

İşaret Dili Videolarından Hizalama ile Ayırık İşaret Çıkarımı

Alignment Based Extraction of Isolated Signs from Sign Language Videos

Pınar Santemiz¹, Oya Aran¹, Murat Saraçlar², Lale Akarun¹

¹Bilgisayar Mühendisliği Bölümü
Boğaziçi Üniversitesi

²Elektrik-Elektronik Mühendisliği Bölümü
Boğaziçi Üniversitesi

{pınar.santemiz, aranya, murat.saracilar, akarun}@boun.edu.tr

Özetçe

Bu çalışmada duraklamasız işaret dili videolarından otomatik ayırık işaret çıkarımı yapan bir yöntem geliştirdik. Aynı işaretin farklı zamanlarda yapılmış örneklerini içeren birden fazla diziyi birbirine hizalayarak işaretlerin başlangıç ve bitiş noktalarını bulduk. İşaret dili videolarında eller ve yüzü bir parçacık süzgeciyle izleyip birer elips oturttuk. Elips parametreleri, ayırık kosinüs dönüşümü ve yönlü gradyan histogramı betimleyicileri kullanarak, dinamik zaman bükmesi ve saklı Markov modeli yöntemleriyle hizalama yaptık ve sistemimizin başarımını Türk işaret dili veri tabanı üzerinde gösterdik.

Abstract

This paper presents a method to extract isolated signs from continuous sign language videos. We use sequences that approximately contain the sign that we are interested in and align the sequences to find the exact start and end frames. We compare different feature extraction methods, different alignment methods, and assess the performance of our system on a database from Turkish sign language.

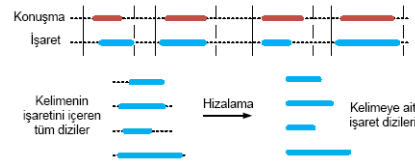
1. Giriş

İşaret dili, konuşma ve duyma engellilerin haberleşmek için kullandıkları, temelinde el hareketleri ve el şekline dayanan fakat bunların yanında yüz mimiklerinin, baş ve vücut hareketlerinin de kullanıldığı görsel bir dildir. İşaretler kavramları belirtir. Her zaman bire bir aynı olmasa da, konuşma dillerindeki kelimelerin bir karşılığı olarak düşünülebilir. İşaret dilleri, işaretlerin basit bir bileşimi olmaktan çok, aynı sözlü diller gibi karmaşık gramer yapısına sahiptir ve her ülke için farklıdır. İşaret dillerinin çevresinde konuşulan sözlü dille benzerlik göstermesi gerekmez ama etkiler görülebilir.

İşaret dilleri sadece ayırık işaretlerden oluşmaz, arka arkaya gelen işaretler cümleleri oluşturur. Normal hızda yapılan bir işaret dili cümlesinde arka arkaya gelen işaretler birbirini etkileyebilir. Dolayısıyla, işaretlerin başında ve sonunda ufak değişimler olabilir. Aynı etki, konuşma dillerinde de görülür.

Bu çalışmadaki amaç işaret dili videolarından otomatik ayırık işaret çıkarımı ve işaret dili tanıma sistemlerinde kullanılacak bir veri tabanının otomatik olarak elde edilme-

