Sahne Tanıma için Çoklu Alan Seçimi Tabanlı Bir Yaklaşım
A Multiple Region Selection Based Approach for Scene Recognition

Ezgi EKİZ, Nazlı İKİZLER CİNBİŞ
Bilgisayar Mühendisliği Bölümü
Hacettepe Üniversitesi
Ankara, Türkiye
{n12124212,nazli}@cs.hacettepe.edu.tr

***Özetçe* —Sahne tanıma problemi, bilgisayarlı görünün sıklıkla çalışılan alanlarından biridir. Bu çalışmada, bu probleme çoklu örnekle sınıflandırma ve pencere seçimi tabanlı bir yöntem önermekteyiz. Önerilen yöntem, öncelikle verilen bir görüntü içerisinden anlamlı ve ayırt edici olabilecek alt pencereler seçmekte, sonrasında çıkarılan bu alt alan bilgilerini Çoklu Örnekle Öğrenme yapısı içinde ele almaktadır. 15-Sahne denektaşı kümesinde yapılan deneyler sonucunda elde edilen sonuçlar, önerilen yöntemin sınıflandırma performansı açısından umut verici olduğunu göstermektedir.**

Anahtar Kelimeler—sahne tanıma, çoklu alan seçimi, öznitelik seçimi, çoklu örnekle öğrenme.

***Abstract*—Scene recognition is a frequently-studied topic of computer vision. In this work, we propose a solution to this problem that involves multiple-region selection and multiple instance classification. In the proposed approach, first, meaningful and discriminative sub-regions are extracted and then, information coming from these regions are considered within a Multiple Instance Learning framework. The results obtained via the tests performed on 15-Scenes benchmark dataset show that the proposed approach is promising for the classification performance.**

Keywords—scene recognition, multi-region selection, feature selection, multiple-instance learning.

I. GİRİŞ

Sahne tanıma, verilen bir görüntünün çekildiği ortama ve içinde bulunduğu sahne bilgisine ait bir semantik etiketin görüntünün görsel içeriği üzerinden otomatik olarak elde edilmesi olarak tanımlanabilir. Bu problem, bilgisayarlı görünün sıklıkla çalışılan problemlerinden biridir. Sahne tanıma, görüntüdeki pek çok farklı etkenden etkilenebilir. Bu etkenler arasında en önemlilerinden biri, sahnenin ayırt edici bir şekilde fotoğraflanmamış olması ve görüntüdeki karmaşıklığın, sahnenin genelini tanımaya engel olmasıdır. Pozlama ve perspektifteki farklılıklar, sahnelerin farklı bakış açılarından tanınmasını zorlaştırmaktadır.

Bu çalışmamızda, bu farklı bakış açılarının yarattığı tanıma problemine çoklu örnekle öğrenme tabanlı bir çözüm yaklaşımı sunmaktayız. Bu amaçla, bu çalışmada verilen bir görüntüden farklı alt alanlar seçerek, çoklu alanlı bir tanıma modellemesi önerilmektedir. Bu bağlamda, öncelikle, verilen bir görüntüden olası sahne bilgisinin ayırt edici olabileceği alt pencereler oluşturulmaktadır. Bu alt pencereler, ileriki bölümlerde anlatılacak kriterlere göre akıllı bir biçimde seçilmekte, sonrasında, her bir görüntü, bu alt pencere örneklemlerinden oluşan bir torba (bag) olarak değerlendirilmektedir. Bu torbalar, sonrasında Çoklu Örnekle Öğrenme (ÇÖÖ) yöntemi ile ele alınmakta, ve ilgili sahne sınıflandırma modelleri oluşturulmaktadır.

