Haber Videolarında Nesne Tanıma ve Otomatik Etiketleme
Object Recognition and Auto-annotation In News Videos

Muhammet Baştan, Pınar Duygulu

Bilgisayar Mühendisliği Bölümü
Bilkent Üniversitesi, Bilkent, Ankara
{bastan, duygulu}@cs.bilkent.edu.tr

Özetçe

Bu çalışmada çok sayıda etiketlenmiş resim ve video içeren arşivlerin varlığından yararlanılarak nesne tanıma problemine yeni bir çözüm önerilmektedir. Nesne tanıma problemi, bir dilden başka bir dile çeviriye benzer sekilde, görsel öğelerin kelimelere çevirisi seklinde ele alınmaktadır. Bu amaçla öncelikle öznitelik uzayında temsil edilen görsel öğeler belli sayıda gruba ayrılır. Daha sonra, elde edilen gruplarla kelimeler arasındaki ilişkiler istatistiksel çeviri yöntemiyle öğrenilir. Son olarak, bir olasılık tablosu şeklinde öğrenilen bu ilişkiler bir resim üzerindeki bölütlerin, ya da bütün resmin belli kelimelerle etiketlenmesinde ve videolardaki konuşmalardan otomatik olarak elde edilen kelimelerin daha doğru video kareleriyle eşlenmesinde kullanılır. Deney sonuçları otomatik konuşma tanıma yöntemi sonucunda elde edilmiş metin bilgisine ve kullanıcılar tarafından girilmiş etiketlere sahip yaklaşık 150 saatlik haber videolarını içeren TRECVID 2004 veri kümesi üzerinde sunulmustur.

Abstract

We propose a new approach to object recognition problem motivated by the availability of large annotated image and video collections. Similar to translation from one language to another, this approach considers the object recognition problem as the translation of visual elements to words. The visual elements represented in feature space are first categorized into a finite set of blobs. Then, the correspondences between the blobs and the words are learned using a method adapted from Statistical Machine Translation. Finally, the correspondences, in the form of a probability table, are used to predict words for particular image regions (region naming), for entire images (auto-annotation), or to associate the automatically generated speech transcript text with the correct video frames (video alignment). Experimental results are presented on TRECVID 2004 data set, which consists of about 150 hours of news videos associated with manual annotations and speech transcript text.

1. Giriş

Gelişen teknoloji ile birlikte resim ve video veritabanlarının boyutları çok büyümüş ve anlamsal düzeyde arama yapabilecek sistemlere ihtiyaç duyulmuştur. Öte yandan, nesne tanıma bilgisayarla görme alanında hala çözülememis zor bir problem olup bu konudaki araştırmalar devam etmektedir. Henüz geniş ölçüde tanıma yapabilecek sistemler olmadığı gibi önerilen sistemler de genelde ancak birkaç sınıf nesneyi çok karmaşık olmayan görüntülerde tanıyabilecek kapasitededir. Bu nedenle varolan sistemler çoğunlukla kelime bazındaki sorgularla yetinmektedir. Bu da resimlerin kelimelerle etiketlenmesini gerektirmektedir. Bu işlemin elle yapılması, veritabanlarının çok büyük olması nedeniyle artık olası değildir. Yakın zamanlı çalışmalar göstermiştir ki çok büyük boyuttaki verinin küçük bir alt kümesinin kullanılarak genele dair bilgilerin öğrenilmesini sağlayacak sistemler daha hızlı ve verimli çözümler sunabilmektedir.

Bir dilden baska bir dile bilgisayarla çeviri yapmakta kullanılan istatistiksel yöntemler görsel veritabanlarının kelimelerle etiketlenmesi problemine uyarlanmış [4, 5, 6]; daha önceden belli kelimelerle etiketlenmiş veri kümelerinin öğrenme aşamasında kullanılması ve sonrasında büyük veritabanlarının otomatik olarak etiketlenmesi sağlanmıştır. Bu sayede nesne tanıma problemine de farklı bir yaklaşım sunulmuştur.

