Sayısal Belgelerde Projeksiyon Ölçek Uzayı ile Tablo Tespiti ve Analizi

Document Table Detection and Analysis Using Projection Scale Space

L. İlham Kalyon ve Yusuf Sinan Akgül
GIT Vision Lab, http://vision.gyte.edu.tr/, Bilgisayar Mühendisliği Bölümü
Gebze Yüksek Teknoloji Enstitüsü, Kocaeli, 41400, Türkiye
ilham.kalyon@gmail.com, akgul@bilmuh.gyte.edu.tr

***Özetçe*—Belge görüntü işleme çalışmalarında yoğun ilgi duyulan alanlardan bir tanesi, belge görüntülerinden tabloların bulunması ve analizi olmuştur. Bu çalışmada, belge görüntülerinden tabloların tespiti ve analizi için özgün yöntemler tanımlanmış ve bu yöntemlerin gerçek tablolar üzerindeki başarımı gösterilmiştir. Geliştirilen ana yöntem, tabloların satır bazında yerel ve bütünsel kısıtlarını aynı anda sağlamak için kullanılan projeksiyon-ölçek-uzayıdır (PÖU). PÖU, belgede kullanılan karakter kümesine, görüntü çözünürlüğüne ve gürültü oranlarına karşı dayanıklıdır ve tablo tespit işlemlerini oldukça etkili bir şekilde yapılabilmektedir. Ayrıca geliştirilen yöntem, sınırları belirli ya da belirsiz tablolar üzerinde çalışabilmekte ve tabloların satır ve sütün analizini de yapabilmektedir. Önerilen yöntem, 130 tablo içeren 105 belge veri kümesi üzerinde test edilmiş ve sistemin başarımı oldukça üst düzeyde olduğu nicel bir şekilde gösterilmiştir.**

Anahtar Kelimeler — tablo tespiti; tablo analizi; belge analizi.

***Abstract*—Detection and analysis of tables on document images has been one of the most researched topics in document image processing. In this study, we define novel methods for the detection and analysis of tables from document images, and show their performance results on realistic table examples. The main method developed is projection-scale-space (PSS), where local and global constraints of the table in row basis are analyzed for consistency. PSS is robust to the character set used in a document, the image resolution and the noise ratio of a document image, and can perform detection operations in a highly effective manner. Furthermore, the method proposed works on tables with and without table borders and is able to analyze rows and columns of tables. The proposed method has been tested on a dataset of 105 documents containing 130 tables and the systems high performance has been stated in quantitative basis.**

Keywords — table detection; table analysis; document analysis.

978-1-4799-4874-1/14/$31.00 ©2014 IEEE

I. GİRİŞ

Tipik bir belge; metin, resim, tablo vb. kısımlardan oluşabilir. Belge üzerindeki değişik bölgelerin algılanması için geometrik düzen analizi ve mantıksal düzen analizi yapılarak belgenin analizi gerçekleştirilebilir[4]. Belge analizinin genel amacı, belge yapısının bulunmasıyla birlikte belge görüntüsünün üzerinde optik karakter tanıma (OCR) ve akıllı karakter tanıma (ICR) işlemleri yapılarak bilgisayar ortamında işlenebilecek elektronik bilgi elde etmektir. Analiz sonucu elde edilen belge bilgisi elektronik ortamda; bilgiyi yeni şekilde sunmak, belge üzerinde değişiklik yapmak, arama işlemi gerçekleştirmek gibi birçok gerekli olan fonksiyonellikleri beraberinde getireceği için belge analizine ihtiyaç duyulmaktadır[8].

