# ÖZET


| | | |
|---|---|---|
| **Üniversite** | : | İstanbul Kültür Üniversitesi |
| **Enstitüsü** | : | Fen Bilimleri Enstitüsü |
| **Dalı** | : | Matematik-Bilgisayar |
| **Programı** | : | Matematik-Bilgisayar |
| **Tez Danışmanı** | : | Yard. Doç. Dr. Levent Çuhacı |
| **Tez Türü ve Tarihi** | : | Yüksek Lisans – Ocak 2014 |


## RESİM TABANLI OSMANLICA BELGELERDE SINIFLANDIRMA


**Ramazan Pehlivan**



Bu çalışmanın amacı resim formatındaki Osmanlıca belgeleri içeriklerine göre sınıflandıran bir model ortaya koymaktır. Bu amaçla resim formatında taranmış Osmanlıca matbu belgelerde, ‘‘Görüntü İşleme’’, ‘‘Kümeleme’’ ve ‘‘Doğal Dil İşleme’’ tekniklerini birlikte kullanarak ‘‘Doküman Sınıflandırma’’ yapan etkin bir sınıflandırma yöntemi önerilmiştir.

Çalışmamızda veri olarak Türkiye Büyük Millet Meclisi (TBMM) Kütüphane ve Arşiv Hizmetleri Başkanlığı’nın resmi web sitesinden alınan Osmanlıca belge örnekleri seçilmiştir. Görüntü işleme teknikleriyle belgeler sayısal forma dönüştürülmüş, ardından satırlar ve satırlardaki kelime ya da harf grupları tespit edilmiş ve her bir harf grubu ayrı birer resim olarak kaydedilmiştir. Resimler arasında kümeleme yapılarak aynı (ya da benzer) harf grupları aynı kümeye atanmıştır. Harf gruplarının ait oldukları küme bilgileri kullanılarak bu belgelerin, birbirini izleyen etiket numaralarını içeren metin formatındaki karşılıkları elde edilmiştir. Bu aşamadan sonra doküman sınıflandırma alanında geçerli bir teknik olan kelime frekans analizi, elde ettiğimiz dönüştürülmüş metin dosyalarında küme frekans analizi olarak uygulanmıştır. Sonuç olarak; resim formatında taranmış Osmanlıca belgeler; semantik analize tabi tutulmadan, belgeyi oluşturan harf gruplarının benzerlik ölçütleri baz alınarak sınıflandırılmıştır.

Proje MATLAB ortamında geliştirilmiş ve bir makine öğrenmesi uygulaması olan WEKA programında sınıflandırma sonuçları elde edilmiştir. Ayrıca aynı veri seti üzerinde kelime frekans analizine dayalı bir doküman sınıflandırma uygulaması da gerçeklenmiştir.



**Anahtar Kelimeler:** Osmanlıca belge, doküman sınıflandırma, resim kümeleme, frekans analizi, satır parçalama, hiyerarşik kümeleme.

## ABSTRACT


| | | |
|---|---|---|
| **University** | : | İstanbul Kültür University |
| **Institute** | : | Institute of Science |
| **Department** | : | Mathematic-Computer |
| **Literature Programme** | : | Mathematic-Computer |
| **Supervisor** | : | Assis. Prof. Dr. Levent Çuhacı |
| **Degree Awarded and Date** | : | MA – January 2014 |


### CLASSIFICATION OF IMAGE-BASED OTTOMAN RECORDS


**Ramazan Pehlivan**



Aim of this work is developing a model which classifies image-formatted Ottoman records by their contents. For this purpose, an effective classification method, which conjunctively uses "Image Processing", "Clustering" and "Natural Language Processing" techniques, is proposed for image-formatted scans of Ottoman printed records.

In our work, Ottoman record samples from the official web page of Turkish Grand National Assembly (TBMM) Library and Documentation Center were used as data. Records were converted into digital form via image processing techniques, then words or letter groups in documents were detected and stored separately as individual pictures. By clustering between these pictures, identical (or similar) letter groups were registered to the same cluster. By using cluster information of letter groups, text-formatted counterparts, which include consecutive label numbers, were obtained for records. After that step, word frequency analysis, which is a valid technique in document classification, was used on converted text files as cluster frequency analysis. As a result, image-formatted scans of Ottoman records were classified based on similarity criteria of constituting letter groups, without using semantic analysis.

Project was developed on MATLAB environment and classification results were obtained by a machine learning application software, WEKA. Another classification method based on word frequency analysis was also implemented using the same data set.



**Keywords:** Ottoman record, document classification, image clustering, frequency analysis, line segmentation, hierarchical clustering