Questioned Document Examination Using CEDAR-FOX

Sargur N. Srihari, Barish Srinivasan, Kartik Desai[1]

Abstract: *Handwriting verification casework often involves comparing the writing in a questioned document with samples of known writing. This paper describes the use of CEDAR-FOX, a software tool for questioned document examination, in a case involving extended writing. The different steps involved from scanning the documents to obtaining a nine-point qualitative measure are described. The various algorithms used, along with a demonstration of its. functionalities on the case are also described. The paper serves two purposes: a guide to using a state of-the-art software system for a quantitative analysis of handwriting, and an introduction to the science and technology of the software.*

Reference: Srihari, S. N., Srinivasan, B., Desai, K. (2004, Vol. 18 – reprinted and reformatted). Questioned Document Examination Using CEDAR-FOX. J. Forensic Document Examination, Vol 28, pp. 15-26.

Keywords: handwriting identification, macro and micro-features, feature extraction, word segmentation, log-likelihood ratio (LLR)

1. Introduction

Writer verification is the task of determining whether two handwriting samples were written by the same or by different writers, an essential task for forensic document examiners (FDE's). In FOE terminology, there are both questioned and known documents. Quite often, several questioned and known documents exist for a particular case. The case presented here consists of the following naturally written detailed request specimens:

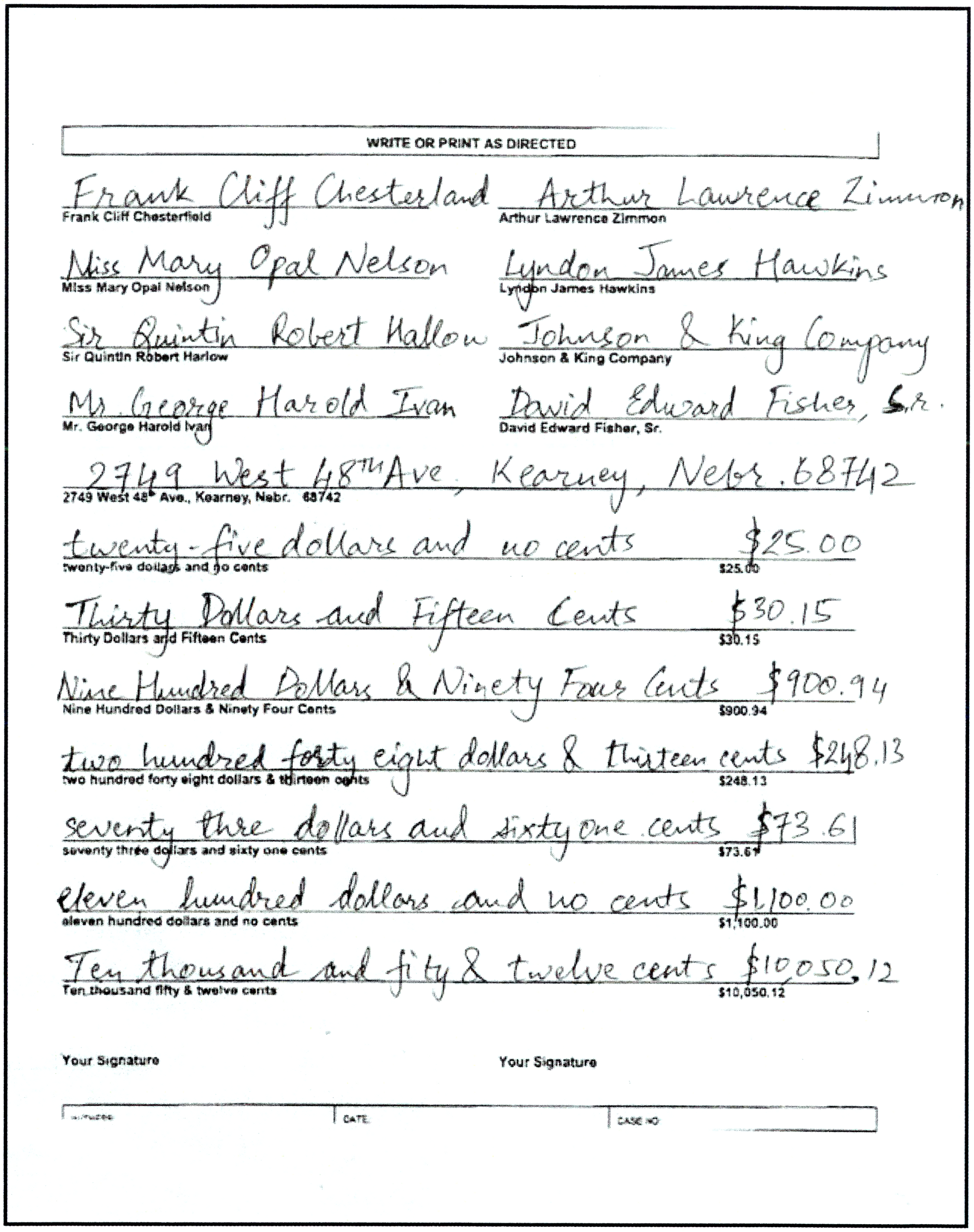
1. Known Documents. The known documents are shown in Figures 1, 2 and 3
2. Questioned Document. The questioned document is shown in Figure 4. This questioned document was indeed written by the same person who wrote the known documents.

The objective of the verification task is to find if the questioned and the known documents were written by the same person. The conclusions of the task are presented in a nine-point scale that is the ASTM standards for opinion terminology currently used by many FDE's in their casework. Computational methods for handwriting analysis have been more recently developed (Plamondon & Lorette, 1989, p 107) (Balacu, Schumaker & Vuurpijl, 2003) (Van-Erp, Vuurpijl, Franke, Schumaker, 2003, p 282) (Srihari, Cha, Arora, Lee, 2002, p 856). When designed as a system they allow conducting large scale and statistically meaningful tests. There are two steps involved in the computational methods for examining questioned and known writings:

1. Document pre-processing and feature extraction.
2. Document comparison.

Both these tasks are extremely complex for the computer to completely automate and hence, CEDAR-FOX software has been designed to be interactive. The rest of the paper is organized as follows. Section 2 describes the various document processing algorithms. Section 3 describes the statistical model used for document comparison, followed by the conclusion in section 4.

