BOBCAT-DI:

Selected Evaluations using the PETS Environment

Army Research Laboratory,
Adelphi MD

from

Laboratory for Language and Media Processing
University of Maryland, College Park, MD, USA

December 20, 2008

Table of Contents

List of Figures

Selected Evaluations using the PETS Environment

1 Introduction

1.1 Background

Over the past five decades, evaluation has become increasingly important as the field of document image understanding has developed. A number of independent evaluations have been run by various academic organizations, all focusing on slightly different problems. In recent years, the University of Maryland has developed a number of tools aimed at supporting generic annotation and evaluation of document (and video) data. A set of three reports is being produced from this one year BOBCAT-DI research and development projects. The three reports include

- A *Segmentation and Evaluation Survey*- designed to identify major algorithms, tools and evaluation methodologies in the community,
- The *PETS Software Description*– a toolkit to evaluate segmentation, line detection and image enhancement algorithms, based on BOBCAT-DI requirements and as a response to the state of the art, and
- *Selected Evaluations using the PETS Environment* (This Document)– Evaluations designed to demonstrate the capabilities of the tools and provide a framework for use in operational environments.

It is the hope that this work will lead to a generic repository for evaluation which host data, tools, algorithms and evaluation results for community wide comparison.

1.2 Project Overview

The DoD Sequoyah Foreign Language Translation Program, managed by the interim Sequoyah Transition Mgt Office (STMO) under PEO IEWS, Ft Monmouth, NJ, is intended to address critical linguist shortfalls in US warfighting and intelligence operations through automated language translation capabilities (speech and text) and to provide document image processing and OCR capabilities for cases when material to be translated is paper or document images. To support unbiased, vendor-neutral assessment of technology candidates prior to field testing and deployment, the STMO has initiated a web-accessible, distributed “Best-of Breed Configurable Active Testbed” (BOBCAT) led and operated by ARL and distributed across NRL and AFRL. Yet to be incorporated into the testbed is the capability to assess OCR and other document image processing (DIP) tools. The STMO as well as the ODNI have tasked ARL with integrating document image (DI) processing assessment into BOBCAT, creating BOBCAT-DI . BOBCAT-DI will be used to assess a variety of document image processing capabilities and tools, with a focus on Arabic and other Southwest Asian languages. The image processing and analysis metrics and methods, particularly as applied to document images obtained from cameras, scanners, etc., is needed to enable assessments that are reliable, robust, and scientifically defensible

1.3 Organization of Report

In this report, we focus on providing the results for selected algorithms developed at the University of Maryland Laboratory for Language and Media Processing. The goal is not to provide a comprehensive evaluation of the field, but rather to demonstrate the ability of PETs and related software.

These algorithms include:

- Page Segmentation
- Clutter Detection and Removal
- Rule Line Removal
- Unsupervised Line Detection
- Zone Classification

These algorithms were funded under by the Department of Defense under and the DARPA MadCat Program under subcontract from BBN, Cambridge Mass.

2 Selected Algorithms

2.1 Overview

The algorithms presented in this report are a collection of segmentation and image analysis algorithms that require detailed and nontraditional evaluation metrics, such as those described in PETS. In the following sections, each algorithm is briefly described, along with evaluation results obtained on representative datasets.

2.2 Page Segmentation with Voronoi++

2.2.1 Introduction

This algorithm presents a dynamic approach to document page segmentation. Current page segmentation algorithms lack the ability to dynamically adapt local variations in the size, orientation and distance of components within a page. Our approach builds upon one of the best algorithms, Kise et. al. work based on Area Voronoi Diagrams, which adapts globally to page content to determine algorithm parameters. In our approach, local thresholds are determined dynamically based on parabolic relations between components, and Docstrum based angular and neighborhood features are integrated to improve accuracy. Zone-based evaluation was performed on four sets of printed and handwritten documents in English and Arabic scripts and an increase of 33% in accuracy is reported

The central idea of the algorithm is creation of Voronoi edges between pairs of connected components using area based Voronoi tessellation. Each edge bisects two points on contours of different components. A physical zone is a fusion of these Voronoi cells, formed by the elimination of Voronoi edges based on two features: 1. Minimum distance and 2. Area Ratio. An edge is deleted if it satisfies the following two criteria: $d(E)\ Td1 < 1$ and $d(E)\ Td2 + ar(E)\ Ta < 1$ where $Td1 < Td2$, Td1 relates to inter-character spacing, Td2 relates to inter-word/line spacing

Docstrum performs transitive closure on within-line components to obtains lines and then on lines to form regions. The thresholds for transitivity are based on the properties of distance and angle of each connected component with its K nearest neighbors. The advantage of Docstrum over the Voronoi based approach is its ‘semi-local’ behavior. Each component looks at K nearest neighbors to make a decision of its association, unlike Voronoi where decision is solely nearest neighbor based. In spite of this, Docstrum has been designed mainly for only text documents.

Voronoi++ technique has the following advantages over Voronoi based page segmentation:

1. Dynamic Distance Threshold: Global thresholds are determined dynamically using local features. This removed the following two problems:
   a. Over segmenting larger fonts
   b. Grouping dissimilar font sizes
2. Combining Docstrum features

    a. Angle: Distances were weighed based on angle between the components in a Gaussian fashion
    b. Nearest neighbor association: Components smaller than frequent text-size were associated with their nearest neighbor to avoid region-formation around diacritic.

