# Document Image Retrieval Based on 2D Density Distributions of Terms with Pseudo Relevance Feedback

Koichi Kise, Yin Wuotang and Keinosuke Matsumoto
Dept. of Computer and Systems Sciences, Osaka Prefecture University
kise@cs.osakafu-u.ac.jp

## Abstract

*Document image retrieval is a task to retrieve document images relevant to a user's query. Most of existing methods based on word-level indexing rely on the representation called "bag of words" which originated in the field of information retrieval. This paper presents a new representation of documents that utilizes additional information about the location of words in pages so as to improve the retrieval performance. We consider that pages are relevant to a query if they contains its terms densely. This notion is embodied as density distributions of terms calculated in the proposed method. Its performance is improved with the help of "pseudo relevance feedback", i.e., a method of expanding a query by analyzing pages. Experimental results on English document images show that the proposed method is superior to conventional methods of electronic document retrieval at recall levels 0.0–0.6.*

## 1. Introduction

Document image databases (DIBs) are databases that store document images obtained mainly by scanning paper documents. Since we have had a huge amount of paper documents and continue to produce them, the technology of DIBs gains importance in the society. An apparent advantage of DIBs is substantial reduction of the space required for storing documents. Another important advantage is a potential for fast and intelligent retrieval of documents, though this has been a subject under study.

So far a lot of retrieval methods have been proposed [1]. Their indexing schemes range from feature-level indexing for layout structure [2] to word-level indexing for OCR'd text [3, 4]. In this paper, we focus our discussion on the word-level indexing.

With this indexing scheme, the central issue has been how to deal with OCR errors. This is because, once the problem of OCR errors is solved, document image retrieval seems equivalent to electronic document retrieval in which extensive experience has been accumulated.

The application of methods for electronic document retrieval indicates that their document representations based on the indexing scheme are also applied to document images. A standard representation is called "bag of words" (BOW) in which all the structure and linear ordering of words are ignored. This means that a lot of information provided by OCR is discarded.

In this paper, we propose a method of document image retrieval based on a new representation called "two dimensional term distributions": We view each page image as a two dimensional distribution of various terms (symbols). The proposed method retrieves pages which *densely* contain terms in a query. This notion originated in the field of electronic document retrieval [5, 6] and was extended as a method of retrieval of Japanese newspaper images by some of the authors [7]. In this paper, we further extend the method by using "pseudo relevance feedback", which is a method of query expansion for improving the performance of retrieval [8]. The experimental results on retrieving English paper documents in the DjVu format [9] show that the proposed method outperforms conventional methods of electronic document retrieval.

## 2. Representation of documents

In the field of information retrieval (IR), a standard document representation is called the "bag of words"(BOW) model. In the BOW model, only the information about words and their frequency is employed; other information such as the word location is ignored. As for document image retrieval, in addition to a BOW model for documents, we can consider a BOW model for pages in which pages are indexed. In this paper, the former is referred to as the "document BOW model" and the latter the "page BOW model".

When we apply OCR to document images, it is natural to obtain layout information in addition to word symbols. Such information includes word bounding boxes, bound-

Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003)
0-7695-1960-1/03 $17.00 © 2003 IEEE

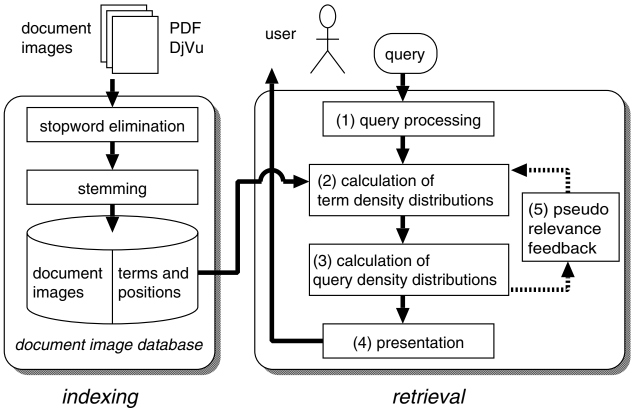


**Figure 1. Processing of the proposed method.**

ing boxes of textlines and zones. In the formats of documents such as PDF and DjVu, this information is recorded with word symbols as hidden text. This means that we have richer information than required for the BOW model, though the layout information is sometimes noisy depending on the quality of images.