[1] Center for Excellence in Document Analysis and Recognition (CE DAR) University at Buffalo, State University of New York Amherst, New York 14228 (srihari@cedar.buffalo.edu)



WRITE OR PRINT AS DIRECTED

Frank Cliff Chesterland
Frank Cliff Chesterfield

Arthur Lawrence Zimmon
Arthur Lawrence Zimmon

Miss Mary Opal Nelson
Miss Mary Opal Nelson

Lyndon James Hawkins
Lyndon James Hawkins

Sir Quintin Robert Hallow
Sir Quintin Robert Harlow

Johnson & King Company
Johnson & King Company

Mr. George Harold Ivan
Mr. George Harold Ivan

David Edward Fisher, Sr.
David Edward Fisher, Sr.

2749 West 48TH Ave, Kearney, Nebr. 68742
2749 West 48th Ave., Kearney, Nebr. 68742

twenty-five dollars and no cents $25.00
twenty-five dollars and no cents $25.00

Thirty Dollars and Fifteen Cents $30.15
Thirty Dollars and Fifteen Cents $30.15

Nine Hundred Dollars & Ninety Four Cents $900.94
Nine Hundred Dollars & Ninety Four Cents $900.94

two hundred forty eight dollars & thirteen cents $248.13
two hundred forty eight dollars & thirteen cents $248.13

seventy three dollars and sixtyone cents $73.61
seventy three dollars and sixty one cents $73.61

eleven hundred dollars and no cents $1,100.00
eleven hundred dollars and no cents $1,100.00

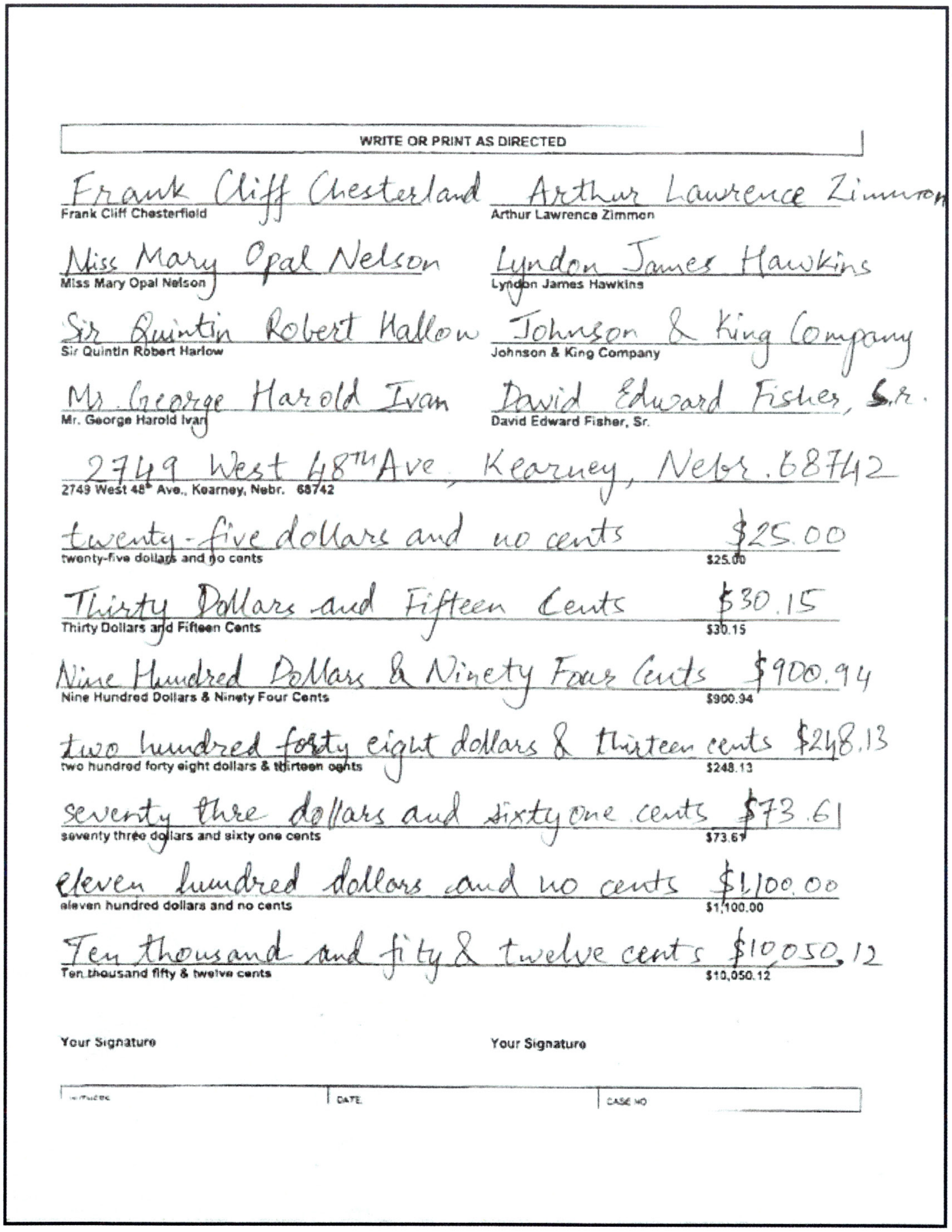
Ten thousand and fity & twelve cents $10050,12
Ten thousand fifty & twelve cents $10,050.12

Your Signature

Your Signature

[illegible] | DATE | CASE NO

Figure 1. Known Document A. Printed text, vertical and horizontal lines are present along with the handwritten text.



WRITE OR PRINT AS DIRECTED

Frank Cliff Chesterland Arthur Lawrence Zimmon
Frank Cliff Chesterfield Arthur Lawrence Zimmon

Miss Mary Opal Nelson Lyndon James Hawkins
Miss Mary Opal Nelson Lyndon James Hawkins

Sir Quintin Robert Hallow Johnson & King Company
Sir Quintin Robert Harlow Johnson & King Company

Mr. George Harold Ivan David Edward Fisher, Sr.
Mr. George Harold Ivan David Edward Fisher, Sr.