3. Word-separation threshold using a more deterministic valley finding algorithm for bimodal histograms
4. Polygonal zones instead of rectangular

Details of this algorithm are available in a technical paper that was submitted to ICDAR 2009 and are available upon request.

2.2.2 Evaluation

A ground-truth text-line is said to lie completely within one detected zone if the area overlap between the two is significant. The drawback of that approach, however, is that if the segmentation algorithm outputs the whole page as one segment, the split and missed errors are ignored and accuracy is higher. In order to avoid this, we compare ground-truth zones with result zones. A result zone is said to be detected, if its foreground pixels overlap with those of ground-truth above a user specified percentage. This is a much stricter evaluation scheme in terms of zone detection. We evaluated the Voronoi++ approach on datasets of printed and handwritten documents in English and Arabic scripts.

We evaluated the Voronoi++ approach on datasets of printed and handwritten documents in English and Arabic scripts. The polygonal regions are outputted in xml format designed for LAMP's GEDI tool. This tool is used to label the data and visualize the segmentation and evaluation results. We compared our approach on 200 document images, half of which were randomly picked from University of Washington III(UW-III) database and results are shown below:

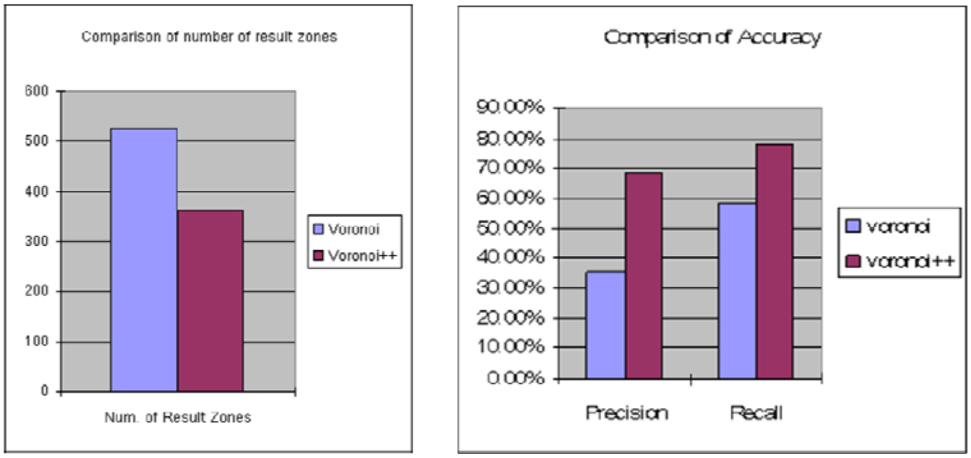
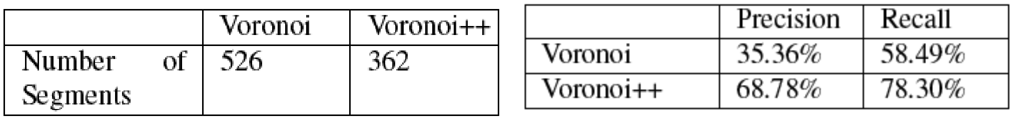


Figure 1: Precision and Recall



	Voronoi	Voronoi++
Number of Segments	526	362

	Precision	Recall
Voronoi	35.36%	58.49%
Voronoi++	68.78%	78.30%

Figure 2: Comparison between Previous and Enhanced Algorithms

The increase in accuracy by 33% also proves that the decrease in number of zones does not increase merge errors.

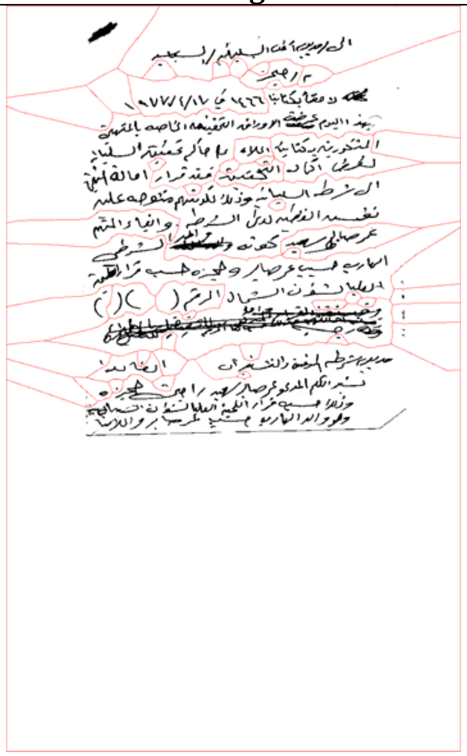
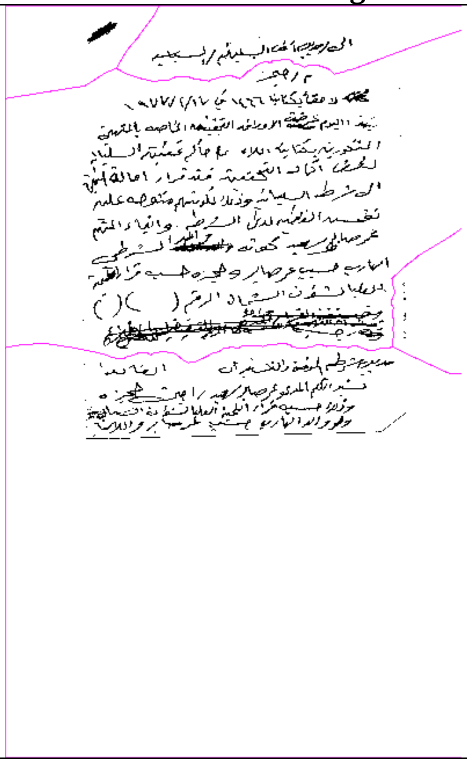
Voronoi Segmentation	Enhanced Voronoi Segmentation
	