sidir. Bu amaç doğrultusunda TRT işitme engelliler haber videoları eşsiz bir kaynak sağlamaktadır. Bu videolarda haber bilgileri üç ayrı ortamda sunulmaktadır: konuşma, işaretleme ve altyazı. Sunucu aynı anda hem konuşmakta hem de söylediği kelimelere denk gelen işaretleri yapmaktadır. Türk işaret dilinin kendine özgü cümle yapısı olduğu düşünüldüğünde, bu videolardaki işaretleme Türk işaret dili olarak değil, işaretlenmiş Türkçe olarak adlandırılabilir. İşaretler ve konuşma birbiriyle örtüştüğünden bir kelimeyi konuşmada aradığımızda bulduğumuz yer aynı zamanda o kelimenin işaretinin de yaklaşık olarak yapıldığı yere denk düşmektedir. İşaretin tam olarak yapıldığı yeri, başlangıç ve bitiş noktalarını bulmak ise ayrı bir problemdir. Aynı kelime konuşmada bir kaç yerde geçebilir. Her söylendiği yerde de aynı işaret yapılmaktadır. Dolayısıyla problemimizi çok sayıda dizi içindeki en uzun ortak alt dizileri bulmak olarak tanımlayabiliriz. Bu problem hizalama teknikleri ile çözülebilir. Çalışmamızda kullandığımız yöntemin akışı Şekil 1'de gösterilmiştir. Bu çalışmada konuşma tanımının önceden gerçekleştirilmiş olduğu ve aranan kelime için kelimenin geçtiği aralıkların girdi olarak sağlandığı kabul edilmiştir. Konuşma tanıma yönteminin ayrıntıları için daha önceki çalışmalarımıza bakılabilir [1].



Şekil 1: Konuşmadan elde edilmiş dizilerden hizalama kullanarak işaretler çıkarılır. Kırmızı çizgiler aranan kelimenin konuşmada bulunduğu yerleri, mavi çizgiler ise kelimenin işaretinin yapıldığı yerleri göstermektedir.

Çoklu dizi hizalama yöntemleri biyolojide DNA dizilerini hizalamak amacıyla uzun süredir kullanılmaktadır [2]. İşaret dilinde de işaret tanıma ve ayırık işaret çıkarımı konularında çoklu dizi hizalama yöntemleri kullanılmıştır [3]. Alon ve arkadaşlarının [3] makalesinde saklı Markov modeli ve ayırık kosinüs dönüşümü yöntemleriyle çoklu işaret dizileri hizalanmış ve bu şekilde ayırık işaretlerin çıkarılmasının

dışında el izleme ve işaret tanıma başarımının da artırılması sağlanmıştır. İşaretleri hizalamak yerine hareket geçiş anlarını modelleyen yöntemler de önerilmiştir [4].

Bu çalışmada TRT işitme engelliler haber videolarının kayıtlarından oluşan bir veri tabanı kullanılmıştır. Bu videolarda birleşik parçacık süzgeci ile el ve yüz takibi yapılmış, bunun sonucunda elde edilen konum bilgisi kullanılarak yeni öznitelikler çıkarılmıştır. Daha sonra bu öznitelikler kullanılarak hizalama yapılmış, videoda geçen hareketlerin başlangıç ve bitiş anları belirlenmiştir.

2. Görüntü Veri Tabanı

Bu çalışmada TRT işitme engelliler haber videolarından elde edilmiş aynı sunucunun sunduğu 15 videoluk bir veri tabanı kullanılmıştır. Bu 2 saatlik veri tabanında toplam 174939 görüntü karesi ve toplam 10318 adet kelime bulunmaktadır. İşaret dili temel alınarak kelimeler gruplandırıldığında ise bu rakam 3498 değişik kelimeye karşılık gelmektedir. Tüm bu kelimelerin gerçek başlangıç ve bitiş anları işaret dili bilen kişilerce işaretlenmiştir. Bu çalışmada hizalama için en fazla örnek içeren 40 kelimeyi seçtik. Her bir kelime için 30 örnek kullandık. Seçilen kelimelerin 20 tanesi tek elle, 20 tanesi çift elle yapılmaktadır. Sunucunun hakim eli sağ el olduğundan tek elle yapılan işaretler sağ elle yapılmıştır.



Şekil 2: (a) Orjinal imge, (b) Birleşik parçacık filtresinde parçacık dağılımı.