2749 West 48th Ave, Kearney, Nebr. 68742
2749 West 48th Ave., Kearney, Nebr. 68742

twenty-five dollars and no cents $25.00
twenty-five dollars and no cents $25.00

Thirty Dollars and Fifteen Cents $30.15
Thirty Dollars and Fifteen Cents $30.15

Nine Hundred Dollars & Ninety Four Cents $900.94
Nine Hundred Dollars & Ninety Four Cents $900.94

two hundred forty eight dollars & thirteen cents $248.13
two hundred forty eight dollars & thirteen cents $248.13

seventy thre dollars and sixty one cents $73.61
seventy three dollars and sixty one cents $73.61

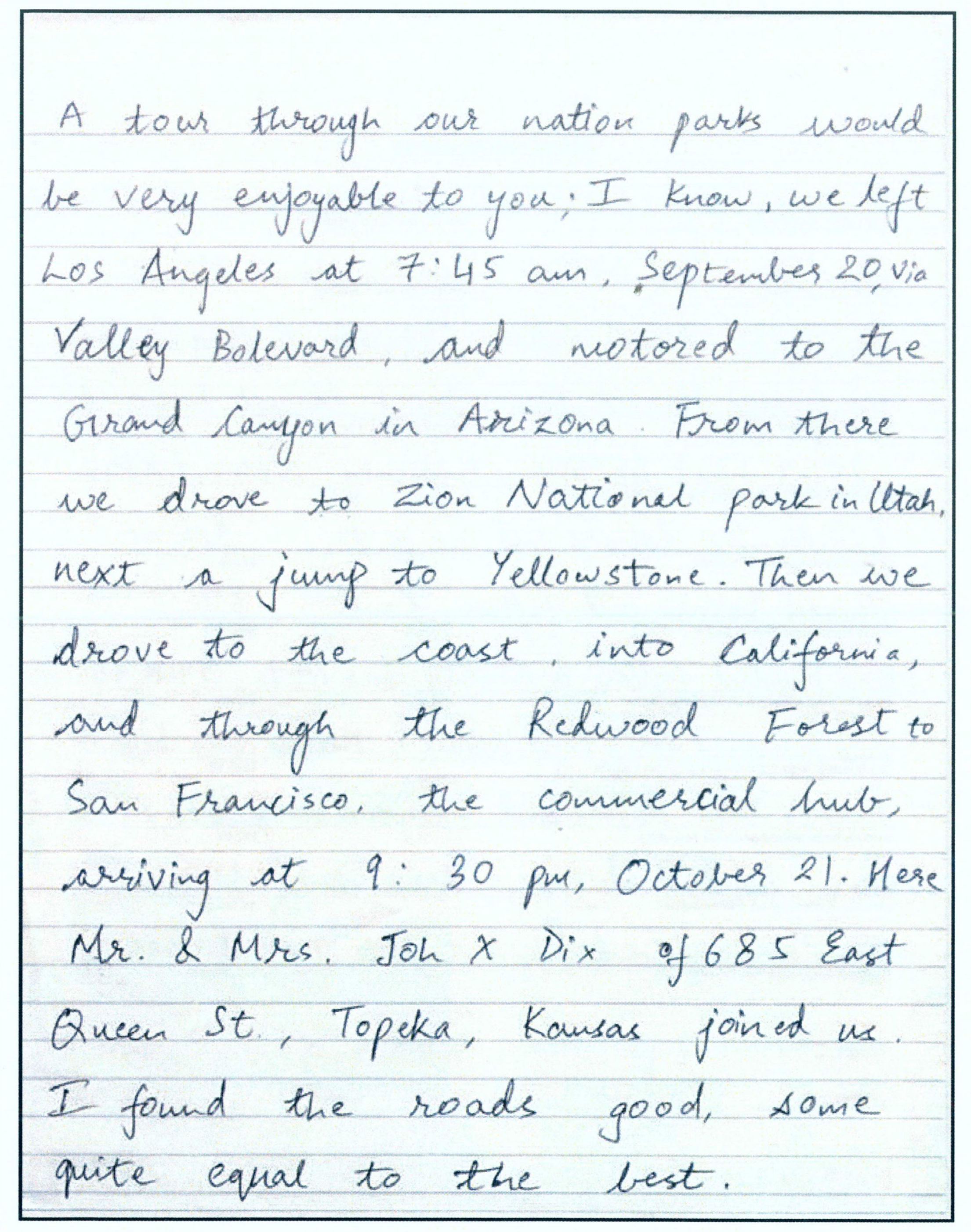
eleven hundred dollars and no cents $1,100.00
eleven hundred dollars and no cents $1,100.00

Ten thousand and fity & twelve cents $10,050.12
Ten thousand fifty & twelve cents $10,050.12

Your Signature Your Signature

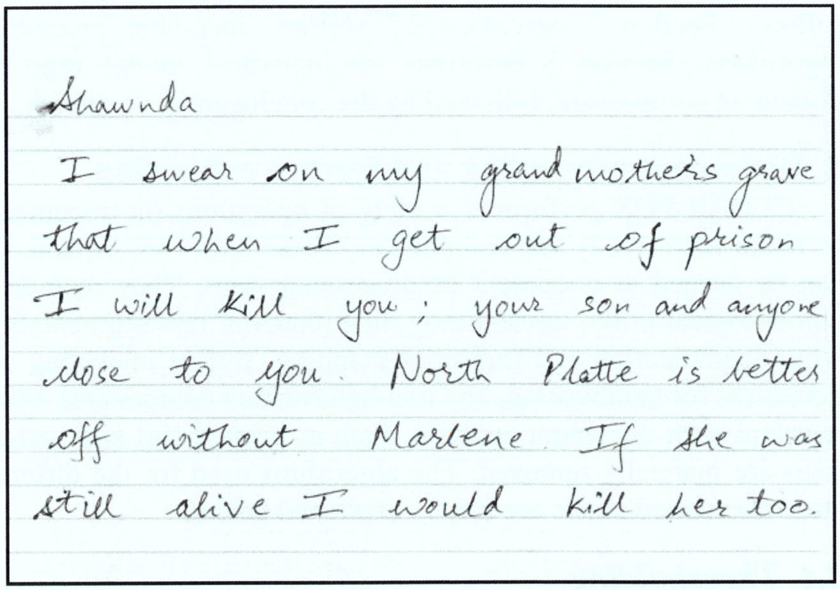
WITNESS | DATE | CASE NO

Figure 2. Known Document B. Printed text, vertical and horizontal lines are present along with the handwritten text.



