Open access • Proceedings Article • DOI:10.1109/DIAL.2004.1263248

Forensic handwritten document retrieval system — Source link

Sargur N. Srihari, Zhixin Shi

Institutions: State University of New York System

Published on: 23 Jan 2004 - Scopus (IEEE Computer Society)

Topics: Document layout analysis, Document retrieval, Visual Word, Word processing and Document clustering

Forensic Handwritten Document Retrieval System

Sargur N SRIHARI and Zhixin SHI
[+]Center of Excellence for Document Analysis and Recognition (CEDAR),
University at Buffalo, State University of New York, Buffalo, USA Srihari@cedar.buffalo.edu

Abstract.

Document storage and retrieval capabilities of the CEDAR-FOX forensic handwritten document examination system are described. The system is designed for automated and semi-automated analysis of scanned handwritten documents. For library creation, the system provides functionalities for (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. For extracting these features the system performs page analysis, page segmentation, line separation, word segmentation and finally recognition of characters and words. The extracted features are used for writer identification by matching against a library built as a database. The system design is driven by questioned document examination with its emphasis on writer identification. Several query modalities are permitted for retrieval: (i) document level: the entire document image is the query; (ii) partial image: a region of interest (ROI) of a document; (ii) word image: which is also called word spotting; (iv) text keyword: the user can type in 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 system has been implemented using Microsoft visual C++ and tested using MySQL database system from MySQL ABTM. It provides as a graphical user interface for forensic document identification, verification and analysis.

1. Introduction

Writer identification has a long history perhaps dating to the origins of handwriting itself. There exist many textbooks describing the methodology employed by forensic document examiners, e.g., Osbom, 1929, Robertson, 1991, Bradford, et. al., 1992, Hilton, 1993, Huber and Headrick, 1999.

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

As other related fields of forensic science, classic forensic handwriting examination is primarily based upon the knowledge and experiences of the forensic expert. Due to the problematic situation of non-objective measurements and non-reproducible decisions, traditional methods have been recently attempted to be supported by computerized semi-automatic and interactive systems. A computer system for retrieving a small set of documents from a larger set, known as the FISH system, has been developed internationally [Franke et.all, 2003].

Modern computer technology gives ways of building large criminal databases from which investigators could retrieve criminal records containing similarities in each of their cases for finding suspects. Digital database or digital library technology for forensic applications plays a very important role in criminal investigations and homeland security. With efficient computer system and vast among of data, forensic investigation has become more efficient and accurate. For various of types of data collected from criminal records or evidences, forensic database are of many types such as fingerprint databases, criminal record databases and multimedia record collections including photos and text documents.

A complete system for forensic handwriting document management, known as CEDAR-FOX, has been developed for handwritten document analysis, identification, verification and document retrieval [Srihari, et. al. 2003]. A digital library for forensic handwritten document is built using the system. Unlike the FISH system, the system design of CEDAR-FOX is driven by writer verification. As a document management system for forensic analysis, it provides user three major functionalities. First it can be used as a document analysis system. Second it can be used for creating a digital library and third it can be used as database management system for document retrieval and writer identifications.

As a document analysis system, CEDAR-FOX provides user a graphic user interface. It can load or scan a handwritten document image. The system will first automatically extract features based on an intensive document image processing and recognition. The user can then use the many graphic tools provided by the system to do document examination. They include viewing automatically extracted document characteristics and do verific ations between different documents.

CEDAR-FOX provides a full range database management tools for creating a handwritten document library. For library creation, the system has functionalities for (i) entering document meta-data, e.g., identification numb er, 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 easy access and retrieval functionalities for adding, mo dification and categorization of the document records in the digital library.

CEDAR-FOX provides document retrieval into query modalities: (i) the entire document image is the query; (ii) partial image: a region of interest (ROI) of a document; (ii) word image: which is also called word spotting; (iv) text keyword: the user can type in 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.