Öğrenme aşaması için gereken, belli kelimelerle etiketlenmiş görsel veri kümeleri günden güne artmaktadır. Örneğin, internette bulunan birçok resim açıklamalara sahiptir. Corel veri kümesi gibi birkaç kelime ile etiketlenmiş fotoğraf arşivleri bulunmakadır. Ayrıca, bu konuda çalışan birçok araştırmacının ortak çabasıyla önemli ölçüde elle etiketlenmiş veritabanları oluşturulmuştur. Haber videolarında anlamsal düzeyde arama yapmayı özendiren TRECVID [1] bunlardan biridir.

Katılımcılar tarafından etiketlenmiş bir grup verinin yanısıra otomatik konuşma tanıma (OKT) yöntemleri [2] sonucu elde edilen metin bilgisi de video referans çerçevelerine (key frame) karşılık gelecek kelimelerin bulunması için kullanılabilmektedir. Ancak, konuşmalarda resimlerdeki nesnelerden çok az ya da hiç behsedilmemesi, nesne adlarıyla yapılacak etiketlemelerde bu veri kümesinin başarı oranının düşmesine sebep olmaktadır.

Bu çalışma, sözü edilen iki tür veri kümesinin kullanılarak

- resimlerin ve resim üzerindeki bölütlerin otomatik olarak etiketlenmesini (otomatik resim etiketleme, bölge etiketleme),
- otomatik konuşma tanıma (OKT) ile elde edilen kelimelerin doğru referans çerçevelerle eşlenmesini ve böylece şekil 1'de gösterilen hizalama problemine (alignment problem) çözüm oluşturulmasını,

... (1) so today it was an energized president **CLINTON** who formally presented his one point seven three trillion dollar budget to the congress and told them there'd be money left over first of the white house a.b.c's sam donaldson (2) ready this (3) morning here at the whitehouse and why not (4) next year's projected budget deficit zero where they've presidental shelf and tell *this* (5) *budget marks the hand of an era and ended decades of deficits that have shackled our economy paralyzed our politics and held our people back.....*

Şekil 1: Videoda hizalama problemi: **CLINTON** adı geçerken (çerçeve 1) görüntüsü verilmediği gibi Clinton konuşurken (çerçeve 5) de adı söylenmemektedir. Dolayısıyla, metin bazlı bir arama sisteminde OKT metinleri kullanılarak yapılacak aramada Clinton yerine sunucunun resmine ulaşılacaktır.

ve sonuç olarak videolar üzerinde daha doğru sonuçlar veren sorgulamaların yapılabilmesini amaçlamaktadır.

2. Görsel öğelerin kelimelerle ilişkilendirilmesi

Bilgisayarlı çeviriden esinlenerek tasarlanan, görsel öğelerin nesne veya kavramlara denk gelen kelimelerle ilişkilendirilmesi metodu detaylı olarak [4]'te anlatılmıştır. Özetle: ilk olarak kullanılacak öznitelikler belirlenip resimler bu özniteliklere göre belli sayıda gruba (blob, visterm) ayrılır (clustering). Daha sonra, elde edilen görsel gruplarla kelimeler arasındaki ilişkiler, iki dil arasında birbirinin çevirisi olan paralel metinlerden yararlanılarak yapılan istatistiksel çeviri yöntemine benzer şekilde [7], öğrenilip bu ilişkileri gösteren bir olasılık tablosu hazırlanır. Son olarak, hazırlanan bu olasılık tablosu resimlerin ya da resim üzerindeki bölütlere ayrılmış bölgelerin kelimelerle etiketlenmesinde; video karelerinin OKT ile elde edilen kelimelerle daha doğru bir şekilde eşlenmesinde kullanılır.
Bu çalışmada, resimler k-means algoritması ile belli sayıda gruplara ayrılmış ve bu gruplarla kelimeler arasındaki ilişkiler Giza++ [3] kullanılarak öğrenilmiştir.
Performans ölçüsü olarak ortalama kelime tahmin oranı (doğru tahmin edilen kelime sayısının elle yapılan etiketlemedeki kelime sayısına oranı), geri getirme yüzdesi (recall) ve kesinlik (precision) değerleri hesaplanmıştır.

