T.C.
BATMAN ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

GÖRÜNTÜYE DAYALI DUDAK OKUMA UYGULAMALARINDA UZAMSAL DUDAK NOKTALARI TEMELLİ YENİ ÖZNİTELİK YAKLAŞIMLARI

Hamdullah TUNG

YÜKSEK LİSANS

Elektrik-Elektronik Mühendisliği Anabilim Dalı

Şubat 2021
BATMAN

TEZ KABUL VE ONAYI

Hamdullah TUNG tarafından hazırlanan "GÖRÜNTÜYE DAYALI DUDAK OKUMA UYGULAMALARINDA UZAMSAL DUDAK NOKTALARI TEMELLİ YENİ ÖZNİTELİK YAKLAŞIMLARI" adlı tez çalışması 15/02/2021 tarihinde aşağıdaki jüri tarafından oy birliği ile Batman Üniversitesi Fen Bilimleri Enstitüsü Elektrik-Elektronik Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS TEZİ olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Doç. Dr. Yılmaz KAYA	……..
Danışman Dr. Öğr. Üyesi Ramazan TEKİN	……..
Üye Dr. Öğr. Üyesi Emrullah ACAR	……..

Yukarıdaki sonucu onaylarım.

Prof. Dr. Şahnaz TİĞREK
Fen Bilimleri Enstitüsü Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

İmza

Hamdullah TUNG

Tarih:

ÖZET

YÜKSEK LİSANS

GÖRÜNTÜYE DAYALI DUDAK OKUMA UYGULAMALARINDA UZAMSAL DUDAK NOKTALARI TEMELLİ YENİ ÖZNİTELİK YAKLAŞIMLARI

Hamdullah TUNG

Batman Üniversitesi
Fen Bilimleri Enstitüsü
Elektrik-Elektronik Mühendisliği Anabilim Dalı

Tez Danışmanı: Dr. Öğr. Üyesi Ramazan TEKİN

2021, 69 Sayfa

Jüri
Doç. Dr. Yılmaz KAYA
Dr. Öğr. Üyesi Ramazan TEKİN
Dr. Öğr. Üyesi Emrullah ACAR

Sosyal bir varlık olan insan, ihtiyaçlarını gidermek için çoğu zaman konuşarak insanlarla iletişime geçmektedir. Konuşma eylemi hem görme ve hem de duyma duyularının ortak kullanımı sonucu gerçekleşmektedir. Konuşmada esnasında sesler üretilirken dudağın aldığı formalar gözle açık bir şekilde izlenebilir. Dudak okuma, sesin duyulmadığı ya da bozuk olduğu durumlarda konuşmayı dudak, yüz ve dilin hareketini çözümleyerek anlama tekniğidir.

Görsel konuşma bilgileri, özellikle ses bozuk veya erişilemez olduğunda, otomatik dudak okumada önemli bir rol oynamaktadır. Ses-görüntü tabanlı dudak okumanın başarısına rağmen, sadece-görüntü tabanlı dudak okumada birbirine benzer dudak hareketlerine sahip sesleri ayırmadaki zorluklardan dolayı oldukça güç bir problemdir. Bu çalışmada, sadece-görsel tabanlı dudak okuma uygulamalarında başarı oranını arttırmak amacıyla birtakım yeni öznitelik yaklaşımları sunulmuştur. Bu çalışmada, konuşmacı-bağımsız ve konuşmacı-bağımlı gerçekleştirilen tahmin uygulamalarında iki ayrı veri seti kullanılmıştır. Bu veri setleri; Latin alfabesindeki 26 harfin beş (5) konuşmacı tarafından yedi (7) kez tekrarlandığı AVLetters2 ve 0-9 arasındaki 10 rakamın altı (6) konuşmacı tarafından dokuz (9) kez

ÖZET (devam ediyor)

tekrarlandığı AVDigits'dir. Öncelikle yüzdeki öğeler ve dudaklar ayrılarak, dudak sınırlarını 20 noktayla işaretlenmiştir. Daha sonra bu uzamsal noktalara dayalı, Merkezi-Öklid-Uzaklık (MÖU), Simetrik-Öklid-Uzaklık (SÖU) ve Komşu-İşaret-Açıları (KİA) isimli öznitelik yaklaşımlarıyla elde edilen özellikler sınıflandırıcılara uygulanmıştır. Son olarak, K-en Yakın Komşu algoritması, Rasgele Orman, Destek Vektör Makinesi isimli sınıflandırma algoritmaları kullanılarak video görüntülerden dudak okuma analizi yapılarak 26 karakter ve 10 rakam tespit edilmeye çalışılmıştır. Yapılan analizler sonucunda en iyi başarı sonuçları AVLetters2 veri seti için RO-MÖU yöntemiyle %45,934 ve AVDigits veri seti için KNN-MÖU yöntemiyle %67,407 olarak bulunmuştur. Bu veri setleri üzerinde sadece-görüntü temelli yapılan diğer çalışmalarla karşılaştırıldığında oldukça yüksek ve başarılı sonuçlar elde edildiği görülmüştür.

Anahtar Kelimeler: Dudak Okuma, Görüntü İşleme, Öznitelik Çıkarma, Uzamsal Öznitelikler

ABSTRACT

MS THESIS

NEW FEATURE APPROACHES BASED ON SPATIAL LIP POINTS IN VISUAL-BASED LIP READING APPLICATIONS

Hamdullah TUNG

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF BATMAN UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE ELECTRİCAL ELECTRONİCS ENGINEERING

Advisor: Assist. Prof. Dr. Ramazan TEKİN

2021, 69 Pages

Jury
Assoc. Prof. Dr. Yılmaz KAYA
Assist. Prof. Dr. Ramazan TEKİN
Assist. Prof. Dr. Emrullah ACAR

As a social being, human beings often communicate with people by talking in order to meet their needs. The act of speaking takes place as a result of the joint use of both sight and hearing. While the sounds are produced during the speech, the forms of the lip can be clearly observed. Lip reading is the technique of understanding speech by analyzing the movement of the lips, face and tongue in cases where the voice is not heard or distorted.

Visual speech information plays an important role in automatic lip reading, especially when the sound is distorted or inaccessible. Despite the success of audio-image-based lip reading, visual-only lip reading is a very difficult problem due to difficulties in distinguishing sounds with similar lip movements. In this study, some new attribute approaches are presented in order to increase the success rate in visual-only lip reading applications. In this study, two separate data sets were used in speaker-independent and

ABSTRACT (continued)

speaker-dependent prediction applications. These data sets; AVLetters2, in which 26 letters in the Latin alphabet are repeated seven (7) times by five (5) speakers, and AVDigits, in which the 10 digits 0-9 are repeated nine (9) times by six (6) speakers. First of all, the facial elements and lips are separated and the lip borders are marked with 20 points. Later, the attribute approaches based on these spatial points, named Center-Euclidean-Distance (CED), Symmetric-Euclidean-Distance (SED) and Neighbor-Points-Angles (NPA), are applied to classifiers. Finally, using the classification algorithms named K-Nearest Neighbor algorithm (KNN), Random Forest (RF), Support Vector Machine (SVM), lip reading analysis was performed from video images to determine 26 characters and 10 numbers. As a result of the analysis, the best success results were found to be 45.934% for the AVLetters2 data set with the RF-CED method and 67.407% for the AVDigits data set using the KNN-CED method. When compared to other visual-only studies on these data sets, it was seen that quite high and successful results were obtained.

Keywords: Lip Reading, Image Processing, Feature Extraction, Spatial Features

ÖNSÖZ

Zorlu bir yüksek lisans eğitimi neticesinde danışman hocam Dr. Öğrt. Üyesi Ramazan TEKİN önderliğinde, büyük bir emek sonucu bu çalışmayı gerçekleştirdik. Literatüre güzel bir katkı sağlayacağını düşündüğüm bu çalışmamda öncelikle emeği geçen Dr. Öğrt. Üyesi Ramazan TEKİN'e, literatür araştırmasında yardımlarını esirgemeyen Araştırma Görevlisi Levent YAVUZ'a ve desteğini hiçbir zaman esirgemeyen Herdem TUNÇ'a sonsuz teşekkürlerimi sunarım. Ayrıca çalışmamda desteklerini hiçbir zaman eksik etmeyen, bana olan güvenlerini hiç kaybetmeyen aileme teşekkür ederim.

Hamdullah TUNG
BATMAN-2021

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR

ADD : Ayrık Dalgacık Dönüşümü

AGM : Aktif Görünüm Modeli

AŞM : Aktif Şekil Modelini

Ç-SGS : Çekirdek Seyrek Gösterim Sınıflandırıcısı

DVM : Destek Vektör Makineleri

GİKT : Görsel-İşitsel Konuşma Tanıma

GMM : Gizli Markov Modelleri

GLCM : Gri Seviye Eş Oluşum Matrisi

HSV : Hue, Saturation, Value

KİA : Komşu İşaret Açıları

KNN : K-En Yakın Komşu

LSDA : Uyarlama Yoluyla Büyük Ölçekli Algılama

LSTM : Uzun Vadeli Hafıza Ağları

MÖU : Merkezi Öklid-Uzaklık

RAT : Regresyon Ağaçları Topluluğu

RO : Rastgele Orman

ROI : Region Of Interest

ROMH : Rastgele Orman Manifold Hizalamasını

SÖU : Simetrik Öklid-Uzaklık

TBA : Temel Bileşen Analizi

ÜOD : Üç Ortogonal Düzlem

YGH : Yönlendirilmiş Gradyanların Histogramı

YİÖ : Yerel İkili Örüntüler

1. GİRİŞ

Konuşmak, hayatımızın büyük bir kısmında yer almaktadır. Bir insan ortalama olarak bir (1) yaşında konuşmaya başlayıp bunu ömrünün sonuna kadar sürdürmektedir. Sosyal bir varlık olan insan, ihtiyaçlarını gidermek için çoğu zaman insanlarla iletişime geçmektedir. Bu iletişim konuşma temelli gerçekleşmektedir. Günlük hayatımızın çoğunda bir şeyler anlatmakta veya bir şeyler dinlemekteyiz. İnsanlar gün içerisinde sürekli iletişim halindedir ve bizler konuşarak yeni bilgiler öğrenmekte, olayları anlamakta ve bilişsel gelişimimizi gerçekleştirmekteyiz. Tüm bunlar konuşma eyleminin insanı insan yapan en belirleyici faktör olduğunu göstermektedir (Kurudayıoğlu, 2003).

Konuşma eylemi basit bir seslendirme işlemi değildir. 17. Yüzyıldan beri konuşmacının yüz hareketlerinde, konuşma hakkında yararlı bilgiler olduğu belgelenmiştir (Matthews et al., 2002). Konuşma eylemi ses ve görüntü ile gerçekleşmektedir. Konuşmada, kullanılan seslerin üretimi esnasında ağzın aldığı şekiller gözle izlenebilir özelliktedirler. Bu da dudak okumayı ortaya çıkarmaktadır. Dudak okuma, sesin anlaşılması mümkün olmadığı durumlarda konuşmayı dudak, yüz ve dilin hareketini çözümleyerek anlama tekniğidir. Konuşma esnasında konuşmacının yüzünü görmek, özellikle sesin anlaşılmasının zor olduğu gürültülü ortamlarda anlaşılırlığı arttırmaktadır.

Dolayısıyla sesin erişilemez ya da bozuk olduğu veya çeşitli işitsel rahatsızlıklar nedeniyle duyu kayıpları bulunan insanlarda algılama işlevlerinin zayıfladığı durumlarda yalnız görsel veriler kullanarak dudak okuma oldukça önemli bir problem olarak ortaya çıkmaktadır. Ancak, konuşma esnasında ağzın aldığı şekilleri ayırt edebilmek güç olduğundan sadece-görüntü temelli dudak okumada yüksek başarım elde etmek oldukça zordur. Örneğin, P ve B harflerini seslendirirken ağız, diş ve çenenin aldığı pozisyon hemen hemen aynıdır. Bu nedenle ses tabanlı yapılan Otomatik Konuşma Tanıma (Automatic Speech Recognition) çalışmalarında yüksek başarı elde edilmesiyle birlikte, yalnızca görüntü tabanlı çalışmalarda kod çözme problemi nedeniyle nispeten daha düşük doğruluk elde edilebilmektedir (Z. Zhou et al., 2014).

Dudak okuma teknolojisi işitme engelli veya işitmede zayıflık bulunan insanların normal dil ifadelerini öğrenme veya anlama becerisini geliştirebilen çözümler sunmaktadır. Bu konuda yapılan çalışmalar, yapay zekâ, bilgi mühendisliği, görüntü işleme sistemleri gibi birçok alanı içermektedir. Dudak okuma uygulamaları, otomatik konuşma tanıma, insan-bilgisayar etkileşimi ve bilgisayarlı görme teknolojilerindeki

hızlı gelişmelerle birlikte, sadece konuşma tanımayı geliştirmek için değil, aynı zamanda insan-makine ve terörle mücadele ara yüzleri gibi yeni alanlarda da kullanılmaktadır (Zheng et al., 2014).

1.1 Dudak Okumanın Gelişim Süreci

Dudak okuma teknolojisi başlangıçta bir fikir olarak 1954'te Sumby tarafından ortaya atılmıştır(Zheng et al., 2014). Sumby, dudak hareketlerinin sesi anlamak için kullanılan görsel bilgi kaynağı olduğunu belirtmiştir. Dudak okuma, dudak hareketinin yorumlanması işlemidir. Aynı zamanda dudak hareketi yoluyla yüz duygularının ifadesiyle ilgili de birtakım bilgiler elde edilebilmektedir.

1984 yılında ilk dudak okuma sistemi IUinois Üniversitesi'nden Petajan ve arkadaşları (Petajan et al., 1988) tarafından kurulmuştur. Bu çalışmada vektör niceleme, dinamik zaman atlama ve yeni bir sezgisel zaman ölçüsü kullanılarak dudak okuma sistemi geliştirmişlerdir. Bu sistem, işitme kanalı için dudak hareketlerini ek bilgi olarak alıp dudak okuma modelini, konuşma tanıma modeliyle bağlamıştır. Sistemde, dudak okuma tanıma modelinin girişi, konuşma tanıma modeli tarafından 2 kanal yoluyla alınmıştır. Böylece optimum koşullarda birden fazla konuşmacının görsel konuşma tanıması gerçekleştirilmiştir. Bu yöntemle gürültülü veya daha fazla konuşan insanın olduğu ortamda bile tanıma oranı artmıştır.

1989 yılında Yuhas ve arkadaşları (Yuhas et al., 1989) tarafından dudak okuma tanıması yapmak için yapay sinir ağları kullanılmıştır. Dudak okuma/tanıma için girdi olarak, tek görsel kanaldan sesli harfler alınıp piksel tabanlı özellik çıkarma yöntemi kullanılmıştır. Bu çalışmada, paralel olarak çalışan birçok işlemci kullanılarak büyük miktarda veri eş zamanlı olarak işlenebilmektedir. Hesaplamayı hızlandırmanın yanı sıra işlemin erken aşamalarında bölümlemeye gerek duyulmamaktadır. Bunun yerine görsel-işitsel yollardan gelen analog sinyaller, gerçek zamanlı olarak akıp doğrudan Çok Geniş Ölçekli Tümleşik (Very Large Scale Integrated) devre uygulamasıyla birleştirilmektedir.

1991 yılında MIT medya laboratuvarlarında dudak okuma ile yüz tanıma özelliği çıkarmak için optik akış yöntemi kullanılmıştır. Sürekli tekrar edilen kelimelerin sadece görsel veriler kullanılarak otomatik olarak tanınması için bir bilgisayar sistemi geliştirilmiştir. Dudak hareketlerinin hızı, kas hareketinin tahmin edilmesine izin veren optik akış verilerinden ölçülebilmektedir. Kas hareketindeki duraklamalar, akışın sıfır hızıyla sonuçlanıp kelime sınırlarını bulmak için kullanılmaktadır. Kas hareketi, söylenen kelimeleri tanımak için kullanılmaktadır (Mase & Pentland, 1991).

1992 yılında Stork ve arkadaşları (Stork et al., 1992) tarafından, California Ricoh araştırma merkezinde dudak okuma tanımlaması yapmak için ilk kez Zaman Gecikmeli Sinir Ağları (Time Delay Neural Network) kullanılmıştır. Bu sistem akustik gürültünün olduğu veya sessiz ortamlarda dudak okumasına olanak sağlanması için geliştirilmiştir.

1993 yılında Goldschen ve arkadaşları (Goldschen et al., 1994) tarafından, konuşmacıya bağlı yeni bir dudak okuma sistemi olarak ilk kez Gizli Markov Modellerini (GMM, Hidden Markov Model) dudak okuma teknolojisinde kullanmışlardır. Bu çalışmada, otomatik konuşma algılamasını geliştirmek için Lipreading olarak bilinen bir görsel-işitsel sistem açıklanmaktadır. Burada dudak okuma sistemi, özellikle bozulmuş akustik koşullar altında doğruluğu iyileştirmek için sesli bir otomatik dudak okuma sistemi ile birlikte kullanılmaktadır. Aynı yıl Texas State University'den Silsbee ve Bovik de (Silsbee & Bovik, 1996) kelime tanıma için vektör kuantalama karakteristiği ve GMM kullanarak izole kelimelere dayanan bir çalışma gerçekleştirmiştir.

1994 yılında, Georgia teknoloji enstitüsünden Rao ve Mersereau (Rao & Mersereau, 1994), dudak konturunun çıkarılması için ilk kez değişken şablonu kullanmıştır. Bu çalışmada, ağzın şeklini modellemek ve otomatik dudak okuma sistemleri tarafından kullanılmak üzere anlamlı boyutlar çıkarmak için bir algoritma açıklanmaktadır. Bu tekniğin bir avantajı, modeli ölçek ve dönüşü telafi edecek şekilde normalleştirme yeteneğidir. Modeli görüntü ile ilişkilendiren bir hata fonksiyonu tanımlanmakta ve hatanın en aza indirilmesi için en uygun modeli vermektedir. Aynı yıl, Carnegie Mellon Üniversitesi'nden Bregler ve Omohundro (Bregler & Omohundro, 1994) GMM ve Yapay Sinir Ağlarını birleştirmiş ve dudak okuma problemine uygulamıştır.

1997'de, İsviçre'de IDIAP Yapay Zeka Enstitüsü'nden Luettin(Luettin, 1997), doğal ışık etkisini azaltıp dudak okuma özelliklerini elde etmek için Aktif Şekil Modelini (AŞM, Active Shape Model) kullanmıştır. Burada yöntem, bir Karhunen-Loéve genişlemesi kullanarak dudak şeklini ve ağız bölgesindeki yoğunlukları sırasıyla temel şekil ve temel yoğunlukların ağırlıklı toplamlarına ayrıştırmaktadır. Yoğunluklar, şekilden bağımsız yoğunluk bilgisi sağlamak için şekil modeliyle deforme olmaktadır. Bu bilgi, model ile görsel arasındaki benzerlik ölçüsüne dayanan görsel aramada kullanılmaktadır. Aynı yıl, Washington Üniversitesi'nden Chiou ve Hwang da (Chiou & Hwang, 1996), özellik çıkarma için Snake modelini kullanmıştır. Modelde yalnızca insan dudaklarının renkli hareketli videosunu kullanarak izole edilmiş kelimeleri tanıyan (akustik veriler olmadan) bir dudak okuma sistemi tasarlanmıştır. Bu dudak okuma

sistemi, Snake, Temel Bileşen Analizi (TBA, Principal Component Analysis) ve GMM'ler kullanarak renkli videolarda dudak okuma gerçekleştirmektedir. Snake algoritması ve TBA, video dizisindeki her görüntüden iki grup görsel özelliği çıkarmak için kullanılmaktadır.