A tour through our nation parks would
be very enjoyable to you; I know, we left
Los Angeles at 7:45 am, September 20, via
Valley Bolevard, and motored to the
Grand Canyon in Arizona. From there
we drove to Zion National park in Utah,
next a jump to Yellowstone. Then we
drove to the coast, into California,
and through the Redwood Forest to
San Francisco, the commercial hub,
arriving at 9:30 pm, October 21. Here
Mr. & Mrs. Joh X Dix of 685 East
Queen St., Topeka, Kansas joined us.
I found the roads good, some
quite equal to the best.

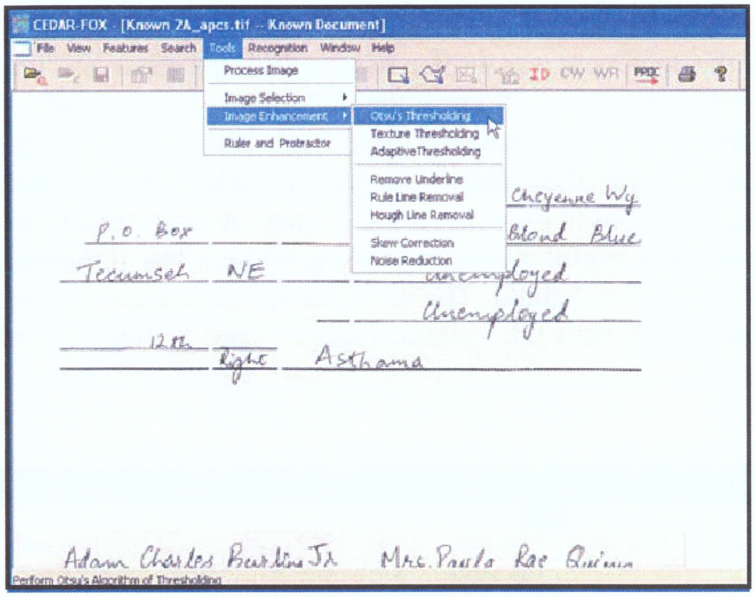
Figure 3. Known Document C. A large handwritten paragraph.



Shawnda

I swear on my grand mother's grave
that when I get out of prison
I will kill you; your son and anyone
close to you. North Platte is better
off without Marlene. If she was
still alive I would kill her too.

Figure 4. Questioned Document. A few lines of text.



CEDAR-FOX - [Known 2A_apcs.tif -- Known Document]

Cheyenne Wy

P.O. Box

Blond Blue

Tecumseh NE

Unemployed

Unemployed

12th

Right

Asthama

Adam Charles Bearline Jr

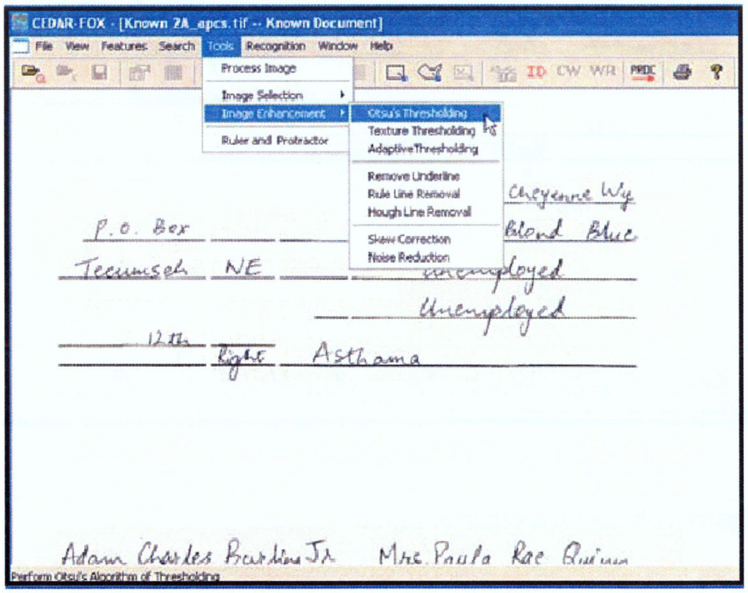
Mrs. Paula Rae Quinn

Perform Otsu's Algorithm of Thresholding

Figure 5. Before Otsu's thresholding.

2. Document processing and feature extraction

CEDAR-FOX performs a variety of operations on documents, to make them ready for comparison. These steps are critical and can be termed as document preprocessing steps. They include in chronological order, thresholding, line removal, line segmentation, word segmentation and transcript mapping. Before analyzing any document for handwriting, any non-handwritten components which interfere with document analysis, such as printed text and vertical lines are manually removed. The algorithms used


CEDAR-FOX - [Known 2A_apcs.tif -- Known Document]

Cheyenne Wy

P.O. Box

Blond Blue

Tecumseh NE

unemployed

Unemployed

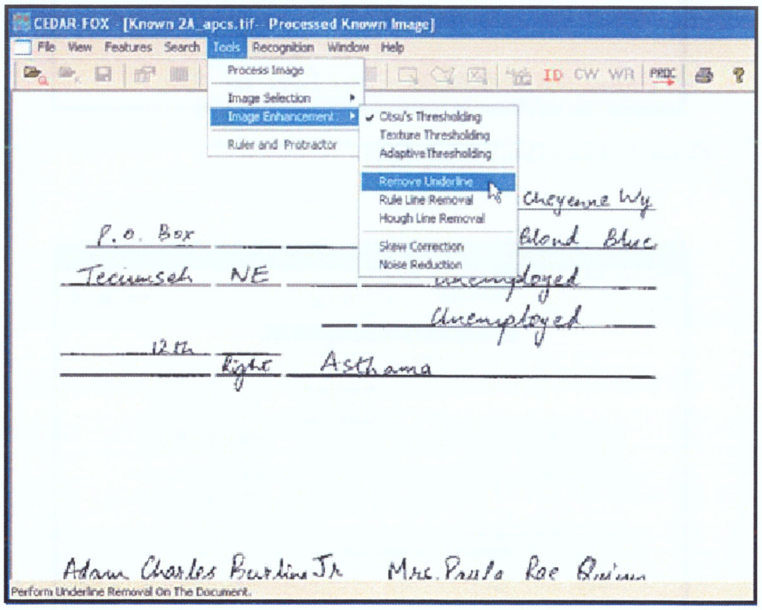
12th

Right Asthama

Adam Charles Burdine Jr Mrs. Paula Rae Quinn

Perform Otsu's Algorithm of Thresholding

Figure 6. After Otsu's thresholding.