Tablo, satırlar ve sütunlar halinde verilerin düzenli şekilde sunulması için kullanılan bir araçtır. Belge üzerinde tabloların tespiti ve tablo analizi için geçmişte birçok çalışma yapılmıştır. Belge üzerinde tablo tespiti yöntemlerinden bir tanesi Shafiat ve Smith [5] tarafından sunulmuştur. Çalışmalarında değişik düzendeki belgeler üzerinde tablo tespiti üzerine yoğunlaşmışlardır. Fang vd. [2], PDF belgelerinde tablo sınır ayırıcının tespiti üzerine yoğunlaşmışlardır. Gatos vd. [9] belge görüntüsü üzerindeki yatay ve dikey doğruların kesişim noktalarının analizi ile tablo tespitini gerçekleştirmişlerdir.

Literatürde görülen çalışmaların ortak özelliği, sayısal belgeler üzerinde uygulanan operatörlerin sadece belli bir ölçekte uygulanmış olmasıdır. Bu çalışmada, birden fazla ölçek özelliğine sahip yeni bir tablo tespit metodu önerilmektedir. Projeksiyon Ölçek Uzayı (PÖU) olarak adlandırdığımız bu teknik, verilen bir tablo belgesi satırının yerel özelliklerinin, daha bütünsel tablo özellikleriyle aynı olduğunu varsayan bir yaklaşımdır. Bu türlü bir yaklaşım için ölçek uzayı tabanlı [10] bir çözümün etkili sonuçlar üreteceği öngörülmektedir. Önerilen PÖU tabanlı tablo tespit tekniği, çoklu ölçek doğası gereği belge çözünürlük değişimleri ve kullanılan karakter kümelerine karşı dayanıklıdır. Tespit edilen tablonun analizi problemine en yaygın yaklaşım; satır ve sütun ayırıcılarının bulunmasıdır.

Yazı ve boşluk gibi ifadelerin konumuna bakılarak genelde ayırıcılar bulunmuştur. Basit yapıdaki tabloların analizi için Bart [1] görüntüde gürültü azaltma işlemi ve OCR motoru sonuçlarının yorumlanması ile tablo analizi problemine yaklaşmıştır. Tablo analizi için sunulan yöntemler hakkında daha detaylı bilgi edinmek için [1] [3] [6] incelenebilir. Bu çalışmada tablo analizi için geliştirilen yöntemde; sınırları belirli ve belirsiz olan tabloların yapısının doğru belirlenmesi için iki çeşit metot aynı anda koşturulmuştur. Metotlardan ilki tablo görüntüsü üzerinde Hough-Doğru-Transformasyonu algoritması ile tablo sınırları var ise sınırların bulunmasıdır. Bulunan sınırlara göre sütun ve satır sayısı belirlenmiş olacaktır. İkinci metot ise, daha çok sınırları doğrularla belirlenmemiş sütun ve satırlar için geliştirilmiştir. Bu metot, sadece tablonun elemanı olabilecek metin ifadelerinin projeksiyonu hesaplanarak piksel yoğunluklarının yorumlanmasıyla sütun ve satırları belirlemektedir. Paralel çalışan iki yöntemin sonuçları karşılaştırılarak tablonun yapısı hakkında sonuca varılmıştır. Bunun nedeni; tablo görüntüsü üzerindeki gürültü, boş bir tablo hücresi olması, tablo ve hücre sınırlarının doğrularla belirlenmemesi durumunda yöntemlerin birbirlerinin eksikliklerini tamamlayarak analizin doğru yapılması gereğidir.

