ABSTRACT

Title of dissertation: CONTENT RECOGNITION AND CONTEXT MODELING FOR DOCUMENT ANALYSIS AND RETRIEVAL

Guangyu Zhu, Doctor of Philosophy, 2009

Dissertation directed by: Professor Rama Chellappa
Department of Electrical and Computer Engineering

The nature and scope of available documents are changing significantly in many areas of document analysis and retrieval as complex, heterogeneous collections become accessible to virtually everyone via the web. The increasing level of diversity presents a great challenge for document image content categorization, indexing, and retrieval. Meanwhile, the processing of documents with unconstrained layouts and complex formatting often requires effective leveraging of broad contextual knowledge.

In this dissertation, we first present a novel approach for document image content categorization, using a lexicon of shape features. Each lexical word corresponds to a scale and rotation invariant local shape feature that is generic enough to be detected repeatably and is segmentation free. A concise, structurally indexed shape lexicon is learned by clustering and partitioning feature types through graph cuts. Our idea finds successful application in several challenging tasks, including content recognition of diverse web images and language identification on documents composed of mixed machine printed text and handwriting.

Second, we address two fundamental problems in signature-based document image retrieval. Facing continually increasing volumes of documents, detecting and recognizing unique, evidentiary visual entities (*e.g.*, signatures and logos) provides a practical and reliable supplement to the OCR recognition of printed text. We propose a novel multi-scale framework to detect and segment signatures jointly from document images, based on the structural saliency under a signature production model. We formulate the problem of signature retrieval in the unconstrained setting of geometry-invariant deformable shape matching and demonstrate state-of-the-art performance in signature matching and verification.

Third, we present a model-based approach for extracting relevant named entities from unstructured documents. In a wide range of applications that require structured information from diverse, unstructured document images, processing OCR text does not give satisfactory results due to the absence of linguistic context. Our approach enables learning of inference rules collectively based on contextual information from both page layout and text features.

Finally, we demonstrate the importance of mining general web user behavior data for improving document ranking and other web search experience. The context of web user activities reveals their preferences and intents, and we emphasize the analysis of individual user sessions for creating aggregate models. We introduce a novel algorithm for estimating web page and web site importance, and discuss its theoretical foundation based on an intentional surfer model. We demonstrate that our approach significantly improves large-scale document retrieval performance.

CONTENT RECOGNITION AND CONTEXT MODELING FOR DOCUMENT ANALYSIS AND RETRIEVAL

by

Guangyu Zhu

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2009

Advisory Committee:
Professor Rama Chellappa, Chair/Advisor
Dr. David S. Doermann, Co-Advisor
Professor Carol Y. Espy-Wilson
Professor Larry S. Davis
Professor David W. Jacobs

Dedication

I dedicate this dissertation to my parents and my wife.

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost, I would like to thank my advisors, Dr. David Doermann and Professor Rama Chellappa, for teaching me so much and pushing me to excellence in academic and professional development. It has been a pleasure to work with and learn from such extraordinary researchers.

I would also like to thank my committee members, Professors Carol Y. Espy-Wilson, Larry Davis, and David Jacobs for serving on the committee and reviewing the dissertation. I was extremely lucky to have such a superstar committee and I will never forget the experience. I am especially grateful to Professor David Jacobs, who gave me valuable advice on my research and education.

My LAMP colleagues have enriched my graduate life in many ways and I received helpful feedback from many of them, including Stefan Jaeger, Yefeng Zheng, Yi Li, Xiaodong Yu, Daniel DeMenthon, Zhe Lin, Daniel Ramsbrock, and Mudit Agrawal. I really enjoyed my collaborations and discussion with these brilliant guys.

I want to give thanks to my co-workers from Yahoo! Labs and IBM Almaden Research Center, who introduced me to challenging research problems and gave me many opportunities early in my career.

I owe my deepest thanks to my parents who have always supported and motivated me. Finally, I thank my wife Yuan Yuan, who gives meaning to everything I do.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

The extent of document analysis has broadened significantly over recent years. The explosive growth of the web has enabled access to much larger and more diverse collections of documents. While the primary purpose of a document remains unaltered – to facilitate the transfer of information [1] – it can be presented in a much larger variety of contexts and formats to readers. The increasing level of diversity and complexity presents numerous challenges for the analysis and retrieval of documents, both in hard copy and electronic form.

In many areas of document image analysis, assumptions based on structured, narrow-domain documents are no longer valid in the presence of complex, heterogeneous content. Interpretations of the content type, language, and genre associated with a document image become essential tasks. A document image processing work flow, as shown in Figure 1.1, would require image content category recognition, indexing, and search capabilities that can be generalized to semi- or unstructured, broad-domain image collections. The goal is no longer limited to the conversion of text to electronic form, but can now extend semantic analysis toward broader content.

For the retrieval of electronic documents, ranking continues to be of utmost importance. The quality of documents presented to web users has significant impact

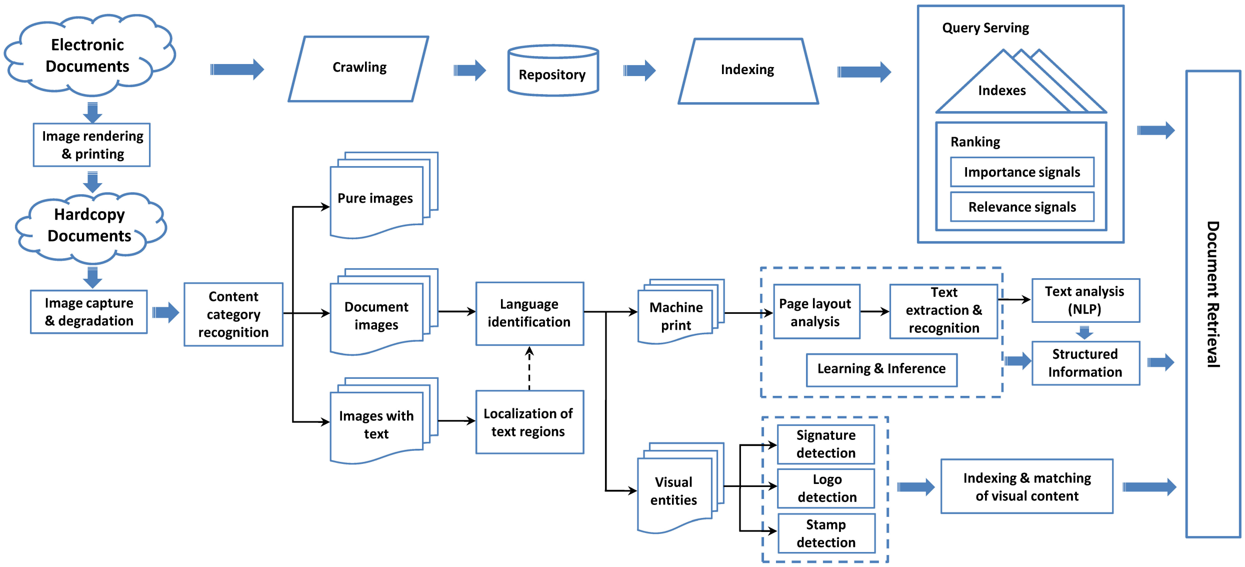


Figure 1.1: Key challenges facing document image analysis and retrieval systems that process diverse document collections.

on their daily lives, yet people have extremely limited spans of attention and have difficulty dealing with the massive quantity of pages on the web. Experiments on users' interaction with the list of ranked results returned by web search engines have shown that a user seldom looks for documents beyond the first page of results [2].

This dissertation presents work in four main areas related to the analysis and retrieval of document images and semi-structured electronic documents. In the following sub-sections, we motivate these problems and present an overview of our approaches.

1.1 Document Content Category Recognition

Many computer vision and image analysis problems begin by obtaining high-level content type interpretation for an image. As shown in Figure 1.1, information about the content category largely determines how we process the image in ensuing tasks. Having recognized text content and the language, a document analysis system can process and index images containing predominantly machine printed text.

We focus on two image content category recognition problems central to heterogeneous document image collections. The first problem lies in the categorization of a general image into different content classes. The second problem arises from language identification for documents that are composed of mixed machine printed text and handwriting, which has been an open research problem. For systems that process diverse multilingual document images, the performance of language identification is crucial for a broad range of tasks — from determining the correct optical character recognition (OCR) engine for text extraction to document indexing, translation, and search. In character recognition, for instance, almost all existing work requires knowledge of the script and/or language of the processed document [3].

In this dissertation, we propose a novel approach for document image content categorization using image descriptors constructed from a lexicon of shape features. We encode local text structures using scale and rotation invariant lexical words, each representing a segmentation-free shape feature that is generic enough to be detected repeatably. We learn a concise, structurally indexed shape lexicon by clustering and partitioning similar feature types through graph cuts. Our approach is extensible and does not require skew correction, scale normalization, or segmentation. We demonstrate our approach on two challenging document image content recognition problems: 1) The classification of 4,500 web images from Google Image Search into three content categories — pure image, image with text, and document image, and 2) Language identification of eight languages (Arabic, Chinese, English, Hindi, Japanese, Korean, Russian, and Thai) on a 1,512 complex document image database composed of mixed machine printed text and handwriting.

1.2 Content-based Document Image Analysis and Retrieval

Visual content, such as signatures [4], logos [5], and stamps [6], present convincing evidence of document source and provide an important form of indexing for document image processing and retrieval. Detecting, segmenting, and matching these free-form objects from clustered background pose unique challenges. Signature detection, for example, is an open document analysis problem [7].

In this dissertation, we study two fundamental problems in signature-based document image retrieval. First, we propose a novel multi-scale approach to jointly detecting and segmenting signatures from document images. Rather than focusing on local features that typically have large variations, our approach captures the structural saliency using a signature production model and computes the dynamic curvature of 2-D contour fragments over multiple scales. This detection framework is general and computationally tractable. Second, we treat the problem of signature retrieval in the unconstrained setting of translation, scale, and rotation invariant deformable shape matching. We propose two novel measures of shape dissimilarity based on anisotropic scaling and registration residual error, and we present a supervised learning framework for combining complementary shape information from different dissimilarity metrics using linear discriminant analysis (LDA). We quantitatively study state-of-the-art shape representations, shape matching algorithms, measures of dissimilarity, and the use of multiple instances as query in document image retrieval. We further compare our matching approach to the state of the art in the task of off-line signature verification.

1.3 Processing Unstructured Document Images

Hard copy documents have been the most common form of information-conveying vehicle over centuries, and people have developed an exceptional capability to extract information from diverse paper sources. Once hard copy documents are captured electronically as scanned images, however, the automated processing of documents with unconstrained layouts and diverse formatting becomes more difficult. In many applications that require such capability, applying traditional language modeling techniques to the stream of OCR text does not produce satisfactory results due to the absence of linguistic context. Extracting structured information, such as named entities, from unstructured document images remains an unsolved research challenge.

In this dissertation, we present a model-based approach for extracting relevant named entities from unstructured document images by combining rich page layout features in the image space with OCR text. We demonstrate our named entity extraction approach in an expense reimbursement system and evaluate its performance on large collections of degraded, real-world receipt images.

1.4 Improving Document Search Ranking

User browsing information, particularly their non-search related activity, reveals important contextual information on the preferences and the intents of web users. A document retrieval system may improve both search ranking performance and user experience by effective leveraging of contextual knowledge gained from general web user behavior data.

In this dissertation, we demonstrate the importance of mining general web user behavior data for improving document ranking and other web search experience, with an emphasis on analyzing individual user sessions to create aggregate models. In this context, we introduce *ClickRank*, an efficient, scalable algorithm for estimating web page and web site importance from general web user behavior data. We lay out the theoretical foundation of ClickRank based on an intentional surfer model and discuss its properties. We quantitatively evaluate its effectiveness for the problem of web search ranking, showing how it contributes significantly to retrieval performance as a novel web search feature. We demonstrate that the results produced by ClickRank for web search ranking are highly competitive with those produced by other approaches, yet achieved with better scalability and substantially lower computational costs. Finally, we discuss novel applications of ClickRank in providing enriched user web search experience, highlighting the usefulness of our approach for non-ranking tasks.

1.5 Outline of the Dissertation

The remainder of this dissertation is organized as follows. Chapter 2 describes our approach to document image content category recognition using a lexicon that is composed of a wide variety of local shape features. We demonstrate its application in two challenging tasks – the categorization of diverse web images and the language identification of document images involving a blend of machine printed text and handwriting. Chapter 3 presents our approaches to signature detection, segmentation, and matching for document image retrieval. Chapter 4 describes our approach to extracting relevant named entities from unstructured document images. We present its successful application in an automated expense reimbursement system. Chapter 5 studies the problem of using contextual information from general user behavior data online to improve web search ranking. The main ideas and contributions of the dissertation are summarized in Chapter 6.

Chapter 2

Image Content Category Recognition

2.1 Introduction

Image content categorization has become a pressing problem in computer vision as we face a phenomenal increase in the diversity of visual content. Content category recognition aims to reduce the semantic gap for ensuing tasks by providing usage-oriented content description that can be utilized in individual applications. For vision systems involving high-volume, complex, and heterogeneous image data, effective high-level content interpretation is essential prior to object detection or object category recognition at a finer level.

In this chapter, we focus on the most pervasive content within documents — text. Once text content and the language are recognized, images containing text can be processed by an optical character recognition (OCR) system and indexed. Toward this end, however, many unsolved challenges to image content category recognition still exist.

First, we consider the recognition of an image's primary content as one of three content categories — pure image (*e.g.*, natural image and human photos), image with text (see examples in Figure 2.1), or document image. This sort of automated content categorization has broad impact in image search, and content-based image indexing and retrieval.

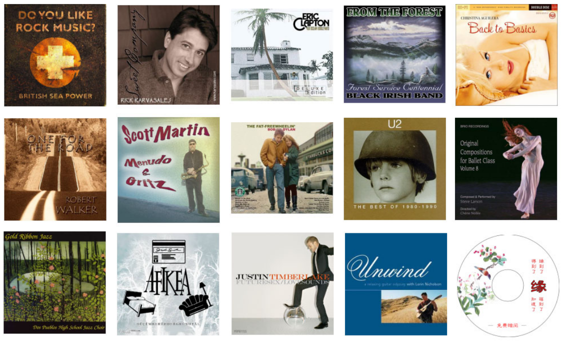
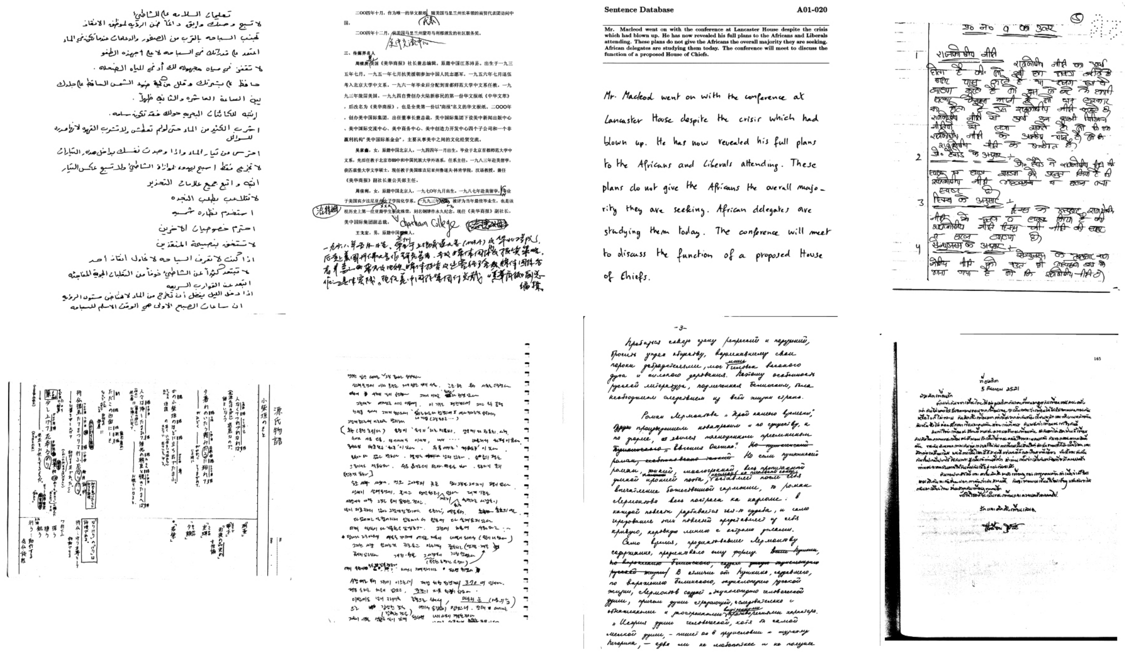


Figure 2.1: Examples of images returned by Google Image Search using the keyword "CD cover".

Second, we study the problem of recognizing the primary language of a document image in an unconstrained setting. This presents a fundamental research challenge for those systems that need to process diverse multilingual document images automatically, such as Google Book Search [8] or an automated global expense reimbursement application [9]. Almost all work to date on OCR requires that the script and/or language of the processed document be known [3]. The performance of language identification is crucial for the success of a wide range of tasks — from determining the correct OCR engine for text extraction to document indexing, translation, and search [10].

Progress in the field of language identification has focused almost exclusively on machine printed text. Document collections, as shown in Figure 2.2, often contain a diverse and complex mixture of machine printed and unconstrained handwritten content, and vary tremendously in font and style. Language identification on document images involving diverse content types, including unconstrained handwriting, is still an open research area [7] and, to our best knowledge, no reasonable solutions have been presented in the literature.



Sentence Database A01-020

Mr. Macleod went on with the conference at Lancaster House despite the crisis which had blown up. He has now revealed his full plans to the Africans and Liberals attending. These plans do not give the Africans the overall majority they are seeking. African delegates are studying them today. The conference will meet to discuss the function of a proposed House of Chiefs.

Mr. Macleod went on with the conference at Lancaster House despite the crisis which had blown up. He has now revealed his full plans to the Africans and Liberals attending. These plans do not give the Africans the overall majority they are seeking. African delegates are studying them today. The conference will meet to discuss the function of a proposed House of Chiefs.

Figure 2.2: Examples from the Maryland multilingual database [11] and the IAM handwriting DB3.0 database [12]. Languages in the top row are Arabic, Chinese, English, and Hindi. In the lower row are Japanese, Korean, Russian, and Thai.

The problem of language identification for handwriting also highlights several common challenges facing category recognition of diverse content. First, handwriting exhibits much larger variability compared to machine printed text. Handwriting variations due to style, cultural, and personalized differences are typical [13], which significantly increase the diversity of shapes found. Text lines in handwriting are curvilinear and the gaps between neighboring words and lines are far from uniform. No well-defined baselines exist for handwritten text lines, even by linear or piecewise-linear approximation [14]. Second, automatic processing of off-line document image content should be robust in the presence of unconstrained document layouts, formatting, and image degradations.

Our approach takes the view that the intricate differences between text can be effectively captured using segmentation-free shape features and structurally indexed shape descriptors. Low-level local shape features serve well for this purpose because they can be detected robustly in practice, without detection or segmentation of high-level entities, such as text lines or words. As shown in Figure 2.3, visual differences between handwriting samples across languages are captured well by different configurations of neighboring line segments, which provide rich description of local text structures.

We propose a novel approach for document image content recognition, using image descriptors built from a lexicon of generic low-level shape features that are translation, scale, and rotation invariant. To construct a structural index among large number of diverse features, we dynamically partition the space of shape primitives by clustering similar feature types. We formulate feature partitioning as a graph cuts problem with the objective of obtaining a concise and globally balanced lexicon index by sampling the training data. Each cluster in the lexicon is represented by an exemplary lexical word, making association of feature type efficient. We obtain competitive document image content categorization performance using a multi-class SVM classifier.

This chapter is structured as follows. Section 2.2 reviews related work. In Section 2.3, we describe our algorithm for learning the shape lexicon and present a document image content recognition approach using the shape lexicon. We discuss experimental results in Section 2.4, and conclude in Section 2.5.

2.2 Related Work

We first present a comprehensive overview of existing approaches on whole-image categorization and script/language identification, respectively. We then highlight work related to our approach on contour-based learning.

2.2.1 Image Content Categorization

Little literature has directly addressed the problem of content recognition for heterogeneous image repositories. However, several approaches based on different motivations have demonstrated good performance in tasks that involve diverse objects. Oliva and Torralba [15] developed a holistic image representation for scene recognition called spatial envelope, which characterizes the dominant spatial structure of a scene using a set of discriminative energy spectrum templates. Another fairly intuitive approach treats blocks of text as texture [16]. One widely used rotation invariant feature for texture analysis is the local binary patterns (LBP) proposed by Ojala *et al*. [17]. LBP captures spatial structure of local image texture in circular neighborhoods across angular space and resolution, and has demonstrated state-of-the-art results in a wide range of whole-image categorization problems involving diverse data [18, 19].

2.2.2 Language Identification

Prior literature on script and language identification has largely focused on the domain of machine printed documents. These works can be broadly classified into three categories — statistical analysis of text lines [20, 21, 22, 23, 24], texture analysis [16, 25], and template matching [26].

Statistical analysis using discriminating features extracted from text lines, including distribution of concavities [20, 22], horizontal projection profile [21, 23], and vertical cuts of connected components [24], has proven effective on homogeneous collection of machine printed documents. These approaches, however, do have major limitations for handwriting. First, they assume uniformity among printed text, and require precise baseline alignment and word segmentation. Freestyle handwritten text lines are curvilinear, and, in general, have no well-defined baselines, even by linear or piecewise-linear approximation [14]. Second, it is difficult to extend these methods to a new language, because they employ a combination of hand-picked and trainable features and a variety of decision rules. In fact, most of these approaches require effective script identification to discriminate between selected subset of languages, and use different feature sets for script and language identification, respectively.

Script identification using rotation invariant texture features, including multichannel Gabor filters [16] and wavelet log co-occurrence features [25], were demonstrated on small blocks of printed text with similar characteristics. However, no results were reported on full-page documents that involve variations in layouts and fonts. Script identification for printed words was explored in [27, 28] using texture features between a small number of scripts.

Template matching approach computes the most likely script by probabilistic voting on matched templates, where each template pattern is of fixed size and rescaled from a cluster of connected components. The script and language identification system developed by Hochberg *et al.* [26] at Los Alamos National Laboratory based on template matching can process 13 machine printed scripts without explicit assumptions of distinguishing characteristics for a selected subset of languages. Template matching is intuitive and delivers state-of-the-art performance when the content is constrained (*i.e.*, printed text in similar fonts). However, templates are not sufficiently flexible to generalize across large variations in fonts or handwriting styles typical to diverse datasets [26]. From a practical view, the system needs to learn the discriminability of each template through labor-intensive training, the extent to which requires tremendous amounts of supervision and further limits applicability.

There exists minimal literature on language identification for handwriting. To the best of our knowledge, the experiments of Hochberg *et al.* [13] on script and language identification of handwritten document images offers the only reference on this topic in the literature. They used linear discriminant analysis based on five simple features of a connected component, including relative centroid location and aspect ratio. The approach is demonstrated to be sensitive to large variations across writers and diverse document content [13]. In their experiments, irregularities in the document, including machine printed text, illustrations, markings, and handwriting in different orientation from main body, were removed manually by image editing from their evaluation dataset.

2.2.3 Contour-based Learning

Learning contour features is an important aspect in many computer vision problems. Ferrari *et al*. [29] proposed scale-invariant adjacent segments (kAS) features extracted from the contour segment network of image tiles, and used them in a sliding window scheme for object detection. By explicitly encoding both geometric and spatial arrangement among the segments, kAS descriptor demonstrates state-of-the-art performance in shape-based object detection, and outperforms descriptors based on interest points and histograms of gradient orientations [30]. However, kAS descriptor is not rotation invariant, because segments are rigidly ordered from left to right. This limits the repeatability of high-order kAS, and the best performance in [29] is reported when using 2AS.

In handwriting recognition literature, one interesting work, also motivated by the idea of learning a feature codebook, is the study on writer identification by Schomaker *et al*. [31]. Writer identification assumes that the language is known beforehand and aims to distinguish between writers based on specific characteristics of handwriting. Schomaker *et al*. used closed-contour features of ink blobs directly, without any shape descriptor. The difference between contour features is computed using Hamming distance. The low-dimensional codebook representation presented in [31] is based on Kohonen self-organizing map (SOM) [32]. Their approach demands good segmentation and is not scale or rotation invariant. To account for size variations, for instance, SOMs need to be computed at multiple scales, which requires large training data and is computationally expensive.

2.3 Recognizing Image Content Using Shape Lexicon

Recognition of diverse visual content needs to account for large variations, because content appears in many forms and contexts. The scope of the problem is intuitively in favor of low-level shape primitives that can be detected repeatably. Rather than focusing on selection of class-specific features, our approach aims to distinguish intricate differences between content types collectively using the statistics of a large variety of generic, geometrically invariant feature types (lexical words) that are structurally indexed. Our emphasis on the generic nature of lexical words provides a different perspective to recognition, which has traditionally focused on finding sophisticated features or visual selection models. This may limit generalization performance.

We explore the kAS contour feature recently introduced by Ferrari *et al.* [29], which consists of a chain of k roughly straight, connected contour segments. Specifically, we focus on the case of triple contour segments, which strike a balance between lower-order contour features that are not properly discriminative and higher-order ones that are less likely to be detected robustly.

2.3.1 Extraction of Contour Feature

We perform computation locally, which means our approach detects features in a highly efficient manner. First, we compute edges using the Canny edge detector [33], which consistently demonstrates good performance on text content and gives precise localization and unique response. Second, we group contour segments by connected components and fit them locally into line segments. Then, within each connected component, we extract every triplet of connected line segments that starts from the current segment. Figure 2.3 provides visualization of the quality of detected contour features by our approach using random colors.

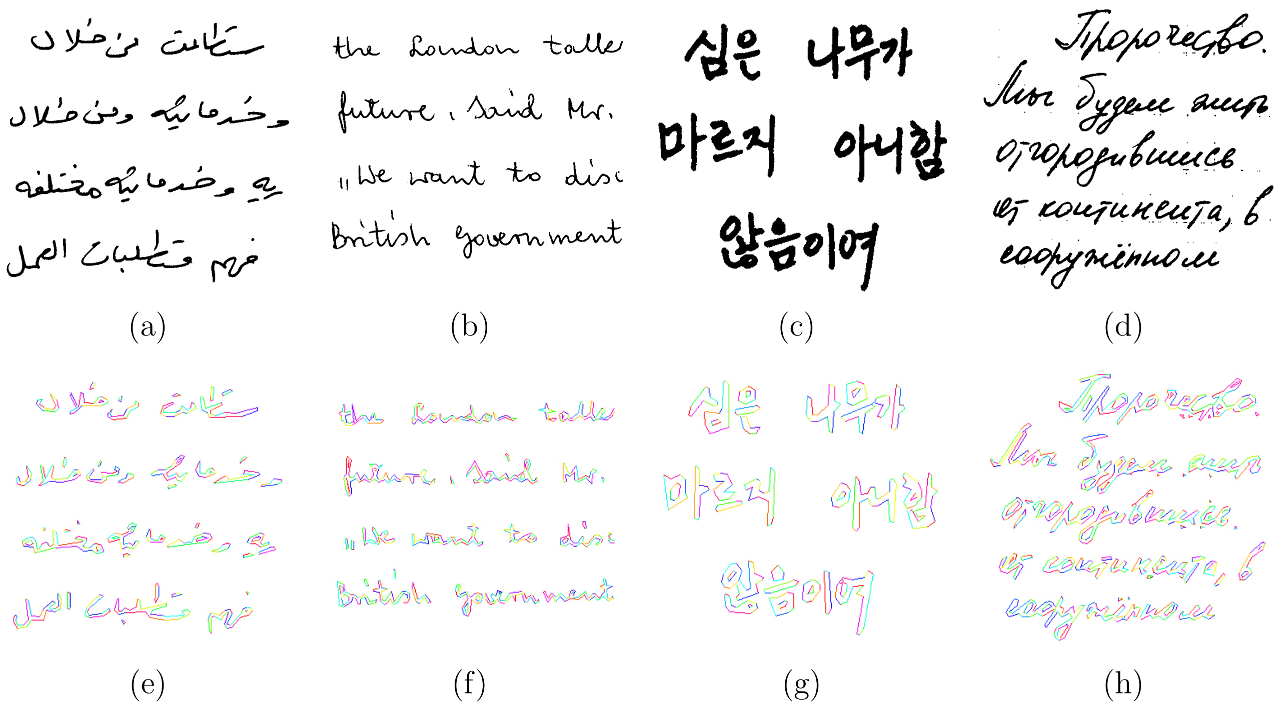
Our feature detection scheme requires only linear time and space in the number of contour fragments n, and is highly parallelizable. It proves much more efficient and stable than [29], which requires construction of contour segment network and depth first search from each segment, leading to $O(n \log(n))$ time on average and $O(n^2)$ in the worst case.

We encode object contours in a translation, scale, and rotation invariant fashion by computing orientations and lengths with reference to the first detected line segment. A contour feature C can be compactly represented by an ordered set of lengths and orientations of c_i for $i \in \{1, 2, 3\}$, where c_i denotes line segment i in C. This is distinct from the motivation of kAS descriptor that attempts to enumerate spatial arrangements of contours within local regions. Furthermore, kAS descriptor does not take rotation invariance into account.

2.3.2 Measure of Dissimilarity

The overall dissimilarity between two contour features can be quantified by the weighted sum of the distances in lengths and orientations. We use the following generalized measure of dissimilarity between two contour features C_a and C_b

$$d(C_a, C_b, \boldsymbol{\lambda}) = \boldsymbol{\lambda}_{\text{length}}^T \mathbf{d}_{\text{length}} + \boldsymbol{\lambda}_{\text{orient}}^T \mathbf{d}_{\text{orient}}, \tag{2.1}$$



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 2.3: Shape differences are captured locally by a large variety of neighboring contour features. (a)-(d) Examples of handwriting from four different languages. (e)-(h) Detected contour features by our approach, each shown in a random color.

where vectors $\mathbf{d}_{\text{length}}$ and $\mathbf{d}_{\text{orient}}$ are composed of the distances between contour lengths and orientations, respectively. $\boldsymbol{\lambda}_{\text{length}}$ and $\boldsymbol{\lambda}_{\text{orient}}$ are their corresponding weight vectors, providing sensitivity control over the tolerance of line fitting. One natural measure of dissimilarity in lengths between two contour segments is their log ratio. We compute orientation difference between two segments by normalizing their absolute value of angle difference to π. In our experiments, we use a larger weighting factor for orientation to de-emphasize the difference in the lengths because they may be less accurate due to line fitting.

