

KELİME HİSTOGRAM MODELİ İLE HİSTOPATOLOJİK GÖRÜNTÜ SINIFLANDIRILMASI

HISTOPATHOLOGICAL IMAGE CLASSIFICATION WITH THE BAG OF WORDS MODEL

Erdem Özdemir¹, Cenk Sökmensüer², Çiğdem Gündüz-Demir¹

¹ Bilgisayar Mühendisliği Bölümü
Bilkent Üniversitesi
erdemo.gunduz@bilkent.edu.tr

² Patoloji AnaBilim Dalı
Hacettepe Tıp Fakültesi
csokmens@hacettepe.edu.tr

ÖZETÇE

En yaygın kanser türlerinden biri olan kolon kanserinin tedavisi erken tanı ile mümkün olabilmektedir. Günümüzde kanser tanısında kolonoskopi, sigmoidoskopi ve stool testi gibi görüntüleme yöntemleri kullanılmakta ise de, en yaygın kullanılan ve geçerli yöntem, dokulardan biyopsi işlemi ile doku kesitlerinin alınması ve bu kesitlerin mikroskop altında incelenmesidir. Ancak bu inceleme, görsel yorumlamaya dayalı olduğundan dolayı, patoloğlar arasında öznel kararların verilmesine neden olabilmekte ve tanı farklılıklarına yol açabilmektedir. Patoloğlar arasındaki kararlardaki değişkenliği azaltma ihtiyacı, bilgisayar yardımı ile biyopsi görüntüleri üzerinde öznitelik çıkarımı ve nesnel kararlar verilmesini sağlayacak algoritmaların geliştirilmesi için çalışmalara yol açmıştır. Bu bildiride, biyopsi görüntü gösterimi için öğreticisiz olarak belirlenen görüntü bölgelerinin birlikteliğini modelleyen histogramların tanımlanması ve bunlar üzerinde çıkarılacak öznitelikler ile kolon doku görüntülerinin otomatik sınıflandırılması için bir yöntem önerilmiştir. Literatürde kelime histogramı (bag-of-words) modeli olarak da bilinen bu yöntem ile, kolon doku görüntüleri üzerinde yaptığımız deneysel çalışmalar, önerilen bu yöntemin benzer yöntemlerle karşılaştırıldığında daha başarılı sonuçlar verdiğini göstermiştir. Bununla beraber, öznitelik tanımlamasında piksel renk yoğunluk değerlerini kullanan yöntemimizin, farklı özniteliklerin beraber kullanılmasıyla daha iyi sonuçlar verme potansiyeli de bulunmaktadır.

ABSTRACT

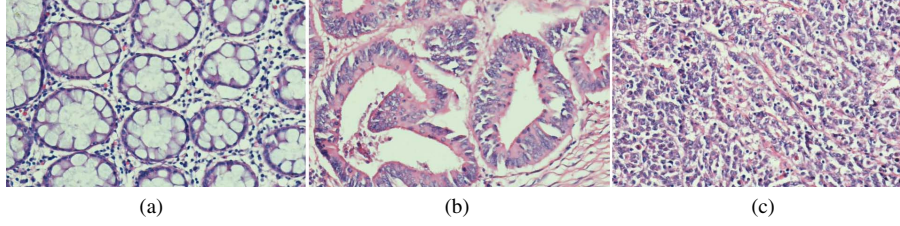
Colon cancer, which is one of the most common cancer type, could be cured with its early diagnosis. In the current practice of medicine, there are many screening techniques such as colonoscopy, sigmoidoscopy, and stool test, however the most effective and most widely used method for cancer diagnosis is to take tissue sections with biopsy and examine them under a microscope. As this examination is based on visual interpretation, it may lead to subjective decisions and diagnostic differences among pathologists. The need of reducing inter-variability in cancer diagnosis has led to studies for extraction of features from biopsy images and development of algorithms that give objective results. In this paper, we propose a method for the

automated classification of a colon tissue image with the features extracted from a histogram that models the existence of image regions determined in an unsupervised way. Experiments on colon tissue images show that the proposed method leads to more successful results compared to its counterparts. Moreover, the proposed method, which uses color intensities for feature extraction, has the potential of giving better results with the use of additional features.