1998 yılında, MIT İnsan-Bilgisayar Etkileşimi Enstitüsü'nden Cootes ve arkadaşları (Cootes et al., 2001), piksel çıkarma yönteminin ve şekil çıkarma yönteminin avantajlarını birleştirmek için Aktif Yüzey Modeli (Active Appearance Models) yöntemini kullanmıştır. Bu çalışmada istatistiksel görünüm modellerini görüntülerle eşleştirmenin yeni bir yöntemi önerilmektedir. Bir eğitim setinden öğrenilen bir dizi model parametresi kontrol modları ve gri seviye varyasyonu, model parametrelerindeki bozulmalar ile indüklenen görüntü hataları arasındaki ilişkiyi öğrenerek verimli bir yinelemeli eşleştirme algoritması oluşturulmaktadır. Aynı yıl, AT&T LABS'den Potamianos ve arkadaşları da (Potamianos et al., 1998), Ayrık Kosinüs Dönüşümü (Discrete Cosine Transform) ve Ayrık Dalgacık Dönüşümü (ADD, Discrete Wavelet Transform) gibi dönüştürme yöntemlerini kullanarak frekans bileşenlerine ait çeşitli özellikleri kullanmıştır.

2000 yılında, Johns Hopkins Üniversitesi'nde toplanan ve dudak okuma problemleriyle ilgilenen araştırmacılar, görsel özellik çıkarma ve dudak okuma konular üzerinde çalışmışlardır. Otomatik dudak okuma için, konuşulan kelimelerin veya ses birimlerinin başarılı bir şekilde tanınmasını sağlayan gelişmiş üç aşamalı, piksel tabanlı görsel esaslı çalışmalar yapmışlardır. İlk aşama, video verilerinin yüksek "enerjili", azaltılmış boyutluluk gösterimini sağlayan bir görüntü sıkıştırma dönüşümüdür. İkinci aşama, az sayıda ardışık görüntü dönüştürülmüş video verilerinin bir birleşimine uygulanan doğrusal diskriminant analizine dayalı veri projeksiyonudur. Üçüncü aşama ise maksimum olabilirlik bir veri dönüşüdür. Bu, doğrusal dönüşüm vasıtasıyla gerçekleşmektedir. Böyle bir dönüşüm, diyagonal kovaryanslı sınıf koşullu Gauss dağılımları varsayımı altında gözlemlenen verilerin olasılığını optimize etmektedir (Potamianos et al., 2000).

2002 yılında, Japonya'nın telekomünikasyon şirketi NTT DoCoMo dünyanın ilk dudak telefonunu geliştirmiştir (Zheng et al., 2014). Bu telefonda, mikrofonun yakınındaki temas sensörleri ağız çevresindeki kasların neden olduğu küçük akım sinyalini tespit edip, bu sinyalleri ses veya SMS metnine dönüştürmüşlerdir. Bu telefon, şimdiye kadar ünlü harfleri tanımlayabilmiştir.

2003 yılında dünyanın en büyük yarı iletken üretici firması olan Intel, bir dudak okuma yazılımı olan işitsel ve Görsel-İşitsel Konuşma Tanıma (GİKT, Audio Visual Speech Recognition)'yi tanıtmıştır. Bu çalışmada görsel-işitsel otomatik konuşma tanımanın ana bileşenlerine iki alanda yeni katkılar sunulmaktadır. Birincisi, bir video ilgi bölgesinin doğrusal görüntü dönüşümlerinin bir kademesine dayanan görsel tasarımı ve ardından görsel-işitsel konuşma entegrasyonu sağlamasıdır. İkincisi ise özellik ve karar füzyon kombinasyonu ile görsel-işitsel konuşmanın eşzamanlı olmayan modelleme ve modalitelerinin güvenilirlik tahminlerini iki modlu tanıma sürecine dâhil etme üzerine yapılan çalışmalardır (Potamianos et al., 2003). Bu yazılım, konuşma yapan kişinin yüz ifadelerini ve dudak hareketini izleyebilmektedir.

2009 yılında, İngiliz bilim adamları, dudak okuma problemini farklı dilleri ayırt edebilecek şekilde geliştirmişlerdir. Bu çalışma, Stephen Cox ve arkadaşları (Cox et al., 2008) tarafından East-Anglia Üniversitesi Bilgisayar Bölümü'nde tamamlanmıştır. Bu çalışma, dudak hareketlerinin istatistiksel modeline dayanmaktadır. Sistem konuşmacının konuştuğu dili yüksek doğrulukla tanımlayabilmektedir.

Microsoft 2013 yılında, dudakları okuyabilen Kinect2'yi tanıtmıştır. Kinect2, kullanıcıdan izlenen vücut verilerini ve kullanıcının kütle merkezinin yanı sıra bir düzlemdeki hareket rotasını analiz etmek için kullanılmaktadır. Değerlendirme süreci üç adımı içermektedir. Birincisi altın standart denge ölçüm platformuyla sezgisel olarak karşılaştırmadır. İkincisi Wii denge panosu ile sayısal analiz düzeyinde karşılaştırma yapıp farklı vücut tiplerine sahip kullanıcıların Vücut Yağ Yüzdesi (Body Fat İndeks) değerlerine göre daha da iyileştirmeler yapmaktır. Üçüncüsü ise Wii denge panosu ile Kinect, kullanıcının ayaklarının göreceli pozisyonunun yanı sıra kullanıcının duruşunun görünümü, kullanıcının çeşitli konumlarıyla karşılaştırmaktır. Derinlik sensörü tarafından elde edilen dudak derinliği bilgisiyle dudak okuma tanıma özelliğine sahip hassas hareket yakalama işlevine sahiptir(Lv et al., 2016).

Dudak okuma teknolojisinin gelişim tarihine bakılınca sadece teorik araştırmada değil, aynı zamanda pratik uygulamada, dudak okuma için yazılım ve donanım hızla geliştiği görülmektedir.

1.2 Dudak Okuma Sistemlerinin Genel Yapısı

Sadece-görüntüye dayalı dudak okuma uygulamalarında temel işlemler ve izlenecek adımlar şekil 1'de gösterilmektedir. Temel olarak girilen verinin tespiti, veride

işlem yapılacak kısımda lokalizasyon, lokalizasyon yapılan kısmı tanıma ve özellik çıkarmayı içermektedir.

Şekil 1: Dudak okuma işlem basamakları

Dudak okuma yerinin tespiti, dudak okuma teknolojisi için en önemli görevlerden biridir. Çünkü temel işlemler, tespit edilen lokal kısım üzerinde yapılacağı için olası bir hatada yöntemini başarımını doğrudan etkilenecektir. Bu alanda ilk yıllarda yapılan çalışmalarda, araştırmacılar genellikle dudak alanını ruj veya reflektörleri boyanmış dudak kullanarak manuel olarak seçerlerdi. Yüz algılama teknolojisinin gelişmesiyle birlikte dudakların tespit edilmesi, bilgisayar sistemlerinde çeşitli algoritmalar sayesinde mümkün hale gelmiştir (Bayrakdar et al., 2016).

Dudak bölgesi tespiti için yapılan çalışmalarda çeşitli yöntemler kullanılmıştır. Örneğin Rao ve Mersereau (Rao & Mersereau, 1994) tarafından yapılan çalışmada, dudak kenarı belirgin yatay kenar özelliğini kullanarak çıkarmışlardır. Ancak bu yöntem ışık, gölge ve sakaldan etkilenmektedir. Lievin ve Luthon (Liévin & Luthon, 2000), cilt ve dudak arasındaki renk farklılıklarını kullanmışlardır. Jun ve Hua ise (Jun & Hua, 2007), kromatiklik kontrastına dayalı uyarlanabilir bir dudak tespit algoritması kullanmıştır. Bu yöntemde renk ve aydınlığı ayırmak için HSV (Hue, Saturation, Value) renk modelini kullanmışlardır.

Dudak bölgesinin tespitinden sonra tespit edilen dudak bölgesinin özelliklerinin çıkarılması, lokalizasyon yapılan kısmı tanıma sürecinin önemli bir parçasını oluşturmaktadır. Etkili ve sağlam karakteristik değerler dudak okuma/tanıma başarım oranını doğrudan etkilemektedir. Günümüzde, dudak özelliği çıkarma yöntemleri, çoğunlukla görsel özellik çıkarma yöntemini benimsemek için kullanılan fonksiyonların geliştirilmesinden oluşmaktadır. Son yıllarda, birçok bilim adamı kullanılan yöntemler için iyileştirme ve yenilik yapmıştır.

Zeliang ve arkadaşları (Zeliang et al., 2012), hesaplamaları azaltmak için dudak bölgesine Maksimum Beklenen ve Temel Bileşen Analizi (Expectation-Maximization Principal Component Analysis) yöntemlerini uygulamıştır. Morade ve arkadaşları (Morade & Patnaik, 2014), ADD ve Uyarlama Yoluyla Büyük Ölçekli Algılamaya (LSDA, Large Scale Detection Through Adaptation) dayalı özellik çıkarma yöntemini

kullanmışlardır. ADD, sadece dudak kısmından belirgin görsel konuşma bilgilerini çıkarmak için kullanılmıştır. LSDA ise özellik boyutunu daha da azaltmak için kullanılmıştır. Kawasaki ve arkadaşları (Kawasaki et al., 2013), dudak okuma performansını arttırmak için konuşmacı ve çevresel adaptasyonu arttırmayı amaçlamışlardır. Lee (Lee, 2014), görsel-konuşma geçişli filtreleme adı verilen otomatik dudak okumanın daha sağlıklı olması için görsel özelliklerin çıkarılmasında kullanılan geçici bir filtreleme tekniği kullanmıştır. Kang ve arkadaşları ise (Kang et al., 2014), 3D hareket yakalama sistemini kullanarak yüz hareket özelliklerinin doğru parametrelerini çıkarmışlardır.

1.3 McGurk ve Füzyon Etkisi

Konuşma normalde tamamen işitsel bir süreç olarak ifade edilmektedir. Fakat konuşma esnasında ortamın gürültüsü, aradaki uzaklık mesafesi veya ses tonunun karşı tarafa efektif aktarılmaması gibi nedenlerden dolayı işitsel yolda oluşan eksiklik görsel süreçle tamamlanmaktadır. Bir konuşmacının yüzünün görüntüsü, konuşma seslerinin algılanmasını etkileyebilmektedir (Paré et al., 2003). Gürültülü ortamda konuşmanın anlaşılabilirliği, dinleyen kişinin, konuşan kişiyi görebildiği zaman artmaktadır (Sumby & Pollack, 1954). Bununla birlikte konuşma esnasında dudağın aldığı birbirine yakın şekiller dudak okuma sürecini zorlaştırabilmektedir. Bu nedenle dudak okuma insanlar tarafından düşük doğrulukta gerçekleşmektedir. Görsel-işitsel konuşma esnasında ortaya çıkan yanılsama üzerine McGurk ve MacDonald (McGurk & MacDonald, 1976) yaptıkları çalışmada, herhangi bir ifadenin ses bileşeni ile başka bir ifadenin görsel bileşeninin bir araya getirilmesi, bu iki farklı ifadenin de dışında farklı üçüncü bir ifadenin algılanmasına neden olduğu gösterilmektedir. Şekilde gösterilen çalışmada soldaki görselde –ba.ba seslendirilmektedir. Ortadaki görselde ise görsel olarak farklı bir sese işitsel olarak –ga,ga eklenmiştir. Bunun sonucunda sağdaki görselde üçüncü bir ses olan –da,da algılanmıştır. Bu yanılsama McGurk etkisi olarak adlandırılmaktadır ve bu etki, dudak okuma işleminin ses ile oldukça ilişkili karmaşık bir problem olduğunu açık bir şekilde ortaya koymaktadır.

Şekil 2: McGurk etkisini göstermek için yapılan çalışma (The Illusions Index, n.d.)

Tiippana tarafından yapılan diğer bir çalışmada (Tiippana, 2014), önce bir ünsüz seslendirilerek kaydedilmiştir. Daha sonra başka bir ünsüz seslendirilen bir görüntüye bu önceki ses eklenmiştir. Konuşma sinyali tek başına tanınmasına rağmen tutarsız görsel bir konuşma ile seslendirildiğinde ikinci uygulamada farklı bir ünsüz olarak algılanmıştır. Bu olay sonucunda üçüncü bir ünsüz algılandığı için füzyon etkisi olarak adlandırılmaktadır.

Dudak okuma teknolojisi işitme engelli veya işitmede zayıflık bulunan insanların normal dil ifadelerini öğrenme veya anlama becerisini geliştirebilen çözümler sunmaktadır. Bu konuda yapılan çalışmalar, yapay zekâ, bilgi mühendisliği, görüntü işleme sistemleri gibi birçok alanı içermektedir. Dudak okuma uygulamaları, otomatik konuşma tanıma, insan-bilgisayar etkileşimi ve bilgisayarlı görme teknolojilerindeki hızlı gelişmelerle birlikte, sadece konuşma tanımayı geliştirmek için değil, aynı zamanda insan-makine ve terörle mücadele ara yüzleri gibi yeni alanlarda da kullanılmaktadır (Zheng et al., 2014).

Bu amaçla bu tez çalışmasında, veri madenciliği ve makine öğrenmesi yöntemleri kullanılarak çeşitli dudak okuma modeli yaklaşımları geliştirilmiştir. Daha önceden yapılmış çalışmalarda paylaşılan yalnızca bir (1) harfin ve rakamın telaffuz edildiği AVLetters2 ve AVDigits isimli video görüntüler içeren veri setleri kullanılmıştır. Bu video görüntülerdeki her bir çerçevedeki konuşmacının dudak bölgesinin sınırları tespit edilerek bu sınırlara yerleştirilen uzamsal işaret noktaları kullanılmıştır. Bu işaret noktalarından, tez çalışmasında önerilen üç farklı uzamsal öznitelik yaklaşımına dayalı

elde edilen öznitelik setleri dudak okuma amacıyla sınıflandırıcı yöntemlere uygulanmıştır. Sınıflandırıcı yöntem olarak, benzer problemlerde sıklıkla kullanılan Rastgele Orman (RO, Random Forest), Destek Vektör Makineleri (DVM, Support Vector Machines) ve K-En Yakın Komşu (KNN, K-Nearest Neighbors) makine öğrenme yöntemleri kullanılarak literatürle kıyasla yüksek dorulukta başarı elde edilmiştir.

2. KAYNAK ARAŞTIRMASI

2.1. Veri Seti ile İlgili Yapılan Çalışmalar

Pei ve arkadaşları (Pei et al., 2013), denetimsiz Rastgele Orman Manifold Hizalamasını (ROMH, Random Forest Manifold Alignment) kullanarak yeni bir dudak okuma yaklaşımı önermişlerdir. Bu çalışmada, sorgu manifoldları ve referans video görüntüler arasındaki benzerlikler kullanılarak dudak okuma tahmini yapılmıştır. Çalışmalar konuşmacı bağımlı ve konuşmacı bağımsız olarak gerçekleştirilmiştir. Bu amaçla bu çalışmada, KinectVS, OULUVS, AVLetters, AVLetters2 ve CUAVEsam veri setleri kullanılmıştır.

Frisky ve arkadaşları (Frisky et al., 2015), video içerik analizi tekniği uygulayarak görsel dudak hareketi tanımaya dayanan bir çalışma yapmışlardır. Çalışmada Negatif-Olmayan Matris Çarpanlara Ayırma (Non Negative Matrix Factorization) ve Çekirdek Seyrek Gösterim Sınıflandırıcısı (Ç-SGS, Kernel Sparse Representation Classifier (K-SRC)) kullanan yeni bir görüntü temelli konuşma tanıma yöntemi önermişlerdir. Bu çalışmada görsel dudak bilgisi içeren videolardan hem yer hem de zamana ilişkin Spatio-temporal özellik tanımlayıcıları kullanarak Üç Ortogonal Düzlem (ÜOD, Three Orthogonal Plane) üzerinden Yerel İkili Örüntüler (YİÖ, Local Binary Patterns) temelli özellikler çıkarmışlardır. Önerdikleri yöntemin performansını ölçmek için AVLetters ve AVLetters2 veri setlerini kullanmışlardır.

Tian ve Ji (Tian & Ji, 2017), görsel-işitsel konuşma tanıma ve dudak okuma için derin öğrenme alanında kullanılan yapay bir tekrarlayan sinir ağı mimarisi olan yardımcı-multimodal Uzun Vadeli Hafıza Ağları (LSTM, Long Short-Term Memory) yöntemini kullanan bir çalışma yapmışlardır. LSTM, uzun süreli bağımlılıklarla başa çıkmak için tasarlanmış bir Tekrarlayan Yapay Sinir ağları (Recurrent Neural Network) mimarisidir (Graves et al., 2005). Standart ileri beslemeli sinir ağlarının aksine, LSTM'nin geri bildirim bağlantıları vardır (Sak et al., 2014). Yalnızca tek veri noktalarını değil, aynı zamanda tüm veri dizilerini işleyebilmektedir. Bu çalışmalarında hem video hem de ses verilerini kullanarak bir GİKT uygulaması gerçekleştirmişlerdir. Bu amaçla çalışmada önerilen yöntem AVLetters, AVLetters2 ve AVDigits olmak üzere 3 veri setine uygulanmıştır.

Bear ve arkadaşları (Bear et al., 2017), Avletters2 veri setini kullanarak konuşmacıya bağlı ve konuşmacıdan bağımsız vizem (visem) sınıflandırıcılar ile dudak okuma modeli geliştirmişlerdir. Konuşmanın yalnızca görsel bilgilerden tanımlanması

olan dudak okumada, telaffuza ait görsel karakteristikler büyük ölçüde konuşmacıya bağlı olmaktadır. Bu çalışmada hem konuşmacı bağımsız hem de konuşmacı bağımlı fonem-vizem haritaları oluşturmak için bir fonem kümeleme yöntemi kullanılmaktadır. Konuşmacıların yüzleri, yalnızca dudakla kombine edilmiş şekil ve görünüm özelliklerinin çıkarıldığı Aktif Görünüm Modelleri (AGM, Active Appearance Model) kullanılarak izlenmiştir.

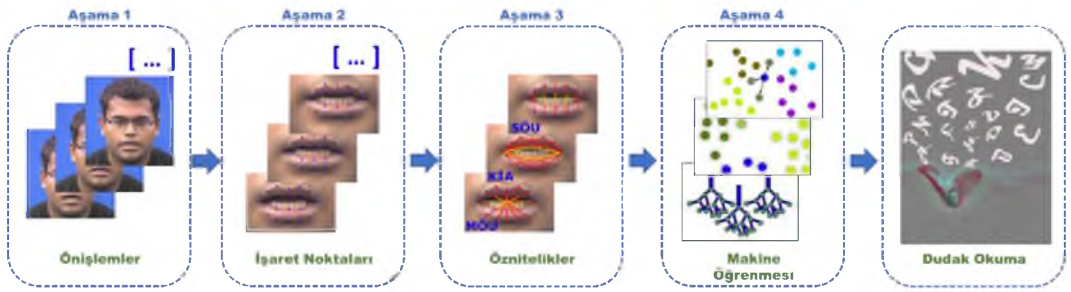
Mattos ve arkadaşları (Mattos et al., 2018), sentetik veriler kullanarak Evrişimli Sinir Ağlarına (Convolutional Neural Network) tabanlı görsel tanıma iyileştirme üzerine çalışma yapmışlardır. Bu çalışmalarında sesbirimlerinin görsel eşdeğeri olan vizemleri tanıma sorununu ele alınmıştır. Bu sorunu, otomatik olarak etiketlenen gerçekçi 3D yüz modellerine dayanan büyük ölçekli sentetik 2D veri kümesi oluşturarak çözmüşlerdir. Bu çalışmalarında Avdigits, AVletters, Cuave ve GRID corpus adlı veri setleri kullanılmıştır.

Kulkarni ve Kirange (Kulkarni & Kirange, 2019), yayın okuma teknikleri üzerine bir araştırma yapmışlardır. Bu çalışmalarında esas olarak derin öğrenme alanına odaklı olarak farklı dudak okuma tekniklerinin ve farklı dil veri setlerinin araştırılmasına odaklanmaktadır. Bu çalışmalarında Avdigits, AVletters, Cuave, GRID corpus IBMVIAVOICE, AusTalk ve MOBIO gibi birçok veri seti kullanılmıştır.