Section 2 describes the functionality provided by CEDAR-FOX for handwritten document analysis. This includes automatic recognition for identification of handwritten characters, feature extraction and verification and identification methodology. Section 3 describes the system capability to prepare handwritten document and other forensic data for the creation of a digital library and the functionality for document retrieval. Section 4 gives results of system experiments and tests. Concluding remarks are presented in Section 5.

2. Recognition and Identification

Research for traditional recognition and analysis of handwritten documents has been emphasized on identification and understanding the content of the documents. The basic obstacle to overcome in optical character recognition (OCR) research is the variations of a same character written by different individuals as well as by same writer at different time. The character recognition features for OCR are designed to minimize as much as possible the distance within these variations for the same characters. While the focus of writer identification is placed on distinguishing the difference between the writings from different writers. The content of the handwritten document becomes less important. The writer's individuality in handwriting is practically assumed and has been adopted in forensic science and practice for many years.

Related to the issue of establishing the individuality is the need for associating a quantitative measure of similarity between two samples. Such a quantitative measure brings in an assurance of repeatability and hence a degree of objectivity. We have developed a theory and tested the theory with experiments for the task of handwriting identification as well as the specific subtask of handwriting verification (one where the task is to determine whether two documents were written by the same individual or by different individuals) [Srihari, et. al., 2002].

Identification of a document is to take the document as a query to compare with many or all of the document data in the digital library. The compressions or matching between the query document and each document in the digital library are performed using many attributes of the documents. Handwriting matching is a two-class problem by measuring a "dis tance" between two document samples and determining whether that distance could be classified as being within the same class or between different classes. For this two-class classification problem we designed and computed two sets of features, known as macro- and micro-features, from the writing elements within the scanned and segmented handwritten images. The macro-features capture the global characteristics of writer's individual writing habit and style. The macro-features are document features therefore they are extracted from the entire document. The micro-features are designed to capture the finer details at character level. To compute the micro-features, there is a need for identifying the handwritten characters, manually or automatically, so that the matching will be done between different documents on the same characters. The identification of the characters is to ensure the comparisons are not of orange-to-apple.

2.1 Character and Word Recognition

Although for the purpose of writer identification, recognition for the content of handwritten document is no longer the top priority, identification and recognition of document components including those belonging to the same character category are always necessary and important. Forensic document examiners use writing characteristics pertaining to certain characters in comparing documents. For example, one of the features they often look for is the lower loop in characters "g" and "y". In the earlier version of our system one of the functionality provided for document examiners is easily cropping out certain character images manually. Then a set of micro-features will be computed for the cropped images for comparison. The manual-cropping tool is useful but it is also limited by its efficiency. Even in a small handwritten document, the number of characters could be too big for the characters being manually cropped. For identifying a questioned document from a large database of documents, automatic identification and recognition of handwritten document components especially isolated characters are very crucial for document examination system designs.

CEDAR-FOX has the capability of using automatic identification and recognition for isolated handwritten characters. For an input handwritten document scanned with resolution 300dpi as a gray-scale image, the image processing starts from converting the image to binary form which has two levels with 0 for black and 255 for background white colors. We adopted the algorithms developed in our earlier developed PENMAN system for page layout analysis [Srihari and Kim, 1997]. The document block segmentation and text line separations are performed followed by text word segmentations. At the same time as the document image undergone these processing, a set of global features known as macro-features at the document level are computed.

As the document text line and word information being available, we estimate the sizes of the handwritten characters using the average height of the text lines. The estimation is also combined with a connected component analysis algorithm, which filters the connected components with appropriate size being possible isolated characters.

Since handwritten document include mostly handwritten cursive script, characters in the words are touching or connected to their neighboring characters. To separate the characters from these words, we implement a word recognizer in the system. The word recognizer utilizes each individual word images, together with the text information of the words, to segment the characters before send them to the character recognizer. The text information for a word can be the content of the word, the geometric information of the word such as the size of the word and the number of characters in the word. There are two ways for the system to get the text information. One is that the system provides the user an easy graphic interface to type in the content of each word at the document registration time as data entry. The other way for getting the contents of the words is through an automatic transcript mapping functionality of the system. This allows a pre-typed transcript being automatically read in and the content of each word also matched with the corresponding word image automatically. Other geometric information is obtained at document image process stage.