1. GİRİŞ

Kanserin en yaygın türlerinden biri olan kolon kanseri için ABD verilerine göre 2010 yılında 100.000'in üzerinde yeni vaka sayısı beklenmektedir ve 2010 yılı için 50.000 üzeri kişinin kolon kanserinden öleceği tahmin edilmektedir [1]. Kolon kanseri tedavisinde erken teşhis önemli yer tutar. Erken teşhis için kolonoskopi gibi tarama amaçlı farklı testler bulunmasına rağmen, kesin tanı için rutin uygulama biyopsi alınması ve bunların mikroskop altında görsel incelenmesidir. Doktorlar, biyopsi görüntülerini incelerken, dokuda yer alan kolon bezlerinin yapılarına bakarak kanserin olup olmadığına karar verirler. Bu bezlerdeki bozulma seviyesi ise kanserin derecesini belirler. Örneğin düşük dereceli kanserde bezlerin oluşturduğu yapı hala seçilebilir durumdadır. Yüksek dereceli kanserde ise bu yapılar tamamen bozulur ve seçilemez hale gelir. Şekil 1'de, normal, düşük dereceli kanserli ve yüksek dereceli kanserli dokular için örnek görüntüler verilmiştir. Kolon bezlerine örnek, bu bezlerin sınırları düşük dereceli kanserli bir görüntü üzerinde çizilerek Şekil 2'de gösterilmiştir.

Histopatolojik görüntülerin sınıflandırılması için literatürde çeşitli yöntemler kullanılmıştır. Bu yöntemler görüntüler üzerinde çıkarılan özniteliklere göre başlıca üçe sınıflandırılabilir. Bunlar renksel/dokusal öznitelik kullananlar [2], morfolojik öznitelik kullananlar [3, 4] ve dokuda yer alan bileşenlerin uzaysal dağılımına dayalı öznitelik kullananlardır [5, 6]. Morfolojik özniteliklerin çıkarımı hücrelerin bölütlenmesini, büyüklük ve şekil özelliklerinin sayılaşdırılmasını içerir [4]. Dokuda bulunan bileşenlerin uzaysal dağılım bilgisini kullanan yöntemler, bu yapıların yerlerinin belirlenmesini, dağılımı ifade etmek için bu bileşenler üzerinde çizge, ağaç, üçgenleme gibi yapıları tanımlamayı



Şekil 1: Kolon doku görüntü örnekleri: (a) normal, (b) düşük dereceli kanserli ve (c) yüksek dereceli kanserli.



Şekil 2: Düşük dereceli kanserli bir görüntü için bez yapıları.

ve bunlardan öznitelik çıkarımını içerir [5, 6]. Bu iki yöntem için temel zorluk, hücrelerin (ve gerekli olan durumlarda diğer bileşenlerin) bölütlenerek yerlerinin tespit edilmesi gerekliliğidir. Renksel/dokusal yöntemlerde bu zorluk olmamasına karşın, bunlar için de iki temel husus göz önüne alınmalıdır. Bunlardan ilki, doku görüntülerinde standart boyamadan (hematoksilin-ve-eosin boyamasından) kaynaklanabilecek renk yoğunluk değerlerindeki olası farklılıktır. Diğer husus ise, görüntülerde bulunan ancak ilgili kanser tanısında etkisi olmayan bölgelerdir. Örneğin, bu bildiri üzerinde çalışılan ve tüm kolon kanserlerinin yüzde 90-95'ini oluşturan kolon adenokanserinde tanı konması için temel olarak bez yapılarına bakılır. Dolayısıyla bez dışındaki yapıların renk yoğunluk değerleri ve dokusal özellikleri yanıltıcı olabilir.

Bu çalışmamızda, biyopsi görüntülerinde bulunabilecek altbölgeleri (kelimeleri) otomatik olarak bulmayı ve bu altbölgelerin bulunma sıklığını gösteren kelime histogramları için öznitelik çıkartarak görüntüleri sınıflandırmayı önermekteyiz. Kelime histogramı (bag-of-words) modeli olarak da bilinen bu yaklaşım doğal dil işleme alanı için önerilmiş ve son zamanlarda farklı bilgisayarla görme problemleri için de kullanılmıştır [7, 8]. Bu yaklaşım medikal görüntülerin gösterimi ve sınıflandırılması için kullanımı oldukça kısıtlıdır. Kelime histogramı yaklaşımında ilk olarak belirli bir kelime dağarcığını oluşturulacak kelimeler tespit edilir; bilgisayarla görme problemlerinde bu kelimeler belirli bölgeleri ifade eden sayısal vektörlerdir. Kelime dağarcığını oluşturan kelimeler bulunduktan sonra, görüntü bu kelimelerin ne kadar sıklıkla geçtiği bilgisıyla ifade edilir.