3. Yüz ve El Takibi

İşaret dilinin doğası gereği, eller hızlı hareket eder ve birbirleriyle ya da yüzle sıklıkla temas halindedirler. Doğal hızında gerçekleştirilen işaret dili videolarında, elleri ve yüzü herhangi bir belirteç kullanmadan başarılı bir şekilde takip edebilmek için kapatmaya ve temasa dayanıklı, ellerin ve yüzün birbirlerine göre pozisyonlarına doğal olmayan önkabuller getirmeyen ve ellerin hızlı hareketini izleyebilen yöntemler geliştirilmelidir. Daha önce yaptığımız çalışmalarda bu tür durumlarda gürbüz izleme yapabilmek için bağımsız ve etkileşimli parçacık süzgeçleri kullanan bir yöntem geliştirmiştik [5]. Bu çalışmada kullandığımız parçacık süzgeci tabanlı yöntemde ise eller ve yüz için bağımsız ve etkileşimli parçacık süzgeçleri yerine ortak tek bir parçacık süzgeci kullanılmaktadır (Şekil 2). Kapatma ya da temas sırasında izlemeyi başarıyla sağlamak için

parçacık içindeki alt parçacıkların birbirlerine göre pozisyonları sınırlanır. Bu sayede iki alt parçacığın birbirine çok yakın olması ancak gerçekten temas ya da kapatma olduğunda mümkün olmaktadır.

Birleşik parçacık süzgeci doğası itibarıyla çok parçacık gerektirmektedir, bu da çalışma süresini uzatır. Bu durumu önlemek için parçacıkların yerini Ortalama Kayma (Mean Shift) yöntemiyle ten renginin daha yoğun olduğu bölgelere kaydırarak. Bu daha az parçacıkla daha iyi başarımlar elde etmemizi sağladı. Bu yöntem ile toplam 15 dakikalık yer gerçekliği çıkarılmış videoda, eller için ortalama %97, yüz için ise %99.99 izleme başarımı sağladık.

4. Öznitelik Çıkarımı

4.1. Önişleme

Ellerin özniteliklerini doğru bir şekilde elde edebilmek için öncelikle el takibi algoritmasının sonucunda elde edilen el bölgesinin içinde ten rengine en yakın pikselden başlayarak bölge büyütme uyguladık. Burada elin belli bir büyüklükle sınırlı olduğunu varsaydık. Bu şekilde her bir el için bir maske elde ettik. Daha sonra bu maskeye elips oturtarak ellerin merkezini, çevresine oturan elipsin minör ve majör eksenlerini ve majör eksenin dönme açısını bulduk.

4.2. Öznitelik Çıkarımı

Hizalama algoritmalarında karşılaştırmak üzere bu el resimlerinden beş öznitelik grubu çıkarttık:

- El şeklinin çevresine oturan elipsin merkez koordinatları ve bu koordinatların değişim vektörleri,
- El şeklinin çevresine oturan elipsin tüm parametreleri, yani merkezi, minör ve majör eksenleri ve döndürme açısı,
- Ardışık iki görüntü karesindeki elips parametrelerinin farklarını bularak elde ettiğimiz elips değişim vektörleri,
- Ayrık kosinüs dönüşümü (DCT),
- Yönlü gradyan histogramı (HOG).

4.3. Elips Parametreleri

El maskelerine oturtduğumuz elips parametrelerini gürültüden arındırmak için eğri oturtarak yumuşattık. Daha sonra farklı videolarda oluşan konum ve ölçek farklarından bağımsız hale getirmek amacıyla merkez noktalarını yüzün merkez noktasına göre öteledik ve tüm değişkenleri 0-1 arasına ölçekledik. Merkez koordinatlarının değişimlerini ise, ardışık iki görüntü karesindeki merkez koordinatlarının farklarını hesaplayarak elde ettik. Son olarak sağ el ve sol el için elde ettiğimiz öznitelikleri art arda bağladık. Böylece elde ettiğimiz ilk öznitelikte sekiz boyutlu, tüm elips parametrelerini kullanarak oluşturduğumuz diğer özniteliklerde ise 10 boyutlu vektörler kullandık.

