Genelleştirilmiş Görsel Kavram Tanıma
Generalized Visual Concept Detection

Ahmet Saracoğlu[1,2], Mashar Tekin[1,3], Ersin Esen[1,2], Medeni Soysal[1,2], K. Berker Loğoğlu[1,4], Tuğrul K. Ateş[1,2], A. Müge Sevinç[1,4], Hakan Sevimli[1,3], Banu Oskay Acar[1] , Ünal Zubari[1] , Ezgi Can Ozan[1,2], A. Aydın Alatan[2]

1. TÜBİTAK Uzay Teknolojileri Araştırma Enstitüsü

2. Elektrik ve Elektronik Mühendisliği Bölümü
Orta Doğu Teknik Üniversitesi

3. Bilgisayar Mühendisliği Bölümü
Orta Doğu Teknik Üniversitesi

4. Enformatik Enstitüsü
Orta Doğu Teknik Üniversitesi

{ahmet.saracoglu, mashar.tekin, ersin.esen, medeni.soysal, berker.logoglu, tugrul.ates, muge.sevinc, hakan.sevimli, banu.oskay, unal.zubari, ezgican.ozan}@uzay.tubitak.gov.tr
alatan@eee.metu.edu.tr

Özetçe

Video arşivlerinin etkin bir şekilde indekslenmesi ve aranması gibi pek çok uygulama için kavram tanıma önemli bir problem olarak durmaktadır. Bu çalışmada, farklı kavramlar için aynı yapının kullanılmasını hedefleyen genelleştirilmiş bir kavram tanıma sistemi önerilmektedir. Sistem içerisinde görsel öznitelik olarak MPEG-7 Belirticileri ve Ölçek Bağımsız Öznitelik Dönüşümü (SIFT) kullanılmıştır. Öznitelikler, k-Ortalamalar ile oluşturulan Kod Tablosu uzayına indirgenmektedir ve sınıflandırma kod tablosundaki dağılım üzerinden gerçekleştirilmektedir. Kavramlara özgü gruplanma çok sayıda kod kelimesi kullanarak sağlanmaktadır. Önerilen kavram tanıma sistemi, öncelikli olarak temel bir kavram için denenmiş ardından ise çeşitli kavramlar için TRECVID 2009 test kümesi üzerinde elde edilen sonuçlar raporlanmıştır. Eğitim verisinin yeterli olduğu durumlarda başarımın oldukça yüksek olduğu gözlenmiştir.

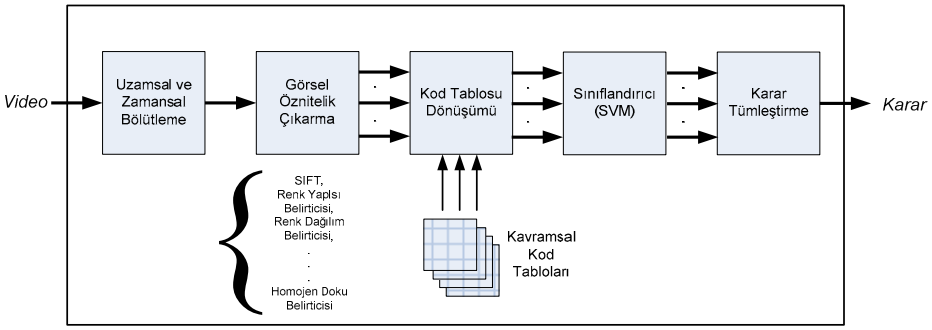
Abstract

For efficient indexing and retrieval of video archives, concept detection stands as an important problem .In this work, a generalized structure that can be used for detection of diverse and distinct concepts is proposed. In the system, MPEG-7 Descriptors and Scale Invariant Transform (SIFT) are utilized as visual features. Furthermore, visual features are transformed by codebooks which are constructed by k-Means clustering. On the other hand, classification is performed on the distribution of visual features over the codebook. Proposed system is firstly tested against an elementary concept. Afterwards for a set of concepts system performance is reported on the TRECVID 2009 test set. It has been observed that with a sufficiently large training set high performance can be achieved with this method.