Bu çalışmamızda, doku görüntüleri, eşit dikdörtgen altbölgelere bölünüp kelimeler bu altbölgelerde bulunan mor, pembe ve beyaz renklerin¹ bulunma sıklığı

¹Hastanelerde standart boyama olan hematoksilin-ve-eosin tekniği, görüntülerdeki piksellerin morumsu, pembemsi ve beyazımsı renk almasına yol açar.

ile tanımlanmıştır. Kullanılan görüntüyü altbölgelere bölme yaklaşımı, görüntüyü bir bütün olarak incelemeye göre, histopatolojik doku görüntüleri için daha başarılı sonuçlar vermektedir. Görüntüyü bir bütün olarak inceleyen yöntemler görüntülerdeki farklı bölgelerin özelliklerini birleştirirken, görüntüyü altbölgelere ayıran yöntemlerse bu farklı bölgeleri ayrı olarak değerlendirip daha iyi bir model oluşturabilmektedir. Ayrıca, histopatolojik doku görüntülerindeki piksellerin öbeklenmesi suretiyle elde edilen üç temel rengin kullanılması, bu görüntülerdeki renk yoğunluk değer farkları ile ilgili olası sorunun azaltılmasına katkı sağlamaktadır. Dolayısıyla uyguladığımız *kelime histogram* yaklaşımı, yukarıda bahsedilen renksel ve dokusal öznitelikleri kullanan yöntemleri etkileyen iki husustan kısmen daha az etkilenmektedir.

2. YÖNTEM

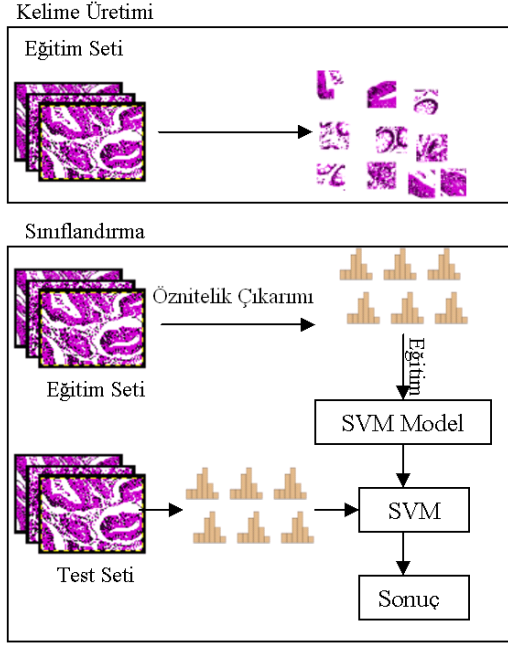
Önerilen sistemde, ilk olarak *kelime histogram* yaklaşımında kullanılacak kelimeler belirlenir. Daha sonra bu kelimelerin görüntülerde bulunma sıklığını tutan histogramlar oluşturulur. Bu histogramlar daha sonra destek vektör makinesinin (DVM) eğitiminde ve görüntülerin sınıflandırılmasında kullanılır. Önerilen sistemin genel çalışması Şekil 3'de gösterilmiştir.

2.1. Kelimelerin Oluşturulması

Kelimelerin oluşturulması için ilk aşama, k-ortalama algoritması kullanılarak her bir görüntünün kendi içinde üçe öbeklenmesidir. Bu öbekler hastanelerde biyopsilerin rutin boyanmasında kullanılan hematoksilin-ve-eosin tekniğinin verdiği morumsu, pembemsi ve beyazımsı piksellere karşılık gelir. Daha sonra, eğitim kümesi içindeki her bir görüntüden N tane rasgele dikdörtgen altbölge seçilir ve bu altbölgelerin öznitelikleri çıkarılır. Bu çalışmada çıkarılan öznitelikler seçilen altbölgenin içine düşen mor, pembe ve beyaz öbeklerin bulunma sıklığıdır. Bu altbölgeler, özniteliklerine bağlı olarak K öbeğe ayrılır ve bulunan K tane öbek vektörü, $V = \{v_1, v_2, \dots, v_K\}$, kelimelerin tanımlanmasında kullanılır. Bu çalışmada, öbekler (kelimeler) her bir görüntüden 100 tane altbölge seçilerek ve k-ortalama algoritması kullanılarak öğrenilmiştir.

