Avrupa Bilim ve Teknoloji Dergisi
Sayı 19, S. 923-934, Ağustos 2020

Araştırma Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 19, pp. 923-934, August 2020

Research Article

Sign2Text: Konvolüsyonel Sinir Ağları Kullanarak Türk İşaret Dili Tanıma

Özer Çelik[1*], Alper Odabaş[2]

[1] Eskişehir Osmangazi Üniversitesi, Fen Edebiyat Fakültesi, Matematik - Bilgisayar Bölümü, Eskişehir, Türkiye (ORCID: 0000-0002-4409-3101)
[2] Eskişehir Osmangazi Üniversitesi, Fen Edebiyat Fakültesi, Matematik - Bilgisayar Bölümü, Eskişehir, Türkiye (ORCID: 0000-0002-4361-3056)

(İlk Geliş Tarihi 3 Haziran 2020 ve Kabul Tarihi 31 Ağustos 2020)

(**DOI:** 10.31590/ejosat.747231)

ATIF/REFERENCE: Çelik, Ö. & Odabaş, A. (2020). Sign2Text: Konvolüsyonel Sinir Ağları Kullanarak Türk İşaret Dili Tanıma. Avrupa Bilim ve Teknoloji Dergisi, (19), 923-934.

Öz

İşaret dili, işitme engellilerin kendi aralarında iletişim kurarken, el hareketlerini ve yüz mimiklerini kullanarak oluşturdukları görsel bir dildir. İşitme engelliler kendi aralarında işaret dili yardımıyla rahatlıkla iletişim kurabilmelerine rağmen hastane gibi kamu kurumlarında, hizmet almaya gidenlerin kendilerini ifade etmekte ve karşılarındakileri anlamakta büyük zorluklar çekmektedirler. İşitme engelli okuma yazma oranı düşüktür. Okuma yazması olanların ise Türk İşaret Dili dilbigisinin farklı olması ve dar kelime dağarcığından dolayı okuduklarını anlamada zorluk yaşamaktadır. Dünya sağlık örgütünün raporlarına göre 2018 yılında Avrupa'da 34 milyon işitme engelli bulunmakta, bu sayının 2050 yılına kadar 46 milyon olması beklenmektedir. Video içerisindeki hareketlerin algılanıp işaret diline çevirme işleminde. Bu çalışmada herhangi bir sensör kullanılmadan işitme engelli bireyler tarafından kamerası karşısında yapılan hareketlerin algılanıp işaret diline çevirme işleminde Konvolüsyonel Yapay Ağlar (CNN: Convolutıion Neural Network) ve Uzun Kısa Süreli Bellek (LSTM: Long Short Term Memory) derin öğrenme teknikleri kullanılmıştır. Öncelikle, kamera aracılığıyla elde edilen veri üzerinde baş bölgesinin tespiti ve eğitime uygun hale getirilmesi, ellerin tespiti ve hareketlerinin takip edilmesi ve kırpma gibi video ön işleme adımları uygulanmıştır. Hazırlanan videoların Konvolüsyonel Yapay Ağlar eğitim modeli için frameler ile eğitimi amaçlanmıştır. Veri seti videoların eğitim aşamasında kullanılması için framelere parçalanmıştır. İşaret dili hareketlerinde öncelikli olarak el ve parmak hareketlerinin tahminlemesi gerçekleştirilir. Sadece el hareketleri için eğitim modeli besleneceği için ten renginin bulunduğu kafa bölgesi tespiti çalışması gerçekleştirilmiştir. Kamera karşısında yapılan 10 rakam ve 29 harfin işaret dili hareketleri ile eğitilen CNN + LSTM modellerinde tahminlemesinde %97 başarı oranı elde edilmiştir. Bu sonuçlar, işitme engelli bireylerin kamera karşısında yaptığı hareketlerin algılanıp metne dönüştürmesinde derin öğrenme yöntemlerinin kullanılabileceğini göstermiştir.

Anahtar Kelimeler: Türk İşaret Dili, CNN, LSTM.

Sign2Text: Turkish Sign Language recognition using Convolutional Neural Networks

Abstract

Sign language is a visual language created by the hearing impaired by using hand gestures and facial expressions while communicating among themselves. Although the hearing impaired can easily communicate with each other with the help of sign language, they have great difficulties in expressing themselves and understanding others in public institutions such as hospitals. The literacy rate for the hearing impaired is low. Those who are literate have difficulty in understanding what they read due to the different grammar of Turkish Sign Language and their narrow vocabulary. According to the reports of the World Health Organization, there are 34 million hearing impaired in Europe in 2018, and this number is expected to be 46 million by 2050. In the process of detecting the movements in the video and converting it into sign language. In this study, Convolutional Artificial Networks (CNN: Convolution Neural Network) and Long Short Term Memory (LSTM: Long Short Term Memory) deep learning techniques were used in the

[*] Sorumlu Yazar: Eskişehir Osmangazi Üniversitesi, Fen Edebiyat Fakültesi, Matematik - Bilgisayar Bölümü, Eskişehir, Türkiye, ORCID: 0000-0002-4409-3101, ozer@ogu.edu.tr

process of detecting the movements made by the hearing impaired individuals against their cameras and converting them into sign language without using any sensors. First of all, video pre-processing steps such as determining the head area and making it suitable for training, detecting and tracking the movements of the hands and cropping were applied on the data obtained through the camera. It is aimed to train the videos prepared with frames for the Convolutional Artificial Networks training model. The data set is divided into frames for the use of videos in the training phase. In sign language movements, hand and finger movements are primarily predicted. Since the training model will be fed only for hand movements, the head region where the skin color is found was determined. A 97% success rate was achieved in the estimation of the CNN + LSTM models, which were trained with the sign language movements of 10 numbers and 29 letters made in front of the camera. These results showed that deep learning methods can be used to perceive the camera movements of hearing impaired individuals and convert them into text.

Keywords: Turkish Sign Language, CNN, LSTM.

1. Giriş

İşaret dili, işitme engellilerin kendi aralarında iletişim kurarken, el hareketlerini ve yüz mimiklerini kullanarak oluşturdukları görsel bir dildir. İşitme engelliler kendi aralarında işaret dili yardımıyla rahatlıkla iletişim kurabilmelerine rağmen hastane gibi kamu kurumlarında, hizmet almaya gidenlerin kendilerini ifade etmekte ve karşılarındakileri anlamakta büyük zorluklar çekmektedirler.

Dünya sağlık örgütünün raporlarına göre 2018 yılında Avrupa'da 34 milyon işitme engelli bulunmakta, bu sayının 2050 yılına kadar 46 milyon olması beklenmektedir. İşitme engelliler yakın zamana kadar toplum içerisinde sosyal hayattan izole tutulmuşlardır. Haunaland (2007) ve Murray (2008) çalışmalarında, bir asırdan fazla bir süredir işitme engellilerin bir araya gelerek spor ve eğlence alanlarında uluslararası faaliyetler içinde bulunmuş olmalarına rağmen, farklı topluluklardan gelen işitme engelliler kendi aralarında bile iletişim kurmakta zorluk çektiklerini belirtmektedirler [1, 2]. Ayrıca Gondon (2005) çalışmasında, dünyada 124'den fazla işaret dilinin bulunduğunu ve bunların birbirine benzer taraflarının olmasına rağmen yine karşılıklı iletişim kurmakta problemler yaşadıklarını belirtmektedir [3].

Dünya Engellilik Raporu, genellikle işitme engellilerin işaret dili tercümesine erişiminde sıkıntılar yaşadığını ortaya koymuştur. 93 ülkeyi kapsayan bir araştırmaya göre 31 ülkede tercüme hizmeti bulunmamakta, 30 ülkede ise tercüme hizmeti için yetkili tercüman sayısı 20 veya daha az sayıda olduğu görülmüştür. İşitme engelliler için erişilebilir formatta olan mevcut bilgi yetersizdir ve iletişim ihtiyaçlarının çoğunu karşılamamaktadır.