Bu çalışmada önerilen yöntemin genel akışı Şekil 1'de gösterilmektedir. Burada görüldüğü üzere, verilen bir görüntüden önce nesne bilgisinin yoğun olarak gözlemlenebileceği ve sahne tanıma için aday olarak nitelendirilebilecek alt pencereler oluşturulmaktadır. Sonrasında, bu aday pencereler, büyüklük, kenar yoğunluğu ve çakışma (overlap) gibi kriterlere göre ayıklanmakta, ve sahne bilgisini daha çok içermesi muhtemel pencereler oluşturulmaktadır. Sonrasında bu pencereler, Çoklu Örnekle Öğrenme yöntemi ile sınıflandırmaya tabi tutulmaktadır.

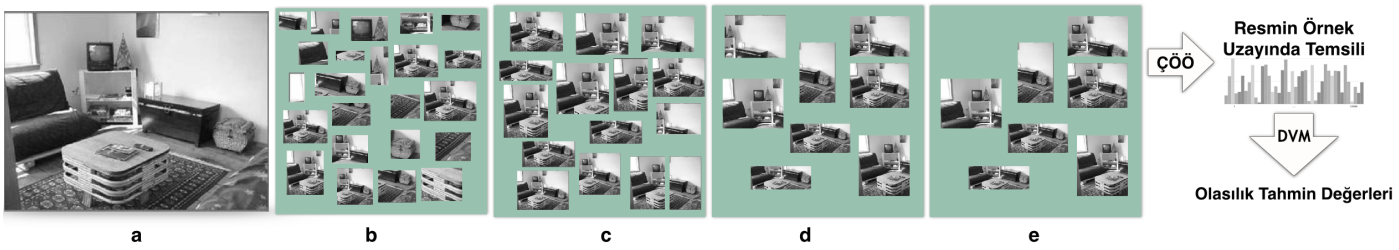
Önerdiğimiz yöntemin, sınıflandırmadaki başarısı, sahne tanıma problemi için önemli bir denektaşı kümesi olan 15-Sahne (15-Scenes) veri kümesi üzerinde değerlendirilmiş, ve önerilen yönteminin sınıflandırma başarımının umut verici olduğu gözlemlenmiştir.

II. İLGİLİ ÇALIŞMALAR

Sahne tanımada yaygın olarak kullanılan yaklaşımlardan birisi, bir sahneye ait resmi bir bütün olarak ifade etmektir, buna örnek olarak oldukça popüler olan GIST [1] özniteliği gösterilebilir. Benzer bir yaklaşım olarak Siagian ve Itti [2] belirginlik (saliency) bilgisini ve resmin genel yapısına ait uzamsal bilgiyi, bir resmin yönelim, parlaklık ve renk gibi çeşitli alt seviyeli öznitelikleri üzerinden oluşturarak çift yönlü bir temsil elde etmektedir. CENTRIST [3] ise, alt seviye piksel komşulukları üzerinden bir Sayım Dönüşüm Dağılımı (Census Transform Histogram) oluşturarak, resme dair uzamsal yapıları ve kaba geometrik bilgiyi ifade etmektedir.

Resimleri bir bütün olarak ifade etmenin yanısıra, içinde bulunan parçalar üzerinden tanımlamaya çalışmak da sahne tanıma problemine bir başka yaklaşımdır. Örneğin Lazebnik

978-1-4673-7386-9/15/$31.00



Şekil 1: Önerilen sahne tanıma yaklaşımının genel akışı. Verilen bir resimden (a) çıkarılan aday alt pencereler (b), büyüklük (c), çakışma (d), kenar yoğunluğu (e) gibi kriterlere göre seçilip, sonrasında çoklu örnekle öğrenme amacı ile görüntüyü ifade etmekte kullanılmaktadır.