For the components that are possible characters, we apply a character recognizer that we designed earlier for recognizing handwritten character using GSC features [Srikantan, et. al., 1996] on these components. The classifier in the recognizer is a k-nearest neighbor classifier. It is trained by using a separate training set of handwritten characters. The recognition results consist a ranked list of character identities each coupled with of a confidence as a measurement for each recognized component being the identity. The ranking is on the confidence values. Therefore the top choice in the recognition result shows the best possible character identity for the component. Components with the low top confidence are filtered out by using a fixed threshold. The rest of the components together with their recognized ids are going to be used for document identification or verifications.

2.2 **Writer Identification**

Writer identification in CEDAR-FOX is based on pairwise handwritten verification recently described [Srihari, et. al., 2002]. Central to handwriting matching is the need for associating a quantitative measure of similarity between two samples. Such a quantitative measure brings in an assurance of repeatability and hence a degree of objectivity. We take handwritten document verification as a two-class classification problem. Two types of features are extracted from the handwritten documents for classification. The first type of features known as micro-features is a binary string extracted from identified and recognized character components. The second type of features known as macro-features are extracted from processing of the document images for global characteristics of writer's individualities.

2.2.1 Micro-features and their similarity measurement

The micro-features are the GSC features we used in identification and recognition of character components. They are obtained by software previously developed for handwriting recognition [Srikantan, et. al., 1996]. The micro-features 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 x 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 converted from the original floating number computations.

Several methods for comparing strings of binary feature vectors representing handwritten characters have been recently evaluated [Zhang and Srihari, 2003]. This has led to the choice of a correlation measure as being the best of binary string matching measure. To measure the similarity between two binary vectors we use the Correlation measure defined in [Tubbs, 1989]. 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:

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

We have collected from a total of 1500 writers 3 documents for each writer. For the same-writer category the writer set has been divided into two sets: 500 for training and 500 for testing. Therefore, for each character we have 3X500 distances between samples belonging to the same writer in each of the training and testing sets. For the different-writer set we have randomly chosen 1500 pairs of documents (from different writers) from the first 500 writers for each of the training and test sets. The distributions of distances in both the same-writer and different-writer categories follow a univariate Gaussian density of the following form:

$$p(x) = \frac{1}{(2\pi)^{1/2}} \exp^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad (2)$$

For training we estimate the parameters of these distributions. The pairs of values mean μ_{SW} and variance σ_{SW} for same-writer category and μ_{DW} and variance σ_{DW} for different-writer category have been computed over the 3000 distances in the training set (for both same-writer and different-writer categories), for each character (micro feature). They define the two densities $p_{SW}(x)$ and $p_{DW}(x)$.

In the testing phase, for each character $c_i, i = 1,..., N$, 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 $j = 1,...,M_i$ possible pairs of samples for a given character. For characters $c_i, i = 1,..., N$ and M_i pairs of samples for each character we estimate the log-likelihood ratio:

$$LLR(micro) = \ln\left(\frac{\prod_{i,j} p_{SW}(d_i^j)}{\prod_{i,j} p_{DW}(d_i^j)}\right) \quad (3)$$

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

2.2.2 Macro-features and their similarity measurement

The macro-features, which represent the entire document, consist 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 pixe ls 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.

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 (3) above resulting in *LLR(macro)* where the distance d_i^j are measured using absolute differences for each real-valued macro features.

2.2.3 Identification decisions

The final discrimination between two documents is done based on *LLR(micro) + LLR(macro).* Writer discriminability is highest when all of the macro-features are used. It is least when only ten numerals (digits) in the handwritten document are used[Zhang, Srihari and S-J Lee, 2003]. According to this discriminability distribution among characters, we apply a weight coefficient from the corresponding distribution value to each computation of the Gaussian density in (2), bigger the coefficient is the stronger the discriminability the corresponding character has. This ends up better system performance by using decision function.

