Information Retrieval System for Handwritten Documents*

Sargur Srihari, Anantharaman Ganesh, Catalin Tomai,
Yong-Chul Shin, and Chen Huang

Center of Excellence for Document Analysis and Recognition (CEDAR)
University at Buffalo, State University of New York, Buffalo, USA
{srihari,aganesh,catalin,ycshin,chuang5}@cedar.buffalo.edu

Abstract. The design and performance of a content-based information retrieval system for handwritten documents is described. System indexing and retrieval is based on writer characteristics, textual content as well as document meta data such as writer profile. Documents are indexed using global image features, e.g., stroke width, slant, word gaps, as well local features that describe shapes of characters and words. Image indexing is done automatically using page analysis, page segmentation, line separation, word segmentation and recognition of characters and words. Several types of queries are permitted: (i) entire document image; (ii) a region of interest (ROI) of a document; (iii) a word image; and (iv) textual. Retrieval is based on a probabilistic model of information retrieval. The system has been implemented using Microsoft Visual C++ and a relational database system. This paper reports on the performance of the system for retrieving documents based on same and different content.

1 Introduction

Methods for indexing and retrieval of scanned handwritten documents are needed for various applications such as historical manuscripts, scientific notes, personal records, as well as criminal records. In each of these applications there is a need for indexing and retrieval based on textual content as well as user-indexed terms. In the forensic application there is a need for searching a database of handwritten documents not only for textual content but also for visual content such as writer characteristics. This paper describes the indexing and retrieval aspects of a system that attempts to provide the full range of functionalities for a digital library of handwritten documents.

Writer identification has a long history perhaps dating to the origins of handwriting itself. Classic forensic handwriting examination is primarily based upon the knowledge and experience of the forensic expert. There exist many textbooks [1–5] describing the methodology employed by forensic document examiners. A computer system for retrieving handwritten documents from a set of documents

* This work was supported in part by the U.S. Department of Justice, National Institute of Justice grant 2002-LT-BX-K007.

S. Marinai and A. Dengel (Eds.): DAS 2004, LNCS 3163, pp. 298–309, 2004.

of forensic interest, known as the Forensic Information System for Handwriting, or the FISH system [6], has been developed by German law enforcement. Also motivated by the forensic application, a handwritten document management system, known as CEDAR-FOX [7, 8], has been developed at CEDAR, whose indexing and retrieval aspects are the subject of this paper.

1.1 CEDAR-FOX System

As a document management system for handwritten documents, CEDAR-FOX provides several functionalities: interactive document analysis, image indexing to create a digital library for content-based retrieval, and use as database management system. For the purpose of indexing based on writer characteristics, features are automatically extracted after several image processing functions followed by character recognition. The user can use interactive graphical tools to assist in obtaining more accurate writer characteristics. A unique aspect of CEDAR-FOX is that image matching is driven by probabilistic writer verification, i.e., whether the query and the document were likely to be written by the same writer.

Several database management tools for creating a handwritten document library are provided: (i) entering document meta-data, e.g., identification number, writer and other collateral information, (ii) creating a textual transcript of the image content at the word level, and (iii) including automatically extracted document level features, e.g., stroke width, slant, word gaps, as well as finer features that capture the structural characteristics of characters and words. The system can be customized to use any commercial or non-commercial database system for the digital library storage. It also provides access and retrieval functionalities for adding, modification and categorization of the document records in the digital library.

Information retrieval can be performed using several query modalities: (i) the entire document image is the query; (ii) partial image query: a region of interest (ROI) of a document or a word image; (iii) text: the user can type in keywords from the words in the documents, and (iv) meta-data: case number, person names, time and the pre-registered keywords such as brief descriptions of the case.

1.2 Organization of Paper

Section 2 describes the indexing aspects of CEDAR-FOX which has two parts: image features and meta data. Section 3 describes the retrieval aspects of the system. Section 4 shows the performance of the system with the original micro and macro features as well as with a new set of cognitive features. Concluding remarks are presented in Section 5.

2 Indexing

Indexing of handwritten document images does not only involve textual information like in the IR counterpart but includes document/writer features and also

meta data regarding the origin of the document, writer profile, type of writing instrument used, etc. Section 2.1 describes the image features and their use in document retrieval.