Marshall vd. (2003), Bungeroth ve Ney (2004), Almohimeed vd. (2011) Arapça ve İngilizce işaret dilleri için metinlerin işaret diline çevirme sistemlerini geliştirmişlerdir. Bu çalışmalarda kural tabanlı ve örnek tabanlı metotlar kullanılmıştır. İşaret dili dil bilgisi kurallarını göz önünde bulundurarak geliştirdikleri sistemlerde metinlerin animasyonlar veya video görüntüleri ile işaret diline çevrilmesi sağlanmıştır [4-6]. Tkashashi ve Kishino (1992), Wang vd. (2006), Shanableh ve Assaleh (2011) çalışmalarında, cihazlar kullanarak işaret dilini metne çevirme üzerine araştırmalarda bulunmuşlardır. Bu cihazların başında Microsoft firmasının üretmiş olduğu Microsoft Kinect cihazı gelmiştir [7-9]. Son yıllardaki çalışmalarda ise Kinect cihazı özelliklerini taşıyan Intel firmasının geliştirmiş olduğu Intel RealSense ve el hareketlerini izleyen daha küçük Leap Motion cihazları da kullanılmıştır.

1.2. Literatür Taraması

İşaret dilinden cümlelerin metne çevirme sistemlerinin geliştirilme aşamasında makine öğrenmesi ve derin öğrenme teknikleri ile birçok çalışma gerçekleştirilmiştir. Starner vd. (1998), Saklı Markov Modeli (HMM) makine öğrenme tekniğini kullanarak, Amerikan İşaret Dili cümlelerini metne çevirme çalışmalarında bulunmuştur [10]. Globel ve Assan (1997) ise %94 lük başarı oranı ile Hollanda İşaret Dili için bir sistem geliştirmiştir. Chai vd. (2013.) ise çalışmasında Çin İşaret Dilinden Çince'ye çevirme üzerinde çalışmalar yapmıştır [12].

Tkashashi ve Kishino (1992), Wang vd. (2006), Shanableh ve Assaleh (2011) çalışmalarında donanımsal cihazlar kullanarak işaret dilinden metne çevirme üzerine çalışmalarda bulunmuşlardır. Donanımsal cihazların başında Microsoft firmasının üretmiş olduğu yaygın kullanılan Microsoft Kinect cihazı gelmiştir. Son yıllardaki çalışmalarda ise Kinect cihazı özelliklerini taşıyan Intel firmasının geliştirmiş olduğu Intel RealSense ve el hareketlerini izleyen daha küçük Leap Motion cihazları da kullanılmıştır.

Haberdar ve Albayrak (2005), Işıkdoğan ve Albayrak (2011), Ketenci vd.(2015) ise Türk işaret dili görüntülerinden Türkçe'ye çevirme sistemleri üzerinden çalışmalarda bulunmuşlardır. Türk İşaret Dili tanıma çalışmalarında makine öğrenme methodlarından Saklı Markov Model (HMM), K-En Yakın Komşu (KNN), Destek Vektör Makineleri (Support Vector Machine- SVM) ve Temel Bileşen Analizi (PCA) ağırlıklı olarak kullanılmıştır. Son yıllardaki çalışmalarda ise derin öğrenme tekniklerin Convolusion Neural Network (CNN) modellerin kullanıldığı gözlemlenmiştir [13-15]. Bu çalışmalardan farklı olarak web kamerasından alınan görüntülerin önce CNN ile tahminlendirilip, daha sonra ise LSTM ile yeni bir model oluşturulmuştur. Böylelikle hareketli olan işaretlerinde doğru tahminlenmesi sağlanmıştır.

2. Materyal ve Metot

Bu çalışmada herhangi bir sensör kullanılmadan, web kamerası karşısında yapılan hareketlerin Konvolüsyonel Yapay Ağlar (Convolutıion Neural Network) ve Uzun Kısa Süreli Bellek (Long Short Term Memory) derin öğrenme teknikleri kullanılmıştır.

2.1. Derin Öğrenme Yöntemi

Konvolüsyonel Yapay Ağlar (CNN: Convolution Neural Networks) ileri beslemeli hayvanların görme merkezinden esinlenerek ortaya çıkan çok katmanlı yapay sinir ağıdır. CNN, görüntüleri girdi olarak içeren problemlerle çalışmak için özel olarak tasarlanmıştır. Facebook ve Google gibi büyük teknoloji şirketler, yüz tanıma ve görsel arama gibi çeşitli amaçlar için çok sayıda konvolüsyonel katmanı olan derin konvolüsyonel sinir ağları kullanırlar. CNN algoritmaları başta görüntü işleme olmak üzere ses ve doğal dil işleme gibi birçok alanda kullanılmaktadır. Konvolüsyonel Yapay Ağlarına (CNN) verilen girişler, her değeri 0 ile 255 arasında değişen bir piksel değerleri dizisidir. Örneğin giriş, 35x35 boyutlarında bir görüntü ise, dizi, 35x35x3 biçiminde 3 boyutlu bir matris oluşturulacaktır. Bir görüntüdeki her piksel, 3 değerle temsil edilir. Bu 3 değer Red Green Blue (RGB), yani kırmızı, yeşil ve mavi yoğunluğunu temsil eder. Görüntü sınıflandırma durumunda, CNN'in işi, bu görüntüyü, yani bir piksel değerleri dizisini, bir girdi olarak almak ve belirli bir sınıfa ait olma ihtimallerini çıkarmaktadır. Tahmin problemlerinde ise, model girdi olarak piksel değerlerini alır ve karşılık gelen çıktı değerlerini tahmin eder. Konvolüsyon Katmanı, CNN'deki ilk katmandır. Bu katman verilen bir girdideki düşük seviyenin yanı sıra yüksek seviye karmaşık özelliklerin tanımlanmasından sorumludur. Konvolüsyonel, ağın geçmişte öğrendiklerine bakarak giriş sinyalini etiketlemeye çalıştığı bir süreçtir. Aktivasyon katmanı, CNN mimarisinin sonuna veya arasına konulan bir katmandır. Aktivasyon katmanı, sinyalin bir katmandan diğerine nasıl aktarıldığını kontrol ederek beynimizdeki nöronların nasıl ateşlendiğini taklit eder. ReLU fonsiyonu, günümüzde sinir ağlarında en yaygın kullanılan aktifleştirme fonksiyonudur. ReLU'nun diğer aktivasyon fonksiyonlarına göre en büyük avantajlarından biri, tüm nöronları aynı anda aktifleşmemesidir. Havuzlama katmanı, bir CNN mimarisindeki Konvolüsyon katmanları arasında görülebilir. Bu katman temel olarak ağdaki uzamsal boyutu giderek azaltır ve aşırı uyumu kontrol ederek ağdaki parametre ve hesaplama miktarını küçültür. Max-havuzlama metodu bir havuzdan yalnızca maksimum değerin alınması esasına dayanır. Girdi boyunca kayan filtrelerin kullanılmasıyla maksimum değerler matrisi oluşturulur. Konvolüsyon katmanından farklı olarak, havuzlama katmanı ağın derinliğini değiştirmez, derinlik boyutunu sabit bırakır [16].

Max-havuzlama sonrası çıktının formülü; n adım sayısı, F filtrenin boyutu, N havuz katmanına giriş boyutu olmak üzere

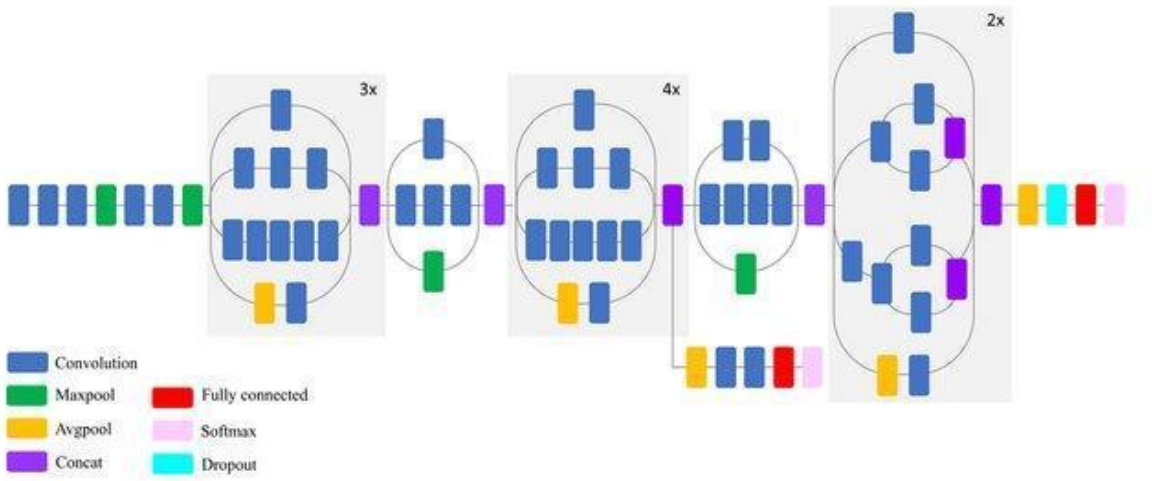
(N - F) / n + 1;

şeklindedir..