1. Giriş

Çoğulortam arşivlerinin indekslenmesi ve etkin bir şekilde metin tabanlı sorgulanabilmesi için kullanılabilecek en önemli bilgi kaynaklarından bir tanesi videonun anlamsal/kavramsal içeriğidir. Günümüzde kullanıcılar arasında hızla yayılan ve sayıları her geçen gün artan video paylaşım sitelerinin varlığı da videonun kavramsal analizini daha da önemli kılmaktadır. Ayrıca, yayın takibi için içeriğin otomatik analizi ve sınıflandırılması da önemli bir ihtiyaç olarak ortaya çıkmaktadır. Bu bağlamda kavramları, "bir nesnenin veya düşüncenin zihindeki soyut ve genel tasarımı" olarak tanımlayabilir ve kavram tanıma işlemini de sisteme tanıtılan/öğretilen kavramların otomatik olarak bulunması olarak nitelendirebiliriz.

Kavram tanıma konusunda oldukça fazla yaklaşım bulunmaktadır. Literatürdeki en başarılı çalışmalardan birisi MediaMill grubunun çalışması olup geliştirdikleri sistem [1], [2] video çekimlerinin zamansal olarak çoklu karelerin seçilmesi ile örneklenmesi ve bu kareler üzerinden elde edilen özniteliklerin sınıflandırılması şeklinde tasarlanmıştır. Öznitelikler, SVM ile sınıflandırılmadan önce daha önceden elde edilen kod tablosu ile dönüştürülmekte ve her bir kareden oluşturulan imge piramitlerinden elde edilmektedir. Sınıflandırma için kullanılan öznitelikler çıkarılma şekillerine göre yerel ve global olmak üzere ikiye ayrılmıştır. Yerel öznitelikler renk uzayını da göz önüne alan SIFT tanımlayıcısı, global öznitelikler ise Wiccest [3] ve Gabor [4] filtrelerinden elde edilen vektörlerdir. Yöntemde öznitelik dönüşümleri için kullanılan kod tablosu her bir öznitelik tipi için ayrı olacak şekilde eğitim kümesinden k-Ortalamalar yöntemi ile oluşturulmaktadır. Çekimlerin kavramsal olarak sınıflandırılması ise çekimden elde edilen bütün özniteliklerin kod tablosundaki dağılımları üzerinden yapılmaktadır. Benzer olarak Chang ve ekibinin çalışmalarında [8] kelime torbası ("bag of words") temelli [5], sınıflandırıcı olarak SVM kullanan bir sistem önerilmektedir. Anahtar karelerde öznitelik çıkarılacak noktalar Gauss Farkı ve Hessian-İlgin yöntemleri ile ayrı ayrı bulunmaktadır. Daha sonra bu noktalardan SIFT tanımlayıcıları hesaplanmakta ve bu vektörler öznitelik olarak kullanılmaktadır. Eğitim kümesinden elde edilen öznitelik kümesinde gerçekleştirilen k-Ortalamalar öbeklemesi ile de



Şekil 1: Genelleştirilmiş Görsel Kavram Tanıma Blok Çizeneği.

görsel kod tablosu elde edilmektedir. Özniteliklerin sınıflandırma işlemi de karelerdeki özniteliklerin bu görsel sözlükteki histogramlarına göre yapılmaktadır. Peng ve ekibi [7] yukarıda anlatılan yöntemleri değerlendirmiş, ancak eğitim veri kümesindeki dengesiz etiket dağılımının sınıflandırıcı eğitiminde sorun yaratmaması için ABU-SVM olarak isimlendirdikleri eğitim yöntemini kullanmışlardır. Eğitim yöntemi, ceza fonksiyonunun değerlerine göre azaltılan negatif örneklere karşılık pozitif örneklerin kopyalanarak arttırılmasını içermektedir. Ceza fonksiyonu da birden fazla SVM'in eğitim etiketlerinden elde edilen farklı kümelerdeki eğitim sonuçlarına göre hesaplanmaktadır. Benzer şekilde TRECVID 2007'de IBM ve Fudan Universiteleri de Seyrek Örneklemeli SVM (USVM) [6] kullanmışlardır.