2.1 Image Features

Image features are computed at the document and the character levels. The document level features are called macro features whereas the character level features are called the micro features. The initial 12 macro features reported in [7] were truly at the document level, meaning they were computed either directly from the entire document (no. of black pixels, threshold, etc.) or on a line-by-line basis and applied to the entire document (no. of exterior contours, no. of interior contours, slope, etc.). We report new macro features that can be used to index images which are holistic - they are computed at the character level and applied to the entire document and normalized. The micro features are the character GSC features which are based on the character's gradient, structure and concavity properties.

Micro-features. The micro-features used in CEDAR-FOX consist of 512 bits corresponding to gradient (192 bits), structural (192 bits), and concavity (128 bits) features. Each of these three sets of features relies on dividing the scanned image of the character into a $4*4$ region. The gradient features capture the stroke flow orientation and its variations using the frequency of the gradient directions, as obtained by convolving the image with a Sobel edge operator, in each of 12 directions and then thresholding the resultant values to yield a 192-bit vector. The structural features representing the coarser shape of the character capture the presence of corners, diagonal lines, and vertical and horizontal lines in the gradient image, as determined by 12 rules. The concavity features capture the major topological and geometrical features including direction of bays, presence of holes, and large vertical and horizontal strokes. All the 512 binary features are converted from the original floating number computations.

Macro-features. Macro-features in CEDAR-FOX [7] represent the entire document. The current implementation of CEDAR-FOX consists of three sets of features: darkness, contour and averaged line-level features. The darkness features, in turn, consist of three features all obtained from the histogram of the gray-scale values in the scanned document image: the number of black pixels in the image, the gray-scale value corresponding to the valley in the histogram that separates the foreground pixels from the background pixels (known as the threshold) and the entropy of the histogram (which is a measure of uncertainty in the distribution). The contour features, six in number, are as follows: the number of components and holes (as measured by the number of interior and exterior contours in the chain-code outline of the handwriting), and slopes in the vertical, negative, positive and horizontal directions. The averaged line-level features consist of average slant and height of characters.

We also develop a new set of cognitive document-level features that capture: (i) the legibility of a handwritten document at character and word level, (ii) the relative proportions existing between the characters, and (iii) the distribution of writing-styles (lexemes) existing in the handwriting of a writer. Some of these features as well as some of the original macro features are illustrated in Fig.1.

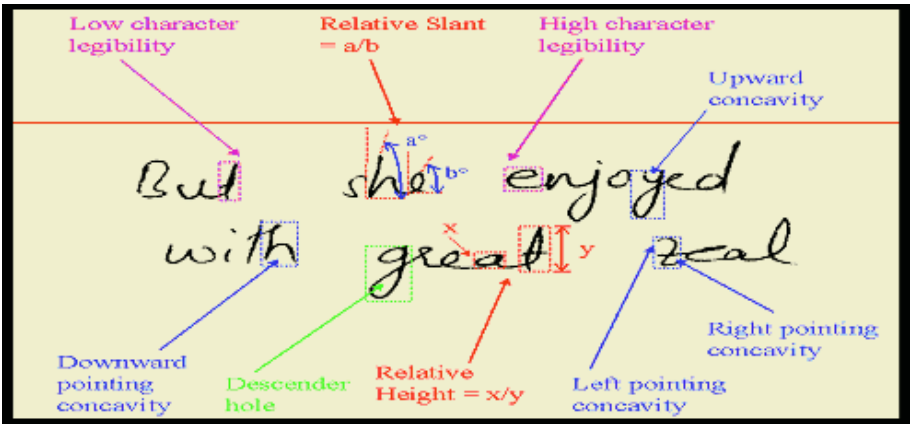

Fig. 1. Macro features used in current and future version of CEDAR-FOX

2.2 Meta-data

Storing indexes for handwritten documents also involves use of data related to the origin of the document, document identification number, style of writing (cursive, handprint) etc., making the task of retrieval much more easy and flexible for the user. For the use of questioned document examiners, who are potential users of such a retrieval system, we decided on the following set of meta-data that describe each document.

1. Case Number - The current document under the case it was associated with
2. Date of Writing - When the document was written
3. Date of Input - Automatically obtained from the system time
4. Writing instrument - Ball-point pen, magic marker, etc.
5. Geographical Area - Area associated with the case or where the document came from
6. Handedness - Handedness of writer, if known
7. Keywords - Any keywords the user wants to associate with the document e.g. Extreme, Threat, Kidnap, etc.
8. Suspect's First & Last Names
9. Victim's First and Last Names
10. Transcript - Transcript for the document obtained either by truthing or a text transcript
11. Path - Path of the file originally stored on disk
12. Image - The scanned document image itself.