3. Document Storage and Retrieval

3.1 Creating Handwritten Digital Library

As a forensic document management system, CEDAR-FOX creates forensic handwritten document digital library 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 document examiners 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.

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 provide easy data entry tool as tables that the users can type the meta-data in. Then the meta-data will be organized as part of a database record to be saved into the digital library.

The original handwritten document image as a gray scale image and a binarized binary image including segmentation information about the words and characters are compressed and to be saved. The original image is kept as its scanned form is for any possible further additional analysis. The binary image is saved for showing the word and character level details.

One of the most important parts of the data is the set of extracted document features. The organization of the features is a tree structure. Starting from the document layout, the character level micro-features are in data cells holding characters together with the image information. A multiply tagged file format called FOX format is designed similar to the tiff images format. The FOX file format allows storing all the extracted information from the meta-data, the document image, the ROI to all the extracted features efficiently for retrieval. The FOX file format can be easily made to be XML compatible.

3.2 Document Retrieval

A forensic digital library is a server on a computer network that can respond to user requests by retrieving relevant data. CEDAR-FOX provides efficient retrieval of such a digital library. Several query modalities are permitted for retrieval. The system has been implemented using Microsoft visual C++ and tested using MySQL database system from MySQL ABTM. It provides as a graphical user interface for forensic document identification, verification and analysis.

i. 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 digital library, 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. 1.)

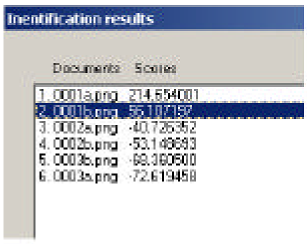

Fig 1. Query result showing a ranked list of documents.

ii. 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. When a ROI is taken as query, the identification will be done only using the features within the selected ROI (Fig. 2.)

iii. Word image: which is also called word spotting

Word spotting is very useful when a word in a document image shown some special features in its shape or contents. User then can crops the word image from the document and run a key word search on image level. All the words of similar shape or content will be returned with similarity rankings (Fig. 3.)

The method of word spotting implemented here [Zhang Srihari and Huang, 2004], based on generalizing the GSC feature extraction for words, is much faster than previously known methods based on dynamic time warping [Rath and Manmatha, 2003].

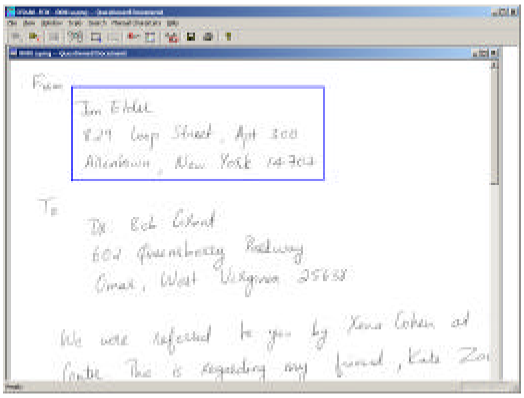

Fig 2. User can select a region of interest (ROI) using a cropping tool.

Identification of the query word image is done by matching the query word image with the word images in the digital library. The matching between two words images use a set word level features that include the shape and content characteristics of the words.

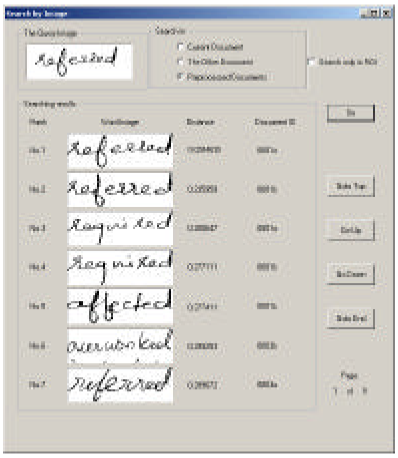

Fig 3. Cropped Word image as query to the digital library; all the similar word images are returned as a ranked list.

iv. 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 Implementation and Experiment

CEDAR-FOX is implemented in a Windows environment. It has a number of interactive features that makes it a useful as a document management tool: image displays [Srihari, et. al. 2003], automatic feature extractors for both manual cropped character images and automatic identified characters. The automatic identified characters are either directly from document component or from segmentation of the words in the documents. For creation of a forensic document digital library, the system provides tools for data entries. All the text information then collected together and saved in a digital library with the document features. A handwritten document digital library is built including 3000 document images from 1000 writers.