Yukarıda açıklanan yöntemler ile benzer şekilde bu çalışmada, kod tablosu ve SVM tabanlı bir sistem önerilmektedir. Önerilen sistemde görsel öznitelikler üç farklı kategoride çıkarılmakta ve MPEG-7 belirticileri ile Ölçek Bağımsız Öznitelik Dönüşümü (SIFT) birlikte kullanılmaktadır. Bildirinin ikinci bölümünde önerilen yöntem detaylı bir şekilde anlatılmaktadır. Üçüncü bölümde ise, TRECVID 2009 çalıştay video veritabanında kısıtlı bir kavram kümesi için elde edilen sonuçlar verilmektedir. Son bölümde de sonuçlar değerlendirilmekte ve gelecek çalışmalar için ön bilgiler verilmektedir.

2. Önerilen Yöntem

Önerilen genelleştirilmiş görsel kavram tanıma sistemi ile, pek çok farklı kavramın aynı yapı ile tanınması hedeflenmektedir. Bu amaçla farklı açılardan görsel bilgiyi temsil eden değişik öznitelikler bir arada kullanılmaktadır. Öznitelikler kod tablosu düzeyine indirgenerek kaba bir özetleme sağlanmaktadır. Kavramlara özgü öbeklenme ise çok sayıda kod kelimesinin bir arada kullanılması ile elde edilmektedir. Sistemin ana yapısı Şekil 1'de gösterilmekdtedir. Şekilde gösterildiği gibi sistem beş ana kısımdan oluşmakta ve sistem bu haliyle video içerisindeki her bir çekim için ayrı kararlar vermektedir. Temel olarak sistem farklı kavramların eğitimi ve tanınması için genelgeçer bir çerçeve sunmaktadır.

2.1. Uzamsal ve Zamansal Bölütleme

Kavram tanımanın ilk adımı olarak video uzamsal ve zamansal olarak parçalara ayrılmaktadır. Zamansal örnekleme, videonun çekimlere ayrılması ve çekimlerden anahtar kareler çıkarılması işlemi olarak özetlenebilir. Video çekimlerini tanımlayan anahtar kareler ise, çekim içerisinden eşit aralıklar ile örneklenmektedir. Uzamsal örneklemede ise anahtar karelerin düzenli ızgaralar ile bölütlenerek öznitelik çıkarmaya hazırlanması gerçekleştirilmektedir. Kullanılan ızgaralar, kareleri 1x1 (karenin tamamı), 2x2 ve 3x3 şeklinde parçalara ayırmaktadır (Şekil 2) ve öznitelikler her bir parça için ayrı ayrı çıkarılmaktadır. Anahtar karenin tamamının yanı sıra, daha küçük bloklarından da özniteliklerin ayrıca çıkarılması ile kavramların taşıyabileceği yerel/bölgesel bilgiler genel öznitelik çıkarmada tamamiyle kaybedilmemiş olmakta ve sınıflandırmada dolaylı da olsa kavramların bölgeselliklerinin de değerlendirilmesini sağlamaktadır.

2.2. Görsel Öznitelik Çıkarma

Özniteliklerin çıkarılması için verinin hazırlanmasından sonra global, yarı-global ve seyrek olmak üzere üç kategoride öznitelik vektörleri çıkarılmaktadır.

Global öznitelikler, karenin tamamından çıkarılan temel özniteliklerden (renk yapısı belirticisi, renk dağılım belirticisi, homojen doku belirticisi vs.) oluşmaktadırlar. Kullanılan öznitelik çıkarma metotları aşağıda kısaca anlatılmaktadır. Yarı-global öznitelikler ise, 2x2 ve 3x3 ızgaralardan çıkarılan yine aşağıda belirtilen özniteliklerdir. Böylelikle, tek bir anahtar karede aynı tip temel öznitelik çeşidinden global ve yarı-global olmak üzere toplam 14 öznitelik vektörü çıkarılmaktadır. Uzamsal örnekleme sayesinde de kare içerisinde kavram ile ilgili olabilecek bölgesel alanlar da ayrıca öznitelikler ile kapsanabilmektedir. Son olarak da karenin tamamından seyrek olarak kategorize edilebilecek, SIFT öznitelikleri çıkarılmaktadır.

Önerilen yöntemde, MPEG-7 belirticilerinden [11], Renk Yapısı Belirticisi (Color Structure Descriptor), Homojen Doku

Şekil 2: Örnek anahtar kare üzerinde global ve yarı-global öznitelik çıkarma alanları.

Belirticisi (Homogeneous Texture Descriptor), Renk Dağılımı Belirticisi (Color Layout Descriptor) ve Kenar Histogramı Belirticisi (Edge Histogram Descriptor) kullanılmıştır. Bunlara ek olarak daha önce belirtildiği gibi SIFT özniteliği de kullanılmaktadır.