The user can store the metadata using our dialog interface to the database, which is activated when a document is open as shown in Fig.2. The feature-based indexes are computed when the document is opened and are automatically stored into the database along with the meta-data. Thus the feature extraction process, which is an online computation task, is done only once and stored in the database. The final representation of the document is thus only the indexes associated with it, which include the meta-data, the features and the textual information associated with the document's transcript. It is to be noted that in contrast with IR, the indexes here are content (textual) as well as cognitively-based (FDE features). The database provides the necessary functionality to do string and numeral based comparisons.

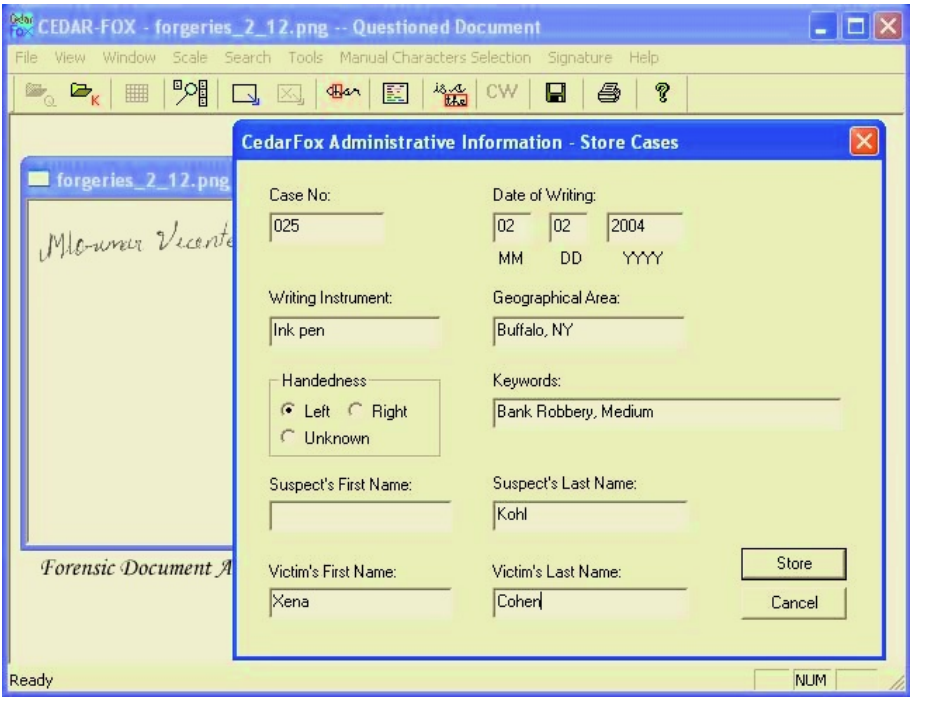

Fig. 2. Dialog interface for database: allows user to enter meta-data for a document

3 Retrieval

Related to the issue of retrieving documents from a database that is relevant to the query supplied is the need for associating a quantitative measure of similarity between two samples. For the task of writer identification, the goal is to take the document as a query to compare with some or all of the document data in the database. The matching between the query document and each document in the

database is performed using many attributes of the documents. The matching may be done using the micro features, macro features or their combination.

3.1 Micro-feature Similarity

To measure the similarity between two characters whose shapes are represented using binary vectors we use the Correlation measure [9]. Given $S_{ij}, i, j \in \{0, 1\}$, the number of occurrences of matches with i in the first feature vector and j in the second feature vector at the corresponding positions, the dissimilarity D between the two feature vectors X and Y is given by the formula:

$$D(X,Y) = \frac{1 - S_{11}S_{00} + S_{10}S_{01}}{2((S_{10} + S_{11})(S_{01} + S_{00})(S_{11} + S_{01})(S_{00} + S_{10}))^{1/2}} \tag{1}$$

The distributions of distances in both the same-writer and different-writer categories follow a univariate Gaussian density. During training the parameters of these distributions are estimated: the mean μ_{SW} and variance σ_{SW} for same-writer category and the mean μ_{DW} and variance σ_{DW} for different-writer. They define the two densities $p_{SW}(x)$ and $p_{DW}(x)$.

During matching, for each character $c_i, i = 1, \cdots, N$, where N is the number of characters considered, we compute the distance d_i^j between the two samples of pair j for that character. We can have $j = 1, \cdots, M_i$ possible pairs of samples for a given character. For characters $c_i, i = 1, \cdots, N$ and M_i pairs of samples for each character we estimate the log-likelihood ratio:

$$LLR(micro) = ln\left(\frac{\prod_{i,j} p_{SW}(d_i^j)}{\prod_{i,j} p_{DW}(d_i^j)}\right) \tag{2}$$

If $LLR(micro) > 0$, we have a same-writer decision, if not we have a different-writer decision.