In this paper, we utilize the bounding boxes (location) of words in pages in addition to symbol information and view pages as two dimensional distributions of words. We consider that this is an initial step towards a more fruitful representation of document images for retrieval.

## 3. Two-dimensional Density Distributions

This section describes the proposed method consisting of two kinds of processing shown in Fig. 1.

### 3.1. Indexing

The process of indexing is applied once when documents are first stored in the document image database. First, terms are extracted from all pages of documents by applying stemming and stopword elimination [1]. Each page is then stored as a collection of terms with their location in the page.

### 3.2. Retrieval

The process of retrieval is composed of five steps shown in Fig. 1. Steps (1)–(4) are the steps that constitute a basic retrieval function. The step (5) is for an extension of the basic functionality to improve the performance. We first describe the details of the basic retrieval function.

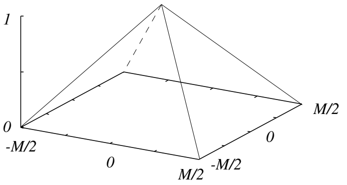


**Figure 2. Window function.**

**(1) Query processing**

A query issued by a user may be represented as a sequence of words. In the first step of retrieval process, this sequence is analyzed to structure the query. For example, a query “color and multimedia documents” is structured in the form ((“color”), (“multimedia”,“documents”)). The inner parentheses indicate that, if they contain more than one term, these terms constitute a *compound word*. The outer parentheses represent the whole query which may contain several compound and normal words.

In the following, a query is represented as $q = (q_1, ..., q_m)$, and $q_i = (q_{i1}, ..., q_{in_i})$ where $q_i$ corresponds to a compound word and $q_{ij}$ is a term (after stemming and stopword elimination) in a compound word $q_i$.

**(2) Calculation of term density distributions**

The parts of page images which densely contain terms in a query may be relevant to a query. A term density distribution represents such information for each term in each page. The definition is as follows.

A term $q_{ij}$ is associated with the weight called *inverse document frequency*. We utilize a common definition: $\mathrm{idf}_{ij} = \log(n/n_{ij})$ where $n$ is the total number of documents in the database and $n_{ij}$ is the number of documents that include the term $q_{ij}$. Let $T^{(p)}_{ij}(x, y)$ be a weighted distribution of a term $q_{ij}$ in a page $p$ defined by:

$$T^{(p)}_{ij}(x, y) = \begin{cases} \mathrm{idf}_{ij} & \text{if } q_{ij} \text{ occurs at } (x, y), \\ 0 & \text{otherwise ,} \end{cases} \quad (1)$$

where $(x, y)$ is the center of the bounding box of a term.

A density distribution $D^{(p)}_{ij}(x, y)$ is a weighted distribution of $q_{ij}$ smoothed by a window $W(x, y)$:

$$D^{(p)}_{ij}(x, y) = \sum_{u,v} W(x - u, y - v) T^{(p)}_{ij}(u, v) \; . \quad (2)$$

In this paper, we utilize a pyramidal function with the window width $M$ shown in Fig. 2.

**(3) Calculation of query density distributions**

The next step is to combine term density distributions to obtain density distributions of a query. First, a density

[1] Stemming is the process of normalizing words by keeping only *word stems*, e.g., from “processes” to “process”. Stopwords are words that convey little meaning such as “a” and “the”. Words employed for indexing after these processes are called terms.

