarlanan ESA tabanlı mimari ile video görüntülerinde bulunan kişilere ait yüzler tespit edilmiş, tanımlanmış ve veri tabanına kayıt edilmiştir. Tanımlanan kişilere ait isim, video içinde geçtiği zaman, görüntü karesi içindeki koordinatları, kare numarası ve tanımlama zamanı bilgileri kayıt altına alınmıştır. Oluşan örnek veri tabanı kayıtları Şekil-11'de gösterilmiştir. Tespit edilen kişilerin ismi, milisaniye cinsinden video içerisinde geçtiği zaman, görüntü karesi içindeki x_1, y_1, x_2 ve y_2 koordinatları, kare numarası ve tespit tarihleri veri tabanına kaydedilmiştir.



Şekil 9. Uygulanan model ve tespit edilen yüz sayısı
(Implemented model and number of detected faces)



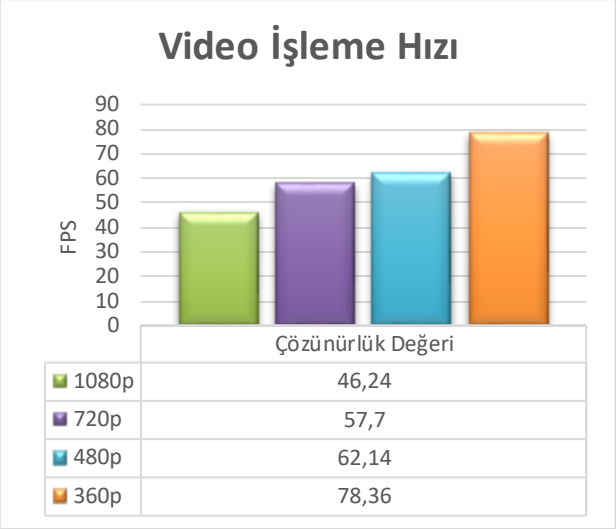
Şekil 10. Uygulanan mimarilerin çalışma süreleri
(Run times of implemented architectures)


```
[ {
    "isim": "Kişi 1",
    "gectigiZaman": 8108,
    "konumu": [57, 336, 97, 297],
    "kareNumarasi": 243,
    "kayitTarihi":"30/11/2021 17:49:09.988"
  }, {
    "isim": "Kişi 2",
    "gectigiZaman": 8208,
    "konumu": [43, 366, 96, 274],
    "kareNumarasi": 246,
    "kayitTarihi": "30/11/2021 17:49:10.030"
  }
]
```


Şekil 11. Video işlemesi sonucunda oluşan veri tabanı kayıtları
(Database records of video processing)

Uygulanan mimari ile yapılan deneylerde, örnek olarak alınan video dosyalarının farklı çözünürlüklerdeki versiyonları aynı algoritma üzerinde çalıştırılmış ve elde edilen sonuçlar karşılaştırılmıştır. Örnek videoların 1080, 720, 480 ve 360 piksel genişliğe sahip boyutlarındaki versiyonları ayrı ayrı çalıştırılmış ve elde edilen sonuçlar Şekil-12'de gösterilmiştir. İşlenen ham video dosyasının FPS (Frame Per Second) değeri 29.97 iken geliştirilen algoritma ile 1080 piksel çözünürlükteki video dosyası ortalama 46.24 FPS ile gerçek zamanlı video görüntüsünden 1,54 kat daha hızlı bir şekilde işlenmiş ve içerisindeki yüzler tanımlanmıştır. Video görüntüsünün çözünürlüğü düşürüldüğünde işleme hızında artış kaydedilmekte ve 360 piksel boyutlarındaki video görüntüsünde 78.36 FPS ile gerçek zamanlı video görüntüsünden 2,61 kat daha hızlı şekilde işlenmiş ve içerisindeki yüzler tanımlanmıştır.

Ayrıca Tablo-1'de deneysel çalışma ve ilgili diğer modellerin elde ettiği başarı oranları verilmiştir.



Şekil 12. Önerilen modelin farklı video çözünürlüklerine ait işleme hızları
(Processing speeds of the proposed model for different video resolutions)

Tablo 1. Uygulanan mimarilerin karşılaştırma tablosu
(Comparison table of implemented architectures)

Karşılaştırma Tablosu	Elde edilen Sonuçlar		
Model	*Kullanılan Veri Seti*	*Doğruluk (%)*	*F1 Skor (%)*
HOG+SVM [1]	Yale	92,00	-
CNN+OPENCV [19]	Kaggle's Medical Mask Dataset	99,00	99,00
CNN [11]	AT&T	98,75	-
MTCNN [3]	Wider Face	95,40	-
SSD+CAFFEMODEL [4]	Pascal Voc	90,70	-
Onerilen Model	**Labeled Faces in the Wild Home**	**99,43**	**99,60**