CEDAR-FOX - [Known 2A_apcs.tif -- Processed Known Image]

Cheyenne Wy

P.O. Box

Blond Blue

Tecumseh NE

unemployed

Unemployed

12th

Right Asthama

Adam Charles Burdine Jr Mrs. Paula Rae Quinn

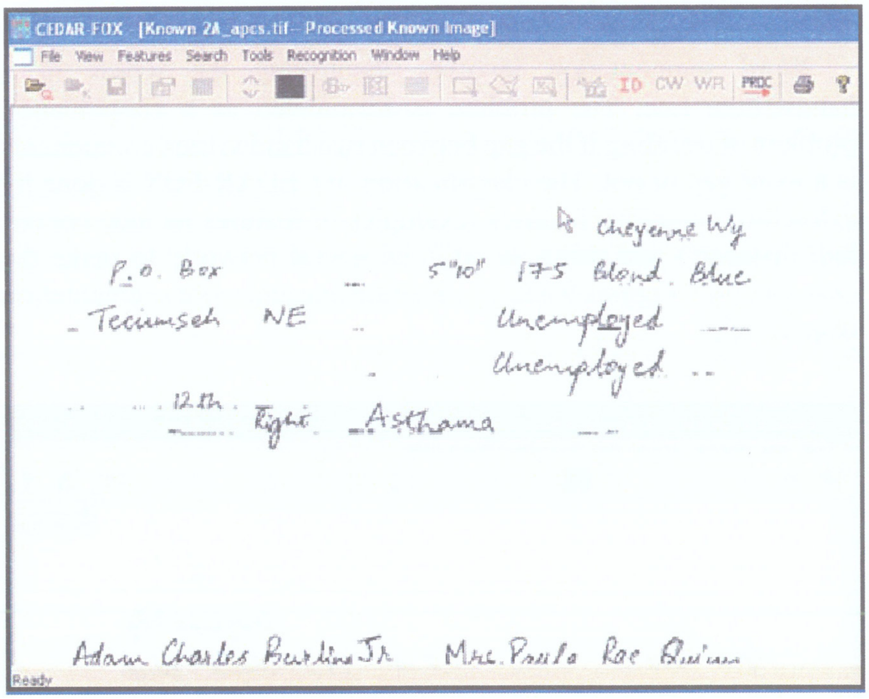
Perform Underline Removal On The Document.

Figure 7. Before Underline Removal.

for the different steps mentioned above are briefly explained below.

2.1 Thresholding

Thresholding converts a gray scale image to binary by determining a value for gray-scale (or threshold) below which the pixel can be considered to belong to the writing, and above which to the background. The operation is useful to separate the foreground (i.e. writing from the background). There is no one universal thresholding algorithm that will work well on different kinds of documents. The system provides for three different thresholding techniques and a method suitable for the task is selected. The thresholding methods used in CEDAR-FOX are Otsu's Thresholding (Otsu, 1979, p.66), Adaptive Thresholding and Texture Thresholding. The Figures 5 and 6 show the


CEDAR-FOX - [Known 2A_apcs.tif-- Processed Known Image]
File View Features Search Tools Recognition Window Help
ID CW WR PROC
Ready

Cheyenne Wy
P.O. Box 5"10" 175 Blond Blue
Tecumseh NE Unemployed
Unemployed
12th Right Asthama
Adam Charles Burlin Jr Mrs. Paula Rae Quinn

Figure 8 After Underline Removal.

case document before and after performing Otsu 's thresholding.

2.2 Line Removal

If the document was written using rule-lined paper, an "underline removal" operation will erase the underlines automatically. Hough transform is used to remove these lines and the system provides for an interface that allows the user to select the threshold that the algorithm uses to remove these lines. A high threshold might remove some useful character strokes, while a low threshhold may leave some lines behind. The user can arrive at a correct threshold by interacting with the system. Figure 7 and 8 illustrate rule line removal.

2.3 Line Segmentation

Line segmentation is an important pre-processing step that separates the handwritten document into lines. The line segmentation algorithm is a very complex and its complete details are described in (Arivazhagan, Srinivasan & Srihari, 2007). Briefly, the steps are highlighted below:

1. An initial set of candidate lines are first obtained.

2. The line drawing algorithm draws lines parallel from left to right and the lines are modeled using bi-variate Gaussian densities1

3. Any obstructing handwritten component is associated with the line above or below by making a probabilistic decision.

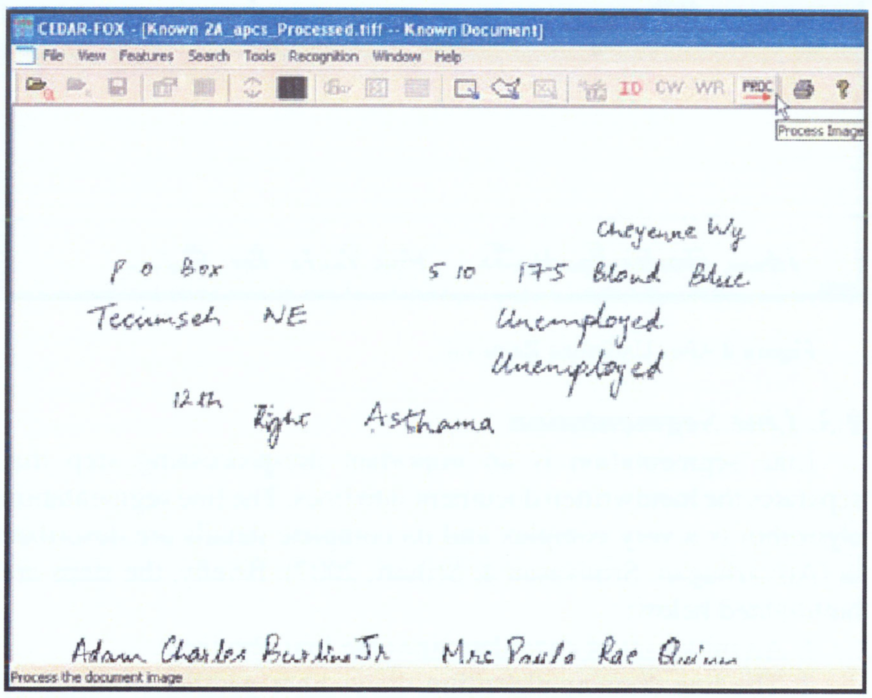
4. The lines are guided by the piece-wise projection profile if available2

2.4 Word Segmentation

Word segmentation involves identifying the words in a handwritten line. The problem is formulated as a classification problem of deciding if the gap between two handwritten components is a word gap or not. The classification in CEDAR-FOX is done by extracting numerous features (examples of features include convex hull distance[3]) and using an artificial neural network[4] to make the classification. Figures 9 and 10 illustrate line and word segmentation respectively.