To match documents based on micro-features, it is necessary to recognize characters, either manually or automatically, so that matching can be performed between the same characters. Character sizes are estimated knowing average height of text lines and word information. The estimates are used to filter connected components, so that those of appropriate size are candidate characters. In the case of cursive writing touching characters are separated using a word recognizer. The word recognizer takes the word image and its text transcription to segment into characters before sending them to the character recognizer. Word transcription is made in one of two ways: user types the content of each word during document registration time using an interface or using automatic transcript mapping functionality. This allows a pre-typed transcript being automatically read in and the content of each word matched with the corresponding word image automatically. Other geometric information is obtained at the document image processing stage.

3.2 Macro-feature Similarity

The macro features are mapped into a distance vector of differences. The distance distributions in both the same-writer and different-writer categories using the macro-features follow the same univariate Gaussian density of the form we used above for that of micro-features. Similar to the training and testing for using micro-features, the distribution coefficients are computed. The verification decision for any pair of documents using macro-features is made also using formula (2) above resulting in $LLR(macro)$ where the distance are measured using absolute differences for each of the real-valued macro features. The degree of match between two documents is measured using the total LLR score as follows: $LLR(micro) + LLR(macro)$.

This approach to similarity measurement is similar to the probabilistic model of information retrieval [10]. In this model, the LLR gives a relevance/similarity measurement for document retrieval.

As a document retrieval system, CEDAR-FOX creates a handwritten document database through its rich set of tools. For each handwritten document image, the system collects all the related information including the document image itself, the features for matching, the region of interest (ROI) that is selected manually by the user and all the possible meta-data. For collecting the information, the system provides a rich set of interactive tools for user to specify any local details in the document image. Among the tools, there is an easy data entry function with which user can type in a transcript of the document to match the image, a easy access tool for fixing automatic segmentation problems by merging or splitting word images and a modification tool for fixing any character recognition problem.

3.3 Query Methods

Based on Meta-data. Meta-data are text data such as identification number, writer and other collateral information. In real forensic applications, there are often text data related to a handwritten document image. They include the time and date the document collected, descriptions about the case, keywords for efficient text search and registration number as identification. Other useful information can be the possible linkage to any known case, know document and the author of the document. The system provides easy data entry tools to be able to add to the database tables the meta-data that the user types in. The meta-data will be then be stored as a record corresponding to the document. Thus a document in terms of a database entity can be considered as a single record or a 'tuple'.

CEDAR-FOX provides efficient retrieval of such a database. Several query modalities are permitted for retrieval. The database functionality has been implemented using MySQL, the database management system by MySQL AB^{TM} and the interface to the database is through the MySQL libraries. From the user point of view, a graphical user interface has been provided which takes as input the fields on which the user wants to query the database of documents.

This query is mainly based on textual information and meta-data. The user can query the database in order to retrieve documents relevant to the meta-data and textual information he uses in the query.

Based on Document Features. Another type of querying is based on the features of the documents and this is where identification comes to play. The process of identification is nothing but querying in which the query consists of a document. Unlike the IR counterpart where the query has to be considered as a pseudo-document for which similar relevant documents are retrieved, here the query is a real document and the task of this information retrieval system is to use the similarity measurements to pull-out documents that it thinks is relevant. In effect, when we give a document as a query, we expect the system to filter out only those documents that it thinks is written by the same writer who wrote the query document. The relevance or similarity of the retrieved document is measured using the similarity metrics presented above in equation (3). The log likelihood ratio is used to decide similarity, just like the probabilistic model of information retrieval does. When the query involves a document rather than textual information entered by the user, there are three possible options the user can use to define what part of the document he wants to use to query the database.

Document Level: The Entire Document Image Is the Query. When the system loads in a document image, it can be directly used as query. For identifying the document from the database, the automatically extracted features are used for the matching. The query returns a ranked list of documents in the library. The scores attached with each document is computed using as much as available information for the query document (Fig.3).