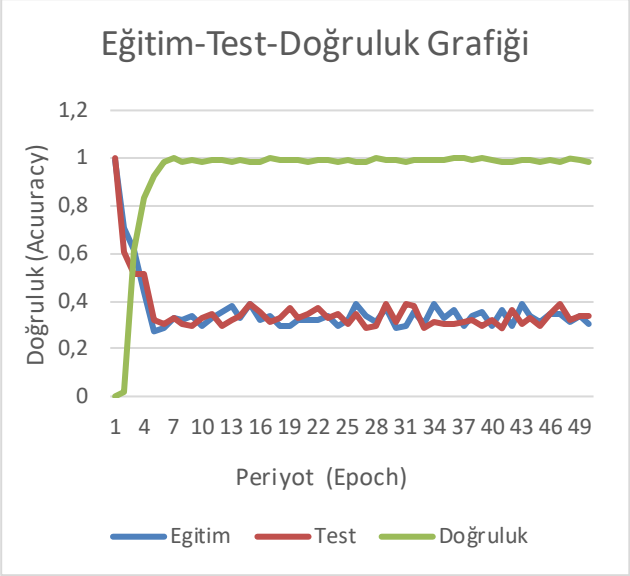
Geliştirilen mimari, görüntü işleme için sıklıkla kullanılan geleneksel ESA modellerine göre daha az sayıda katmana sahiptir. 28 katmana sahip geliştirilen ESA modeli ile görüntü karelerinin işlenme hızı arttırılmıştır. Geliştirilen mimaride ön işleme aşamasında oluşturulan ön bellekler sayesinde disk okuma-yazma işlemlerinden kaynaklı gecikmeler ortadan kaldırılmıştır. Ayrıca oluşturulan algoritmada kullanılan paralel programlama teknikleri ile görüntü işleme ve karşılaştırma aşamalarındaki gecikme seviyesi en aza indirgenmiştir. Bu sayede geleneksel yüz tanıma uygulamalarına göre %33'e varan performans artışı kaydedilmiştir. Geliştirilen mimaride yüz tespiti için veri setinde bulunan görüntülerin %60'lık kısmı eğitim aşamasında %40'lık kısmı ise test aşamasında kullanılmıştır. Geliştirilen ESA modeli için her bir verinin modelden geçiş sayısı (Epoch) 50, aynı anda eğitilecek veri sayısı (Batch size) ise 16 olarak seçilmiştir. Ayrıca eğitim ve test aşamasında kullanılan veriler daha doğru sonuçlar alabilmek için her aşamada kendi grubu içerisinde karıştırılmıştır (Shuffle). ESA modeli için kullanılan bu değerler hem literatürdeki kullanıma hem de model geliştirilmesi aşamasında elde edilen sonuçlara göre en optimal değerler olarak belirlenmiştir. Bu parametreler Tablo-2'de gösterilmiştir. ESA modelinin belirlenen bu parametreler ile oluşan Eğitim-Test-Doğruluk grafiği Şekil-13'te gösterilmiştir.

Tablo 2. ESA modeli için kullanılan parametreler
(Parameters used for the CNN model)

ESA Parametre Tablosu	Kullanılan Değerler	
	Parametre	*Değer*
	Periyot (Epoch)	50
	Periyot Veri Sayısı (Batch Size)	16
	Karıştırma (Shuffle)	Evet

Geliştirilen mimarinin farklı çözünürlüklere sahip video görüntülerini gerçek video zamanından daha hızlı bir şekilde işlediği tespit edilmiştir. Uygulanan mimari ile video görüntülerinde tespit edilen kişilerin doğru bir şekilde tanımlanması ve kayıt altına alınması sağlanmıştır. Geliştirilen mimari ile taranan 6863 kareye sahip örnek video dosyasında bulunan toplam 4941 yüz örüntüsünden 4902'si başarılı bir şekilde tespit edilmiştir. Video karelerinde bulunan 39 yüz örüntüsü tespit edilememiştir.

Tablo-3'te oluşturulan mimari ile elde edilen yüz tespitine ait karmaşıklık matrisi verilmiştir. Yüz tanımlama aşamasında tespit edilen yüzler başarılı bir şekilde tanımlanmış ve veri tabanına kayıt edilmiştir.



Şekil 13. Önerilen modelin eğitim, test ve doğruluk grafiği
(The training, loss and accuracy graph of the proposed model)

Elde edilen başarılı yüz ve kişi tespitine ait örnek görüntü Şekil-14'te gösterilmiştir.

Şekil 14. Örnek görüntü içerisinde tespit edilen yüzler
(Faces detected in the sample image frame)

Tablo 3. Önerilen modelin karmaşıklık matrisi
(The confusion matrix of the proposed model)

Karmaşıklık Matrisi		Tahmin	
		Pozitif	*Negatif*
Gerçek	*Pozitif*	4902	39
	Negatif	0	1961

