J Forensic Sci, July 2002, Vol. 47, No. 4
Paper ID JFS2001227_474
Available online at: www.astm.org

Sargur N. Srihari,[1] *Ph.D.; Sung-Hyuk Cha,*[2] *Ph.D.; Hina Arora,*[3] *M.E.; and Sangjik Lee,*[4] *M.S.*

Individuality of Handwriting

ABSTRACT: Motivated by several rulings in United States courts concerning expert testimony in general, and handwriting testimony in particular, we undertook a study to objectively validate the hypothesis that handwriting is individual. Handwriting samples of 1500 individuals, representative of the U.S. population with respect to gender, age, ethnic groups, etc., were obtained. Analyzing differences in handwriting was done by using computer algorithms for extracting features from scanned images of handwriting. Attributes characteristic of the handwriting were obtained, e.g., line separation, slant, character shapes, etc. These attributes, which are a subset of attributes used by forensic document examiners (FDEs), were used to quantitatively establish individuality by using machine learning approaches. Using global attributes of handwriting and very few characters in the writing, the ability to determine the writer with a high degree of confidence was established. The work is a step towards providing scientific support for admitting handwriting evidence in court. The mathematical approach and the resulting software also have the promise of aiding the FDE.

KEYWORDS: forensic science, document analysis, feature extraction, handwriting identification, handwriting individuality

Introduction

Analysis of handwritten documents from the viewpoint of determining the writer has great bearing on the criminal justice system. Numerous cases over the years have dealt with evidence provided by handwritten documents such as wills and ransom notes. Handwriting has long been considered individual, as evidenced by the importance of signatures in documents. However, the individuality of writing in handwritten notes and documents has not been established with scientific rigor, and therefore its admissibility as forensic evidence can be questioned.

Writer individuality rests on the hypothesis that each individual has consistent handwriting that is distinct from the handwriting of another individual. However, this hypothesis has not been subjected to rigorous scrutiny with the accompanying experimentation, testing, and peer review. Our objective was to make a contribution towards this scientific validation.

The task involved setting up a methodology for validating the hypothesis that everyone writes differently. The study is built on recent advances in developing machine learning algorithms for recognizing handwriting from scanned paper documents. Software for recognizing handwritten documents has many applications, including sorting mail with handwritten addresses. Handwriting recognition focuses on interpreting the message conveyed, such as determining the town in a postal address, which is done by averaging out the variation in the handwriting of different individuals. On the other hand, the task of establishing individuality focuses on determining those very differences. However, both tasks involve processing images of handwriting and extracting features.

Legal Motivation

Our study was motivated by several rulings in United States courts that pertain to the presentation of scientific testimony in general and handwritten document examination testimony in particular. Six such rulings and their summaries are as follows:

- *Frye v. United States* (1), decided 1923: Expert opinion based on a scientific technique is inadmissible unless the technique is generally accepted as reliable in the relevant scientific community.
- *Daubert, et al. v. Merrell Dow Pharmaceuticals* (2), decided June 28, 1993: To admit expert opinion based on scientific technique in court, the technique needs to be established based on testing, peer review, error rates, and acceptability. *Daubert* is considered to be a landmark ruling in that it requires the judge to perform a gate-keeping function before scientific testimony is admitted.
- *U.S. v. Starzecpyzel* (3), decided April 3, 1995: (i) Forensic document examination expertise is outside the scope of *Daubert*, which established reliability standards for scientific expert testimony; (ii) forensic document examination testimony is admissible as nonscientific or skilled testimony; (iii) possible prejudice deriving from possible perception by jurors that forensic testimony met scientific standards of reliability did not require exclusion of testimony.
- *General Electric Co., et al. v. Joiner et al.* (4), decided December 15, 1997: Expert testimony that is both relevant and reliable must be admitted, and testimony that is irrelevant or unreliable must be excluded. Further, a weight-of-evidence methodology, where evidence other than expert testimony is admitted, is acceptable.
- *Kumho Tire Co., Ltd., et al. v. Carmichael et al.* (5), decided March 23, 1999: The reliability standard (does the application of the principle produce consistent results?) applies equally well to scientific, technical and other specialized knowledge.
- *United States v. Paul* (6), decided May 13, 1999: Handwriting analysis qualifies as expert testimony and is therefore admis-

[1] University distinguished professor, Department of Computer Science and Engineering and Director, Center of Excellence for Document Analysis and Recognition, University at Buffalo, State University of New York, Buffalo, NY 14228.