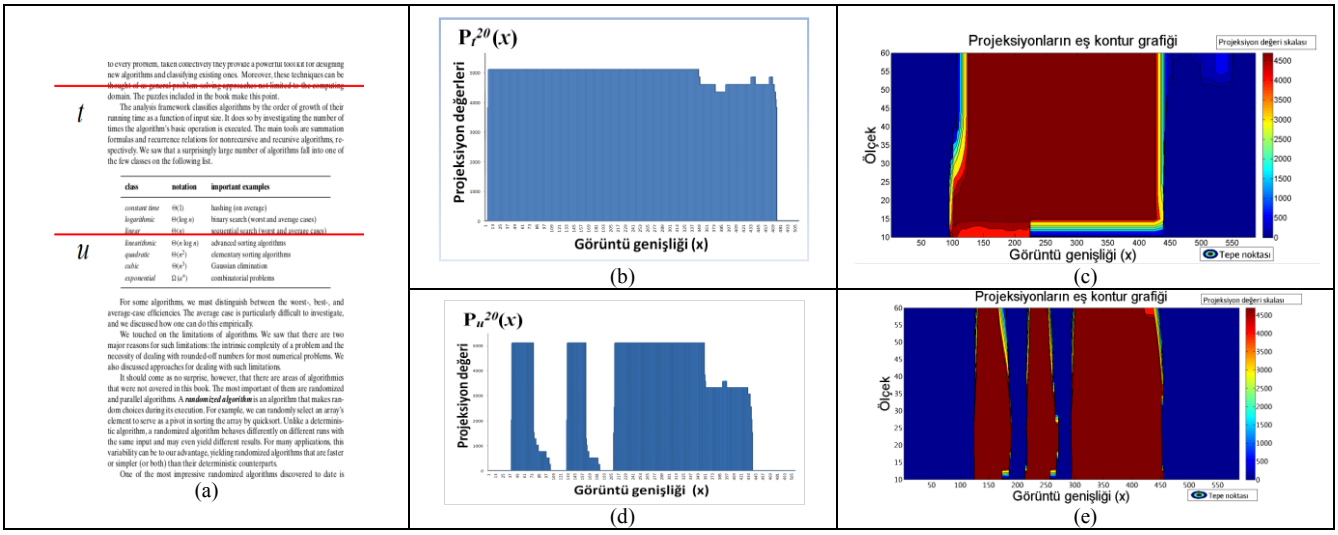
Bu bildirinin arta kalan kısmında sunulan yöntemin detayları anlatılmıştır. II. Bölümde; tablo tespitinde kullanılacak olan projeksiyon operatörü ve ölçek uzayı anlatılmaktadır. III. Bölümde; PÖU ile tablo tespiti ve tablo analizi sistemi açıklanmıştır. Son olarak sistem üzerinde yapılan deneyler ve sonuç kısmına yer verilmiştir.

II. PROJEKSİYON OPERATÖRÜ VE ÖLÇEK UZAYI

Projeksiyon operatörü, verilen bir ikili belge görüntüsünün bir satırı için belli bir ölçekte yerel projeksiyonunu üretir ve şu formülle hesaplanabilir:

$$P_i^n(x) = \sum_{j=i-n/2}^{i+n/2} I(x, j) \qquad (1)$$

Formülde, I ikili belge görüntüsünü, i projeksiyon operatörünün üzerinde çalıştığı satırı, n ise kullanılan ölçeği göstermektedir.Projeksiyon operatörü kullanılarak oluşturulan projeksiyon ölçek uzayı ise verilen bir belge görüntüsünün aynı satırı üzerinde tüm ölçekler üzerinde oluşturulmuş projeksiyon operatörlerinin birleştirilmesiyle elde edilir. Gerçek dünya nesneleri farklı ölçeklerde farklı özellikleri ile görünmektedirler. Nesnelerin ölçeğe bağlı olarak farklı şekillerde görünmelerinden kaynaklanabilecek hataları ortadan kaldırmak için ölçek uzaya ihtiyaç duyulmaktadır [10]. Ölçek uzayı görüntünün tüm ölçeklerdeki değerini dikkate almaktadır. Bu nedenle tablo tespiti için geliştirilen yöntemde kullanılmıştır. Ölçek uzaydaki iki boyutlu görüntünün bir boyutlu projeksiyonu, görüntü üzerindeki piksel yoğunluklarının yorumlanması için hesaplanmıştır. Projeksiyon dağılımına bakıldığında projeksiyon vektöründeki değerlerin yerel tepe noktaları oluştuğu görülür. Ölçek uzaylardaki projeksiyonlarda hesaplanan tepe nokta sayıları aslında o uzayda bulunan belge görüntüsü bölgesinde olan sütun sayılarını ifade etmektedir. Sunulan yöntemde projeksiyon vektörlerindeki yerel tepe noktalarının, bütünsel tepe noktasıyla olan tutarlılığı (sütun sayısı tutarlılığı) incelenerek tablo tespiti gerçekleştirilmiştir. Şekil 1'de örnek belge görüntüsü üzerinde ele alınan iki satır Şekil 1.a'da gösterilmiştir. Şekil 1.b, $P_t^{20}(x)$ için üretilmiş iki boyutlu projeksiyon grafiğini göstermektedir. PÖU kullanarak elde edilen ve tüm ölçekleri içeren uzay üç boyutlu bir uzaydır (görüntü genişliği, ölçek, projeksiyon yüksekliği boyutları). 3B PÖU uzaylarını göstermek için eş kontur (iso contour) gösterme tekniği kullanılarak, t satırı için elde edilen sonuç Şekil 1.c de gösterilmiştir. Buna göre t satırı için tüm ölçeklerde sadece bir tepe noktası tespit edilmektedir ve t