2.3.3 Learning the Shape Lexicon

We extract a large number of lexical words by sampling from training images, and construct an indexed shape lexicon by clustering and partitioning the lexical words. A lexicon provides a concise structural organization for associating large varieties of low-level features, and is efficient because it enables comparison to far fewer feature types.

2.3.3.1 Clustering Lexical Words

Prior to clustering, we compute the distance between each pair of lexical words and construct a weighted undirected graph $\mathcal{G} = (V, E)$, in which each node on the graph represents a word. The weight on an edge connecting two nodes C_a and C_b is defined as a function of their distance

$$w(C_a, C_b) = \exp\left(-\frac{d(C_a, C_b)^2}{\sigma_d^2}\right), \tag{2.2}$$

where we set parameter σ_d to 20 percent of the maximum distance among all pairs of nodes.

We pose feature clustering as a spectral graph partitioning problem, for which we seek to group the set of vertices V into disjoint sets $\{V_1, V_2, \ldots, V_K\}$, such that by the measure defined in (2.1) the dissimilarity among the vertices in a set is low, and that between different sets is high.

More concretely, let the $N \times N$ symmetric weight matrix for all the vertices be W, where $N = |V|$. We define the degree matrix D as an $N \times N$ diagonal matrix, and i-th element $d(i)$ along the diagonal satisfies $d(i) = \sum_j w(i,j)$. We use an $N \times K$ matrix X to represent a graph partition, *i.e.*, $X = [X_1, X_2, \ldots, X_K]$, where each element of matrix X is either 0 or 1. We can show that the feature clustering formulation that seeks globally balanced graph partitions is equivalent to the normalized cuts criterion [34], and can be written as

$$\text{maximize } \epsilon(X) = \frac{1}{K} \sum_{l=1}^{K} \frac{X_l^T W X_l}{X_l^T D X_l}, \tag{2.3}$$

$$\text{subject to } X \in \{0,1\}^{N \times K}, \text{ and } \sum_j X(i,j) = 1. \tag{2.4}$$

Minimizing normalized cuts exactly is NP-complete. We use a fast algorithm [35] for finding its discrete near-global optimum, which is robust to random initialization and converges faster than other clustering methods.

2.3.3.2 Organizing Features in the Lexicon

For each cluster, we select the feature instance closest to the cluster's center as the exemplary lexical word. This ensures that an exemplary word has the smallest sum of squared distance to the other features within the cluster. In addition, each exemplary word is associated with a cluster radius, which is defined as the maximum distance from the cluster center to all the other features within the cluster. The constructed shape lexicon $\mathcal{L}$ is composed of all exemplary lexical words.

Figure 2.4 shows the 25 most frequent exemplary lexical words for Arabic, Chinese, English, and Hindi, learned from 10 documents of each language. Distinguishing features between languages, including cursive style in Arabic, 45 and 90-degree transitions in Chinese, and various configurations due to long horizontal lines in Hindi, are learned automatically. Each row in Figure 2.4(e)-(h) lists exam-

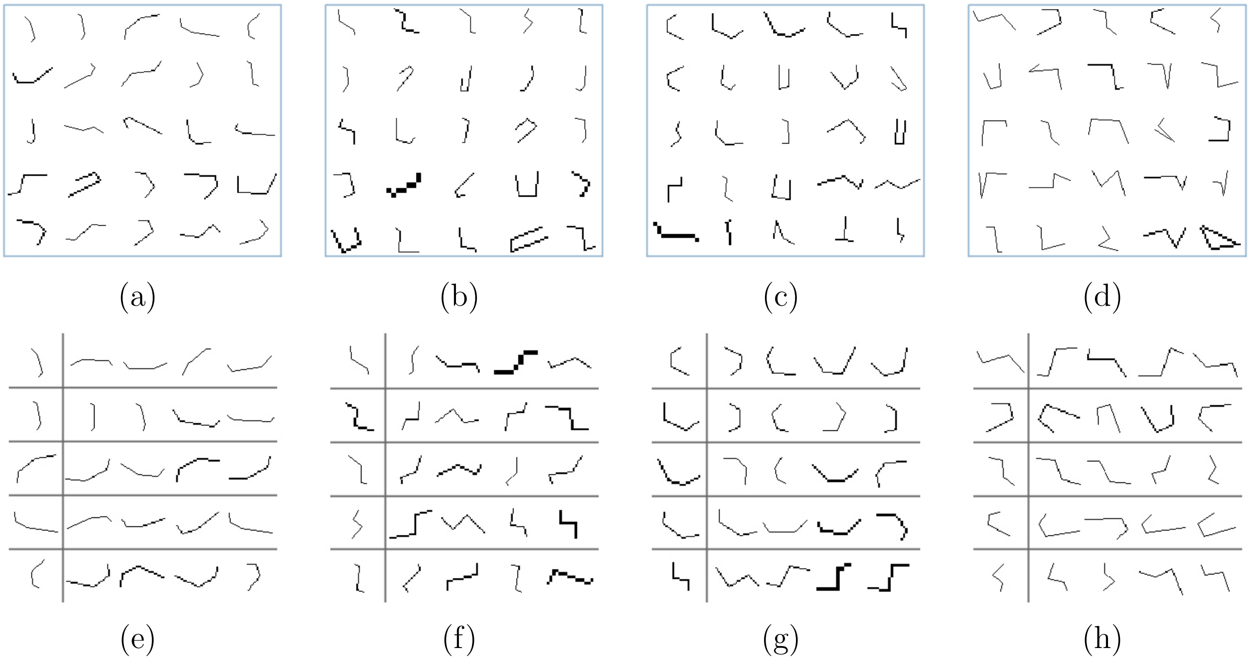


Figure 2.4: The 25 most frequent exemplary lexical words in (a) Arabic, (b) Chinese, (c) English, and (d) Hindi document images capture the distinct features of different languages. (e)-(h) show lexical words in the same cluster as the top 5 exemplary lexical words for each language, ordered by ascending distances to the center of their clusters. Scaled and rotated versions of feature types are clustered together.

ples of lexical words in the same cluster, ordered by ascending distances to the center of their associated clusters. By clustering, translated, scaled and rotated versions of feature types are grouped.

Since each lexical word represents a generic local shape feature, a majority of lexical words should intuitively appear in images across content categories, even though their frequencies of occurrence deviate significantly. In our experiments, we find that 95.1% and 92.6% of lexical words in natural images also appear in images with text and document images, respectively. In addition, 86.3% of lexical word instances appear in document images across all eight languages.

2.3.4 Constructing the Image Descriptor

We construct a shape descriptor for each image, which provides statistics of the frequency at which each feature type occurs. For each detected lexical word W from the test image, we compute the nearest exemplary word C_k in the shape lexicon. We increment the descriptor entry corresponding to C_k only if

$$d(W, C_k) < r_k, \tag{2.5}$$

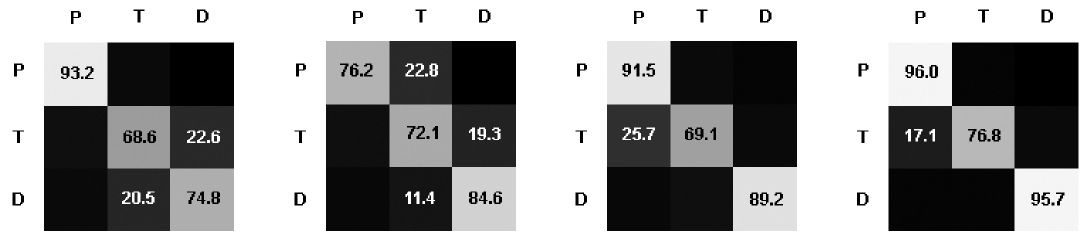
where r_k is the cluster radius associated with the exemplary lexicon word C_k. This quantization step ensures that unseen features that deviate considerably from training features are not used for image description. In our experiments, we found that only less than 2% of the contour features cannot be found in the shape lexicon learned from the training data.

2.4 Experimental Results

2.4.1 Image Content Category Recognition

2.4.1.1 Dataset

To evaluate our approach for image content category recognition, we construct a 4,500-image dataset by crawling web images from the Google Image search engine using a wide variety of keywords. Figure 2.1 shows some examples of images with text returned by using the text keyword "CD cover". All the images are automatically downloaded by a script. Duplicate and junk images are manually inspected and removed to reduce the proportion of unrelated images.



	P	T	D
P	93.2		
T		68.6	22.6
D		20.5	74.8

(a) Mean diagonal = 78.9%

	P	T	D
P	76.2	22.8	
T		72.1	19.3
D		11.4	84.6

(b) Mean diagonal = 77.6%

	P	T	D
P	91.5		
T	25.7	69.1	
D			89.2

(c) Mean diagonal = 83.3%

	P	T	D
P	96.0		
T	17.1	76.8	
D			95.7

(d) Mean diagonal = 89.6%

Figure 2.5: Confusion tables for image content category recognition using (a) Spatial envelope [15], (b) LBP [17], (c) kAS [29], (d) Our approach. (P: Pure image, T: Image with text, D: Document image)

2.4.1.2 Overview of the Experiments

We compare our approach with spatial envelope [15], local binary patterns (LBP) [17], and the state-of-the-art kAS descriptor [29], which are well-known approaches based on different views of whole-image characterization. Spatial envelope uses a holistic image representation without attempting to exploit localized information such as shape, whereas LBP is based on rotation-invariant texture analysis. Since 2AS gives the best performance among different kAS [29], we use it as the benchmark for kAS.

We train a multi-class SVM classifier [36] on only 100 randomly selected images from each category and use it to test the rest of the images in the collection. For easy comparison, we set the dimensions of the image descriptor to 90 for both kAS and our approach in the following experiments.

2.4.1.3 Results and Discussions

The confusion tables for spatial envelope, LBP, kAS, and our approach appear in Figure 2.5. Spatial envelope demonstrates performs well in recognizing pure images, but is not very effective for text content. Texture-based LBP produces balanced results for all three content types. Our approach obtains the best performances for recognizing each content class, with a respectable mean diagonal of 89.6%. The only notable confusion occurs when distinguishing between image with text and pure image.

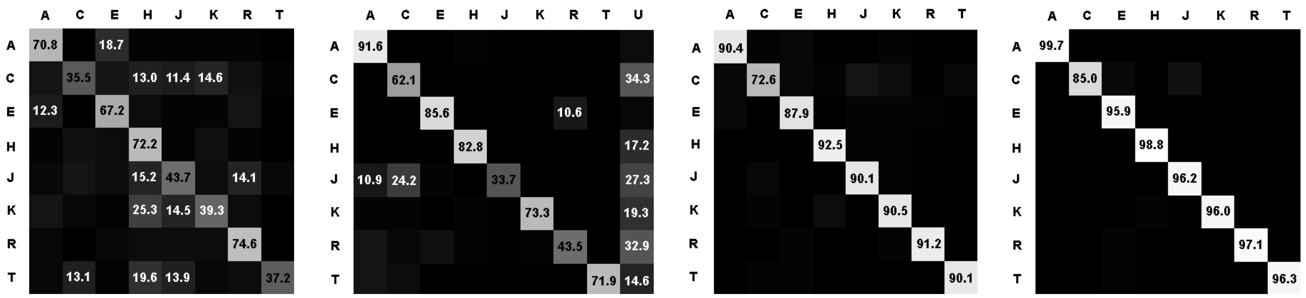
2.4.2 Language Identification

2.4.2.1 Dataset

We use 1,512 document images of eight languages (Arabic, Chinese, English, Hindi, Japanese, Korean, Russian, and Thai) from the University of Maryland multilingual database [11] and the IAM handwriting database DB3.0 [12] (see Figure 2.2) for evaluation on language identification. Both databases are large public real-world collections, containing the source identity of each image in the ground truth. This enables us to construct a diverse dataset that closely mirrors the true complexities of heterogeneous document image repositories in practice.

2.4.2.2 Overview of the Experiments

We compare our approach with the state-of-the-art language identification system [26], which is based on template matching. We also include LBP and kAS in this experiment since they have demonstrated reasonable performance on diverse text contents. For effective comparison, we used multi-class SVM classifiers trained on the same pool of randomly selected handwritten document images from each



(a) Mean diagonal = 55.1%

(b) Mean diagonal = 68.1%

(c) Mean diagonal = 88.2%

(d) Mean diagonal = 95.6%

Figure 2.6: Confusion tables for language identification using (a) LBP [17], (b) Template matching [26], (c) *k*AS [29], (d) Our approach. (A: Arabic, C: Chinese, E: English, H: Hindi, J: Japanese, K: Korean, R: Russian, T: Thai, U: Unknown)

language class in the following experiments, for LBP, *k*AS, and our approach, respectively. We further evaluated the generalization performance of our approach as the size of training data varied.

2.4.2.3 Results and Discussions

The confusion tables for LBP, template matching, *k*AS, and our approach appear in Figure 2.6. Our approach demonstrates excellent results on all the eight languages, with a mean diagonal of 95.6% and a standard deviation of 4.5%. Table 2.1 lists all entries in the confusion table of our approach for the eight languages. *k*AS, with a mean diagonal of 88.2%, is also effective. Neither *k*AS nor our approach has difficulty generalizing across large variations, such as font types or handwriting styles, as evident from their relatively small standard deviations along diagonal entries in the confusion tables shown in Figure 2.7.

The performance of template matching varies significantly across languages, with 68.1% mean diagonal and 20.5% standard deviation along diagonal. One big

Table 2.1: Confusion table of our approach for the 8 languages.

	A	C	E	H	J	K	R	T
A	99.7	0.3	0	0	0	0	0	0
C	1.4	85.0	4.0	1.0	6.7	1.0	0.7	0.2
E	1.6	0	95.9	0.2	0	1.1	0.6	0.6
H	0.2	0.2	0	98.8	0.8	0	0	0
J	0	1.3	1.0	0.2	96.2	1.3	0	0
K	0	0.8	0.1	1.9	0.5	96.0	0.5	0.1
R	0.5	0	2.0	0	0	0	97.1	0.4
T	0	0.3	1.6	0.9	0.6	0.3	0	96.3

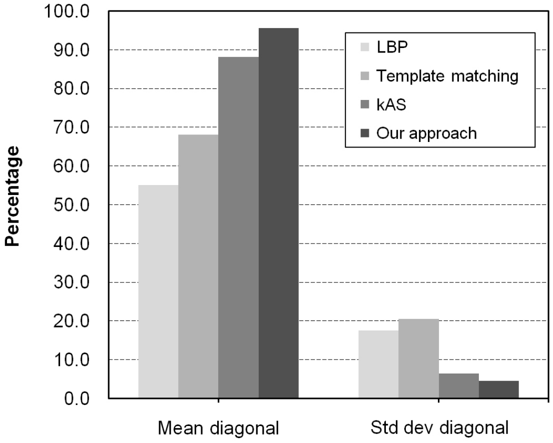
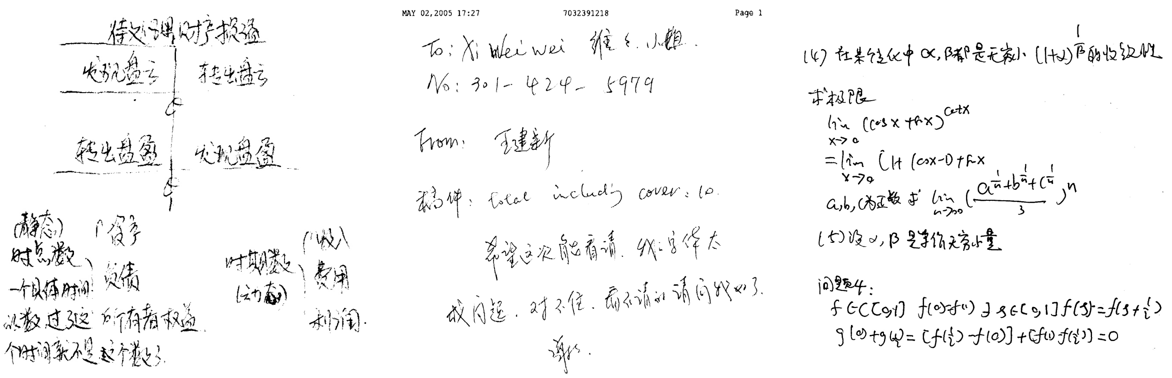


Figure 2.7: Comparison of language identification performances.

confusion of template matching occurs between Japanese and Chinese, given a document in Japanese may contain varying number of Kanji (Chinese characters). Templates are not flexible for identifying discriminative partial features, and the bias in voting decision toward the dominant candidate causes less frequently matched templates to be ignored. Another performance lowering source of error comes from undetermined cases (see the *unknown* column in Figure 2.6(b)), where probabilistic voting cannot decide between languages with roughly equal votes.

Texture-based LBP could not effectively recognize differences between languages on a diverse dataset because distinctive layouts and unconstrained hand-


MAY 02,2005 17:27 7032391218 Page 1

To: Xi Wei wei

No: 301-424-5979

From: 王建新

Figure 2.8: Examples of error cases in Chinese handwritten documents.

writing exhibit irregularities that are difficult to capture using texture, and its mean diagonal is only 55.1%.

Figure 2.7 quantitatively compares the overall language identification performances of different approaches by the means and standard derivations of diagonal entries in their confusion tables. Shape-based approaches, including our approach and *k*AS, show higher recognition rates and smaller performance derivations, compared to those based on models of textures and templates.

Analysis of the language identification errors made by our approach provides further insights. Among all eight languages, recognition of Chinese handwriting proved the most challenging task. As shown in Figure 2.7, the Chinese language identification performances for all four approaches were significantly lower compared to the other languages. We show several typical error cases in Figure 2.8, where Chinese handwritten documents are recognized incorrectly. These error cases include document images with severe degradations, second-generation documents captured from low-resolution source copies (*e.g.*, photocopy of a fax document), and documents containing a significant amount of other content.

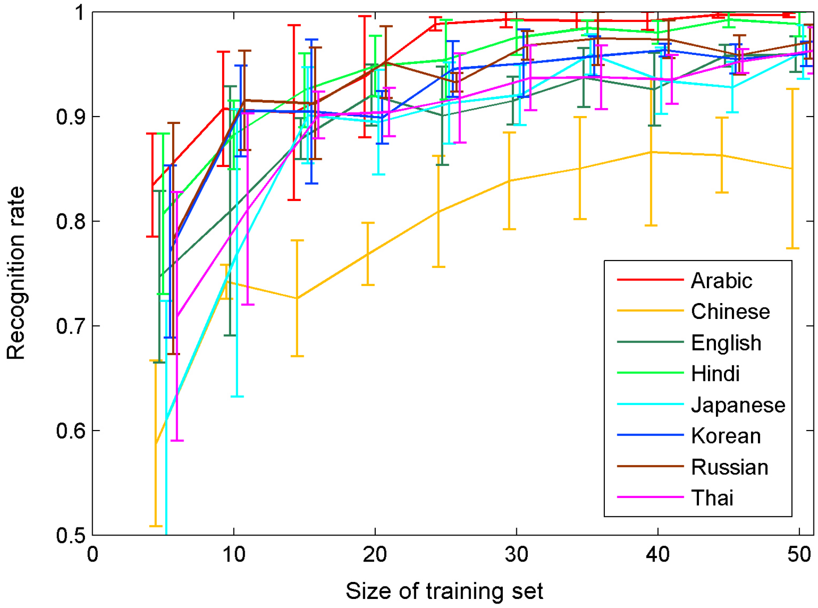


Figure 2.9: Recognition rates of our approach for different languages as the size of training data varies. Our approach achieves excellent performance even using a small number of document images per language for training.

Good generalization helps determine the success of document analysis systems that need to process diverse, unconstrained data. Figure 2.9 shows the recognition rates of our approach as the size of training set varies. We observe highly competitive language identification performance on this challenging dataset even when using a small amount of training data per language class. This demonstrates the effectiveness of generic low-level shape features when mid or high-level vision representations may not be generalized or flexible enough for the task.

Our results on language identification are encouraging, as the training requires considerably less supervision. Our approach needs only the class label of each training image, and does not require prior skew correction, scale normalization, or segmentation.

2.5 Summary

In this chapter, we proposed a novel approach for document image content categorization using a lexicon composed of a wide variety of local shape features. Each lexical word represents a characteristic structure generic enough to be detected repeatably and segmentation free. The lexicon provides a principled approach to structurally indexing and associating a vast number of feature types, and is learned from training data with little supervision. Our approach is extensible and does not require constructing explicit content models. In two challenging real world document image content recognition problems involving large-scale, highly variable image collections, our approach demonstrated excellent results and outperformed other state-of-the-art techniques. Our future work will be directed toward refining and evaluating the approach by further incorporating spatial co-occurrences of lexical words using a secondary lexicon.

Chapter 3

Signature Detection and Matching

3.1 Introduction

Searching for relevant documents from large complex document image repositories presents a central problem in document image analysis and retrieval. One approach recognizes text in the image using an OCR system, and applies text indexing and query. This solution is primarily restricted to machine printed text content because state-of-the-art handwriting recognition is error prone and limited to applications with a small vocabulary, such as postal address recognition and bank check reading [7]. In broader, unconstrained domains, including searching historic manuscripts [37] and processing languages where character recognition is difficult [38], image retrieval has demonstrated better results.

As unique and evidentiary entities in a wide range of business and forensic applications, signatures provide an important form of indexing that enables effective exploration of large heterogeneous document image collections. Given an abundance of documents, searching for a specific signature is a highly effective way of retrieving documents authorized or authored by an individual [39]. In this context, handwriting recognition is suboptimal, because of its prohibitively low recognition rates and the fact that the character sequence of a signature is often unrelated to the personal identity it represents. More importantly, as a number of studies [40, 41]

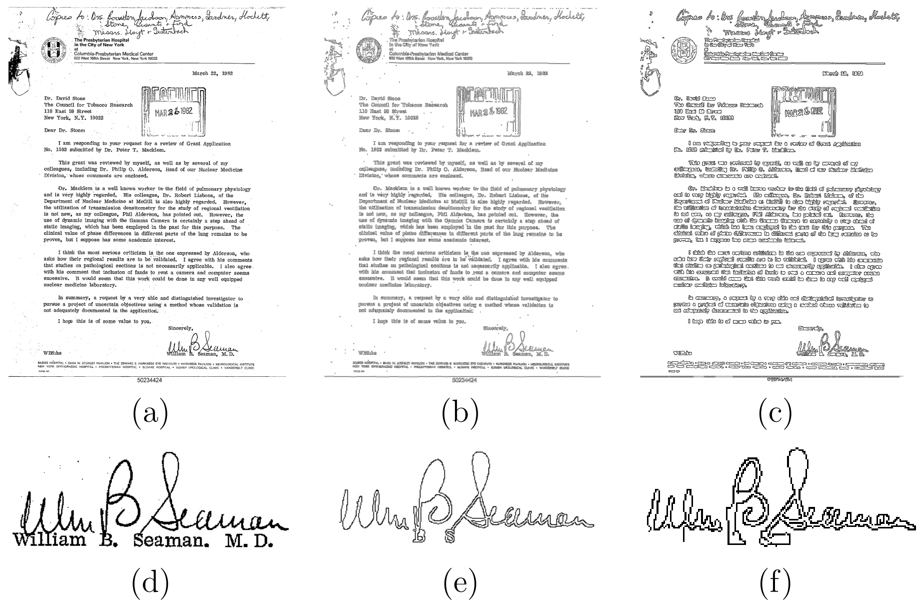


Figure 3.1: The prominence of a signature is perceived across scales whereas background text merges into blobs at coarser scales. (a) A document image example. (b) and (c) are edges computed at the scale of 1.0 and 2.0, respectively. (d) Detected signature region without segmentation. (e) and (f) are detected and segmented signatures by our approach at the two scales.

demonstrated, signatures are highly stylistic in nature and are best described by their graphic style.

Detecting, segmenting, and matching deformable objects such as signatures are important and challenging problems in computer vision. In the following subsections, we address detection, segmentation, and matching in the context of signature-based document image retrieval and present an overview of our approach.

3.1.1 Signature Detection and Segmentation

Detecting free-form objects is challenging in a number of aspects. First, detection needs to be robust in the presence of large intra-class variations and cluttered backgrounds. Second, the contours of these complex objects are fragmented 2-D signals in real off-line images, for which reliably recovering the order of points in handwriting is generally difficult [42]. Moreover, recognition and retrieval techniques require well segmented objects to minimize the effects of outliers during matching. Detecting signatures from documents offers one such example that further involves diverse layout structures and complex mixture of machine printed text, handwriting, diagrams, and other elements. Signature detection and segmentation remains an open research problem [43].

Prior research on off-line signatures has focused almost exclusively on signature verification and identification to perform authentication [44, 45, 40, 41, 46, 47, 48] in the context of biometrics. For signature verification, the problem lies in deciding whether a sample signature is genuine by comparing it with stored reference signatures. Signature identification is essentially a writer identification problem, which has an objective to find the signer of a sample given a database of signature exemplars from different signers. Most studies published to date assume the availability of good detection and segmentation [7].

The problem of signature detection and segmentation pivots on solving signature-based document indexing and retrieval. Equally important, a solution will benefit off-line signature verification and identification in a range of domains. In addition, the ability to detect signatures robustly and extract them intact from volumes of documents is highly desirable for many business and government applications.

We propose a new multi-scale approach to detect and extract signatures jointly from document images. Rather than focusing on local features, we treat signatures as symbols that exhibit characteristic structural saliency in our multi-scale detection framework. We employ a novel saliency measure based on a signature production model, which captures the dynamic curvature in a signature without recovering its temporal information. The signature detection approach can also be applied to on-line handwritten notes, where the trajectories of the pen are readily available.

Our detection framework is general and has the advantage of not embeding explicit assumptions on local features of signatures, such as the granulometric size distributions [44] or stroke-level features [40]. Therefore, it performs robustly against many forms of variations in shape-based object detection problems, and is generally applicable despite language differences.

3.1.2 Signature Matching for Document Image Retrieval

Detection and segmentation produce a set of 2-D contour fragments for each detected signature. Given a few available query signature instances and a large database of detected signatures, signature matching needs to find the most similar signature samples from the database. By constructing the list of best matching signatures, we effectively retrieve the set of documents authorized or authored by the same person.

We treat a signature as a non-rigid shape, and represent it by a discrete set of 2-D points sampled from the internal or external contours on the object. A 2-D point feature offers several competitive advantages compared to other compact geometrical entities used in shape representation. It relaxes the strong assumption that the topology and the temporal order need to be preserved under structural

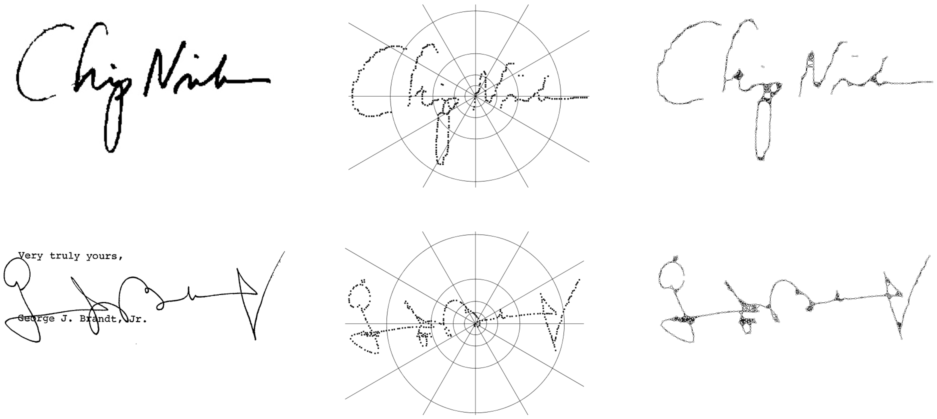


Figure 3.2: Shape contexts [55] and local neighborhood graphs [56] constructed from detected and segmented signatures. First column: Examples of signatures. Second column: Shape contexts descriptors constructed at a point, which provides a large-scale shape description. Third column: Local neighborhood graphs capture local structures for non-rigid shape matching.

variations or clustered background. For instance, two strokes in one signature sample may touch each other, but remain well separated in another. These structural changes, as well as outliers and noise, are generally challenging for shock-graph based approaches [49, 50], which explicitly use the connection between points. In earlier studies [51, 52, 53, 54], a shape is represented as an ordered sequence of points. This 1-D representation is well-suited for signatures collected online using a PDA or Tablet PC. For unconstrained off-line handwriting, however, it is generally difficult to recover their temporal information from real images due to large structural variations [42]. Represented by a 2-D point distribution, a shape is more robust under structural variations while carrying general shape information. As shown in Figure 3.2, the signature's shape is captured by a finite set $\mathcal{P} = \{P_1, \ldots, P_n\}, P_i \in \mathbb{R}^2$, of n points, which are sampled from edge pixels computed by an edge detector.

We use two state-of-the-art non-rigid shape matching algorithms for signature matching. The first method is based on the representation of shape contexts, introduced by Belongie *et al.* [55]. In this approach, a spatial histogram defined as *shape context* is computed for each point, which describes the distribution of the relative positions of all remaining points. Prior to matching, the correspondences between points are first solved through weighted bipartite graph matching. Our second method uses the non-rigid shape matching algorithm proposed by Zheng and Doermann [56], which formulates shape matching as an optimization problem that preserves local neighborhood structure. This approach contains an intuitive graph matching interpretation, where each point represents a vertex. Two vertices are considered connected in the graph if they are neighbors. So, finding the optimal match between shapes is equivalent to maximizing the number of matched edges between their corresponding graphs under a one-to-one matching constraint. Computationally, [56] employs an iterative framework for estimating the correspondences and the transformation. In each iteration, graph matching is initialized using the shape context distance, and subsequently is updated through relaxation labeling [57] for more globally consistent results.

Treating an input pattern as a generic 2-D point distribution broadens the space of dissimilarity metrics and enables effective shape discrimination using the correspondences and the underlying transformations [58]. We propose two novel shape dissimilarity metrics that quantitatively measure anisotropic scaling and registration residual error, and present a supervised training framework for effectively combining complementary shape information from different dissimilarity measures by linear discriminant analysis (LDA).

The contributions of this dissertation are twofold:

1. We propose a novel multi-scale approach to detect and segment signatures jointly from documents with diverse layouts and complex backgrounds. We treat a signature generally as an unknown grouping of 2-D contour fragments, and solve for the two unknowns — identification of the most salient structure and its grouping, using a signature production model that captures the dynamic curvature of 2-D contour fragments without recovering the temporal information.