vd.nin uzamsal piramit eşleme [4] (spatial pyramid matching) yaklaşımı, resmi farklı seviyelerde eşit alt parçalara ayırıp, bu alt parçaların temsillerini birlikte kullanarak, resmi bir bakıma alt alanları cinsinden ifade etmektedir ve bu yaklaşım, bir çok diğer çalışmanın da altyapısında kullanılmaktadır. Örnek olarak, Xiao vd. [5] bu temsil biçimini bir çok farklı öznitelik ve Destek Vektör Makineleri (DVM) çeşidi ile birlikte kullanmaktadır. Ayrıca, [4] çalışmasının resmi eşit alt alanlara bölme yaklaşımına farklı bir yorum olarak, Jiang vd. [6] rastgele bir çok uzamsal bölütlendirme oluşturmakta ve bunların hangilerinin temsil edici olabileceğine karar vermeye çalışmaktadır. Benzer biçimde, [7] çalışması da bir çok bölütlendirme arasından hangilerinin daha ayırt edici olduğuna karar verirken, bu bölütlemelerin konumlarını gizli değişken olarak ele almaktadır. Resimleri parçalar üzerinden ifade etmeye örnek olan bir başka yaklaşım, Torralba vd.nin [8] resimleri bölütlendirme tabanlı bir kök alanı ve bu kök alanın etrafındaki bir çok yardımcı alan üzerinden modeller oluşturarak ifade eden çalışmasıdır. Pandey vd. de [9] sahne tanımaya yardımcı olabilecek nesnelerin yerlerini gizli DVM (latent SVM) yardımıyla bulmaya çalışmaktadır.

Sahne tanıma problemine bir başka yaklaşım ise sahneleri bir nesne kümesi ya da orta düzeyli resim parçaları cinsinden ifade etmektir. Bu anlamda en temel yaklaşım olarak, Nesne Bankası [10] (Object Bank) yöntemi resimleri daha önceden öğrenilmiş nesne modellerine verdiği tepkiler üzerinden ifade etmektedir. Daha önceden öğrenilmiş nesneye dayalı modellerin kullanılmasına alternatif popüler bir yaklaşım ise resimlerin ait oldukları sahneye dair ayırt edici orta-düzey özniteliklerin öğrenilmesidir [11-14]. Bu metodlardaki temel yaklaşım, bir nesnenin görünümünün çeşitli nedenlerden dolayı değişiklik gösterebileceği ve bu nedenle nesneleri daha önceden tanımlanmış tek bir modelle ifade etmek yerine sahip oldukları parçalar cinsinden ifade etmenin daha uygun olduğudur.

III. YÖNTEM

Önerilen yöntem üç ana adımdan oluşmaktadır. İlk olarak anlamlı alt pencere çıkarımı yapılmakta ve bunların belli kriterlere göre elemesi yapılarak her bir resim alt pencerelerden oluşan bir torba olarak ifade edilmektedir. Sonrasında, bu alt pencere torbaları üzerinde benzerlik tabanlı kodlama yapılarak, torbalar Çoklu-Örnekle Öğrenme yapısı içinde benzerlik uzayına taşınmakta ve bu uzay üzerinde DVM kullanılarak sınıflandırma modelleri öğrenilmektedir. Aşağıda, her bir adım detaylı olarak anlatılmıştır.

A. Alt Pencerelerin Seçilmesi

Bu kısımda resimlerden sahne tanıma problemi için anlamlı olabilecek, bir başka deyişle, resmin ait olduğu sahnenin karakterini yansıtabilecek alt pencerelerin çıkarılması amaçlanmaktadır. Bunun için öncelikle Seçmeli Arama Algoritması [11] kullanılarak yüksek miktarda aday alt pencere çıkarımı yapılmaktadır. Bu algoritma, resimlerde bulunan nesneleri bulmak için resimleri yüksek bölütlemeye (oversegmentation) tabi tutmakta ve elde ettiği bu küçük bölgelerin benzerlikleri üzerinden hiyerarşik bir yapı oluşturmaktadır. Buradaki benzerlikler bir çok alt düzey öznitelik (örneğin renk, doku ve parlaklık gibi) arasında çeşitli uzaklık metriklerinin birlikte kullanılması ile hesaplanmaktadır, dolayısıyla bahsedilen hiyerarşi oluşturulurken bir nesneye ait olan bölgelerin birden fazla açıdan ön plana çıkarılması ve anlamlı alan olarak işaretlenmesi olasılığı yükselmektedir. Seçmeli Arama Algoritması, elde edilen hiyerarşinin farklı seviyelerini kullanarak büyüklük bakımından çeşitlilik gösteren bir çok alt pencere ortaya çıkarmaktadır. Bu noktada, Seçmeli Arama algoritmasının oluşturduğu aday alt pencere sayısı oldukça fazladır ve boyuttaki bu büyüklük, etkili bir sınıflandırma yapmayı engellemektedir. Bizim yöntemimizde, bu algoritma tarafından bulunan aday alt pencereler arasından, sahneyi ifade edebilecek şekilde büyük ve anlamlı alanlar seçilmektedir.