Tam bağlı katman, CNN mimarisinin son katmanıdır. Bu katmanda, nöronlar önceki katmanlardan gelen bütün aktivasyonlarla tam bağlantıya sahiptir. Amacı, 3 boyutlu birime giriş yapmak ve N'nin sınıf sayısı olduğu bir N boyutlu vektör çıkarmaktır.

2.1.1. CNN ile transfer öğrenime

Transfer öğrenme, önceden eğitilmiş bir modelin yeni bir problem üzerinde kullanılmasıdır. CNN ağlarını az veriyle eğitilmesini sağladığı için Transfer Öğrenme, Derin Öğrenme alanında çok popülerdir. Dünyadaki sorunların çoğu, bu gibi karmaşık modelleri eğitmek için milyonlarca etiketli veri noktasına sahip değildir. Transfer öğrenmede, önceden eğitilmiş bir Makine Öğrenimi modelinin bilgisi farklı fakat ilgili bir problemi uygulanır. Örneğin, bir görüntünün bir sırt çantası içerip içermediğini tahmin etmek için basit bir sınıflandırıcı çalıştırılırsa, eğitim sırasında kazandığı bilgiyi güneş gözlüğü gibi diğer nesneleri tanımak için kullanılabilir. Transfer öğrenme, çoğunlukla çok fazla miktarda hesaplama gücü gerektiren Bilgisayar Görme ve Doğal Dil İşleme alanlarında kullanılır. Önceden eğitilmiş birçok transfer öğrenme modelleri vardır. ImageNet 22.000 nesne kategorisine ait 1,2 milyon eğitim görüntüsünün bulunduğu bir veri setidir. Model oluşturmak için 100.000 test görüntüsü ve 50.000 doğrulama görüntüsü kullanılmıştır. Bu çalışmada, ImageNet “Büyük Görsel Tanıma Mücadelesi” için eğitilmiş olan InceptionV3 modeli kullanılmıştır [16].

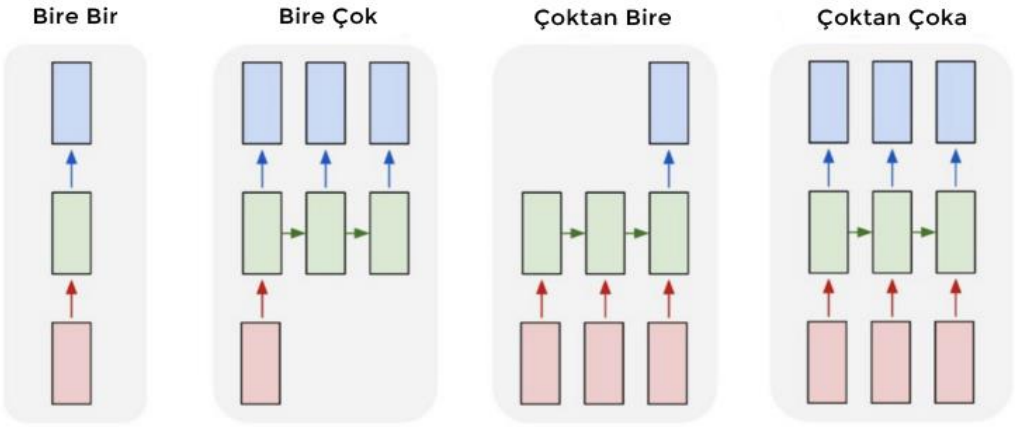


Şekil 1. InceptionV3 modeli mimarisi

InceptionV3 modeli, ağın aynı modülünde 1 × 1, 3 × 3 ve 5 × 5 konvolüsyonlar hesaplanarak uygulanan "çok seviyeli bir özellik çıkarıcı" olarak tasarlanmıştır. Ağ, konvolüsyonlardan elde edilen sonucun kanal boyutu boyunca istifleneceği ve daha sonra ağdaki tabakaya besleneceği şekilde inşa edilmiştir. Şekil 1'de, InceptionV3 modelinin mimarisi gösterilmektedir.

2.1.2. Tekrarlayan Sinir Ağları

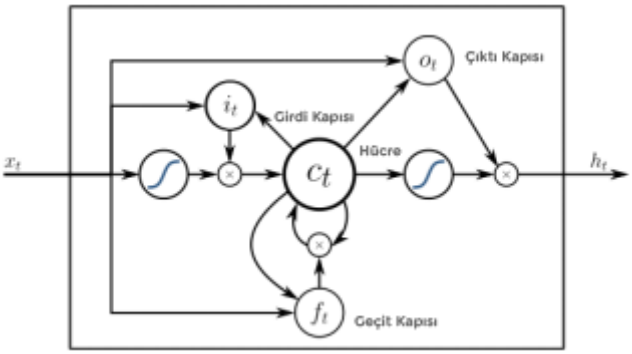
Transfer öğrenme, önceden eğitilmiş bir modelin yeni bir problem üzerinde kullanılmasıdır. CNN ağlarını az veriyle eğitilmesini sağladığı için Transfer Öğrenme, Derin Öğrenme alanında çok popülerdir. Dünyadaki sorunların çoğu, bu gibi karmaşık modelleri eğitmek için milyonlarca etiketli veri noktasına sahip değildir. Transfer öğrenmede, önceden eğitilmiş bir Makine Öğrenimi modelinin bilgisi farklı fakat ilgili bir problemi uygulanır. Örneğin, bir görüntünün bir sırt çantası içerip içermediğini tahmin etmek için basit bir sınıflandırıcı çalıştırılırsa, eğitim sırasında kazandığı bilgiyi güneş gözlüğü gibi diğer nesneleri tanımak için kullanılabilir. Transfer öğrenme, çoğunlukla çok fazla miktarda hesaplama gücü gerektiren Bilgisayar Görme ve Doğal Dil İşleme alanlarında kullanılır. Önceden eğitilmiş birçok transfer öğrenme modelleri vardır. ImageNet 22.000 nesne kategorisine ait 1,2 milyon eğitim görüntüsünün bulunduğu bir veri setidir. Model oluşturmak için 100.000 test görüntüsü ve 50.000 doğrulama görüntüsü kullanılmıştır. Bu çalışmada, ImageNet "Büyük Görsel Tanıma Mücadelesi" için eğitilmiş olan InceptionV3 modeli kullanılmıştır [16].



Şekil 2. RNN Mimarileri

2.1.2.1. Uzun Kısa Süreli Bellek Ünitesi (Long Short -Term Memory Units)

Alman araştırmacılar Hochreiter ve Schmidhuber (1997) tarafından kaybolan gradyan sorununa bir çözüm olarak "Long Short Term Memory Units" LSTM'ler önerilmiştir [17]. LSTM birimleri, tekrarlayan bir sinir ağının (RNN) katmanları için bir yapı birimidir. LSTM birimlerinden oluşan bir RNN'ye genellikle LSTM ağı denir. Ortak bir LSTM birimi bir hücreden, bir giriş geçidinden, bir çıkış geçidinden ve bir unutma geçidinden oluşur. Hücre, rastgele zaman aralıklarındaki değerleri "hatırlamaktan" sorumludur. Dolayısıyla LSTM'de "hafıza" kelimesi vardır. Üç geçitten her biri, çok katmanlı (veya ileriye dönük) bir sinir ağında olduğu gibi "geleneksel" bir yapay nöron olarak düşünülebilir. Yani, ağırlıklı bir toplamın bir aktivasyonunu (bir aktivasyon fonksiyonu kullanarak) hesaplar. Sezgisel olarak, LSTM'nin bağlantılarından geçen değer akışının düzenleyicileri olarak düşünebilen "kapı" ifadesi vardır. Bu kapılar ve hücre arasında çeşitli bağlantılar vardır. Kapılar ve hücre bağlantıları Şekil 3'da ayrıntılı olarak verilmiştir [16].