2. We treat signature recognition and retrieval in the unconstrained setting of non-rigid shape matching, and propose two new measures of shape dissimilarity that correspond to i) the amount of anisotropic scaling, and ii) the registration residual error. We demonstrate robust signature-based document image retrieval, and comprehensively evaluate different shape representations, shape matching algorithms, measures of dissimilarity, and the use of multiple signature instances in overall retrieval accuracy.

We structure this chapter as follows: The next section reviews related work. In Section 3.3, we present our multi-scale structural saliency approach to signature detection and segmentation in detail. Section 3.4 introduces a structural saliency measure for capturing the dynamic curvature under a signature production model. In Section 3.5, we describe our signature matching approach and present methods to combine different measures of shape dissimilarity and multiple query instances for effective retrieval with limited supervised training. We discuss experimental results on real English and Arabic document datasets in Section 3.6, and conclude in Section 3.7.

3.2 Related work

3.2.1 Structural Saliency and Contour Grouping

Detecting salient structures from images is an important task in many segmentation and recognition problems. The *saliency network* proposed by Sha'ashua and Ullman [59] offers a well-known approach to extracting salient curves by jointly solving the two aspects of this problem iteratively, *i.e.*, identifying salient structures and grouping contours. The saliency function defined in [59] monotonically increases with the length and decreases with the total squared curvature of the evaluated curve. To reduce the exponential search space, the saliency network assumed that an optimal solution has a recurrent structure, which they called *extensibility*, so that searching by dynamic programming in the exponential space of possible curves takes polynomial time. However, greedily reducing the solution space by such a recurrent formulation involves hard decisions at each step, and theoretically, a single mistake can result in the convergence to a wrong solution.

Alter and Basri [60] presented a comprehensive analysis of the saliency network and derived its $O(k^2N^2)$ time complexity, where N is the total number of pixels and k is the number of neighboring elements considered in forming a locally connected network. They demonstrated that the salient network has a few serious problems due to the extensibility assumption, and the convergence rates vary significantly depending on the object structure. These limitations are difficult to overcome without fundamentally changing the computation.

A body of literature exists on contour grouping and contour-based learning in computer vision. Here we highlight the work more closely related to ours, which includes Parent and Zucker's [61] work using relaxation methods, and Guy and Medioni's [62] work using voting patterns. Elder and Zucker [63] developed a method for finding closed contours using chains of tangent vectors. Williams and Jacobs [64] and Williams and Thornber [65] discussed contour closure using stochastic completion fields. Shotton *et al.* [66] demonstrated a learning-based method for object detection using local contour-based features extracted from a single image scale. Such an approach, however, is inherently restricted to closed contour shapes, which have an explicit ordering of points.

3.2.2 Shape Matching

Rigid shape matching has been approached in a number of ways with intent to obtain a discriminative global shape description. Approaches using silhouette features include Fourier descriptors [67, 68], geometric hashing [69], dynamic programming [70, 52], and skeletons derived using Blum's medial axis transform [71]. Although silhouettes are simple and efficient to compare, they have limits as shape descriptors because they ignore internal contours and are difficult to extract from real images [72]. Other approaches, such as chamfer matching [73] and the Hausdorff distance [74], treat the shape as a discrete set of points extracted using an edge detector. Unlike approaches that compute correspondences, these methods do not enforce pairing of points between the two sets. While they work well under selected subset of rigid transformations, they cannot be extended to handle non-rigid transformations. [75, 76] present a general survey on classic rigid shape matching techniques.

Matching for non-rigid shapes needs to consider unknown transformations that are both linear (*e.g.*, translation, rotation, scaling, and shear) and non-linear. One comprehensive framework for shape matching in this setting is to estimate iteratively the correspondence and the transformation. The iterative closest point (ICP) algorithm introduced by Besl and McKay [77] and its extensions [78, 79, 80] provide a simple heuristic approach. Assuming two shapes align roughly, the nearest-neighbor in the other shape is assigned as the estimated correspondence at each step. This estimate of the correspondence is then used to refine the estimated affine or piece-wise-affine mapping, and vice versa. While ICP is fast and guaranteed to converge to a local minimum, its performance degenerates quickly in the presence of large non-rigid deformation or a significant amount of outliers [81]. Chui and Rangarajan [82] developed an iterative optimization algorithm to estimate the correspondences and the transformation jointly, using thin plate splines as a generic parameterization of a non-rigid transformation. Joint estimation of correspondences and transformation leads to highly non-convex optimization, which is solved using the softassign and deterministic annealing.

3.2.3 Document Image Retrieval

Rath *et al*. [83] demonstrated retrieval of handwritten historical manuscripts by using images of handwritten words to query un-labeled document images. The system compares word images based on Fourier descriptors computed from a collection of shape features, including the projection profile and the contours extracted from the segmented word. Srihari *et al*. [84] developed a signature matching and retrieval approach by computing correlation of gradient, structural, and concavity features extracted from fixed-size image patches. It achieved 76.3% precision using a collection of 447 manually cropped signature images from the Tobacco-800 database [85, 86]. These approaches are not translation, scale or rotation invariant.

3.3 Multi-scale structural saliency

3.3.1 Theoretical framework

We consider the identification of salient structure and the grouping of its structural components separately. There are clear motivations for decoupling these two unknowns, as opposed to solving them jointly. First, we have a broader set of functions to use as measures of saliency. For object detection, saliency measures that fit high-level knowledge of the object give more globally consistent results than jointly optimizing a fixed set of low-level vision constraints. Having identified the salient structures, the problem of grouping becomes simpler based on constraints such as proximity and good continuation. Second, we can effectively formulate structural saliency across image scales, as opposed to single-scale approaches such as the saliency network. Multi-scale detection is important for non-rigid objects like signatures, whose contours can be severely broken due to poor ink condition and image degradations. Last, multi-scale saliency computation generates detection hypotheses at the natural scale where grouping among a set of connected components becomes structurally obvious. These provide a unified framework for object detection and segmentation that produces meaningful representation for object recognition and retrieval.

From a computational point of view, computing saliency using connected components makes the computation tractable and highly parallelizable. Our serial implementation runs in $O(N)$, where N is the total number of edge points. This is significantly faster than the saliency network approach that has $O(k^2N^2)$ time complexity. We also explore document context to improve detection. The idea is to estimate the length and inter-line spacing of text lines and use the information to locate the bottom or end of a document, where signatures are more likely to appear. In our evaluation, we show results of signature detection on whole documents, as well as by exploration of document context.

3.3.2 Signature detection and segmentation

In this section, we describe a structural saliency approach to signature detection by searching a range of scales $\mathcal{S} = \sigma_1, \sigma_2, \cdots, \sigma_n$. We select the initial scale σ_1 based on the resolution of the input image. We define the multi-scale structural saliency for a curve Γ as

$$\Phi(\Gamma) = \max_{\sigma_i \in \mathcal{S}} f(\Phi_{\sigma_i}(\Gamma_{\sigma_i}), \sigma_i), \tag{3.1}$$

where $f\colon \mathbb{R}^2 \rightarrow \mathbb{R}$ is a function that normalizes the saliency over its scale, and Γ_{σ_i} is the corresponding curve computed at the scale σ_i. Using multiple scales for detection relaxes the requirement that the curve Γ be well connected at a specific scale.

Detection at a particular scale σ_i proceeds in three steps. First, we convolve the image with a Gaussian kernel G_{σ_i}, re-sample it using the Lanczos filter [87] at the factor d_{σ_i}, and compute its edges using the Canny edge detector [33]. This effectively computes a coarse representation of the original image in which small gaps in the curve are bridged by smoothing followed by re-sampling (see Figure 3.1 on page 31). Second, we form connected components on the edge image at the scale σ_i, and compute the saliency of each component using the measure presented in Section 3.4, which characterizes its dynamic curvature. We define the saliency of a connected component $\Gamma_{\sigma_i}^k$ as the sum of saliency values computed from all its pairs of edges. Third, we identify the most salient curves and use a grouping strategy based on proximity and curvilinear constraints to obtain the rest of the signature parts within their neighborhood. Our joint detection and segmentation approach considers identifying the most cursive structure and grouping it with neighboring elements in two steps. By locating the most salient signature component, we effectively focus our attention on its neighborhood. Subsequently, a complete signature is segmented from background by grouping salient neighboring structures.

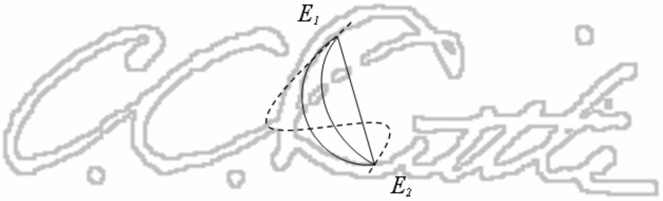


Figure 3.3: Among the large number of geometric curves passing the two random end points E_1 and E_2 on a signature, few are realistic (solid curves). An example of unrealistic curve is shown in dotted line.

3.4 Measure of saliency for signatures

In this section, we consider the problem of recognizing the structural saliency of a 2-D off-line signature segment using a signature production model. As shown in Figure 3.3, among the infinite number of geometric curves that pass two given end points E_1 and E_2 on a signature, minimal are realistic. The author's wrist is highly constrained in the degrees of freedom while signing the document. Furthermore, a signature segment rarely fits locally to a high-order polynomial, as shown by the dotted curve in Figure 3.3.

We propose a signature production model that incorporates two degrees of freedom in the Cartesian coordinates. We assume that the pen moves in a cycloidal fashion with reference to a sequence of shifting baselines when signing. The local baseline changes as the author's wrist adjusts its position with respect to the document. Within a short curve segment, we assume that the baseline remains unchanged. In addition, the locus of the pen maintains a proportional distance from the local center point (*focus*) to the local baseline (*directrix*), which imposes an additional constraint that limits the group of second-order curves to ellipses. A signature fragment thus can be equivalently viewed as concatenations of small elliptic

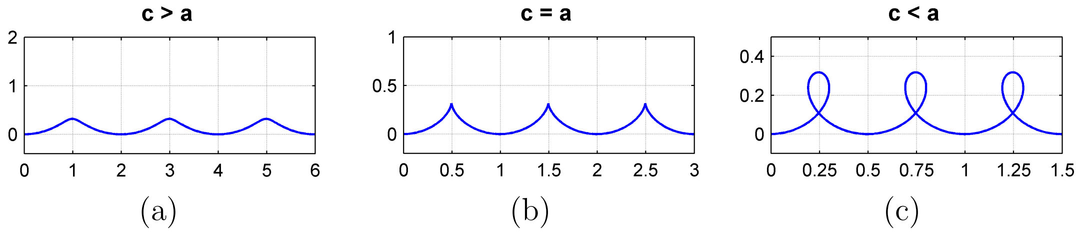


Figure 3.4: (a) Curtate cycloid, (b) cycloid, and (c) prolate cycloid curves generated when the speeds of writing in the horizontal baseline and the vertical direction both vary sinusoidally.

segments. In a similar spirit, Saint-Marc *et al*. [88] have used piece-wise quadratic B-splines to approximate complex-shaped contours for symmetry detection.

Our hypothesis on cycloidal writing is motivated by Hollerbach's [89] oscillation theory of handwriting, who discovered that embedded oscillations coupled along the horizontal and vertical directions produce letter forms that closely resemble handwriting samples. In fact, when the signature baseline aligns to the x axis, velocities in the x and y directions are

$$\begin{aligned} v_x &= a\,\sin(\omega_x t + \phi_x) + c \\ v_y &= b\,\sin(\omega_y t + \phi_y) \end{aligned} \tag{3.2}$$

where a and b are horizontal and vertical velocity amplitudes, ω_x and ω_y are the horizontal and vertical frequencies, ϕ_x and ϕ_y are the horizontal and vertical phases, t is the time variable, and constant c is the average speed of horizontal movement. Without loss of generality, consider the case of $a = b$, $\omega_x = \omega_y$ and $\phi_x - \phi_y = \pi/2$. We can show that for different values of a and c, the resulting curves are curtate cycloid, cycloid, and prolate cycloid, as shown in Figure 3.4, respectively.

We model piecewise segments of a signature by a family of second-order curves that satisfy constraints imposed by the signature production. In addition, we use the local gradient directions approximated at the two end points, which can be viewed as soft constraints on the segment of the curve imposed by the global structure of the signature instance. In Cartesian coordinates, the family of a quadratic equation in two variables x and y is always a conic section. Figure 3.5 demonstrates that the directionst of the gradients at the two edge points greatly limit the inference on local curve segment to a family of conics, under the second-order signature production model.

We can formalize this intuition geometrically. For a pair of edge points E_1 at (x_1, y_1) and E_2 at (x_2, y_2), we obtain estimates of their local gradients $N_1(p_1, q_1)$ and $N_2(p_2, q_2)$ through edge detection. For definiteness, we suppose both E_1 and E_2 point into the angular section between the tangent lines, as shown in Figure 3.5.

$$\begin{aligned} p_1(x_2 - x_1) + q_1(y_2 - y_1) &> 0 \quad \text{and} && (3.3a) \\ p_2(x_2 - x_1) + q_2(y_2 - y_1) &< 0 && (3.3b) \end{aligned}$$

The two tangent lines at E_1 and E_2 are normal to their local gradients and are given by

$$t_1(x, y) \equiv p_1(x - x_1) + q_1(y - y_1) = 0 \tag{3.4}$$

and

$$t_2(x, y) \equiv p_2(x - x_2) + q_2(y - y_2) = 0. \tag{3.5}$$

The straight line $l(x, y)$ that passes through E_1 and E_2 can be written as

$$l(x, y) \equiv \begin{vmatrix} x & y & 1 \\ x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \end{vmatrix} = 0. \tag{3.6}$$

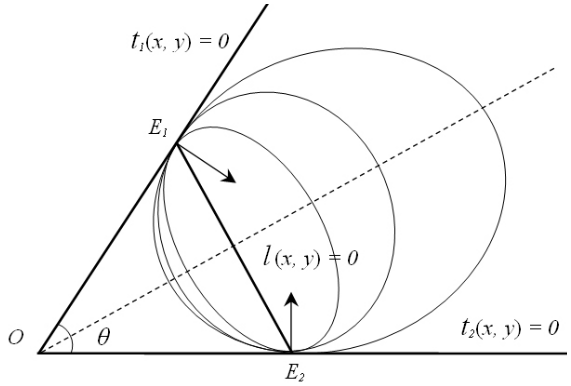


Figure 3.5: Conic sections inferred by a pair of edge points.

Note that $t_1(x, y)$, $t_2(x, y)$ and $l(x, y)$ are all first-order linear functions in x and y. The family of second-order curves bounded within the angular section between $t_1(x, y)$ and $t_2(x, y)$ can be expressed as

$$C(x, y) \equiv l^2(x, y) - \lambda t_1(x, y)t_2(x, y) = 0. \tag{3.7}$$

In the canonical form, that is

$$ax^2 + 2hxy + by^2 + 2gx + 2fy + c = 0, \tag{3.8}$$

where parameters a, b, c, f, g, h are first-order linear functions in λ, and the parameter set $(x_1,\ y_1)$, $(x_2,\ y_2)$, $(p_1,\ q_1)$, $(p_2,\ q_2)$.

Given a parameter set $(x_1,\ y_1)$, $(x_2,\ y_2)$, $(p_1,\ q_1)$, $(p_2,\ q_2)$, which is equivalent to fixing the set of three straight lines $t_1(x, y)$, $t_2(x, y)$ and $l(x, y)$, we can theoretically analyze how parameter λ for $\lambda \in [0, +\infty)$ affects the curvature of the bounded quadratic curve defined by (3.7). When $\lambda = 0$, Equation (3.7) degenerates into the straight line $l(x, y)$, and the total squared curvature of the bounded segment is strictly zero. When λ monotonically increases from 0 within certain range ($0 < \lambda < \lambda_0$), the curve segment bounded by E_1 and E_2 exhibits increasingly more curvature. This happens because the second-order curves governed by (3.7) for $\lambda \in [0, \lambda_0)$ are ellipses with monotonically increasing eccentricities. As $\lambda \rightarrow \lambda_0$, the center of the ellipse recedes to infinity, so that the ellipse tends to a parabola at $\lambda = \lambda_0$. When $\lambda > \lambda_0$, the conic of (3.7) becomes a hyperbola. Eventually as $\lambda \rightarrow +\infty$, the hyperbola degenerates into the two intersected straight lines $t_1(x, y)$ and $t_2(x, y)$. We prove in Appendix A that λ_0 is given by

$$\lambda_0 = \frac{4[p_1(x_2 - x_1) + q_1(y_2 - y_1)]}{(p_1 q_2 - p_2 q_1)} \times \frac{[p_2(x_1 - x_2) + q_2(y_1 - y_2)]}{(p_1 q_2 - p_2 q_1)}. \tag{3.9}$$

Equation (3.9) provides a theoretical second-order upper bound of the dynamic curvature, given the parameter set $(x_1,\ y_1)$, $(x_2,\ y_2)$, $(p_1,\ q_1)$, $(p_2,\ q_2)$ that fits the signature production model. We use λ_0 as the saliency value $\Lambda_{\sigma_i}(E_i, E_j)$ for a pair of points at scale σ_i. When $a = |OE_1| = |OE_2|$, it is straightforward to show that the right hand side of (3.9) is $4a^2$. This result allows us to normalize saliency over scale, whereas the scale interpretation of most published saliency measures surveyed in [65] are largely unclear. Obviously, our saliency measure is translation and rotation invariant as it only uses local gradient directions.

The saliency of a curve $\Gamma_{\sigma_i}^k$ at scale σ_i is defined as the sum of saliency values computed from all pairs of points on it and is written as

$$\Phi_{\sigma_i}(\Gamma_{\sigma_i}^k) = \sum_{E_i, E_j \in \Gamma_{\sigma_i}^k} \Lambda_{\sigma_i}(E_i, E_j). \tag{3.10}$$

It measures the likelihood of elliptic segment fitting given a set of 2-D points, and the computation does not require the temporal order among points.

The analysis so far applies only to the continuous case. To account for the discretization effect, we impose two additional conditions. First, the absolute values of the two functions on the left hand side of Equation (3.3a-b) must be strictly larger than ϵ. Second, the denominator term in (3.9) must be strictly large than ϵ. In our experiments, we use $\epsilon = 0.1$. For robustness, we weight the saliency contribution by the gradient magnitude of the weaker edge.

Separating saliency detection from grouping significantly reduces the level of complexity. Let the total number of edge points be N and the average length of connected components be L_c. Saliency computation for each component in Equation (3.10) requires $O(L_c^2)$ time on average. Therefore, the overall computation is of order $(N/L_c) \times L_c^2 = NL_c$. Since L_c is effectively upper bounded by prior estimate of signature dimensions and the range of searched scales n is limited, they can be considered as constants. The complexity in saliency computation is linear in N. Gaussian smoothing and connected component analysis both require $O(N)$ time. The total complexity in the signature detection algorithm is therefore $O(N)$.

3.5 Matching and Retrieval

3.5.1 Measures of Shape Dissimilarity

Before we introduce two new measures of dissimilarity for general shape matching and retrieval, we first discuss existing shape distance metrics. Each of these dissimilarity measures captures certain shape information for effective discrimination. In the next sub-section, we describe how to combine these individual measures effectively with limited supervised training, and present our evaluation framework.

Several measures of shape dissimilarity have demonstrated success in object recognition and retrieval, including the thin-plate spline bending energy D_{be} and the shape context distance D_{sc}. As a conventional tool for interpolating coordinate mappings from $\mathbb{R}^2$ to $\mathbb{R}^2$ based on point constraints, the thin-plate spline (TPS) is commonly used as a representation of non-rigid transformation [90]. The TPS interpolant $f(x, y)$ minimizes the bending energy

$$\iint_{\mathbb{R}^2} \left[(\frac{\partial^2 f}{\partial x^2})^2 + 2(\frac{\partial^2 f}{\partial x \partial y})^2 + (\frac{\partial^2 f}{\partial y^2})^2 \right] dx\, dy \tag{3.11}$$

over the class of functions that satisfy the given point constraints. Equation (3.11) imposes smoothness constraints to discourage non-rigidities that are too arbitrary. The bending energy D_{be} [82] measures the amount of *non-linear* deformation to warp the shapes into alignment and provides physical interpretation. However, D_{be} measures only the deformation beyond an affine transformation, and its functional in (3.11) is zero if the undergoing transformation is purely affine.

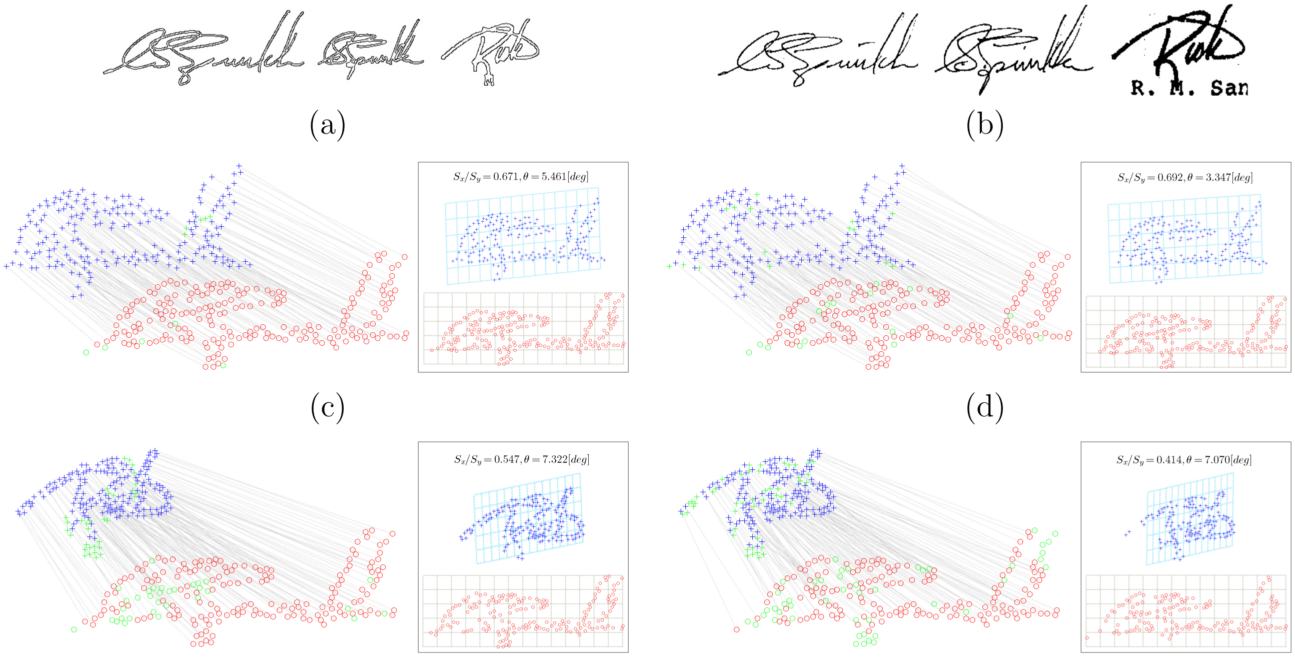
The shape context distance D_{sc} between a template shape $\mathcal{T}$ composed of m points and a deformed shape $\mathcal{D}$ of n points is defined in [55] as

$$D_{sc}(\mathcal{T}, \mathcal{D}) = \frac{1}{m} \sum_{t \in \mathcal{T}} \arg \min_{d \in \mathcal{D}} C(T(t), d) + \frac{1}{n} \sum_{d \in \mathcal{D}} \arg \min_{t \in \mathcal{T}} C(T(t), d), \tag{3.12}$$

where $T(.)$ denotes the estimated TPS transformation and $C(., .)$ is the cost function for assigning correspondence between any two points. Given two points, t in shape $\mathcal{T}$ and d in shape $\mathcal{D}$, with associated shape contexts $h_t(k)$ and $h_d(k)$, for $k = 1, 2, \ldots, K$, respectively, $C(t, d)$ is defined by χ^2 statistic as

$$C(t, d) \equiv \frac{1}{2} \sum_{k=1}^{K} \frac{[h_t(k) - h_d(k)]^2}{h_t(k) + h_d(k)}. \tag{3.13}$$

We introduce a new measure of dissimilarity D_{as} that characterizes the amount of anisotropic scaling between shapes. Anisotropic scaling is a form of affine trans-


R. M. San
$S_x/S_y = 0.671, \theta = 5.461[deg]$
$S_x/S_y = 0.692, \theta = 3.347[deg]$
$S_x/S_y = 0.547, \theta = 7.322[deg]$
$S_x/S_y = 0.414, \theta = 7.070[deg]$

(a) (b)

(c) (d)

(e) (f)

Figure 3.6: Anisotropic scaling and registration quality effectively capture shape differences. (a) Signature regions without segmentation. The first two signatures are from the same person, whereas the third is from someone else. (b) Detected and segmented signatures by our approach. Second row: matching results of first two signatures using (c) shape contexts and (d) local neighborhood graph, respectively. Last row: matching results of first and third signatures using (e) shape contexts and (f) local neighborhood graph, respectively. Corresponding points identified by shape matching are linked and unmatched points are shown in green. The computed affine maps are shown in the figure legends.

formation that involves change to the relative directional scaling. As shown in Figure 3.6, the stretching or squeezing of the scaling in the computed affine map effectively captures global mismatch in shape dimensions among registered points, even in the presence of large intra-class variation.

We compute the anisotropic scaling between two shapes by estimating the ratio of the two scaling factors S_x and S_y in the x and y directions, respectively. A TPS transformation can decompose into a linear part corresponding to a global affine alignment, together with the superposition of independent, affine-free deformations (or principal warps) of progressively smaller scales [90]. We ignore the non-affine terms in the TPS interpolant when estimating S_x and S_y. The 2-D affine transformation is represented as a 2×2 linear transformation matrix $\mathbf{A}$ and a 2×1 translation vector $\mathbf{T}$

$$\begin{pmatrix} u \\ v \end{pmatrix} = \mathbf{A} \begin{pmatrix} x \\ y \end{pmatrix} + \mathbf{T}, \tag{3.14}$$

where we can compute S_x and S_y by singular value decomposition on matrix $\mathbf{A}$.

We define D_{as} as

$$D_{as} = \log \frac{\max(S_x, S_y)}{\min(S_x, S_y)}. \tag{3.15}$$

Note that we have $D_{as} = 0$ when merely isotropic scaling is involved (*i.e.*, $S_x = S_y$).

We propose another distance measure D_{re} based on the registration residual errors under the estimated non-rigid transformation. To minimize the effect of outliers, we compute the registration residual error from the subset of points that has been assigned correspondence during matching, and ignore points matched to the dummy point *nil*. Let function $M\colon \mathbb{Z}^+ \to \mathbb{Z}^+$ define the matching between two point sets of size n representing the template shape $\mathcal{T}$ and the deformed shape $\mathcal{D}$. Suppose t_i and $d_{M(i)}$ for $i = 1, 2, \ldots, n$ denote pairs of matched points in shape $\mathcal{T}$ and shape $\mathcal{D}$, respectively. We define D_{re} as

$$D_{re} = \frac{\sum_{i:M(i) \neq nil} ||T(t_i) - d_{M(i)}||}{\sum_{i:M(i) \neq nil} 1}, \tag{3.16}$$

where $T(.)$ is the estimated TPS transformation and $||.||$ is the Euclidean norm.

3.5.2 Shape Distance

After matching, we compute the overall shape distance for retrieval as the weighted sum of individual distances given by all the measures: shape context distance, TPS bending energy, anisotropic scaling, and registration residual errors.

$$D = w_{sc}D_{sc} + w_{be}D_{be} + w_{as}D_{as} + w_{re}D_{re}. \tag{3.17}$$

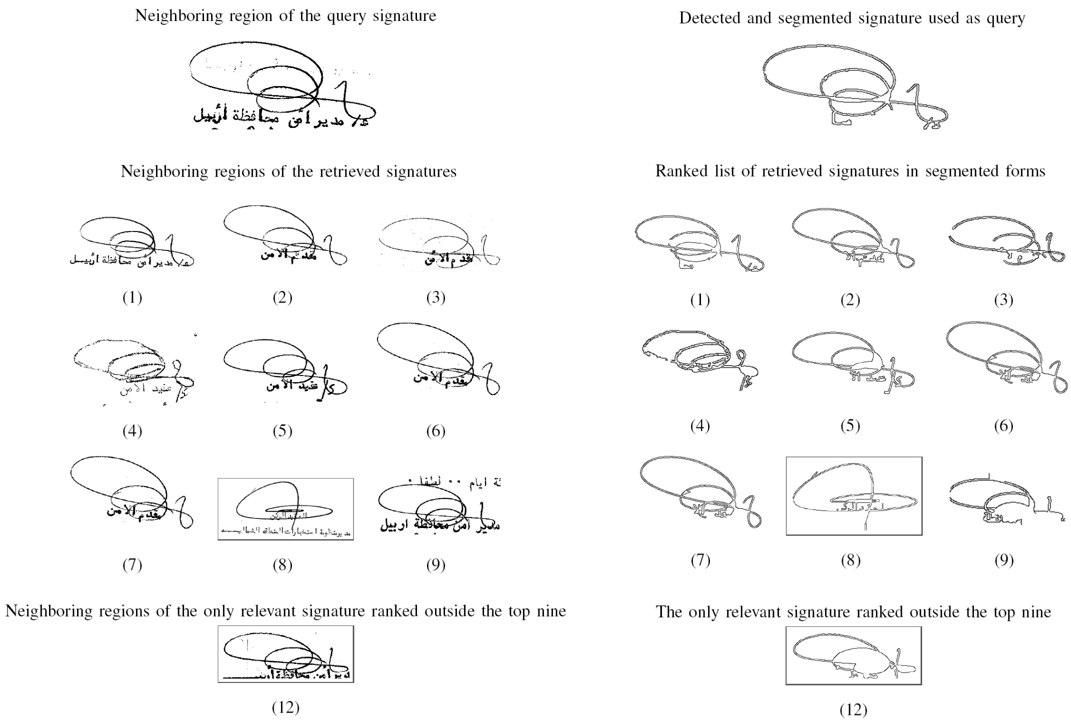
The weights in (3.17) are optimized by LDA, using only a small amount of training data.