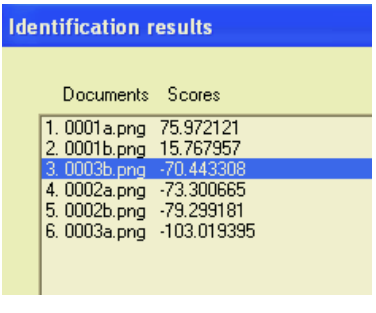

Fig. 3. Query result showing a ranked list of documents

Partial Image: A Region of Interest (ROI) of a Document. A document image may include many text or graphical objects. User often needs to specify a local region of the most interest. Using a system cropping tool, user can easily crop a rectangular region and use the ROI as his query.

Word Level Image: Any Word in the Document. Any word from the query document can be cropped and compared against other documents in the database. The result will be a ranking of he words in other documents that are most similar with the query. The original document containing the retrieved words can easily be obtained from the database.

Based on Text Keyword. The user can type in any keywords ranging from the words in the documents, case number, person names, time and the pre-registered keywords such as brief descriptions of the case. The text identification is done by matching between the query text words and the text in the digital library. The matching considers the priority of the information represented by the words. The distance measure is edit distance based.

4 Results and Analysis

4.1 Experimental Framework

The document image set consists of 3000 samples written by 1000 writers, where each writer wrote three samples of a preformatted letter. From each document images of numerals and alphabets were extracted and represented as 512-bit binary vectors. The distance between two character images is given by a real-valued similarity distance between the coprresponding feature vectors. Real-valued features (macro) were also extracted for the documents.

The writer identification performance is evaluated for two scenarios: (i) same-content - the documents being compared have the same text content and (ii) different-content scenario - the documents being compared have different text content. This distinction is important, since in most cases the documents to be compared are presumed to have different content.

Same Content. In this scenario, the 3000 document set is divided into a training set of 2000 documents and a test set of 1000 documents. Therefore, each writer has two documents in the train set and one document in the test set.

Different Content. For each of the 3000 documents, an imaginary center line is computed and used to split the document into an upper and lower half. The features described before are extracted from both halves. Two similar sets of 2000 and 1000 document images are built randomly selecting half-images for each writer and assigning them to the sets. In the comparison process we ensure that only different halves are being compared. Some of the old macro features (entropy, threshold, number of black pixels and average height) could not be computed for this scenario and were simply not taken into consideration in the final combination mix.

For both scenarios we have used a weighted version of the k-nn classifier for identification.

4.2 Results

The document management system CEDAR-FOX on which these experiments are conducted is able to perform document retrieval at the document image, document word image or text keyword level. Document image retrieval, for which the experimental results presented here have been obtained, provides the user with an automatic and efficient way to identify a questioned document from a large set of known documents in the database. Each document image in the test set is compared with every document image in the train set. The comparison is done in the feature space, using some or all of the previously described features.

We have estimated both the individual and accumulated writer identification performance: (i) individual performance of each feature, (ii) accumulated performance of previously proposed features, newly proposed ones and the character-level features (extracted from all 62 characters for the same content scenario and from the common characters only in the different content scenario). We previously reported results for 12 macro features in [7] and shown the performance to be 99% for 2 writers and about 59% for 900 writers. Along with 10 character features the identification performance was shown to be 87%.

The individual and accumulated writer identification performance results obtained using the newly proposed features are presented in Table 1. As expected, the same-content performance is much better than the different-content performance. We note that to obtain these values we identify the writer of the top retrieved document image as our result. If we retrieve the top 5 or top 10 document images (out of the set of 2000) or we combine these features with others the performance becomes significantly better (almost 70%).

Table 1. Individual and accumulated writer identification performance of proposed features for same and different content

Features	Identification	
	Same Content	Different Content
Lexeme	16.36	3.98
Character	9.01	7.05
Word Legibility	1.33	0.20
Inter-Character Distance	0.72	0.41
Char Relative Height	3.48	0.41
Char Relative Slant	2.56	1.02
Proposed Features	35.62	11.34

Fig.5 presents the retrieval results for accumulated features in the same-content and different-content scenarios. The first plot diplays the precision-recall curves obtained for 50 features (old+new + numerals + some lower-case characters) when varying the number of top images retrieved from 1 to 20. For the second plot we maintain fixed the number of top retrieved images (top 10) and vary the number of features used (1-50). As we can see, the writer identification performance in the same-content case goes about 90% when the top result is returned. These results clearly depict the usefulness of our system in querying the