Şekil 3. Basit bir LSTM yapısı

2.1.3. Derin Öğrenme Metodunda Video Ön İşleme Ve Eğitim

Video içerisindeki hareketlerin algılanıp işaret diline çevirme işleminde derin öğrenmek teknikleri kullanılmıştır. Günümüzde makine öğrenmesi ve derin öğrenme tekniklerinin açık kaynak kodlu Python dili ve R programlama diller popüler olarak kullanılmaktadır. Python windows, macintosh ve linux platformlarında kullanılmaktadır.

TensorFlow, derin öğrenme modellerini uygulamak için farklı işlevler sağlayan Python kütüphanesidir. Tensorflow makine öğrenimi ve derin sinir ağları araştırması yürütmek amacıyla Google'ın Google Brain ekibi tarafından 2015 yılında geliştirilmiştir. Grafikteki düğümler, matematiksel işlemleri temsil ederken, grafik kenarları aralarında iletilen çok boyutlu veri dizilerini (tensörler) temsil eder. Esnek mimari, hesaplamayı tek bir API ile bir masaüstü, sunucu veya mobil cihazdaki bir veya daha fazla CPU'ya veya GPU'ya dağıtmasına imkan verir

Bu çalışmada TensorFlow kütüphanesiyle transfer öğrenme modellerinden InceptionV3 modelini kullanılmıştır. InceptionV3, çok sayıda görüntü türünü ayırt edebilen milyonlarca parametreye sahip devasa bir görüntü sınıflandırma modelidir. Veri setimiz 10 adet rakam ve 29 adet harf videolarından oluşmaktadır. Bunlar dörk farklı kişiden 5 er adet video ile toplam 200 videodan oluşmaktadır. Veri setini oluşturmak için tek el sabit, çift elli ve sadece tek parmak farklılıklarıyla sınıflandırma yapılacak kelimeler seçilmiştir.

Hazırlanan videoların Konvolüsyonel Yapay Ağlar (CNN) eğitim modeli için frameler ile eğitimi amaçlanmıştır. Veri seti videoların eğitim aşamasında kullanılması için framelere parçalanmıştır. İşaret dili hareketlerinde öncelikli olarak el ve parmak hareketlerinin tahminlemesi gerçekleştirilir. Sadece el hareketleri için eğitim modeli besleneceği için ten renginin bulunduğu kafa bölgesi tespiti çalışması gerçekleştirilmiştir. Tüm framelerde tespit edilen kafa bölgesi siyah çerçeve ile kapatılmıştır.

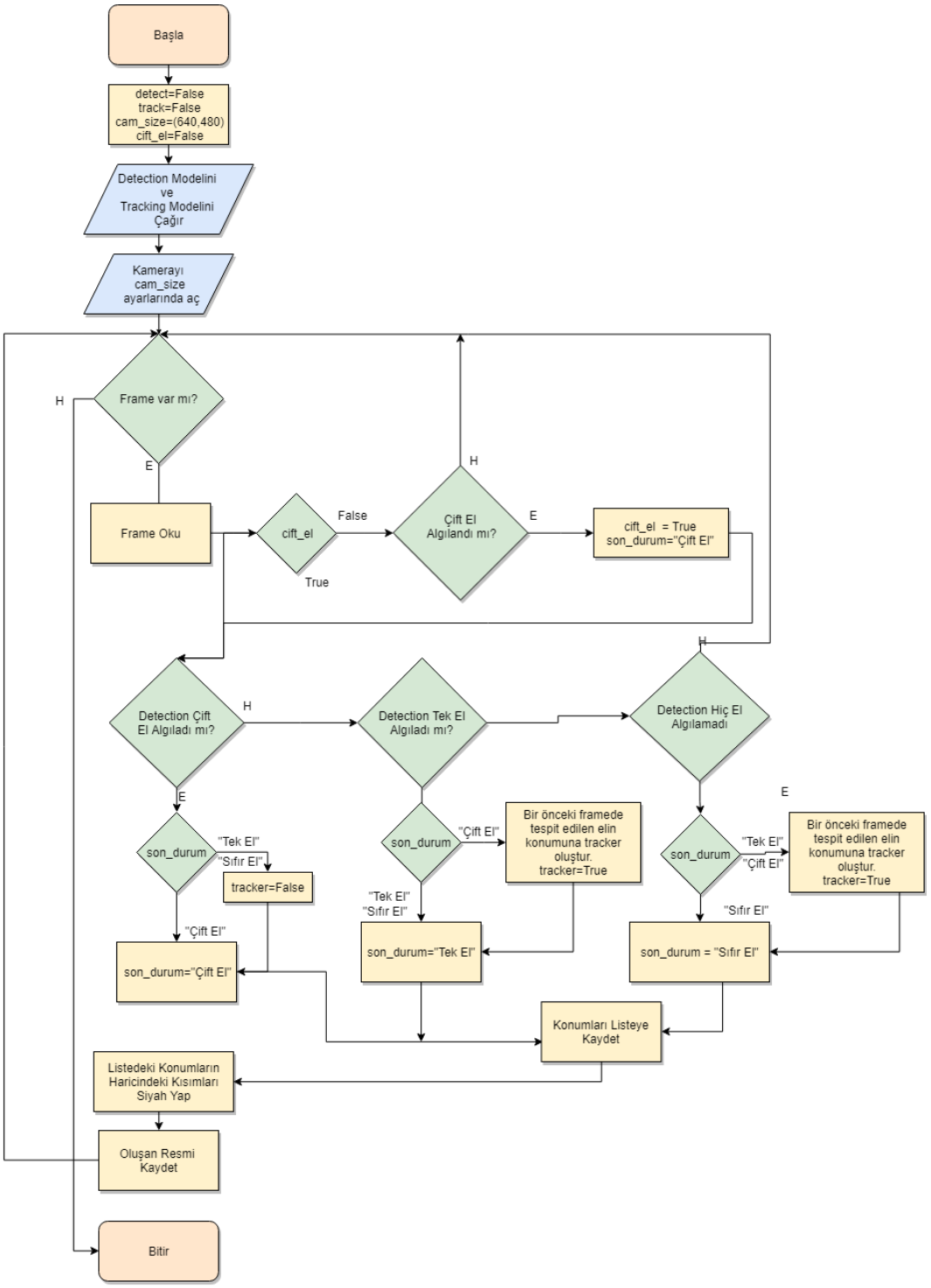
İşaret dilinde el hareketleri önem arz ettiği için eğitimde sadece el bölgesi alınmıştır. El dışındaki alanların işaretlere olumsuz etkisi olacağından bu bölgeler temizlenmiştir. Görüntü yakalamada temel amaç resmin tamamının değil yalnızca işe yarayan kısımlarını alınmasıdır. El algılama için Object Detection metodu kullanılmıştır.

Günümüzde nesne takibi ile çok başarılı çalışmalar yapılmaktadır. Fakat el tespiti nesne takibine benzemek ile beraber işaret dili hareketlerinde el sabit şekilde kalmadığı için farklı şekiller almaktadır. Veri seti için oluşturulan videolar framelere parçalanarak el tespiti veri seti olarak kullanılmıştır. Etiketleme işlemi "LabelImg" aracını kullanarak yapılmıştır.

İşaret dili hareketlerinde farklı şekiller aldığı için el tespiti yapılamamaktadır. Bu problemi çözmek için el tespit ve takibi birlikte kullanılmıştır. Sürekli olarak el tespiti metodu çalıştırılmaktadır. El tespiti modelinin el hareketinin algılanmadığı framelerde veya sadece tek elin tespit edildiği durumlarda el tespiti devreye girmektedir. Bu durumda Tablo 1'deki algoritma işlenmektedir. Şekil 4'de ellerin tespit ve takip algoritması verilmiştir.