The retrieval performance of a single query instance may depend largely on the instance used for the query [91]. In practice, it is often possible to obtain multiple signature samples from the same person. So, we can use them as an equivalence class to achieve better retrieval performance. When multiple instances $q_1, q_2, \ldots, q_k$ from the same class $\mathcal{Q}$ are used as queries, we combine their individual distances $D_1, D_2, \ldots, D_k$ into one shape distance as

$$D = \min(D_1, D_2, \ldots, D_k). \tag{3.18}$$

3.5.3 Evaluation Methodology

We use two commonly cited measures, average precision and R-precision, to evaluate the performance of each ranked retrieval. First, we make precise the intuitions of these evaluation metrics, which emphasize the retrieval ranking differently. Given a ranked list of documents returned in response to a query, average precision (AP) is defined as the average of the precisions at each relevant hit. It rewards retrieval systems that rank relevant documents higher and at the same time penalizes


Neighboring region of the query signature

Detected and segmented signature used as query

Neighboring regions of the retrieved signatures

Ranked list of retrieved signatures in segmented forms

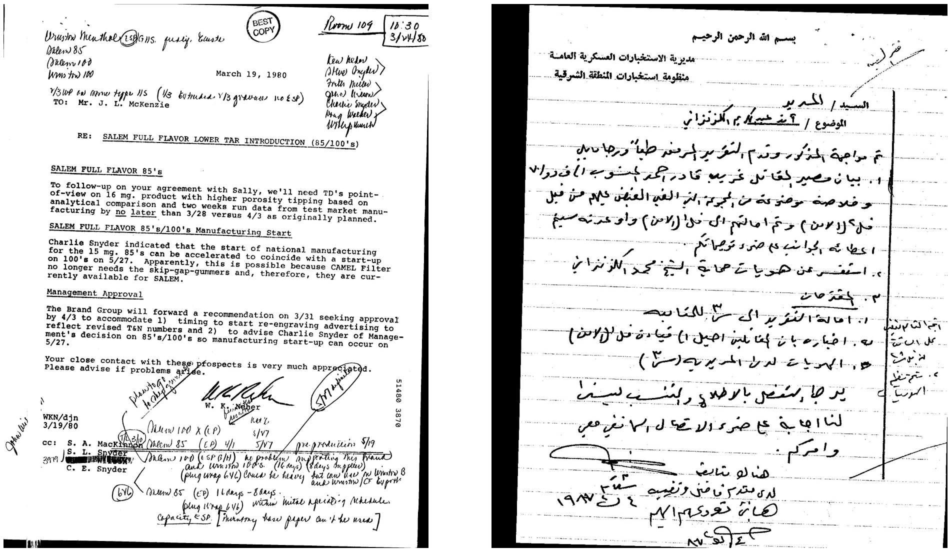
Neighboring regions of the only relevant signature ranked outside the top nine

The only relevant signature ranked outside the top nine

Figure 3.7: A signature query example. Among the total of nine relevant signatures, eight appear in the top nine of the returned ranked list, giving average precision of 96.0%, and R-precision of 88.9%. The irrelevant signature is highlighted with a bounding box. Left: signature regions in the document. Right: detected and segmented signatures used in retrieval.

those that rank irrelevant ones higher. AP effectively combines the precision, recall, and relevance ranking, and is considered a stable and discriminating measure of the quality of retrieval engines [91]. R-precision (RP) for a query i is the precision at the rank $R(i)$, where $R(i)$ is the number of documents relevant to query i. It de-emphasizes the exact ranking among the retrieved relevant documents and is more useful given a large number of relevant documents.

Figure 3.7 shows a query example, in which eight of the nine total relevant signatures are among the top nine and one relevant signature is ranked 12. For this query, $AP = (1{+}1{+}1{+}1{+}1{+}1{+}1{+}8/9{+}9/12)/9 = 96.0\%$, and $RP = 8/9 = 88.9\%$.



Winston Menthol (ESP) GHS. purify. Taste
Salem 85
Salem 100
Winston 100

BEST COPY

Room 109 | 10:30 3/24/80

Ken Heller
Steve Snyder
Fritz Miller
John Weiss
Charlie Snyder
Doug Walker
W.H. Upchurch

March 19, 1980

1/3 WP on some type HS (1/3 extruded 1/3 gravure no ESP)

TO: Mr. J. L. McKenzie

RE: SALEM FULL FLAVOR LOWER TAR INTRODUCTION (85/100's)

SALEM FULL FLAVOR 85's

To follow-up on your agreement with Sally, we'll need TD's point-of-view on 16 mg. product with higher porosity tipping based on analytical comparison and two weeks run data from test market manufacturing by no later than 3/28 versus 4/3 as originally planned.

SALEM FULL FLAVOR 85's/100's Manufacturing Start

Charlie Snyder indicated that the start of national manufacturing for the 15 mg. 85's can be accelerated to coincide with a start-up on 100's on 5/27. Apparently, this is possible because CAMEL Filter no longer needs the skip-gap-gummers and, therefore, they are currently available for SALEM.

Management Approval

The Brand Group will forward a recommendation on 3/31 seeking approval by 4/3 to accommodate 1) timing to start re-engraving advertising to reflect revised T&N numbers and 2) to advise Charlie Snyder of Management's decision on 85's/100's so manufacturing start-up can occur on 5/27.

Your close contact with these prospects is very much appreciated. Please advise if problems arise.

plan to go to skip gummers

W. K. Neher

5/27 definite

Final test results

WKN/djn
3/19/80

John Weiss

cc: S. A. MacKinnon
S. L. Snyder
3979 [redacted]
C. E. Snyder

Salem 100 X (EP) Rec'd 5/27
T/L 3/10 Salem 85 (EP) 4/1 5/27 pre-production 5/19
Salem 100 (ESP G/H) no problem supporting this brand and Winston 100's. (16 days) (8 days supplied)
(plug wrap 646) Could be heavy but can use on Winston B and Winston/CF export
646 Salem 85 (EP) 16 days - 8 days.
(plug wrap 646) within initial operating schedule.
Capacity ESP. [Inventory base paper can't be used]

51480 3870

بسم الله الرحمن الرحيم

مديرية الاستخبارات العسكرية العامة
منظومة استخبارات المنطقة الشرقية

السيد / المدير
الموضوع / [illegible] الكزنزاني

تم مواجهة المذكور وقدم التقرير المرفق طياً ورجائي
١. بيان مصير المقاتل غريب قادر احمد المنسوب الى فدوا ٧٠
وخلاصة موضوعه من المجموعة التي القي القبض عليهم من قبل
فل ٢ (الامن) وتم احالتهم الى فل ١ (الامن) والوعد منه سيتم
اعطائه الجوانب في ضوء توصياتكم.
٢. استفسر عن هويات جماعة الشيخ محمد الكزنزاني
٣. المقترحات
١. احالة التقرير الى س ٣ للمتابعة
ب. اخباره بان المقاتلين الصيل ١) قيادة فل ١ (الامن)
ج. الهويات لدى المديرية (س ٣)
يرجى التفضل بالاطلاع والتنسيب لسيدنا
للاجابة في ضوء الاتصال ١٧/٢ تحريري
وامركم.

هذا لو تابلت
لدى مقدم ركن فاضل ونفيسه ٣/٣
٤ / ٤ في ١٩٨٧
هاتف تعود لهم الاكم
٤/ ٨٧

[illegible]
٢. سيتم تنظيم الهويات

Figure 3.8: Examples from the Tobacco-800 database [85, 86] (left) and the Maryland Arabic database [11] (right).

3.6 Experiments

3.6.1 Datasets

To evaluate detection and matching in signature-based document image retrieval, we used two large collections of real world documents—the *Tobacco-800* database [86] and the *University of Maryland Arabic* database [11]. Tobacco-800 is a public subset of the complex document image processing (CDIP) collection [85] and has been used in TREC 2006 and 2007 evaluations. It was constructed from 42 million pages of English documents (in 7 million multi-page TIFF images) released by tobacco companies, and was originally hosted at UCSF [92]. Tobacco-800 is a realistic dataset for document analysis and retrieval as these documents were collected and scanned using a wide variety of equipments over time. The Maryland Arabic dataset consists of 166, 071 Arabic handwritten business documents.

Table 3.1: Summary of the English and Arabic evaluation datasets.

	Tobacco-800	*Maryland Arabic*
Document Types	Printed/handwritten	Mostly handwritten
Total Pages	1290	169
Resolution (in DPI)	150–300	200
Labeled Signatures	900	149
Number of Signers	66	21

Figure 3.8 shows some examples from the two datasets.

We tested our approach using all the 66 and 21 signature classes in Tobacco-800 and Maryland Arabic datasets, respectively, among which the number of signatures per person varies from 6 to 11. The overall performance across all queries is computed quantitatively in mean average precision (MAP) and mean R-precision (MRP), respectively.

3.6.2 Multi-scale Signature Detection

We organize our experiments on signature detection as follows. First, we use the detection probability P_D and the false-alarm probability P_F as evaluation metrics. P_D and P_F represent the two degrees of freedom in a binary hypothesis test, and they do not involve *a priori* probabilities of the hypothesis. To factor in the "*quality*" of detection, we consider a signature *correctly detected* and *complete* if the detected region overlaps with more than 75% of the labeled signature region. We declare a *false alarm* if the detected region does not overlap with more than 25% of any labeled signature. Since the number of signatures varies significantly

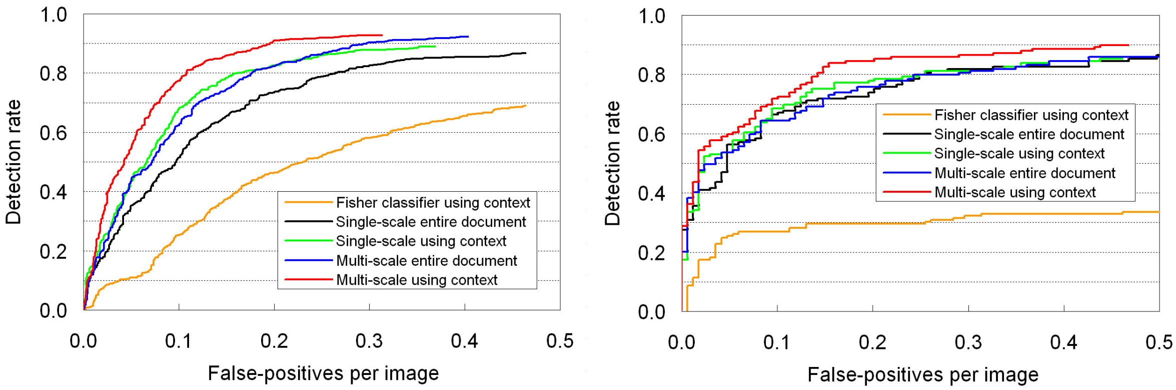


Figure 3.9: Signature detection ROC curves for the Tobacco-800 database (left) and the Maryland Arabic database (right).

across the documents in practice, we assume no prior knowledge on the distribution of signatures per document.

Figure 3.9 shows the receiver operating characteristic (ROC) curves for the Tobacco-800 and Maryland Arabic datasets. A Fisher classifier using size, aspect ratio, and spatial density features serves as a baseline for comparison, with all other procedures remaining the same in the experiment. We use two scale levels in multi-scale detection experiments. Parameters involved in obtaining the ROC curves, including the detection threshold in saliency and estimates of signature dimensions, are tuned on 10 documents. We use the following approach to compute each operating point on an ROC curve, and sort the list of detected candidates from the entire test set by their saliencies. To plot a new point, we move down the ranked list by one and look at the portion of the ranked list from the top to the current rank, which is equivalent to lowering the decision threshold gradually. The entire sets of ROC curves computed using this scheme, as shown in Figure 3.9, are densely packed and include every operating point.

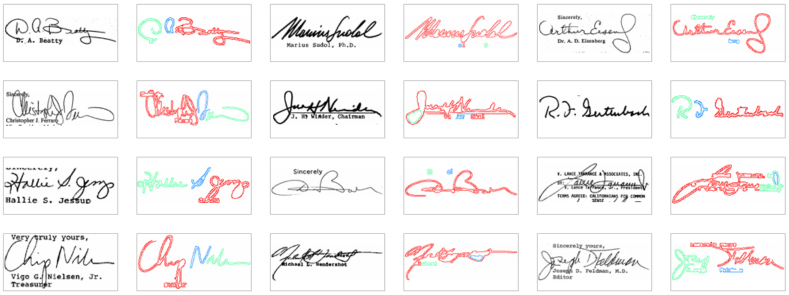


Figure 3.10: Examples of detected signatures from the Tobacco-800 database, together with their saliency maps. The top three most salient parts are shown in red, green, and blue, respectively.

The multi-scale saliency approach obtains the best overall detection performance on both English and Arabic datasets. Using document context, our multi-scale signature detector achieves 92.8% and 86.6% detection rates for the Tobacco-800 and Maryland Arabic datasets, at 0.3 false-positives per image (FPPI). Exploring context is more effective on machine printed documents, given the uniformity of geometric relationships among text lines. The operating point of 0.1 FPPI offers an example. Context alone gives an average increase of 13.6% in detection accuracy on Tobacco-800, compared to only 5.1% on the Maryland Arabic database. The improvements by using multi-scale are 5.8% and 4.0% on the Tobacco-800 and the Maryland Arabic datasets, respectively. This demonstrates the advantage of multi-scale approach on datasets that capture more diversity, such as the Tobacco-800 database.

Second, we test our saliency measure's ability to discriminate signatures from other handwriting. The handwritten Maryland Arabic dataset serves better, because local features become clearly less discriminative as evident from the poor performance of the Fisher classifier.

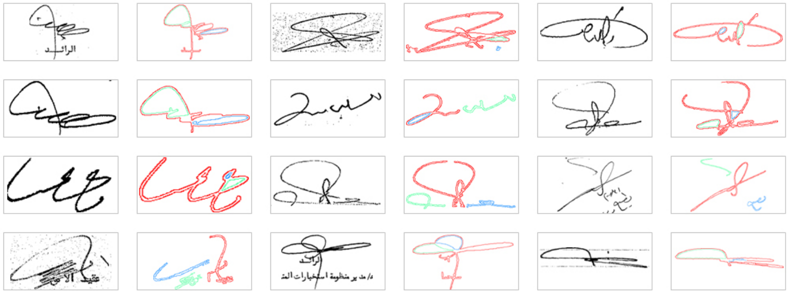
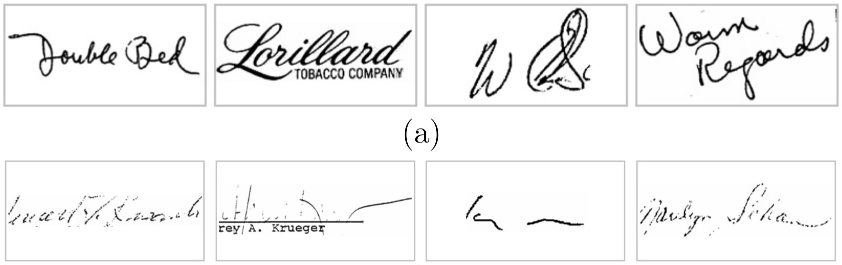


Figure 3.11: Examples of detected signatures from the Maryland Arabic database, together with their saliency maps. The top three most salient parts are shown in red, green, and blue, respectively.

Figures 3.10 and 3.11 show samples of detected signatures from Tobacco-800 and Maryland Arabic datasets, with their saliency maps. We delineate the top three most salient parts in red, green, and blue, respectively. In our experiment, a cursive structure is normally more than an order of magnitude more salient than printed text of the same dimensions. However, we did find rare instances of printed text among false positives as shown in Figure 3.12(a), with comparable saliencies to signatures because of their highly cursive font styles. A limitation of our proposed method derives from when the detected and segmented signature contain a few touching printed characters.

For better interpretation of the overall detection performance, we summarize key evaluation results. On Tobacco-800, 848 signatures of 900 labeled signatures are detected by the multi-scale saliency approach using document context. Among detected signatures, 83.3% are complete. The mean percentage area overlap with the groundtruth is 86.8%, with a standard deviation of 11.5%. As shown in Figures 3.10 and 3.11, using connected components for detection gives structurally and perceptually meaningful output quality. Furthermore, it does not necessarily limit



(a)

(b)

Figure 3.12: Examples of (a) false alarms and (b) missed signatures from the Tobacco-800 database.

the detection probability when used in a multi-scale framework. In fact, these figures approximate the word segmentation performance for machine printed text of a leading commercial OCR product [93] on Tobacco-800 documents. The results on the Maryland Arabic dataset are also encouraging, as the collection consists mainly of unconstrained handwriting in complex layouts and backgrounds. In addition, we have conducted field tests of using an ARDA-sponsored dataset composed of 32,706 document pages in 9,630 multi-page images. Among the top 1,000 detections, 880 are real signatures.

3.6.3 Discussion on Signature Detection

On the saliency maps, an edge detector generates two parallel contour lines from a stroke as both are local maxima in gradient magnitude. A ridge detector can generate more compact segmentation output since it produces only one thin curve in response. However, a ridge detector [94] performs significantly worse in signature detection in our experiments. The Canny edge detector provides good localization that guarantees accurate estimation of local gradient directions, and performs more robustly under structural variations.

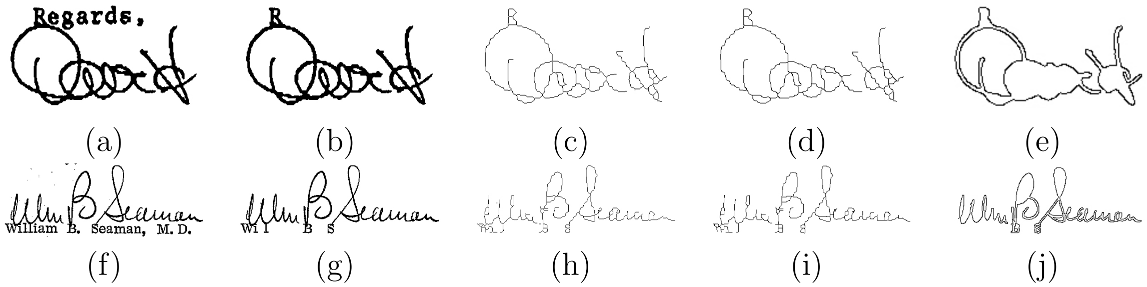
Examples of false positives from the Tobacco-800 database are shown in Figure 3.12(a), which include handwriting. The classification between signature and handwriting is sometimes not well posed by studying shape alone. Highly cursive handwriting may not have any obvious visual differences from signatures, as shown in Figure 3.12(a). Using geometric information cannot effectively resolve such intricacies in real documents because these handwritings are primarily annotations created in an ad hoc manner. Semantics and content layer information is required to solve the ambiguity in this case.

Figure 3.12(b) gives examples of false negatives. These missed signatures are severely broken, and a step edge operator like Canny could not detect contours, even on a coarse scale. As shown on most signatures, however, using multiple scales for detection partially overcomes the limitations of a connected-components-based approach by relaxing the requirement that the contour fragment be well connected at a particular scale. This improvement is more evident on the Tobacco-800 dataset, which contains a considerable number of highly degraded images at low resolutions.

3.6.4 Signature Matching for Document Image Retrieval

3.6.4.1 Shape Representation

We compare shape representations computed using different segmentation strategies in the context of document image retrieval. In particular, we consider skeleton and contour, which are widely used mid-level features in computer vision and can be extracted with relative robustness.



(a) (b) (c) (d) (e)

(f) (g) (h) (i) (j)

Figure 3.13: Skeletons and contours computed from signatures. The first column are labeled signature regions in the groundtruth. The second column are signature layers extracted from labeled signature regions by the baseline approach [95]. The third and fourth columns are skeletons computed by Dyer and Rosenfeld [96] and Zhang and Suen [97], respectively. The last column are salient contours of actual detected and segmented signatures from documents by our approach.

For comparison, we developed a baseline signature segmentation approach by removing machine printed text and noise from labeled signature regions in the groundtruth using a Fisher classifier [95]. To improve classification, the baseline approach models the local contexts among printed text using Markov Random Field (MRF). We implemented two classical algorithms—one by Dyer and Rosenfeld [96] and the other by Zhang and Suen [97], which compute skeletons from the signature layer extracted by the baseline approach. Figure 3.13 shows layer subtraction and skeleton extraction results by the baseline, as compared to the salient contours of detected and segmented signatures from documents.

In this experiment, we sample 200 points along the extracted skeleton and salient contour representations of each signature. We use the faster shape context matching algorithm to solve for correspondences between points on the two shapes and compute all the four shape distances using D_{sc}, D_{be}, D_{as}, and D_{re}. To remove any bias in all retrieval experiments, each query signature is removed from the test set in that query.

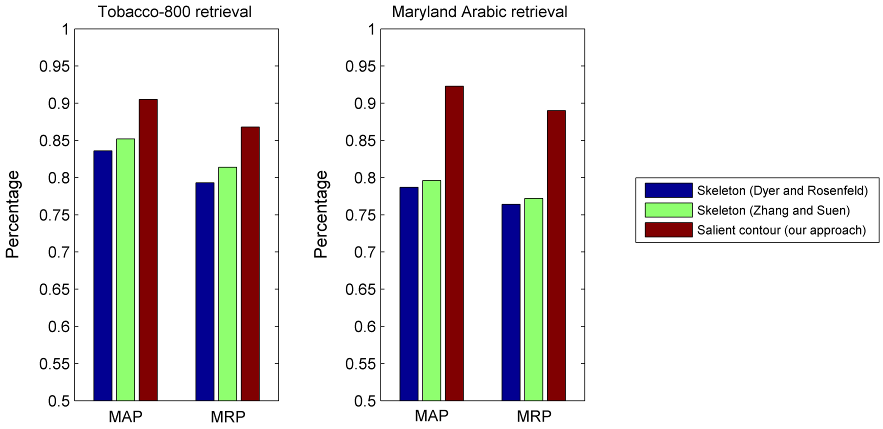


Figure 3.14: Document image retrieval performances using different shape representations on the two datasets measured by mean average precision (MAP) and mean R-precision (MRP).

Document image retrieval performance using different shape representations is shown in Figure 3.14. Salient contours computed by our detection and segmentation approach outperform the skeletons that are extracted directly from labeled signature regions on both datasets. As illustrated by the third and fourth columns in Figure 3.13, thinning algorithms are sensitive to structural variations among neighboring strokes and noise. In contrast, salient contours provide a globally consistent representation by weighting the more structurally important shape features. This advantage in retrieval performance shows clearly on the Maryland Arabic dataset, in which signatures and background handwriting are closely spaced.

3.6.4.2 Shape Matching Algorithms

We developed signature matching approaches using two non-rigid shape matching algorithms—shape contexts and local neighborhood graph, and evaluated their retrieval performances on salient contours. We use all four measures of dissimilarity

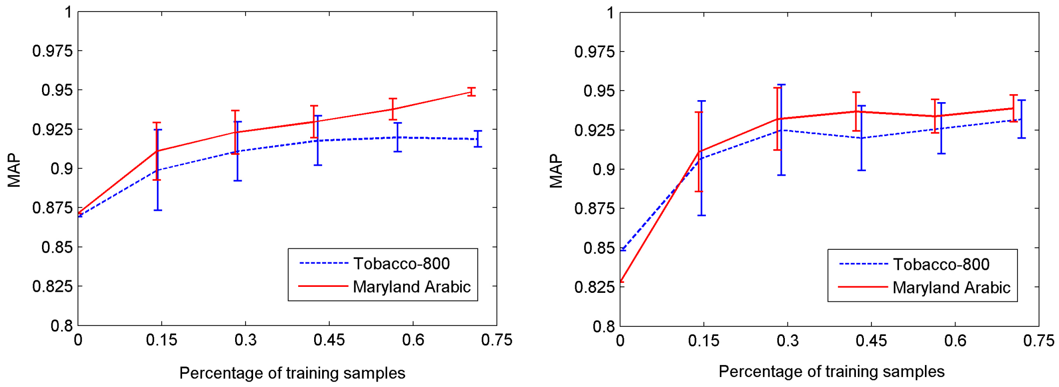


Figure 3.15: Document image retrieval using single signature instance as query using shape contexts [55] (left) and local neighborhood graph [56] (right). The weights for different shape distances computed by the four measures of dissimilarity can be optimized by LDA using a small amount of training data.

D_{sc}, D_{be}, D_{as}, and D_{re} in this experiment. The weights of different shape distances are optimized by linear discriminant analysis, using randomly selected subset of signature samples as training data. Figure 3.15 shows retrieval performances measured in MAP for both methods as the size of training set varies. A special case in Figure 3.15 occurs when no training data is used. In this case, we simply normalize each shape distance by the standard deviation computed from all instances in that query, effectively weighting every shape distance equally.

A significant increase in retrieval performance is observed with a fairly small amount of training data. Both shape matching methods are effective with no significant difference. In addition, the performances of both methods measured in MAP deviates less than 2.55% and 1.83%, respectively, when different training sets are randomly selected. These demonstrate the generalization performance of representing signatures by non-rigid shapes and counteracting large variations among unconstrained handwriting through geometrically invariant matching.

Table 3.2: Retrieval performance using different measure of shape dissimilarity on the Tobacco-800 database.

Measure of Dissimilarity	MAP	MRP
D_{sc}	66.9%	62.8%
D_{as}	61.3%	57.0%
D_{be}	59.8%	55.6%
D_{re}	52.5%	48.3%
$D_{sc} + D_{be}$	78.7%	74.3%
$D_{sc} + D_{as} + D_{be} + D_{re}$	90.5%	86.8%

3.6.4.3 Measures of Shape Dissimilarity

Table 3.2 summarizes the retrieval performance using different measures of shape dissimilarity on the larger Tobacco-800 database. The results are based on the shape context matching algorithm, as it demonstrates smaller performance deviation in previous experiment. We randomly select 20% of signature instances as training data and use the rest for test.

The most powerful single measure of dissimilarity for signature retrieval is the shape context distance (D_{sc}), followed by the affine transformation based measure (D_{as}), the TPS bending energy (D_{be}), and the registration residual error (D_{re}). By incorporating rich global shape information, shape contexts can discriminate even under large variations. Moreover, the experiment shows that measures based on transformations (affine for linear and TPS for non-linear transformation) are also effective. The two proposed measures of shape dissimilarity D_{sc} and D_{be} improve the retrieval performance considerably, increasing MAP from 78.7% to 90.5%. This demonstrates that we can significantly improve the retrieval quality by combining

Table 3.3: Retrieval performance using multiple signature instances on the Tobacco-800 database.

Measure of Dissimilarity	MAP	MRP
One	90.5%	86.8%
Two	92.6%	88.2%
Three	93.2%	89.5%

effective complementary measures of shape dissimilarity through limited supervised training.

3.6.4.4 Multiple Instances as Query

Table 3.3 summarizes the retrieval performances using multiple signature instances as an equivalent class in each query on the Tobacco-800 database. The queries consist of all the combinations of multiple signature instances from the same person, giving even larger query sets. In each query, we generate a single ranked list of retrieved document images using the final shape distance between each equivalent class of query signatures and each searched instance defined in Equation (3.17). As shown in Table 3.3, using multiple instances steadily improves the performance in terms of both MAP and MRP. The best results on Tobacco-800 is 93.2% MAP and 89.5% MRP, when three instances are used for each query.

3.6.5 Off-line Signature Verification

We quantitatively compare our signature matching approach with several state-of–the-art off-line signature verification approaches [44, 40, 48] on the Sabourin

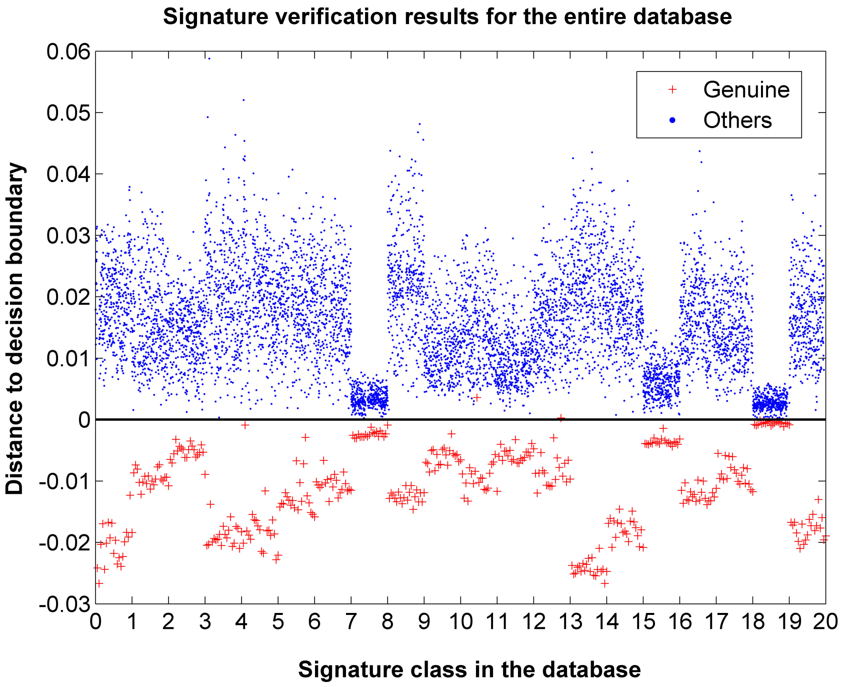


Figure 3.16: Our approach robustly discriminates genuine signatures (red cross) from forgeries (blue dots) across all signature classes in the Sabourin signature verification database [44].

off-line signature database [44], which has been used in [98, 99, 44, 40]. This database of 800 signature images contains genuine and random forged signatures of 20 classes, with 40 signatures per class. Historically, results on the Sabourin signature database are reported in terms of false rejection rate (FRR) of genuine signatures, and false acceptance rate (FAR) of forged signatures, with parameters trained to minimize total errors. We follow the same evaluation protocol as in [44, 40], where the first 20 signatures in each class are used for training and the remaining 20 signatures for test.

Our approach exceeded the state-of-the-art performance on off-line signature verification on the Sabourin database without any explicit assumption of local features, such as granulometric size distributions in [44] and sets of stroke-level features

Table 3.4: Signature verification results on the Sabourin off-line signature database [44].

Approach	FRR	FAR
Sabourin *et al.* [44]	1.64%	0.10%
Guo *et al.* [40]	1.66%	1.10%
Shanker and Rajagopalan [48]	8.75%	3.26%
Our approach	0.50%	0.21%

in [40]. The 0.50% FRR and 0.21% FAR is the best published off-line signature verification result on this database using the standard experimental protocol. As shown in Figure 3.16, our approach effectively separates genuine signatures and forgeries across all the 20 signature classes. The performance of [44] in Table 3.4 occurs by combining four classifiers using multiple classification schemes. The best single classifier reported in [44] gives 1.55% FRR and 0.15% FAR at 25 iterations. This experiment also demonstrates the significant difference in performance between our approach and [48], which employs an improved dynamic time warping algorithm on 1-D projection profile. Projection profile feature provides a limited ability to discriminate shape and has demonstrated sensitivity to changes in the baseline direction. Heuristic alignment, such as finding the orientation which gives the minimum length horizontal projection [48], is not robust under large structural variations in off-line signatures. Treating signatures in the unconstrained setting of a 2-D deformable shape, our approach is more robust to large intra-class variations and provides better generalization performance.