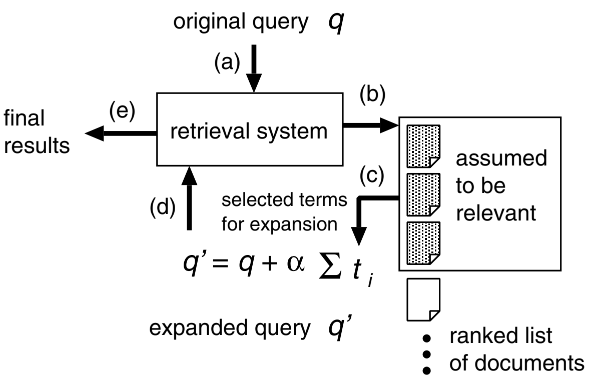
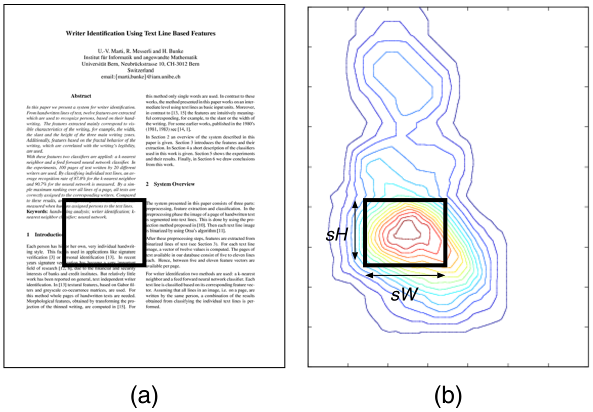


**Figure 3. Pseudo relevance feedback.**



Writer Identification Using Text Line Based Features

U.-V. Marti, R. Messerli and H. Bunke
Institut für Informatik und angewandte Mathematik
Universität Bern, Neubrückstrasse 10, CH-3012 Bern
Switzerland
email:{marti,bunke}@iam.unibe.ch

Abstract

*In this paper we present a system for writer identification. From handwritten lines of text, twelve features are extracted which are used to recognize persons, based on their handwriting. The features extracted mainly correspond to visible characteristics of the writing, for example, the width, the slant and the height of the three main writing zones. Additionally, features based on the fractal behavior of the writing, which are correlated with the writing's legibility, are used.*
*With these features two classifiers are applied: a k-nearest neighbor and a feed forward neural network classifier. In the experiments, 100 pages of text written by 20 different writers are used. By classifying individual text lines, an average recognition rate of 87.8% for the k-nearest neighbor and 90.7% for the neural network is measured. By a simple maximum ranking over all lines of a page, all texts are correctly assigned to the corresponding writers. Compared to these results, a[illegible] measured when hu[illegible]s assigned persons to the text lines.*
**Keywords:** *handw[illegible]ing analysis; writer identification; k-nearest neighbor cl[illegible]sifier; neural network.*

1 Introductio[illegible]

Each person has hi[illegible]r her own, very individual handwriting style. This fac[illegible]s used in applications like signature verification [3] or [illegible]rsonal identification [13]. In recent years signature ve[illegible] field of research [12, 8], due to the financial and security interests of banks and credit institutes. But relatively little work has been reported on general, text independent writer identification. In [13] textural features, based on Gabor filters and greyscale co-occurrence matrices, are used. For this method whole pages of handwritten texts are needed. Morphological features, obtained by transforming the projection of the thinned writing, are computed in [15]. For this method only single words are used. In contrast to these works, the method presented in this paper works on an intermediate level using text lines as basic input units. Moreover, in contrast to [13, 15] the features are intuitively meaningful corresponding, for example, to the slant or the width of the writing. For some earlier works, published in the 1980's (1981, 1983) see [14, 1].

In Section 2 an overview of the system described in this paper is given. Section 3 introduces the features and their extraction. In Section 4 a short description of the classifiers used in this work is given. Section 5 shows the experiments and their results. Finally, in Section 6 we draw conclusions from this work.

2 System Overview

[illegible]he system presented in this paper consists of three parts: [illegible]preprocessing, feature extraction and classification. In the [illegible]preprocessing phase the image of a page of handwritten text [illegible]s segmented into text lines. This is done by using the pro[illegible]ection method proposed in [10]. Then each text line image [illegible]s binarized by using Otsu's algorithm [11].

[illegible]fter these preprocessing steps, features are extracted from [illegible]inarized lines of text (see Section 3). For each text line [illegible]mage, a vector of twelve values is computed. The pages of [illegible]ext available in our database consist of five to eleven lines [illegible]ach. Hence, between five and eleven feature vectors are [illegible]vailable per page.

For writer identification two methods are used: a k-nearest neighbor and a feed forward neural network classifier. Each text line is classified based on its corresponding feature vector. Assuming that all lines in an image, i.e., on a page, are written by the same person, a combination of the results obtained from classifying the individual text lines is performed.