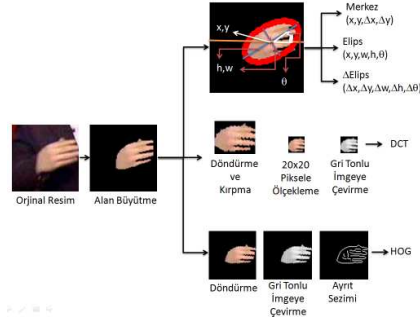
4.4. Ayrık Kosinüs Dönüşümü Parametreleri (DCT)

Ayrık kosinüs dönüşümünü hesaplamadan önce el resmini, bulunan elipsin döndürme açısına göre döndürdük. Daha sonra bu

resmi elin sınırlarından kırıp 20x20 piksel boyutuna ölçekledik ve gri tonlu imgeye çevirdik. Zigzag tarama yöntemiyle sıralayarak elde ettiğimiz 400 boyutlu vektörlerin DC katsayısını atıp ilk 200 elemanını seçtik. Daha sonra bu vektörlerin ana bileşenler analizi yöntemiyle boyutlarını her bir el için 50'ye düşürdük. Vektör boyutunun çok büyük olması nedeniyle sadece hakim elin (sağ el) özniteliklerini kullandık. Deneylerimizde diğer elin özniteliklerini kullanmamanın başarımı etkilemediğini gözlemledik.

4.5. Yönlü Gradyan Histogramı Parametreleri (HOG)

Yönlü gradyan histogramını hesaplarken, el resmini 80x80 piksellik bir karenin ortasına oturtup bulunan elipsin döndürme açısına göre resmi döndürdük. Daha sonra resmi gri tonlu imgeye çevirip Canny ayırma algoritmasıyla ayrıtları elde ettik ve ayrıtların bulunduğu piksellerdeki gradyanları hesapladık. Son olarak bu gradyanların histogramını çıkarttık ve tüm vektörleri 0-1 arası ölçekledik. Sağ el ve sol el için elde ettiğimiz vektörleri art arda bağlayarak her bir görüntü karesi için 18 boyutlu vektörler elde ettik.



Şekil 3: Öznitelik çıkarımı yöntemleri.

5. Hizalama

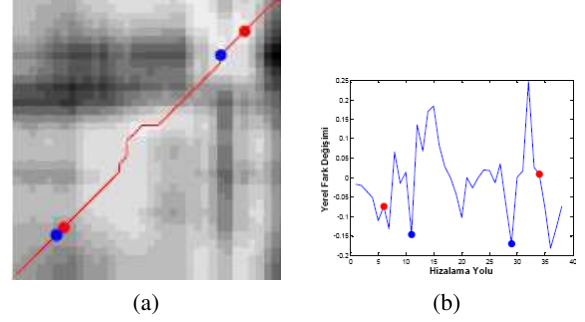
Konuşma tanıma kullanılarak belirlenen video parçalarını baştan ve sondan beş görüntü karesi genişleterek hizalamada kullandık. İşaretler kelimelerle tam eşzamanlı olmadığından elimizdeki video parçaları işaretin sadece bir kısmını içerebilir. Genişletme işlemiyle elimizdeki video parçasının aradığımız işaretin tamamını içermesini sağlamaya çalıştık. Elde ettiğimiz öznitelikleri kullanarak dinamik zaman bükmesi ve saklı Markov modeli algoritmalarını karşılaştırdık.

5.1. Dinamik Zaman Bükmesi (DTW)

Dinamik zaman bükmesi algoritmasında her bir kelime için tüm örnekleri ikili gruplar halinde hizaladık. Hizalama sırasında video parçalarının ortalarındaki bir pencere içinde birbirine en çok benzeyen öznitelikleri bulup, bu noktadan ileri ve geriye doğru hizalama yaptık. Bunun sonucunda her bir video parçası ikilisi arasında hizalama yolu elde ettik.

Bu hizalama yollarını ve örnekler arasında oluşan yerel fark matrislerini kullanarak her bir örnek için 29 aday başlangıç ve bitiş pozisyonu bulduk. Aday noktalarını bulurken öncelikle hizalama yolu üzerine düşen yerel fark değerlerini inceledik.