Fernandez-Lopez ve Sukno (Fernandez-Lopez & Sukno, 2018) ise derin öğrenme ile ilgili olarak yaptıkları çalışmalarında otomatik dudak okuma ile ilgili hem uygulanan yöntemler hem de kullanılan veri setleri ile ilgili kapsamlı bir çalışma gerçekleştirmişlerdir. Bu çalışmada, son yıllarda otomatik dudak okuma alanındaki araştırmaları gözden geçirerek, geleneksel olarak adlandırılan önceki yaklaşımlardan uçtan uca otomatik dudak okuma mimarilerine doğru ilerleme vurgulanmıştır. Bu çalışmada harf, kelime ve sözcük gibi oldukça geniş birçok veri seti ele alınmıştır.

3. MATERYAL VE YÖNTEM

Dudak okuma uygulamaları, kademeli yapıda çeşitli hesaplama süreçlerinden oluşmaktadır. Dolayısıyla sistem çeşitli seviyelerde bölümlere ayrılarak her bölümde yapılacak işlemleri ayrı ayrı değerlendirmek sistemin bütününü daha iyi tanımlamaktadır. Bu tez çalışması kapsamında dudak okuma için önerilen sistem şeması Şekil 3'te gösterilmiştir. Önerilen sistem 4 aşamadan oluşmaktadır. Her aşamada gerçekleştirilen işlemler aşağıda özetlenmiştir.



Şekil 3: Dudak okuma için kullanılan işlem basamakları

Aşama 1: Kullanılan veri setinden görüntü girişi ile başlayan süreç, görüntülerde yüz tespiti ve ardında yüz öğelerinin belirlenmesi ile devam etmektedir. Video görüntülerdeki her bir çerçevede yüz ve yüzdeki diğer öğelerin tespiti bu aşamada yapılmaktadır.

Aşama 2: Sonraki aşamada dudak bölgesi belirlenerek dudak konturları üzerinde işaret noktaları belirlenmektedir. Bu aşamada her bir çerçevede tespit edilen iç ve dış dudak sınırları toplam 20 nokta ile işaretlenmektedir.

Aşama 3: Bu aşamada ise işaret noktaları kullanılarak Merkezi-Öklit-Uzaklık, Simetrik-Öklit-Uzaklık ve Komşu-İşaret-Açıları isimli uzamsal özellikler hesaplanarak sınıflandırıcıya uygun öznitelik girişlerine dönüştürülmektedir.

Aşama 4: Son aşamada çıkarılan öznitelikler, KNN, RO, DVM gibi sınıflandırma yöntemlerine uygulanarak sınıflandırılma yapılmakta ve telaffuz edilen dudak tahmini edilmektedir.

Tüm işlemler intel i7-3517 işlemci ve 8 GB RAM'e sahip bir kişisel bilgisayar üzerinde gerçekleştirilmiştir. Video görüntülerden yüz tespiti ve dudak işaretlerinin tespit işlemleri ve öznitelik çıkarım süreçleri Python programlama dili kullanılarak elde edilmiştir. Sınıflandırma işlemleri KNIME ile gerçekleştirilmiştir.

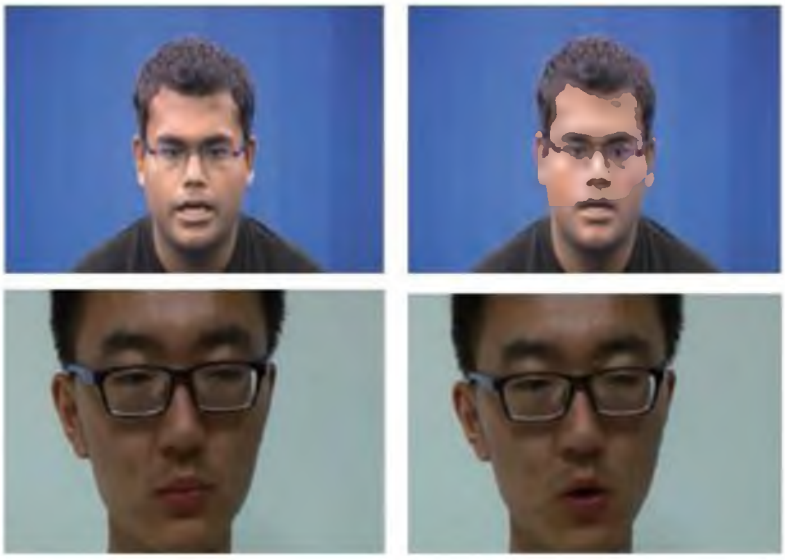
3.1. Kullanılan Veri Setleri

Bu çalışma için, AVletters2 isimli izole harflerden ve AVDigits isimli izole rakamlardan oluşan hizalanmış görsel-işitsel verilerden oluşan veri setlerinden faydalanılmıştır. AVletters2 veri seti, Cox ve arkadaşları (Cox et al., 2008) tarafından oluşturulan AVLetters (Matthews et al., 2002) veri setinin bir uzantısıdır. AVLetters2 veri seti, her harf için yedi tekrar olmak üzere beş kişi tarafından söylenen "A" harfinden "Z" harflerine kadarki 26 harfi içermektedir. AVDigits ise Hu ve arkadaşları (Hu et al., 2016) tarafından oluşturulmuş, 0 ile 9 arasındaki 10 rakamın her birinin 6 kişi tarafından 9 kere tekrar edilmesine ait kayıtlar içermektedir. Aşağıdaki Tablo 1'de her iki veri setiyle ilgili detaylı bilgiler sunulmaktadır.

Tablo 1: Kullanılan veri setlerinin detayları

	AVLetters2	**AVDigits**
Video kayıt içeriği	'A' ile 'Z' arasındaki harfler	'0' ile '9' arasındaki rakamlar
Konuşmacı sayısı	5	6
Tekrar sayısı	7	9
Telaffuz sayısı	910	540
Çerçeve boyutu	1920 × 1080	1920 × 1080

Veri setinde bulunan ses ve video içerikleri ham verilere bölünmüştür. Bu veri setindeki görüntüler, yüksek tanımlı kameralarla 1920 * 1080 RGB kullanılarak renkli olarak kaydedilmiştir (Yuan et al., 2018). AVletters2 ve AVDigits veri kümelerindeki bazı konuşmacılara ait örnek görüntüler Şekil 4'te gösterilmiştir.



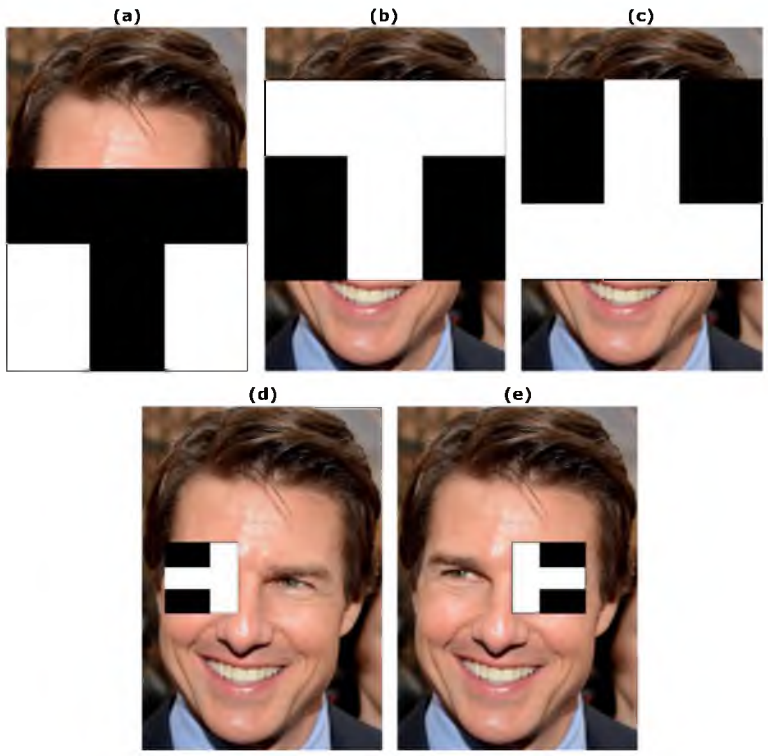
Şekil 4: AVLetter2 ve AVDigits veri setlerindeki konuşmacılara ait örnek görüntüler

3.2. Yüz Tespiti

Öncelikle girdiğimiz verideki yüz bölgesini tespit etmek gerekmektedir. Çalışmamızda yüz tespiti yapmak için OpenCV adlı kütüphane kullanılmıştır. OpenCV gerçek zamanlı bilgisayar görüsü uygulamalarında kullanılan açık kaynak kodlu bir kütüphanedir. Intel Mikroişlemci Araştırma Laboratuvarında geliştirilmiştir. Windows ve Linux adlı işletim sistemleri tarafından desteklenen bu kütüphane kodlama sistemi geliştirilerek birçok işletim sistemi tarafından kullanılmaktadır (Yu et al., 2004) . Bu kütüphane C++ arayüzünü kullanıp 2500'den fazla algoritma içermektedir (Culjak et al., 2012).

3.2.1. Haar Algoritması

OpenCV kütüphanesi yüz tanımak için Haar dalgacıklarından faydalanmaktadır. Yaygın olarak kullanılan bu algoritma gri tonlamalı görüntülerde, siyah renkli olan pikselleri toplayıp beyaz renkli olan piksellerin toplam değerinden çıkarmaktadır (Sharifara et al., 2014). Elde edilen sonuç eşik değeriyle karşılaştırılmaktadır. Eğer sonuç beklentiyi karşılıyorsa özellik amaca uygun olarak değerlendirilmektedir. Bu dalgacık gösteriminde beyaz renk bölgesi "alanı ekle", karanlık renk bölgesi ise "alanı çıkar" olarak yorumlanmaktadır (Cuimei et al., 2017).



Şekil 5: Haar algoritmasının yüzdeki uzuvlara uygulanması

Şekil 5'da Haar algoritmasının yüzdeki uzuvlardan özellik çıkarmak için uygulanmasını göstermektedir. Resimde 5 farklı küçük resim bulunmakta olup her küçük resimde farklı bölgelerin özelliği çıkarılmaktadır. Şekil 5(a)'daki siyah alan, insanların geneline dağılmış yüz özelliklerinin (gözler, kaşlar, burun, ağız) T şeklinde görülebildiğini, beyaz alan ise insanların yanaklarını göstermektedir. Şekil 5(b)'de T şeklinde ve açık renkte olan bölgede kaşlar ve alnın özelliği çıkarılmaktadır. Şekil 5(c)'de, açık rengi içeren ters T şeklindeki bölgesinin burnundan daha büyük göz ve kaş bölgesinin, gri değerinin daha büyük olduğunu göstermektedir, Şekil 5(d) ve Şekil 5(e)'de T şeklindeki alanda gözlerin karşılaştırıldığını göstermektedir.

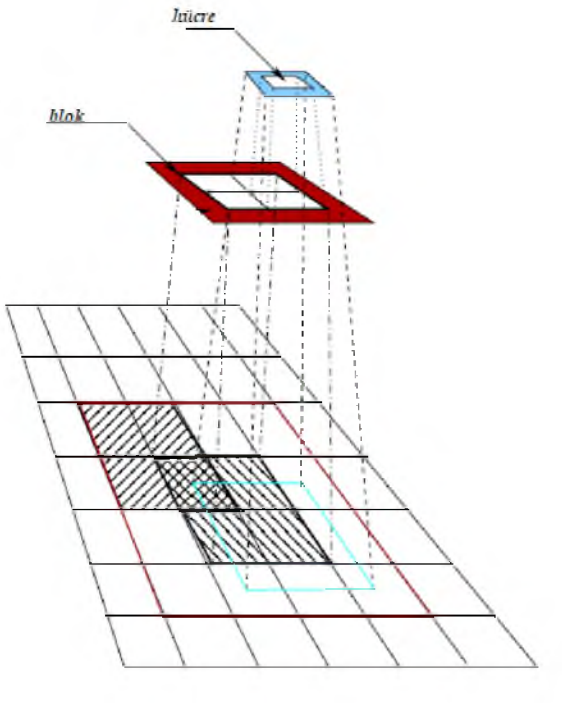
3.3. Dudak Tespiti

Bu aşamada ise dudak tespiti yapılmaktadır. Dudak tespiti yapmak için Dlib adlı kütüphane kullanılmıştır. Dlib, C++ zengin bir ortam sağlamayı amaçlayan, hem mühendisleri hem de bilim adamlarını hedefleyen açık kaynaklı bir kütüphanedir (King, 2009). Dlib kütüphanesi üzerinden yüz tanıma (face recognition), yüz tespiti (face detection) ve yüz modelleme (facial landmarks) gibi uygulamalar yapılmaktadır. Ayrıca yapay sinir ağları, akışlar, grafik kullanıcı arayüzleri, veri yapıları, makine öğrenimi,

görüntü işleme vb. yazılım bileşenleri içermektedir (Gaidar & Yakimov, 2019). Dlib, çoklu görüntülü yüz tanımada düşük verimlilik ve düşük doğruluk sorunlarını çözmeyi, çok görüntülü bir yüz tanıma sistemi kurmayı ve uygun bir izleme şeması keşfetmeyi amaçlamaktadır (H. Zhou et al., 2018).

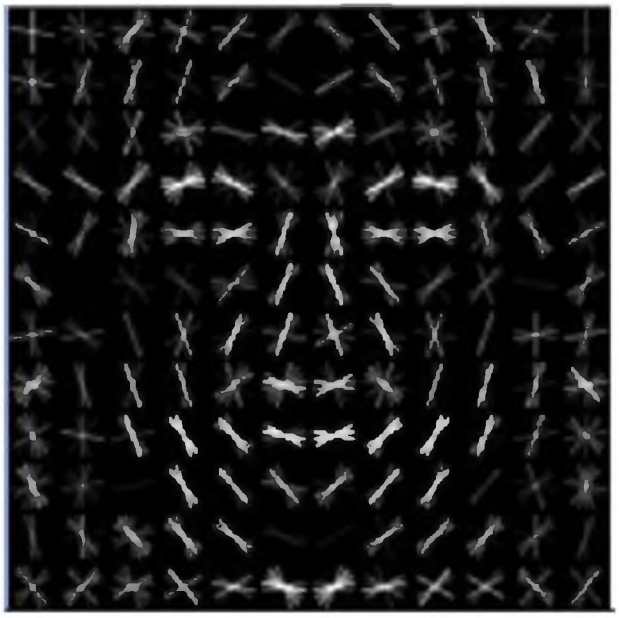
3.3.1. Yönlendirilmiş Gradyanların Histogram (YGH)

Dlib'deki görüntü işleme, Yönlendirilmiş Gradyanların Histogramı (YGH, Histogram of Oriented Gradients) ve DVM algoritmaları üzerine kurulmuştur (Gaidar & Yakimov, 2019). Dlib, yüz dedektörü olarak Evrişimli Sinir Ağlarına sahip bir Maksimum Sınır Nesne Dedektörü kullanmaktadır. Kütüphanesindeki YGH tabanlı yüz dedektörü, nispeten yüksek hızı ve kabul edilebilir doğruluğu nedeniyle yüz algılama için uygulanmaktadır (Lu et al., 2019). YGH tanımlayıcısı çalışma prensibi, uç/kenar bilgilere dayanmaktadır (Dahmane & Meunier, 2011). Yani, her bir pencere bölgesi, kenar yönelimlerinin yerel dağılımı ve karşılık gelen gradyan büyüklüğü ile tanımlanmaktadır. Lokal dağılım, algılama penceresinin hücre adı verilen bir dizi küçük bölgeye bölünmesiyle oluşturulan, yönlendirilmiş gradyanların yerel histogramı ile açıklanmakta olup, burada her bir hücre için kenar gradyanının büyüklüğü tüm pikseller için hesaplanmaktadır. Daha iyi değişmezlik sağlamak için, yerel YGH, bir komşu hücre bloğu üzerinde normalleştirilmektedir. YGH tanımlayıcısının özellik seviyeleri aşağıda Şekil 6'de örneklendirilmiştir.



Şekil 6: YGH algoritmasının sınıflandırma basamakları (Dahmane & Meunier, 2011)

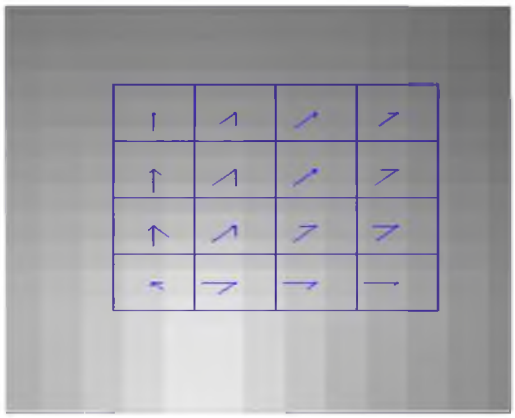
YGH algoritması nesne tanıma için tekil noktaların tanımlayıcıları, bir fotoğraftaki yüzleri tespit etmek için kullanılmaktadır. YGH-tanımlayıcı, DVM sınıflandırıcısının veya bir sinir ağının kullanılmasına izin veren çok boyutlu bir vektöre bir görüntü dönüşümüdür (Voronov et al., 2019). YGH tanımlayıcılarının sınıflandırılması için genellikle DVM kullanılır. Yüzün YGH yapısının bir örneği Şekil 7'te gösterilmektedir.



Şekil 7: Çok sayıda yüz görüntüsünden oluşturulan YGH yüz deseni (Wikipedia, n.d.)

3.3.1.1.YGH tanımlayıcı hesaplama

Görüntüler hücreler adı verilen küçük bağlantılı bölgelere bölünür ve her hücre için bir kenar yönelimleri histogramı hesaplanır. İlk adımda renk normalizasyonu ve gama düzeltmesi gerçekleştirilir. Renk normalizasyon işleminde kullanılacak görüntüler gri tonlamaya dönüştürülmektedir. YGH özellik tanımlayıcısı, bir görüntünün farklı boyutlardaki yamalarında hesaplanmaktadır. Çoklu ölçeklerdeki yamalar birçok görüntü konumunda analiz edilir. Bunu yapabilmek için analiz edilen yamaların sabit bir en-boy oranına sahip olması gerekmektedir. Şekil 8'de 4x4 boyutunda örnek bir yama gösterilmektedir. Yamanın her hücresi gradiyentlerin yönünü göstermektedir (Mallick, 2016).



Şekil 8: YGH tanımlayıcı ile 4x4 boyutundaki yamanın gösterimi

Bir YGH tanımlayıcısını hesaplamak için önce yatay ve dikey gradyanları hesaplamak gerekmektedir. Bu, görüntünün aşağıdaki Şekil 9'da gösterilen çekirdeklerle filtrelenmesiyle elde edilmektedir.

-1	0	1

-1
0
1

Şekil 9: Gradyan histogram filtre çekirdekleri

Ardından aşağıdaki denklem (1) ve (2)'de gardyanların yönü ve büyüklüğü hesaplanmaktadır. Denklem 1'de büyüklüğü gösterilmektedir. Denklem 2'de ise yönü görülmektedir.

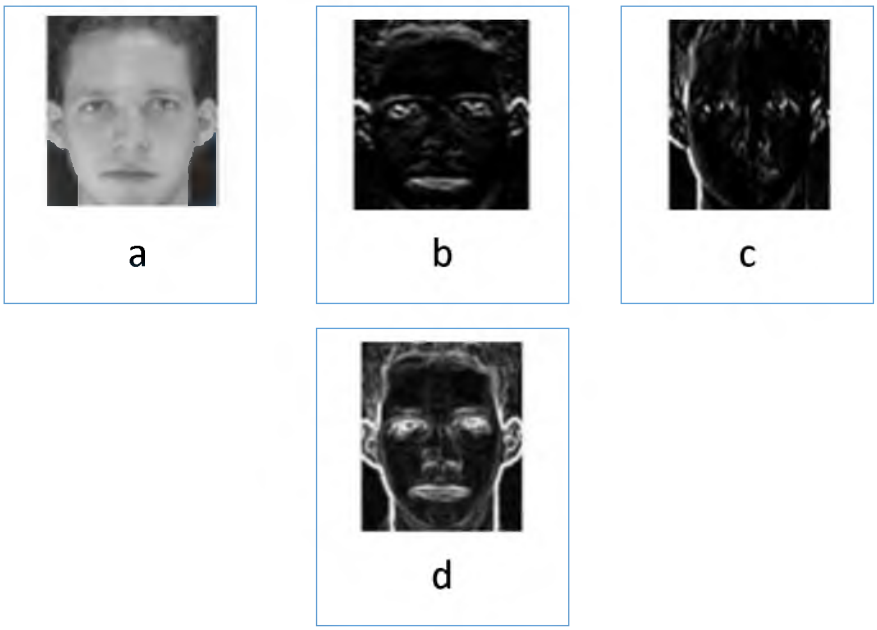
$$g = \sqrt{gx^2 + gy^2} \qquad (1)$$

$$\theta = arctan\frac{gx}{gy} \qquad (2)$$

Buradaki düşünce, yerel gradyan yoğunluğunun ve yönünün dağılımının yerel nesne görünümünü ve şeklini tanımlayabilmesidir (Chen et al., 2014).