3.7 Summary

In this chapter, we proposed a novel signature detection and segmentation approach based on the view that object detection can be a process that aims to capture the characteristic structural saliency of the object across image scales. This differs from the common object detection framework that focuses on sets of local properties. The results on signature detection using multi-language document image collections show that our approach is effective in the presence of large variations and clustered backgrounds. In one advantage of using multi-scale saliency for joint detection and segmentation, it provides a general framework for which detection and segmentation degrades gracefully as the problem becomes more challenging. In addition, detected and segmented outputs are both structurally and perceptually meaningful for matching and recognition.

To handle large structural variations robustly, we treated the signature in the general setting of a non-rigid shape and demonstrated document image retrieval using state-of-the-art shape representations, measures of shape dissimilarity, shape matching algorithms, and multiple instances as query. We quantitatively evaluated these techniques in challenging retrieval tests, each composed of a large number of classes but a relatively small numbers of signature instances per class. We also demonstrated our matching techniques in off-line signature verification and achieved competitive results.

Chapter 4

Learning Document Context

4.1 Introduction

Obtaining structured information from unstructured image sources presents a grand document analysis challenge. In a wide range of applications requiring such capability, traditional language modeling techniques does not produce satisfactory results on the stream of OCR text, despite the character recognition performance on quality machine printed documents has improved steadily [100, 3].

Processing of receipts offers a classic example, where the unstructured nature of the document images present critical challenges in several aspects. First, text on receipt documents consists predominantly of terse streams of nouns. The lack of linguistic context, such as punctuation and language constructs, makes both syntactic and semantic analysis diffcult. Second, the output text from OCR systems does not contain useful page layout information, such as spatial block region and font type, which is readily available in the source image. Third, the text often contains high error rates (typically more than 6% at the character level) because receipts are often generated by low-resolution printers (e.g. dot matrix printers) and are kept in less than ideal physical conditions. These source image degradations are challenging to recover and have significant impact on the performance of character recognition and downstream tasks.

We present a model-based approach for extracting relevant named entities from unstructured document images, which enables learning of inference rules collectively based on contextual information from both page layout features in the image space and OCR text. We demonstrate the named entity extraction approach in an automated expense reimbursement system [9], which consists of (1) an electronic submission infrastructure that provides multi-channel image capture, transport, and storage of documents; (2) an document image analysis engine that extracts relevant named entities from unstructured document images; and (3) automation of auditing procedures.

Our contributions in this work include (1) a model-based approach to solving the named entity extraction and question answering aspects of the problem jointly; (2) a formal framework for efficient probabilistic inference by learning contextual dependencies using both page layout and text features; and (3) a demonstration of the effectiveness of integrating rich feature sets available in the image space, when OCR text reveals limited structural information.

The remainder of this chapter is organized as follows: The next section provides background information on the emerging application of automated expense reimbursement. Section 4.3 presents an overview of the system architecture and components. Section 4.4 describes our approach to extracting relevant named entities from diverse, unstructured document images. Section 4.5 discusses experimental results on collections of real-world receipts. We review related work in Section 4.6, and conclude in Section 4.7.

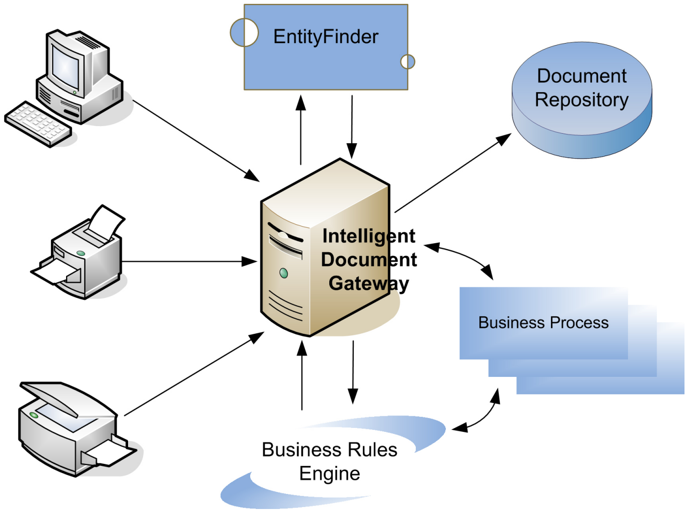


Figure 4.1: System architecture of our automated expense reimbursement system.

4.2 Motivations and Background

Expense reimbursement is a time consuming and labor intensive process for organizations of all sizes. Even though policies and regulations defining the process vary across organizations and industries, corporate expense reimbursement faces a set of common challenges. The solution requires technical innovations in the following three key areas.

(1) *A generalized paper-free framework for capturing, transporting, and storing paper documents in digital image form*

In spite of progress made with electronic tools, paper consumption in the office is growing, and paper continues to inhibit business process innovation. Expense reporting provides an example where paper receipts continue to generate unnecessary costs and delays even though the organization has access to web-based applications. The problem arises from the requirements to hold receipts to prove the validity of submitted reimbursement claims. Currently, without any pervasive mechanism for electronic submission, paper receipts are mailed for centralized processing, with printed cover sheets. To protect against the risk of loss in the mail, the package is often copied, which still creates more paper. Even worse than handling the expanding amount of paper, time to reimbursement remains at the speed of mailed package and manual processing rather than at the speed of electronic transaction.

A fast technology shift in the printing industry from analog copiers to consolidated high-resolution digital multi-function devices (MFDs) enables us to close the paper-digital gap by using these image capturing and transporting channels. In our system, submission of paper paper receipts requires a few easy steps: walking to the office MFD, authenticating yourself with your intranet password, selecting the appropriate menu option on the touch screen, and hitting the "big green button" to scan, and submitting receipts singly.

(2) *Extraction of relevant named entities from diverse receipt images with unconstrained layouts and formatting*

It is important to distinguish between the standard named entity recognition (NER) problem [101, 102, 103] and the one present in this situation. In our task, the query to the relevant named entity in each category equals a question. We look to find the unique answer that best answers the question using the presented context. For instance, given a receipt document, we can ask for the name of the merchant. If more than one merchant entity exists, the system needs to resolve such ambiguity and return most relevant answer using all available cues collectively.

Effective solution to both the entity extraction and question answering (QA) requires integrating interdependent mixture of features from page layout and language content. This is a research area has relatively little work in the literature. In addition, we prefer a formal model-based approach that can be trained on new data.

(3) *Automation of auditing procedures that enables an organization to perform expense validation with minimum human interaction*

Many organizations have a limited ability to audit volumes of expense reports, as it requires dedicated auditors to examine incoming receipts manually and judge their accuracy from the associated report. This labor-intensive approach often causes an organization to downgrade their internal requirements for the percentage of submissions audited. For some organizations with have a high percentage of employees requiring travel, controlling costs in expense processing generally requires a lower rate of oversight than would be desired.

4.3 System Overview

In this following sub-sections, we give an overview of the expense reimbursement system and describe each component in its client-server architecture, as shown in Figure 4.1. The system's client may be any computer, multi-function device (MFD), fax machine, or other electronic device, which has built-in capability to transmit native image files. The application server running an IBM Intelligent Document Gateway (IDG) server directly interacts with the centralized document image repository, the named entity extraction engine (EntityFinder), the dynamic business rule engine server, and back-end business processes.

4.3.1 Electronic Submission

Our system provides multi-channel image capture, transport, and storage of paper documents. On the client side, users have several options to submit paper receipts or scanned receipt images securely to the document server:

1. Multi-function devices (MFDs): The touch screen displays a customized user interface for directions of how to scan and submit paper receipts, once the user successfully authenticate by providing passwords.

2. Web-based client: The user uploads document image files directly from their computer to the server through a light-weight client. This web-based application allows the user to provide additional information associated with the submitted receipt document, including its language set and personal remainder.

3. Desktop print-job: This option allows the user to submit receipt documents in native image format through their computer's printer queue.

The meta-data transmitted along with a submitted image file includes the type of the document and an identifier that links the submitted receipt document to the corresponding expense claim. The receipt image and its associated meta-data are encrypted prior to transmission to the document gateway server via the corporate or pubic networks.

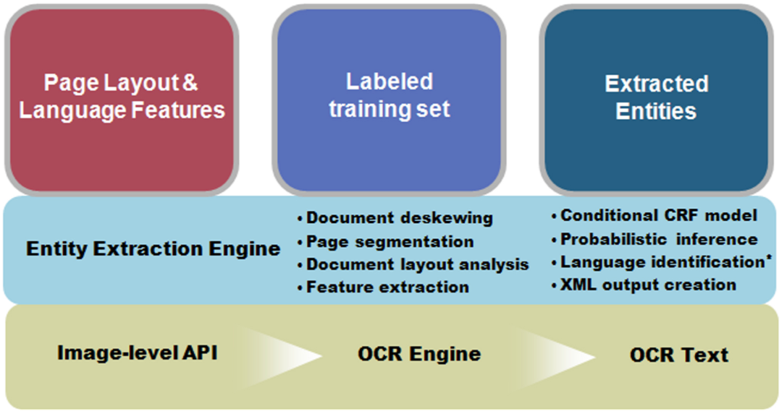


Figure 4.2: Architecture of the document image analysis module – EntityFinder.

4.3.2 Named Entity Extraction

The capability to extract relevant named entities from documents is integrated into a document image analysis module called EntityFinder, as shown in Figure 4.2. At the lower level, EntityFinder handles images at their native formats (e.g. multi-page TIFF images) and provides support for higher-level functions, including document layout analysis and feature extraction, through interfaces with the OCR engine libraries. We present our approach to relevant named entity extraction in Section 4.4.

4.3.3 Automated Auditing

Extracting relevant named entities from unstructured document images opens many possibilities for business process automation. Our system uses a business rules engine to analyze the extracted data for relevancy within the context of automated expense auditing, and it activates actions based on the result of the rule execution. The set of business rules are defined in XML and are dynamically configurable in the live system. The auditing actions include verification of extracted entities from receipts with reference to their corresponding entries in the expense reimbursement claim, flagging instances of potential fraud, adherence to prescribed organizational policies (*e.g.*, meal limit). Automation of these routine procedures enables a significantly higher rate of auditing and a much shorter processing time between submission and compensation, bringing tangible productivity gain and cost savings to the organization.

4.3.4 Document Archival

Once all business rules for automated auditing have been executed and the resulting external business processes have been initiated, the set of extracted named entities, along with the source document, are stored in centralized repositories. The task aims to conform to business and legal requirements for document retention, for future data analysis needs of the organization, and for auditing control purpose. Document archival is governed by a configuration file associated with the document process. This includes the type of repository adapter to use, a link to the server, associated authentication data, and the descriptive information, including table and column information for database access. The design allows multiple repositories as required by a given process, and can be flexibly adapted to organization-specific archival requirements.

4.4 Extraction of Relevant Named Entities

4.4.1 Task

We consider the task of extracting transaction-related named entities (NEs) from receipt documents. Different levels of complexities are involved in extracting NEs of diverse natures. The limited variation in NEs like transaction amount, date, credit card number, and merchant phone number can be handled effectively using regular expressions, in combination with rules. In this chapter, we focus on the task of finding the set of NEs with arbitrarily large variation (*e.g.*, the merchant) and present an approach to extracting these challenging NEs by exploiting context collectively from the page layout in the source image and its OCR text.

The application imposes three major requirements:

1. Handle document images effectively, given unconstrained layouts and formatting, since the system must be able to process all kinds of receipts.

2. Provide the most likely answer to each NE as inferred collectively from the document's context.

3. Should not rely on large external dictionaries, as it is not economical to create and maintain such dictionaries. Furthermore, NEs on receipts commonly appear in various abbreviated forms that are difficult to enumerate. In fact, even with this constraint lifted, the NE extraction task is not trivial, but presents a different set of problems. The challenges involved in improving NER performance using external dictionaries are discussed in [104].

4.4.2 Our Approach

The structural information derived collectively from page layout and language features is important to the NER task on unstructured documents like receipts, where OCR text stream may not be sufficient. Figure 4.3 shows a receipt from a Union 76 gas station. The string "76" itself is most likely to be a number when it appears without context. However, people can determine that its reference to a merchant by examining the document layout and linguistic elements collectively.

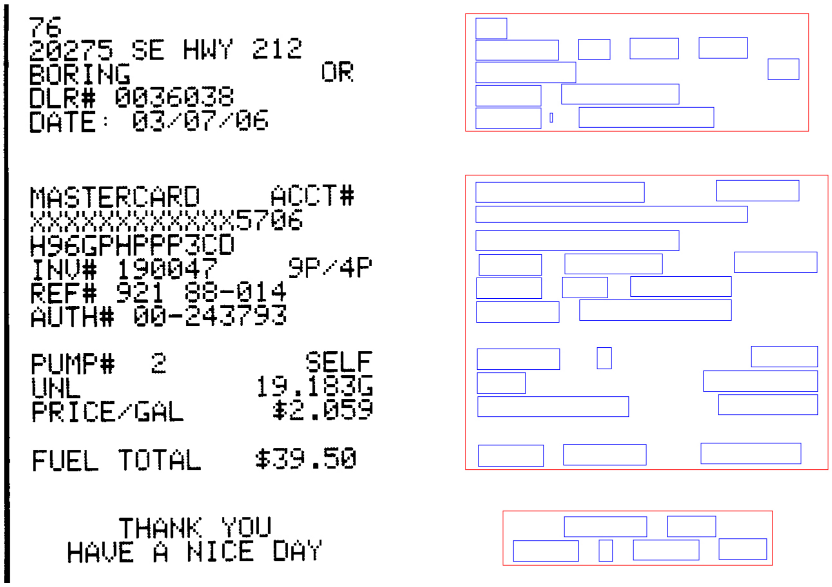
A receipt document with unconstrained page layout and formatting still conveys structural information in two aspects:

- Many semantically related entities are placed geometrically within spatial proximities, even if their structure within the region does not seem obvious.
- The sequence of decomposed regions and the combination of layout and linguistic features within the regions reveals important contextual information.

Our approach to extracting relevant NEs involves (1) decomposing the document image spatially into regions by page segmentation; and (2) learning the inference rules collectively in a discriminative conditional framework using the contextual information from page layout and text features.

4.4.2.1 Page Segmentation

Two page segmentation strategies can be employed to divide a general document image into homogeneous regions. In one strategy, a page segmentation algorithm is used. Representative page segmentation approaches from the document



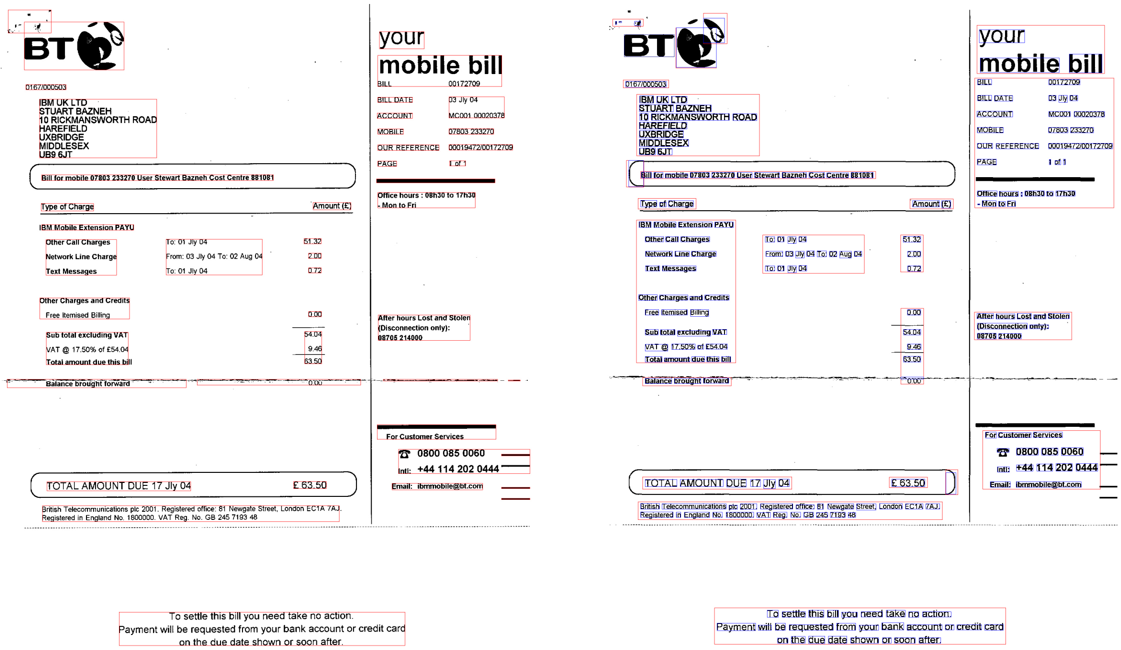
```
76
20275 SE HWY 212
BORING          OR
DLR# 0036038
DATE: 03/07/06


MASTERCARD    ACCT#
XXXXXXXXXXXX5706
H96GPHPPP3CD
INV# 190047     9P/4P
REF# 921 88-014
AUTH# 00-243793

PUMP#  2         SELF
UNL            19.183G
PRICE/GAL       $2.059

FUEL TOTAL    $39.50

     THANK YOU
  HAVE A NICE DAY
```

Figure 4.3: Page and word segmentation results of a receipt image.

image analysis community include the Docstrum algorithm by O'Gorman [105] and the Voronoi diagram-based algorithm by Kise et al. [106]. Another approach uses the OCR engine through low-level function calls. Figure 4.4 shows the page segmentation results by the Docstrum algorithm[1] and the OCR engine, respectively. Each segmented region is plotted using a red bounding box. For better visualization of content within each segmented region, we show the word segmentation results in the right sub-figure of Figure 4.4 using blue bounding boxes.

Using the OCR engine to segment a document page directly has a few practical advantages, as opposed to a stand-alone algorithm. First, it makes region-level attributes easily accessible. Most leading commercial OCR packages offer region-level classification capabilities on machine-printed documents, which allow regions con-

[1] Docstrum is a bottom-up page segmentation algorithm that can work on document images with non-Manhattan layout and arbitrary skew angles. It has limited capability to handle non-text regions and text zones with irregular font sizes and spacing, and it tends to fragment them.



0167/000503

IBM UK LTD
STUART BAZNEH
10 RICKMANSWORTH ROAD
HAREFIELD
UXBRIDGE
MIDDLESEX
UB9 6JT

your
mobile bill

BILL	00172709
BILL DATE	03 Jly 04
ACCOUNT	MC001 00020378
MOBILE	07803 233270
OUR REFERENCE	00019472/00172709
PAGE	1 of 1

Bill for mobile 07803 233270 User Stewart Bazneh Cost Centre 881081

Type of Charge		Amount (£)
IBM Mobile Extension PAYU		
Other Call Charges	To: 01 Jly 04	51.32
Network Line Charge	From: 03 Jly 04 To: 02 Aug 04	2.00
Text Messages	To: 01 Jly 04	0.72
Other Charges and Credits		
Free Itemised Billing		0.00
Sub total excluding VAT		54.04
VAT @ 17.50% of £54.04		9.46
Total amount due this bill		63.50
Balance brought forward		0.00

Office hours : 08h30 to 17h30
- Mon to Fri

After hours Lost and Stolen
(Disconnection only):
08705 214000

For Customer Services
☎ 0800 085 0060
Intl: +44 114 202 0444
Email: ibmmobile@bt.com

TOTAL AMOUNT DUE 17 Jly 04 £ 63.50

British Telecommunications plc 2001. Registered office: 81 Newgate Street, London EC1A 7AJ.
Registered in England No. 1800000. VAT Reg. No. GB 245 7193 48

To settle this bill you need take no action.
Payment will be requested from your bank account or credit card
on the due date shown or soon after.

0167/000503

IBM UK LTD
STUART BAZNEH
10 RICKMANSWORTH ROAD
HAREFIELD
UXBRIDGE
MIDDLESEX
UB9 6JT

your
mobile bill

BILL 00172709
BILL DATE 03 Jly 04
ACCOUNT MC001 00020378
MOBILE 07803 233270
OUR REFERENCE 00019472/00172709
PAGE 1 of 1

Bill for mobile 07803 233270 User Stewart Bazneh Cost Centre 881081

Type of Charge Amount (£)
IBM Mobile Extension PAYU
Other Call Charges To: 01 Jly 04 51.32
Network Line Charge From: 03 Jly 04 To: 02 Aug 04 2.00
Text Messages To: 01 Jly 04 0.72
Other Charges and Credits
Free Itemised Billing 0.00
Sub total excluding VAT 54.04
VAT @ 17.50% of £54.04 9.46
Total amount due this bill 63.50
Balance brought forward 0.00

Office hours : 08h30 to 17h30
- Mon to Fri

After hours Lost and Stolen
(Disconnection only):
08705 214000

For Customer Services
☎ 0800 085 0060
Intl: +44 114 202 0444
Email: ibmmobile@bt.com

TOTAL AMOUNT DUE 17 Jly 04 £ 63.50

British Telecommunications plc 2001. Registered office: 81 Newgate Street, London EC1A 7AJ.
Registered in England No. 1800000. VAT Reg. No. GB 245 7193 48

To settle this bill you need take no action.
Payment will be requested from your bank account or credit card
on the due date shown or soon after.

Figure 4.4: Page segmentation results by the Docstrum algorithm [105] (left) and by the OCR engine [93] (right).

taining text, tables, and graphics to be identified and processed accordingly. Second, it facilitates feature extraction from the segmented regions, including character-level attributes such as the coordinates of character borders on the image grid, font information, and recognition confidence. Last, using OCR for page segmentation removes the tedious step of training the free parameters involved in a stand-alone algorithm. Packaged OCR products provide a convenient black-box solution, in which the engine parameters have been tuned for performance over large collections of documents. At the post-processing stage, the OCR engine can use preliminary recognition results to improve page segmentation in an iterative fashion. Additional information, including consistency in the font style and spatial alignment of segmented regions, helps improve overall page segmentation performance, and tends to produce more structurally meaningful results, even for a highly degraded input image. Past empirical studies [107] also observe this, where representative page segmentation algorithms are evaluated against built-in page segmentation functions provided by several OCR products.

4.4.2.2 Learning From Document Context

We use conditional random fields (CRFs) as the framework for learning inference rules based on features at multiple levels of granularity and multiple modalities. Clear motivations exist for using a discriminative model, such as CRFs. First, CRFs relax the strict conditional independence assumptions of observations in generative models like hidden Markov models (HMMs) for ensuring tractable inference [108]. This allows CRFs more flexibility to integrate complex, overlapping and non-independent feature sets that operate at multiple levels of granularity and multiple modalities. Second, modeling of conditional probabilities devotes model resources directly to the relevant task of label inference. It generally requires fewer labeled observation sequences, which leads to better generalization performance given limited training data. In addition, CRFs avoid the label bias problem exhibited by maximum entropy Markov models (MEMMs) and other discriminative Markov models based on directed graphical models [109]. A few studies have shown that CRFs outperform both MEMMs and HMMs on a number of real-world language related sequence labeling tasks [109, 110, 111, 112].

4.4.3 Feature Selection

4.4.3.1 Practical Constrains

Optical character recognition on machine-printed characters has emerged as an industrial-strength technology since its phenomenal advances from early 1990s. However, OCR accuracy still lacks in comparison to a second-grade child [3]. It is important to place the strength and weakness of OCR technology in perspective, and understand the factors involved that significantly affect performance. These insights provide useful guidelines for selecting feature sets that can be extracted relatively reliably given the practical constraints imposed by a targeted application.

The most successful application domain of OCR technology occurs with machine-printed characters. During the last decade, the acceptance rates of form readers on hand-printed digits and constrained alphanumeric fields have risen significantly. The relatively low recognition errors in these constrained domains reflect the complexities involved in classifying a novel pattern under such a limited variation in the data set [113]. In contrast, recognition of unconstrained off-line human handwriting and multi-lingual recognition among a variety of scripts are more challenging problems, and they remain active research frontiers.

The accumulated imaging degradations have a significant effect on OCR performance. Typical imaging defects in the printing process include blotchy characters caused by dot-matrix printer ribbons, and faint impressions resulting from worn ribbons and printer cartridges. The scanning process also introduces various imperfections. Digital scanning involves sampling both horizontally and vertically on the image grid. Desirable sampling rates by OCR are beyond 300 dots per inch (dpi). Although commercial packages can work at as low as 150 dpi by interpolating a low-resolution image to the preferred dpi, this generally leads to significant increase in recognition errors. Significant image degradation also occurs when storing an image in binary format by applying thresholds to separate foreground content. Using gray-scale and color scans of the image captures more detailed information for pattern recognition and reduces the error rate. Most high-end MFDs provide these functions.

4.4.3.2 Feature Set

We use a rich combination of page layout and text features for NE extraction. The feature extraction process can be viewed as a set of binary-valued functions defined on the appropriate feature space that output either 1 or 0 based on the presence or absence of the corresponding feature. The conditional nature of CRFs enables effective learning from these discrete-valued, interdependent features, which may have extremely complex joint probabilities.

Page Layout Features As shown in Figure 4.4, noise speckles and graphic elements in the document, including logos, lines, and region borders, may not be reliably classified and segmented, and thus be given to the machine-printed text recognizer in error. We simply discard those segmented text regions, where the majority of text is un-recognizable or suspicious. We use the following collection of page layout features extracted from each segmented region:

- Variation of font size and font face within the region
- Presence of the largest font on the entire page
- Presence of bold font face in the region
- Whether the text block is horizontally aligned to the center
- Whether the text block horizontally aligned to the left

Text Features Word tokens that are logically or semantically related to a NE are useful and relatively robust features for extracting the NE. In fact, current OCR systems commonly use the technique for character-level error correction that makes explicit use of context at word level, by choosing a common letter n-gram over a rare one [3]. This allows for improvement in recognition performance at word level, even if the quality of the image remains poor. We organize word tokens into equivalent groups. For example, "Inc." and "Companies" are grouped together. The following text features are also used:

- Capitalization of words
- Mixed cases
- Frequent appearance of digit characters (0-9)
- Presence of special characters (/, –, #, -, *, $, £)
- Presence of special patterns ('s)

Named Entity Features The orthogonality between NEs can be effective features for probabilistic inference. A region of text containing a credit card number is less likely to contain the name of the merchant. We use the set of orthogonally related NEs as features. These include addresses, phone numbers, credit card numbers, dates, and monetary amounts.

4.4.3.3 Region Labels

We use a compact set of labels to categorize the ordered list of regions obtained by page segmentation. This is based on our observation that context change along the sequence of regions is frequent, making the inference of label more effective among neighboring regions. For extracting the merchant, we use three labels of regions, NON_DATA, MERCHANT_DATA, and TRANS_DATA. NON_DATA represents a region that does not contain details of a transaction or any association with a merchant. The MERCHANT_DATA denotes a region that contains a merchant. TRANS_DATA region includes details of a transaction.

4.4.4 Relevant NE Extraction with CRFs

A conditional random field can be viewed as an undirected graphical model, and be used to compute the conditional probability of labels on designated output nodes $\mathbf{Y}$, when globally conditioned on $\mathbf{X}$, the random variable representing observation sequences. We construct a conditional model $p(\mathbf{y}|\mathbf{x})$ from paired observation and labeled sequences, and do not explicitly model the marginal $p(\mathbf{x})$.

We use CRFs with a linear chain structure. Given an instance of observed input sequence $\mathbf{x}$, the probability of a label sequence $\mathbf{y}$ is defined in [109] as

$$p(\mathbf{y}|\mathbf{x}) \propto \exp\left(\sum_{t=1}^{T}\left(\sum_{k}\lambda_k f_k(y_{t-1}, y_t, \mathbf{x}, t) + \sum_{k}\mu_k g_k(y_t, \mathbf{x}, t)\right)\right), \qquad (4.1)$$

where $f_k(y_{t-1}, y_t, \mathbf{x}, t)$ is a transition feature function of the entire observation sequence and labels at positions t and $t-1$; $g_k(y_t, \mathbf{x}, t)$ is a state feature function of the label at position t and the observation sequence. More compactly, the probability of a label sequence $\mathbf{y}$ given the observation sequence $\mathbf{x}$ is given by

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k F_k(y_{t-1}, y_t, \mathbf{x}, t)\right), \qquad (4.2)$$

where $Z(\mathbf{x})$ is a normalization factor and

$$Z(\mathbf{x}) = \sum_{s \in \mathcal{S}^T}\exp\left(\sum_{t=1}^{T}\sum_{k}\lambda_k F_k(y_{t-1}, y_t, \mathbf{x}, t)\right). \qquad (4.3)$$

Assuming the training data $\mathcal{D} = \{\mathbf{x}^i, \mathbf{y}^i\}_{i=1}^{N}$ are independent identically distributed, the product of Equation (4.2) over all training sequences as a function of the parameters λ is the likelihood function. Maximum likelihood training chooses parameters values so that the log-likelihood is maximized. For a CRF, the log-likelihood is a concave function, guaranteeing convergence to the global maximum.

$$L = \sum_{i=1}^{N}\log p(\mathbf{y}^i|\mathbf{x}^i). \qquad (4.4)$$

Likelihood maximization can be performed efficiently using a quasi-Newton method, such as L-BFGS [114], which approximates the second derivative of the likelihood by keeping a running, finite window of previous first-derivatives. L-BFGS can be treated as a black-box optimization procedure, requiring only the first derivative of the function to be optimized.

Table 4.1: Summary of the two real-world receipt image collections.

	Collection 1	Collection 2
Total images	145	283
Number of characters	$71,316$	$522,320$
Character error rate	6.03%	9.48%
Image resolution	200-300 dpi	150-200 dpi
Country of origin	US	UK

Let $\mathbf{y}^i$ be the state path up to position T on instance i of the labeled training sequence. The first-derivative of the log-likelihood function is given by

$$\frac{\delta L}{\delta \lambda_k} = \sum_{i=1}^{N}\sum_{t=1}^{T} F_k(y_t^i, y_{t-1}^i, \mathbf{x}_t^i) - \sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{y,y'} F_k(y, y', \mathbf{x}_t^i) p(y, y'|\mathbf{x}^i). \qquad (4.5)$$

Intuitively, when the state paths chosen by the CRF parameters match the state paths from the labeled sequence, the derivative given in Equation (4.5) becomes zero.

4.5 Results and Discussion

4.5.1 Datasets

We used two large real-world receipt collections provided by IBM World Wide Reimbursement Center for training and testing, which contain binary scanned receipt images from IBM internal business units and IBM Global Services customers. These two collections provide realistic examples because the paper receipts were gathered and scanned over time, using a variety of equipment. Characteristics of the two datasets are summarized in Table 4.1.