2.2.1. Renk Yapısı Belirticisi

Renk histogramını temel almasına karşın ufak bir yapısal pencere kullanarak yapısal renk dağılımlarını da nitelendirmektedir, ayrıca çözünürlük değişimlerinden de etkilenmemektedir.

2.2.2. Homojen Doku Belirticisi

İmgenin dokusal içeriğini Gabor filtrelerini kullanarak nitelendirmektedir. Gabor filtreleri ile seçilen belirli frekans bantlarındaki enerjinin ortalaması ve değişintileri belirticiyi oluşturmaktadır.

2.2.3. Renk Dağılım Belirticisi

Öznitelik, imgedeki renk içeriğinin uzamsal dağılımını temsil etmektedir. Özniteliğin çıkarılmasında Ayrık Kosinüs Dönüşümünden faydalanılmaktadır.

2.2.4. Kenar Histogramı Belirticisi

Yerel ve global kenar dağılımlarını nitelendirmektedir. Kenar dağılımları ise daha önceden belirlenen beş farklı kenar kategorisine (yatay, dikey, köşegen, ters köşegen ve yönsüz) göre gerçekleştirilmektedir.

2.2.5. Ölçek Bağımsız Öznitelik Dönüşümü

SIFT öznitelik çıkarma yöntemi [10], video kareleri üzerinde çeşitli ilgi noktalarının bulunmasını ve bu noktalardan ayırt edici ölçek ve dönme bağımsız öznitelik vektörlerinin çıkarılmasını kapsamaktadır. Literatürde özellikle obje tanıma ve sahne eşleme gibi çeşitli uygulamalarda kullanılmaktadır; aydınlatma değişikliklerine, ilgin bozulumuna, gürültü ve benzeri bozulumlara karşı gürbüzlüğü kavram tanıma için uygunluğunu arttırmaktadır.

2.3. Kod Tablosu Dönüşümü ve Sınıflandırma

Sınıflandırma işleminden hemen önce çekimlerden elde edilen bütün öznitelikler daha önceden hazırlanan kod tabloları kullanılarak dönüştürülmektedir. Dönüşüm sayesinde çekimin tamamını tanımlayan tek bir tanımlayıcı öznitelik vektörü elde edilmektedir; böylelikle sınıflandırma öncesi boyut indirgenmekte ve ayrıca hesaplama karmaşıklığında önemli bir kazanım sağlanmaktadır. Dönüşümde kullanılan kod tabloları, sınıflandırılacak/sezilecek her bir kavram ve her bir öznitelik tipi için ayrı olacak şekilde hazırlanmaktadır. Bununla birlikte bahsedilen bu kod tabloları eğitim kümesindeki öznitelik vektörlerinin k-Ortalamalar metodu ile belirlenen topak merkezlerinden oluşturulmaktadır. Kod tabloları dönüşümü sırasında, bir çekimde çıkarılan tek bir öznitelik tipindeki (örn.: global-homojen doku, yarı global-renk dağılımı ya da seyrek-SIFT vb.) bütün vektörler en yakın oldukları topak merkezine eşlenmekte ve özniteliklerin topak uzayındaki dağılımlarının histogramı çıkarılmaktadır.

Yukarıda bahsedilen yöntemle sınıflandırılacak çekim için çoklu sayıda öznitelik elde edilmiş olmaktadır; sınıflandırma işlemi ise bahsedilen histogramlar üzerinde çift sınıf problem olarak, kavramın varlığı ve yokluğu şeklinde gerçekleştirilmektedir. Sınıflandırıcı olarak sistemde SVM kullanılmaktadır ve karar aşağıdaki fonksiyon sayesinde elde edilmektedir;

$$y(x) = \sum_k w_k l_k K(x, x_k) + b \qquad (1)$$

(1)'de x test edilen vektörü, x_k eğitim kümesinden vektörü $K(.)$ çekirdek fonksiyonunu, l_k ise x_k vektörünün sınıfını ve w_k da x_k örneğinin ağırlığını belirtmektedir. Kullanılan çekirdek tercihi ön deneyler sonucunda en iyi öğrenme performansına göre yapılmıştır ve önerilen sistemde radyal taban fonksiyonu (2) kullanılmaktadır. Radyal taban fonksiyonundaki $d(x, y)$ öznitelik uzayında tanımlanan her hangi bir uzaklık metriğini tanımlamaktadır.