Şekil 1. Örnek belge görüntüsü üzerinde ele alınan iki satır (a)'da gösterilmiştir. (b) t satırının ölçek değeri 20 olan uzaydaki projeksiyon grafiğidir. (c) t satırının farklı ölçeklerdeki projeksiyonunun eş kontur grafiğidir (PÖU grafiği). (d) u satırının ölçek değeri 20 olan uzaydaki projeksiyon grafiğidir.(e) u satırının farklı ölçeklerdeki projeksiyonunun eş kontur grafiğidir (PÖU grafiği).

satırı etrafında birden fazla sütundan oluşan bir tablo olmadığı anlaşılır. Aynı işlemler *u* satırı için tekrarlandığında, Sekil 1.d ve Şekil 1.e elde edilmektedir. Şekil 1.e'nin incelenmesiyle, PÖU uzayındaki ölçeklerin çoğunda, 3 tepe noktasına sahip oldukları görülmüştür. Buna göre, verilen herhangi bir yerel ölçek için elde edilen tepe noktası sayısı, bütünsel ölçekler için de rahatlıkla elde edilebilmektedir.

III. PÖU İLE TABLO TESPİTİ VE TABLO ANALİZİ

Şekil 2'de sistemin akış diyagramı verilmiştir. Belge üzerinde yapılan ön işlem sonrası tablo tespiti yapılmıştır. Ardından tablonun analizi yapılarak satır ve sütun sayıları bulunmuştur.

A. Görüntü üzerinde ön işlem yapılması

Sistem projeksiyon hesaplarına bağlı olduğu için ön işleme ihtiyaç duyulmuştur. Analizlerin doğru yapılması için uygulanan ana işlemler *Gauss-Smoothing,* ve *Morfoloji* fonksiyonları olan *Erosion* ve *Dilate* transformasyonlarıdır.