Document retrieval uses several query models. Retrieval can be done using a document image, any word in a document or text keywords that user type in. Document retrieval provides user an efficient way to identify a questioned document from among a large set of known documents in the digital library automatically. Our recent experiment result showed the performance for writer identification using a document set from 975 writers is as follow: using only macro features, the identification rate is around 60%; using macro feature plus micro features extracted only from 10 characters, the identification rate is 89.1%; using macro features and micro features extracted from all 62 classes, the identification rate is 97.94%.

5 Conclusion

We have described a document analysis and management system for computer processing of handwritten documents. It incorporates a system for handwriting identification based on learning from the differences between exemplars of the same individual and between those of different individuals. Handwriting recognition and writer identification are coupled by utilizing recognition results in identification and by using identification results to retrieve and employ writer characteristics in recognition. The CEDARFOX system for handwriting retrieval has a number of user-definable features and it is capable of producing confidence values in matching. Ongoing work at CEDAR will refine each of the functionalities described.

Acknowledgment

We would like to thank our collaborators in the Forensic

Document Analysis project group, in particular, Yong Shin, Catalin Tomai, Chen Huang, Anantharaman Ganesh, Harish Srinivasan, Gautam Srihari and Meenakshi Kalera.

References

Osborn, A.S. (1929), *Questioned Documents,* Nellon Hall Pub.

Robertson, E. W. (1991), *Fundamentals of Document Examination,* Burnham Inc Pub.

Bradford R. R., and Bradford R. B., (1992), *Introduction to Handwriting Examination and Identification,* Burnham Inc Pub.

Hilton, O, (1993), *Scientific examination of questioned documents,* CRC Press Inc.

Huber, R.A., and Headrick, A.M., (1999), *Handwriting Identification::Facts and Fundamentals*, Boca Roton: CRC Press.

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.

Srihari, S. N., B. Zhang, C. Tomai, S-J. Lee, Z. Shi and Y.C. Shin, (2003), A *system for handwriting matching and recognition*, Proceedings of the Symposiu m on Document Image Understanding Technology (SDIUT 03), Greenbelt, MD.

Srihari, S.N., S-H Cha, H. Arora and S. Lee, (2002), *Individuality of Handwriting*, Journal of Forensic Sciences, 44(4): 856-72.

Srihari, S.N. and Kim, G., (1997), *PENMAN: A system for reading unconstrained handwritten page images*, Proceedings of the Symposium on Document Image Understanding Technology (SDIUT 97), Annapolis, MD, pp. 142-153.

Zhang, B., and Srihari, S. N., (2003), *Binary vector dissimilarity measures for handwriting identification*, T. Kanungo, Smith, E. H. B., Hu, J. and Kantor, P.B., eds., *Document Recognition and Retrieval X*, Bellingham, WA: SPIE, 5010: 28-38.

Srikantan. G, Lam S-W., AND Srihari, S. N., (1996), *Gradient-based contour encoding for character recognition*, Pattern Recognition, 29: 1147-60.

Tubbs J.D., (1989), *A note on binary template matching*. Pattern Recognition, 22(4): 359-365.

Zhang, B., Srihari, S. N. and S-J Lee., (2003), *Individuality of handwritten characters,* Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, pp. 1086-1090.

Zhang, B., Srihari, S. N. and Huang C., (2004), *Word Image Retrieval Using Binary Features* Accepted by SPIE, Document Recognition and Retrieval XI, San Jose, California, USA, Jan 18-22, 2004

Rath, T.M. and Manmatha, R., (2003) *, Features for Word Spotting in Historical Manuscripts***,** Proceedings of the Seventh International Conference on Document Analysis and Recognition (ICDAR 2003), Edinburgh, pp. 218-222