Şekil 10'da Dadi ve Pillutla (Dadi & Pillutla, 2016) tarafında yapılmış çalışmadan alıntılanan örnek bir gradyan gösterilmektedir. X gradyanı dikey çizgilerde ve y gradyanı yatay çizgilerde belirginleşmektedir. Şekil 10(a)'da örnek bir yüz şekli gösterilmektedir. Şekil 10(b)'de y gradyanlarının belirginleşmesine örnek gösterilmektedir. Şekil 10(c)'de ise dikey çizgiler belirginleşmektedir. Yoğunlukta keskin bir değişiklik olan her yerde

gradyanlar belirginleşmektedir. Fakat keskin değişiklik olmayan pürüzsüz bölgelerde gradyan belirginleşmesi görülmemektedir.

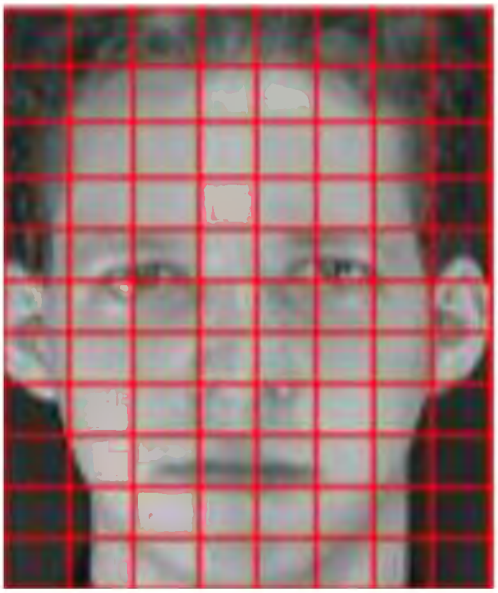


Şekil 10: Örnek gradyanlar

Gradyan görüntüleri, arka plan gibi gereksiz bilgileri çıkarıp ana görüntüye odaklanmaktadır (Dadi & Pillutla, 2016). Her pikselde gradyanların bir büyüklüğü ve yönü bulunmaktadır. Bir pikseldeki gradyan büyüklüğü, üç kanalın gradyan büyüklüğünün maksimumudur. Açı ise maksimum gradyan'a karşılık gelen açıdır.

3.3.1.2.Yönlendirme gruplaması

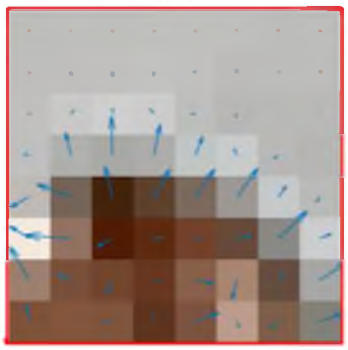
Gradyan hesaplamasında elde edilen değerlerin sayısına bağlı olarak, hücre içindeki her piksel, oryantasyona dayalı bir histogram kanalı için ağırlıklı bir oy kullanmaktadır (Dadi & Pillutla, 2016). Hücreler, radyal veya dikdörtgen şeklinde olabilmektedir. Kullanılan görüntüler 8×8 hücrelere bölünüp ve her hücre için bir gradyan histogramı hesaplanmaktadır (Singh et al., 2020). Bir görüntünün bir yamasını tanımlamak için bir özellik tanımlayıcı kullanmanın önemli nedenlerinden biri, kompakt bir temsil sağlamasıdır (Mallick, 2016). 8×8 görüntü yaması, $8x8x3 = 192$ piksel değerleri içermektedir. Kanallar $0 - 360°$ veya $0 - 180°$'e yayılmıştır. Dalal ve Triggs (Dalal & Triggs, 2005), işaretsiz gradyanlarla birlikte kullanılan histogramın en iyi performansı gösterdiğini gözlemlemiştir. Aşağıdaki Şekil 11'de hücrelere ayrılmış örnek bir görüntüyü göstermektedir.



Şekil 11: Hücrelere ayrılmış örnek bir yüz (Dadi & Pillutla, 2016)

Şekil 12'de Singh ve arkadaşları (Singh et al., 2020) tarafından yapılan çalışmadan alıntılanan şekilde, örnek bir görüntünün yaması gösterilmektedir. Şekil 12 (a)'da görüntü yaması, Şekil 12(b)'de ve Şekil 12(c)'de ise yamaya ait gradyanlar sırasıyla genlik ve yönü gösterilmektedir. Yamada kullanılan oklar gradyanın yönü ve büyüklüğünü göstermektedir. 8×8 hücrelerde ise gradyentleri temsil eden ham sayıları gösterilmektedir. Açılar 0 ile 180 derece arasında numaralandırılmıştır.

(a)



(b)

2	3	4	4	3	4	2	2
5	11	17	13	7	9	3	4
11	21	23	27	22	17	4	6
23	99	165	135	85	32	26	2
91	155	133	136	144	152	57	28
98	196	76	38	26	60	170	51
165	60	60	27	77	85	43	136
71	13	34	23	108	27	48	110

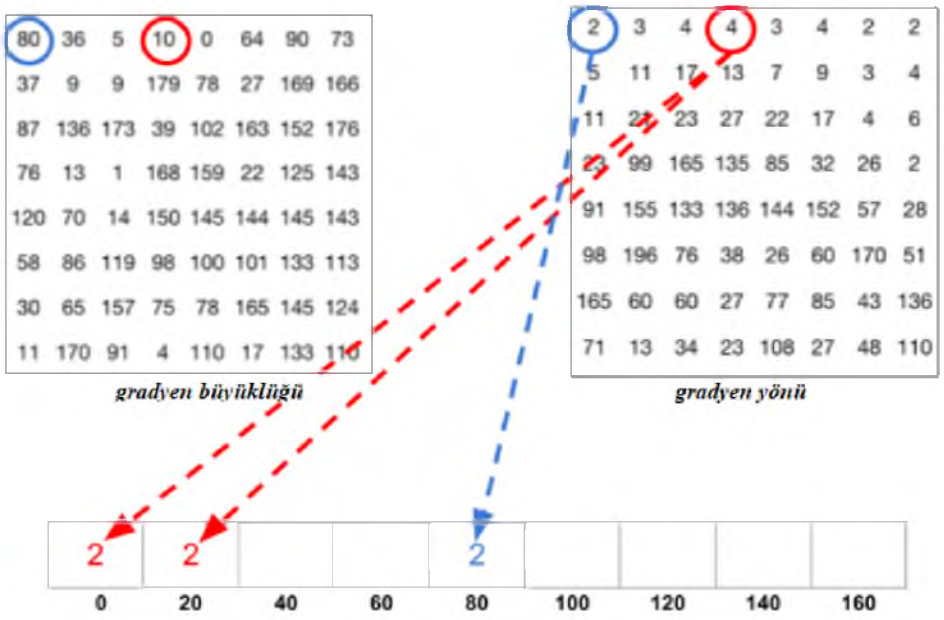
(c)

80	36	5	10	0	64	90	73
37	9	9	179	78	27	169	166
87	136	173	39	102	163	152	176
76	13	1	168	159	22	125	143
120	70	14	150	145	144	145	143
58	86	119	98	100	101	133	113
30	65	157	75	78	165	145	124
11	170	91	4	110	17	133	110

Şekil 12: Örnek görüntüden çıkarılan yama ve gradyanlar

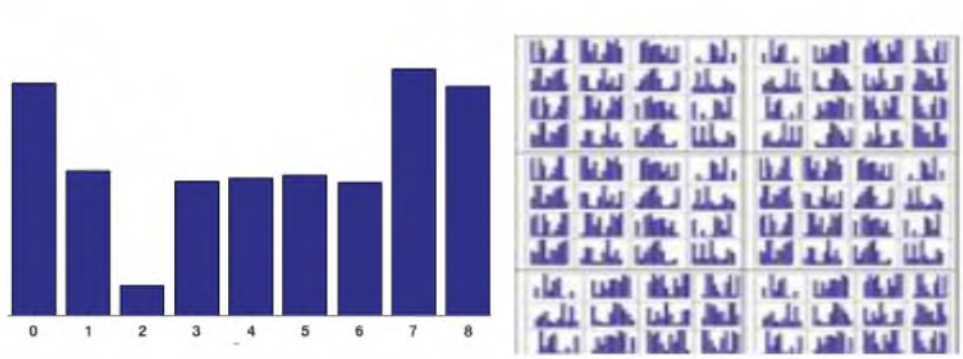
Sonraki aşamada $0, 20, 40 \ldots 160$ açılarına karşılık gelen 9 bölme içeren histogram oluşturulmaktadır. Şekil 13'de buna örnek gösterilmektedir. Her büyüklüğün yönü, kendi karşılığı olan kutucuğa eklenmektedir. Tabloda 160 dereceye kadar derecelenmektedir. 160-180 arasındaki değerler 0 derece ve 160 dereceye eşit orantılı

olacak şekilde dağılmaktadır. Şekil 13'de mavi kutucukla belirtilen 80 değerinin yönü, kendi büyüklüğünün karşılığı olan 80 kutucuğuna eklenmektedir. Kırmızı ile belirtilen 10 değeri ise kendisine ait bir kutucuk olmadığı için sağında ve solunda bulunan 0 ve 20 kutucuklarına orantılı olarak eklenmektedir. Sağında ve solunda bulunan kutucuklara eşit mesafede olduğu için yön değeri olan 4,0 ve 20 kutucuklarına 2'şer olarak eklenmektedir. Bu şekilde tüm değerler gerekli kutucuklara doldurulup 9 bölmeli histogram oluşturulmuş olur.



Şekil 13: Histogramların oluşturulması için yapılan çalışmaya örnek (Mallick, 2016)

Daha sonra 8×8 hücrelerdeki tüm piksellerin katkıları, 9 bölmeli histogramı oluşturmak için toplanmaktadır. Hücrelerden oluşan bir bloğa ait oluşturulan tüm histogramlar birleştirilerek büyük bir histogram oluşturulmaktadır. Toplam oluşturulduktan sonra Şekil 14'deki gibi örnek bir tablo oluşmaktadır. Şekil 14(a)'da bir hücreye ait grafik görülmektedir. Şekil 14(b)'de ise görüntüde kullanılan tüm hücrelerin tablolarının bir arada oluşan görüntüsü gösterilmektedir.



(a) (b)

Şekil 14: Histogram blokları. (a) Tek bir hücre için örnek histogram (Mallick, 2016), (b) görüntüde kullanılan tüm hücreleri temsil eden toplam histogramların gösterimi (Dadi & Pillutla, 2016)

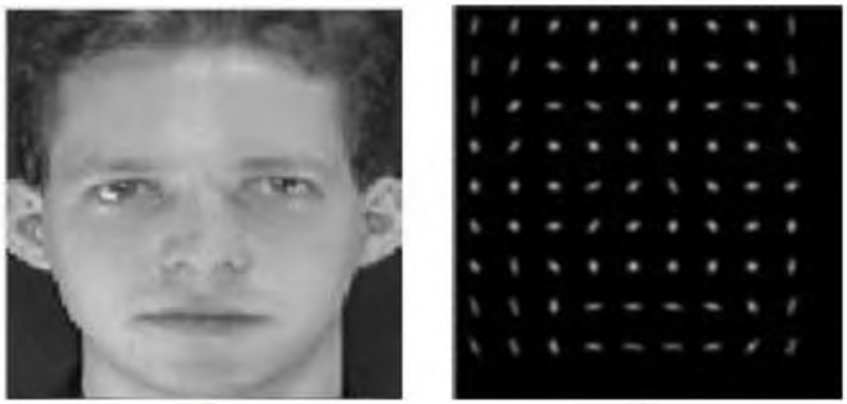
3.3.1.3.Normalizasyon

Bir görüntünün gradyanları, genel ışığa duyarlıdır. Çalışmalarda yapılan tanımlayıcıların aydınlatma varyasyonlarından bağımsız olmasını istenmektedir. Işık değişimlerinden etkilenmemeleri için histogram üzerinde yapılan çalışmalara "normalleştirme" denilmektedir. Bu işlemler yapıldıktan sonra narmalizasyon işlemine geçilmektedir. Verilen bir bloktaki tüm histogramları içeren normalize edilmemiş vektör "v" olarak isimlendirilmektedir. Normalleştirme faktörü aşağıdaki gibi hesaplanır:

$$\nu = \frac{V\kappa}{\sqrt{\|V\kappa\|^2+1}} \qquad (3)$$

burada $V\kappa$, blok için birleşik histograma karşılık gelen bir vektördür ve ν, son bir YGH özelliği olan normalleştirilmiş bir vektördür.

YGH ile yapılan çalışmalar sonunda Şekil 15'deki gibi örnek bir görüntü elde edilmektedir. Burada Şekil 15(a)'da kullanılan örnek görüntüden YGH histogramı kullanarak Şekil 15(b)'deki şekil elde edilmiştir.



Şekil 15: YGH dönüşümü. (a) Kullanılan örnek görüntü, (b) YGH kullanarak çıkarılan görüntü (Dadi & Pillutla, 2016)

3.4. Yüz İşaretçilerinin Tespiti

Yüzdeki işaretlerin tespiti, öncelikle görüntüdeki yüzün ve yüzdeki bölgelerin tespit edilmesi gibi önemli bir aşamayı gerektirir. Yüz tespiti için, önceki bölümlerde anlatılan haar-cascade algoritması (ya da Viola ve Jones dedektör) (Viola & Jones, 2001) veya Yönlendirilmiş Gradyanların Histogramı (Dalal & Triggs, 2005) gibi yöntemler, ya da günümüzde sıklıkla adını duyduğumuz derin öğrenme tabanlı algoritmalar kullanabilir. Burada önemli olan yüzü sınırlayan sınırların tespit edilmesidir. İkinci aşamada ise yüzdeki öğeleri işaretlemek gerekmektedir. Bu amaçla geliştirilmiş çeşitli işaretleyici algoritmalar bulunmaktadır. Bu algoritmalar çoğunlukla, ağız, sağ ve sol kaş, sağ ve sol göz, burun ve çene gibi öğeleri işaretlemektedirler. Bizim için önemli olan öğe bu çalışmada ağız bölgesidir. Çünkü harflerin telaffuzu sırasında dudak bölgesinin çeşitli formalara girdiği ve bunun her harf için değişkenlik gösterdiği düşünülmektedir. Dolayısıyla harflerin telaffuzu sırasında, ağız sınırlarını gösteren işaretçilerin her bir video çerçevesinde uzamsal koordinatlarının birbirilerine göre değişimleri bizlere harfleri sınıflandırmada oldukça faydalı öznitelikler sağlayabilir. Bu çalışmada bu amaçla, doğrusal-DVM sınıflandırıcı ile birleştirilmiş YGH özniteliklerle yüz tespiti yapıldıktan sonra dudak sınırları üzerindeki işaretleri tespit etmek için Kazemi ve Sullivan (Kazemi & Sullivan, 2014) tarafından önerilen Regresyon Ağaçları Topluluğu (Ensemble of Regression Trees) olarak isimlendirilen yöntem kullanılmıştır. Bu yöntemde kaskad/kademeli çoklu regresyon yapısı kullanılmıştır. Aşağıda Kazemi ve Sullivan (Kazemi & Sullivan, 2014) tarafından önerilen yöntemin detayları verilmiştir.

3.4.1. Regresyon ağaçları topluluğu

Regresyon ağaçları topluluğu (RAT), yüz öğelerinin manuel olarak etiketlendiği ((x,y) koordinatları işaretlenerek) bir eğitim veri seti kullanılmaktadır. Daha sonra bu eğitim veri seti kullanılarak bir Regresyon Ağaçları Topluluğu (RAT, Ensemble of Regression Trees) yüz işaretlerini tespit etmek için piksel yoğunlukları esasına göre eğitilmektedir.

3.4.2. Kaskad regresörler

Kaskad yapısında bulunan her bir regresör $r_t(\cdot,\cdot)$, I görüntüsünden ve geçerli tahmin $\hat{S}^{(t)}$ den elde ettiği güncelleme vektörünü geçerli tahmine $\hat{S}^{(t)}$'e ekleyerek tahmini daha da iyileştirir (Kazemi & Sullivan, 2014):

$$\hat{S}^{(t+1)} = \hat{S}^{(t)} + r_t\left(I, \hat{S}^{(t)}\right) \tag{4}$$

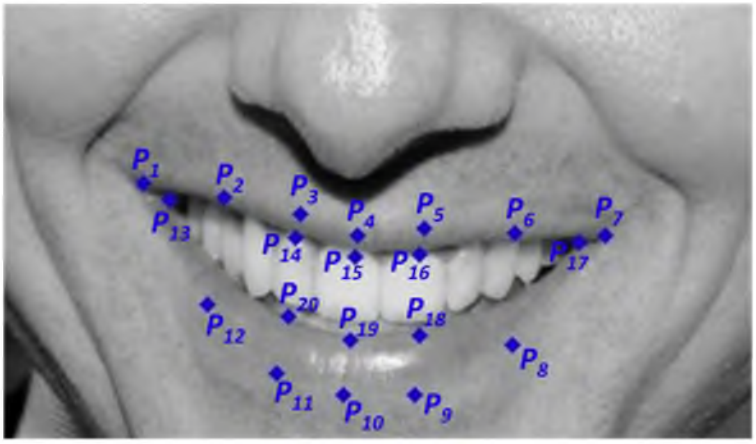
burada x_i I görüntüsündeki $i.$ yüz işaretinin (x, y) koordinatları olmak üzere, $\mathrm{x}_i \in \mathbb{R}^2$ dir. Buna göre, I görüntüsündeki p adet yüz işaretine ait koordinatlar $S = \left(\mathrm{x}_1^T, \mathrm{x}_2^T, \ldots, \mathrm{x}_p^T\right)^T \in \mathbb{R}^{2p}$ olmak üzere S'nin geçerli tahmini $\hat{S}^{(t)}$ ile ifade edilmektedir.

Kaskad yapısındaki kritik nokta, r_t regresörünün tahminlerini, I'den hesaplanan ve mevcut şekil tahmini $\hat{S}^{(t)}$'ye göre indekslenen piksel yoğunluğu değerleri gibi özelliklere dayalı olarak yapmasıdır. Bu, sürece bir tür geometrik değişmezlik getirir ve kaskad ilerledikçe, yüzdeki kesin bir anlamsal konumun indekslendiğinden daha emin olunabilir. Bu yöntemde, her bir r_t'yi eğitmek için, (Hastie et al., 2009)'da açıklandığı gibi bir kare hata kaybı toplamı ile gradyan ağacı güçlendirme (gradient tree boosting) algoritmasını kullanılmıştır. Yöntemle ilgili detaylar Kazemi ve Sullivan (Kazemi & Sullivan, 2014) tarafından yapılan çalışmada açıklanmıştır.

3.5. Uzamsal Öznitelikler

Bu çalışma kapsamında önerilen uzamsal öznitelikler aşağıdaki Şekil 16'te görüldüğü gibi her bir video görüntüdeki çerçevelerde dudak üstünde tespit edilen 20 adet işaret noktası kullanılarak elde edilmiştir. Telaffuz kaydını içeren video görüntüdeki çerçevelerden elde edilen bu öznitelikler bir araya getirilerek her bir kayıt için öznitelik vektörü oluşturulmuştur. AVLetters2 ve AVDigits veri setlerinde yer alan sırasıyla toplam 910 ve 540 video görüntü kaydındaki çerçeve sayısı değişkenlik göstermektedir. AVLetters2 ve AVDigits veri setlerinde en yüksek çerçeve sayıları sırasıyla 67 ve 64

çerçeve olup, daha az sayıda çerçeve içeren kayıtların son çerçeve öznitelikleri tekrarlanarak öznitelik vektörlerinin uzunlukları eşit hale getirilmiştir.