[2] Assistant professor, Pace University, Pleasantville, NY 10570.

[3] Research scientist, IBM, Endicott, NY.

[4] Doctoral candidate, Department of Computer Science and Engineering, University at Buffalo, State University of New York, Buffalo, NY 14228.

Received 6 July 2001; and in revised form 30 Oct. 2001, 10 Dec. 2001; accepted 19 Dec. 2001; published 14 June 2002.

> sible under the *Daubert* guidelines. It further states that if the witness qualifies as an expert on handwriting analysis, such testimony could assist the jury. Furthermore, the ability of the jury to perform the same visual comparisons as the expert "cuts against the danger of undue prejudice from the mystique attached to expert."

These high court rulings point to the need for a scientific study: (i) to validate the hypothesis that handwriting is individual, and (ii) to validate procedures used in establishing writer identity by experimentation and statistical analysis to establish error rates. Our study is an effort to establish the individuality of handwriting. The approach taken utilizes automated techniques derived from those used by experts.

Overview of Study

There are two variances of concern when comparing handwriting: within the handwriting of the same individual and between the handwritings of two individuals. These two variances are seen when several individuals are asked to write the same word many times (Fig. 1). Intuitively, the *within-writer* variance (the variation within a person's handwriting samples) is less than the *between-writer* variance (the variation between the handwriting samples of two different people). The goal of this study was to establish this intuitive observation in an objective manner.

The study consisted of three phases: data collection, feature extraction, and statistical analysis to establish the discriminative power of handwriting. In the data collection phase, representative samples of handwriting were collected. The feature extraction phase was to obtain handwriting attributes that would enable the writing style of one writer to be discriminated from the writing style of another writer. The validation phase was to associate a statistical confidence level with a measure of individuality.

The study pertains to natural handwriting and not to forgery or disguised handwriting. Examination of handwritten documents for forensic analysis is different from recognition of content, e.g., reading a postal address, or in attempting to assess personality (also known as graphology).

Handwriting Samples

Our objective was to obtain a set of handwriting samples that would capture variations in handwriting between and within writers. This meant that we would need handwriting samples from multiple writers, as well as multiple samples from each writer. The handwriting samples of the sample population should have the following properties (loosely based on *Ref* 7): (i) they are sufficient in number to exhibit normal writing habits and to portray the consistency with which particular habits are executed, and (ii) for comparison purposes, they should have similarity in texts, in writing circumstances, and in writing purposes.

Several factors may influence handwriting style, e.g., gender, age, ethnicity, handedness, the system of handwriting learned, subject matter (content), writing protocol (written from memory, dictated, or copied out), writing instrument (pen and paper), changes in the handwriting of an individual over time, etc. For instance, we decided that document content would be such that it would capture as many features as possible. Only some of these factors were considered in the experimental design. The other factors will have to be part of a different study. However, the same experimental methodology can be used to determine the influence factors not considered.

There were two design aspects to the collection of handwriting samples: content of the handwriting sample and determining the writer population.

Source Document

A source document in English, which was to be copied by each writer, was designed for the purpose of this study (Fig. 2*a*). It was concise (156 words) and complete in that it captured all characters (letters and numerals) and certain character combinations of interest. In the source document, each letter occurred in the beginning of a word in upper case and lower case and in upper case in the middle and end of a word (a total of 104 combinations). The number of occurrences in each position of interest in the source text is shown in Table 1. In addition, the source document also contained punctuation, all ten numerals, distinctive letter and numeral combinations (ff, tt, oo, 00), and a general document structure that allowed