2.2. Görüntülerin Kelimelerle İfade Edilmesi

Kelimeler öğrenildikten sonraki aşama, verilen her bir görüntüyü altbölgelerine ayırmak ve bu altbölgelerin karşılık geldiği kelimeleri bularak görüntüyü sınıflandırmaktır. Bunun için, verilen görüntü X , pencere boyutuna göre $m \times n$



Şekil 3: Yöntem

tane dikdörtgen altbölgeye ayrılır. Görüntüde bulunan i numaralı altbölge, mor, pembe ve beyaz piksellerin bulunma sıklığını gösteren b_i vektörüyle ifade edilir; buna göre X görüntüsü, $X = \{b_1, b_2, \dots, b_{m \times n}\}$ olarak yazılır. Dikdörtgen altbölgelerini önceden öğrenilmiş kelimelerle eşleştiren ve aşağıdaki denklemde verilmiş $\Theta : B \rightarrow V$ fonksiyonu yardımıyla $X = \{\Theta(b_1), \Theta(b_2), \dots, \Theta(b_{m \times n})\}$ olarak yazılır.

$$\Theta(b_k) = \underset{j}{\operatorname{argmin}} \left(\sum_{i=1}^n (b_{ki} - v_{ji})^2 \right), \forall v_j \in V \quad (1)$$

Verilen bu denklemde b_{ki} gösterimi k numaralı altbölgenin i numaralı özniteliliğine ve v_{ji} gösterimi j numaralı öbek vektörünün i numaralı özniteliliğine karşılık gelmektedir. Θ fonksiyonun verdiği değerler kullanılarak $X = \{u_t | u_t \in V\}$ olarak ifade edilir ve X için bulunan u_k kelimelerinin öbek sıklığı hesaplanıp X görüntüsü sayısal bir vektöre çevrilmiş olur.

Hesaplanan bu vektörlerin öğrenilmesi için, eğitim kümesinde yer alan görüntülerden bu vektörler üretilir ve bu vektörler çok sınıflı doğrusal kernel kullanan destek vektör makinesinin eğitiminde kullanılır. Test kümesi için sınıflandırma, destek vektör makinesi (DVM) tarafından öğrenilmiş model ile yapılır.

3. DENEYSEL ÇALIŞMALAR

Deneylerimiz 258 hastadan alınmış 3236 doku görüntüsü üzerinde gerçekleştirilmiştir. Dokular hematoksilin-eosin tekniği boyanmış ve görüntüleri $20 \times$ mikroskop objektifiyle alınmıştır. Bu görüntüler normal, düşük dereceli kanserli ve yüksek dereceli kanserli olmak üzere üç sınıfa ayrılmıştır. Eğitim ve test kümelerindeki sınıf dağılımı Tablo 1'de verilmiştir.

Tablo 1: Normal, düşük dereceli kanserli ve yüksek dereceli kanserli görüntülerin eğitim ve test kümelerindeki dağılımı.

	Eğitim kümesi	Test kümesi
Normal	510	491
Düşük dereceli kanserli	859	844
Yüksek dereceli kanserli	275	257
Toplam	1644	1592

Tablo 2: Karşılaştırmada kullanılan algoritmaların parametreleri (destek vektör makinesi parametresi C hariç)

Öbeklenmiş altbölge istatistikleri	Bölge büyüklüğü = {10, 20, 40, 80}
Yoğunluk histogramları	Bölme sayısı = {4, 8, 16, 32}
Alt bölge tabanlı yoğunluk histogramları	Bölme sayısı = {4, 8, 16, 32}
Birlikte-bulunma matrisleri	Bölme sayısı = {4, 8, 16, 32}
Alt bölge tabanlı birlikte-bulunma matrisleri	Bölme sayısı = {4, 8, 16, 32}
	Uzaklık = {5, 10, 20, 40}
	Bölme sayısı = {4, 8, 16, 32}
	Uzaklık = {5, 10, 20, 40}
	Bölge büyüklüğü = {10, 20, 40, 80}

Önerdiğimiz *Kelime histogram* tabanlı yöntem üç tane model parametresi kullanmaktadır. Bunlardan ilki altbölge büyüklüğünü belirleyen dikdörtgen kenar uzunluğu D (altbölgeler kare olarak seçilmiştir), ikincisi kelime sayısını belirleyen K ve üçüncüsü destek vektör makinesinin doğrusal kernel fonksiyon parametresi C 'dir. Bu parametreler eğitim kümesi üzerinden 3-kat çapraz doğrulama ile belirlenmiştir. Bu yöntemle parametre değerleri $D = \{10, 20, 40, 80\}$, $K = \{10, 20, 40, 80\}$ ve $C = \{1, 2, \dots, 9, 10, 20, \dots, 90, 100, 150, \dots, 950, 1000\}$ arasında $D = 10$, $K = 20$ ve $C = 50$ olarak seçilmiştir.