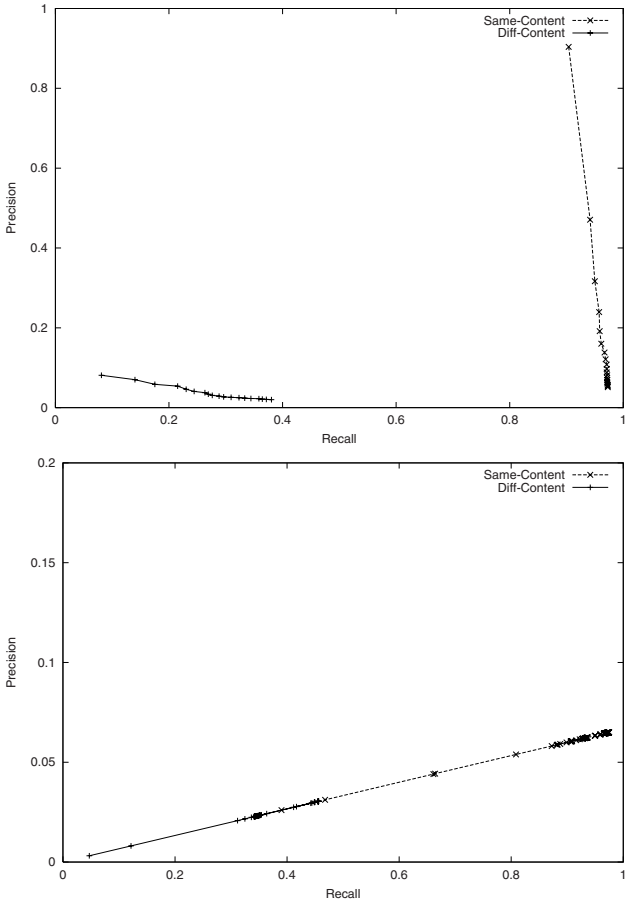

Fig. 4. Precision-recall curves obtained for same-content/different-content scenarios for (a) fixen number of features (50) and variable number of top choices considered (1-50) and (b) variable number of features (1-50) and fixed number of top choices considered (15)

large document collection to return one or more documents that match closely with the given query document.

We can also observe that: (i) the performance of the proposed features is highly dependent on the type of comparison scenario (same-content vs. different-content); (ii) while individually some features perform much better than others, each one brings its own contribution to a superior accumulated performance.

5 Conclusion

We have described the information retrieval aspects of a document analysis and management system for handwritten documents. Writer characteristics as well as document content and meta data can be used for retrieval. The performance

of the system using a large set of document and character-level features has been presented.

Acknowledgments

CEDAR-FOX is the result of the effort of many students and researchers at CEDAR. Others who have contributed to the system implementation and its testing are Zhixin Shi, Harish Srinivasan, Bandi Kartik Reddy, Meenakshi Kalera, Devika Kshirsagar, Siyuan Chen, Aihua Xu, Vinay Shah and Vivek Shah.

References

1. Osborn, A.S. (1929), *Questioned Documents*. Nellon Hall Pub.
2. Robertson, E. W. (1991), *Fundamentals of Document Examination*, Burnham Inc Pub.
3. Bradford R. R. and Bradford R. B. (1992), *Introduction to Handwriting Examination and Identification*, Burnham Inc Pub.
4. Hilton, O. (1993), *Scientific examination of questioned documents*, CRC Press Inc.
5. Huber, R.A. and Headrick, A.M. (1999), *Handwriting Identification: Facts and Fundamentals*, Boca Roton: CRC Press.
6. Franke, K., Schomaker, L., Vuurpijl, L., and Giesler, St. (2003), *FISH-new: A common ground for computer-based forensic writer identification*. Proceedings of the Third European Academy of Forensic Science Triennial Meeting, Istanbul, Turkey, p. 84.
7. Srihari, S. N., S-H Cha, Arora, H. and Lee, S. (2002), *Individuality of Handwriting*, Journal of Forensic Sciences, 44(4): 856-72.
8. Srihari, S. N., Zhang, B., Tomai, C., Lee, S-J., Shi, Z., and Shin, Y. C. (2003), *A system for hand-writing matching and recognition*, Proceedings of the Symposium on Document Image Understanding Technology (SDIUT 03), Greenbelt, MD.
9. Zhang, B. and Srihari, S. N. (2003), *Binary vector dissimilarity measures for handwriting identification*, Kanungo, T., Smith, E. H. B., Hu, J. and Kantor, P.B., eds., Document Recognition and Retrieval X, Bellingham, WA: SPIE, 5010: 28-38.
10. Sparck Jones, K. (1998), *A Probabilistic Model of Information Retrieval: Development and Status*, Technical Report, Computer Laboratory, University of Cambridge, UK.