Algoritma 1 Pencere eleme algoritması: Eğer iki pencereden küçük olanı (S_j) büyük olan (S_i) ile büyük oranda kesişiyorsa küçük pencere kümeden çıkartılmaktadır.

```
S ← alana göre büyükten küçüğe sıralanmış pencere kümesi
for i ← 1 to size(S) − 1 do
    for j ← i + 1 to size(S) do
        if area(S_i ∩ S_j) ≥ area(S_i) × θ then
            S ← S − S_j
            j ← j − 1
```

Şekil 1'de Seçmeli Arama Algoritması tarafından bulunan pencerelerden sahne tanımaya uygun olanların nasıl seçildiği gösterilmektedir. Görüldüğü gibi, ortaya çıkan çok sayıda pencere arasından yalnızca belirli büyüklükte olanlar seçildiğinde, elimizde kalan kümede birbirine oldukça benzeyen pencereler bulunmaktadır. Bu noktada hem gereksiz bilgiyi ortadan kaldırmak, hem de elde edilen alt pencere sayısını olabildiğince azaltmak adına, birbirine oldukça benzeyen pencerelerden bir tanesinin elenmesini sağlayacak bir algoritma kullanılmıştır (Algoritma 1). Bu algoritma, eğer

iki pencerenin kesişimine ait alan büyük olan pencerenin büyük bir yüzdesini kapsıyorsa (θ), küçük olan pencereyi seçili alt pencereler kümesinden çıkarmaktadır.

Seçmeli Arama Algoritması kullanılarak elde edilen alt pencereler, yeterli büyüklük ve farklılıkta olsalar bile, bazı pencerelerin içerdikleri boş alanlardan dolayı (örneğin gökyüzü, kırsal kesim, vb. gibi) yeterli görsel bilgiyi sağlamadıkları söylenebilir. Bu tip yeterli görsel içerik sağlamayan pencereleri elemek için, çalışmamızda ek olarak resimdeki kenar bilgilerini kullanan bir eleme yöntemi ele alınmıştır. Bu amaçla, Dollar vd.nin kenar bulma yöntemi [12] kullanılarak resimlerdeki kontür bilgisi, bir başka deyişle herhangi bir kenara ait olan piksellerin kümesi (E) elde edilmiş ve bu bilgi Denklem 1'deki biçimde, bir resimdeki (I) kenar piksellerinin yoğunluğunu (KY) bulmak için kullanılmıştır.

$$KY = \frac{|E|}{|I - E|} \tag{1}$$

Bu denklemde $|E|$ alt pencere içinde bulunan kenar piksel sayısını, $|I - E|$ ise kenar olmayan alanların piksel sayısını ifade etmektedir. Böylece, belirli bir yoğunluğun altında değere sahip olan pencereler, kümeden elenmektedir.