$$K(x, y) = e^{-Cd(x,y)} \qquad (2)$$

2.4. Karar Tümleştirme

Son olarak da her bir öznitelik üzerinden yapılan sınıflandırma sonuçları tümleştirilmektedir. Önerilen sistemde karar tümleştirme aşamasında buluşsal kurallar ile mantıksal çarpım/toplam kullanılmaktadır. Buluşsal kurallar ön deneyler sonucunda kavram bazında özniteliklerin SVM ile olan performanslarına göre belirlenmiştir ve temelde basit ağırlıklı toplamaya dayanmaktadır.

3. Deneyler

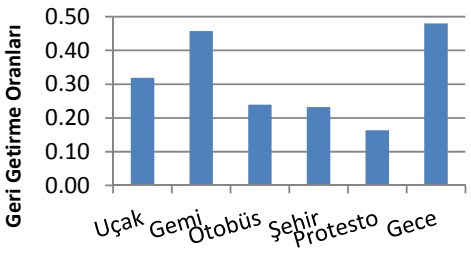
Önerilen yöntemin performansının denenmesi için iki farklı test senaryosu değerlendirilmiştir. Senaryolardan ilkinde videolardan ziyade anahtar karelerdeki tek bir kavrama ait tanıma performansı irdelenmiştir. İkinci durumda ise, geniş bir video veritabanında birden fazla kavramın bu sistem ile tanınma performansı gözlenmiştir.

Senaryoların ilkinde *İç Mekan* kavramının tanınması amaçlanmıştır. Bu kavramın eğitimi için Quattoni'nin [11]'deki çalışmasında kullanılan veritabanından rastgele seçilen 961 imge pozitif örnek kümesi (Şekil 3) oluşturulmuştur. Negatif örneklerin kümesi ise farklı kaynaklardan toplanan 913 imgeden oluşmaktadır. Bununla birlikte, test için 648 pozitif, 313 negatif örnekten oluşan bir küme hazırlanmıştır. Yalnızca anahtar karelerin kullanılmasından dolayı zamansal bölütleme yapılmamış ve ilgi noktalarından çıkarılan öznitelikler 128 öbekten oluşan bir kod tablosu ile dönüştürülmüştür. Karar tümleştirme adımı ise yapılan ön deneyler sonucunda belirlenen ağırlıklı toplama ile gerçekleştirilmiştir. Test kümesinde önerilen sistem ve belirtilen konfigürasyon ile %94.29 geri getirme (recall) ve %97.14 kesinlik (precision) performansına ulaşılmıştır.

Şekil 3: "İç Mekan" kavramına ait örnek anahtar kareler.

Şekil 5: Sırasıyla Şehir, Gece ve Protesto kavramlarına ait örnek anahtar kareler.

Gerçekleştirilen ikinci deney senaryosunda TRECVID 2009 Üst Seviye Öznitelik Çıkarma [9] çağrısında kullanılan 280 saatlik toplam 61384 çekimden oluşan, MPEG-1 kodlanmış CIF boyutunda özgün video veritabanı kullanılmıştır. Veritabanını oluşturan videolar Hollanda televizyon kanallarında yayınlanan eğitim ve belgesel programlarından oluşmaktadır. Tanınacak kavramlar ise TRECVID 2009 çağrısında bulunan 20 kavram arasından, *Protesto, Şehir, Gece, Hareket halindeki uçak, gemi ve otobüs* deneyler için seçilmiştir. Sınıflandırıcıların eğitimi ve kod tablolarının oluşturulması için TRECVID 2008 videoları ve çekim bazındaki ilgili mutlak doğrulukları kullanılmıştır. Toplam 980 adet pozitif mutlak doğruluk verisi ve aynı sayıda rastgele seçilen negatif örnek eğitim için kullanılmıştır. Kavramlar ile ilgili örnek kareler Şekil 5'te görülebilir. Her ne kadar test kümesi 280 saatlik videoya sahip olsa da verinin tamamında etiketler mevcut değildir. Bu nedenle geri getirme ve kesinlik değerlerinin yanı sıra eksik etiket durumunda geri getirme performansının değerlendirilmesi için literatürde sıklıkla tercih edilen *çıkarımsal ortalama kesinlik* (infAP) [13] hesaplanmıştır. Deneylerde MPEG-7 öznitelik uzayına ait kod tablosu 128 öbek ve SIFT öznitelik uzayı 256 öbekli kod tablosu, her bir kavram için ayrı ayrı oluşturulmuştur. Her bir kavram için elde edilen 2000 çekimlik sonuçların oranları ve infAP değerleri sırasıyla Şekil 4 ve Şekil 6'da gösterilmektedir.