3. Veri kümesi ve öznitelikler

Deneylerde, her yıl NIST (Amerikan Standartlar Enstitüsü) tarafından düzenlenen TRECVID yarışması [1] katılımcılarına verilen veri kümelerinden 2004 yılına ait, 150 saatlik CNN ve ABC haber videolarından oluşan TRECVID 2004 veri kümesi kullanılmıştır. Videolardan çıkartılan referans çerçeve resimleri, bu resimlere denk gelen, katılımcıların ortak çalışmasıyla belli sayıda kelime ile yapılan etiketlemeler (manual annotation), LIMSI [2] tarafından otomatik konusma tanıma (OKT) yöntemiyle elde edilen zaman bazlı metinler (ASR text) de sağlandı. OKT ile elde edilen metinler anahtar kelimelerin dışında birçok kelime (fiil, sıfat, ek, vb) içerdiği için önişleme sonucunda yalnızca nesnelere veya kavramlara denk gelen isimler anahtar kelimeler (keyword) olarak kullanılmıştır.
Haber videoları herbiri ayrı bir haberden bahseden hikayelerden (news stories) oluşur. Bu hikayeler, hikaye bölütleme (story segmentation) metotları kullanılarak elde edilebilir. NIST tarafından sağlanan ve her hikayenin başlangıç ve bitiş zamanını milisaniye cinsinden gösteren veriler kullanılarak her hikayenin içerdigi referans çerçeveler ve onlara denk gelen OKT ile elde edilen kelimeler eğitim ve test aşamalarında kullanılmıştır.
Referans çerçeveler, genel renk (RGB, HSV, LUV) ve ayrıt (egde) histogramları; referans çerçevelerin bölündüğü 5X7'lik ızgaralar (grid) da renk (RGB, HSV, LUV) ortalama ve standart sapma değerleri ile, doku (Gabor) gibi özniteliklerle temsil edildi.

4. Deney sonuçları

Bu çalışma genel olarak iki ana kısımdan oluştuğu için deney sonuçları iki ayrı bölüm halinde sunulacak olup daha çok ikinci kısım üzerinde yoğunlaşılacaktır.

4.1. Otomatik etiketleme

Bu kısımdaki deneylerde TRECVID 2004 veri kümesine ait 92 video kullanılmıştır. Videolara ait referans çerçeveler elle 614 nesne ve kavram adıyla etiketlenmiş olup yanlış yazılan ve frekansı düşük kelimelerin elenmesiyle geriye 62 tane anahtar kelime kalmıştır. Sonucu verilen deneylerde resimler 5X7'lik ızgaralara bölünmüş, renk (RGB ortalama, standart sapma) ve doku (Gabor) ile temsil edilip k-means kullanılarak 1000 gruba ayrılmıştır. Test kümesi üzerindeki performans hesabı için, tahmin edilen kelimeler gerçek olanlarla otomatik olarak karşılaştırılmış; ortalama kelime tahmin performansı 0.29, en az 1 kere tahmin edilen kelimeler için kesinlik ve geri getirme oranları da sırasıyla 0.18 ve 0.33 olarak elde edilmiştir.
Resim gerçekte `n` kelime ile etiketlenmişse performans hesaplarında sadece tahmin edilen ilk `n` kelime dikkate alınmıştır. Bazı resimlerin içinde daha fazla nesne olmasına rağmen sadece 1 ya da çok az kelime ile etiketlenmiş olması, tahmin edilen kelimeye ait nesnenin resimde olmasına rağmen gerçek etikette bulunmaması (örnek: şekil 2'de ikinci sıradaki ilk resimde **sky** olmasına rağmen etikette yer almıyor) gibi sebepler otomatik olarak hesaplanan performansın olduğundan daha düşük görünmesine sebep olmaktadır.
Şekil 2'de bazı otomatik etiketleme örnekleri verilmiştir. Sonuçlar göstermiştir ki, resimlerden oluşan veri kümeleri için herhangi bir etiket olmadığında, otomatik etiketleme sonucu elde edilen kelimeler daha iyi erişim için kullanılabilir.
Şekil 3'te resim üzerindeki bölütler otomatik olarak etiketlenmiş olup `female-news-person`, `female-face`, `studio-setting`, `graphics` gibi kelimeler doğru olarak tahmin edilebilmiştir. Resim üzerinde bölütlere ayrılmış alanların bu şekilde etiketlenmesi nesne tanıma olarak kabul edilebilir.