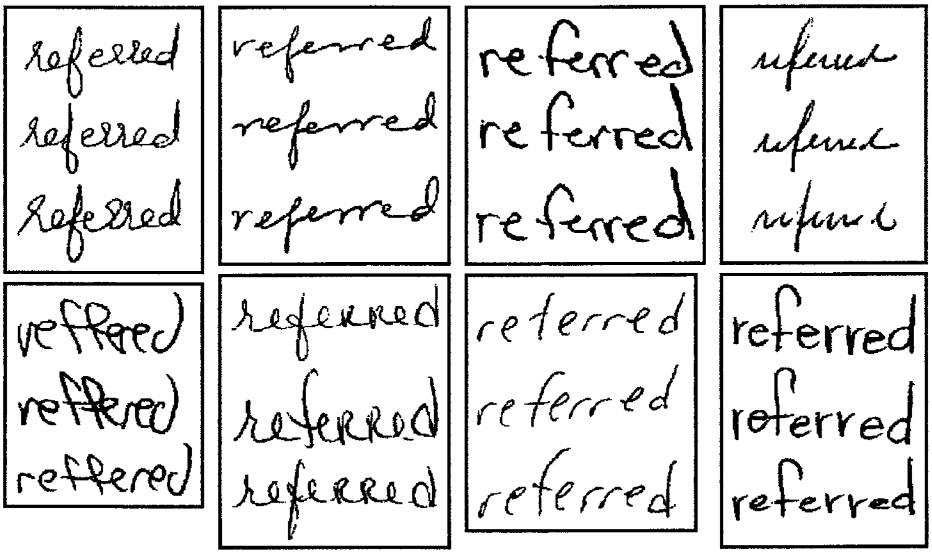
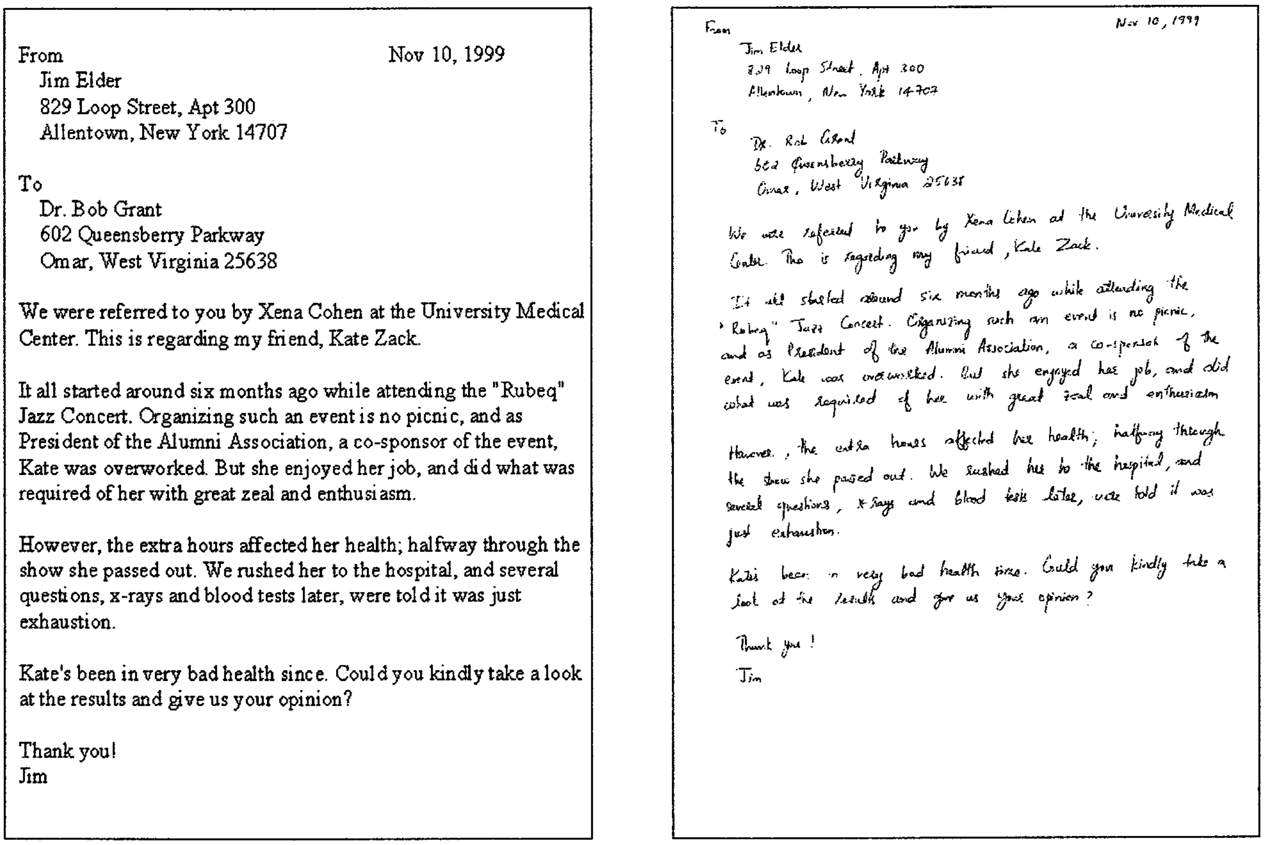


FIG. 1—*Variability in handwriting: samples provided by eight writers (boxed), each of whom wrote the same word three times.*



From Nov 10, 1999
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Bob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm.