(a) (b)

**Figure 4. Finding terms for expansion.**

distribution $D_i^{(p)}(x,y)$ of a compound word $q_i$ is obtained by multiplying those of terms $q_{ij}$ in $q_i$:

$$D_i^{(p)}(x,y) = \prod_j D_{ij}^{(p)}(x,y) \ . \tag{3}$$

This means that a compound word exists if all of its terms exist.

The density distributions $D^{(p)}(x,y)$ of a query is likewise obtained:

$$D^{(p)}(x,y) = \sum_i D_i^{(p)}(x,y) \ . \tag{4}$$

Summation is utilized since parts of pages may be relevant to a query if some of its compound words occur frequently.

**(4) Presentation**

Based on the query density distributions, a score $\mathrm{score}(p)$ of a page $p$ is obtained:

$$\mathrm{score}(p) = \max_{x,y} D^{(p)}(x,y) \ . \tag{5}$$

Finally, the user receives a list of pages sorted according to the score. Note that query density distributions help the user browse the results of retrieval, since relevant parts can be found as the parts whose density is high.

### 3.3. Pseudo relevance feedback

An important limitation of the above basic function is that retrieval performance for some queries is not high enough. This is due to query terms less effective to find relevant documents.

In the field of IR, researchers have tackled a similar problem. A well-known method for the problem is called "pseudo relevance feedback"(PRF). Figure 3 illustrates a process of PRF for electronic documents. First, an original query is utilized to obtain an initial list of documents sorted according to the degree of relevance to the query (Figs. 3 (a) and (b)). Next, some top-ranking documents are *assumed* to be relevant and terms that characterize those documents are extracted (c). Then, the original query is expanded using the terms (d). Finally, the expanded query is utilized to improve the ranking (e).

We introduce PRF for retrieving document images based on the density distributions. The key is how to obtain terms for expansion. We consider that the terms contained in the parts whose density is high would be useful for expansion.

The details are as follows. First, top $N_p$ pages are selected from the initial list of pages. Then terms are selected from the pages based on query density distributions.

Figure 4(b) shows contour lines of a query density distribution for Fig. 4(a). In order to extract terms, we use a small window of size $sW \times sH$ whose center is at:

$$(x^*, y^*) = \arg\max_{(x,y)} D^{(p)}(x,y) \ . \tag{6}$$

Then we extract terms that are not in the original query and whose center is in the small window.

Let $\{t_k\}$ be a set of extracted terms from the top $N_p$ pages, and $E_k^{(p)}(x,y)$ be the term density distribution of a term $t_k$ in a page $p$ obtained by Eq. (2). The density distribution $\hat{D}^{(p)}(x,y)$ of the expanded query is given by:

$$\hat{D}^{(p)}(x,y) = D^{(p)}(x,y) + \alpha \sum_k E_k^{(p)}(x,y) \ . \tag{7}$$

where $\alpha$ is a parameter. The results of PRF are obtained by ranking pages according to the scores given by $\hat{D}^{(p)}$.

## 4. Experimental comparison

### 4.1. Methods for comparison

We compared the proposed method experimentally with the following representative methods of electronic document retrieval: the simple vector space model (VSM) and the latent semantic indexing (LSI).

The VSM is a standard way of retrieving electronic documents [8]. In this method, we utilized the TF-IDF weighting scheme with log-TF and log-IDF. As the representation of documents, we employed both the document BOW model and the page BOW model, which are referred to as VSM(doc) and VSM(page), respectively.

The latent semantic indexing is an extension of the VSM. LSI employs the singular value decomposition (SVD) of a term-by-document matrix [8]. Reducing the dimensionality of the original term-by-document matrix based on the result of SVD enables us to uncover *latent* semantic relation among terms. This often results in obtaining better ranking of documents. The parameter of this method is the dimension $\kappa$. We also applied methods with two representations that are indicated as LSI(doc) and LSI(page).

In addition, we utilized the proposed method without PRF for comparison. This is referred to as 2D-DD; the proposed method (with PRF) is 2D-DD/PRF.

### 4.2. Test collection

In order to make experiments on the proposed method, it is necessary to employ a test collection of document images with bounding boxes of words. To our knowledge, however, there is no such collection. Thus we prepared it using the DjVu files of the Proc. ICDAR 2001 [9].