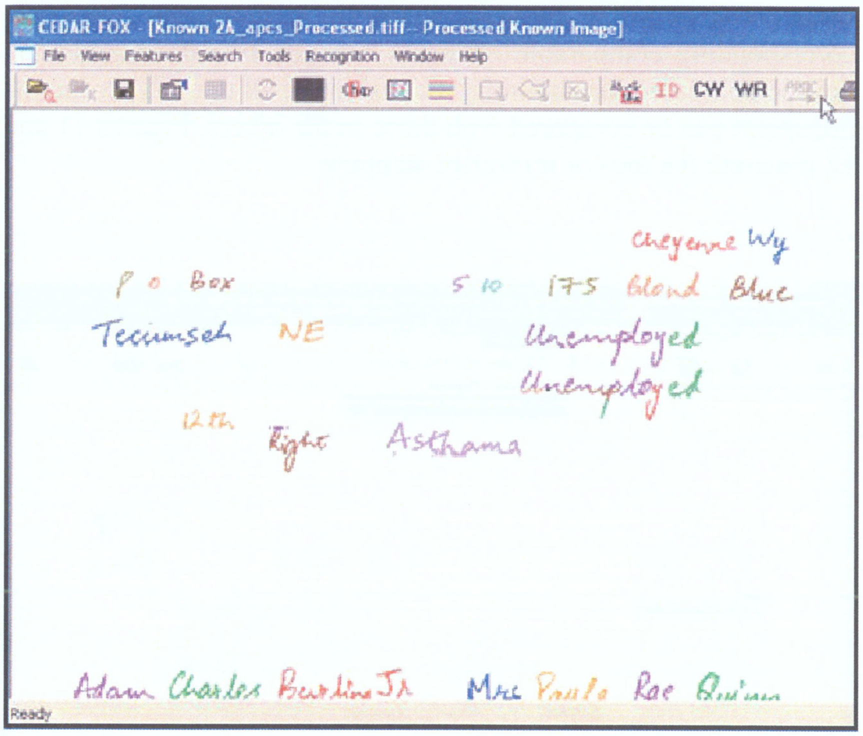
2.5 Transcript Mapping

Transcript mapping performs ground truth assignment by using a text file that contains the


CEDAR-FOX - [Known 2A_apcs_Processed.tiff -- Known Document]
File View Features Search Tools Recognition Window Help
ID CW WR PROC
Process Image
Process the document image

Cheyenne Wy
P o Box 5 10 175 Blond Blue
Tecumseh NE Unemployed
Unemployed
12th Right Asthama
Adam Charles Burdine Jr Mrs Paula Rae Quinn

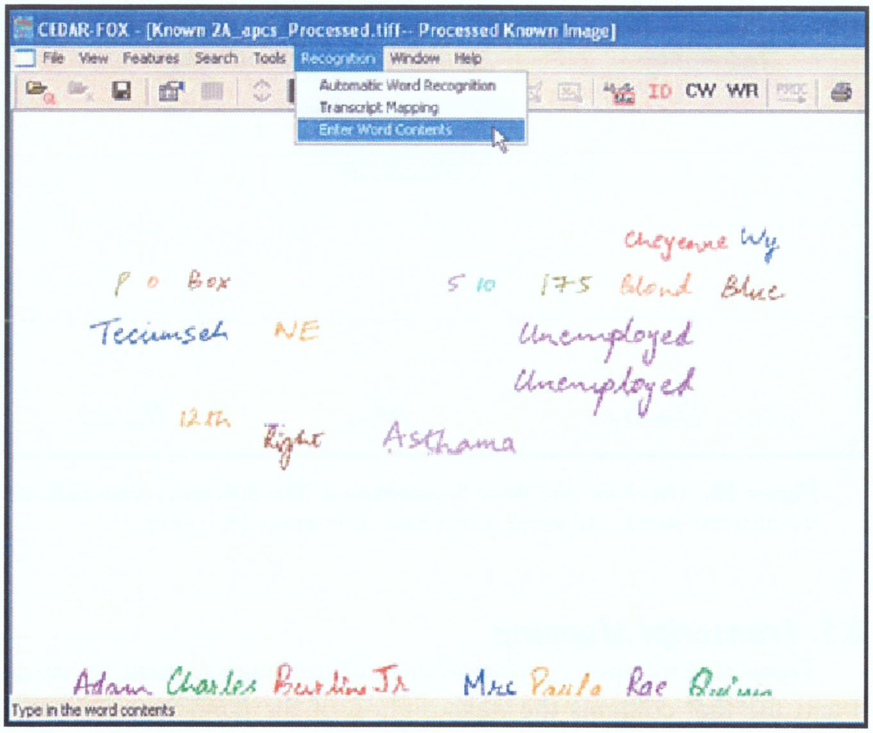
Figure 9 Before line and word segmentation.