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim

(a)

From Nov 10, 1999
Jim Elder
829 Loop Street, Apt 300
Allentown, New York 14707

To
Dr. Rob Grant
602 Queensberry Parkway
Omar, West Virginia 25638

We were referred to you by Xena Cohen at the University Medical Center. This is regarding my friend, Kate Zack.

It all started around six months ago while attending the "Rubeq" Jazz Concert. Organizing such an event is no picnic, and as President of the Alumni Association, a co-sponsor of the event, Kate was overworked. But she enjoyed her job, and did what was required of her with great zeal and enthusiasm

However, the extra hours affected her health; halfway through the show she passed out. We rushed her to the hospital, and several questions, x-rays and blood tests later, were told it was just exhaustion.

Kate's been in very bad health since. Could you kindly take a look at the results and give us your opinion?

Thank you!
Jim

(b)

FIG. 2—*Handwriting exemplar: a) source document to be copied by writers, and b) a digitally scanned handwritten sample provided by writer.*

TABLE 1—*Positional frequency of occurrence of letters in text.*

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
Init	4	2	4	1	1	1	1	1	1	2	3	1	1	1	2	2	1	1	1	2	1	1	3	1	1	1
	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o	p	q	r	s	t	u	v	w	x	y	z
Init	17	4	1	1	6	1	2	9	4	2	1	2	2	1	6	2	1	5	8	14	1	1	8	1	3	1
Mid	33	2	8	6	59	4	5	20	32	1	3	14	3	35	36	4	1	30	19	25	18	7	5	2	2	2
Term	5	2	1	21	20	3	3	5	1	0	3	5	2	7	5	1	1	12	15	17	2	1	2	1	8	1

extracting macro-document attributes such as word and line spacing, line skew, etc. Forensic literature refers to many such documents, including the *London Letter* and the *Dear Sam Letter* (8). We set out to capture each letter of the alphabet in upper and lower case in the initial, middle, and terminal positions of a word. This creates a total of 104 possibilities (cells) for each of the 26 letters in the alphabet. A measure of how "complete" the source text is given by the expression: (104–Number of empty cells)/104. While our source text scores 99% on this measure, the London Letter scores only 76%.

Each participant (*writer*) was required to copy the source document three times in his/her most natural handwriting, using plain, unlined sheets, and a medium black ballpoint pen, which we provided. The repetition was to determine, for each writer, the variation of handwriting from one occasion to the next.

Writer Population

We made the writer population as representative of the U.S. population as possible. Statistical issues in determining the writer population are: the number of samples needed to make statistically valid conclusions and the population distribution needed to make conclusions that apply to the U.S. population, which are issues in the design of experiments (9).

Randomness—If the samples are random, then every individual in the U.S. should have an equal chance of participating in the study. We attempted to make our sample population as random as possible. Sample handwriting was obtained by contacting participants in person, by mail, by advertising the study with the use of flyers and internet newsgroups, and by manning a university booth. For geographic diversity, we obtained samples by contacting schools in three states (Alaska, Arizona, and New York) and communities in three states (Florida, New York, and Texas) through churches and other organizations.

Sample Size—The sample population should be large enough to enable drawing inferences about the entire population through the observed sample population. The issue of large enough is related to sampling error, the error that results from taking one sample instead of examining the whole population, i.e., how close is an estimate of a quantity based on the sample population to the true value for the entire population?

Public opinion polls that use simple random sampling specify using a sample size of about 1100, which allows for a 95% confidence interval, with a margin of error of 0.03 (10). Higher precision levels would entail a larger number of samples. Our database has a sample size of about 1500, and our results are therefore subject to such a margin of error.

Representativeness—The sample population should be representative of the U.S. population. For instance, since the U.S. population consists of an (approximately) equal number of males and females, it would be unwise to perform the study on a sample population consisting of only males and expect the conclusions of the study to apply to the entire U.S. population (especially in the absence of any scientific evidence that proves or disproves the association between handwriting and gender). The sample was made representative by means of a stratified sample with proportional allocation (9).