3.1. Karşılaştırma Yöntemleri

Önerilen yöntem beş farklı yöntemle karşılaştırılmıştır. Bu karşılaştırmalarda, sınıflandırıcı olarak doğrusal kernelli DVM kullanılmıştır. İlgili yöntemlerin parametreleri 3-kat çapraz doğrulama ile belirlenmiş ve bu parametreler ile alabilecekleri değerler C parametresi dışında Tablo 2'de listelenmiştir. C parametresi için yukarıda verilen değerler kullanılmıştır.

Öbeklenmiş altbölge istatistikleri: Bu yöntemde mor, pembe ve beyaz renklere piksel bazında öbeklenmiş olan görüntüler dikdörtgen altbölgelere bölünür ve bu bölgeler için yöntemimizde kullandığımız öznitelikler çıkarılır. Daha sonra, çıkarılan öznitelikler öbeklenerek kelime tanımlamak ve bu kelimelerin histogramını kullanmak yerine, bu özniteliklerin tüm görüntü üzerinden ortalaması alınır.

Yoğunluk histogramları: Gri seviyedeki tüm görüntü için yoğunluk değerlerinin histogramı bulunur. Bu histogram üzerinde hesaplanan ortalama, standart sapma, basıklık ve çarpıklık değerleri öznitelik olarak kullanılır [9]. Yoğunluk değerlerindeki gürültünün etkisini azaltmak için histogram çıkarılırken değerler N bölmeyle ayrılır.

Alt bölge tabanlı yoğunluk histogramları: Bu yöntem, görüntüyü dikdörtgen altbölgelere ayırır ve her bir altbölge

Tablo 3: Test Sonuçları Doğruluk Oranları (%)

	Normal	Düşük dereceli	Yüksek dereceli	Genel
<i>Öbeklenmiş altbölge istatistikleri</i>	87.4	86.0	86.5	86.7
<i>Yoğunluk histogramları</i>	75.4	72.7	72.8	73.5
<i>Altbölge tabanlı yoğunluk histogramları</i>	75.2	77.0	79.4	76.9
<i>Birlikte-bulunma matrisleri</i>	76.4	83.9	79.4	80.8
<i>Altbölge tabanlı birlikte-bulunma matrisleri</i>	82.9	87.6	87.0	85.8
<i>Kelime histogramları</i>	94.3	84.0	86.4	88.4

için yukarıda verilen yoğunluk histogram özneliklerini çıkarır. Daha sonra tüm altbölgelerden çıkarılan özneliklerin ortalamasını alarak tüm görüntüyü nicilemek için kullanır.

Birlikte-bulunma matrisleri: Enerji, entropi, kontrast, homojenlik, korelasyon, benzemezlilik, ters diferansiyel moment ve maksimum olasılık öznelikleri, gri seviye görüntü üzerinde tanımlanan birlikte bulunma matrisi üzerinden hesaplanır [2, 10]. Deneylerimizde birlikte bulunma matrisi sekiz farklı yön için hesaplanmış ve her yön için bulunan matrislerin ortalaması alınarak yön bilgisinden bağımsız, ortalama birlikte bulunma matrisi M elde edilmiştir. Daha sonra ilgili öznelikler M matrisi üzerinden hesaplanmıştır.

Altbölge tabanlı birlikte-bulunma matrisleri: Bu yöntemin bir önceki yöntemden tek farkı, özneliklerin tüm görüntü yerine, dikdörtgen altbölgelere ayrılmış bölgeler için hesaplanmasıdır. Tüm görüntünün ifade edilebilmesi için, altbölge özneliklerinin ortalaması alınarak kullanılmıştır.