We used the first two fifths from each dataset for training and the rest for test. Groundtruth labels were created by first running a rule-based heuristic on the sequence of segmented regions. Human judgment was then employed to correct mistakes in the heuristic labeling.

4.5.2 Evaluation and Discussion

We evaluate performance with the widely used precision-recall metrics. Throughout our evaluation, we define recall as the ratio of the number of NEs correctly extracted to the number of NEs that are physically present in the collection. Precision is the ratio of the number of NEs correctly extracted divided by the total number of NEs extracted in the category. The F-Measure (or F1 Measure) is computed by $(2 \times \text{Precision} \times \text{Recall})/(\text{Precision} + \text{Recall})$ [115].

Table 4.2 summarizes the overall NE extraction performances on the two collections of receipt images, respectively. On both datasets, using CRF to extract the merchant significantly outperformed rule-based heuristic approach. This is encouraging because improvements on a rule-based system require constant changes to the code base, while improvements on the CRF system generally require only defining new features and retraining the model. In fact, we observe improvement in performance after modifying the word token sets to reflect locale difference of the two databases.

The impact of recognition errors on rule-based NE extraction approaches is evident. Almost all the errors made by the heuristic on simple NEs were caused

Table 4.2: Named entity extraction performances on the two databases.

	Collection 1			Collection 2		
Entity	Precision	Recall	F-Measure	Precision	Recall	F-Measure
Date	100.0	85.2	92.0	96.3	78.2	86.3
Credit card #	98.8	76.5	86.2	92.4	69.5	79.3
Expense total	100.0	97.7	98.8	89.1	91.7	90.4
Phone #	95.3	78.1	85.8	84.8	66.1	74.3
Address	95.4	82.7	88.6	82.3	58.6	68.5
Merchant (by heuristic)	63.2	56.8	59.8	58.5	49.7	53.7
Merchant (by CRF)	73.8	70.5	72.1	67.2	62.9	65.0
Merchant (improvement)	10.6	13.8	12.3	8.7	13.2	11.3

by text errors. Image scan on a higher-resolution device or a device that supports gray-scale or color format can minimize this problem by effectively containing the recognition errors. In addition, a paper document undergoes the image capturing (scanning) process only once in our system, in contrast to some receipts in the evaluation datasets that are second-generation copies (*e.g.*, a scanned image of a previously faxed document). The figures in Table 4.2 provide a realistic estimate of the practical performance, given improvements in digital printing technologies.

Extracting relevant NEs using CRF proves more robust to character recognition errors because it leverages collective information from presented document context. For instance, the 9.48% character error rate on Collection 2 translates to a 39.2% word level error rate for a five-letter word. Nevertheless, we achieved 65.0% F-score on merchants using CRF on the dataset.

4.6 Related Work

Named entity recognition (NER) is an essential task in deriving structured information from un-structured sources. Historically, it has been defined on un-structured text due to its roots in Message Understanding Conferences [101]. NER capabilities have been demonstrated using un-structured text corpora from a wide range of domains, including identifying personal names and company names in newswire text [102], identifying titles and authors in on-line publications [103, 116], and identifying gene and protein names in biomedical publications [117, 118]. More recently, unsupervised NER results are reported on a massive corpus of domain-independent text from the web [119].

The vast majority of NER systems employ rule-based or machine-learning-based approaches. Examples of rule-based systems in the literature include [120, 121]. Machine-learning-based approaches can be further divided into two main categories: classifier-based and Markov-model-based. Common choices of classifiers include decision trees, naive Bayes, and support vector machines (SVMs). In addition, studies in [122] and [123] have used classifier combination techniques in NER tasks. Markov-based models, including hidden Markov models (HMMs) [124], maximum entropy Markov models (MEMMs) [125], and Conditional Random Fields (CRFs) [109], are well suited to problems that involve sequential analysis. More recently, CRFs have also been extended to computer vision problems, including region classification [126] and human motion recognition [127].

4.7 Summary

In this chapter, we proposed a model-based approach for extracting relevant named entities from unstructured documents, which enables learning about inference rules collectively based on contextual information from both page layout and text features. We applied our approach to the processing of diverse real-world receipt documents in an expense reimbursement system, and demonstrated that it brought significant improvement to named entity extraction. Our approach delivers better generalization performance and is shown to be more robust to recognition errors than approaches that rely solely on linguistic features. These results demonstrate the importance of jointly interpreting the context of the document when text itself reveals limited structural information.

Chapter 5

Mining Session Context

5.1 Introduction

The understanding of users' behavior presents a major challenge in improving the ranking performance and the usability of a document retrieval system. One common solution used by search engines is to record and analyze user activities on search result pages or within query sessions. These search-related logs provide indication of relevance judgement and can be useful as training data to improve numerous search capabilities [128], including query refinement and suggestion [129, 130, 131], learning ranking functions [132, 133, 134, 135], image search [136], and targeted advertising [137].

Analysis of search-centric user behavior data sources has several inherent limitations. First, search-oriented activity represents a minimal fraction (less than 5%) of a user's behavior online [138], even taking into account their post-query browsing trails. This may significantly affect the document coverage. Second, typical search engine logs contain user activities only within the scope of interactions with search result pages [139], despite the fact that the use of contextual information can improve the user's search experience. Third, user activities on search result pages or within search sessions introduce a strong bias toward dated documents with a high ranking [140], since users seldom search beyond the first page [2]. Consequently, the learned ranking will not converge to an optimal ranking because documents that are fairly recent or initially ranked low will rarely be presented to users.

By effectively incorporating information on *all* web user activities, search engines gain insights into user preferences and intents, and improve both retrieval performance and user experience. First, analysis of all user actions provides a more robust estimate of a user's perceived importance associated with web pages and sites [141]. Second, search engines face the challenge of prioritizing and adapting their computing resources under practical constraints in crawling, indexing, and query processing [142]. In this context, the relative attention a web page receives from all users provides an intuitive and user-centric optimization criterion, and responds to evolving user behaviors. As a large amount of web content emerges and refreshes within a shorter time interval than a typical crawling and indexing cycle of a search engine [143], discovering popular content and adapting crawling schedules based on the degree of usage may prove an effective and agile policy. Finally, another challenging area for search engines is the access to the deep (or invisible) web – the fraction of the web that is dynamically generated and not directly accessible to automated crawlers [144]. Their coverage can substantially improve by leveraging large-scale user browsing history, which collectively reveals hidden URLs, providing gateways to their contents.

As a logical unit of user web experience, a web session contains rich information on the user's preferences and intents within an actionable time frame [138]. Web session representation and interpretation is a particularly important subject for the web community because it applies across all activities.

In this dissertation, we focus on using the large amount of knowledge gained from computational analysis of general user browsing behavior by: (1) leveraging information on all browsing actions; and (2) developing models that incorporate rich contextual information within web sessions. Our main contribution is ClickRank, a novel algorithm we propose for estimating web page and web site importance, which is based on these two key notions. ClickRank first estimates a local importance value for every page or site in each user browsing session, based on the implicit preference judgments of the user in the session context. It then aggregates these local values over all sessions of interest to construct a global ranking.

We evaluate ClickRank in three important areas of web search. Our first experiment tests the traditional task of web site ranking, where we show that results from ClickRank are competitive against state-of-the-art approaches, including PageRank [145] and the recently proposed BrowseRank [141], and are obtained at significantly lower computational costs. In the second experiment, we demonstrate the novelty and effectiveness of ClickRank in web page ranking with several hundred state-of-the-art web search features, including those computed from page visit counts and the link structure of the web graph. In this large-scale test, we formulate the task of learning the optimal ranking model as an additive regression problem using gradient-boosted decision trees, and quantify the variable importance of ClickRank in direct comparison with other quality features. Finally, we demonstrate ClickRank in a system that mines and presents recent, popular pages to web search users as dynamic quicklinks in search result summaries.

We structure the remainder of this chapter as follows. Section 5.2 reviews related work. In Section 5.3, we present important characteristics of general web sessions, and describe in details our approach to session mining by incorporating contextual information in session representation. In Section 5.4, we introduce the ClickRank algorithm, and describe how we combine it with other useful web search signals to improve web search by learning an optimal ranking model. We comprehensively evaluate ClickRank in three web search applications in Section 5.5, and conclude in Section 5.6.

5.2 Related Work

PageRank [145], HITS [146], and TrustRank [147] are representative link analysis algorithms for computing authoritative sources, using the link structure of the web graph, and have been widely used as measures of relative importance of web pages. The well-known PageRank algorithm, for instance, considers a link from a source page to another as an explicit endorsement of the destination page in perceived page quality, and uses only the static link structure of the web as input. Based on the assumption of a random surfer model and the first-order Markov process, PageRank computes the stationary probability distribution for the web link graph iteratively, resulting in the world's largest matrix computation [148].

A number of problems are commonly associated with link analysis algorithms. First, intent drives user browsing behaviors, which significantly deviate from the random surfer model upon which PageRank is based. A recent study on real network traffic [149] demonstrated that user visitation patterns differ considerably from that approximated by the uniform surfing behavior model used in PageRank. Second, static modeling of the link structure favors old pages, because a recent page is less likely to be linked within a short period of time, even if the page is of good quality. Third, link structures are prone to manipulation as adversarial links can be generated to inflate ranking artificially compared to good links that typically originate in manual editing. Last, as the web grows at an explosive speed[1], computing page importance at the web scale by link analysis becomes computationally expensive [151], even through various optimization schemes [152, 153].

Minimal study has been accomplished on general non-search browsing data. Prior literature related to sessions [154, 131, 155, 156, 157] focuses almost exclusively on search trails within query sessions. However, as we will present in Section 5.3, search-related activities account for less than 5% of overall user activity online. Analysis of web sessions in a general setting broadens the user behavior models with richer contextual information from the entire spectrum of actions, and the analysis is key to new web search applications that aim to provide enriched user search experience centered around users' interests.

A new page importance ranking algorithm called BrowseRank [141] has recently been proposed, and it makes two significant contributions. First, it uses the more reliable input of user behavior data, computing a user browsing graph, rather than a web link graph. Second, BrowseRank models the random walk on the user

[1] While the first Google index in 1998 had 26 million pages, this number officially reached 1 trillion mark as of July 25, 2008 [150].

browsing graph by a continuous-time Markov process. BrowseRank has shown better ranking performance compared with link analysis algorithms at the expense of higher computational costs.

Study on search trails within query sessions is the subject of a few recent works. Dramatic differences in user interaction behaviors with a search engine are reported in [154]. The idea of using popular end points in search trails as query-dependent feature is studied in [131] to improve web search interaction. A recent study [155] shows improved retrieval quality using post-search browsing activities over alternative data sources that contain only the end points of search trails or clickthrough logs. Also, the study suggests that post-search browsing behavior logs provide strong signal for inferring document relevance for future queries.

5.3 Mining Web Sessions

We define a web session as a logical unit of time-ordered user browsing activities, representing a single span of user interactions with a web browser. The concept of session in our study is generalized to all categories of web activities, while studies related to search log or search clickthrough data consider a session simply as a set of search queries and largely ignore all other activity.

5.3.1 Session Identification

A user's browsing history is commonly accessible from several sources, such as the ISP or other gateway to the web [149] or clients installed on the user's environment [154]. In this study, we use information logged by the Yahoo! Toolbar, a browser add-on that assists users with quick access to various web tasks. The toolbar logs user activities for a subset of users who opted for this data collection during installation.

Each log entry is a tuple of {*cookie*, *timestamp*, *URL*, *referral URL*, *event attribute list*}. The cookie serves as a unique, anonymous client identifier that expires and refreshes after a pre-defined time period. The URL identifies the page being accessed, and the referral URL is the URL from which the user access the current URL. The event attribute list comprises various metadata associated with the activity. For the experiments in this dissertation, the browsing data consists of more than 30 billion anonymous events, across millions of unique Yahoo! users, collected over six months in 2008.

To segment web activities into sessions, we first use the referral URL → current URL structure to reconstruct the entire chain of browsing activities per user. This scheme ensures that, for users who are multitasking (*e.g.*, those having multiple browser windows or tabs open), we group activities associated with different tasks into separate sessions rather than interleaving them together. Next, we partition the time-ordered user events using two boundary conditions. First, we start a new session from the current event if more than 30 minutes of inactivity occured between the current event and its immediately preceding event. Second, a new session starts if the current event entry does not have a referral URL. This typically happens when the user launches a new web browser, or clicks on a link in a non-browser source (*e.g.*, in a text file).

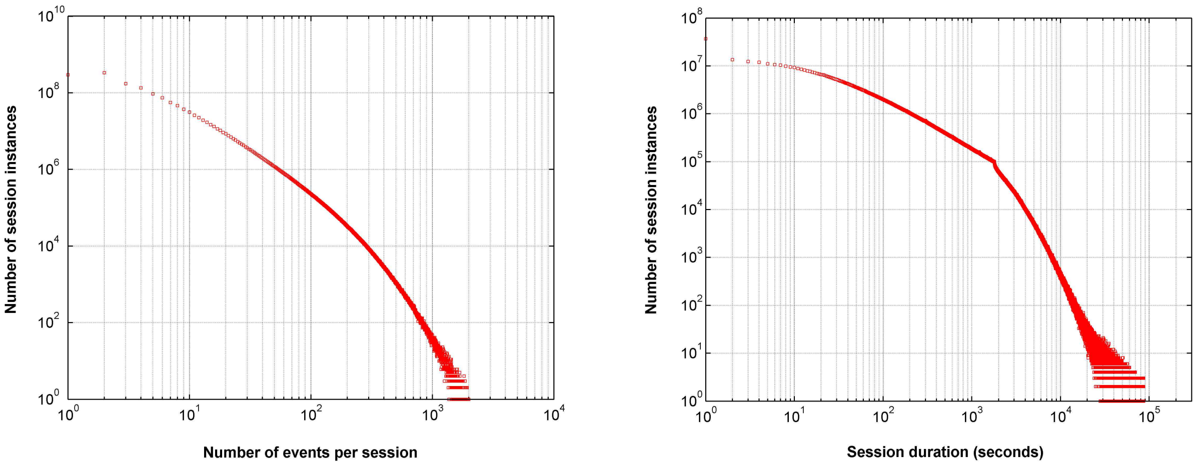


Figure 5.1: Distribution of session lengths (left) and session durations (right) in general web user behavior logs.

Our session segmentation approach requires only one-pass scanning over the data. This may seem a simple mechanism. However, a recent study on finding logical sessions from query logs [156] has shown that in the vast majority (92%) of cases, a session segmentation method based on timeout threshold gives identical scores to an advanced and computationally expensive algorithm [156], when both are compared with human judged sessions using the objective Rand index [158]. For the small fraction of remaining sessions that are difficult for the advanced algorithm, the timeout-based method performs at a merely marginal degradation of 1.4%.

5.3.2 Session Characteristics

Table 5.1 summarizes the key characteristics of general web sessions. Figure 5.1 shows the probability distributions of the number of events in a session and session duration, respectively. The number of events in a web session approximately follows a power law distribution. Its mean and standard derivation are 9.1 and 24.5,

Table 5.1: Key characteristics of general web sessions.

Average events per session	9.1
Standard deviation of events per session	24.5
Average session duration (seconds)	420.3
Standard deviation of session duration (seconds)	1068.0
Sessions per user per day	15.5
Percentage of search sessions	4.85%

respectively, demonstrating that a general web session contains significantly richer activity context and diversity than a search session, which reportedly consists of an average of five events [155]. In addition, search sessions (those containing at least one query sent to one of the major web search engines) constituted 4.85% of overall sessions, signaling that focusing on them may lead to a biased view in downstream analysis [159]. The session duration graph in Figure 5.1 shows two different power law behaviors across the timeout threshold of 1,800 seconds. On average, a web session lasts 420.3 seconds, with the standard deviation of 1,068 seconds, demonstrating its short-to-medium time range coverage of user activities.

5.3.3 Session Clustering

Mining user sessions at the web scale is particularly important for learning and recognizing user behavior patterns associated with structured intents. We employ several clustering approaches to discover web sessions driven by different intents and to learn their statistical characteristics. Due to space constraint, we focus our discussion on one representative clustering effort.

In this experiment, we mapped each interpretable URL to one of five intents categories—search, email, information/reference, rich content (*e.g.*, social networking and multimedia), and shopping. We computed the histogram representation of a session by the distribution of number of events over these intent categories. While certain temporal information is clearly discarded, we will see in the next section that this histogram representation preserves adequate discriminating power for the clustering purpose and remains compact for the large amount of data.

To associate a visit reliably to each URL with an intent type, we used human categorizations of the top 1,200 most popular web sites. While the coverage achieved this way was reasonable at 41% for all events, we augmented these categorizations using heuristics that map from URLs to likely intents. For example, URLs of the format `shopping.*.com/*` were mapped to shopping intent, and so on.

Within each session, a browsing event was labeled either as unknown, or assigned to one of these five intent categories described above. We then computed the distribution of events over the six intent labels (*i.e.*, including the unknown class), and discarded those sessions that contained more than 80% of unknown events, as they could not be reliably clustered. Finally, we smoothed each normalized intent histogram by evenly distributing the weight associated with the unknown class to the other five histogram bins.

The final session histogram is a seven dimensional feature vector. The first five dimensions correspond to the normalized intent histogram, with their sum equal to 100. The last two dimensions correspond to the number of events in the session and the session duration in seconds, respectively.

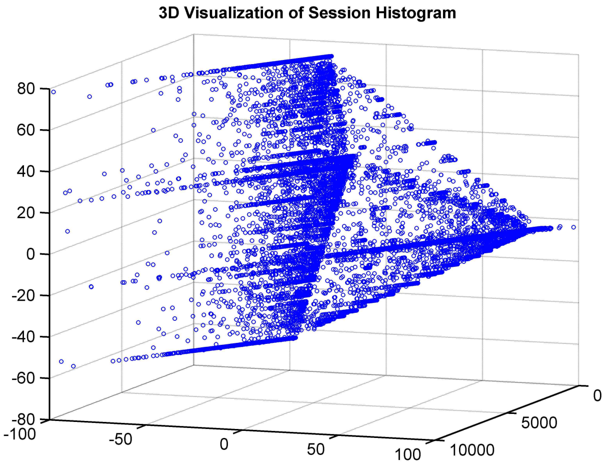


Figure 5.2: Visualization of session histograms in 3D by dimensionality reduction using principle component analysis.

To gain further insights on the spread in session histograms, we used principle component analysis (PCA) to reduce the dimensionality. PCA seeks projections onto a low-dimensional linear subspace that best preserves the data scatter in a least-squares sense [113]. The 3D view of session histogram shown in Figure 5.2 demonstrates the heterogeneity as the histogram data covers a broad continuum of activity space. Among the first six significant eigenvalues, the first eigenvalue is dominant.

5.3.4 Session Interpretation

A meaningful interpretation to sessions offers the key to understanding the context of activities on general, unconstrained user behavior data. Table 5.2 summarizes the unsupervised session histogram clustering results, using k-means algorithm with $k = 10$. These clusters are ordered according to the cluster size. Significant features that clearly indicate cluster attributes in Table 5.2 are highlighted.

Table 5.2: Unsupervised clustering of session histograms reveals web user browsing patterns. Significant features associated with each cluster are highlighted in bold.

Feature	Entire Data	1	2	3	4	5	6	7	8	9	10
Dimension	Mean / Std. Dev.	29.8%	16.6%	14.3%	11.9%	11.0%	4.7%	4.6%	3.5%	2.1%	1.5%
Search	23.63 / 37.71	0.35	**98.43**	1.19	2.35	2.33	**56.18**	**41.52**	**52.23**	6.46	0.09
Mail	16.81 / 34.98	0.07	0.66	**97.25**	0.39	0.42	1.29	**51.79**	0.70	9.79	0.07
Information	12.26 / 30.85	0.04	0.27	0.39	1.03	**96.50**	**24.58**	2.65	0.50	5.97	0.02
Content	34.31 / 45.69	**99.42**	0.37	0.64	0.45	0.36	0.64	0.95	**45.25**	**60.51**	**99.54**
Shopping	12.85 / 31.60	0.08	0.24	0.41	**95.67**	0.29	**16.92**	2.60	0.86	16.84	0.06
Events	9.06 / 24.53	11.14	2.89	5.66	6.25	5.33	4.24	5.38	4.26	7.84	**151.68**
Duration	420.30 / 1067.99	532.49	261.4	303.85	235.78	298.91	228.40	455.58	218.01	439.78	**4237.65**

Various intent-driven web browsing patterns emerge from clusters' statistical properties. The top five clusters correspond to coherent sessions of rich content browsing, search, email, shopping, and information, respectively. For instance, the center of cluster 1, with 29.8% of entire data, contains 99.42% rich content browsing. Typically, these are user interactions with social networking sites such as Facebook and MySpace. Its cluster-wise standard deviation of 2.82% along this feature dimension is significantly smaller than the standard deviation of 45.69% for the entire data.

Clusters revealing more sophisticated user behaviors are also evident in Table 5.2. These interesting patterns include browsing web search results without a click (cluster 2), collecting information during shopping (cluster 6), visits to rich content web site through navigational queries (cluster 8), and prolonged activities in social networking sites (cluster 10, note the average session duration).

These observations demonstrate that even a simple approach to session representation — as distributions over high-level event categories — can provide the search engine with valuable information, such as the distribution over the types of content that users are likely to access (useful for crawling scheduling as well as for ranking purposes). If we apply a filter to the entire set of sessions and preserve only those containing search queries, we can further observe which queries lead to a particular session type (e.g., a shopping session) and optimize the user experience accordingly.

5.4 Using Browsing Information for Web Search

In this section we present a novel web search ranking algorithm, ClickRank, that combines different notions of user preferences mined from browsing sessions. The ClickRank algorithm provides a robust estimate of the importance of web pages and web sites without explicitly constructing a web graph. Its relatively low computational cost makes it particularly useful for web search ranking purposes. We also describe how ClickRank can be incorporated with a large set of other ranking features for learning a ranking model.

5.4.1 ClickRank

A web session contains several contextual indicators of user preferences among the visited web pages. Intuitively, users tend to browse content that they perceive as important in the context of their informational need. This makes the dwell time on a web page an important endorsement of the user's interest level. The click order within a general trail of user activities is also important: accessing one web page before another in the session may be interpreted as the user's preference signal. ClickRank aims to combine these signals to determine a local importance value for each page within a session, and then aggregate the importance values over all sessions of interest.

We start by computing local importance values within each session. The ClickRank of a web page p_i in a given web session s_j is a function of several indicators within the session context—the dwell time on the page, the page load time, the rank of p_i in the ordered set of all visited URLs, and the frequency of occurrence in the session. We define the local ClickRank function as

$$ClickRank(p_i, s_j) = \sum_{p_i \in s_j} w_r(i, s_j) w_t(p, s_j) I(p = p_i), \tag{5.1}$$

where $w_r(i, s_j)$ is a weight function induced on the rank of the event i in session s_j, and $w_t(p, s_j)$ is a weight function computed from the set of temporal attributes associated with the browsing of page p_i. $I()$ denotes the indicator function.

We define the weight function $w_r()$ for an event i in rank $r(i)$ of a session s_j with a total of n_j events as

$$w_r(i, s_j) = \frac{2\,(n_j + 1 - r(i))}{n_j(n_j + 1)}, \tag{5.2}$$

where $r(i) \in \{1, \ldots, n_j\}$ and $w_r(i, s_j)$ is a monotonically decreasing function w.r.t. the rank of the event within a session i. The function choice for $w_r()$ is motivated by measurements of implicit user preference judgements through eye tracking experiments [2], which show decreasing relative attention devoted to ordered clicks in navigational and informational tasks.

Lemma 5.1. *The weight function $w_r()$ has the following properties:*

(a) $w_r(i,s)$ *is a monotonically decreasing function w.r.t. the rank* $r(i)$ $(1 \leq r(i) \leq n)$ *of the event in session* s*, and* $w_r(i-1,s) - w_r(i,s) = w_r(i,s) - w_r(i+1,s)$*,*

(b) $\sum_{i=1}^{n} w_r(i,s) = 1$*, where* n *is the number of events in session* s*.*

For a set of web sessions $\mathcal{S} = (s_1, \ldots, s_k)$ across users and over a period of time, we aggregate the ClickRank values as

$$ClickRank(p,\mathcal{S}) = AGGR_{s\in\mathcal{S}}\ [ClickRank(p,s)], \tag{5.3}$$

where $ClickRank(p,s)$ is the local ClickRank function defined in (5.1) given an instance of observed sessions, and $AGGR$ denotes an aggregation function, such as summation or averaging, over all sessions of interest. In the following experiments, we use summation as the aggregation function.

Finally, the ClickRank of a web site w for a set of sessions $\mathcal{S}$ is simply the sum of the ClickRank values of all pages in $\mathcal{S}$ that are part of the site: $ClickRank(w,\mathcal{S}) = \sum_{p\in w} ClickRank(p,\mathcal{S})$. Note that using a sum implicitly models both the importance (as evidenced by ClickRank values of individual pages) and the size of the web site – the amount of pages it comprises.

5.4.2 Theoretical Analysis

The formulation of ClickRank has a theoretical interpretation based on an intentional surfer model. A web session can be viewed as a logical sequence of hops through the hyperlink structure of the web. At each step, a user selects what she judges as most relevant as the next click, based on a variety of features such as the attractiveness of content in the context of the user's activity and her prior knowledge. The user further indicates her interest through various temporal attributes, such as the time devoted to the page or whether it was visited multiple times. This process continues throughout the duration of the session, until the user starts another journey on the web.

More concretely, the local ClickRank function defines a random variable $X_j : \Omega \rightarrow \mathbb{R}_0^+$ associated with the web page p_j in the event of a logical sequence of web clicks. The mean and variance of the random variable X_j are non-negative and finite.

Lemma 5.2. *For any bounded, non-negative weight function $w_t()$, $E(X_j) < \infty$ and $var(X_j) < \infty$.*

Proof. Let $w_t()$ be bounded by a non-negative constant A. Then,

$$\begin{aligned}
E(X_j) &= \sum_{p_j \in s} w_r(i,s) w_t(p,s) P(p = p_j, s) \\
&= \sum_{p_j \in s} w_r(i,s_j) w_t(p,s) P(s) P(p = p_j | s) \\
&= \sum_{p_j \in s} w_r(i,s_j) w_t(p,s) P(s) \\
&\leq \sum_{p_j \in s} w_r(i,s_j) A \\
&\leq A \sum_{p_j \in s} w_r(i,s_j) \\
&\leq A
\end{aligned}$$

by part (b) of Lemma 5.1. Similarly, we can show that $var(X_j) \leq A^2$. □

The following convergence property of ClickRank defines its asymptotic behavior over increasing volume of empirical data. Furthermore, we can establish bounds on a ClickRank-induced function in a probabilistic setting by Markov's inequality and its corollaries [160].

Theorem 5.1 (Convergence property)**.** *Let* $\{X_j^1,\ X_j^2,\ \ldots,\ X_j^k\}$ *be the sequence of random variables associated with the web page* p_j *over observed sessions* $\mathcal{S} = (s_1, \ldots, s_k)$*, and assume they are independent and identically distributed. Then*

$$\frac{1}{k}\sum_{i=1}^{k} X_j^i \to E(X_j) \text{ almost surely, as } k \to \infty.$$

Proof. X_j^i are non-negative, independent identically distributed random variables with finite means. The result follows directly from the strong law of large numbers [161]. □

Theorem 5.2 (Markov's inequality)**.** *Let* $f : \mathbb{R} \to [0, +\infty)$ *be a non-negative function, then*

$$\mathbb{P}[f(X) \geq a] \leq \frac{E(f(X))}{a} \quad \text{for all } a > 0.$$

Corollary 5.3. *If* $f : \mathbb{R} \to [0, +\infty)$ *is a non-negative function taking values bounded by some number* M*, then*

$$\mathbb{P}[f(X) \geq a] \geq \frac{E(f(X)) - a}{M - a} \quad \text{whenever } 0 \leq a < M.$$

Corollary 5.4 (Chebyshev's inequality)**.** *Let* X *be a random variable with expected value* μ *and finite variance* σ^2*, then*

$$\mathbb{P}[|X - \mu| \geq a] \leq \frac{\sigma^2}{a^2} \quad \text{if } a > 0.$$

Simply put, as the volume of the web browsing sessions analyzed by ClickRank reaches a sizable sample of the entire web traffic, the rank computed by ClickRank for each page converges to its true rank according to a usage criterion.

5.4.3 Application to Web Search Ranking

As a query-independent feature, ClickRank can be incorporated into a document ranking process in several ways [162]. One particular framework that has recently become prominent is the *learning to rank* approach to information retrieval, which aims to apply machine-learning algorithms to derive a ranking function from data. In a machine-learned ranking framework, a large variety of features are used to model a query and a document. Query features can be its length or frequency in a search log, and document features can be term statistics or, in the case of web documents, the number of incoming HTML links. Machine-learned ranking provides a convenient approach for quantitatively evaluating the effectiveness of ClickRank as a novel feature in addition to a large collection of existing ranking features.

We formulate the task of learning the ranking model for web search as an optimization problem. Our goal is to find a ranking function $f^*(\mathbf{x})$ that maps a set of input random variables corresponding to features $\mathbf{x} = \{x_1, \ldots, x_n\}$ to an output random variable y representing the relevance score, such that the expected value of the loss function $\Psi(y, f(\mathbf{x}))$ is minimized over the joint distribution of $(y, \mathbf{x})$

$$\begin{aligned} f^*(\mathbf{x}) &\equiv \arg\min_{f(\mathbf{x})} E_{y,\mathbf{x}} \Psi(y, f(\mathbf{x})) && (5.4) \\ &= \arg\min_{f(\mathbf{x})} E_{\mathbf{x}}[E_y(\Psi(y, f(\mathbf{x})))|\mathbf{x}]. && (5.5) \end{aligned}$$

We compute the optimal ranking model using the numerical optimization framework of gradient boosting [163], since analytical solution cannot be derived generally for $f(\mathbf{x})$ and Ψ in Equation 5.4. Gradient boosting employs functional regression that expresses the solution to the ranking function as additive expansion of M parameterized functions

$$f^*(\mathbf{x}) = \sum_{i=0}^{M} f_m(\mathbf{x}) \equiv \sum_{i=0}^{M} \beta_m h(\mathbf{x}; \mathbf{a}_m), \tag{5.6}$$

where $f_0(\mathbf{x})$ is an initial guess, and $[f_m(\mathbf{x})]_1^M$ are incremental functions (or "boosts"). In Equation(5.6), each incremental function $f_m(\mathbf{x})$ can be further factored as the product of a base learner $h(\mathbf{x}; \mathbf{a}_m)$ and corresponding coefficient β_m.