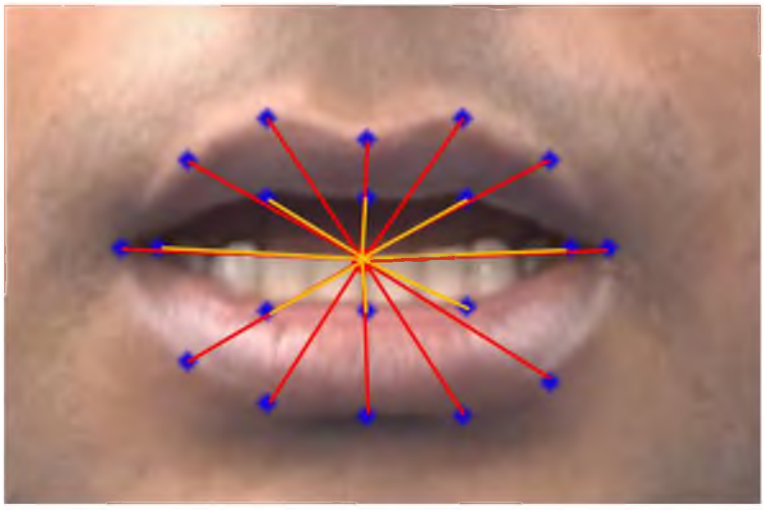


Şekil 16: Dudak sınırları üzerinde tespit edilen 20 işaret noktası

Aşağıda detayları verilen öznitelik çıkarım yaklaşımları Şekil 16'teki uzamsal işaret noktalarına göre elde edilmiştir. Bunlardan 12 tanesi dış dudak ($P_i,\ i = 1, 2, \ldots, 12$) ve 8 tanesi iç dudak ($P_i,\ i = 13, 14, \ldots, 20$) üzerinde olmak üzere toplam 20 işaret noktası bulunmaktadır.

3.5.1. Merkezi Öklid-Uzaklık (MÖU)

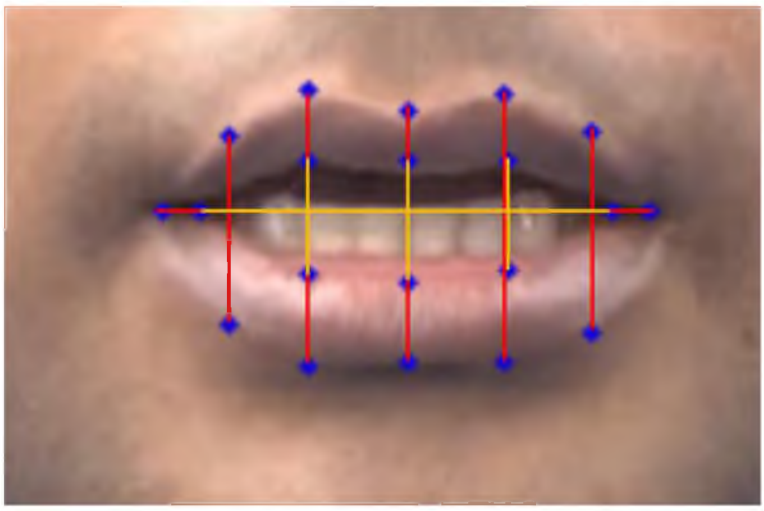
Bu yaklaşımda, Şekil 17'te gösterildiği gibi sol dış kenar (P_1) ile sağ dış kenar (P_7) işaretler arasındaki tam orta noktanın uzamsal koordinatları hesaplanarak tüm diğer dudak işaretleri ile arasındaki Öklit-Uzaklığı (Denklem 6) hesaplanmış ve elde edilen öznitelikler sınıflandırıcıya verilmiştir. Her bir çerçeve için 20 öznitelik olmak üzere AVLetters2 ve AVDigits veri setleri ait her bir video görüntü kaydı için sırasıyla toplam $67\ x\ 20 = 1340$ ve $64\ x\ 20 = 1280$ öznitelik elde edilmiştir.



Şekil 17: Merkezi Öklit-Uzaklık ölçeğine göre öznitelikler

3.5.2. Simetrik Öklid Uzaklık (SÖU)

Bu yaklaşımda, Şekil 18'da görüldüğü gibi dudağın üst ve alt dikey simetrik işaret noktaları arasındaki Öklit-Uzaklığı (Denklem 6) hesaplanmış ve elde edilen öznitelikler sınıflandırıcıya verilmiştir. Her bir çerçeve için 10 öznitelik olmak üzere AVLetters2 ve AVDigits veri setleri ait her bir video görüntü kaydı için sırasıyla toplam $67\ x\ 10 = 670$ ve $64\ x\ 10 = 640$ öznitelik elde edilmiştir.

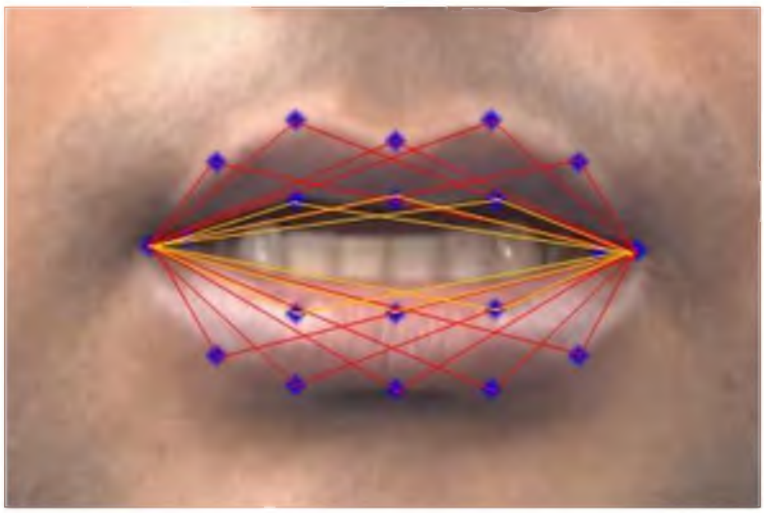


Şekil 18: Simetrik Öklit-Uzaklık ölçeğine göre öznitelikler

3.5.3. Komşu İşaret Açıları (KİA)

Bu yaklaşımda, Şekil 19'de görüldüğü gibi sol dış kenar (P_1) ile sağ dış kenar (P_7) arasında kalan dudağın dış sınırları üzerinde yer alan 10 işaret noktasıyla ve sol iç kenar

(P_{13}) ile sağ iç kenar (P_{17}) arasında kalan dudağın iç sınırları üzerinde yer alan 6 işaret noktasıyla oluşan açılar Denklem 7'e göre radyan cinsinden hesaplanarak elde edilen öznitelikler sınıflandırıcıya verilmiştir. Her bir çerçeve için 16 öznitelik olmak üzere AVLetters2 ve AVDigits veri setleri ait her bir video görüntü kaydı için sırasıyla toplam $67\ x\ 16 = 1072$ ve $64\ x\ 16 = 1024$ öznitelik elde edilmiştir.



Şekil 19: Komşu İşaret Açıları yöntemine göre öznitelikler

3.6. Gri Seviye Eş Oluşum Matrisi Öznitelikleri

Tez çalışmasında görüntü işlemede bilinen ve sıklıkla kullanılan öznitelik çıkarım yöntemlerinden biri olan GLCM ile elde edilen öznitelikler karşılaştırma amacıyla ayrıca kullanılmıştır.

3.6.1. Gri seviye eş oluşum matrisi

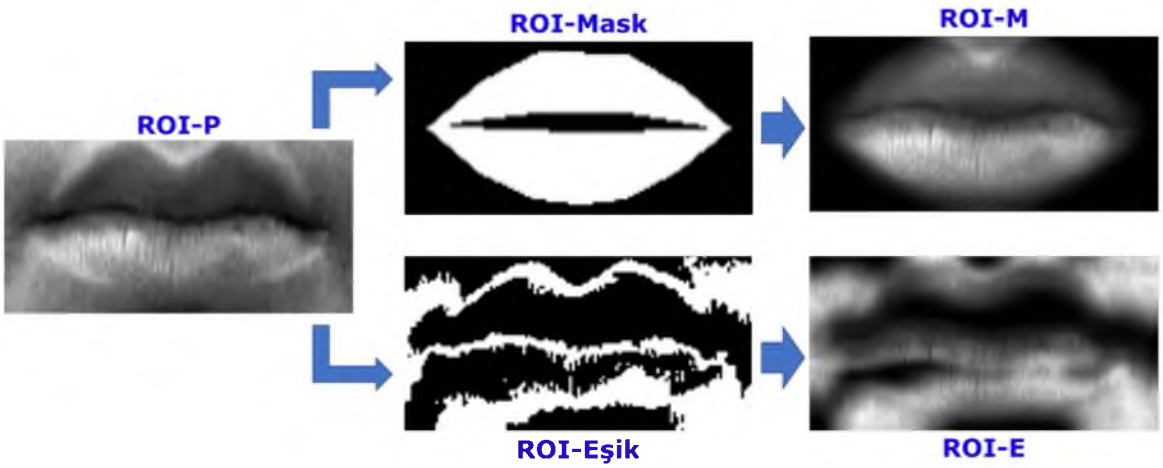
Gri Seviye Eş Oluşum Matrisi (GLCM, Gray Level Co-Occurence Matrices), Haralick (Haralick et al., 1973) tarafından geliştirilmiş olan istatistiksel görüntü özniteliklerinin hesaplanmasında kullanılan bir doku analiz yöntemidir. Görüntü dokularını analiz etmek ve GLCM'e dayalı özellik çıkarma için ikinci dereceden istatistikler kullanılmaktadır. Bu görüntü özelliklerini ikili piksel grupları arasındaki gri seviyesi değişimlerinin farklı kombinasyonları ile ilişkili olarak elde etmektedir. İlk piksel referans olarak, ikinci piksel ise komşu piksel olarak bilinmektedir. GLCM, bir görüntüdeki piksel parlaklık değerlerinin ne sıklıkla bir örüntü meydana getirdiğini gösteren bir matristir. GLCM aşağıdaki denklemle ifade edilebilir (Xian, 2010),

$$P_{\mu}(i,j) \quad (i,j = 0,1,2,\dots,N-1) \tag{4}$$

burada i ve j piksellerin gri seviyeleri, N gri yeğinlik seviyelerini, μ iki pikselin konum ilişkisini ifade eder ($\mu = (\Delta x,\ \Delta y)$). Farklı μ değerleri, işlenen iki piksel seçiminde aralığı ve yönü belirler. GLCM matrisi, genel olarak piksel aralığına, piksel yönüne (0º, 45º, 90º ve 135º) ve gri-seviye yeğinlik seviyesine bağımlıdır (Ondimu & Murase, 2008).

3.6.2. GLCM özniteliklerinin elde edilmesi

GLCM özniteliklerinin elde edilmesi amacıyla, belirlenen dudak işaretleri kullanılarak aşağıda Şekil 20'de görüldüğü gibi ilgilenilen bölge (Region of Interest, ROI) olan dudak bölgesinin dikdörtgen kesit penceresi (ROI-P), bu kesitin maskelenmiş örneği (ROI-M) ve eşiklenmiş örneği (ROI-E) çıkarılmıştır.



Şekil 20: GLCM için dudak bölgesi örnekleri

Her bir video görüntüdeki çerçevelerden elde edilen örnekler için, GLCM öznitelikleri mesafe=1 ve 0, $\pi/4$, $\pi/2$, $3\pi/4$ radyal açılara bağlı olarak kontrast (contrast), benzemezlik (dissimilarity), homojenlik (homogeneity), enerji (energy), korelasyon (correlation) ve açısal ikinci moment olmak üzere toplam 6 ölçek değerleri hesaplanmıştır. Buna göre bir çerçeve için 24 öznitelik olmak üzere AVLetters2 ve AVDigits ver isetlerine ait her bir video görüntü kaydı için sırasıyla toplam $67\ x\ 24 = 1608$ ve $64\ x\ 24 = 1536$ öznitelik elde edilmektedir.

3.7. Öznitelikler Hesaplanırken Kullanılan Yöntemler

Öznitelik ölçütleri elde edilirken Öklid uzaklık ölçütü ve noktalar arası radyal açı ölçütü kullanılmıştır. Bu ölçütlere ilişkin detaylar aşağıda sunulmuştur.

3.7.1. Öklid uzaklık ölçütü

Öklid uzaklık ölçütü, Öklid uzayında iki nokta arasında hesaplanan uzaklık ölçüsüdür (Gower, 1985). Bu tez çalışmasında önerilen MÖU ve SÖU öznitelik çıkarım yöntemleri Öklid-Uzaklık ölçütünü kullanmaktadırlar. Öklid uzaklık ölçütü genel olarak aşağıdaki gibi ifade edilebilir,

$$d(i,j) = \sqrt{\sum_{k=1}^{p} \left(X_{ik} - X_{jk}\right)^2} \tag{5}$$

burada $d(i,j)$, her biri p düzlemle ifade edilen iki nokta (X_i ve X_j noktaları) arasındaki Öklid uzaklığını ifade etmektedir. Bu durumda (x,y) koordinat noktaları ile ifade edilen A ve B noktaları arasındaki Öklid uzaklığı aşağıdaki gibi ifade edilebilir,

$$d(A,B) = \sqrt{(x_A - x_B)^2 + (y_A - y_B)^2} \tag{6}$$

3.7.2. Açı ölçütü

Tez çalışması kapmasında önerilen üçüncü öznitelik yaklaşımı olan KİA, üç nokta arasındaki radyal açı ölçüsünü kullanmaktadır. Öklid uzayında A, B ve C noktalarının oluşturduğu iki vektör ($\overrightarrow{AB}$ ve $\overrightarrow{BC}$) arasında kalan θ açısı aşağıdaki gibi hesaplanabilir,

$$\theta = \cos^{-1}\left(\frac{\overrightarrow{AB} \cdot \overrightarrow{BC}}{\left\|\overrightarrow{AB}\right\| \left\|\overrightarrow{BC}\right\|}\right) \tag{7}$$

burada $\overrightarrow{\ }$ öklid vektörünü, $\left\|\overrightarrow{\ }\right\|$ vektörün uzunluğunu ve $\cdot$ ise iç-çarpımı ifade etmektedir.

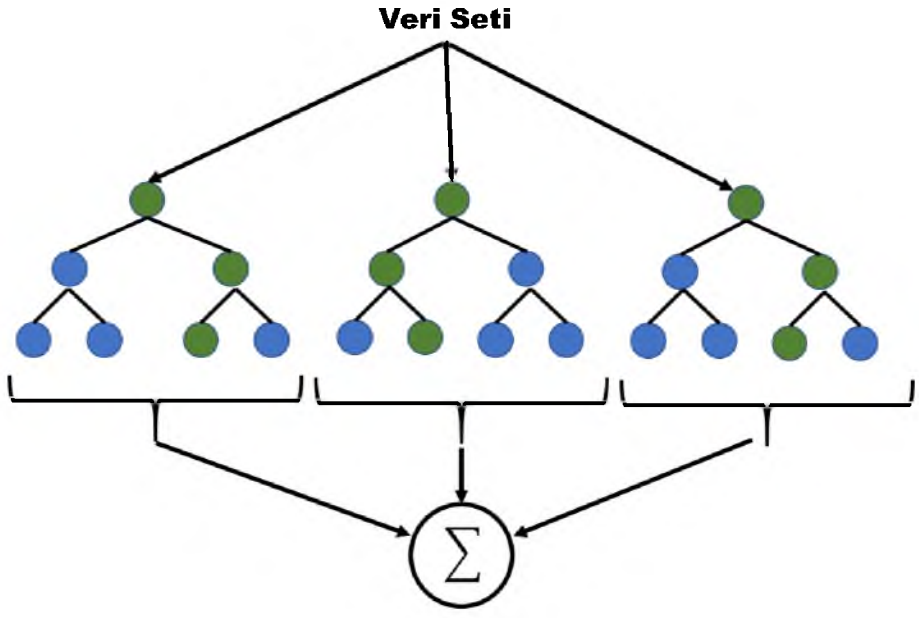
3.8. Sınıflandırma Yöntemleri

Bu çalışmada, önerilen öznitelik yaklaşımlarının performanslarını karşılaştırmak için literatürde sıklıkla kullanılan üç farklı sınıflandırma yöntemi kullanılmıştır. Bu sınıflandırma yöntemleri Rastgele Orman, Destek Vektör Makineleri ve K-En Yakın Komşu yöntemleri olup aşağıda bunlarla ilgili detaylı bilgiler sunulmuştur. Tüm

sınıflandırma sonuçları K-fold değeri 10 şeçilerek çapraz doğrulama uygulanarak elde edilmiştir.

3.8.1. Rastgele Orman (RO)

Rastgele Orman bir ağaç topluluğudur. Her bir karar ağacının birbirinden bağımsız olarak örneklenen rastgele bir vektör kullanılarak üretildiği ve her ağacın bir girdi vektörünü sınıflandırmada en popüler sınıf belirlenirken bir birim oya sahip olduğu bir sınıflandırıcı topluluğudur (Pal, 2005). Buna göre bağımsız ormanlarda en çok ortaya çıkan sonuç geçerli sonuç olarak kabul edilmektedir. Böylece sonuç, bir ağaç değil birçok ağacın ortak kararı ile belirlenmekte ve sonuç yüksek doğrulukta tahmin edilebilmektedir (Breiman, 2001). Rastgele ormanlar veya rastgele karar ormanları algoritması, sınıflandırma dışında regresyon veya eksik gözlem tahmini gibi problemlerde de kullanılabilmektedirler. Eğitim aşamasında çok sayıda karar ağacı oluşturarak problemin tipine göre sınıflandırma veya regresyon tahmini yapabilen grupsal bir yapıya sahip bir yöntemdir. Bu yöntemde karar ağaçları bir araya gelerek şekil 21'daki gibi karar ormanı oluşturmaktadır.



Şekil 21: RO sınıflandırma yönteminin çalışma prensibi

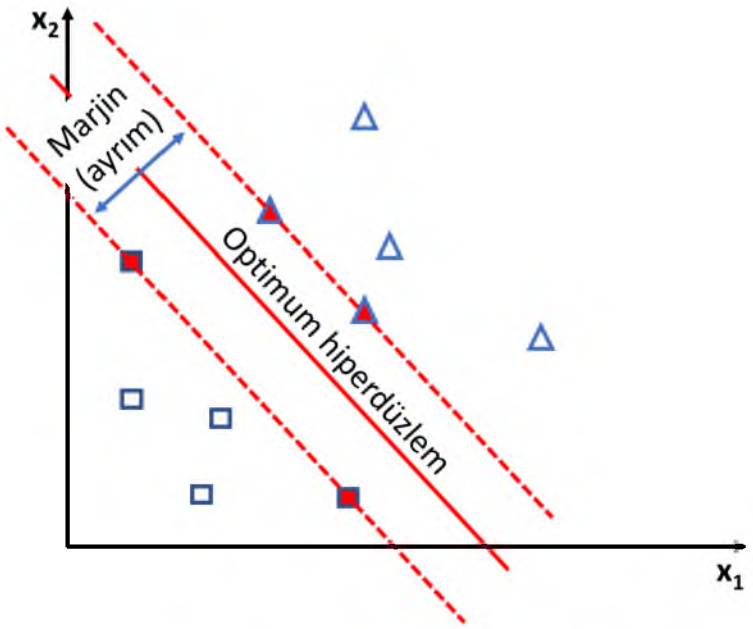
Topluluk makine öğrenmesi algoritmalardan biri olan RO yönteminde, Çantalama (Bagging) yöntemi ile Rastgele Alt-Uzay (Random Sub-Space) yöntemi bir arada kullanılmaktadır (Akman, 2010). Çantalama algoritmasının nedeni, alt karar ağaçlarına

ait eğitim veri setleri seçilirken varyans'ın küçük tutulmasını sağlamaktır. Karar ağaçları oluşturulurken CART (Classification And Regression Tree) algoritması kullanırken, düğümler bölünürken gini index kullanılır.