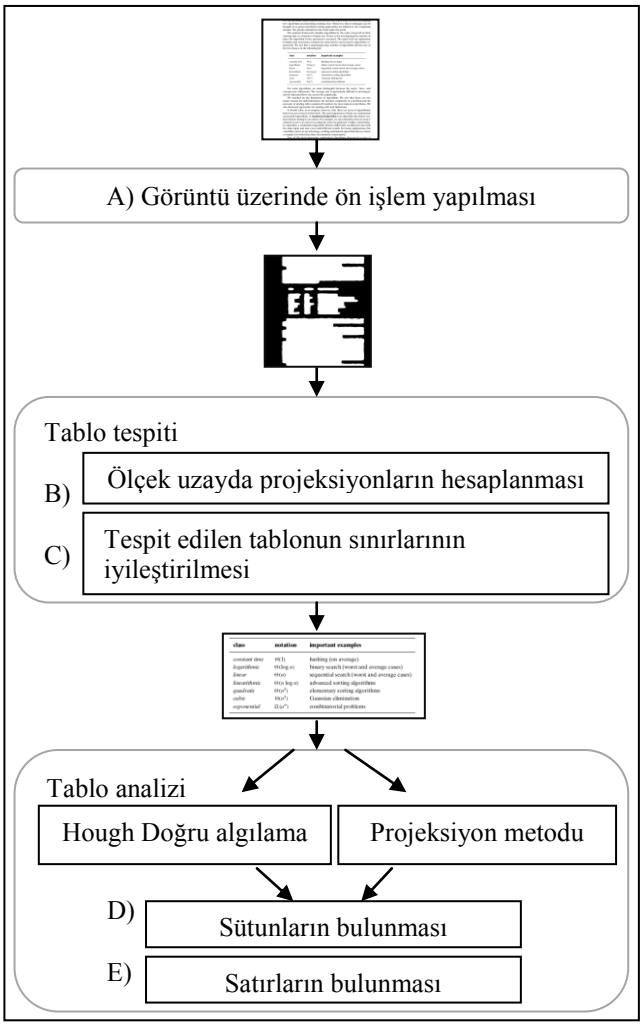
1. Tablo tespiti

B. Ölçek uzayda projeksiyonların hesaplanması

Belge görüntüsünün çözünürlük değerinden etkilenmeyen gürbüz bir yöntem geliştirmesi, tablo tespitinde ölçek uzaya bakılmasının nedenidir. Tablolar, belge görüntüsü üzerinde ön işlem yapılması sonrası tespit edilmiştir. Tablo tespitinde; Belge görüntüsünün her satırı için ölçek uzayda eşitlik (1) ile projeksiyonu hesaplanmıştır. Ölçek uzaylarındaki satırların projeksiyon vektöründeki değerlerin dağılımından tepe nokta sayıları hesaplanmıştır. Tepe nokta sayısı ele alınan satırın kaç sütuna sahip olduğu bilgisini vermektedir. Şekil 1.d'de, Şekil 1.a'da gösterilen *u* satırının, 20 ölçek değerindeki projeksiyonundan 3 sütuna sahip olduğu bilgisi görülmektedir. Ölçek uzaylarında, satır sayılarının uyumuna göre satırların tablo oluşturduğu söylenebilmektedir. Farklı ölçeklerde, her satırın yerel tepe noktaları bu şekilde hesaplanmıştır. PÖU ile farklı ölçeklerde hesaplanan örnek satırların grafiği Şekil 1.c ve 1.e'de gösterilmiştir. Ölçek uzaylarda tespit edilen tablolar arasında tablo aralığı maksimum bulunan ölçek ile analize devam edilmiştir.

C. Tespit edilen tablonun sınırlarının iyileştirilmesi

Tespit edilen tablo sınırları içersinde tabloya ait olmayan paragraf metni gibi ifadeler bulunabilir. Sınırların doğru belirlenmesi için tablo iki parça; üst ve alt kısım olarak ayrı incelenmiştir. Üst sınırının belirlenmesi için üst yarının projeksiyonu hesaplanmıştır. Metin vb. ifadeleri barındıran sınırların projeksiyonunun tepe nokta sayısı, gerçek tablonun sütun sayısına eşit olmayacaktır. Projeksiyonlar ile hesaplanan tepe nokta sayısı, tablonun sütun sayısına eşit olana kadar tablo sınırı ilerletilmiştir. Tepe nokta sayısının sütun sayısına eşit olduğu konum doğru sınır olarak belirlenmiştir. Tablonun alt sınırının belirlenmesi üst sınırın belirlenmesiyle aynı mantığa dayalıdır.



Şekil 2. Sistemin akış diyagramı

2. Tablo analizi

Giriş bölümünde açıklandığı gibi, tablolar üzerinde bahsedilen doğru algılama ve projeksiyon metotları paralel çalıştırılarak, sonuçlarının karşılaştırılması ile yapısı belirlenmiştir. Tablo analizinde hücre sınırları yaklaşık olarak hesaplanmıştır. Analiz sonucunda esas olarak tablo sütun ve satır sayısı bulunmuştur.