In parameter estimation, gradient boosting sequentially fits a parameterized function to current residuals by least-squares criterion at each iteration

$$y_{im} = -\left[\frac{\partial\Psi(y_i, f(\mathbf{x}_i))}{\partial\Psi(f(\mathbf{x}_i))}\right]_{f(\mathbf{x})=f_{m-1}(\mathbf{x})} \tag{5.7}$$

and

$$\mathbf{a}_m = \arg\min_{\mathbf{a},\beta} \sum_{i=1}^{N} [y_{im} - \beta h(\mathbf{x}_i; \mathbf{a})^2], \tag{5.8}$$

where N is the number of training samples. The optimal coefficient β_m is computed by line search

$$\beta_m = \arg\min_{\beta} \sum_{i=1}^{N} \Psi\left(y_i, f_{m-1}(\mathbf{x}_i) + \beta h(\mathbf{x}_i; \mathbf{a}_m)\right). \tag{5.9}$$

We use a decision trees as the base learner $h(\mathbf{x}; \mathbf{a}_m)$ in (5.6), where it is parameterized by the splitting variables and corresponding split points. At each iteration m, a decision tree partitions the entire feature space into disjoint regions $[R_{lm}]_{l=1}^{L}$ and predicts according to the region that contains the observed feature vector $\mathbf{x}$ as

$$h(\mathbf{x}; [R_{lm}]_1^L) = \sum_{i=1}^{L} y_{lm} I(\mathbf{x} \in R_{lm}). \tag{5.10}$$

Since a decision tree predicts a constant value y_{lm} within each region R_{lm}, we

Algorithm 5.1 Gradient Tree Boosting

```
$f_0(\mathbf{x}) = \arg\min_\gamma \sum_{i=1}^{M} \Psi(y_i, \gamma)$
for $m = 1$ to $M$ do
    $y_{im} = -\left[\frac{\partial\Psi(y_i, f(\mathbf{x}_i))}{\partial\Psi(f(\mathbf{x}_i))}\right]_{f(\mathbf{x})=f_{m-1}(\mathbf{x})}, i = 1, N$
    $\{R_{lm}\}_1^L =$ L-terminal node tree $\left(\{y_{im}, \mathbf{x}_i\}_1^N\right)$
    $\gamma_{lm} = \arg\min_\gamma \sum_{\mathbf{x}_i \in R_{lm}} \Psi\left(y_i, f_{m-1}(\mathbf{x}_i) + \gamma\right)$
    $f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu \cdot \sum_{l=1}^{L} \gamma_{lm}\mathbf{I}(\mathbf{x} \in R_{lm})$
end for
```

can rewrite (5.9) as

$$\gamma_{lm} = \arg\min_{\gamma} \sum_{\mathbf{x}_i \in R_{lm}} \Psi\left(y_i, f_{m-1}(\mathbf{x}_i) + \gamma\right). \tag{5.11}$$

We add a regularization term, the shrinkage parameter $0 < \nu < 1$ to control the learning rate of the update procedure

$$f_m(\mathbf{x}) = f_{m-1}(\mathbf{x}) + \nu \cdot \gamma_{lm} I(\mathbf{x} \in R_{lm}). \tag{5.12}$$

We use Algorithm 5.1 to estimate ranking model parameters for decision trees with L-terminal nodes. Gradient boosted decision trees (GDBT) produce competitive, highly robust, interpretable procedures in regression and classification [163], and are particularly useful for settings with large amounts of data and a dense feature space.

5.4.4 Relation to Graph-based Models

ClickRank has a number of advantages compared to approaches that estimate the web page authority from explicit graph formulations, such as PageRank and BrowseRank. First, ClickRank is data driven and does not embed assumptions on the traversing scheme over the web. Second, it is significantly more computationally efficient: local ClickRank values are inexpensive to calculate and can be derived independently for each session. This makes ClickRank well-suited to distributed computing (*e.g.*, the MapReduce framework [164] that we used to implement the experiments in this dissertation). It is also memory friendly. Furthermore, addition of new data requires only incremental computation of local ClickRank values on the newly logged web sessions and combination with those from existing sessions, rather than re-computation of the entire model (such as would be needed by PageRank and BrowseRank). This is particularly important for the processing of web-scale user browsing information, which changes constantly.

5.5 Experiments

We demonstrate the effectiveness of ClickRank algorithm in three core aspects of web search—site ranking, page ranking, and mining new, popular web pages. In our experiments, we assume that the dwell time on a page and the page load time are two independent random processes and define the temporal weight function in (5.1) as

$$w_t(p, s) = (1 - e^{-\lambda_1 t_d}) e^{-\lambda_2 t_l} I\left(t(p) \in \mathcal{T}\right), \tag{5.13}$$

where t_d and t_l are the normalized dwell time on the page, and page load time w.r.t. the entire session. $t(p)$ is the timestamp of the event, and $\mathcal{T}$ denotes the time range.

In the following experiments, we used the same six months of aggregate user browsing logs collected from the Yahoo! toolbar. In total, the data comprises of

Table 5.3: Top-ranked sites computed by different algorithms.

Rank	PageRank	BrowseRank	ClickRank
1	adobe.com	myspace.com	yahoo.com
2	wordpress.com	msn.com	google.com
3	w3.org	yahoo.com	myspace.com
4	miibeian.gov.cn	youtube.com	live.com
5	statcounter.com	live.com	youtube.com
6	phpbb.com	facebook.com	facebook.com
7	baidu.com	google.com	msn.com
8	php.net	ebay.com	friendster.com
9	microsoft.com	hi5.com	pogo.com
10	mysql.com	bebo.com	aol.com
11	mapquest.com	orkut.com	microsoft.com
12	cnn.com	aol.com	wikipedia.org
13	google.com	friendster.com	ebay.com
14	blogger.com	craigslist.org	craigslist.org
15	paypal.com	google.co.th	hi5.com
16	macromedia.com	microsoft.com	go.com
17	jalbum.net	comcast.net	ask.com
18	nytimes.com	wikipedia.org	google.co.th
19	simplemachines.org	pogo.com	comcast.net
20	yahoo.com	photobucket.com	orkut.com

more than 3.3 billion web sessions. These sessions contain 16.3 million unique web sites, and 3.1 billion unique web pages.

5.5.1 Site Ranking

We computed the ClickRank for each web site and ordered them by this value. We list the top-ranking 20 sites computed with ClickRank and compare them to those computed by PageRank[2] and BrowseRank[3]. Results are listed in Table 5.3, following the same convention used in [141].

On the task of site ranking, our results confirm the same finding reported in [141], which states that link analysis algorithms like PageRank have a strong bias

[2] Using the web link graph as constructed by the Yahoo! crawler.

[3] This list is included from the reported list in [141] on a total of 5.6 million web sites.

toward sites with higher degree of inlinks and do not necessarily reflect the degree of actual usage. This is a fundamental limitation of the web link graph, from which PageRank and other link-based authority estimation algorithms are derived.

The computed site ranked lists by both ClickRank and BrowseRank algorithms are surprisingly similar, with a total of 18 overlapping entries among the top 20 sites. Both ranked lists correlate better with web users' informational need compared to PageRank, as they are both computed with user behavior data. Some ranking differences between BrowseRank and ClickRank in this table can be attributed to their data source. BrowseRank is computed with a set of users who installed the Live toolbar, and are presumably users of live.com and msn.com services; similarly, ClickRank is computed with a set of Yahoo! users.

One key difference between the results produced by ClickRank and BrowseRank is that ClickRank consistently ranks the starting point of user's web experience higher. One of the major search engines, `ask.com`, does not even appear among top 20 sites produced by BrowseRank.

ClickRank has a significantly lower computational cost than PageRank or BrowseRank. ClickRank requires only one pass through the data and does not require building intermediate graphs and solving stationary probability distributions. This allows for rapid adaptation of ClickRank values to new content: as noted earlier, new browsing information that is collected does not require recomputation over the entire data. The overall running time of our implementation of ClickRank algorithm in ranking of the 16.3 million web sites in this section and 3.1 billion web pages for the page ranking test in the next section are 56 minutes and 1 hour 32

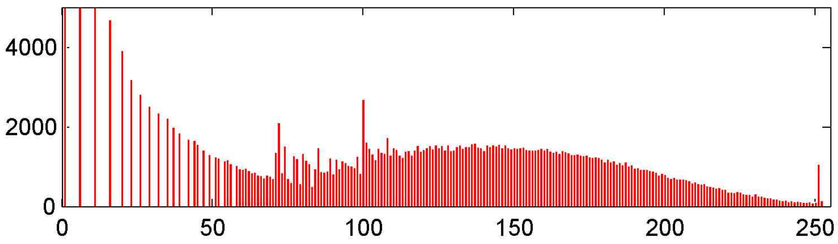


Figure 5.3: Distribution of discretized ClickRank scores over a large collection of judged documents.

minutes, respectively, using the map-reduce framework on 300 Hadoop [165] nodes. To our best knowledge, these are the best published run times for page importance ranking on a web scale.

In a realistic, production-grade search engine environment, it is important to minimize the footprint of every relevance feature used by the ranking model so latency and memory requirements are met. Often, float numeric values are compressed or discretized into a small dynamic range that can be represented with as few bits as possible. To this end, and to evaluate the ranking performance of ClickRank as deployed in a production system, we quantize the computed ClickRank score for each web page into an unsigned byte within the range of $[0, \ldots, 255]$. The distribution of these values are shown in Figure 5.3.

5.5.2 Page Ranking

5.5.2.1 Evaluation Methodology

We comprehensively evaluated the performance of ClickRank in conjunction with several hundred features used in commercial search engines. To gain further insights, we quantified the search improvement from ClickRank with a state-of-the-art baseline system, and measured its relative variable significance against this large pool of ranking features. This evaluation scheme gives more realistic, quantitative results, in contrast to common published evaluations, using limited feature set as baseline. For instance, [141] employs the single feature of BM25 [166] as the relevance baseline in their comparison.

We used discounted cumulative gain (DCG) and normalized discounted cumulative gain (NDCG), two widely used search engine relevance measures [167], to quantitatively evaluate ranking performances. Given a query and the ranked list of returned documents, the $DCG(K)$ score for the query is computed as

$$DCG(K) = \sum_{k=1}^{K} \frac{g_k}{\log_2(1+k)}, \tag{5.14}$$

where g_k is the weight for the document at rank k. A five-grade score is assigned to each document based on its degree of relevance.

We trained ranking models using gradient boosted decision trees on the baseline system with all existing features, and on the alternative system that includes one additional ClickRank feature, respectively. Training and test data is partitioned through cross-validation. We used identical parameter settings in all the following comparison experiments. Table 5.4 provides a high-level summary of this experiment.

To quantify the relative importance S_i of each ranking feature x_i, we used the following measure of variable importance for decision trees [168]

$$S_i^2 = \frac{1}{M} \sum_{m=1}^{M} \sum_{n=1}^{L-1} \frac{w_l w_r}{w_l + w_r} (\overline{y}_l - \overline{y}_r)^2 I(v_t = i), \tag{5.15}$$

Table 5.4: Summary of the page ranking experiment.

Total number of queries	9.041
Number of affected queries	7,341
Total number of judged documents	429,985
Number of documents with ClickRank scores	268,929
Cross-validation split ratio	3

where v_t is the splitting variable at the non-terminal node n, $\overline{y}_l$, $\overline{y}_r$ are the means of the regression responses in the left and right subtrees respectively, and w_l, w_r are the corresponding sums of the weights.

5.5.2.2 Data Preparation

We used a set of 9,041 randomly sampled queries from a search log. For each query, 5–20 web pages have been independently judged by a panel of editors and assigned with one of the five relevance scores.

5.5.2.3 Results and Discussion

The usage of ClickRank as an additional relevance feature brings 1.02%, 0.97%, 1.11%, and 1.331% web search improvements in $DCG(1)$, $DCG(5)$, $DCG(10)$, and $NDCG$, respectively, on top of a state-of-the-art ranking model—a model already incorporating hundreds of features derived from content (*e.g.*, anchor, title, body, and section), from the link structure of the web, from search engine query logs, and from other sources. The reported gains are statistically significant.

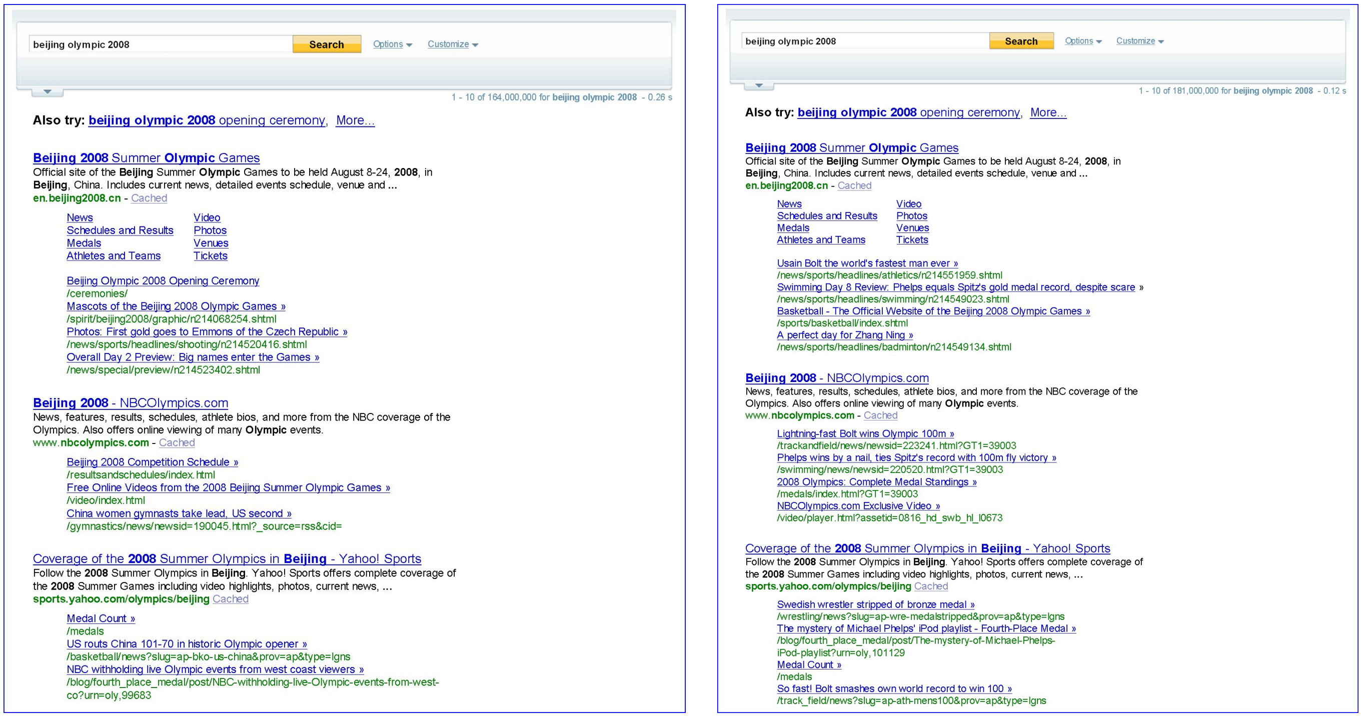
The gains in retrieval performance from ClickRank on top of a competitive

Table 5.5: ClickRank delivers statistically significant improvements over a state-of-the-art baseline.

Query Length	Number of Queries	Affected Queries	Improvements in $DCG(1)$	$DCG(5)$	$DCG(10)$	$NDCG$	Significance Test p-value
1	1484	1232	0.448%	0.713%	1.002%	0.378%	5.33×10^{-2}
2	2992	2450	0.560%	0.993%	1.121%	1.071%	4.65×10^{-4}
3	2153	1722	1.618%	1.076%	1.406%	2.177%	1.10×10^{-4}
4+	2412	1937	0.918%	0.861%	0.784%	1.433%	1.61×10^{-5}
All	9041	7341	1.020%	0.966%	1.105%	1.331%	9.98×10^{-5}

search engine are summarized in Table 5.5. These are substantial improvements in the context of commercial web search: our strong baseline incorporates a feature set of several hundred signals tuned over a long period of time. In addition, 81.2% of more than 9,000 queries are affected in the alternative experiment, demonstrating the generality of ClickRank. Furthermore, we observe higher improvements on long queries in Table 5.5, which are typically much more challenging for search engines. We show search improvements across different query lengths in Table 5.5.

We experimented with several variants of ClickRank and observed that it consistently ranks among the top features in variable significance as calculated by Equation (5.15). For example, in the experiment shown in Table 5.5, the ClickRank feature is ranked 25th in variable importance among several hundred other features, significantly higher than the highest-ranking feature derived from page visit counts (ranked 56th) and a feature based on a propagation of authority through the link graph (ranked 108th). These results demonstrate the significance of session-based web importance estimation and show that ClickRank captures novel user preference knowledge not identified through other modeling techniques.



beijing olympic 2008

1 - 10 of 164,000,000 for **beijing olympic 2008** - 0.26 s

Also try: **beijing olympic 2008** opening ceremony, More...

Beijing 2008 Summer **Olympic** Games
Official site of the **Beijing** Summer **Olympic** Games to be held August 8-24, **2008**, in **Beijing**, China. Includes current news, detailed events schedule, venue and **...**
en.beijing2008.cn - Cached

News	Video
Schedules and Results	Photos
Medals	Venues
Athletes and Teams	Tickets

Beijing Olympic 2008 Opening Ceremony
/ceremonies/
Mascots of the Beijing 2008 Olympic Games »
/spirit/beijing2008/graphic/n214068254.shtml
Photos: First gold goes to Emmons of the Czech Republic »
/news/sports/headlines/shooting/n214520416.shtml
Overall Day 2 Preview: Big names enter the Games »
/news/special/preview/n214523402.shtml

Beijing 2008 - NBCOlympics.com
News, features, results, schedules, athlete bios, and more from the NBC coverage of the Olympics. Also offers online viewing of many **Olympic** events.
www.**nbcolympics.com** - Cached

Beijing 2008 Competition Schedule »
/resultsandschedules/index.html
Free Online Videos from the 2008 Beijing Summer Olympic Games »
/video/index.html
China women gymnasts take lead, US second »
/gymnastics/news/newsid=190045.html?_source=rss&cid=

Coverage of the **2008** Summer Olympics in **Beijing** - Yahoo! Sports
Follow the **2008** Summer Olympics in **Beijing**. Yahoo! Sports offers complete coverage of the **2008** Summer Games including video highlights, photos, current news, **...**
sports.yahoo.com/olympics/beijing Cached

Medal Count »
/medals
US routs China 101-70 in historic Olympic opener »
/basketball/news?slug=ap-bko-us-china&prov=ap&type=lgns
NBC withholding live Olympic events from west coast viewers »
/blog/fourth_place_medal/post/NBC-withholding-live-Olympic-events-from-west-co?urn=oly,99683

(a) Quicklinks for August 10, 2008.

beijing olympic 2008

1 - 10 of 181,000,000 for **beijing olympic 2008** - 0.12 s

Also try: **beijing olympic 2008** opening ceremony, More...

Beijing 2008 Summer **Olympic** Games
Official site of the **Beijing** Summer **Olympic** Games to be held August 8-24, **2008**, in **Beijing**, China. Includes current news, detailed events schedule, venue and **...**
en.beijing2008.cn - Cached

News	Video
Schedules and Results	Photos
Medals	Venues
Athletes and Teams	Tickets

Usain Bolt the world's fastest man ever »
/news/sports/headlines/athletics/n214551959.shtml
Swimming Day 8 Review: Phelps equals Spitz's gold medal record, despite scare »
/news/sports/headlines/swimming/n214549023.shtml
Basketball - The Official Website of the Beijing 2008 Olympic Games »
/sports/basketball/index.shtml
A perfect day for Zhang Ning »
/news/sports/headlines/badminton/n214549134.shtml

Beijing 2008 - NBCOlympics.com
News, features, results, schedules, athlete bios, and more from the NBC coverage of the Olympics. Also offers online viewing of many **Olympic** events.
www.**nbcolympics.com** - Cached

Lightning-fast Bolt wins Olympic 100m »
/trackandfield/news/newsid=223241.html?GT1=39003
Phelps wins by a nail, ties Spitz's record with 100m fly victory »
/swimming/news/newsid=220520.html?GT1=39003
2008 Olympics: Complete Medal Standings »
/medals/index.html?GT1=39003
NBCOlympics.com Exclusive Video »
/video/player.html?assetid=0816_hd_swb_hl_l0673

Coverage of the **2008** Summer Olympics in **Beijing** - Yahoo! Sports
Follow the **2008** Summer Olympics in **Beijing**. Yahoo! Sports offers complete coverage of the **2008** Summer Games including video highlights, photos, current news, **...**
sports.yahoo.com/olympics/beijing Cached

Swedish wrestler stripped of bronze medal »
/wrestling/news?slug=ap-wre-medalstripped&prov=ap&type=lgns
The mystery of Michael Phelps' iPod playlist - Fourth-Place Medal »
/blog/fourth_place_medal/post/The-mystery-of-Michael-Phelps-iPod-playlist?urn=oly,101129
Medal Count »
/medals
So fast! Bolt smashes own world record to win 100 »
/track_field/news?slug=ap-ath-mens100&prov=ap&type=lgns

(b) Quicklinks for August 16, 2008.

Figure 5.4: Dynamic quicklinks discovered using ClickRank by recency ranking.

5.5.3 Mining Dynamic Quicklinks

Many commercial web search engines supply a set of "quicklinks" – direct access links to certain pages within the site, in addition to the search result itself. Typically, these quicklinks are pointers to frequently visited destinations within the host, mined from query or clickthrough logs. This method, however, has two major limitations. First, query logs do not contain user activities beyond the scope of interactions with search engines, which account for the vast majority (more than 95%, as shown earlier) of real web traffic. Second, results computed from query logs have a strong bias toward old navigational links within the site since they receive more clicks within the visibility range of search engines.

We demonstrate a novel application of ClickRank for discovering and displaying dynamic quicklinks in web search results through recency ranking. The idea is to adapt the time range for the indictor function in Equation (5.13) w.r.t. the content refresh rate found by web crawlers. In addition to normal search results, the system displays highly ranked web pages computed by ClickRank as quicklinks to the user.

Figure 5.4 shows search results with discovered quicklinks by the system in response to the query of "beijing olympic 2008" on two days during the summer Olympic Games in 2008, using the time range of 24 hours. Quicklinks mined by ClickRank are displayed alongside of the most frequently clicked navigational links. The quicklinks effectively capture the event highlights, while the most frequently clicked navigational links remain unchanged. The quicklink results by ClickRank are more meaningful in suggesting content that are of potential interest to web users than those that reflect the structural property of the web site.

5.6 Summary

In this chapter, we expanded the use of general web user browsing information for discovering session models driven by structured user intents, and proposed user preference models that incorporate rich session context for web search ranking. We presented characteristics of general web sessions and revealed interesting user behavior patterns mined from sessions. We introduced *ClickRank*, an efficient, scalable algorithm for estimating web page and web site importance based on user preference judgments mined from session context. ClickRank is based on a data-driven intentional surfer model, and is empirically shown to be an effective and novel ranking feature even on top of a highly competitive baseline system employing hundreds of ranking features. We also discussed the advantages of ClickRank compared to existing importance ranking approaches. ClickRank is efficient to compute, delivering highly competitive ranking results compared to the state-of-the-art models based on web graphs. We also demonstrated a promising application that mines dynamic quicklinks for enhancing web user experience. These promising results on data-driven user behavior modeling highlight the prominent role of mining web user behavior in the understanding of user intents and in next-generation web search.

Chapter 6

Summary of Contributions

This dissertation has addressed two critical challenges in analysis and retrieval of unstructured, broad-domain document collections—recognition of diverse content and the use of broad contextual knowledge. Our key contributions include:

1. Obtaining high-level content interpretation for an image offers an important starting point in many computer vision and image analysis problems. We proposed a novel approach for document image content categorization, using image descriptors constructed from a lexicon of shape features. We encoded local text structures using scale and rotation invariant lexical words and learned a concise, structurally indexed shape lexicon. Our approach is extensible and does not require skew correction, scale normalization, or segmentation. In two challenging problems—content category recognition of diverse web images and language identification of documents with mixed machine printed text and handwriting—our approach's performance exceeded the state of the art.

2. We explored a new direction of detecting and matching evidentiary visual content for document indexing and retrieval. Our study on signature-based document image retrieval addressed two important problems. First, we proposed a novel multi-scale approach to detect and segment signatures jointly

from documents. Our approach captures the structural saliency using a signature production model and is computationally tractable. Second, we treated the problem of signature retrieval in the unconstrained setting of translation, scale, and rotation invariant deformable shape matching. We proposed two novel measures of shape dissimilarity based on anisotropic scaling and registration residual error, and presented a supervised learning framework for combining complementary shape information from different dissimilarity metrics. Our approach demonstrated state-of-the-art performance in the tasks of signature matching and signature verification.

3. We presented a model-based approach for recognizing relevant named entities from unstructured document images by combining rich page layout features in the image space with OCR text. We demonstrated our named entity extraction approach in an automated expense reimbursement system and evaluated its performance on large collections of real receipt images.

4. We proposed a computational framework for incorporating contextual knowledge gained from general web user behavior data for improving ranking and other web search experience, with the objective of constructing aggregate models by analyzing individual user sessions. We introduced ClickRank, an efficient, scalable algorithm for estimating web page and web site importance and laid out its theoretical foundation. ClickRank is shown to contribute significantly as a novel web search feature.

6.1 Possible Extensions

Some suggested additional work has already been discussed at the end of relevant chapters as extensions that follow naturally from the technical content. In this section, we consider topics somewhat further afield from the specific topics in this dissertation but are a continuation of the underlying ideas.

Content Categorization of Printed Text and Handwriting The approach presented in Chapter 2 worked well in identifying the primary script or language of a document in the presence of mixed printed text and handwriting. The abstraction of shape by line segment features and the representation of geometrically invariant descriptors were shown to be effective on diverse text content. In principle, our shape lexicon approach can be naturally extended to the problem of categorizing text and handwriting, given a region of text. We are collaborating with another Ph.D. student on this topic and the preliminary results are promising. We have found that the problem of identifying machine printed text and handwriting becomes extremely challenging when the region contains minimal amount of text. The limited features observed might be too sparse for reliable classification. It is useful to include additional features, such as those derived from the uniformity among text. These cases require further investigation.

Indexing Scheme for Visual Content A versatile indexing scheme for visual content has been a fundamental challenge in content-based image retrieval. Because image data is complex and high dimensional, obtaining finer semantic interpretation for compact indexing often presents a difficult challenge. In Chapter 3, we used a set of feature points extracted from signatures for matching and retrieval, which is a simple form of indexing. From the retrieval point of view, it is helpful to include a blend of other useful information in the index, including the text within the context and image features (*e.g.*, dimensions, location, and color information). A flexible engineering design for visual content indexing facilitates query processing and contributes to the success of content-based image retrieval in practice.

Customized Web Search Using Context The experiments in Chapter 5 demonstrated two important results. First, a high percentage of sessions are associated with structured actions (*e.g.*, checking emails, looking for information, and shopping). Second, using models learned from aggregate user behavior data significantly improves search quality for a large number of queries. The natural extensions would be 1) how to predict a structured intent in the current query, and 2) how to develop models from user behavior data to improve customized search results for the structured intent. The answers to both questions involve learning of models from aggregate user behaviors (for coverage and freshness) and individual behavior history.

Appendix A

Proof of Theorems

Theorem 3.1 Given two intersecting lines $t_1(x, y)$ and $t_2(x, y)$ in the forms of $p_1(x - x_1) + q_1(y - y_1) = 0$ and $p_2(x - x_2) + q_2(y - y_2) = 0$, respectively. Let $L(x, y)$ be the straight line that passes through points (x_1, y_1) and (x_2, y_2). Then, the quadratic equation $C(x, y) \equiv l^2(x, y) - \lambda t_1(x, y)t_2(x, y) = 0$ represents a conic section and the range of λ for which $C(x, y)$ is an ellipse is $0 < \lambda < \lambda_0$, where

$$\lambda_0 = \frac{4[p_1(x_2 - x_1) + q_1(y_2 - y_1)]}{(p_1q_2 - p_2q_1)} \times \frac{[p_2(x_1 - x_2) + q_2(y_1 - y_2)]}{(p_1q_2 - p_2q_1)}.$$

Proof. A quadratic equation $C(x, y)$ in two variables x and y can be written as

$$C(x, y) = ax^2 + 2hxy + by^2 + 2gx + 2fy + c = 0, \tag{A.1}$$

where a, b, c, f, g, h are linear functions in λ and the parameter set (x_1, y_1), (x_2, y_2), (p_1, q_1), (p_2, q_2). We can rewrite equation (A.1) as

$$\mathbf{x}^T\mathbf{A}\mathbf{x} + 2\mathbf{f}^T\mathbf{x} + c = 0, \tag{A.2}$$

where $\mathbf{x} = (x, y)^T$, $\mathbf{f} = (g, f)^T$ and the 2×2 matrix $\mathbf{A}$ is

$$\mathbf{A} = \begin{pmatrix} a & h \\ h & b \end{pmatrix}. \tag{A.3}$$

When $C(x, y)$ is an ellipse with center $\mathbf{x}_0$ at $(x_0, y_0)^T$, (A.2) can be written as

$$(\mathbf{x} - \mathbf{x}_0)^T \mathbf{A} (\mathbf{x} - \mathbf{x}_0) = d, \tag{A.4}$$

i.e.,

$$\mathbf{x}^T\mathbf{A}\mathbf{x} - 2\mathbf{x}_0^T\mathbf{A}\mathbf{x} + \mathbf{x}_0^T\mathbf{A}\mathbf{x}_0 - d = 0. \tag{A.5}$$

Comparing (A.2) with (A.8), we have that

$$\mathbf{A}\mathbf{x}_0 = -f,$$

giving

$$\mathbf{x}_0 = -\mathbf{A}^{-1}\mathbf{f}, \tag{A.6}$$

and

$$d = \mathbf{f}^T\mathbf{A}^{-1}\mathbf{f} - c. \tag{A.7}$$

Equation (A.5) represents a parabola if $\mathbf{A}$ is singular, since the center $\mathbf{x}_0$ is at infinity. It represents an ellipse if $\mathbf{A}$ is positive definite. When $\lambda = 0$, the matrix $\mathbf{A}$ is singular. As λ increases, $\mathbf{A}$ becomes positive definite first, and then becomes singular again at certain positive value λ_0. For λ larger than λ_0, matrix $\mathbf{A}$ is indefinite, and equation (A.5) becomes a hyperbola. Thus, the range of λ_0 for which (A.5) is an ellipse is the interval $0 < \lambda < \lambda_0$. We can find λ_0 by solving

$$det(\mathbf{A}) = \begin{vmatrix} a & h \\ h & b \end{vmatrix} = 0, \tag{A.8}$$

where

$$a(\lambda) = (y_1 - y_2)^2 - \lambda p_1 p_2, \tag{A.9}$$

$$b(\lambda) = (x_2 - x_1)^2 - \lambda q_1 q_2, \tag{A.10}$$

$$h(\lambda) = (y_1 - y_2)(x_2 - x_1) - \frac{1}{2}\lambda(p_1 q_2 + p_2 q_1). \tag{A.11}$$

One root of (A.8) is zero, and the other root is strictly positive. The positive root is the required λ_0 and is given by

$$\lambda_0 = \frac{4[p_1(x_2 - x_1) + q_1(y_2 - y_1)]}{(p_1q_2 - p_2q_1)} \times \frac{[p_2(x_1 - x_2) + q_2(y_1 - y_2)]}{(p_1q_2 - p_2q_1)}. \qquad \square$$

Lemma 5.1 The weight function $w_r()$ has the following properties:

(a) $w_r(i, s)$ is a monotonically decreasing function w.r.t. the rank $r(i)$ $(1 \leq r(i) \leq n)$ of the event in session s, and $w_r(i-1, s) - w_r(i, s) = w_r(i, s) - w_r(i+1, s)$,

(b) $\sum_{i=1}^{n} w_r(i, s) = 1$, where n is the number of events in session s.