CEDAR-FOX - [Known 2A_apcs_Processed.tiff-- Processed Known Image]
File View Features Search Tools Recognition Window Help
ID CW WR PROC
Ready

Cheyenne Wy
P o Box 5 10 175 Blond Blue
Tecumseh NE Unemployed
Unemployed
12th Right Asthama
Adam Charles Burdine Jr Mrs Paula Rae Quinn

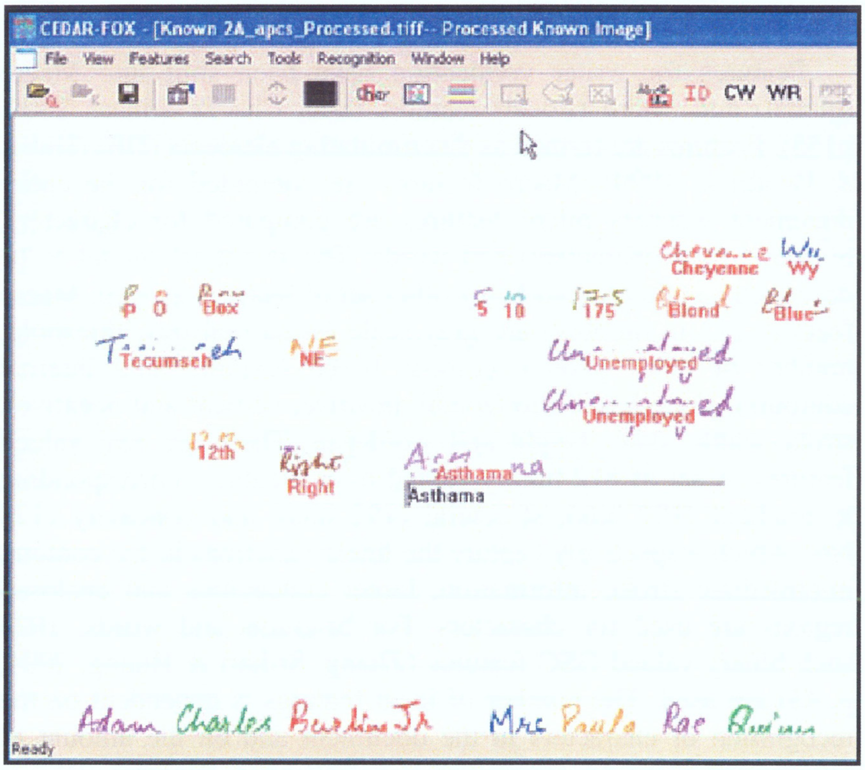
Figure 10. After Line and Word Segmentation. The different colors indicate the different words. Adjacent words have different color coding.

transcription of the handwritten image. It is especially useful when a number of documents share the same text. Quite often in FDE case work, different subjects are required to handwrite the same content. The content is usually similar to that of the known documents shown earlier in figures 1, 2 and 3. In such cases, the transcript of the content to be handwritten is available. The transcript mapping algorithm finds



Type in the word contents

Figure 11. Before Transcript Mapping.



Ready

Figure 12. After Transcript Mapping.

the best word level alignment between the transcript and the handwritten image and associates, a text in the transcript to a word image in the handwritten document. Once this is done, character images can now be extracted from the word image by segmenting the word image. After this, corresponding comparable elements between two documents can now be compared for similarity. For example, all the "a's" from one document can be compared with those in the others. Figures 11 and 12 illustrate the idea of transcript mapping.

3. Statistical model for writer verification

All the documents are processed using CEDAR-FOX as described in the previous section and the handwriting features are compared. When more than one known document is available as is in this case, the handwriting features for each of the known documents is compared individually with the questioned document. The opinion of each of the individual comparisons is combined, as the final step.

The four major components of the verification model include: (i) Identifying discriminating elements; (ii) Mapping from feature to distance space by using similarity measure; (iii) Parametric modeling of the distance space distribution; and (iv) Computing a 9 point strength of evidence. Each of the four is briefly described below.

3.1 Features -Discriminating elements

Features for writer verification have been split into Macro (global) and Micro (Local) features (Lee, Cha & Srihari, 2002, p l55). Features are termed as discriminating elements (DE) (Huber & Headrick, 1999). Macro features are computed for the entire document whereas micro features are computed for characters, bi-grams (two characters) and words. The statistical model to be described can be used with any other set of features as well. Macro features (13 in number) are gray-scale based (entropy, threshold, number of black pixels), contour based (external and internal contours), slope-based (horizontal, positive, vertical and negative), stroke-width, slant, height and word-gap. These are real valued features. A set of 512 binary-valued micro-features corresponding to gradient (192 bits), structural (192 bits), and concavity (128 bits) which respectively capture the finest variations in the contour, intermediate stroke information, larger concavities and enclosed regions are used for characters. For bi-grams and words, 1024 such binary valued GSC features (Zhang, Srihari & Huang, 2004, p.45) are used. The number of local features is dependent on the recognition of characters in the document and on the amount of information in the document (full page/half page, ect).

3.2 Distance space distribution

Once a set of features described above are available for two documents, they can be compared.

The comparison results in mapping from feature space to distance space. The macro features are real valued and hence the mapping to distance space is just the absolute difference between the two feature values. The similarity between binary valued feature vectors for local features can be calculated using a number of different measures such as Hamming distance, Euclidean distance and etcetera. A detailed description of similarity measures for binary features is described in (Zhang & Srihari, 2003, p. 28). After much experimentation (Zhang & Srihari, 2003, p. 28), the correlation similarity was decided as the best similarity measure.

3.3 Parametric model

The distribution in distance space is modeled using known probability density functions (pdf). Assuming that similarity data can be acceptably represented by Gaussian or Gamma distributions, PDFs of distances conditioned upon the same-writer and different writer categories for a single feature x have the parametric forms conditioned on variables $p_s(x) \sim N(\mu_s,\sigma_s^2)$, $p_d(x) \sim N(\mu_d,\sigma_d^2)$ for the Gaussian case, and $p_s(x) \sim Gam(\alpha_s,\beta_s)$, $p_s(x) \sim Gam(\alpha_{d,} \beta_d)$ for the Gamma case. Estimating μ and σ from samples using the usual maximum likelihood estimation, the parameters of the gamma distribution are calculated as $a = \mu^2/\sigma^2$ and $b = \sigma^2/\mu$. Since the distribution is positive, it is intuitive to model them with the use Gamma distributions in general. But there are some exceptions. For macro features, we model both categories by Gamma distribution as $p_s(x) \sim Gam(\alpha_s,\beta_{s,})$ $p_s(x) \sim Gam(\alpha_d,\beta_d)$. For micro-features,while the "same-writer" category is modeled as $p_s(x) \sim Gam(\alpha_s,\beta_s)$ Gamma distribution, the "different-writer" is modeled as $p_d(x) \sim N(\mu_d, \sigma_d^2)$ Gaussian distribution. Once the distributions are modeled, the learning phase is complete. A new pair of unseen documents when compared results in N distance values, one for each feature compared. For example, if there were 10 common characters in the two documents, and zero bigrams and zero words common, the value of N would be 13 Macro + 10 Micro-features = 23. It's evident that the nature of the document affects the number of micro features but not the macro features. For one distance value x_i, i $\varepsilon=\{1...N\}$, the Likelihood Ratio (LR) is given as $LR(x_i) = p_s(x)/p_d(x)$. Considering the features as independent, we can have the LR for *N* of them as $\prod^N_i LR(x_i) = \prod^N_i p_s(x_i)/p_d(x_i)$. For ease with computer precision, the Log Likelihood Ratio (LLR) is used instead of LR and is given as $LLR= \Sigma^N_i \log p_s(x_i) - \log p_d(x_i)$.