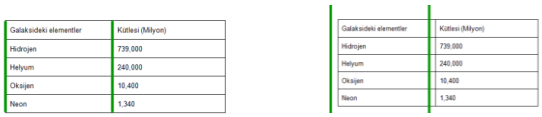
D. Sütunların bulunması

Hough-Doğru algılama yöntemiyle tablo üzerindeki sütunların bulunması için, tablo üzerinde dikey çizgilerin algılanması sağlanmıştır. Projeksiyon yönteminde ise tablo görüntüsünün bir boyutlu projeksiyonu yatayda hesaplanmıştır. Şekil 3'te tablo sınırları belirli olan tablo örneği verilmiştir. Şekil 3.a'da sütun sınırları Hough-Doğru algılama metodu, şekil 3.b'de projeksiyon yöntemiyle bulunmuştur. Şekil 3'teki tablonun sınırları belirsiz olsaydı Hough-Doğru algılama yöntemi sınırları bulamazken projeksiyon yöntemi sütunları bulacaktır. İki

metodun sonuçları ile tablonun yapısının belirlenmesi bu yüzdendir.

E. Satırların bulunması

Satırların bulunmasında, sütunların bulunmasından farklı olarak Hough-Doğru algılama yönteminde tablo üzerinde yatay çizgilerin algılanması sağlanmıştır. Projeksiyon yönteminde ise tablo görüntüsünün bir boyutlu projeksiyonu düşeyde hesaplanmıştır.



Galaksideki elementler	Kütlesi (Milyon)
Hidrojen	739,000
Helyum	240,000
Oksijen	10,400
Neon	1,340

(a) (b)

Şekil 3. Sınırları belirli olan örnek tablo üzerinde (a) Hough-Doğru algılamanın, (b) projeksiyon metodunun ürettiği sütun analizi sonuç görüntüleridir.

IV. DENEYLER

Belge görüntüleri üzerinde tablo tespiti ve tablo analizi için oluşturulan sistemi test etmek için veri kümeleri oluşturulmuştur. Veri kümeleri; kitap, rapor ve internet sayfası gibi ortamlarda kullanılan belge ve tablo örneklerini içermektedir. Belge görüntüleri üzerinde tablo tespiti için geliştirilen yöntemi sayısal olarak doğrulamak için yöntemin çalışması sonucu elde edilen tablo koordinat bilgileri ile manüel hesaplanan tablo koordinat bilgileri karşılaştırılmıştır. Deney, 105 belge görüntüsü içerisinde bulunan toplamda 130 tablo içeren veri kümesi üzerinde yapılmıştır. Deneyde tek bir ölçek değerleriyle elde edilen sonuçlar ile PÖU sonucu karşılaştırılmıştır. PÖU ile tablo tespitinin daha gürbüz bir yöntem olduğu görülmektedir.

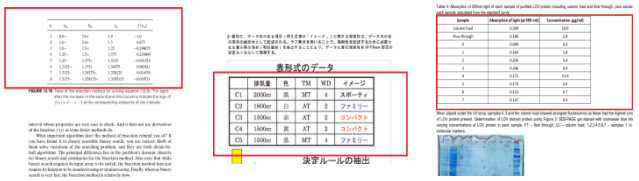
Ölçek	Doğru pozitif oranı	Yanlış pozitif oranı	Kesinlik (precision)	Geri çekme (recall)	F-Skor
10	0.623	0.484	0.562	0.623	0.590
20	0.846	0.215	0.797	0.846	0.820
30	0.853	0.153	0.847	0.853	0.849
40	0.861	0.115	0.881	0.861	0.870
50	0.853	0.138	0.860	0.853	0.856
PÖU	**0.892**	**0.092**	**0.906**	**0.892**	**0.899**