3.8.2. Destek Vektör Makineleri (DVM)

Destek Vektör Makineleri Vapnik ve arkadaşları tarafından geliştirilen bir yöntemdir (Vapnik, 2013). Şekil 22'da gösterildiği gibi, DVM'ler özellik uzayındaki doğrusal olarak ayrılabilen iki eğitim örneği sınıfını en uygun şekilde ayırmak için kullanılmaktadır (Mavroforakis & Theodoridis, 2006). Yapısal risk minimizasyonu ve istatiksel öğrenme teorisine dayanan bu yöntem uygulaması kolay esnek bir algoritmadır (Özcan, 2020). Bir düzlemde bulunan iki ya da daha fazla grup uygun noktalarda çizilecek sınır çizgileriyle birbirlerinden ayrıştırılabilir. Bu sınır çizgilerinin konumu öyle ayarlanmalıdır ki ayrıştırılacak grup elemanlarına en uzak olan noktalarda bulunmalıdırlar. DVM algoritması bu sınır çizgilerinin konumlarının nasıl tespit edileceği konusunda devreye girmektedir.

DVM'ler denetimli öğrenme sınıfına girmektedir. Sınıflandırma problemleri için yaygın olarak kullanılan bir istatistiksel öğrenme modelidir (Pauly & Sankar, 2015). Verileri analiz edip sınıflandırma için ikili yanıtlar üretirler. Doğrusal DVM, eğitim aşamasında veriler arasında bir hiperdüzlem çizerek yalnızca iki sınıf problemini sınıflandırmaktadır (Chen et al., 2014). Hiperdüzlem, destek vektörleri olarak işlev gören veri noktaları alt kümesiyle karakterize edilmektedir (Kumar & Kishore, 2017). Bu teknik, örnek kümesindeki olumlu örnekleri olumsuz örneklerden ayırmaya çalışan iki sınıfta bir sınıflandırma yapmaktadır. Yöntem daha sonra pozitif örnekleri negatif örneklerden ayıran hiperdüzlemi bulup en yakın pozitif ve negatif arasındaki marjın maksimum olmasını sağlamaktadır (Bougharriou et al., 2017). Şekil 22'da iki boyutlu bir problemde optimum hiperdüzlem ve marjin çizimini göstermektedir.

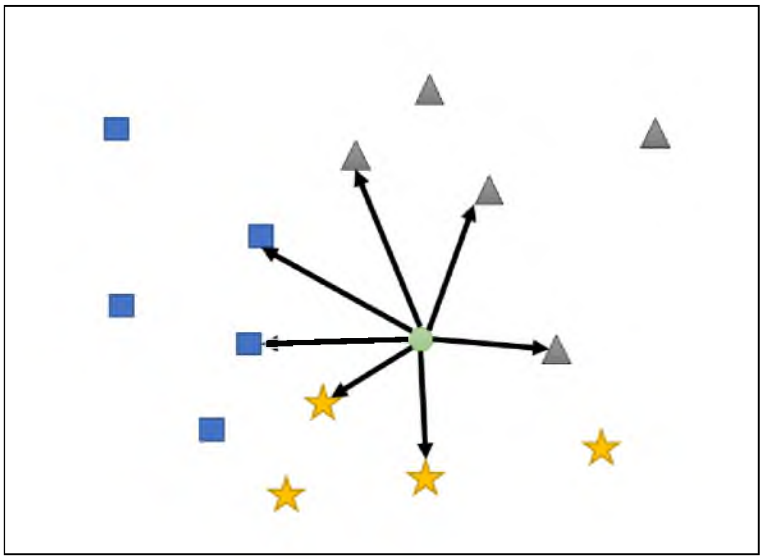


Şekil 22: Doğrusal ayrılabilen optimal hiperdüzlem

DVM'lere iyi bir genelleme performansı, sınıflandırma hatası ve sınıflandırıcıların kapasitesi üzerine sınırlar getirilerek elde edilmektedir (Yau et al., 2008). DVM'ler, sadece doğrusal problemlerde değil ayrıca doğrusal olmayan problemlerde de kullanılabilir. Bu durumda doğrusal bir hiperdüzlem çizilemez. Veriler, DVM eğitimi sırasında doğrusal olmayan haritalama yoluyla daha yüksek boyutlu bir Hilbert uzayına yansıtılmaktadır. Doğrusal olmayan veriler, yüksek boyutlu özellik uzayında çekirdek işlevler kullanılarak doğrusal olarak ayrılabilir. Bu çekirdek işlevlere Lineer, Polinom, RBF (Radial Basic Function) ve KMOD gibi çekirdek yöntemler örnek olarak verilebilir (Ayat et al., 2005).

3.8.3. K-En Yakın Komşu (KNN)

K-en Yakın Komşu algoritması, basit uygulaması ve üstün performansı nedeniyle veri madenciliği ve makine öğrenimi uygulamalarında yaygın olarak kullanılmaktadır (Zhang et al., 2017). KNN yöntemi ilk olarak bir test örneği için en yakın eğitim örneğini seçip ve ardından en yakın eğitim örnekleri arasında ana sınıfla birlikte test örneğini tahmin etmektedir (Deng et al., 2016). Aşağıda Şekil 23'te örnek bir test verisi için sınıflandırma işlemine ait şematik gösterim sunulmuştur.



Şekil 23: KNN sınıflandırma yöntemine ait örnek bir gösterim

Bu işlem basamakları şu şekilde yapılmaktadır. Algoritmada öncelikle bir K değeri belirlenir. Bu değer, verilen bir noktaya en yakın komşuların sayısıdır. Uzaklık fonksiyonları yardımı ile verilen noktanın (yeni veri), mevcut verilere göre tek tek uzaklığı hesaplanır (Özcan, 2020).

3.9. Performans Ölçütleri

Çalışmada kullanılan yaklaşımların performansını ortaya koymak için doğruluk (Accuracy), kesinlik (Precision), duyarlılık/hassasiyet (Recall/Sensitivity), Özgüllük/ Seçicilik (Specificity) ve f-ölçütü gibi performans ölçütleri kullanılmıştır. Bu performans ölçütlerinin hesaplanması için aşağıdaki ifadeler kullanılmaktadır.

$$Doğruluk(Accuracy) = \frac{TP + TN}{TP + TN + FP + FN}$$

$$Kesinlik(Precision) = \frac{TP}{TP + FP}$$

$$Duyarlılık/Hassasiyet\ (Recall/\text{Sensitivity}) = \frac{TP}{TP + FN} \quad (8)$$

$$\text{Özgüllük}\ (Specificity) = \frac{TN}{FP + TN}$$

$$F - \text{Ölçütü}(F - Measure) = 2\frac{Recall * Precision}{Recall + Precision}$$

Bu denklemlerde T, F, P ve N sırasıyla doğruyu, yanlışı, pozitif ve negatif kavramlarını ifade etmektedir. Örneğin, TP doğru sınıflandırılan pozitif örnek sayısını; FN ise yanlış sınıflandırılan negatif örnek sayısını göstermektedir.

Doğruluk (Accuracy): Modelinin başarısını tespit için kullanılan en popüler ve basit yöntemdir. Bu oran doğru sınıflandırılmış (TP+TN) örnek sayısının, toplam örnek sayısına (TP+TN+FP+FN) oranı olarak tanımlanmaktadır.

Kesinlik (Precision): Modelin pozitif olayları pozitif tahmin etmede ne kadar iyi olduğunu ölçer. Pozitif olarak etiketlenen örneklerin sayısının (TP) pozitif olarak sınıflandırılmış toplam örneklere (TP+FP) oranıdır.

Duyarlılık/Hassasiyet (Recall/Sensitivity): Modelin pozitif sınıftaki olayları tespit etmeye ne kadar uygun olduğunu ölçer. Pozitif olarak etiketlenmiş örneklerin (TP) gerçekten pozitif olan örneklerin (TP+FN) toplam sayısına oranıdır. Hassasiyet (Sensitivity) benzer şekilde hesaplanıp, modelin pozitif olayları tespit etmede ne kadar iyi olduğunu ölçer.

Özgüllük (Specificity): Modelin pozitif sınıfa atamada ne kadar doğru olduğunu ölçer. Negtif olarak etiketlenmiş örneklerin (TN) gerçekten negatif olan örneklerin (TN+FP) toplam sayısına oranıdır.

F-Ölçütü: Kesinlik (Precision) ve duyarlılık (Recall) metrikleri birbiriyle ilişkilidir ve ikidi birleştirilerek bunların harmonik ortalamsı olan F-Ölçütünü verirler. Sistemin, kesinlik veya duyarlılık yönüne doğru optimize edilmesinde kullanılmaktadır.

4. ARAŞTIRMA SONUÇLARI VE TARTIŞMA

Her bir veri seti için, RO, DVM ve KNN ile 10 çapraz-doğrulama sonucu elde edilen konuşmacı bağımlı ve konuşmacı bağımsız sonuçlar aşağıdaki tabloda sunulmuştur. Rastgele Orman Bölümleme-kriteri olarak Gini indeks kullanılmıştır. DVM için $power = 1$, $gamma = 1$ ve $bias = 1$ parametre değerlerine sahip polinom kernel kullanılmıştır. KNN için denemeler sonucunda en iyi sonucu verdiğinden $k = 1$ alınmıştır.

Çapraz doğrulama Tabakalı örnekleme (stratified sampling) şeklinde gerçekleştirilmiştir. Tabakalı örnekleme, buradaki problemde olduğu gibi, veri kümesinin alt-gruplara ayrılabildiği ve örnekleme sırasında her alt-grup örneğinin temsil edilmesi gereken durumlarda kullanılan bir örnekleme yöntemidir.

4.1. Konuşmacı bağımlı Sonuçlar

Konuşmacı bağımlı dudak okuma uygulaması, her bir konuşmacıya ait veriler izole edilerek elde edilen alt veri seti kullanılarak elde edilmiştir. Buna göre aşağıdaki Tablo 2 ve Tablo 3'te sırasıyla AVLetters2 ve AVDigits veri setleri için her bir konuşmacının performansları sunulmuştur. Bu tablolarda önerilen MÖU, SÖU ve KİA öznitelik yaklaşımları ile birlikte karşılaştırma amacıyla GLCM tabanlı ROI-P, ROI-M ve ROI-E öznitelikleri ve bunların RO, DVM ve KNN sınıflandırıcılar kullanılarak 10 kat çapraz-doğrulama ile elde edilmiş performans sonuçları tablo halinde sunulmuştur.

İlk uygulama olarak Tablo 2'de, AVLetters2 veri setine ait konuşmacı bağımlı 5 farklı konuşmacı için ayrı ayrı çeşitli yöntemler kullanılarak elde edilen öznitelikler RO, DVM ve KNN sınıflandırıcılarına girilerek ölçülen çeşitli performans parametreleri sunulmuştur.

Tablo 2: AVLetters2 konuşmacı bağımlı başarı performansları (Konuşmacılar K_i, i = 1,2,3,4,5).

Yöntem	K_1	K_2	K_3	K_4	K_5
RO-MÖU	**51,648**	72,527	63,187	**40,659**	39,011
RO-SÖU	48,352	**77,473**	**65,934**	37,912	39,560
RO-KİA	42,857	69,781	60,440	34,066	37,363
RO-ROI-P	37,363	63,187	51,648	35,165	42,308
RO-ROI-M	27,473	41,758	46,154	25,275	27,473
RO-ROI-E	35,165	62,637	61,538	39,560	37,363
DVM-MÖU	43,956	69,231	56,593	36,264	45,055
DVM-SÖU	37,363	65,385	55,495	29,121	42,308
DVM-KİA	38,462	74,176	57,143	39,011	**47,253**
DVM -ROI-P	36,813	54,945	43,407	34,615	39,560
DVM -ROI-M	18,681	40,659	20,879	20,330	24,725

DVM -ROI-E	39,011	46,154	54,945	36,813	35,714
KNN-MÖU	32,418	67,582	53,297	23,077	17,582
KNN-SÖU	36,264	65,934	52,747	25,824	29,121
KNN-KİA	30,769	62,088	50,549	32,418	20,330
KNN -ROI-P	24,725	45,604	39,560	32,967	17,582
KNN -ROI-M	6,044	33,516	14,835	17,582	8,242
KNN -ROI-E	14,835	40,110	47,802	29,670	23,626

Yukarıdaki Tablo 2'de görüldüğü gibi AVLetters2 veri seti için, K_1 ve K_4 konuşmacıları için en iyi yöntemin RO-MÖU olduğu ve sırasıyla 51,648 ve 40,659 başarı performansı elde edilmiştir. K_2 ve K_3 konuşmacıları için ise en iyi yöntemin RO-SÖU olduğu ve sırasıyla 77,473 ve 65,934 başarı performansı elde edilmiştir. Son olarak K_5 konuşmacısı için en iyi başarı performansı DVM-KİA yöntemiyle 47,253 olarak elde edilmiştir.

Sonraki uygulamada ise aşağıdaki Tablo 3'te görüldüğü gibi, AVDigits veri setine ait konuşmacı bağımlı 6 farklı konuşmacı için çeşitli yöntemler kullanılarak elde edilen öznitelikler RO, DVM ve KNN sınıflandırıcılara girilerek ölçülen çeşitli performans parametreleri sunulmuştur.

Tablo 3: AVDigits konuşmacı bağımlı başarı performansları (Konuşmacılar K_i, i = 1,2,3,4,5,6).

Yöntem	K_1	K_2	K_3	K_4	K_5	K_6
RO-MÖU	**84,444**	58,889	72,222	77,778	73,333	54,444
RO-SÖU	68,889	62,222	57,778	63,333	67,778	48,889
RO-KİA	65,556	**67,778**	63,333	65,556	58,889	46,667
RO-ROI-P	72,222	56,667	48,889	63,333	62,222	**63,333**
RO-ROI-M	76,667	46,667	42,222	64,444	63,333	42,222
RO-ROI-E	56,667	51,111	48,889	62,222	50,000	61,111
DVM-MÖU	82,222	57,778	67,778	**84,444**	72,222	56,667
DVM-SÖU	64,444	55,556	61,111	71,111	56,667	47,778
DVM-KİA	78,889	55,556	65,556	83,333	71,111	51,111
DVM -ROI-P	58,889	48,889	44,444	52,222	53,333	58,889
DVM -ROI-M	63,333	32,222	38,889	63,333	55,556	23,333
DVM -ROI-E	45,556	41,111	37,778	46,667	37,778	62,222
KNN-MÖU	81,111	62,222	68,889	83,333	**74,444**	45,556
KNN-SÖU	67,778	58,889	**73,333**	73,333	63,333	55,556
KNN-KİA	66,667	54,444	64,444	78,889	57,778	38,889
KNN -ROI-P	64,444	52,222	36,667	68,889	61,111	56,667
KNN -ROI-M	62,222	37,778	35,556	54,444	48,889	25,556
KNN -ROI-E	47,778	36,667	31,111	61,111	33,333	52,222

Yukarıdaki Tablo 3'de görüldüğü gibi AVDigits veri seti için, K_1 konuşmacısı için RO-MÖU (ve DVM-MÖU, KNN-MÖU), ve K_2 konuşmacısı için RO-KİA (ve KNN-MÖU) yöntemleriyle sırasıyla 84,444 ve 67,778 başarı değerleri elde edilmiştir. K_3

konuşmacısı için KNN-SÖU (ve RO-MÖU), K_4 konuşmacısı DVM-MÖU için (ve DVM-KİA, KNN-MÖU) yöntemleriyle sırasıyla 73,333 ve 84,444 başarı değerleri elde edilmiştir. Son olarak K_5 konuşmacısı için KNN-MÖU (ve RO-MÖU, DVM-MÖU, DVM-KİA) ve K_6 konuşmacı için RO-ROI-P (ve DVM-ROI-E, RO-ROI-E) yöntemleriyle sırasıyla 74,444 ve 63,333 başarı değerleri elde edilmiştir.

4.2.Konuşmacı bağımsız Sonuçlar

Aşağıdaki Tablo 4 ve Tablo 5 de sırasıyla AVLetter2 ve AVDigits veri setleri için farklı öznitelik yöntemleri ve çeşitli sınıflandırıcılarla elde edilmiş konuşmacı bağımsız performanslar sunulmuştur. Bu tablolarda, önerilen MÖU, SÖU ve KİA öznitelik yaklaşımları ile birlikte karşılaştırma amacıyla GLCM tabanlı ROI-P, ROI-M ve ROI-E öznitelikleri ve bunların RO, DVM ve KNN sınıflandırıcılar kullanılarak 10 kat çapraz-doğrulama ile elde edilmiş performans sonuçları tablo halinde sunulmuştur.

İlk olarak aşağıdaki Tablo 4'de AVLetters2 veri setine ait konuşmacı bağımsız, farklı yöntemlerle elde edilen öznitelikler RO, DVM ve KNN gibi sınıflandırıcılara girilerek ölçülen çeşitli performans parametreleri sunulmuştur.

Tablo 4: AVLetters2 konuşmacı bağımsız başarı performansları

Yöntem	Doğruluk	Hata %	Kappa	#Doğru	#Yanlış
RO-MÖU	**45,934**	**54,066**	**0,438**	**418**	**492**
RO-SÖU	43,956	56,044	0,417	400	510
RO-KİA	41,978	58,022	0,397	382	528
RO-ROI-P	39,67	60,33	0,373	361	549
RO-ROI-M	28,462	71,538	0,256	259	651
RO-ROI-E	40,549	59,451	0,382	369	541
DVM-MÖU	31,319	68,681	0,286	285	625
DVM-SÖU	21,978	78,022	0,189	200	710
DVM-KİA	28,462	71,538	0,256	259	651
DVM -ROI-P	26,813	73,187	0,239	244	666
DVM -ROI-M	14,725	85,275	0,113	134	776
DVM -ROI-E	23,516	76,484	0,205	214	696
KNN-MÖU	41,429	58,571	0,391	377	533
KNN-SÖU	41,648	58,352	0,393	379	531
KNN-KİA	40,989	59,011	0,386	373	537
KNN -ROI-P	31,758	68,242	0,29	289	621
KNN -ROI-M	16,154	83,846	0,128	147	763
KNN -ROI-E	31,099	68,901	0,283	283	627

Yukarıdaki Tablo 4'te görüldüğü gibi AVLetters2 için RO-MÖU yöntemi ile elde edilen konuşmacı bağımsız başarı performansı 45,934 olarak bulunmuştur. Buna göre

AVLetters2 veri setinde mevcut 910 örnekten 418 doğru tahmin edilirken 492 örnek yanlış tahmin edilmiştir. Her bir harf tahminine ait detaylı bilgi aşağıdaki karışıklık matrisinde sunulmuştur. Bu başarı performansı, literatürde AVLetters2 veri seti üzerinde yapılan çalışmaların performanslarının üzerinde bir başarıdır.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
A	18	1	0	0	1	2	0	0	2	0	0	0	0	1	0	0	0	3	2	2	1	0	0	1	1	0
B	2	21	1	0	1	0	0	0	0	0	1	0	0	0	0	7	0	0	0	0	0	1	0	0	1	0
C	3	2	6	2	4	0	1	0	1	3	0	0	0	0	0	0	0	0	0	3	1	2	1	1	0	5
D	1	1	4	6	8	0	0	0	0	0	3	1	0	1	0	3	0	0	0	4	0	2	0	0	0	1
E	0	1	1	3	18	0	0	2	0	1	0	1	0	2	0	0	0	0	1	2	0	0	0	1	0	2
F	0	1	0	0	0	27	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	1	0	1
G	1	0	0	1	0	0	15	0	0	14	0	1	0	0	0	0	0	1	0	0	0	1	1	0	0	0
H	1	0	0	1	0	1	1	15	0	0	2	1	0	0	0	0	0	1	4	1	2	0	1	4	0	0
I	2	0	0	2	0	1	0	0	17	0	0	2	0	0	1	0	0	7	1	1	0	0	0	0	0	1
J	1	1	1	0	0	1	13	0	0	12	2	0	0	0	1	0	0	0	0	0	0	1	0	0	1	1
K	2	2	0	2	2	0	1	1	0	1	8	3	0	3	0	0	0	0	0	3	0	1	0	4	0	2
L	1	0	0	0	3	1	1	1	1	0	4	12	2	0	0	0	0	1	3	0	0	0	0	4	0	1
M	0	0	0	0	0	7	0	0	1	1	0	3	20	0	0	0	0	0	0	0	0	1	1	1	0	0
N	2	0	0	0	4	1	0	0	0	0	4	4	1	8	0	2	0	0	4	0	0	0	1	4	0	0
O	0	0	0	0	0	0	0	0	0	0	0	0	1	0	27	0	1	1	0	0	5	0	0	0	0	0
P	0	10	0	0	1	1	0	0	0	0	1	0	0	2	0	13	0	0	1	0	0	3	1	0	2	0
Q	0	0	1	0	0	0	0	0	1	1	0	0	0	0	2	0	12	2	0	0	15	0	1	0	0	0
R	0	0	0	0	0	0	1	0	3	2	0	0	0	0	5	0	0	23	0	0	1	0	0	0	0	0
S	1	2	0	1	2	4	0	3	0	1	0	3	1	6	0	1	0	0	3	0	0	0	1	5	0	1
T	0	0	4	6	4	0	1	2	0	1	2	0	0	1	0	3	0	0	0	8	1	0	0	0	0	2
U	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	6	0	0	0	26	0	0	0	1	0
V	0	2	2	0	2	0	3	0	0	1	1	0	0	0	0	1	0	0	0	0	0	20	1	0	0	2
W	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0	0	1	1	30	0	0	0
X	2	0	0	0	2	1	0	2	0	0	2	1	1	5	0	0	0	0	9	0	0	0	1	9	0	0
Y	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2	0	0	0	1	0	0	31	0
Z	1	0	7	1	2	0	1	0	1	1	2	0	0	1	0	0	0	0	1	2	0	0	2	0	0	13

Şekil 24: AVLetters2 RO-MÖU yöntemi konuşmacı bağımsız karışıklık matrisi

Yukarıdaki Şekil 24'de görüldüğü gibi B-P sesleri, C-D-E-T-Z sesleri, F-M sesleri, G-J sesleri, N-S-X sesleri, Q-U sesleri arasındaki telaffuz benzerliği nedeniyle harflerin kendi arasında karıştırıldığı ve yanlış sınıflandırıldığı görülmektedir. Bu tüm sınıflandırma başarısının da düşmesine neden olmaktadır. Bu seslerin telaffuzu sırasındaki dudak hareketleri oldukça benzer olduğundan elde edilen uzamsal öznitelikler de benzer olduğu düşünülmektedir. Aşağıda Tablo 5'te her bir harf için detaylı başarı istatistikleri verilmiştir.