Figure 3: Example Results

2.3 Clutter Detection and Removal

2.3.1 Introduction

This section describes a clutter detection and removal algorithm for complex document images. The distance transform based approach is independent of clutter's position, size, shape and connectivity with text. Features are based on a new technique called 'nth erosion' and clutter elements are identified with an SVM classifier. Removal is restrictive, so text attached to the clutter is not deleted in the process. The method was tested on a mix of degraded and noisy, machine-printed and handwritten Arabic and English text documents. Results show pixel-level accuracies of 97.5% and 95% for clutter detection and removal respectively. This approach was also extended with a noise detection and removal model for documents having a mix of clutter and salt-n-pepper noise.

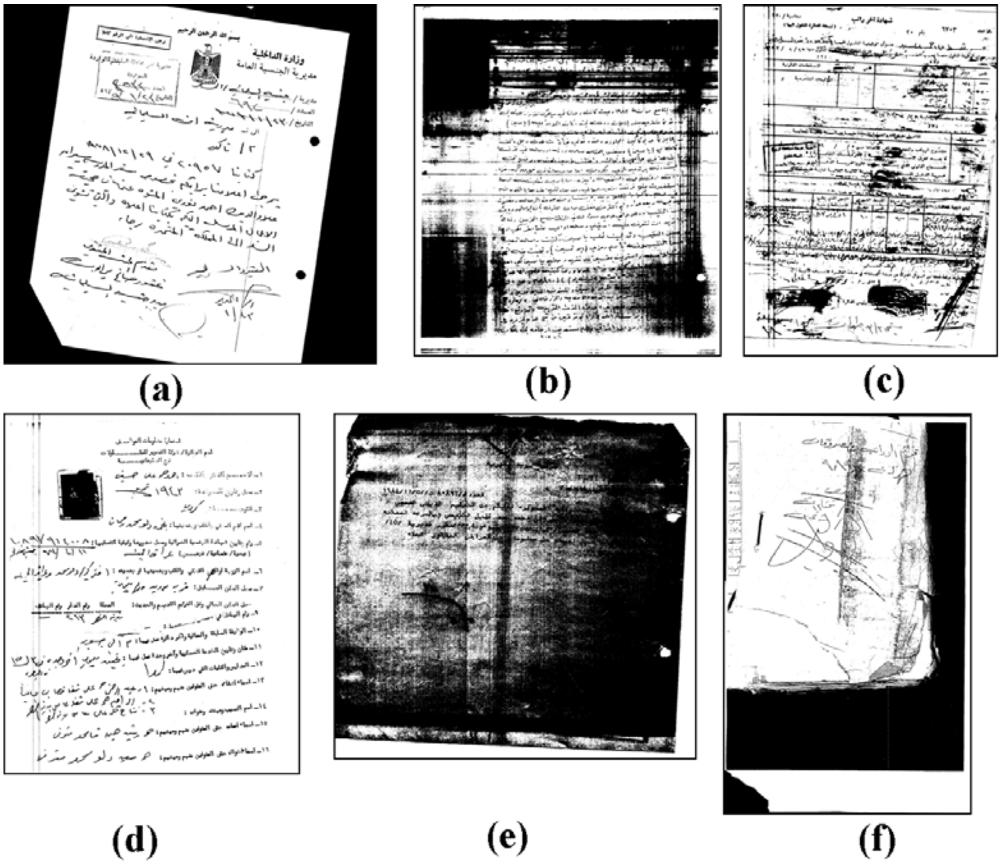


Figure 4: Examples of Images with Clutter/Noise

2.3.2 Clutter Detection

Clutter is an assemblage of non-informative foreground pixels in binary images and can result from numerous sources. While some forms of clutter like punched holes ink seeps ink blobs and copier borders typically are present before the scanning process, marginal noise results from the scanning of bound or skewed documents (see figure above) because of the gap between the gutter and scanner or between edges of paper and scanner bed. Degraded documents when scanned and binarized overlay the background patterns, formed due to degradation, on the foreground, whereas pictures and figures not binarized appropriately may give rise to clutter as well. Clearly, clutter is predominantly independent and irregular. One of the major issues with clutter is its connectivity with text. In the case of ruled line documents with clutter, a single connected component connecting clutter, ruled lines and text may appear. Complete removal of the connected component in such cases may result in tremendous loss of content. As far as we know, there has been no collective work on the detection and removal of clutter, without removing the attached text, in binary document images. Fan, Wang and Kay detect and remove marginal noise regions based on three assumptions of shape, length and position. The technique does fairly well at removing only the marginal noise without the attached text. By contrast, our technique achieves the same on all forms of clutter, without these three assumptions. Our approach is independent of clutter's position, size, shape and connectivity with text. It is also independent on the inclusion of any other type of noise. Clutter thickness is much more than maximum text stroke width present in the document, whereas thickness of ruled-lines, salt-n-pepper, stray-marks, bleed-through, blur etc. are of the order of text-stroke width. It is interesting to note that this property of clutter differentiates it from other types of noise and text. We assume clutter's thickness is greater than twice the text's maximum stroke width present in the document. Hence, thinning the foreground pixels to half the thickness of the maximum thick clutter, will erode all other text and other noise pixels from the document and will leave behind only a core of each clutter element. On the other hand, in the absence of any clutter, text strokes will be thinned to half their maximum width, maintaining a textlike pattern (albeit broken). This process of thinning the image by half of the maximum depth foreground is called half erosion. It can be computed as follows:

1. Perform distance transform DP over the set P on the binary image I, as illustrated in Equation 2
2. Calculate the maximum value dtMax = max p_P (DP)
3. Set all pixels with DP (p) < dtMax=2 to background.

The half-eroded image IHE is obtained. The features are extracted from this half-eroded image for clutter detection and can be classified easily as having clutter or not

2.3.3 Clutter Removal

Once the document image is classified as having clutter noise, the components from the half-eroded image IHE, called HE-cores or endo-clutter, are mapped to their corresponding components in the original image I. Resulting image ICC has only these detected clutter components in their original sizes.

The challenge is to remove only the clutter from these components, without deleting the text attached to it. Distance transform on a clutter component yields distance contours, with distances increasing inwards. Each distance contour contains pixels having same distance value. At distance dth, from clutter component's boundary, text branches are separated from the clutter body. This distance is called the exo-clutter distance, and the clutter component left inside is called sarco-clutter SC.. If we now dilate the sarco-clutter by exoclutter distance, we obtain the clutter in original shape with text-branches removed. Knowing the correct exo-clutter distance is the key. It is typically approximately the textstroke width, but due to various stroke widths (font sizes) present in the document, we do not make the assumption of calculating it from average or most frequent stroke-width. Only minimal erosion of clutter component is preferred (equal to exo-clutter distance), as excessive erosion tends to lose clutter's shape and hence dilation thereafter won't perfectly recover it. The number of distance contours from the endo-clutter, passing through each distance contour on the clutter-component, increases sharply at exo-clutter distance. This is due to the fact that text-branches protrude out of the shape maintained by sarco and endo-clutter. Moving inwards from the boundary of clutter-component, there is a sharp fall in number of distance contours at exo-clutter distance. This function is a monotonically decreasing function. f0(i) is the rate of change of the function, which slows down at dth. If g(i) = f00(i), dth is the index of first maxima of g(x). d di (g(i)) = 0; d2 di2 (g(i)) < 0 (5) It is not important that endo-clutter should maintain the exact shape of the clutter. The point of first sudden drop in the number of contours can predict the exo-clutter distance. The depth of the dip is proportional to the length of the text-branch. Once the exo-clutter distance dth is obtained, shrinking and expanding the clutter-component by this distance, gets the clutter out from its text-branch.

Details of this algorithm are available in a technical paper that was submitted to ICDAR 2009 and are available upon request.

2.3.4 Results

We evaluated the clutter detection and removal approach on datasets of printed and handwritten documents in English and Arabic scripts from five different sources using ImageDIFF. The dataset contains a representative set of 50 images with all forms of clutter. For clutter detection, each image is labeled as clean or noisy and reported accuracy is 97.5%. For clutter removal algorithm, we use an xml-based LAMP's GEDI tool for run-length labeling and visualization. The evaluation was based on a pixel-level criteria, as the percentage of noise pixels removed. The reported accuracy is 95%. The Figure below shows the clutter removal results of images above.