B. Özniteliklerin Çıkarılması

Bir önceki adımda elde edilen alt pencerelerin içerdiği bilgiyi ifade edebilmek için iki farklı görsel öznitelik kullanılmıştır. Bunlardan birincisi GIST [1] özniteliği, elde edilen pencerelerin genel geometrik yapısını ifade etmek için kullanılmıştır. Kullanılan ikinci öznitelik olan HOG2 $\times$ 2 [5] ise, kelime torbaları (bag-of-words) yapısında elde edilen bir özniteliktir. Bu özniteliği elde edebilmek için resimler 8×8 piksellik hücrelere bölünmüş ve birbiri ile örtüşen 2×2'lik hücrelerin HOG özniteliklerinin peşpeşe eklenmesi ile bir öznitelik uzayı elde edilmiştir. Bu uzayda k=300 olacak biçimde *k*-ortalamalar yöntemi kullanılmış ve elde edilen kümelerin merkez noktaları ile bir kütüphane oluşturulmuştur. Çalışmamızda elde edilen pencereler, 3 farklı uzamsal piramit seviyesi için elde edilen bu kütüphanedeki görsel kelimeler cinsinden ifade edilmektedir.

C. Çoklu Örnekle Öğrenme

Elde edilen alt pencerelerden hepsinin olmasa da, bir kısmının ait olduğu sahneler ile ilgili bilgi verdiği varsayılabilir. Bu nedenle, bu alt pencerelerin hangileri olduğunu bulmak ve bir resmi birden fazla alt penceresine dayanarak ifade edebilmek için Çoklu Örnekle Öğrenme yaklaşımı kullanılmıştır. Bu yaklaşımda, her resim bir torba olarak ele alınmakta ve o resmin her alt penceresi bu torbanın bir örneği olarak ifade edilmektedir. Pozitif olarak işaretlenen bir torbanın en az bir pozitif örnek içerdiği varsayılırken, negatif torbalarda bulunan bütün örneklerin de negatif olduğu varsayılmaktadır.

Pozitif bir torba içerisinde bulunan örneklerden hangisinin pozitif olduğu bilinmediği için, torbaların daha önce elde edilen öznitelikler yerine içerdikleri örnekler cinsinden ifade edilmesi, aynı sınıfa ait torbalar arasında paylaşılan benzer örneklerin ön plana çıkarılması için uygun bir yaklaşımdır. Bu amaçla [13] çalışmasındaki Çoklu Örnekle Öğrenme yaklaşımı probleme uygun olarak kullanılmıştır. Burada, her torba, bütün torbalarda bulunan örnekler cinsinden ifade edilmektedir. Bu amaçla bir torba (B_i) ve örnek (x^k) arasındaki benzerliğin hesaplanması gerekmektedir, böylece bir torba bütün örneklere olan benzerliği üzerinden ifade edilebilir (Denklem (2)). İlgili denklemde, $D(x_{ij}, x^k)$, B_i torbasının j'inci örneği ile x^k örneği arasındaki uzaklıktır ve çalışmamızda χ^2 uzaklığı kullanılarak elde edilmiştir; σ ise bant genişliği parametresi olup, uzaklıkların ortalaması alınarak hesaplanmaktadır.

$$s(x^k, B_i) = \max_j \exp\left(-\frac{D(x_{ij}, x^k)}{\sigma^2}\right) \tag{2}$$

Böylelikle, her torba tüm torbalarda bulunan örneklere (toplam N adet) olan benzerliği cinsinden ifade edildiğinde, yeni bir kodlama biçimi/öznitelik elde edilmektedir:

$$m(B_i) = [s(x^1, B_i), \ldots, s(x^N, B_i)]^T \tag{3}$$

Torbaların tüm örneklere olan benzerlikleri cinsinden ifade edilmesi ile, elde edilen bu yeni öznitelik üzerinden sahne tanımlama amaçlı sınıflandırıcıların öğrenilmesi mümkün hale gelmektedir. Bu amaçla, her sınıf için bir DVM modeli öğrenilmiş ve bir test resmine bu modeller arasından en yüksek skoru veren modele ait sınıfın etiketi atanmıştır. Bahsedilen bu skorlar aynı zamanda daha sonra sonradan birleştirme amacı ile de kullanılmaktadır (bkz: Bölüm IV).