Tablo 5: AVLetters2 RO-MÖU konuşmacı bağımsız doğruluk istatistikleri

	TP	FP	TN	FN	Duyarlılık (Recall)	Kesinlik (Precision)	Hassasiyet (Sensitivity)	Özgüllük (Specificity)	F-ölçütü
A	18	20	855	17	0,514	0,474	0,514	0,977	0,493
B	21	23	852	14	0,600	0,477	0,600	0,974	0,532
C	6	21	854	29	0,171	0,222	0,171	0,976	0,194
D	6	19	856	29	0,171	0,240	0,171	0,978	0,200
E	18	36	839	17	0,514	0,333	0,514	0,959	0,404
F	27	20	855	8	**0,771**	**0,574**	**0,771**	**0,977**	**0,659**
G	15	23	852	20	0,429	0,395	0,429	0,974	0,411
H	15	11	864	20	0,429	0,577	0,429	0,987	0,492
I	17	10	865	18	0,486	0,630	0,486	0,989	0,548
J	12	29	846	23	0,343	0,293	0,343	0,967	0,316
K	8	24	851	27	0,229	0,250	0,229	0,973	0,239
L	12	20	855	23	0,343	0,375	0,343	0,977	0,358
M	20	11	864	15	0,571	0,645	0,571	0,987	0,606
N	8	22	853	27	0,229	0,267	0,229	0,975	0,246
O	27	10	865	8	**0,771**	**0,730**	**0,771**	**0,989**	**0,750**
P	13	19	856	22	0,371	0,406	0,371	0,978	0,388
Q	12	8	867	23	0,343	0,600	0,343	0,991	0,436
R	23	18	857	12	0,657	0,561	0,657	0,979	0,605
S	3	26	849	32	0,086	0,103	0,086	0,970	0,094
T	8	18	857	27	0,229	0,308	0,229	0,979	0,262
U	26	27	848	9	0,743	0,491	0,743	0,969	0,591
V	20	14	861	15	0,571	0,588	0,571	0,984	0,580
W	30	12	863	5	**0,857**	**0,714**	**0,857**	**0,986**	**0,779**
X	9	26	849	26	0,257	0,257	0,257	0,970	0,257
Y	31	6	869	4	**0,886**	**0,838**	**0,886**	**0,993**	**0,861**
Z	13	19	856	22	0,371	0,406	0,371	0,978	0,388

Yukarıdaki Tablo 5 incelendiğinde RO-MÖU konuşmacı bağımsız sonuçlar, en başarılıdan daha az başarılıya doğru sırasıyla Y, W, O ve F gibi harflerin oldukça fazla doğrulukta ve oldukça düşük yanlışla tahmin edilebildiği göstermektedir. Diğer taraftan S, C, D, K, N, X ve T gibi harflerin ise oldukça düşük oranda doğru tahmin edildiği aksine yanlış tahmin sayısının oldukça yüksek olduğu görülmektedir.

Son olarak aşağıdaki Tablo 6'da AVDigits veri setine ait konuşmacı bağımsız farklı yöntemlerle elde edilen öznitelikler RO, DVM ve KNN gibi sınıflandırıcılara girilerek ölçülen çeşitli performans parametreleri sunulmuştur.

Tablo 6: AVDigits konuşmacı bağımsız başarı performansları

Yöntem	Doğruluk	Hata %	Kappa	#Doğru	#Yanlış
RO-MÖU	58,519	41,481	0,539	316	224
RO-SÖU	55,741	44,259	0,508	301	239
RO-KİA	59,630	40,370	0,551	322	218
RO-ROI-P	55,185	44,815	0,502	298	242
RO-ROI-M	46,296	53,704	0,403	250	290
RO-ROI-E	43,519	56,481	0,372	235	305
DVM-MÖU	46,667	53,333	0,407	252	288
DVM-SÖU	32,593	67,407	0,251	176	364
DVM-KİA	38,333	61,667	0,315	207	333
DVM -ROI-P	40,926	59,074	0,344	221	319
DVM -ROI-M	27,407	72,593	0,193	148	392
DVM -ROI-E	21,481	78,519	0,128	116	424
KNN-MÖU	**67,407**	**32,593**	**0,638**	**364**	**176**
KNN-SÖU	62,222	37,378	0,580	336	204
KNN-KİA	57,963	42,037	0,533	313	227
KNN -ROI-P	53,148	46,852	0,479	287	253
KNN -ROI-M	39,074	60,926	0,323	211	329
KNN -ROI-E	42,963	57,037	0,366	232	308

Yukarıdaki Tablo 6'da görüldüğü gibi AVDigits için KNN-MÖU yöntemi ile elde edilen konuşmacı bağımsız başarı performansı 67,407 olarak bulunmuştur. Buna göre AVDigits veri setinde mevcut 540 örnekten 364 doğru tahmin edilirken 176 örnek yanlış tahmin edilmiştir. Her bir rakam tahminine ait detaylı bilgi aşağıdaki karışıklık matrisinde sunulmuştur. Bu başarı performansı, literatürde AVDigits veri seti üzerinde yapılan çalışmaların performanslarının üzerinde bir başarıdır.

	0	1	2	3	4	5	6	7	8	9
0	50	0	0	1	2	0	1	0	0	0
1	4	32	6	5	3	3	0	0	0	1
2	3	1	46	0	3	0	1	0	0	0
3	7	1	1	33	4	0	4	3	0	1
4	4	0	5	3	37	1	2	2	0	0
5	2	1	2	1	3	32	3	3	1	6
6	3	0	0	5	1	0	32	6	4	3
7	3	0	0	4	0	0	6	32	6	3
8	1	0	0	4	0	0	4	3	38	4
9	1	1	0	0	0	2	5	6	7	32

Şekil 25: AVDigits KNN-MÖU konuşmacı bağımsız karışıklık matrisi

Yukarıdaki Şekil 25'te görüldüğü gibi AVDigits veri seti için, en yüksek doğruluk sıfır (0) ve daha sonra iki (2) rakamı için elde edildiği görülmektedir. Aşağıdaki Tablo 7'deki başarı istatistikleri tablosunda bu iki rakam için FN değerleri sırasıyla 4 ve 8 dir. Diğer rakamların ise çeşitli yanlış rakamlarla karıştırıldığı ve yanlış sınıflandırıldığı görülmektedir. Bu rakamların Tablo 7'deki FN değerlerinin daha yüksek olduğu ve 16-22 değerleri arasında değiştiği görülmektedir. Bu tüm sınıflandırma başarısının da düşmesine neden olmaktadır. Aşağıda Tablo 7'de her bir rakam için detaylı başarı istatistikleri verilmiştir.

Tablo 7: AVDigits KNN-MÖU konuşmacı bağımsız doğruluk istatistikleri

	TP	FP	TN	FN	Duyarlılık (Recall)	Kesinlik (Precision)	Hassasiyet (Sensitivity)	Özgüllük (Specificity)	F-ölçütü
0	50	28	458	4	**0,926**	**0,641**	**0,926**	**0,942**	**0,758**
1	32	4	482	22	0,593	0,889	0,593	0,992	0,711
2	46	14	472	8	**0,852**	**0,767**	**0,852**	**0,971**	**0,807**
3	33	23	463	21	0,611	0,589	0,611	0,953	0,600
4	37	16	470	17	0,685	0,698	0,685	0,967	0,692
5	32	6	480	22	0,593	0,842	0,593	0,988	0,696
6	32	26	460	22	0,593	0,552	0,593	0,947	0,571
7	32	23	463	22	0,593	0,582	0,593	0,953	0,587
8	38	18	468	16	0,704	0,679	0,704	0,963	0,691
9	32	18	468	22	0,593	0,640	0,593	0,963	0,615

Yukarıdaki Tablo 7 incelendiğinde KNN-MÖU konuşmacı bağımsız sonuçlar, 0 ve 2 rakamlarının oldukça yüksek doğrulukta ve oldukça düşük yanlışla tahmin edilebildiği göstermektedir. Bu rakamları yüksek başarı performansından düşük performansa sırasıyla 8, 4 ve 3 rakamlarının takip ettiği görülmektedir. Diğer geri kalan rakamların ise nispeten düşük doğrulukta tahin edildiği görülmektedir.

Yukarıda listelenen sonuçlar literatürde var olan çalışmalarla karşılaştırıldığında önerilen öznitelik çıkarım yöntemlerinin yalnızca-görüntü kullanılarak gerçekleştirilen dudak okuma uygulamalarında oldukça başarılı olduğu görülmüştür. Aşağıda Tablo 8'de bu tez çalışmasında önerilen öznitelik çıkarım yaklaşımlarının diğer daha önce yapılmış çalışmalarda elde edilen sonuçlarla kıyaslaması yapılmak üzere listelenmiştir.

4.3. Sonuçların karşılaştırılması

Bu çalışmada önerilen uzamsal öznitelik yaklaşımları ile elde edilen yalnızca-görüntü uygulamalarına ait başarı oranlarının mevcut çalışmalarla kıyaslaması aşağıdaki Tablolarda sunulmuştur.

AVLetters2 veri setin ile ilgili uygulanan yöntemler ve bu yöntemlerin konuşmacı bağımsız başarı oranları Tablo 6'da listelenmiştir.

Tablo 8: AVLetters2 konuşmacı bağımsız başarı performans karşılaştırmaları

Çalışma	Öznitelik	Sınıflandırıcı	Doğruluk
Cox vd. (Cox et al., 2008)	Aktif Görünüm Modelleri (AGM)	Modeli Gizli Markov (GMM)	% 8,3
Liu vd. (Liu et al., 2015)	Geometrik Tabanlı Şekil Değişmez Özellikler (geo-SIF)	Bayes Kombinasyon Füzyonu	**% 64,38**
Frisky vd. (Frisky et al., 2015)	Yerel İkili Örüntü-Üç Ortogonal Düzlem (YİÖ-ÜOD) XYXTYT Özellik Düzlemi	Çekirdek Seyrek Gösterim Sınıflandırıcısı (Ç-SGS)	% 25,90
Frisky vd. (Frisky et al., 2015)	Yerel İkili Örüntü-Üç Ortogonal Düzlem (YİÖ -ÜOD) XTYT Özellik Düzlemi	Çekirdek Seyrek Gösterim Sınıflandırıcısı (Ç-SGS)	% 24,2
Hu vd. (Hu et al., 2016)	Tekrarlayan Temporal Multimodal Sınırlı Boltzmann Makineleri (RTMRBM)		% 31,21
Petridis vd. (Petridis et al., 2020)	Ham dudak bölgesi (raw-ROI) Geri beslemeli Network bottleneck özellikler	B-LSTM	% 42,6
Petridis vd. (Petridis et al., 2020)	Fark dudak bölgesi (dif-ROI) Geri beslemeli Network bottleneck özellikler	B-LSTM	% 32,2
Petridis vd. (Petridis et al., 2020)	Ham ve fark dudak bölgesi (raw+dif-ROI) Geri beslemeli Network bottleneck özellikler	B-LSTM	% 37,8
Bu çalışma (RO-MÖU)	Merkezi Öklit-Uzaklık (MÖU)	Rastgele Orman (RO)	**% 45,934**

Yukarıdaki Tablo 8'de AVLetters2 veri seti için, konuşmacı bağımsız yapılan çalışmaların başarı performansları kıyaslandığında, RO-MÖU ile elde edilen çalışmadaki doğruluk oranı, bu alanda yapılan çalışmalarla kıyaslandığında birçoğundan yüksek başarı oranına sahip olduğu görülmektedir. Sadece Liu ve arkadaşlarının yaptığı çalışmanın başarı oranı, RO-MÖU ile elde edilen başarı oranından yüksek olduğu görülmektedir. Buda yapılan çalışmada ortalama literatür başarısının üzerinde bir başarı elde edildiğini göstermektedir.

Benzer şekilde AVDigits veri setin ile ilgili uygulanan yöntemler ve bu yöntemlerin konuşmacı bağımsız başarı oranları Tablo 9'da listelenmiştir.

Tablo 9: AVDigits konuşmacı bağımsız başarı performans karşılaştırmaları

Çalışma	Öznitelik	Sınıflandırıcı	Doğruluk
Hu vd. (Hu et al., 2016)	Tekrarlayan Temporal Multimodal Sınırlı Boltzmann Makineleri (RTMRBM)		% 40,66
(Srivastava & Salakhutdinov, 2012)	Multimodal Derin Boltzmann Makineleri (MDBM)		%55
(Ngiam et al., 2011)	Multimodal Derin Otomatik Kodlayıcılar (MDAE)		%66.74
Bu çalışma (KNN-MÖU)	Merkezi Öklit-Uzaklık (MÖU)	K-En Yakın Komşu (KNN)	**% 67,407**

Yukarıdaki Tablo 9'da AVDigits veri seti için, konuşmacı bağımsız yapılan çalışmaların başarı performansları kıyaslandığında, KNN-MÖU ile elde edilen çalışmadaki doğruluk oranı, bu alanda yapılan çalışmalarla kıyaslandığında en yüksek başarı oranına sahip olduğu görülmektedir.

AVLetters2 ve AVDigits veri setleriyle ilgili uygulanan yöntemler ve bu yöntemlerin konuşmacı bağımlı başarı oranları Tablo 10'da listelenmiştir.

Tablo 10: AVLetters2 ve AVDigits konuşmacı bağımlı başarı performans karşılaştırmaları

Veri Seti	Çalışma	Öznitelik	Sınıflandırıcı	Doğruluk
AVLetters2	Huang ve Kingsbury (Huang & Kingsbury, 2013)	Çok Kanallı Derin Ağlar (MDBN)		%54,10
	Pei vd. (Pei et al., 2013)	ROMH		**% 91,80**
	Frisky vd. (Frisky et al., 2015)	Yerel İkili Örüntü-Üç Ortogonal Düzlem (YİÖ-ÜOD) XYXTYT Özellik Düzlemi	Çekirdek Seyrek Gösterim Sınıflandırıcısı (Ç-SGS)	**%87,12**
	Frisky vd. (Frisky et al., 2015)	Yerel İkili Örüntü-Üç Ortogonal Düzlem (YİÖ-ÜOD) XTYT Özellik Düzlemi	Çekirdek Seyrek Gösterim Sınıflandırıcısı (Ç-SGS)	**% 89,02**
	Bear ve Harvey (Bear & Harvey, 2017)	fonem-vizem esaslı Aktif Görünüm Modelleri (AGM)	Gizli Markov Model (GMM)	% 36,53
	Bu çalışma	Simetrik Öklit Uzaklığı (SÖU)	Rastgele Orman (RO)	**%77,473**
AVDigits	Bu çalışma	Merkezi Öklit Uzaklığı (MÖU)	Rastgele Orman (RO)	**%84,44**
		Merkezi Öklit Uzaklığı (MÖU)	Destek Vektör Makineleri (DVM)	**%84,44**

Yukarıda Tablo 10'da ise AVLetters2 ve AVDigits için konuşmacı bağımlı yapılan çalışmaların başarı performanslarındaki oranlar görülmektedir. Bizim çalışmamızda AVLetter veriseti için RO-SÖU yönteminde ve AVDigits veri seti için ise RO-MÖU ve DVM-MÖU yöntemlerinde en yüksek başarı oranları elde edildiği ve literatür başarısına yakın performans değerleri elde edildiği görülmektedir.

Literatürde yapılan diğer benzer veri setleri üzerinde yapılan çalışmaları incelendiğinde Zhao ve arkadaşlarının harf içeriğine sahip AVletter adlı veri seti ile yaptıkları çalışmada %43.5 başarı elde etmişlerdir (Z. Zhou et al., 2014). Christoph Bregler ve arkadaşlarının yaptığı, harf ve kelime düzeyinde dudak okuma içeren çalışmalarında, harf düzeyinde yaptıkları çalışmada %50.2 başarı elde etmişlerdir

(Bregler et al., 1993). Çalışmamızda kullandığımız veri setinden farklı harf düzeyinde veri setleri ile yapılan çalışmalarla bizim harf düzeyinde yaptığımız çalışmalar kıyaslandığında kayda değer bir başarı elde ettiğimiz görülmektedir. Elde ettiğimiz başarı oranı olan %77,473, bu çalışmalardaki başarıdan daha yüksek olduğu görülmektedir. Fu ve arkadaşlarının rakam içeriğine sahip AVICAR adlı veri seti ile yaptıkları çalışmada %37.9 başarı elde etmişlerdir (Fu et al., 2008). Vanegas ve arkadaşlarının rakam içeriğine sahip M2VTS adlı veri seti ile yaptığı çalışmada ise %74.5 başarı elde edilmiştir (Vanegas et al., 1999). Çalışmamızda kullandığımız veri setinden farklı rakam düzeyinde veri setleri ile yapılan çalışmalarla bizim rakam düzeyinde yaptığımız çalışmalar kıyaslandığında burada da kayda değer bir başarı elde edildiği görülmektedir. Elde ettiğimiz başarı oranı olan %84,44, bu çalışmalardaki başarıdan çok daha yüksek olduğu görülmektedir. Bu tez çalışmasında kullanılan veri setleri üzerinde literatürde yapılan bu çalışmalarla kıyaslandığında önerilen yöntemlerin çok daha başarılı olduğu görülmüştür.