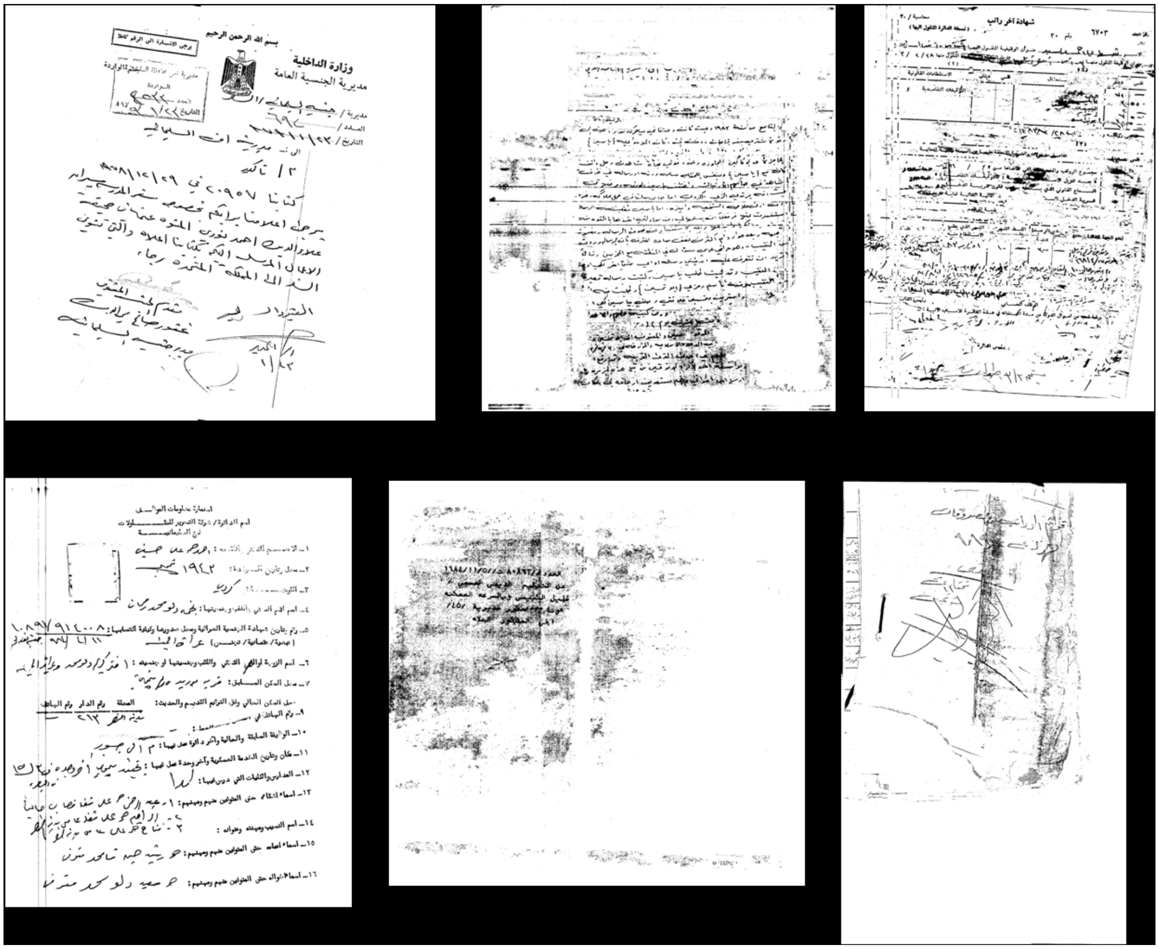


Figure 5: Resulting Cleaned Image

The following image show some additional results.

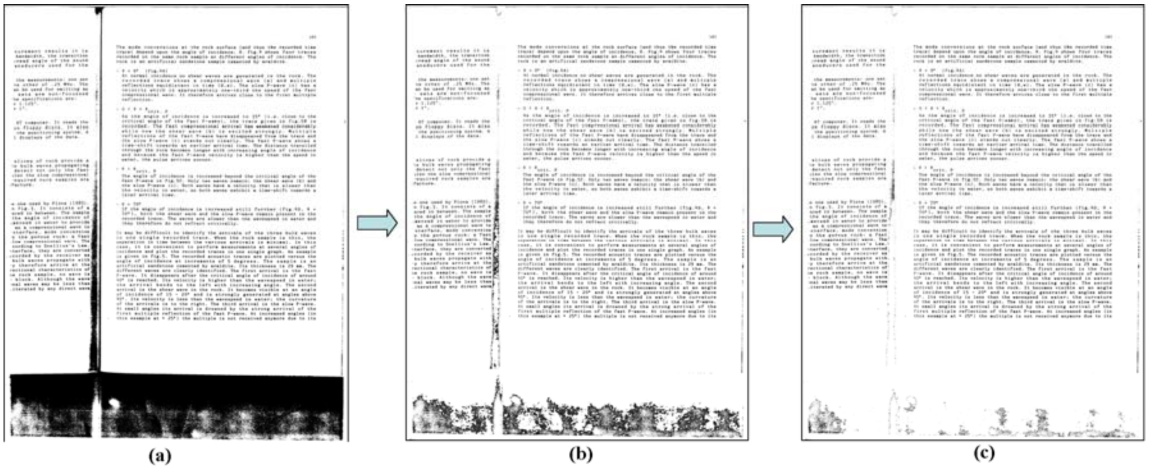


Figure 6: Example Showing Clutter/Spray Removal

2.4 Rule Line Removal

2.4.1 Introduction

The purpose of this task is to develop an algorithm for automatically removing the background rule lines in order to improve the overall quality of the document image prior to further processing steps, such as Optical Character Recognition (OCR). Two objectives are taken into consideration when designing our rule line removal algorithm. First, the algorithm must not depend on explicitly detecting rule lines prior to removing them. Methods that depend on detecting the rule lines (e.g. using Hough Transform) are prone to making estimation errors and usually have many parameters that are difficult to tune for a large distribution of document pages. Therefore, we need to design an algorithm that performs well on documents with and without rule lines. Second, the algorithm designed must not degrade the quality of the textual connected components when applied to non-rule lines pixels.