Tablo 1. El tespitinden el takibine geçiş süreci

Elin Şuanki Konumu	**Elin Geçmiş Konumu**	**Açıklama**
Çift El	*Tek El*	Çift el algılandığında problem yoktur. Sadece ellerin son lokasyon tutulması yeterli olacaktır.
	Algılama Yok	
Tek El	*Çift El*	O an algılanan el ile El tespit modelinden gelen son lokasyonları kıyaslanıp hangi elin kaybolduğunun belirlenmesi gerekmektedir. Daha sonra belirlenen ele izleyici (tracker) eklenmiştir.
	Algılama Yok	O an algılanan el ile Trakingden gelen son lokasyonları kıyaslanıp hangi elin algılandığını belirlenmesi gerekmektedir. Daha sonra eski trackerları sıfırlayıp yeni belirlenen ele tracker konulmuştur.
Algılama Yok	*Çift El*	Son lokasyonları detection'dan alınıp iki adet yeni tracker oluşturulması gerekmektedir.
	Tek El	Son lokasyonları trackerdan alıp, eski trackerları sıfırlayıp yeni iki adet tracker oluşturulması gerekmektedir.

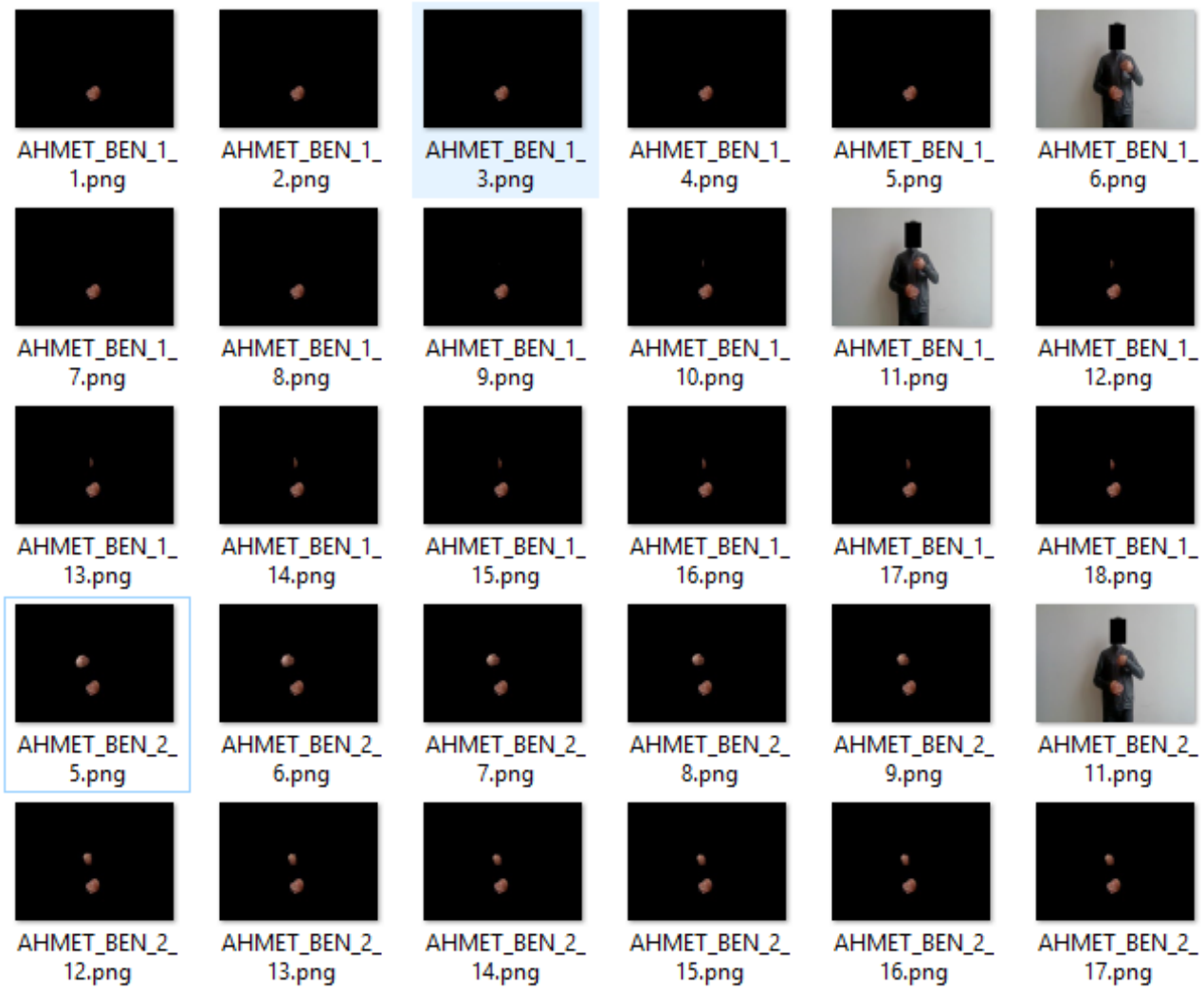


Şekil 4. Ellerin tespiti ve takibi algoritmasının akış diyagramı

Ten tespiti ile kafa, kol ve el bölgelerinin bulunduğu framelerin alınması hedeflenmiştir. Bir önceki adımda kafa bölgesi siyah çerçeve ile kapatıldığı için ten tespiti ile el ve kol bölgeleri dışındaki bütün pikseller silinmiştir. Kişilere göre ten rengi değiştiği için belirlenen alt ve üst renk aralığındaki pikseller seçilmiştir.

Videolardaki ten tespiti için alt aralık (0, 140, 77), üst aralık ise (255, 173, 127) olarak RGB değerleri belirlenmiştir. Bu değerler ortam ışığına göre değişiklik gösterebilir.

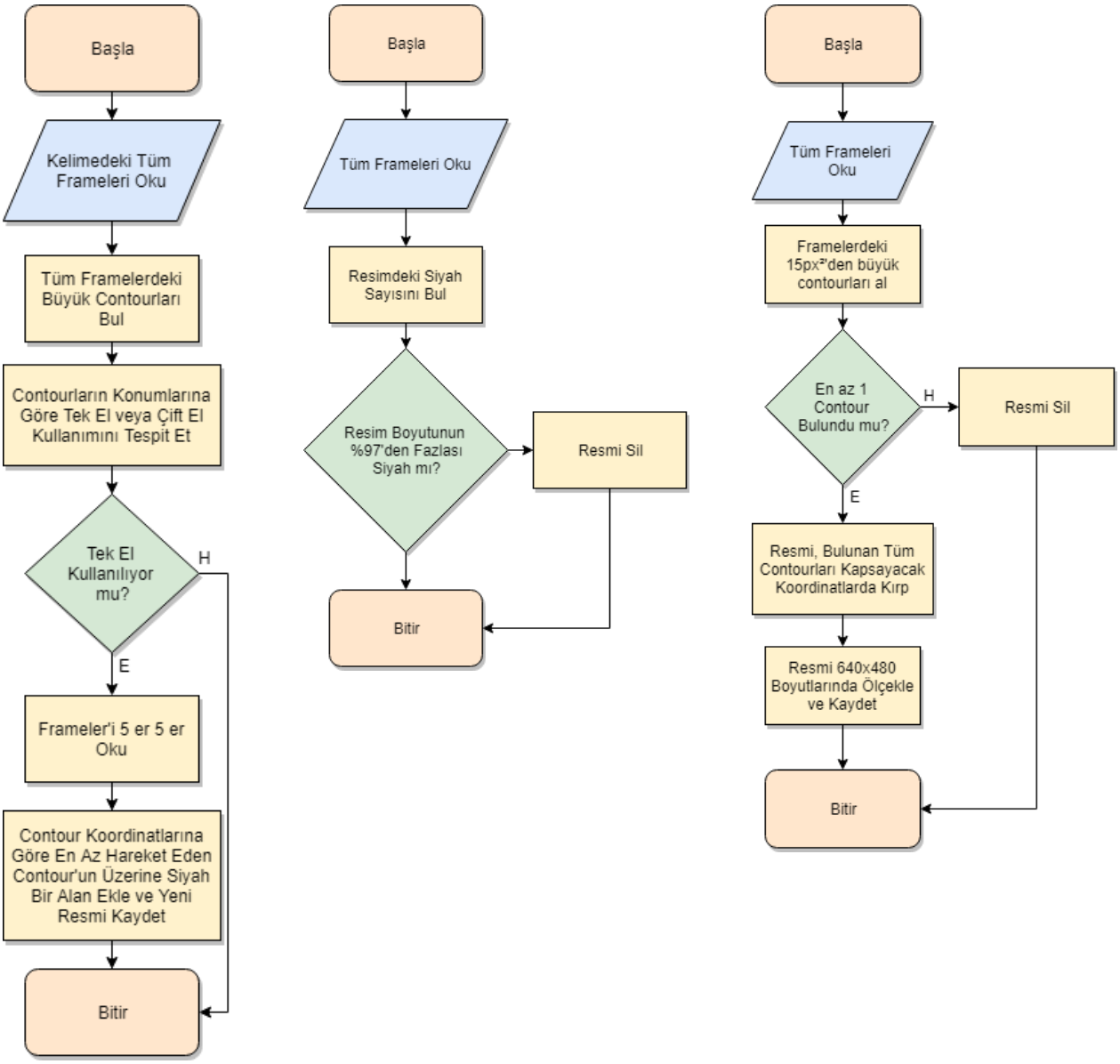
Daha sonra Detection ve tracking algoritmalarında herhangi bir sonuç üretmeyen (tespit edilememiş) frameler devre dışı bırakılmalıdır. Şekil 5'da elin tespit ettiği ve edilemediği frameler yer almaktadır.