4.2. OKT metinlerinin kullanılışı ve hizalama problemi

Bu kısımda sonuçları verilen deneylerde OKT metinleri ve haber hikayelerinin başlama ve bitiş zamanları LIMSI ve NIST

	
studio-setting graphics female-news-person male-news-subject person	people basketball
female-news-person studio-setting people male-face graphics person scene-text	people graphics basketball female-news-nerson scene-text male-news-subject studio-setting
	
water-body boat	forest male-news-subject female-face person graphics
sky graphics water-body building boat person male-news-person	people person graphics male-face greenery scene-text female-face

Şekil 2: Otomatik etiketleme sonuçları. Asıl kelimeler üstte, tahmin edilen ilk 7 kelime altta verilmiştir.

studio-setting female-news-person

468,359,213: female-face 300,225: female-news-person 167,272,346,443: graphics 202,429,320,43,46,79: studio-setting 223,475,317: male-face	104,404: person 81,299: scene-text 437: people 61: flag 319: basketball

Şekil 3: Resim üzerinde bölütlerin etiketlenmesi (region labeling) örneği.

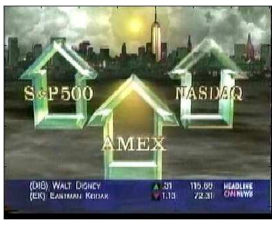
tarafından sağlanan TRECVID 2004 haber videolarından 111'i eğitim, 110'u da test için kullanılmıştır. OKT metinlerindeki kelimeler ön işlemeyle sadece isimler elde edilmiş, 300'den daha az frekansa sahip isimlerin elenmesiyle geriye 251 kelime kalmıştır. Her referans çerçeve, renk (RGB) ve doku (Canny) genel histogramı ile temsil edilip k-means ile 1000 gruba ayrılmıştır. Ayrıca her resimde kaç tane insan yüzü olduğu bilgisi de eğitimde kullanılmıştır.

Eğitim sonrası elde edilen olasılık tablosu kullanılarak test kümesindeki resimler için kelimeler tahmin edilmiş; şekil 4'te de gösterildiği gibi genel renk özelliklerinden ayırdedilebilecek `hava durumu`, `spor`, `borsa` gibi haberlerle ilgili resimler doğru olarak tahmin edilmiştir.

Şekil 5'te haber hikayeleri için kelimeler tahmin edilmiş; yine kullanılan özniteliklerle ayırdedilebilecek resimlerin bulunduğu `hava durumu`, `borsa`, `spor` gibi konularla ilgili haber hikayeleri için oldukça başarılı tahminler yapılabilmiştir. Asıl OKT metinleriyle karşılaştırıldığında haber hikayeleri için ortalama kelime tahmin performansı 0.17, kelime başına ortalama geri getirme yüzdesi 0.16, kesinlik değeri ise 0.20 olmuştur.

Tasarlanan sistem sayesinde OKT metinleri olmasa bile istenen nesne ya da kavramlarla ilgili resimlere ulaşmak mümkün olabilmektedir. Şekil 6, `sport` kelimesiyle farklı spor sahneleri arasındaki ilişkilerin sistem tarafından öğrenilebildiğini göstermektedir. Yine şekil 7'deki örnekte `snow`, `night`, `office` gibi, kullanılan resim öznitelikleri ile ayırdedilebilecek sahnelerle kelimeler arasındaki ilişkiler başarıyla öğrenilebilmiştir.

Önerilen sistemin videolardaki OKT metinleri ile resimler arasındaki hizalama problemine çözüm olabileceğini gösteren bir örnek şekil 8'de verilmiştir. OKT metininde sunucu ile eşlenen `clinton` kelimesi, resimlerle kelimeler arasındaki ilişkiler önerilen sistemle öğrenildiğinde en yüksek olasılıkla 3. sıradaki doğru resimle eşlenebilmiştir.

	
temperature weather forecast	point nasdaq stock
	
sport time game	jenning people evening

Şekil 4: Bazı resimler için OKT metinlerinden elde edilen kelimelerle yapılan eğitim sonnrasında tahmin edilen en yüksek olasılıklı 3 kelime.