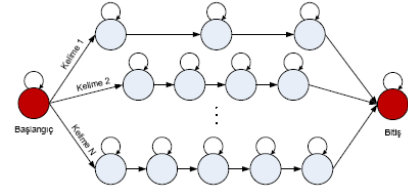
Bu değerlerin değiştiği noktalar arasından yolun ortasına en yakın değerleri başlangıç ve bitiş aday noktası olarak seçtik. Bu şekilde bir aday nokta belirlenemediği durumlarda ise hizalama yolunun içindeki en uzun köşegenin sınırlarını başlangıç ve bitiş noktası olarak seçtik. Bulduğumuz aday noktalarının ortalamasını gerçek başlangıç ve bitiş pozisyonu olarak seçtik.



Şekil 4: Dinamik zaman bükmesi ile hizalama. (a) İki video arasındaki yerel hata matrisi ve hizalama yolu, mavi nokta bulunan pozisyon, kırmızı nokta gerçek pozisyon (b) Hizalama yolu üzerindeki yerel hata değişimi değerleri.

5.2. Saklı Markov Modeli (HMM)

Saklı Markov modeli algoritmasında ise her bir kelimeyi süresiyle orantılı bir durum sayısı kullanarak soldan-sağa HMM ile modelledik. 10 katlı çarpaz geçiş yapılarak 10 adet eğitim ve geçiş kümesi elde ettik. Böylece her bir geçiş kümesinde üç örnek, her bir eğitim kümesinde ise 27 örnek oldu.



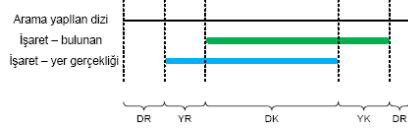
Şekil 5: Saklı Markov modeli ile hizalama.

Bu modeli ilk ve son durumu tüm kelime örnekleri için ortak olacak şekilde eğittik. Böylece her kelime için ilk durumun bittiği çerçeveyi başlangıç, son durumun başladığı çerçeveyi ise bitiş noktası olarak kabul ettik. Modelimizi eğitirken ilk ve son duruma girmeye zorladık.

6. Sonuçlar

Sistemin başarısını ölçmek için dört hata fonksiyonu kullandık: doğruluk oranı, kesinlik oranı, geri-getirme oranı ve sezim hatası. Bu fonksiyonları hesaplamak için bulduğumuz işareti yer gerçekliği verisiyle ve arama yaptığımız diziyi karşılaştırarak doğru kabul (DK), doğru red (DR), yanlış kabul (YK) ve yanlış red (YR) değerlerini elde ettik (Şekil 7).

Doğruluk oranını hesaplarken doğru kabul ve doğru red



Şekil 6: Bulunan işaret için aranan dizi ve yer gerçekliği verisine göre doğru kabul (DK), doğru red (DR), yanlış kabul (YK) ve yanlış red (YR) değerleri.

değerlerinin toplamının doğru kabul, doğru red, yanlış kabul ve yanlış red değerlerinin toplamına oranını bulduk (Denklem 1). Kesinlik oranını doğru kabul değerinin doğru kabul ve yanlış kabul değerlerinin toplamına oranı şeklinde (Denklem 2), geri getirme oranını ise doğru kabul değerinin doğru kabul ve yanlış red değerlerinin toplamına oranı şeklinde hesapladık (Denklem 3).

$$\text{Doğruluk} = \frac{DK + DR}{DK + DR + YK + YR} \quad (1)$$

$$\text{Kesinlik} = \frac{DK}{DK + YK} \quad (2)$$

$$\text{Geri-Getirme} = \frac{DK}{DK + YR} \quad (3)$$

Alon ve arkadaşlarının [3] önerdiği şekilde bu oranlar %50 ve üzerinde olduğunda sezimimizi doğru kabul ettik. Buna göre elde ettiğimiz sonuçlar Tablo 1’de verilmiştir. HMM ile farklı öznelıklar için, farklı eşik değerleri ile elde ettiğimiz doğruluk oranları ise Şekil 7’de verilmiştir. Sezim hatasını hesaplarken ise, beş görüntü karesine kadar yapılan hatalarda sezimimizi doğru kabul ettik. Buna göre elde ettiğimiz sonuçlar ise Tablo 2’de verilmiştir.

Tablo 1: Doğruluk, kesinlik ve geri getirme oranlarına göre başarımlar.