Şekil 5. Frame listesi

Elin tespit edilemeyen framelerinin eğitimi olumsuz etkilememesi için Şekil 6'da yer alan algoritmaya göre silme işlemi gerçekleştirilmiştir.

Bu aşamada kelimenin hareketi tek el ile yapılıyorsa kullanılmayan diğer sabit el silinmektedir. Şekil 6.a'da verildiği gibi sabit olan sol el silinmiştir. Şekil 6.b'de ise çift elde hareketli olduğu için herhangi bir silme işlemi gerçekleştirilmemiştir. Bu aşamada gerçekleştirilen algoritmanın akış diyagramı Şekil 6'da yer almaktadır.



(a) Sabit ellerin silinme algoritma şeması

(b) Siyah frame silme algoritma şeması

(c) El boyutu kadar frame kırpılması algoritma şeması

Şekil 6. Kullanılan algoritma şemaları

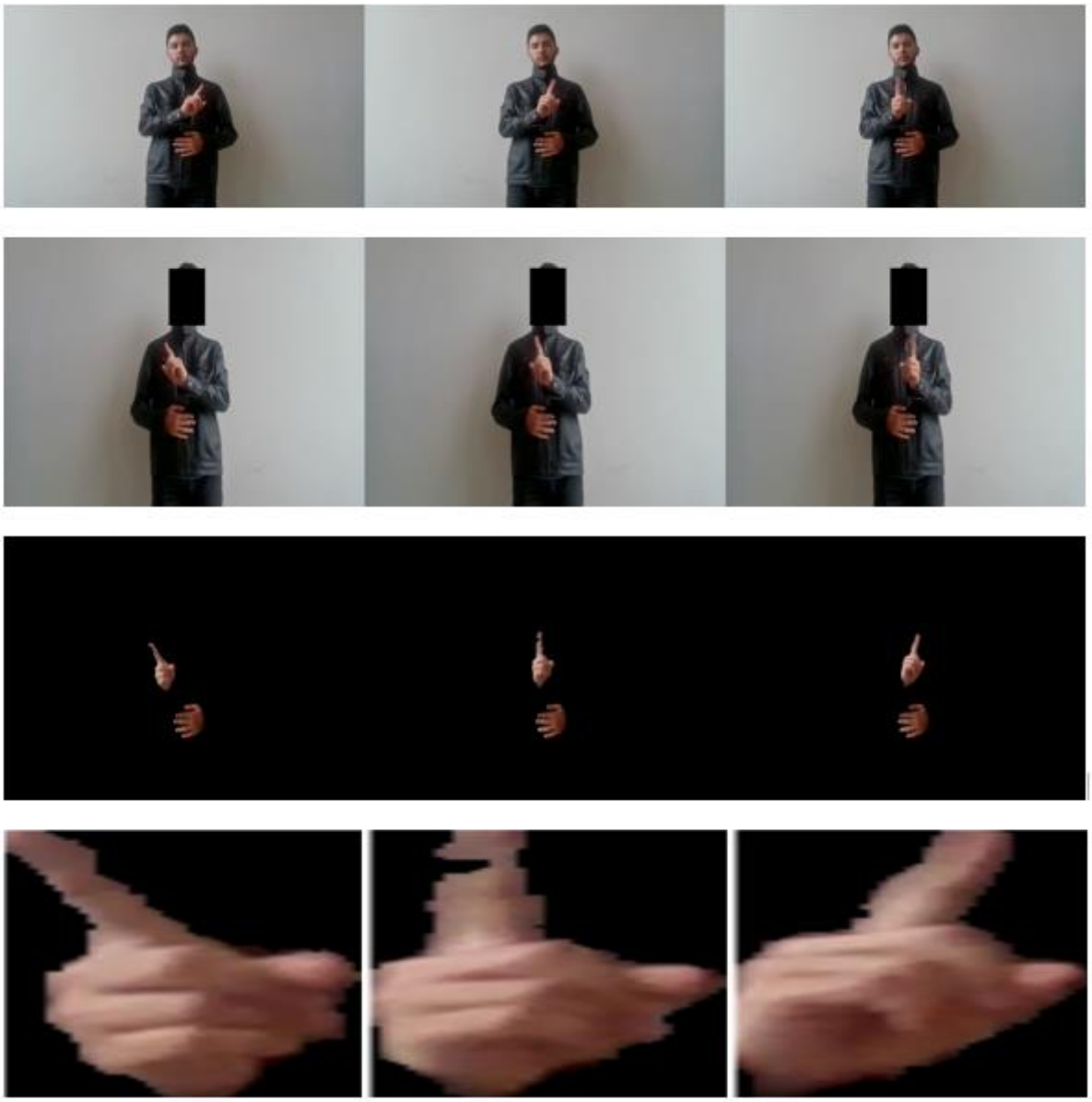
Ellerin birlikte sabit oldukları noktalarda, ilgili framelerin silinmesi için Şekil 6'da verilen algoritma geliştirilmiştir. Veri setinde ön planda olması istenilen kısım el olduğundan, büyük bir siyahlık ortasında küçük bir el kullanmak yerine el bölgesi Şekil 6.c'de yer alan algoritmaya göre kırpılıp tekrar boyutlandırılmıştır. CNN algoritmasında en boy değerlerinin belli bir oranda sabit tutulması gerektiğinden, tüm resimler 640x480 olarak boyutlandırılır. Şekil 7'de tek ve çift el hareketli iken işaretlerin görünümleri yer almaktadır.

(a) Tek el hareketli görüntünü kırpılmış görünümü

(b) Çift el hareketli görüntünü kırpılmış görünümü

Şekil 7. (a) ve (b) Sabit el boyutuna uygun el kırpılmış görüntüsü

CNN için hazırlanan framelerin eğitim sonrası softmax sonuçlarının bir sonraki LSTM modelinde giriş değerleri olacağından, LSTM ağı için eşit boyutta veri ile beslenmesi gerekmektedir. İşaret dili hareketi kısa olan videoların son frameleri çoğaltılarak eşitlendirilir. Bütün ön işleme süreçlerini aşağıdaki Şekil 8'te yer almaktadır.