2.4.2 Approach: SVDD

Page rule lines represent a distinct data domain versus other document content types, such as text, logos and signatures. The classic approach in such cases is using a multi-class classifier. However, multi-class classifiers are not suitable for this problem because (1) the required large training sets from different classes and (2) the limitations in currently used voting schemes, as we have pointed out in [1] and (3) the objective is to identify pixels that belong to the rule line class versus other pixel types, rather than actually identifying other pixel types.

Consequently, we will use data domain description methods for characterizing page rule lines. The objective is to completely describe the data domain while rejecting outlier data. For this purpose, we used Support Vector Data Description (SVDD) to describe the observations extracted from rule lines. SVDD is inspired by Support Vector Machines (SVM) and it describes the data domain using a hypersphere in a higher dimensional feature space. The hypersphere is characterized by the support vectors of the data domain.

Binary document images restrict feature extraction choices. Therefore, in order to obtain a pseudo-gray scale representation of the document image, we compute the directional gradient of the Distance Transform (DT) of the binary image. The magnitude of the gradient is discarded and the direction of the gradient is use a pseudo gray scale representation.

The texture of the pseudo gray image will be used to distinguish between pixels that belong to page rules lines versus text pixels. Texture features are extracted using a Gabor Filter Bank.

2.4.3 Approach using Linear Subsapces

The Figure below shows a zoomed segment of a rule line in the document shown above. Due to various sources of noise (e.g. scanning and binarization), the shape of the line at different areas varies considerably, which will be reflected in the computed feature vectors. In order to identify rule line pixels, we need to construct a model that can robustly describe the varying nature of the line shape.

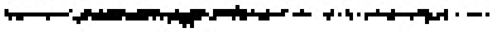


Figure 7: Example of broken Rule Line

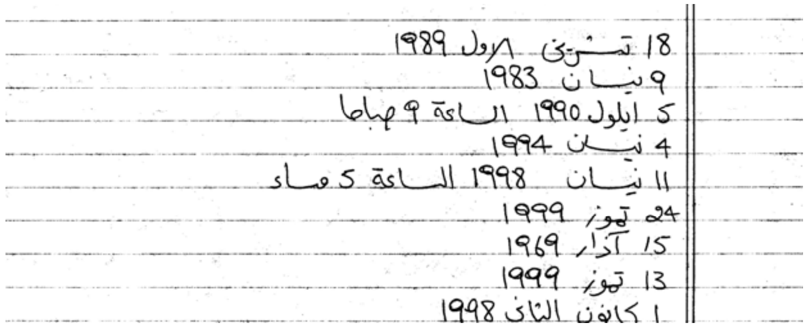

18 تشرين ا/ول 1989
9 نيسان 1983
5 ايلول 1990 الساعة 9 صباحا
4 نيسان 1994
11 نيسان 1998 الساعة 5 مساء
24 تموز 1999
15 آذار 1969
13 تموز 1999
1 كانون الثاني 1998

Figure 8: Text Touching Rule Lines

In printed documents, this problem can be solved using a classic binary classifier, which models the features of line pixels as one class and the features of the foreground text as another class. However, this approach is not suitable for handwritten documents because of wide variations of the writers styles. Many data description method have been proposed to address this problem, such as the Support Vector Data Description (SVDD). Above, we have used to SVDD to model rule lines and achieved a median harmonic mean of 75% on our test set. However, two difficulties arise when using SVDD and other kernel-based methods. First, a $O(N^2)$ kernel matrix must be computed (and later inverted), where N is the number of training examples. Consequently, very large training sets cause many computational, numerical and memory problems. Second, SVDD-based methods cannot be incrementally updated. If new training samples become available, the training process must be totally repeated, rather than adapting the existing model.

Subspace methods present an intriguing alternative to solve data description problems. They have been frequently used to solve various computer vision and image processing problems. In order to construct a subspace model for the rule lines, we extract feature vectors using central moments and histogram properties, from a set of training images containing only rule lines, as shown in Figure below. Feature vectors are incrementally projected on the subspace and the reconstruction error is computed. If the reconstruction error is smaller than a threshold, then the subspace is capable of representing the feature vector. Otherwise, the residual is normalized and the subspace is augmented.

Figure 9: Original Page Source

In order to evaluate our rule line removal algorithm, we built a semi-synthetic dataset of 50 document images. The 50 images are generated as a cross product between five rule line-only images and 10 text-only images from the MADCAT LDC dataset.

Figure 4 illustrates the precision-recall characteristics of the rule line removal algorithm using the subspace method, using different threshold values. The figure shows that the method can achieve approximately 90% precision and recall. Also, we compute the F1 harmonic mean in order to combine the precision and recall into one metric. Figure 5 shows that we can achieve an F1 score of approximately 85%, compared to the 75% achieved using the SVDD method.