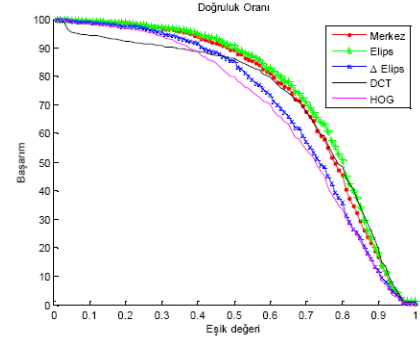
		Merkez	Elips	△Elips	DCT	HOG
DTW	Doğruluk	0.85	0.86	0.86	0.86	0.85
	Kesinlik	0.80	0.81	0.80	0.81	0.81
	Geri-getirme	0.70	0.70	0.70	0.70	0.68
HMM	Doğruluk	0.89	0.91	0.86	0.87	0.80
	Kesinlik	0.86	0.88	0.82	0.85	0.79
	Geri-getirme	0.86	0.86	0.78	0.71	0.81

Tablo 2: Sezim hatasına göre başarımlar.

		Merkez	Elips	△Elips	DCT	HOG
DTW	Baş	0.61	0.59	0.60	0.60	0.60
	Son	0.76	0.79	0.77	0.79	0.78
HMM	Baş	0.81	0.83	0.79	0.70	0.77
	Son	0.80	0.82	0.69	0.83	0.71

7. Vargılar

Bu çalışmada, duraklamasız işaret dili videolarından birden fazla video parçasını hizalayarak otomatik ayrık işaret çıkarımı yapan bir yöntem geliştirdik. Elips parametreleri, ayrık kosinüs dönüşümü ve yönlü gradyan histogramı betimleyicileri kullanarak, dinamik zaman bükmesi ve saklı Markov Modeli yöntemleriyle hizalama sonuçlarını karşılaştırdık. Buna göre saklı Markov modeli ile yaptığımız hizalamada daha iyi sonuç elde ettik. En iyi sonuca ise elips parametrelerinin tümünü kullandığımızda ulaştık ve %91 başarı elde ettik. Çalışmalarımızda, ortak alt diziyi bulmak için daha kısa süreli kelimelerde hata yapma olasılığının arttığını gördük. İşaretler arası geçiş belirsiz olduğundan, başlangıç ve bitiş noktaları tam olarak belli değildir. Bu nedenle 3-5 çerçevelik bir hata kabul edilebilir bir sonuçtur. Şekil 7’de görüldüğü gibi eşik değerimizi %70’e çıkardığımızda bile %70 üzerinde bir başarı elde edilmektedir.



Şekil 7: Farklı öznelıkların değişik eşik değerlerine göre doğruluk oranları.

8. Teşekkür

Bu çalışma Tübitak 107E021 nolu proje tarafından desteklenmektedir.

9. Kaynakça

- [1] Oya Aran, Ismail Ari, Pavel Campr, Erinc Dikici, Marek Hruz, Siddika Parlak, Lale Akarun, and Murat Saraclar, “Speech and sliding text aided sign retrieval from hearing impaired sign news videos,” *Journal on Multimodal User Interfaces*, vol. 2, no. 1, pp. 117–131, 2008.
- [2] Cédric Notredame, “Recent progresses in multiple sequence alignment: a survey,” *Pharmacogenomics*, vol. 3, no. 1, pp. 131–144, January 2002.
- [3] Jonathan Alon, Quan Yuan, and Stan Sclaroff, “A unified framework for gesture recognition and spatiotemporal gesture segmentation,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99, no. 1, 2008.
- [4] Gaolin Fang, Wen Geo, and Debin Zhao, “Large-vocabulary continuous sign language recognition based on transition-movement models,” *IEEE Transactions on Systems, Man and Cybernetics - Part A: Systems and Humans*, vol. 37, no. 1, pp. 1–9, 2007.
- [5] Oya Aran and Lale Akarun, “Etiklesimli parçacık süzgeci yöntemi ile kapatmaya dayanıklı yüz ve el takibi,” in *IEEE Sinyal İşleme Uygulamaları Konferansı*, 2008.