We defined queries based on the titles of the oral sessions shown in the table of contents of the proceedings. The complete list of the queries with their structure is shown in Table 1 [2]. Documents are determined to be relevant to a query if they are assigned to the corresponding session. Statistics about the test collection are shown in Table 2.

### 4.3. Evaluation

A standard way of evaluating the retrieval performance is to compute recall and precision [8]. Let $X$ and $Y$ be a set of retrieved documents and a set of relevant documents for a query, respectively. The recall and the precision for a query are defined by $R = |X \cap Y|/|Y|$ and $P = |X \cap Y|/|X|$, respectively. In addition, it is sometimes convenient to utilize a single value summary of retrieval performance. The mean average precision over all relevant documents [10] is one of such measures.

In order to determine the values of parameters in a less biased manner, we applied cross-validation as follows. First, queries were divided into four sets $Q_1 = \{1, ..., 4\}$, $Q_2 = \{5, ..., 9\}$, $Q_3 = \{10, ..., 14\}$, and $Q_4 = \{15, ..., 19\}$. The values of parameters for a test set $Q_i$ was determined as the values that maximized the mean average precision for the rest of query sets (learning sets). Table 3 lists the ranges of parameters from which the best values were selected.

**Table 1. Queries**

| query id | terms and their structure |
|---|---|
| 1 | ((“ handwriting”, “ classifiers”)) |
| 2 | ((“ image”, “ processing”)) |
| 3 | ((“ OCR”), (“ document”, “ understanding” )) |
| 4 | ((“ handwriting”, “ features”),<br>(“ writer”, “ identification” )) |
| 5 | ((“ evaluation”), (“ datasets”)) |
| 6 | ((“ word”, “ recognition”)) |
| 7 | ((“ classifiers”), (“ learning”)) |
| 8 | ((“ graphics”, “ recognition”)) |
| 9 | ((“ document”, “ analysis”, “ systems”),<br>(“ check”, “ processing”),<br>(“ postal”, “ automation”)) |
| 10 | ((“ applications”)) |
| 11 | ((“ online”, “ handwriting”)) |
| 12 | ((“ tables”)) |
| 13 | ((“ classification”), (“ language”, “ models”)) |
| 14 | ((“ forms”)) |
| 15 | ((“ segmentation”)) |
| 16 | ((“ web”, “ documents”), (“ video”)) |
| 17 | ((“ information”, “ retrieval”),<br>(“ word”, “ spotting”)) |
| 18 | ((“ color”), (“ multimedia”,“ documents”)) |
| 19 | ((“ image”, “ quality”), (“style”)) |

**Table 2. Statistics about the test collection.**

| | | |
|---|---|---|
| no. of terms | 7022 | |
| no. of doc. | 129 | |
| no. of pages | 673 | |
| ave. doc. len. | 1507 | terms |
| ave. page. len. | 289 | terms |
| no. of queries | 19 | |
| ave. query len. | 2.37 | terms |
| ave. no. of rel. doc. per query | 6.79 | |

### 4.4. Results

The results are shown in Table 4 and Fig. 5. The following are the summary of the results.

- From a viewpoint of the mean average precision, LSI(page) was the best among the conventional methods [3].

[2] In defining the queries we skipped some sessions such as “Non-Latin Alphabets”, and merged some sessions whose contents overlap such as “Document Analysis Systems and Check Processing” and “Postal Automation and Check Processing”.

[3] This indicates that the physical structure of “pages” available in document images is effective. We consider that pages play a role of passages in passage retrieval, which often improves the retrieval performance[6].

**Table 3. Parameters.**

| method and param. | | range |
|---|---|---|
| LSI(doc) | $\kappa$ | 10~ 120 step 10 |
| LSI(page) | $\kappa$ | 10~ 600 step 10 |
| 2D-DD | $M$ | 100 ~ 700 step 100 |
| 2D-DD/PRF | $M$ | 600 |
| | $N_p$ | 4 ~ 7 step 1 |
| | $\alpha$ | 0.002 ~ 0.01 step 0.002<br>0.01 ~ 0.1 step 0.01 |
| | $sW$ | 50 ~ 62 step 4 |
| | $sH$ | 14 ~ 30 step 4 |