Yukarıda bahsedilen kodlama biçiminde bir torbanın ifade edilmesinde onun örnek uzayındaki boyutlara en yakın örneği kullanıldığı için, sınıflara ait modellerin öğrenilmesi ile (bir başka deyişle elde edilen örnek uzayında tanımlı boyutların sınıflandırmada sahip oldukları ağırlıkların öğrenilmesi ile) torbanın içerisindeki hangi örneğin o torbanın sınıflandırılmasına en fazla katkıyı yaptığı incelenebilir. Bu katkı ($g(x_{ij^*})$), öğrenilen DVM ağırlıklarının ((w_k^*)) ilgili örneğe ait benzerlik değeri (x_{ij}) ile çarpılması ile hesaplanabilir:

$$g(x_{ij^*}) = \sum_{k \in \mathcal{I}_{j^*}} \frac{w_k^* s(x^k, x_{ij^*})}{m_k} \tag{4}$$

burada m_k, i torbası içerisinde k örnek uzayı boyutuna minumum uzaklığı veren örneklerin sayısı olup, $\mathcal{I}_j^*$ de ($\mathcal{I}$ bütün örnek uzayı boyutlarını ifade etmek üzere); i torbasındaki örneklerin hangi örnek uzayı boyutuna en yakın olduklarını tutan listedir.

Şekil 2'de, *mutfak* ve *oturma odası* sınıfları için en yüksek katkıyı yapan pencerelere örnekler verilmiştir. Burdaki örneklerde görülebileceği gibi, resimlerde daha karakteristik bilgi içeren alt pencereler üzerinden sınıflandırma yapılmaktadır.

IV. DENEYLER

Deneylerde kullanılan 15-Sahne veri kümesinde [4], 15 farklı iç ve dış mekan sahnesine ait toplam 4485 adet siyah beyaz resim bulunmaktadır. Bu resimler arasından, [5] çalışmasında kullanılan birinci bölütlemeye uygun olarak, her sınıf için 100 adet resim pozitif olarak seçilip öğrenme sırasında kullanılmış, geri kalanları ise test amaçlı kullanılmıştır. Seçmeli arama algoritması için, asgari bölüt büyüklüğü 50 piksel, alt pencerelerin azami büyüklüğü resmin 0.3'ü ve alt pencerelerin örtüşme oranı (θ) 0.7

Şekil 2: Resimlerin etiketlendirilmesinde en yüksek paya sahip olan alt pencerelere örnekler. Üst sıra, *mutfak* sınıfına ait örnekleri, alt sıra *oturma odası* sınıfına ait örnekleri göstermektedir.

Tablo I: Farklı pencere seçme yöntemleri ile sahne sınıflandırma sonuçları.

Yöntem	HOG2×2	GIST	Sonradan Birleştirme
DVM	75.37	66.36	78.72
PTST	80.63	70.51	82.25
PTST-Kenar	80.31	71.47	**82.50**

olarak belirlenmiştir. Kenar yoğunluğuna göre alt pencereler seçilirken ise eşik değeri 0.5 olarak seçilmiştir.

Kullanılan yöntemlerin HOG2×2 ve GIST öznitelikleri ile elde ettikleri doğruluk değerleri Tablo I'de görülebilir. Burada, DVM ile belirtilen yöntem, pencere oluşturma ve seçim adımları olmaksızın, resmin tamamından çıkartılan özniteliklerin lineer DVM'ler kullanılarak sınıflandırılmasını ifade etmektedir. PTST ile ifade edilen yöntem, önerilen Pencere Tabanlı Sahne Tanıma yöntemi, PTST-Kenar ise, pencere seçim bölümünde kenar bilgisinin ek olarak kullanıldığı yöntemdir. Bu üç yöntemin sınıflandırıcılarının oluşturulması esnasında, hepsi için eğitim kümesi üzerinde 5-katmanlı çapraz geçerleme yöntemi ile parametre optimizasyonu yapılmıştır.