Proof. (a) Clearly $w_r(i, s)$ is a monotonically decreasing function because $r(i)$ is strictly increasing, and for $2 \leq i \leq (n-1)$

$$w_r(i-1, s) - w_r(i, s) = w_r(i, s) - w_r(i+1, s) = \frac{2}{n(n+1)}. \tag{A.12}$$

(b)
$$\begin{aligned}\sum_{i=1}^{n} w_r(i, s) &= \sum_{i=1}^{n} \frac{2(n+1-r(i))}{n(n+1)} \\ &= n \cdot \frac{2(n+1)}{n(n+1)} - \frac{2}{n(n+1)} \sum_{i=1}^{n} r(i) \\ &= 2 - \frac{2}{n(n+1)} \frac{n(n+1)}{2} \\ &= 1. \quad \square\end{aligned}$$

Bibliography

[1] D. Doermann, E. Rivlin, and A. Rosenfeld, "The function of documents," *Int. J. Computer Vision*, vol. 16, pp. 799–814, 1998.

[2] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay, "Accurately interpreting clickthrough data as implicit feedback," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005, pp. 154–161.

[3] S. V. Rice, G. Nagy, and T. A. Nartker, *Optical Character Recognition: An Illustrated Guide to the Frontier*. Norwell, Massachusetts: Kluwer Academic Publishers, 1999.

[4] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger, "Multi-scale structural saliency for signature detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[5] G. Zhu and D. Doermann, "Automatic document logo detection," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2007, pp. 864–868.

[6] G. Zhu, S. Jaeger, and D. Doermann, "A robust stamp detection framework on degraded documents," in *Proceedings of the Document Recognition and Retrieval Conference*, vol. 6067, 2006, pp. 1–9.

[7] R. Plamondon and S. N. Srihari, "On-line and off-line handwriting recognition: A comprehensive survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 63–84, 2000.

[8] L. Vincent, "Google Book Search: Document understanding on a massive scale," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2007, pp. 819–823.

[9] G. Zhu, T. J. Bethea, and V. Krishna, "Extracting relevant named entities for automated expense reimbursement," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2007, pp. 1004–1012.

[10] G. Zhu, X. Yu, Y. Li, and D. Doermann, "Language identification for handwritten document images using a shape codebook," *Pattern Recognition*, vol. 42, no. 12, pp. 3184–3191, 2009.

[11] Y. Li, Y. Zheng, D. Doermann, and S. Jaeger, "Script-independent text line segmentation in freestyle handwritten documents," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 8, pp. 1313–1329, 2008.

[12] U. Marti and H. Bunke, "The IAM-database: An English sentence database for offline handwriting recognition," *Int. J. Document Analysis and Recognition*, vol. 5, pp. 39–46, 2006, Available at http://www.iam.unibe.ch/~fki/iamDB/.

[13] J. Hochberg, K. Bowers, M. Cannon, and P. Kelly, "Script and language identification for handwritten document images," *Int. J. Document Analysis and Recognition*, vol. 2, no. 2-3, pp. 45–52, 1999.

[14] X. Li, Y.-Y. Wang, and A. Acero, "Learning query intent from regularized click graphs," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 339–346.

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

[16] T. Tan, "Rotation invariant texture features and their use in automatic script identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 7, pp. 751–756, 1998.

[17] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[18] M. Heikkila and M. Pietikainen, "A texture-based method for modeling the background and detecting moving objects," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 657–662, 2006.

[19] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, 2006.

[20] A. Spitz, "Determination of script and language content of document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 3, pp. 235–245, 1997.

[21] J. Ding, L. Lam, and C. Suen, "Classification of oriental and European scripts by using characteristic features," in *Proceedings of the International Conference on Document Analysis and Recognition*, 1997, pp. 1023–1027.

[22] D.-S. Lee, C. R. Nohl, and H. S. Baird, *Language Identification in Complex, Unoriented, and Degraded Document Images*, ser. Document Analysis Systems II, J. J. Hull and S. L. Taylor, Eds. River Edge, New Jersey: World Scientific Publishing Co. Pte. Ltd, 1998.

[23] C. Suen, S. Bergler, N. Nobile, B. Waked, C. Nadal, and A. Bloch, "Categorizing document images into script and language classes," in *Proceedings of the International Conference on Document Analysis and Recognition*, 1998, pp. 297–306.

[24] S. Lu and C. Tan, "Script and language identification in noisy and degraded document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 14–24, 2008.

[25] A. Busch, W. Boles, and S. Sridharan, "Texture for script identification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1720–1732, 2005.

[26] J. Hochberg, P. Kelly, T. Thomas, and L. Kerns, "Automatic script identification from document images using cluster-based templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 2, pp. 176–181, 1997.

[27] H. Ma and D. Doermann, "Word level script identification on scanned document images," in *Proc. Document Recognition and Retrieval (SPIE)*, 2004, pp. 124–135.

[28] S. Jaeger, H. Ma, and D. Doermann, "Identifying script on word-level with informational confidence," in *Proceedings of the International Conference on Document Analysis and Recognition*, 2005, pp. 416–420.

[29] V. Ferrari, L. Fevrier, F. Jurie, and C. Schmid, "Groups of adjacent contour segments for object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 1, pp. 36–51, 2008.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[31] L. Schomaker, M. Bulacu, and K. Franke, "Automatic writer identification using fragmented connected-component contours," in *Proceedings of the International Workshop on Frontiers in Handwriting Recognition*, 2004, pp. 185–190.

[32] T. Kohonen, *Self-Organization and Associative Memory*, 3rd ed. New York, New York: Springer-Verlag New York, Inc., 1989.

[33] J. Canny, "A computational approach to edge detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 8, no. 6, pp. 679–697, 1986.

[34] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 8, pp. 888–905, 2000.

[35] S. Yu and J. Shi, "Multiclass spectral clustering," in *Proceedings of the IEEE International Conference on Computer Vision*, 2003, pp. 11–17.

[36] C.-C. Chang and C.-J. Lin, *LIBSVM: A library for support vector machines*, Online, Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2001.

[37] T. Rath and R. Manmatha, "Word image matching using dynamic time warping," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. 521–527.

[38] J. Chan, C. Ziftci, and D. Forsyth, "Searching off-line Arabic documents," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2006, pp. 1455–1462.

[39] G. Zhu, Y. Zheng, and D. Doermann, "Signature-based document image retrieval," in *Proceedings of the European Conference on Computer Vision*, vol. 3, 2008, pp. 752–765.

[40] K. Guo, D. Doermann, and A. Rosenfeld, "Forgery detection by local correspondence," *Int. J. Pattern Recognition and Artificial Intelligence*, vol. 15, no. 4, pp. 579–641, 2001.

[41] B. Fang, C. Leung, Y. Tang, K. Tse, P. Kwokd, and Y. Wonge, "Off-line signature verification by the tracking of feature and stroke positions," *Pattern Recognition*, vol. 36, no. 1, pp. 91–101, 2003.

[42] D. Doermann and A. Rosenfeld, “Recovery of temporal information from static images of handwriting,” *Int. J. Computer Vision*, vol. 15, no. 1-2, pp. 143–164, 1995.

[43] G. Nagy, “Twenty years of document image analysis in PAMI,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 38–62, 2000.

[44] R. Sabourin, G. Genest, and F. Preteux, “Off-line signature verification by local granulometric size distributions,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 9, pp. 976–988, 1997.

[45] M. Munich and P. Perona, “Continuous dynamic time warping for translation-invariant curve alignment with applications to signature verification,” in *Proceedings of the IEEE International Conference on Computer Vision*, 1999, pp. 108–115.

[46] M. Kalera, S. N. Srihari, and A. Xu, “Off-line signature verification and identification using distance statistics,” *Int. J. Pattern Recognition and Artificial Intelligence*, vol. 18, no. 7, pp. 1339–1360, 2004.

[47] S. Chen and S. N. Srihari, “Use of exterior contours and word shape in off-line signature verification,” in *Proceedings of the International Conference on Document Analysis and Recognition*, 2005, pp. 1280–1284.

[48] A. Shanker and A. Rajagopalan, “Off-line signature verification using DTW,” *Pattern Recognition Letters*, vol. 28, no. 12, pp. 1407–1414, 2007.

[49] K. Siddiqi, A. Shokoufandeh, S. Dickinson, and S. Zucker, “Shock graphs and shape matching,” *Int. J. Computer Vision*, vol. 35, no. 1, pp. 13–32, 1999.

[50] T. Sebastian, P. Klein, and B. Kimia, “Recognition of shapes by editing their shock graphs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 5, pp. 550–571, 2004.

[51] L. Latecki, R. Lakamper, and U. Eckhardt, “Shape descriptors for non-rigid shapes with a single closed contour,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2000, pp. 424–429.

[52] E. Petrakis, A. Diplaros, and E. Milios, “Matching and retrieval of distorted and occluded shapes using dynamic programming,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 11, pp. 1501–1516, 2002.

[53] T. Sebastian, P. Klein, and B. Kimia, “On aligning curves,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 1, pp. 116–124, 2003.

[54] H. Ling and D. Jacobs, “Shape classification using the inner-distance,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 2, pp. 286–299, 2007.

[55] S. Belongie, J. Malik, and J. Puzicha, “Shape matching and object recognition using shape contexts,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 4, pp. 509–522, 2002.

[56] Y. Zheng and D. Doermann, “Robust point matching for non-rigid shapes by preserving local neighborhood structures,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 4, pp. 643–649, 2006.

[57] A. Rosenfeld, R. Hummel, and S. Zucker, “Scene labeling by relaxation operations,” *IEEE Trans. System, Man and Cybernetics*, vol. 6, no. 6, pp. 420–433, 1976.

[58] G. Zhu, Y. Zheng, D. Doermann, and S. Jaeger, “Signature detection and matching for document image retrieval,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 11, pp. 2015–2031, 2009.

[59] A. Sha’ashua and S. Ullman, “Structural saliency: The detection of globally salient structures using a locally connected network,” in *Proceedings of the IEEE International Conference on Computer Vision*, 1988, pp. 321–327.

[60] T. Alter and R. Basri, “Extracting salient curves from images: An analysis of the saliency network,” *Int. J. Computer Vision*, vol. 27, no. 1, pp. 51–69, 1998.

[61] P. Parent and S. Zucker, “Trace inference, curvature consistency, and curve detection,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 8, pp. 823–839, 1989.

[62] G. Guy and G. Medioni, “Inferring global perceptual contours from local features,” *Int. J. Computer Vision*, vol. 20, no. 1-2, pp. 113–133, 1996.

[63] J. Elder and S. Zucker, “The effect of contour closure on the rapid discrimination of two-dimensional shapes,” *Vision Res.*, vol. 33, no. 7, pp. 981–991, 1993.

[64] L. Williams and D. Jacobs, “Stochastic completion fields: A neural model of illusory contour shape and salience,” *Neural Computation*, vol. 9, pp. 849–870, 1997.

[65] L. Williams and K. Thornber, “A comparison of measures for detecting natural shapes in cluttered backgrounds,” *Int. J. Computer Vision*, vol. 33, no. 2-3, pp. 81–96, 2000.

[66] J. Shotton, A. Blake, and R. Cipolla, “Contour-based learning for object detection,” in *Proceedings of the IEEE International Conference on Computer Vision*, 2005, pp. 503–510.

[67] C. Zahn and R. Roskies, “Fourier descriptors for plane closed curves,” *IEEE Trans. Computing*, vol. 21, no. 3, pp. 269–281, 1972.

[68] C. Lin and R. Chellappa, “Classification of partial 2-D shapes using Fourier descriptors,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 9, no. 5, pp. 686–690, 1987.

[69] Y. Lamdan, J. Schwartz, and H. Wolfson, “Object recognition by affine invariant matching,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1988, pp. 335–344.

[70] J. Gorman, R. Mitchell, and F. Kuhl, “Partial shape recognition using dynamic programming,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 2, pp. 257–266, 1988.

[71] D. Sharvit, J. Chan, H. Tek, and B. Kimia, “Symmetry-based indexing of image database,” *J. Visual Commun. and Image Representation*, vol. 9, no. 4, pp. 366–380, 1998.

[72] G. Mori, S. Belongie, and J. Malik, "Efficient shape matching using shape contexts," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 11, pp. 1832–1837, 2005.

[73] G. Borgefors, "Hierarchical chamfer matching: A parametric edge matching algorithm," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 10, no. 6, pp. 849–865, 1988.

[74] D. Huttenlocher, R. Lilien, and C. Olson, "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 9, pp. 850–863, 1993.

[75] S. Loncaric, "A survey of shape analysis techniques," *Pattern Recognition*, vol. 31, no. 8, pp. 983–1001, 1998.

[76] R. Velkamp and M. Hagedoorn, *State of the art in shape matching.* Technical Report UU-CS-1999-27, Utrecht University, Netherlands, 1999.

[77] P. Besl and H. McKay, "A method for registration of 3-D shapes," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 14, no. 2, pp. 239–256, 1992.

[78] Z. Zhang, "Iterative point matching for registration of free-form curves and surfaces," *Int. J. Computer Vision*, vol. 13, no. 2, pp. 119–152, 1994.

[79] J. Feldmar and N. Anyche, "Rigid, affine and locally affine registration of free-form surfaces," *Int. J. Computer Vision*, vol. 18, no. 2, pp. 99–119, 1996.

[80] T. Wakahara and K. Odaka, "Adaptive normalization of handwritten characters using global/local affine transform," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 12, pp. 1332–1341, 1998.

[81] S. Gold, A. Rangarajan, C. Lu, S. Pappu, and E. Mjolsness, "New algorithms for 2-D and 3-D point matching: Pose estimation and correspondence," *Pattern Recognition*, vol. 31, no. 8, pp. 1019–1031, 1998.

[82] H. Chui and A. Rangarajan, "A new point matching algorithm for non-rigid registration," *Computer Vision and Image Understanding*, vol. 89, no. 2-3, pp. 114–141, 2003.

[83] T. Rath, R. Manmatha, and V. Lavrenko, "A search engine for historical manuscript images," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2004, pp. 369–376.

[84] S. N. Srihari, S. Shetty, S. Chen, H. Srinivasan, C. Huang, G. Agam, and O. Frieder, "Document image retrieval using signatures as queries," in *Proceedings of the International Conference on Document Image Analysis for Libraries*, 2006, pp. 198–203.

[85] G. Agam, S. Argamon, O. Frieder, D. Grossman, and D. Lewis, *The Complex Document Image Processing (CDIP) test collection*, Online, Illinois Institute of Technology, 2006, http://ir.iit.edu/projects/CDIP.html.

[86] D. Lewis, G. Agam, S. Argamon, O. Frieder, D. Grossman, and J. Heard, "Building a test collection for complex document information processing," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 665–666.

[87] K. Turkowski, *Filters for common resampling tasks*, A. S. Glassner, Ed. San Diego, California: Academic Press Professional, Inc., 1990.

[88] P. Saint-Marc, H. Rom, and G. Medioni, "B-spline contour representation and symmetry detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 11, pp. 1191–1197, 1993.

[89] J. M. Hollerbach, "An oscillation theory of handwriting," Ph.D. dissertation, Massachusetts Institute of Technology, 1978.

[90] F. Bookstein, "Principle warps: Thin-plate splines and the decomposition of deformations," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 11, no. 6, pp. 567–585, 1989.

[91] C. Buckley and E. Voorhees, "Evaluating evaluation measure stability," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2000, pp. 33–40.

[92] *The Legacy Tobacco Document Library (LTDL)*, Online, University of California, San Francisco, 2007, http://legacy.library.ucsf.edu/.

[93] *ABBYY FineReader SDK 8.0*, ABBYY Corp, 2006, http://www.abbyy.com/.

[94] T. Lindeberg, "Edge detection and ridge detection with automatic scale selection," *Int. J. Computer Vision*, vol. 30, no. 2, pp. 77–116, 1996.

[95] Y. Zheng, H. Li, and D. Doermann, "Machine printed text and handwriting identification in noisy document images," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 3, pp. 337–353, 2004.

[96] C. Dyer and A. Rosenfeld, "Thinning algorithms for gray-scale pictures," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 1, no. 1, pp. 88–89, 1979.

[97] T. Zhang and C. Suen, "A fast parallel algorithm for thinning digital patterns," *Comm. ACM*, vol. 27, no. 3, pp. 236–239, 1984.

[98] R. Sabourin, R. Plamondon, and L. Beaumier, "Structural interpretation of handwritten signature images," *Int. J. Pattern Recognition and Artificial Intelligence*, vol. 8, no. 3, pp. 709–748, 1994.

[99] J. Drouhard, R. Sabourin, and M. Godbout, "A neural approach to off-line signature verification using directional PDF," *Pattern Recognition*, vol. 29, no. 3, pp. 415–424, 1996.

[100] S. Mori, C. Y. Suen, and K. Yamamoto, "Historical review of OCR research and development," vol. 80, no. 7, pp. 1029–1058, 1992.

[101] N. Chinchor, "MUC-7 named entity task definition," in *Proceedings of the Message Understanding Conference*, 1997, pp. 1–21, Available at http://www.itl.nist.gov/iad/894.02/related_projects/muc/proceedings/ne_task.html.

[102] A. Borthwick, J. Sterling, E. Agichtein, and R. Grishman, "Exploiting diverse knowledge sources via maximum entropy in named entity recognition," in *Proceedings of the Sixth Workshop on Very Large Corpora*, 1998, pp. 152–160.

[103] S. Lawrence, C. L. Giles, and K. Bollacker, “Digital libraries and autonomous citation indexing,” *IEEE Computer*, vol. 32, no. 6, pp. 67–71, 1999.

[104] W. Cohen and S. Sarawagi, “Exploiting dictionaries in named entity extraction: Combining semi-Markov extraction processes and data integration methods,” in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2004, pp. 89–98.

[105] L. O’Gorman, “The document spectrum for page layout analysis,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 11, pp. 1162–1173, 1993.

[106] K. Kise, A. Sato, and M. Iwata, “Segmentation of page images using the area Voronoi diagram,” *Computer Vision and Image Understanding*, vol. 70, no. 3, pp. 370–382, 1998.

[107] S. Mao and T. Kanungo, “Empirical performance evaluation methodology and its application to page segmentation algorithms,” *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, no. 3, pp. 242–256, 2001.

[108] H. M. Wallach, *Conditional Random Fields: An Introduction*, Technical Report MS-CIS-04-21, University of Pennsylvania, Available at http://repository.upenn.edu/cis_reports/22/, 2004.

[109] J. Lafferty, A. McCallum, and F. Pereira, “Conditional random fields: Probabilistic models for segmenting and labeling sequence data,” in *Proceedings of the International Conference on Machine Learning*, 2001, pp. 282–289.

[110] D. Pinto, A. McCallum, X. Wei, and W. Bruce, “Table extraction using conditional random fields,” in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2003, pp. 235–242.

[111] F. Sha and F. Pereira, “Shallow parsing with conditional random fields,” in *Proceedings of the Conference of the North American Chapter of the ACL on Human Language Technology*, 2003, pp. 134–141.

[112] S. Sarawagi and W. W. Cohen, “Semi-markov conditional random fields for information extraction,” in *Proceedings of the Annual Conference on Neural Information Processing Systems*, 2004, pp. 1185–1192.

[113] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*, 2nd ed. New York, New York: John Wiley and Sons, Inc., 2000.

[114] R. H. Byrd, J. Nocedal, and R. B. Schnabel, “Representations of quasi-Newton matrices and their use in limited memory methods,” *Mathematical Programming*, vol. 63, pp. 129–156, 1994.

[115] H. T. Ng, C. Y. Lim, and J. L. T. Koo, “Learning to recognize tables in free text,” in *Proceedings of the Annual Meeting of the ACL*, 1999, pp. 443–450.

[116] A. K. McCallum, K. Nigam, J. Rennie, and K. Seymore, “Automating the construction of internet portals with machine learning,” *Information Retrieval*, vol. 3, no. 2, pp. 127–163, 2000.

[117] R. Bunescu, R. Ge, R. J. Mooney, E. Marcotte, and A. K. Ramani, *Extracting Gene and Protein Names from Biomedical Abstracts*, Unpublished Technical Note, University of Texas, Austin, Available at http://www.cs.utexas.edu/users/ml/publication/ie.html, 2002.

[118] K. Humphreys, G. Demetriou, and R. Gaizauskas, "Two applications of information extraction to biological science journal articles: Enzyme interactions and protein structures," in *Proceedings of the Pacific Symposium on Biocomputing*, 2000, pp. 502–513.

[119] O. Etzioni, M. Cafarella, D. Downey, A.-M. Popescu, T. Shaked, S. Soderland, D. S. Weld, and A. Yates, "Unsupervised named-entity extraction from the web: An experimental study," *Art. Intell.*, vol. 165, no. 1, pp. 91–134, 2005.

[120] K. Fukuda, T. Tsunoda, A. Tamura, and T. Takagi, "Toward information extraction: Identifying protein names from biological papers," in *Proceedings of the Pacific Symposium on Biocomputing*, 1998, pp. 707–718.

[121] M. Narayanaswamy, K. E. Ravikumar, and K. Vijay-Shanker, "A biological named entity recognizer," in *Proceedings of the Pacific Symposium on Biocomputing*, 2003, pp. 427–438.

[122] X. Carreras, L. Marquez, and L. Padro, "Named entity extraction using AdaBoost," in *Proceedings of the Conference on Computational Natural Language Learning*, 2002, pp. 167–170.

[123] R. Florian, A. Ittycheriah, H. Jing, and T. Zhang, "Named entity recognition through classifier combination," in *Proceedings of the Conference on Computational Natural Language Learning*, 2003, pp. 168–171.

[124] M. Collins, "Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2002, pp. 1–8.

[125] A. McCallum, D. Freitag, and F. Pereira, "Maximum entropy Markov models for information extraction and segmentation," in *Proceedings of the International Conference on Machine Learning*, 2000, pp. 591–598.

[126] X. He and M. C.-P. R. Zemel, "Multiscale conditional random fields for image labeling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2004, pp. 695–702.

[127] C. Sminchisescu, A. Kanaujia, Z. Li, and D. Metaxas, "Conditional models for contextual human motion recognition," in *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2, 2005, pp. 1808–1815.

[128] E. Amitay and A. Broder, "Introduction to special issue on query log analysis: Technology and ethics," *ACM Trans. Web*, vol. 2, no. 4, Article 18, pp. 1–2, 2008.

[129] D. Beeferman and A. Berger, "Agglomerative clustering of a search engine query log," in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2000, pp. 407–416.

[130] R. Jones, B. Rey, O. Madani, and W. Greiner, “Generating query substitutions,” in *Proceedings of the World Wide Web Conference*, 2006, pp. 387–396.

[131] R. W. White, M. Bilenko, and S. Cucerzan, “Leveraging popular destinations to enhance web search interaction,” *ACM Trans. Web*, vol. 2, no. 3, Article 16, pp. 1–30, 2008.

[132] W. Cohen, R. Shapire, and Y. Singer, “Learning to order things,” *J. Art. Intell. Res.*, vol. 10, pp. 243–270, 1999.

[133] T. Joachims, “Optimizing search engines using clickthrough data,” in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2002, pp. 133–142.

[134] F. Radlinski and T. Joachims, “Query chains: Learning to rank from implicit feedback,” in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2005, pp. 239–248.

[135] E. Agichtein, E. Brill, and S. Dumais, “Improving web search ranking by incorporating user behavior information,” in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2006, pp. 19–26.

[136] N. Craswell and M. Szummer, “Random walks on the click graph,” in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2007, pp. 239–246.

[137] F. Radlinski, A. Broder, P. Ciccolo, E. Gabrilovich, V. Josifovski, and L. Riedel, “Optimizing relevance and revenue in ad search: A query substitution approach,” in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 403–410.

[138] G. Zhu and G. Mishne, “Mining rich session context to improve web search,” in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2009, pp. 1037–1046.

[139] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz, “Analysis of a very large web search engine query log,” *SIGIR Forum*, vol. 33, no. 1, pp. 6–12, 1999.

[140] F. Radlinski and T. Joachims, “Active exploration for learning rankings from clickthrough data,” in *Proceedings of the ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 2007, pp. 570–579.

[141] Y. Liu, B. Gao, T.-Y. Liu, Y. Zhang, Z. Ma, S. He, and H. Li, “BrowseRank: Letting web users vote for page importance,” in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2008, pp. 451–458.

[142] R. Baeza-Yates, C. Castillo, F. Junqueira, V. Plachouras, and F. Silvestri, “Challenges in distributed information retrieval,” in *Proceedings of the IEEE Conference on Data Engineering*, 2007, pp. 6–20.

[143] C. Olston and S. Pandey, “Recrawl scheduling based on information longevity,” in *Proceedings of the World Wide Web Conference*, 2008, pp. 437–446.

[144] B. He, M. Patel, Z. Zhang, and K. C.-C. Chang, “Accessing the deep Web,” *Commun. ACM*, vol. 50, no. 5, pp. 94–101, 2007.

[145] L. Page, S. Brin, R. Motwani, and T. Winograd, *The PageRank Citation Ranking: Bringing Order to The web*, Technical Report, Stanford InfoLab, Available at http://ilpubs.stanford.edu:8090/422/, 1999.

[146] J. Kleinberg, “Authoritative sources in a hyperlinked environment,” *J. ACM*, vol. 46, no. 5, pp. 604–632, 1999.

[147] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen, “Combating web spam with TrustRank,” in *Proceedings of the International Conference on Very Large Data Bases*, 2004, pp. 576–587.

[148] C. Moler, *The World’s Largest Matrix Computation*, Online, 2002, http://www.mathworks.com/company/newsletters/news_notes/clevescorner/oct02_cleve.html.

[149] M. R. Meiss, F. Menczer, S. Fortunato, A. Flammini, and A. Vespignani, “Ranking web sites with real user traffic,” in *Proceedings of the ACM Conference on Web Search and Data Mining*, 2008, pp. 65–76.

[150] Google, *We Knew The Web Was Big*, Online, 2008, http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.

[151] A. N. Langville and C. D. Meyer, “Deeper inside PageRank,” *Internet Math.*, vol. 1, no. 3, pp. 335–400, 2005.

[152] A. Broder, R. Lempel, F. Maghoul, and J. Pedersen, “Efficient PageRank approximation via graph aggregation,” in *Proceedings of the World Wide Web Conference*, 2004, pp. 484–485.

[153] F. McSherry, “A uniform approach to accelerated PageRank computation,” in *Proceedings of the World Wide Web Conference*, 2005, pp. 575–582.

[154] R. W. White and S. M. Drucker, “Investigating behaviorial variability in web search,” in *Proceedings of the World Wide Web Conference*, 2007, pp. 21–30.

[155] M. Bilenko and R. W. White, “Mining the search trails of surfing crowds: Identifying relevant websites from user activity,” in *Proceedings of the World Wide Web Conference*, 2008, pp. 51–60.

[156] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna, “The query-flow graph: Model and applications,” in *Proceedings of the ACM Conference on Information and Knowledge Management*, 2008, pp. 609–618.

[157] D. Downey, D. Liebling, and S. Dumais, “Understanding the relationship between searchers’ queries and information goals,” in *Proceedings of the ACM Conference on Information and Knowledge Management*, 2008, pp. 449–458.

[158] W. M. Rand, “Objective criteria for the evaluation of clustering methods,” *J. Am. Statist. Assoc.*, vol. 66, no. 336, pp. 846–850, 1971.

[159] A. Mowshowitz and A. Kawaguchi, "Bias on the Web," *Commun. ACM*, vol. 45, no. 9, pp. 56–60, 2002.

[160] R. D. Yates and D. J. Goodman, *Probability and Stochastic Processes*, 2nd ed. Hoboken, New Jersey: John Wiley and Sons, Inc., 2005.

[161] G. Grimmett and D. Stirzaker, *Probability and Random Processes*, 3rd ed. Oxford: Oxford University Press, 2001.

[162] N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor, "Relevance weighting for query independent evidence," in *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval*, 2005, pp. 416–423.

[163] J. H. Friedman, "Greedy function approximation: A gradient boosting machine," *Ann. Statist.*, vol. 29, no. 5, pp. 1189–1232, 2001.

[164] J. Dean and S. Ghemawat, "MapReduce: Simplified data processing on large clusters," in *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, 2004, pp. 137–150.

[165] Hadoop, *Open Source Implementation of MapReduce*, Online, 2007, http://lucene.apache.org/hadoop/.

[166] K. S. Jones, S. Walker, and S. E. Robertson, "A probabilistic model of information retrieval: Development and comparative experiments (parts 1 and 2)," *Inform. Proces. Manage.*, vol. 36, no. 6, pp. 779–840, 2000.

[167] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of IR techniques," *ACM Trans. Inf. Syst.*, vol. 20, no. 4, pp. 422–446, 2002.

[168] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees.* Boca Raton, Florida: CRC Press, 1984.