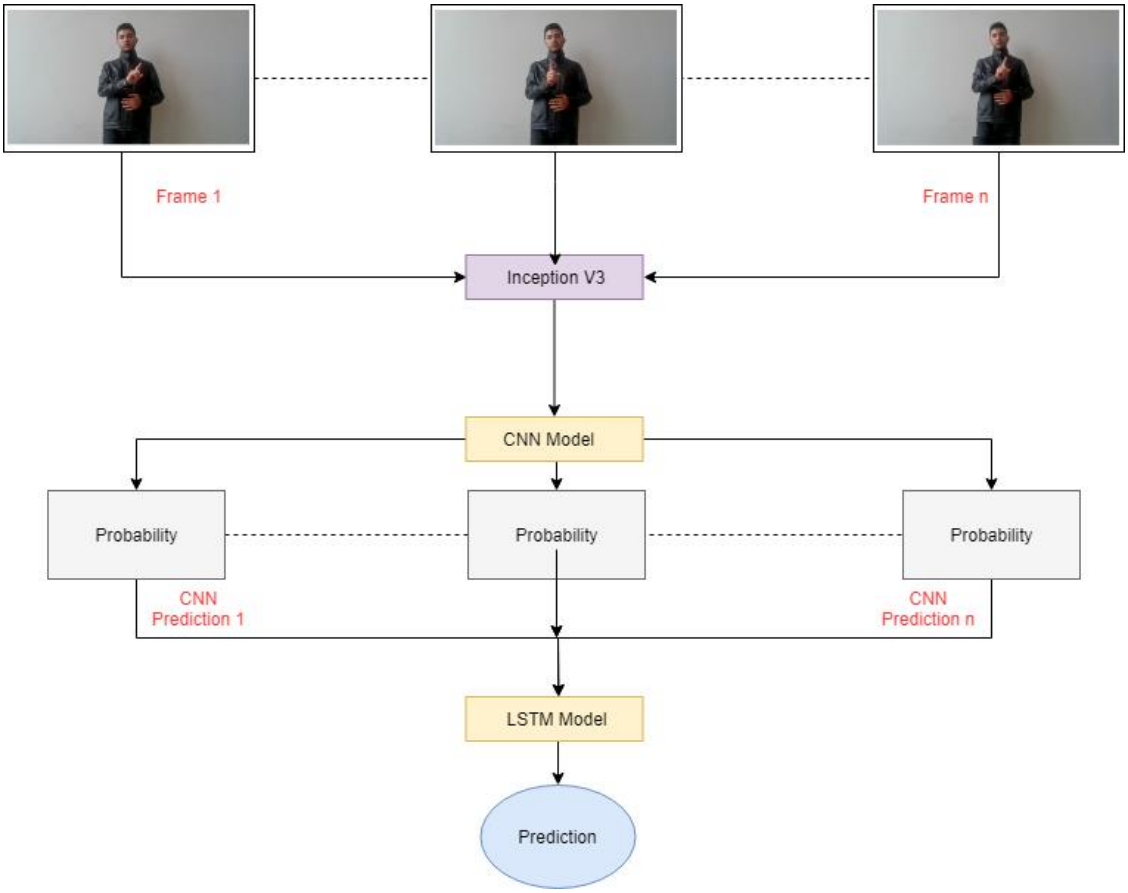


Şekil 8. Hareketli el bölgesinin kırpılması ve boyutlandırması sonrası frameler

Video ön işleme aşamalarından elde edilen frameler ile CNN InceptionV3 modeli ile eğitim gerçekleştirilmiştir. Softmax katmanına veri setindeki kelimeler verilmiştir. Eğitim aşamasından sonra oluşan modele, her bir frame tahminlemesi yapıtırılmaktadır.

Sabit işaretler için bu tahminleme başarılı sonuçlar vermesine rağmen, hareketli işaretlerde hareketin tamamı işareti ifade ettiği için maalesef iyi sonuçlar elde edilememektedir. Ön işlemeden gelen veriler CNN modeli ile tahminlendirip, gelen tahminler ise Şekil

9'da görüldüğü LSTM ile eğitilerek model oluşturulmuştur. Böylelikle hareketli işaretlerin bütün sıralı frameler ile tahminleme gerçekleştirilmiştir.



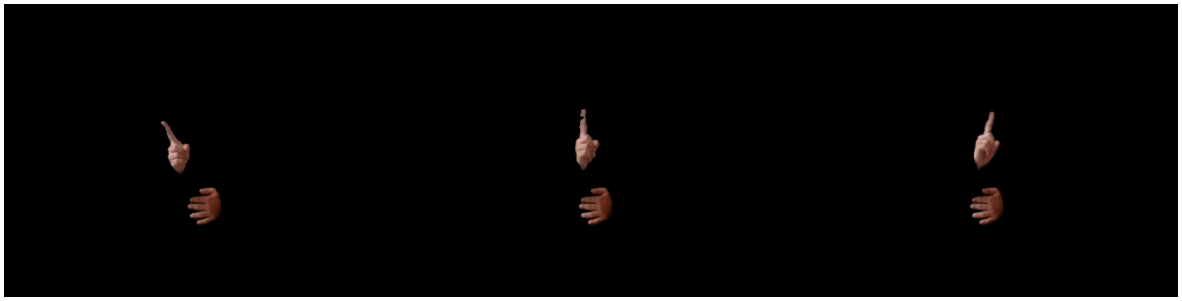
Şekil 9. Derin öğrenme yöntemi çalışma modeli

Bu metodda Lenova Yoga 720 diz üstü bilgisayar kamerası kullanılarak veri seti oluşturulmuş ve testler gerçekleştirilmiştir. Konuyla ilgili detaylı bilgilere yazarın doktora tezinden ulaşılabilir [16].

3. Araştırma Bulguları

Video ön işlem aşamasında gelinen son nokta için farklı eğitim ve testler yapılmıştır. Bu testlerin sonuçları sırasıyla aşağıda yer almaktadır.

İlk test videodan çıkartılan framelerin el detection ve tracking sonucunda elde edilen görüntülerinin eğitim ve tahminlemesi üzerine yapılmıştır. Videodan elde edilen görüntü Şekil 10'da verilmiştir. Tablo 2'de eğitim bilgilerine yer verilmiştir.



Şekil 10. Eğitime verilen frame görünümü

Tablo 2: İlk test eğitim bilgileri

Görüntü Özelliği	Sadece ellerin bulunduğu orijinal boyut
Eğitim adım sayısı	4.000
Toplam frame sayısı	41.790
Eğitimde kullanılan kelime sayısı	39
Doğru tahmin sayısı	21

Yanlış tahmin sayısı	18
Test Başarı Oranı	%53.8

Kelimelerin doğruluklarının tahminlemesi Tablo 3'te verilmiştir.

Tablo 3: İlk test tahminleme sonuçları

0	1	2	3	4	5	6	7	8	9
✓	✓	✗	✓	✗	✓	✗	✗	✗	✗

A	B	C	Ç	D	E	F	G	Ğ	H	I	İ	J	K	L	M	N	O	Ö	P	R	S	Ş	T	U	Ü	V	Y	Z
✓	✓	✗	✗	✗	✗	✓	✗	✓	✗	✗	✓	✓	✓	✓	✓	✗	✗	✓	✓	✓	✓	✗	✗	✓	✓	✗	✓	✓

İkinci test çalışmasında ise veri setindeki kelimelerin el ve parmak hareketleri belirgin olmasından dolayı el bölgesi kesilip 640x480 oranında büyültülerek eğitime alınmıştır. Ayrıca sabit elli olan işaretlerde ise sabit el de temizlenmiştir. Şekil 11'te eğitime verilen frame yer almaktadır.

Şekil 11. İkinci test için eğitime verilen frame

Eğitim özellikleri test sonuçları Tablo 4'te verilmiştir. Sadece hareketli elin büyültülerek eğitime verilmesiyle başarı oranı % 91'e yükselmiştir. 39 kelime'den sadece 3 kelime tahminlemesinde hata tespit edilmiştir. 0 rakamı işareti yerine O harfini, I harfi yerine 1 rakamını, O harfi yerine de 0 rakamı olarak yanlış tahminlenmiştir. Yanlış tahminlenen harf ve rakamların işaretleri birbirine çok benzemektedir.