5. SONUÇLAR VE ÖNERİLER

5.1. Sonuçlar

Bu tez çalışmasında, oldukça zor problemlerden bir olan ve sadece-görüntü verilerinin kullanıldığı dudak okuma problemine yönelik üç farklı uzamsal öznitelik yaklaşımı önerilmiştir. Bu öznitelik yaklaşımlarından elde edilen öznitelikler RO, DVM ve KNN yöntemleriyle sınıflandırılarak dudak hareketlerinden video kaydında telaffuz edilen harf tahmin edilmeye çalışılmıştır.

Görüntüler kayıtlarındaki her bir çerçevede öncelikle yüz ve daha sonra yüzdeki odak noktamız olan dudak sınırları tespit edilerek, bu sınırlarda yer alacak ve dudak okuma sırasında hareketleri izlenecek işaret noktaları en doğru şekilde yerleştirilmeye çalışılmıştır. Bu işaret noktalarının, her bir telaffuza ait video kaydında sahip oldukları değişimler uzamsal olarak ölçümlenerek üç farklı yaklaşımla öznitelik seti oluşturulmuştur.

Öznitelik setleri RO, DVM ve KNN sınıflandırıcılarla sınıflandırılarak sadece-görüntü ile dudak okuma gerçekleştirilmeye çalışılmıştır. Bu sınıflandırıcılarla ve öznitelik yaklaşımlarıyla AVLetters2 için konuşmacı bağımsız elde edilen en yüksek başarı oranları sırasıyla RO-MÖU=45,934, DVM-MÖU=31,319 ve KNN-SÖU=41,648 olarak elde edilmiştir. Bu sonuçlara göre MÖU öznitelik yaklaşımı ve RO sınıflandırıcının AVLetters2 veri seti için en başarılı yöntem olduğu görülmüştür. AVDigits için ise konuşmacı bağımsız elde edilen en yüksek başarı oranları sırasıyla RO-KİA=59,630, DVM-MÖU=46,667 ve KNN-MÖU=67,407 olarak elde edilmiştir. Bu sonuçlara göre MÖU öznitelik yaklaşımı ve KNN sınıflandırıcının AVDigits veri seti en başarılı yöntem olduğu görülmüştür.

Konuşmacı bağımlı başarı oranlarında ise konuşmacıdan konuşmacıya en iyi yöntem değişkenlik göstermiştir. Buna göre AVLetters2 için, 1 ve 4 ID'li konuşmacılar da en yüksek başarı RO-MÖU yöntemiyle sırasıyla 51,648 ve 40,659 olarak elde edilmiştir. Diğer taraftan, 2 ve 3 ID'li konuşmacılar da en yüksek başarı RO-SÖU yöntemiyle sırasıyla 77,473 ve 65,934 olarak elde edilmiştir. Son olarak 5 ID'li konuşmacı için ise en yüksek başarı DVM-KİA yöntemiyle 47,253 olarak elde edilmiştir. AVDigits için ise en yüksek başarı oranları, 1, 2, 3, 4, 5, 6 ID'li konuşmacılarda sırasıyla RO-MÖU ile 84,444, RO-KİA ile 67,778, DVM-MÖU ile 84,444, KNN-MÖU ile 74,444 ve RO-ROI-P ile 63,333 olarak elde edilmiştir.

Farklı konuşmacılar için en yüksek başarı oranlarının büyük değişkenlik göstermesi ve farklı yöntemlerin (öznitelik+sınıflandırıcı) farklı kullanıcılarda daha iyi performanslar sergilemesi, başarı performansı üzerinde hem öznitelik yaklaşımlarının hem de konuşmacıların oldukça etkili olduğunu göstermektedir. Konuşmacıların farklı dudak morfolojileri ve farklı aksan ya da mimiklere sahip oldukları düşünüldüğünde başarı oranlarının konuşmacı bağımlı olması beklenen bir durum olarak düşünülebilir. Fakat bununla birlikte elde edilen sonuçlar önerilen öznitelik yaklaşımlarının da başarı üzerinde önemli bir etkisi olduğu görülmektedir.

5.2. Öneriler

Sadece-Görsel dudak okuma uygulamalarında önerilen öznitelik yaklaşımlarının ve çalışmada kullanılan sınıflandırıcı yöntemin başarılı bir şekilde kullanılabileceği düşünülmektedir. Bu çalışmanın ileriki safhalarında veri seti daha da genişletilerek konuşmacı bağımlılığı düşürülebilir. Ayrıca sadece harf içeren veri setleri yanında kelime ya da cümle içeren veri setleri de dâhil edilerek eğitim kümesi genişletilebilir.

Kelime veya cümle içeren veri setleri kullanılırken, sistemi eğitmek için video kaydında harf bölütleme ve harf tespiti konusuyla ilgili ayrı çalışmalar yapılabilir. Bu sayede eğitim seti zenginleştirilerek sistem daha güçlü hale getirilebilir.

Ayrıca benzer yöntemler dil tanıma ya da konuşmacı problemlerine de uygulanabilir. Bunun yanında farklı diller için özgün veri setleri oluşturularak bu diller özelinde de dudak okuma çalışmaları yapılabilir.

KAYNAKLAR

Akman, M. (2010). Veri madenciliğine genel bakış ve random forests yönteminin incelenmesi: sağlık alanında bir uygulama. *Unpublished Master's Thesis). Ankara University, Ankara*.

Ayat, N.-E., Cheriet, M., & Suen, C. Y. (2005). Automatic model selection for the optimization of SVM kernels. *Pattern Recognition*, *38*(10), 1733–1745.

Bayrakdar, S., Akgün, D., & Yücedağ, İ. (2016). Yüz ifadelerinin otomatik analizi üzerine bir literatür çalışması. *Sakarya Üniversitesi Fen Bilimleri Enstitüsü Dergisi*, *20*(2), 383–398.

Bear, H. L., Cox, S. J., & Harvey, R. W. (2017). Speaker-independent machine lip-reading with speaker-dependent viseme classifiers. *ArXiv Preprint ArXiv:1710.01122*.

Bear, H. L., & Harvey, R. (2017). Phoneme-to-viseme mappings: the good, the bad, and the ugly. *Speech Communication*, *95*, 40–67.

Bougharriou, S., Hamdaoui, F., & Mtibaa, A. (2017). Linear SVM classifier based HOG car detection. *2017 18th International Conference on Sciences and Techniques of Automatic Control and Computer Engineering (STA)*, 241–245.

Bregler, C., Hild, H., Manke, S., & Waibel, A. (1993). Improving connected letter recognition by lipreading. *1993 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, 557–560.

Bregler, C., & Omohundro, S. M. (1994). Surface learning with applications to lipreading. *Advances in Neural Information Processing Systems*, 43–50.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Chen, J., Chen, Z., Chi, Z., & Fu, H. (2014). Facial expression recognition based on facial components detection and hog features. *International Workshops on Electrical and Computer Engineering Subfields*, 884–888.

Chiou, G. I., & Hwang, J.-N. (1996). Lipreading by using snakes, principal component analysis, and hidden Markov models to recognize color motion video. *IEEE Transactions on Image Processing*.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *23*(6), 681–685.

Cox, S. J., Harvey, R. W., Lan, Y., Newman, J. L., & Theobald, B.-J. (2008). The

challenge of multispeaker lip-reading. *AVSP*, 179–184.

Cuimei, L., Zhiliang, Q., Nan, J., & Jianhua, W. (2017). Human face detection algorithm via Haar cascade classifier combined with three additional classifiers. *2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, 483–487.

Culjak, I., Abram, D., Pribanic, T., Dzapo, H., & Cifrek, M. (2012). A brief introduction to OpenCV. *2012 Proceedings of the 35th International Convention MIPRO*, 1725–1730.

Dadi, H. S., & Pillutla, G. K. M. (2016). Improved face recognition rate using HOG features and SVM classifier. *IOSR Journal of Electronics and Communication Engineering*, *11*(4), 34–44.

Dahmane, M., & Meunier, J. (2011). Emotion recognition using dynamic grid-based HoG features. *Face and Gesture 2011*, 884–888.

Dalal, N., & Triggs, B. (2005). Histograms of oriented gradients for human detection. *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, *1*, 886–893.

Deng, Z., Zhu, X., Cheng, D., Zong, M., & Zhang, S. (2016). Efficient kNN classification algorithm for big data. *Neurocomputing*, *195*, 143–148.

Fernandez-Lopez, A., & Sukno, F. M. (2018). Survey on automatic lip-reading in the era of deep learning. *Image and Vision Computing*, *78*, 53–72.

Frisky, A. Z. K., Wang, C.-Y., Santoso, A., & Wang, J.-C. (2015). Lip-based visual speech recognition system. *2015 International Carnahan Conference on Security Technology (ICCST)*, 315–319.

Fu, Y., Yan, S., & Huang, T. S. (2008). Classification and feature extraction by simplexization. *IEEE Transactions on Information Forensics and Security*, *3*(1), 91–100.

Gaidar, A. I., & Yakimov, P. Y. (2019). Real-time fatigue features detection. *Journal of Physics: Conference Series*, *1368*(5), 52017.

Goldschen, A. J., Garcia, O. N., & Petajan, E. (1994). Continuous optical automatic speech recognition by lipreading. *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, *1*, 572–577.

Gower, J. C. (1985). Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and Its Applications*, *67*, 81–97.

Graves, A., Fernández, S., & Schmidhuber, J. (2005). Bidirectional LSTM networks for

improved phoneme classification and recognition. *International Conference on Artificial Neural Networks*, 799–804.

Haralick, R. M., Shanmugam, K., & Dinstein, I. H. (1973). Textural features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, *6*, 610–621.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Histogram of oriented gradients – Wikipedia. (n.d.). Retrieved January 14, 2021, from https://de.wikipedia.org/wiki/Histogram_of_oriented_gradients

Hu, D., Li, X., & Lu, X. (2016). Temporal multimodal learning in audiovisual speech recognition. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3574–3582.

Huang, J., & Kingsbury, B. (2013). Audio-visual deep learning for noise robust speech recognition. *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 7596–7599.

Jun, H., & Hua, Z. (2007). A real time lip detection method in lipreading. *2007 Chinese Control Conference*, 516–520.

Kang, D. W., Choi, J. S., Bae, J. H., Shin, Y. H., Lee, J. H., Choi, J. B., & Tack, G. R. (2014). Recognition of Korean Monosyllabic Speech Using 3D Facial Motion Data. *The 15th International Conference on Biomedical Engineering*, 569–572.

Kawasaki, T., Ukai, N., Takumi, S., Tamura, S., & Hayamizu, S. (2013). Improvement of Lip Reading Performance in Real Environments Using Speaker and Environmental Adaptation. *2013 2nd IAPR Asian Conference on Pattern Recognition*, 346–350.

Kazemi, V., & Sullivan, J. (2014). One millisecond face alignment with an ensemble of regression trees. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1867–1874.

King, D. E. (2009). Dlib-ml: A machine learning toolkit. *The Journal of Machine Learning Research*, *10*, 1755–1758.

Kulkarni, A. H., & Kirange, D. (2019). Artificial intelligence: a survey on lip-reading techniques. *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, 1–5.

Kumar, K. V. V, & Kishore, P. V. V. (2017). Indian classical dance mudra classification using HOG features and SVM classifier. *International Journal of Electrical & Computer Engineering (2088-8708)*, *7*(5).

KURUDAYIOĞLU, A. G. M. (2003). Konuşma eğimi ve konuşma becerisini geliştirmeye yönelik etkinlikler. *Türklük Bilimi Araştırmaları*, *13*, 287.

Lee, J.-S. (2014). Visual-speech-pass filtering for robust automatic lip-reading. *Pattern Analysis and Applications*, *17*(3), 611–621.

Liévin, M., & Luthon, F. (2000). A hierarchical segmentation algorithm for face analysis. application to lipreading. *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532)*, *2*, 1085–1088.

Liu, H., Zhang, X., & Wu, P. (2015). Regression based landmark estimation and multi-feature fusion for visual speech recognition. *2015 IEEE International Conference on Image Processing (ICIP)*, 808–812.

Lu, J., Zhang, X., Yang, X., & Unwala, I. (2019). Intelligent in-vehicle safety and security monitoring system with face recognition. *2019 IEEE International Conference on Computational Science and Engineering (CSE) and IEEE International Conference on Embedded and Ubiquitous Computing (EUC)*, 225–229.

Luettin, J. (1997). *Visual speech and speaker recognition*. University of Sheffield.

Lv, Z., Penades, V., Blasco, S., Chirivella, J., & Gagliardo, P. (2016). Evaluation of Kinect2 based balance measurement. *Neurocomputing*, *208*, 290–298.

Mallick, S. (2016). Histogram of oriented gradients. *Learn OpenCV*, *6*.

Mase, K., & Pentland, A. (1991). Automatic lipreading by optical-flow analysis. *Systems and Computers in Japan*, *22*(6), 67–76.

Matthews, I., Cootes, T. F., Bangham, J. A., Cox, S., & Harvey, R. (2002). Extraction of visual features for lipreading. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *24*(2), 198–213.

Mattos, A. B., Oliveira, D. A. B., & da Silva Morais, E. (2018). Improving CNN-based viseme recognition using synthetic data. *2018 IEEE International Conference on Multimedia and Expo (ICME)*, 1–6.

Mavroforakis, M. E., & Theodoridis, S. (2006). A geometric approach to support vector machine (SVM) classification. *IEEE Transactions on Neural Networks*, *17*(3), 671–682.

McGurk, H., & MacDonald, J. (1976). Hearing lips and seeing voices. *Nature*, *264*(5588), 746–748.

Morade, S. S., & Patnaik, S. (2014). Lip reading using dwt and lsda. *2014 IEEE International Advance Computing Conference (IACC)*, 1013–1018.

Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). *Multimodal deep learning*.

Ondimu, S. N., & Murase, H. (2008). Effect of probability-distance based Markovian texture extraction on discrimination in biological imaging. *Computers and Electronics in Agriculture*, *63*(1), 2–12.

Özcan, T. (2020). Yığınlanmış Özdevinimli Kodlayıcılar ile Göğüs Kanserinin Sınıflandırılması ve Klasik Makine Öğrenme Metotları ile Performans Karşılaştırması. *Erciyes Üniversitesi Fen Bilimleri Enstitüsü Fen Bilimleri Dergisi*, *36*(2), 151–160.

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. https://doi.org/10.1080/01431160412331269698

Paré, M., Richler, R. C., ten Hove, M., & Munhall, K. G. (2003). Gaze behavior in audiovisual speech perception: The influence of ocular fixations on the McGurk effect. *Perception & Psychophysics*, *65*(4), 553–567.

Pauly, L., & Sankar, D. (2015). Detection of drowsiness based on HOG features and SVM classifiers. *2015 IEEE International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, 181–186.

Pei, Y., Kim, T.-K., & Zha, H. (2013). Unsupervised random forest manifold alignment for lipreading. *Proceedings of the IEEE International Conference on Computer Vision*, 129–136.

Petajan, E., Bischoff, B., Bodoff, D., & Brooke, N. M. (1988). An improved automatic lipreading system to enhance speech recognition. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 19–25.

Petridis, S., Wang, Y., Ma, P., Li, Z., & Pantic, M. (2020). End-to-end visual speech recognition for small-scale datasets. *Pattern Recognition Letters*, *131*, 421–427.

Potamianos, G., Graf, H. P., & Cosatto, E. (1998). An image transform approach for HMM based automatic lipreading. *Proceedings 1998 International Conference on Image Processing. ICIP98 (Cat. No. 98CB36269)*, 173–177.

Potamianos, G., Neti, C., Gravier, G., Garg, A., & Senior, A. W. (2003). Recent advances in the automatic recognition of audiovisual speech. *Proceedings of the IEEE*, *91*(9), 1306–1326.

Potamianos, G., Verma, A., Neti, C., Iyengar, G., & Basu, S. (2000). A cascade image transform for speaker independent automatic speechreading. *2000 IEEE*

International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No. 00TH8532), *2*, 1097–1100.

Rao, R. A., & Mersereau, R. M. (1994). Lip modeling for visual speech recognition. *Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers*, *1*, 587–590.

Sak, H., Senior, A. W., & Beaufays, F. (2014). *Long short-term memory recurrent neural network architectures for large scale acoustic modeling*.

Sharifara, A., Rahim, M. S. M., & Anisi, Y. (2014). A general review of human face detection including a study of neural networks and Haar feature-based cascade classifier in face detection. *2014 International Symposium on Biometrics and Security Technologies (ISBAST)*, 73–78.

Silsbee, P. L., & Bovik, A. C. (1996). Computer lipreading for improved accuracy in automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, *4*(5), 337–351.

Singh, S., Singh, D., & Yadav, V. (2020). Face Recognition Using HOG Feature Extraction and SVM Classifier. *International Journal*, *8*(9).

Srivastava, N., & Salakhutdinov, R. (2012). Learning representations for multimodal data with deep belief nets. *International Conference on Machine Learning Workshop*, *79*, 3.

Stork, D. G., Wolff, G., & Levine, E. (1992). Neural network lipreading system for improved speech recognition. *[Proceedings 1992] IJCNN International Joint Conference on Neural Networks*, *2*, 289–295.

Sumby, W. H., & Pollack, I. (1954). Visual contribution to speech intelligibility in noise. *The Journal of the Acoustical Society of America*, *26*(2), 212–215.

The Illusions Index - The Illusions Index. (n.d.). Retrieved January 14, 2021, from https://www.illusionsindex.org/index.php?option=com_content&view=article&id=70&catid=8

Tian, C., & Ji, W. (2017). Auxiliary Multimodal LSTM for Audio-visual Speech Recognition and Lipreading. *ArXiv Preprint ArXiv:1701.04224*.

Tiippana, K. (2014). What is the McGurk effect? *Frontiers in Psychology*, *5*(JUL), 725. https://doi.org/10.3389/fpsyg.2014.00725

Vanegas, O., Tokuda, K., & Kitamura, T. (1999). Location normalization of HMM-based lip-reading: experiments for the M2VTS database. *Proceedings 1999 International*

Conference on Image Processing (Cat. 99CH36348), *2*, 343–347.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Viola, P., & Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, *1*, I–I.

Voronov, V., Strelnikov, V., Voronova, L., Trunov, A., & Vovik, A. (2019). Faces 2D-recognition and identification using the HOG descriptors method. *Conference of Open Innovations Association, FRUCT*, *24*, 783–789.

Xian, G. (2010). An identification method of malignant and benign liver tumors from ultrasonography based on GLCM texture features and fuzzy SVM. *Expert Systems with Applications*, *37*(10), 6737–6741.

Yau, W. C., Kumar, D. K., & Chinnadurai, T. (2008). Lip-reading technique using spatio-temporal templates and support vector machines. *Iberoamerican Congress on Pattern Recognition*, 610–617.

Yu, Q., Cheng, H. H., Cheng, W. W., & Zhou, X. (2004). Ch OpenCV for interactive open architecture computer vision. *Advances in Engineering Software*, *35*(8–9), 527–536.

Yuan, Y., Tian, C., & Lu, X. (2018). Auxiliary loss multimodal GRU model in audio-visual speech recognition. *IEEE Access*, *6*, 5573–5583.

Yuhas, B. P., Goldstein, M. H., & Sejnowski, T. J. (1989). Integration of acoustic and visual speech signals using neural networks. *IEEE Communications Magazine*, *27*(11), 65–71.

Zeliang, Z., Xiongfei, L., & Chengjia, Y. (2012). An effective parameter estimation algorithm of the visual language features [j]. *International Journal of Digital Content Technology and Its Applications*, *6*, 69–76.

Zhang, S., Li, X., Zong, M., Zhu, X., & Cheng, D. (2017). Learning k for knn classification. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *8*(3), 1–19.

Zheng, G.-L., Zhu, M., & Feng, L. (2014). Review of lip-reading recognition. *2014 Seventh International Symposium on Computational Intelligence and Design*, *1*, 293–298.

Zhou, H., Chen, P., & Shen, W. (2018). A multi-view face recognition system based on cascade face detector and improved Dlib. *MIPPR 2017: Pattern Recognition and*

Computer Vision, *10609*, 1060908.

Zhou, Z., Zhao, G., Hong, X., & Pietikäinen, M. (2014). A review of recent advances in visual speech decoding. *Image and Vision Computing*, *32*(9), 590–605.