OKT : center headline thunderstorm morning line move state area pressure chance shower lake head monday west end weekend percent temperature gulf coast tuesday
Tahmin : weather thunderstorm rain temperature system shower west coast snow pressure

OKT : check peace york morning charge dollar share nasdaq market issue percent consumer month
Tahmin : market stock york nasdaq street check point yesterday record share

OKT : night game sery story
Tahmin : game headline sport goal team product business record time shot

Şekil 5: Bazı haber hikayeleri için OKT metinleri kullanılarak yapılan tahminlerde en yüksek olasılıklı 10 kelime.

Şekil 6: `sport` kelimesinin ilk 2'de tahmin edildiği resimler.

Şekil 7: Sırasıyla `snow`, `night` ve `office` kelimelerinin ilk 7'de tahmin edildiği resimler.

5. Tartışma ve Sonuçlar

Bu çalışmada bilgisayarla çeviriden uyarlanarak görsel öğelerin kelimelere çevrilmesini amaçlayan bir sistem geliştirilmiştir. Önerilen sistemde, resim üzerinde bölütlere ayrılmış bölgelerin

OKT : (1) home washington president clinton (2) office president state department (3) deal

Şekil 8: Clinton ile ilgili 3 resimden oluşan bir haberde her resme denk gelen OKT metinleri gösterilmiştir. OKT metnine göre Clinton aslında sunucunun olduğu ilk resimle eşlenmektedir. Resimler ile kelimeler arasındaki ilişkiler öğrenilip `clinton` kelimesi ile yapılan bir aramada ise `clinton` kelimesi en yüksek olasılıkla 3. resimle eşleşmektedir.

etiketlenmesi nesne tanıma, bütün bir resmin belli kelimelerle otomatik etiketlenmesi resim veri kümelerine erişim, OKT metinlerinin daha doğru video kareleriyle eşlenmesi de geniş video arşivlerine OKT metinleri yardımıyla daha etkin erişimi sağlamaya yönelik çözümler sunmaktadır.
Videolarda hareket eden nesneler de önemli bilgiler taşır. Dolayısıyla, elde edilecek hareket bilgileri, nesnelerin isimlerle eşlenmesine benzer şekilde fiiller ile eşlenebilir. Böylece videolar üzerinde daha zengin içerikli aramalar yapılabilir.
Önerilen çeviri metodu çok sayıda isim ile yüzlerin eşlenebilmesi için de yeni bir yaklaşım olarak düşünülebilir.

6. Teşekkür

Bu çalışma TÜBİTAK Kariyer 104E065 ve TÜBİTAK 104E077 nolu projeleri tarafından desteklenmiştir.

7. Kaynakça

[1] TRECVID, TREC Video Retrieval Evaluation, http://www-nlpir.nist.gov/projects/trecvid.

[2] J.L. Gauvain and L. Lamel and G. Adda, "The LIMSI Broadcast News Transcription System", Speech Communication, Vol. 37, p 89-108, 2002.

[3] Giza++, http://www.fjoch.com/GIZA++.html.

[4] K. Barnard and P. Duygulu and N. de Freitas and D. A. Forsyth and D. Blei and M. Jordan, "Matching words and pictures", Journal of Machine Learning Research, Vol. 3, p 1107-1135, 2003.

[5] P. Duygulu and K. Barnard and N.d. Freitas and D. A. Forsyth, "Object recognition as machine translation: learning a lexicon for a fixed image vocabulary", Seventh European Conference on Computer Vision (ECCV), Vol. 4, p 97-112, 2002.

[6] P. Virga and P. Duygulu, "Systematic Evaluation of Machine Translation Methods for Image and Video Annotation", The Fourth International Conference on Image and Video Retrieval (CIVR 2005), Singapore, 2005.

[7] I. D. Melamed, Empirical Methods for Exploiting Parallel Texts, MIT Press, Cambridge Massachusetts, 2001.