5. SONUÇ VE DEĞERLENDİRME (CONCLUSION AND EVALUATION)

Üretilen video veri boyutlarının her geçen gün artması, çeşitliliğinin ve kaynaklarının artması; video görüntülerinin sınıflandırılmasını, etiketlenmesini, kişi veya zaman bazında kategorize edilmesi gerekliliğini kaçınılmaz hale getirmiştir. Bu bağlamda geliştirilen mimari ile farklı alanlarda oluşan bu ihtiyacın karşılanması, hızlı ve etkin bir çözüm sağlanması amaçlanmıştır. Geliştirilen mimaride ESA tabanlı model ile yüz tespiti, tanımlaması ve zamana göre işaretlenmesi sağlanmıştır. Çalışmada veri kümesinde bulunan etiketlenmiş kişilere ait yüz görüntüleri kullanılmıştır. Yapılan deneysel çalışmalarda farklı çözünürlüğe sahip video dosyaları ilgili video görüntüsünün gerçek zamanından 1,54 ile 2,61 kat arasında daha hızlı şekilde işlenmiş ve tanımlanan kişiler veri tabanına başarılı bir şekilde kayıt edilmiştir. Yaklaşık 30 FPS hıza sahip ham video görüntüleri yüksek performanslı grafik işlemci desteği ile ortalama 78,36 FPS değerlerine ulaşan hızlarda yüz tespiti ve tanımlaması yapılmıştır. Veri kümesinde bulunan etiketlenmiş kişiler uygulanan ESA modeli ile %99,43 oranında doğru tanımlama ile başarılı sonuç vermiştir. Yine aynı modellerin çalışma süreleri önerilen mimari ile kıyaslanmış ve aynı video görüntüsü üzerinde 1,84 ile 5,77 kat arasında değişen hız artışı kaydedilmiştir.

Önerilen ESA tabanlı model, literatürde yer alan MTCNN, OPENCV-CNN, HOG+SVM, SSD-CAFFEMODEL modellerine göre daha yüksek performans göstermiş olup, daha yüksek doğruluk oranına sahiptir. Gelecek çalışmalarda önerilen model ile video görüntüleri içinde kişi sorgulama sistemi oluşturulması hedeflenmektedir.

KAYNAKLAR (REFERENCES)

[1] H.S. Dadi & G. M. Pillutla, "Improved face recognition rate using HOG features and SVM classifier", *IOSR Journal of Electronics and Communication Engineering*, 11(4), 34-44,2018.

[2] O. Yıldız, "Derin öğrenme yöntemleriyle dermoskopi görüntülerinden melanom tespiti: Kapsamlı bir çalışma", Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, 34(4), 2241-2260, 2019.

[3] K. Zhang, Z. Zhang, Z. Li, & Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks", *IEEE Signal Processing Letters*, 23(10), 1499-1503, 2016.

[4] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, & A. C. Berg, "Ssd: Single shot multibox detector", *European conference on computer vision*, 21-37, 2016.

[5] WS. McCulloch and W. Pitts., "A logical calculus of the ideas immanent in nervous activity", *The bulletin of mathematical biophysics*, 5(4), 115–133, 1943.

[6] S. Hochreiter and J. Schmidhuber, "Long short-term memory", *Neural computation*, 9(8), 735–1780, 1997.

[7] S. Lohr , "The age of big data", *New York Times*, 11(2012), 2012.

[8] P. Viola and M. Jones , "Rapid object detection using a boosted cascade of simple features", in *Computer Vision and Pattern Recognition. Proceedings of the 2001 IEEE Computer Society Conference (CVPR)* on, vol. 1, 511–518, 2001.

[9] E. Özbaysar and E. Borandağ , "Vehicle plate tracking system", **in 2018 26th Signal Processing and Communications Applications Conference (SIU)**, IEEE, 1–4, 2018.

[10] X. Luo, R. Shen, J. Hu, J. Deng, L. Hu & Q. Guan, "A deep convolution neural network model for vehicle recognition and face recognition", *Procedia Computer Science*, 107, 715-720, 2017.

[11] K. B. Pranav, & J. Manikandan, "Design and Evaluation of a Real-Time Face Recognition System using Convolutional Neural Networks", *Procedia Computer Science*, 171, 1651-1659, 2020.

[12] Z. Mahmood, N. Muhammad, N. Bibi, & T. Ali, "A review on state-of-the-art face recognition approaches", *Fractals*, 25(02), 2017.

[13] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments", *Technical Report*, 07-49, 2007.

[14] S. Sharma, K. Shanmugasundaram, & S. K. Ramasamy, "FAREC—CNN based efficient face recognition technique using Dlib", **In International Conference on Advanced Communication Control and Computing Technologies (ICACCCT)**, 192-195, IEEE, 2016.

[15] F. Schroff, D. Kalenichenko, & J. Philbin, "Facenet: A unified embedding for face recognition and clustering", **In Proceedings of the IEEE conference on computer vision and pattern recognition**, 815-823, 2015.

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

[17] H. Jiang, & E. Learned-Miller, "Face detection with the faster R-CNN", **12th IEEE international conference on automatic face & gesture recognition**, 650-657, IEEE, 2017.

[18] I. Kalinovskii, & V. Spitsyn, "Compact convolutional neural network cascade for face detection", **In Proceedings of the 10th Annual International Scientific Conference on Parallel Computing Technologies (PCT)**, 1576,375–387, 2016.

[19] K. M. Sagayam, "CNN-based Mask Detection System Using OpenCV and MobileNetV2", **In 2021 3rd International Conference on Signal Prosessin and Communication (ICPSC)**, 115-119, 2021.