Tablo 4: İkinci eğitim test sonuçları

0	1	2	3	4	5	6	7	8	9
✗	✓	✓	✓	✓	✓	✓	✓	✓	✓

A	B	C	Ç	D	E	F	G	Ğ	H	I	İ	J	K	L	M	N	O	Ö	P	R	S	Ş	T	U	Ü	V	Y	Z
✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✗	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

4. Sonuç

İşaret dilinden cümlelerin metne çevirme sistemlerinin geliştirilme aşamasında makine öğrenmesi ve derin öğrenme teknikleri ile birçok çalışma gerçekleştirilmiştir. Starner vd., Saklı Markov Modeli (HMM) makine öğrenme tekniğini kullanarak, Amerikan İşaret Dili cümlelerini metne çevirme çalışmalarında bulunmuştur [10]. Grobel ve Assan ise %94 lük başarı oranı ile Hollanda İşaret Dili için bir sistem geliştirmiştir [11]. Chai vd. ise çalışmasında Çin İşaret Dilinden Çince'ye çevirme üzerinde çalışmalar yapmıştır [12]. Tkashashi ve Kishino, Wang vd., Shanableh ve Assaleh (2011) çalışmalarında donanımsal cihazlar kullanarak işaret dilinden metne çevirme üzerine çalışmalarda bulunmuşlardır. Donanımsal cihazların başında Microsoft firmasının üretmiş olduğu yaygın kullanılan Microsoft Kinect cihazı gelmiştir. Son yıllardaki çalışmalarda ise Kinect cihazı özelliklerini taşıyan Intel firmasının geliştirmiş olduğu Intel RealSense ve el hareketlerini izleyen daha küçük Leap Motion cihazları da kullanılmıştır. Haberdar ve Albayrak, Işıkdoğan ve Albayrak, Ketenci vd. ise Türk işaret dili görüntülerinden Türkçe'ye çevirme sistemleri üzerinden çalışmalarda bulunmuşlardır. Türk İşaret Dili tanıma çalışmalarında makine öğrenme metotlarından Saklı Markov Model, KNN, SVM ve PCA ağırlık olarak kullanılmıştır. Son yıllardaki çalışmalarda ise derin öğrenme tekniklerin CNN modellerin kullanıldığı gözlemlenmiştir [13-15]. Bu çalışmalardan farklı olarak web kamerasından alınan görüntülerin önce CNN ile tahminlendirilip, daha sonra ise LSTM ile yeni bir model oluşturulmuştur. Böylelikle hareketli olan işaretlerinde doğru tahminlenmesi sağlanmıştır.

İşitme engelli okuma yazma oranı düşüktür. Okuma yazması olanların ise Türk İşaret Dili dilbigisinin farklı olması ve dar kelime dağarcığından dolayı okuduklarını anlamada zorluk yaşamaktadır. Bu çalışmada işaret dilinden metne çevirme işlemi için web kamerası ile derin öğrenme methodları üzerinde çalışılmıştır. El ve parmak hareketlerinin belirgin olarak kelimenin işaretini gösteren alfabede iyi sonuçlar alınmıştır. Sadece CNN ile videodan gelen tek frame'in tahminlemesi sabit hareketli işaretler için başarılı olmasına karşı, hareketli işaretlerde ise maalesef iyi sonuçlar alınamamıştır. Testin başarı oran %53.8 olmuştur. Hareketli işaretlerde işareti birden fazla frame'deki görüntü ifade ettiği için CNN ile sadece tek frame ile sonuç alınmaktadır. Bu problemi gidermek için CNN ile her bir frame tahminlenerek, sonuçlar sıralı olarak LSTM modeli ile tahminlenmiştir. Böylelikle hareketli işaretlerin tahminlemesindeki sorun giderilmiştir. Son resim ön işleme adımları gerçekleştirilip CNN + LSTM modellerinde tahminleme başarı oranı %97 elde edilmiştir. Bu şekilde, işitme engelli bireylerin kamera karşısında yaptığı hareketleri algılayıp metne dönüştürme çalışması tamamlanmıştır. İşaret dilinde yaklaşık 4000 kelime bulunmaktadır. Bütün kelimeler için oluşturulacak eğitim modellerinde benzer hareketler çok olacağı için başarı oranları düşecektir. İleriki çalışmalarda daha fazla veri seti ile sonuçların alınması hedeflenmektedir.

Kaynakça

Kaynaklar yazılırken APA formatında yazılmasına ve bir sonraki sayfadan başlamasına dikkat ediniz.

[1] Haualand H., 2007, The two week village, The significance of sacred occasions for the deaf community, In Benedicte Ingstad & Ssuan R., Whyte, ed., Disability in local and global worlds, 33-55, Berkeley: University of California Press.Joachims, T. (1999, June). Transductive inference for text classification using support vector machines. In Icml (Vol. 99, pp. 200-209).

[2] Murray J. J., 2008, Coequality and transnational studies: understanding deaf lives, In H.-D. L. Bauman (ed.) Open your eyes, Deaf studies talking, 100-110.. London: University of Minnesota Press.

[3] Gordon R. G., Jr. ed., 2005, Ethnologue: Languages of the World, Fifteenth edition, Dallas TX: SIL International.

[4] I. Marshall É. S., A prototype text to British Sign Language (BSL) translation system, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 2, Sapporo, Japan: Association for Computational Linguistics, 2003, pp. 113-116. [6].

[5] Bungeroth J., Ney H., Statistical sign language translation, In Proc. of the Workshop on Representation and Processing of Sign Languages (LREC2004), pages 105–108, Lisbon, Portugal, 2004.

[6] Almohimeed A., Wald M., Damper R. I., 2011, Arabic Text to Arabic Sign Language Translation System for the Deaf and HearingImpaired Community, In, EMNLP 2011: The Second Workshop on Speech and Language Processing for Assistive Technologies (SLPAT), Edinburgh, UK, Scotland, pp. 101-109.

[7] Takahashi T., Kishino F., 1992, A hand gesture recognition method and its application, Systems and Computers in Japan, 23 (3), 38-48.

[8] Wang H., Leu M., Oz C., C., 2006, American Sign Language recognition using multidimensional Hidden Markov Models, Journal of Information Science and Engineering, 22 (5), 1109-1123.

[9] Shanableh T., Assaleh K., 2011, User-independent recognition of Arabic sign language for facilitating communication with the deaf community, Digital Signal Processing: A Review Journal, 21 (4), 535-542.

[10] Starner T., Weaver J., Pentland A., "Real-time American Sign Language Recognition using Desk and Wearable Computer based Video", IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, pp. 1371–1375, 1998.

[11] Grobel K., Assan M., "Isolated Sign Language Recognition using Hidden 65 Markov Models", IEEE International Conference on Computational Cybernetics and Simulation, Systems, Man, and Cybernetics, pp. 162–167, 1997.

[12] Chai X., Li G., Chen X., Zhou M., Wu G., Li H., "VisualComm: A Tool to Support Communication Between Deaf and Hearing Persons with the Kinect", 15th International ACM SIGACCESS Conference on Computers and Accessibility, p. 76, 2013.

[13] Haberdar H., 2005, Saklı Markov Modelleri Kullanılarak Görüntüden Gerçek Zamanlı Türk İşaret Dili Tanıma Sistemi, Bilgisayar Mühendisliği, Yüksek Lisans Tezi, Yıldız Teknik Üniversitesi.N. El-Makky et al., Sentiment analysis of colloquial Arabic tweets, 2015.

[14] Işikdoğan F., Albayrak S., 2011, June, Automatic recognition of Turkish fingerspelling, In Innovations in Intelligent Systems and Applications (INISTA), 2011 International Symposium on (pp. 264-267), IEEE.

[15] Ketenci S., Kayikçioğlu T., Gangal A., 2015, May, Recognition of sign language numbers via electromyography signals, In Signal Processing and Communications Applications Conference (SIU), 2015 23th (pp. 2593-2596), IEEE.

[16] Celik O, 2019, An artificial intelligence based remote communication system for hearing impaired. Ph.D. thesis, Eskisehir Osmangazi University.

[17] Hochreiter S., Schmidhuber J., 1997, Long short-term memory. Neural computation, 9(8), 1735-1780.