3.4 Strength of Evidence

Once the LLR was computed as discussed above, the next step was to map it to a 9 point qualitative scale (Zhang, Srihari, 2003,p. 28). This scale corresponds to the strength of evidence that is associated with the LLR value. The 9 point scale is decided based on the following information: (i) LLR value; (ii) the amount of information compared in each of the two documents (line/half page/full page, etc.); (iii) the nature of content present in the document (same/ different content); (iv) features used for comparison and; (v) the error rate of the model used.

This scale follows the 9 point scale from the ASTM terminology [1-Identified as same, 2-Highly probable same, 3-Probably did, 4-Indications did, 5-No conclusion, 6-Indications did not, 7-Probably did not, 8-Highly Probable did not and 9-Identified as an elimination] (ASTM Standard E1658-04 "Standard Terminology for Expressing Conclusion of Forensic Document Examiners," ASTM International, West Conshohocken, PA). For the example discussed in this paper, the comparison of the known document A and questioned document gave a LLR value of 18.08 and opinion "Probably did". Comparison of the known document B and questioned document gave a LLR value of 19.32 and an opinion "Probably did". Finally, the comparison of the known document C and questioned document gave a LLR value of 39.33 and opinion "Identified as same". The combined opinion was found to be "Highly probable did". The opinion of the system: "Highly probable did" is indeed a correct conclusion to the case-work considered.

4. Discussion

CEDAR-FOX is an interactive software system to assist the document examiner in comparing handwriting samples. The system has capabilities for both handwriting verification and signature verification. This paper has presented a description of each of the different processing steps only for the writer verification task. A case example was used to demonstrate the steps.

The steps in preparing a handwriting sample for comparison are grey-scale to black-and-white thresholding, printed line removal and entering the ground-truth. The system computes features from two such processed specimens. Based on differences between the two feature sets, the system produces a score. The score, known as the log-likelihood ratio (LLR), is the natural logarithm of the ratio of the probability of being written by the same writer and the probability of being written by different writers. The necessary probabilities are determined by the system based on having been trained on previously observed samples. A positive score indicates that the system favors the same writer hypothesis and a negative score favors the different writer hypothesis. The score itself can be discretized by CEDAR-FOX into a nine-point scale analogous to the opinion expressed by the document examiner according an ASTM testing standard. In general, the system's verification accuracy when presented with at least half page of handwriting has been shown to be 97% approximately (Srihari, Huang, Srinivasan, 2007). When the amount of writing present is smaller, the accuracy is lower- reaching about 90% when only one line of writing is present.

5. References

5. References

Arivazhagan, M., Srinivasan, H., Srihari, S.N. (2007) A statistical approach to handwritten line segmentation. Proceedings, SPIE: Document Recognition and Retrieval IX, San Jose, California, pp. 6500T 1-11.

Balacu, M., Schumaker, L., Vuurpijl, L. (2003). Writer identification using edge-based directional features. International Conference on Document Analysis and Recognition (ICDAR).

Huber, R.A, Headrick, A.M. (1999). Handwriting Identification: Facts and Fundamentals. CRC press, Boca Raton, Florida.

Lee, S., Cha, S., Srihari, S.N. (2002). Combining macro and micro features for writer identification. SPIE: Document and Retrieval IX, San Jose, California, pp. 155-166.

Otsu, N. (March 1979). A threshold selection method from gray level histograms. IEEE Trans. Systems, Man and Cybernetics, 9, pp. 62-66.

Plamondon, R., Lorette, G. (1989). Automatic signature verification and writer identification – state of the art. Pattern Recognition, pp. 22, 107-131.

Srihari, S.N., Cha, H., Arora, H., Lee, S. (2002). Individuality of handwriting. J. Forensic Science, 47(4), pp. 856-872.

Srihar, S.N., Huang, C., Srinivason, H. (2008). On the Discriminability of Handwriting of Twins. J. Forensic Sciences, 53(2), pp. 430-446.

Srinivasan, H., Kabra, S., Srihari, S.N. (2007). On computing strength of evidence for writer verification. Proceedings, International Conference on Document analysis & recognition (ICDAR), Curitiba, Brazil, IEEE Computer Society Press, pp. 844-848.

Van Erp, M., Vuurpijl, L., Franke, K, Schumaker, L. (2003). The Wanda measurement tool for forensic document examination. Proceedings, 11th International Graphonomics Society Conference, pp. 282-285.

Zhang, B., Srihari, S.N., Huang, C. (2004). Word image retrieval using binary features. SPIE: E.H.B. Smith, J. Hu, and J. Allan, eds., Document Recognition and Retrieval XI 5296, pp. 45-53.

Zhang, B., Srihari, S.N. (January 2003). Binary vector dissimilarity measures for handwriting identification. Proceedings, SPIE: Document Recognition and Retrieval X, Santa Clara, California, pp. 28-38.

Endnotes

[1] In statistics and probability theory, Gaussian density functions appear as the density function of the Normal distribution which is used to model quantitative phenomena in natural sciences.

[2] A projection profile is a count of the number of foreground pixels across the width of the document - the count taken for every row (1 pixel wide) of the document.

[3] In mathematics, the convex hull of a set of points is the smallest convex set containing the points.

[4] An artificial neural network is a computational model loosely based on biological neural networks for learning and classification problems.

[5] A simple way to compare two values is to take their absolute difference. The result is always positive and it can be termed as a distance. Similarly one could take the Euclidean distance instead of the absolute difference. All these distance measures, map two inputs in feature space to one output in distance space.