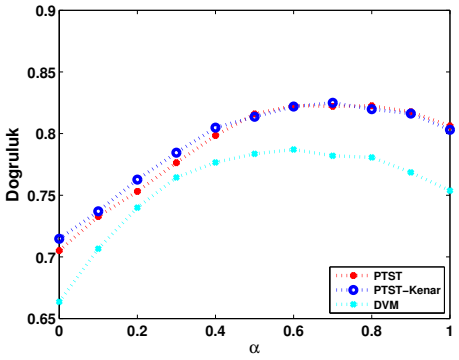
Farklı öznitelikler kullanılarak elde edilen sınıflandırma modellerin içerdikleri bilgiyi birlikte kullanabilmek adına, sonradan birleştirme (*late fusion*) yöntemi kullanılmaktadır. Bu yöntemde, bir özniteliğe ait modellerin bir test resmine ait etiketi belirlerken ortaya koydukları olasılık tahminleri kullanılmaktadır. Birinci özniteliğe (HOG2×2) ait modellerin verdiği olasılık değerleri ile (c_1), ikinci özniteliğe (GIST) ait modellerin verdiği olasılık değerleri (c_2), α ağırlık parametresi kullanılarak şu şekilde

$$c_f = \alpha c_1 + (1-\alpha)c_2 \tag{5}$$

lineer olarak birleştirilmektedir. Bu birleştirme sırasında kullanılan ağırlık parametresi α'nın etkisi, Şekil 3'te görülebilir. Görüldüğü gibi genel olarak, seçilen pencereler üzerinde Çoklu Örnekle Öğrenme yaklaşımı, resmin bütünün tek başına sınıflandırılmasına göre daha başarılı sonuçlar vermektedir. Buna ek olarak, kenar bilgisini pencere seçimi dahilinde kullanmak sınıflandırmaya katkı sağlamaktadır.

V. SONUÇLAR VE DEĞERLENDİRME

Bu çalışmamızda, sahne tanıma problemine görüntüden çıkartılan aday alt pencerelerin seçimi vasıtasıyla oluşturulan torba yapısı üzerinden Çoklu Örnekle Öğrenme gerçekleştiren



Şekil 3: Doğruluğun ağırlık parametresi α'ya göre değişimi.

bir yöntem önermekteyiz. Önerilen yöntemin sınıflandırma başarımı, 15-Sahne denektaşı kümesinde doğrulanmış, ve önerilen yöntemin sahne tanıma açısından umut verici olduğu görülmüştür. Alt pencere seçiminde, kenar bilgisi yoğunluğunun kullanılmasının daha verimli sonuçlar oluşturduğu gözlemlenmiştir.

TEŞEKKÜR

Bu çalışma TÜBİTAK tarafından 112E149 no'lu Kariyer projesi kapsamında desteklenmiştir.

KAYNAKÇA

[1] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *International Journal of Computer Vision*, vol. 42, no. 3, pp. 145–175, 2001.

[2] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 300–312, Feb. 2007.

[3] J. Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1489–1501, 2011.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006, pp. 2169–2178.

[5] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "SUN database: Large-scale scene recognition from abbey to zoo," in *CVPR*, 2010, pp. 3485–3492.

[6] Y. Jiang, J. Yuan, and G. Yu, "Randomized spatial partition for scene recognition," in *ECCV*, 2012, pp. 730–743.

[7] F. Sadeghi and M. F. Tappen, "Latent pyramidal regions for recognizing scenes," in *ECCV*, 2012, pp. 228–241.

[8] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in *CVPR*, 2009, pp. 413–420.

[9] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in *ICCV*, 2011, pp. 1307–1314.

[10] L. Li, H. Su, E. P. Xing, and F. Li, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in *NIPS*, 2010, pp. 1378–1386.

[11] J. R. R. Uijlings, K. E. A. van de Sande, T. Gevers, and A. W. M. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[12] P. Dollár and C. L. Zitnick, "Fast edge detection using structured forests," *PAMI*, 2015.

[13] Y. Chen, J. Bi, and J. Z. Wang, "MILES: multiple-instance learning via embedded instance selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, 2006.