Şekil 4: TRECVID 2009 test verisinde elde edilen geri getirme oranları. Ortalama geri getirme oranı %31.50 olarak hesaplanmıştır.

4. Sonuçlar

Bu çalışmada video görüntülerinden kavram tanıma için kod tablosu ve SVM tabanlı bir sistem önerilmiştir. Sistemin performansının değerlendirilmesi için iki farklı test senaryosu denenmiştir. Her ne kadar İç Mekan kavramı için umut verici sonuçlar alınsa da TRECVID 2009 gibi geniş bir veri kümesinde düşük performans gözlemlenmiştir. TRECVID 2009 test kümesindeki düşük performansı, kullanılan kod tablolarının verinin karmaşıklığına göre düşük kalan boyutuna ve az sayıdaki eğitim kümesi elemanına (örn: *otobüs* kavramı için eğitim kümesinde 47 çekim bulunmaktadır) bağlayabiliriz. Performansı artırmak için eğitim verisini artırmanın yanı sıra bazı kavramlar için geçerli olmak üzere ses verisinin kullanılması da gelecek çalışmalarda hedeflenmektedir.

5. Kaynakça

[1] C. G. M. Snoek, M. Worring, J.-M. Geusebroek, D. C. Koelma, F. J. Seinstra, and A. W. M. Smeulders. The semantic pathfinder: Using an authoring metaphor for generic multimedia indexing. IEEE Trans. PAMI, 28(10):1678–1689, 2006.

[2] C. G. M. Snoek, K. E. A. van deSande, et. al. The MediaMill TRECVID2008 semantic video search engine, 2008.

[3] C. Bovik, M. Clark, and W. S. Geisler. Multichannel texture analysis using localized spatial filters. IEEE Trans. PAMI, 12(1):55–73, 1990.

[4] J.-M. Geusebroek. Compact object descriptors from local colour invariant histograms. In BMVC, Edinburgh, UK, 2006.

[5] Y.-G. Jiang, C.-W. Ngo, J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval", ACM CIVR, 2007.

[6] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets", ECML, Sep. 2004.

[7] Y. Peng, Z. Yang, J. Yi, L. Cao, H. Li, and J. Yao, "Peking University at TRECVID 2008: High Level Feature Extraction" in TRECVID Proceedings, 2008.

[8] S. Chang, J. He, Y. Jiang, A. Yanagawa, and E. Zavesky, E. El Khoury, C. Ngo, "Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search," in TRECVID Proceedings, 2008.

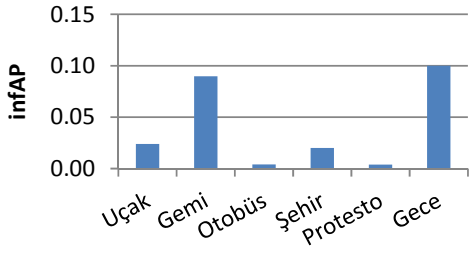
[9] "TREC Video Retrieval Evaluation Home Page," [Çevrimiçi]. http://www-nlpir.nist.gov/projects/trecvid/ [Son Erişim: 15 Ekim 2009].

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," Int'l J. Computer Vision, Nov. 2004.

[11] B. S. Manjunath, P. Salembier, Thomas Sikora, "Introduction to MPEG-7 Multimedia Content Description Interface,".

[12] A. Quattoni, and A.Torralba. Recognizing Indoor Scenes. IEEE CVPR, 2009.

[13] E. Yilmaz ve J. A. Aslam. "Estimating average precision with incomplete and imperfect judgments." In Proceedings of the 15th ACM International Conference on Information and Knowledge Management, 2006.



Şekil 6: TRECVID 2009 test verisinde elde edilen infAP performansı.