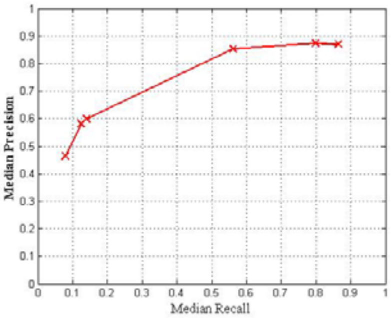


Figure 10: Precision/Recall using Subspace Method

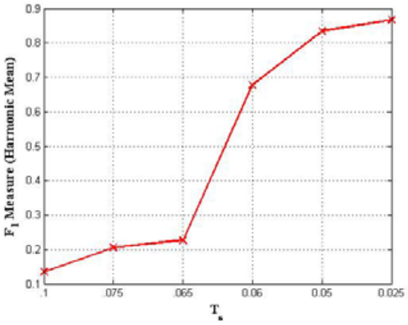


Figure 11: F1 Sore for Previous Figure

2.5 Unsupervised Text Line Detection

2.5.1 Introduction

Automatic detection of text lines in Arabic handwritten documents is a fundamental step prior to any further Optical Character Recognition (OCR) or automatic translation. The purpose of this task is to develop a robust, unsupervised text line detection algorithm for Arabic handwritten documents. Such an algorithm must have two main characteristics – (1) the designed algorithm must accurately detect text lines in different orientations and (2) the designed algorithm must detect text lines in near real-time.

2.5.2 Affinity Propagation

Affinity Propagation (AP) [2] is a recently developed unsupervised data clustering technique. AP depends on passing messages between data points. Two types of messages are passed between data points – (1) *availability* messages indicate the availability of a given data point to serve as a cluster center for the neighboring data points and (2) *responsibility* messages indicate the willingness of a certain data point to be a member of a specific cluster. These two types of messages are passed based on the *similarity* between pairs of data points.

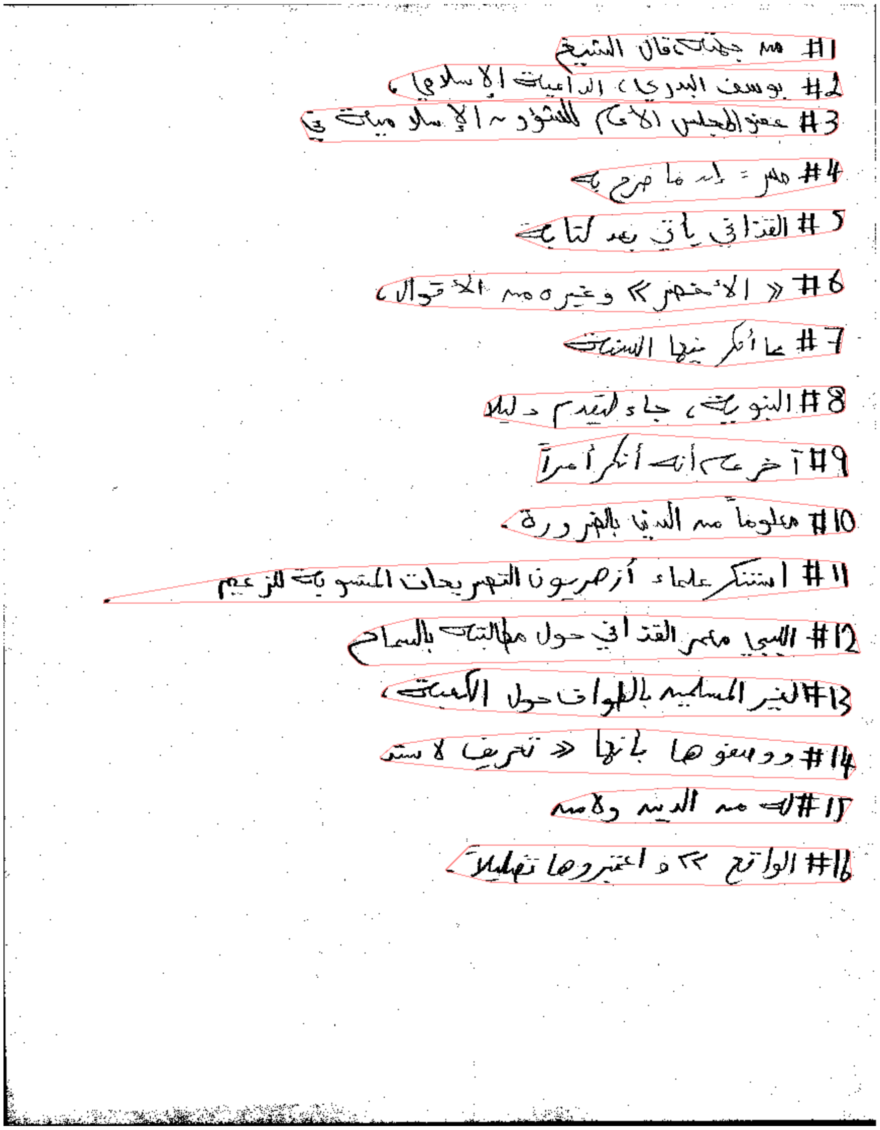
AP has two main advantages. First, the number of cluster need not be *a priori* specified. Second, AP is very flexible because it is independent on the method used to compute the pair-wise similarities.

2.5.3 Approach

We adopt an unsupervised, hierarchical clustering methodology to automatically detect text lines in multiple orientations in Arabic handwritten documents. Text lines are detected in two steps. First, the orientations of the document's connected components are used to compute a pair-wise similarity matrix using exponential kernels. AP is applied to the orientation similarity matrix to cluster the document into a number of text areas with homogenous orientations.