**Table 4. Mean average precision.**

| | | |
|---|---|---|
| VSM(doc) | 0.3154 | |
| VSM(page) | 0.2902 | (−8.0%) |
| LSI(doc) | 0.3011 | (−4.5%) |
| LSI(page) | 0.3372 | (+6.9%) |
| 2D-DD | *0.3440* | (+9.1%) |
| 2D-DD/PRF | **0.3551** | (+12.6%) |

( ) : difference to the VSM (doc)

- The proposed method (2D-DD/PRF) was the best among the methods and 2D-DD was the second best.
- As shown in Fig. 5, at recall levels 0.0 – 0.6, 2D-DD/PRF yielded the best results.
- At recall levels 0.3 – 1.0, 2D-DD/PRF outperformed 2D-DD.
- At recall levels more than 0.6, LSI(page) was the best method.

From these results, 2D-DD/PRF can be considered to be the best choice among the tested methods for the users who are interested in the results at lower recall levels up to 0.6.

## 5. Conclusion

We have presented a new method of document image retrieval based on two dimensional density distributions of terms. The proposed method employs "pseudo relevance feedback" to expand an original query so as to improve the performance of retrieval. From the experimental results on document images, it has been shown that the proposed method is superior to the conventional methods. Future work includes further investigation with a larger number of documents and queries as well as to improve the way of selecting terms for expansion.

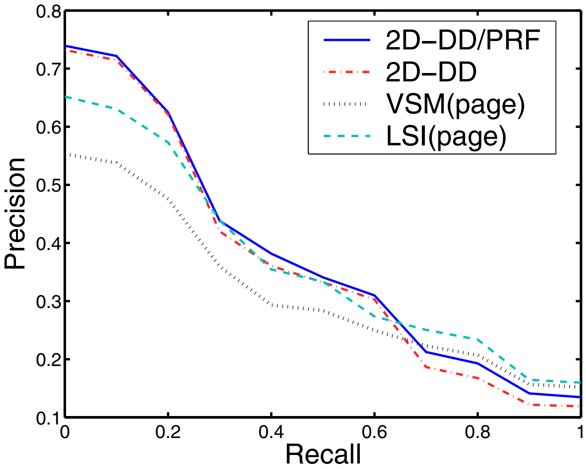


**Figure 5. Recall-precision graph.**

## Acknowledgment

This research was supported in part by Grant-in-Aid for Scientific Research (C) and (B) from Japan Society for the Promotion of Science (No.14580453, No.14380182).

## References

[1] Doermann, D.: The Indexing and Retrieval of Document Images: A Survey, *Computer Vision and Image Processing*, Vol. 70, No. 3, pp.287–298, 1998.

[2] Hu, J., Kashi, R. and Wilfong, G., Document Image Layout Comparison and Classification, *Proc. 5th ICDAR*, pp.285–288, 1999.

[3] Taghva, K., Borsack, J. and Condit, A., Evaluation of Model-Based Retrieval Effectiveness with OCR Text, *ACM Trans. Information Systems*, Vol.14, No.1, pp.64–93, 1996.

[4] Jones, G.J.F. and Lam-Adesina, A.M., Examining the Effectiveness of IR Techniques for Document Image Retrieval, *Proc. SIGIR Workshop on Information Retrieval and OCR: From Converting Content to Grasping Meaning*, 2002.

[5] Kurohashi, S., Shiraki, N. and Nagao, M., A Method for Detecting Important Descriptions of a Word Based on Its Density Distribution in Text, *Trans. Information Processing Society of Japan*, Vol.38, No.4, pp.845–853, 1997 [In Japanese].

[6] Kise, K., Junker, M., Dengel, A. and Matsumoto, K., Experimental Evaluation of Passage-Based Document Retrieval, in *Proc. ICDAR'01*, pp.592–596, 2001.

[7] Kise, K., Tsujino, M. and Matsumoto, K., Spotting Where to Read on Pages — Retrieval of Relevant Parts from Page Images, in *Proc. DAS'02*, pp.388–399, 2002.

[8] Baeza-Yates, R. and Ribeiro-Neto, B., *Modern Information Retrieval*, Addison-Wesley Pub. Co., 1999.

[9] `http://icdar.djvuzone.org/`

[10] `ftp://ftp.cs.cornell.edu/pub/smart/`