Tablo 1. Tablo tespiti başarı sonuçları

Tablo görüntüleri üzerine satır ve sütun analizi için geliştirilen yöntem 45 tablo görüntüsü içeren farklı bir veri kümesi üzerinde de ayrıca test edilmiştir. Sütun analizi hata oranı %7, satır analizi hata oranı %10 olarak hesaplanmıştır. Hatalar örnek tabloların satır ve sütunlarının çok bitişik olmasından kaynaklanmıştır. İlerde tepe-nokta bulma algoritmasının geliştirilmesiyle daha iyi sonuçlar elde edilebilecektir. Literatürdeki çalışmalar farklı veri kümeleri üzerinde çalıştırıldığı için geliştirilen sistemin literatürdeki çalışmalar ile birebir karşılaştırılması yapılamamıştır. Fakat, üzerinde uzun süre çalışılmış araştırmaların [2][6] F-skor değerlerinin %95 civarında olması, henüz ilk önerilme aşamasında olan tekniğimizin diğer tekniklerle uyumlu olduğunu göstermiştir.

V. SONUÇLAR

Bu çalışmada, sınırları belirli ya da belirsiz olan tabloların, belge görüntüleri üzerinde tespit edilmesi ve analizi için metotlar geliştirilmiştir. Metotlar, projeksiyonların yorumlanmasına dayalıdır. Tablo tespitinde belge görüntüsünün ölçek uzayları değerlendirilerek gürbüz bir sistem geliştirilmiştir. Hatalı sonuçlar genellikle gürültü ve çok sütunlu tablo içeren görüntü örneklerinde görülmüştür. Bunun nedeni belge görüntüleri üzerinde projeksiyon değerlerinin tepe nokta sayısının doğru hesaplanamamasından kaynaklanmıştır. Sistem belge görüntüsü üzerinde OCR motoru ile metin analizi işlemleri gerçekleştirmediği için performansı yüksektir. Bu çalışma finansal bir firmanın fatura analizinin elektronik ortamda yapılması için yola çıkılan bir çalışmadır. İleride fatura belgelerinde tablo analizi üzerine çalışılacaktır.



Şekil 4. Farklı karakter kümeleri ve şekil içeren belgeler üzerinde tablo tespiti sonucu elde edilen görüntüler.

KAYNAKÇA

[1] Evgeniy Bart, "Parsing Tables by probabilistic modeling of perceptual cues", *IAPR International Workshop on Document Analysis Systems*, 2012.

[2] J. Fang, L. Gao, K. Bai, R. Qui, X. Tao and Z. Tang, "A Table Detection Method for Multipage PDF Documents via Visual Separators and Tabular Structures", *International Conference on Document Analysis and Recognition*, 2011.

[3] H. Hamza, Y. Belaid, A. Belaid, "A case-based reasoning approach for invoice structure extraction", 2007.

[4] S. Maoa, A. Rosenfelda, T. Kanungob, "Document Structure Analysis Algorithms: A Literature Survey", *Electronic Imaging* 2003. International Society for Optics and Photonics, 2003.

[5] F. Shafiat, R. Smith, "Table Detection in Heterogeneous Documents", *DAS '10 Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, 2010.

[6] Y. Belaid, A. Belaid, "Morphological Tagging Approach in Document analysis of Invoices", *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on* Vol. 1. IEEE, 2004.

[7] Y. Y. Tang, C. Y. Suen, C. D. Yan, and M. Cheriet, "Document Analysis and Understanding: A Brief Survey", *in Proceedings of the ICDAR*, 1991.

[8] Glenn A. Bowen, (2009) "Document Analysis as a Qualitative Research Method", *Qualitative Research Journal*, 9.2 (2009): 27-40.

[9] B. Gatos, D. Danatsas, I. Pratikakis and S. J. Perantonis, "Automatic table detection in document images", *Proc. Intl. Conf. Advances Patt. Recog.* pp. 609-618, 2005.

[10] Lindeberg, Tony (2008). "Scale-space". Encyclopedia of Computer Science and Engineering (Benjamin Wah, ed), John Wiley and Sons IV: 2495–2504.