For each of the detected text areas, a pair-wise similarity matrix is computed using exponential kernels using the projections perpendicular to the orientation. AP is applied using the similarity matrix and text lines are detected. The results computed using ImageDiff are:

Precision	Recall	F1 score
88.2%	**94.15%**	**91.06%**



#1 من جهته قال الشيخ
#2 يوسف البدري، الداعية الإسلامي،
#3 عضو المجلس الأعلى للشئون الإسلامية في
#4 مصر - إنه ما صرح به
#5 القذافي يأتي بعد كتابه
#6 « الأخضر » وغيره من الأقوال
#7 ما أنكر فيها السنة
#8 النبوية، جاء ليقدم دليلا
#9 آخر على أنه أنكر أمرا
#10 معلوما من الدين بالضرورة.
#11 استنكر علماء أزهريون التصريحات المنسوبة للزعيم
#12 الليبي معمر القذافي حول مطالبته بالسماح
#13 لغير المسلمين بالطواف حول الكعبة
#14 ووصفوها بأنها « تحريف لا يستر
#15 له من الدين ولا سند
#16 الواقع »، واعتبروها تضليلا.

Figure 12: Sample Image

2.5.4 References

[1] Wael Abd-Almageed and David Doermann, “Text Line Detection in Arabic Handwritten Documents,” To be submitted to International Conference on Document Analysis and Recognition, Barcelona, Spain, July 2009.

[2] B. Frey and D. Dueck,” Clustering by Passing Messages Between Data Points,” Science, pp. 972—976, 2007.

2.6 Document Zone Classification

2.6.1 Introduction

Identifying the content type of document zones is a fundamental component of modern document analysis systems. For example, identifying zone type allows the application of content-specific algorithms and can improve Optical Character Recognition (OCR) by providing domain knowledge. More importantly, zone type identification is crucial to indexing and retrieval of large document databases.

Broadly speaking, content analysis algorithms can be classified as one of three main approaches -- (1) type-specific detection, (2) page classification and (3) zone classification. Type-specific approaches, emphasize finding specific types of zones, such as text regions, logos, mathematical expressions and tables. Page classification approaches, assume the content of the entire page is of a single type (e.g. title page or index page) and a classifier is used to determine the page content. Finally, zone classification approaches assume that the page is segmented into zones with independent content types. Low level image features are extracted from each zone and a statistical classifier is used to label the different zones into one of possible content types (e.g. text, math, etc.).

2.6.2 Approach

Classically, multi-class classification problems have been treated by either constructing a number of one-against-all binary classifiers or constructing a number of one-against-one binary classifiers, with the latter method reported to be the most successful one. A voting scheme is then used to determine the required class label. One-against-one method suffers a principal limitation. If the observation being tested does not belong to either of the two classes on which the classifier is trained, a vote will be incorrectly cast, biasing the final classification outcome. To overcome this limitation, we used a novel approach by constructing hybrid of the two approaches.

We construct a number of two-against-all classifiers, which will be used to determine if a vote will be cast to a given class. Based on the decision of the two-against-all classifier, a regular one-against-one classifier is used to label the test sample and cast the vote. This mechanism prevents casting incorrect votes if the test sample does not belong to either of the two classes modeled by the one-against-one classifier.

2.6.3 Evaluation

We applied our new approach to the University of Washington (UW) data set. The dataset contains 1690 document images with a total of 24531 zones. We considered 10 different zone types -- -- chemical drawing, small text and symbols, drawing, halftone, logo or seal, map, math, ruling, table and large text. Using SVM as the underlying binary classifier, hybrid classifier achieves 97.3% classification accuracy. To our knowledge, the best reported performance on this dataset is 98.45% of Wang et al. [2]. However, the UW data set is significantly unbalanced with 87.9% small text samples and 0.065% logo and seal samples and 0.057% map samples, which skews the classifier performance. The approach of Wang et al. [2] has not been tested on a balanced data set.

In order to further assess our proposed algorithm, we eliminated small text, logo and seal and map classes from the dataset which leaves a balanced data set of seven zone classes. The hybrid classifier achieves a comparable 96.6% accuracy. No result is available for a similar experiment from [2]. Moreover, similar experiments show that the hybrid classification scheme out-performs the classic one-against-one scheme. The following table summarizes the results.

	1-vs-1	Wang et al. [2]	Hybrid
Unbalanced	93.1%	98.45	97.3%
Balanced	88.2%	N/A	96.6%

2.6.4 Publications

[1] Wael Abd-Almageed, Mudit Agrawal, Wontaek Seo and David Doermann, “Document Zone Classification using Partial Least Squares and Hybrid Classifiers,” Accepted, International Conference on Pattern Recognition, Tampa, Florida, December, 2008.

[2] W. Wang, I. T. Phillips and R. M. Haralick, “Document Zone Content Classification and Its Evaluation”, Pattern Recognition, 39(1), 2006.

3 Summary and Conclusions

The algorithms demonstrated here are representative of the types of algorithms that PETS can evaluate. They differ in zone type, zone shape and pixel level composition. Detailed reports are included with the datasets and software.