3.2. Test Sonuçları

Test kümesinin üzerinde alınmış doğruluk oranları, her bir sınıf için ayrı ayrı olmak üzere tüm yöntemler için Tablo 3’de verilmiştir. Sonuçlardan görüldüğü üzere *kelime histogramu* yaklaşımı karşılaştırmada kullanılan diğer renksel ve dokusal yöntemlere göre daha başarılı sonuçlar vermektedir. Ayrıca görüntüleri altbölgelere ayırıp öznelik çıkaran yöntemlerin görüntüyü bir bütün olarak ele alan yöntemlere göre daha başarılı oldukları da görülmektedir. McNemar testine göre önerilen yöntemin yoğunluk histogramları, altbölge tabanlı yoğunluk histogramları, birlikte-bulunma matrisleri ve altbölge tabanlı birlikte bulunma matrisleri yöntemleri ile arasındaki fark istatistiksel olarak anlamlıdır ($p < 0.05$). Önerilen yöntem ile öbeklenmiş altbölge istatistikleri yöntemi arasında fark normal doku sınıflandırılması için istatistiksel olarak anlamlıdır.

4. SONUÇ

Bu çalışmada histopatolojik doku görüntülerinin analizi ve tanı konulmasında *kelime histogramu* yönteminin kullanılması önerilmektedir. Önerilen bu yöntem, verilen doku görüntüsü için kelime tanımlamaya ve bu kelimelerin histogramını kullanmaya dayanır. Deneysel çalışmalarda, kolon doku görüntüleri için elde edilen sonuçlar umut vericidir. Bu yöntemde kelimeler oluşturulurken, eğitim kümesinden rasgele dikdörtgen altbölgeler seçilip bu bölgeler öbeklere ayrılmıştır. İleride bu bölgelerin rasgele seçilmesi yerine sonuçları artıracak şekilde belirli bölgelerden seçilmesi üzerine çalışılabilir. Ayrıca bu yöntemde görüntü üzerindeki kelimelerin birbirleriyle olan uzaysal ilişkileri modellenmemektedir. İleride bu kelimelerin

birbirleriyle olan uzaysal ilişkilerinin modellenmesi üzerine çalışmak da mümkündür.

5. KAYNAKÇA

- [1] A. Jemal, R. Siegel, J. Xu, E. Ward, “Cancer Facts & Figures 2010,” American Cancer Society, 2010.
- [2] A.N. Esgiar, R.N.G. Naguib, B.S Sharif., M.K. Bennett, A. Murray, “Microscopic image analysis for quantitative measurement and feature identification of normal and cancerous colonic mucosa,” *IEEE T. Inf. Technol. Biomed.*, vol. 2, no. 3, pp. 197-203, 1998.
- [3] A. Tabesh, M. Teverovski, H.Y. Pang, V.P. Kumar, D. Verbel, A. Kotsianti, O. Saidi, “Multifeature prostate cancer diagnosis and Gleason grading of histological images,” *IEEE Trans. Med. Imaging*, vol. 26, no. 10, pp. 1366-1378, 2007.
- [4] W. Wang, J.A. Ozolek, G.K. Rohde, “Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images,” *Cytometry Part A*, vol. 77A, no. 5, pp. 485-494, 2010.
- [5] B. Weyn, G. van de Wouwer, S. Kumar-Singh, A. Van Daele, P. Scheunders, E. van Marck, W. Jacob, “Computer-assisted differential diagnosis of malignant mesothelioma based on syntactic structure analysis,” *Cytometry*, vol.35, pp. 23-29, 1999.
- [6] D. Altunbay, C. Cigir, C. Sokmensuer, C. Gunduz-Demir, “Color graphs for automated cancer diagnosis and grading,” *IEEE Trans. Biomed. Eng.*, Évol. 57, no. 3, pp. 665-674, 2010.
- [7] Fei-Fei, L.; Perona, P.; , “A Bayesian hierarchical model for learning natural scene categories”, *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on* , vol.2, no., pp. 524- 531 vol. 2, 20-25 June 2005
- [8] Sivic, J.; Russell, B.C.; Efros, A.A.; Zisserman, A.; Freeman, W.T.; , “Discovering objects and their location in images”, *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on* , vol.1, no., pp. 370- 377 Vol. 1, 17-21 Oct. 2005
- [9] M. Wiltgen, A. Gerger, J. Smolle, “Tissue counter analysis of benign common nevi and malignant melanoma,” *Int. J. Med. Inform.*, vol.69, pp.17-28, 2003.
- [10] R.M. Haralick, “Statistical and structural approaches to texture,” *P. IEEE*, vol.67, pp.786-804, 1979.