We divided the population into a predetermined number of subpopulations, or *strata*. The strata do not overlap, and they constitute the whole population so that each sampling unit belongs to exactly one stratum. We drew independent probability samples from each stratum, and we then pooled the information to obtain overall population estimates. The stratification was based on U.S. census information (1996 projections).

Proportional allocation was used when taking a stratified sample to ensure that the sample reflects the population with respect to the stratification variable, and the sample is a miniature version of the population. In proportional allocation, so called because the number of sampled units in each stratum is proportional to the size of the stratum, the probability of selection is the same for all strata. Thus, the probability that an individual will be selected to be in the sample is the same as in a simple random sample without stratification, but many of the bad samples that could occur otherwise cannot be selected in a stratified sample with proportional allocation. The sample size again turns out to be about 1000 for a 95% confidence interval, with a margin of error of 0.03.

A survey described above would allow drawing conclusions only about the general U.S. population and not any subgroup in particular. In order to draw any conclusions about the subgroups, we would need to use allocation for specified precision within data. This would entail having 1000 in each cell of the cross classification.

From the census data, we obtained population distributions pertaining to gender, age, ethnicity, level of education, and country of origin; we also obtained a distribution for handedness from (11). Based on this information, a proportional allocation was performed for a sample population of 1000 across these strata. Among these variables, only gender, age, and ethnicity can be considered as strata (by definition). Due to the limited amount of census data on other combinations, we were unable to stratify across handedness and level of education.

Each writer was asked to provide the following writer data, enabling us to study the various relationships: gender (male, female), age (under 15 years, 15–24 years, 25–44 years, 45–64 years, 65–84 years, 85 years and older), handedness (left, right), highest level of education (high school graduate, bachelors degree and higher), country of primary education (if U.S., which state), ethnicity (Hispanic, white, black, Asian/Pacific Islander, American Indian/Eskimo/Aleut), and country of birth (U.S., foreign).

The details (actual/target) of the distribution for a sample size of 1568 writers are given in Table 2. The strata are sometimes under-represented (actual < target) or over-represented (actual > target). Parameters considered in addition to strata shown in Table 2 are handedness and country of origin—Male: handedness (right, left): 382/429, 61/61, and country of origin (U.S., foreign): 373/451, 71/39; Female: handedness (right, left): 1028/461, 95/49, and country of origin (U.S., foreign): 1026/469, 98/41.

There may be other relevant strata that could have been considered, such as the system of writing learned (e.g., the Palmer method), country in which writing was learned, etc. We were constrained by the limited information we have on these distributions. Moreover, a perfect sample (a scaled-down version of the population that mirrors every characteristic of the whole population) cannot exist for complicated populations. Even if it did exist, we would not know it was a perfect sample without measuring the whole population.

Handwriting Attributes (Features)

Our approach to studying the handwriting of different individuals was to scan the samples into a computer and then automatically obtain handwriting attributes for further study.

Scanning and Image Segmentation

Each handwritten document was scanned and converted into a digitized image using a desktop black and white scanner. The resolution of scanning was 300 dpi, and the resulting images were stored as gray-scale images of discrete pixels (each pixel value can vary from 0 to 255, where 0 is pure black, and 255 is pure white). After all handwritten documents were digitally scanned, the gray-scale image was converted to a pure black and white (or binary) image by using a binarization algorithm. The method of binarization

TABLE 2—*Writer population distribution in handwriting database (actual and target): male population size: 444/490, female population size: 1124/510. The population was stratified over gender, age, ethnicity, education, and handedness.*

Ethnicity/ Gender	White Female	White Male	Black Female	Black Male	API Female	API Male	AIEA Female	AIEA Male	Hispan Female	Hispan Male
Age/Total	872/371	333/359	103/64	36/56	38/16	31/14	19/5	4/5	91/54	40/56
12–14	49/17	25/16	2/4	2/4	1/1	2/1	0/0	0/0	22/4	16/4
15–24	158/66	111/64	25/15	13/13	16/4	18/2	4/1	1/2	22/13	10/14
25–44	252/140	76/136	31/25	8/22	12/6	7/6	11/3	2/1	34/24	11/24
45–64	267/87	69/85	24/13	10/11	6/4	2/3	3/1	1/1	7/10	1/10
65–84	139/56	50/55	20/6	3/5	3/1	2/1	1/0	0/0	6/3	2/4
85 ~	7/5	2/5	1/1	0/1	0/0	0/1	0/0	0/1	0/0	0/0

NOTE: The numbers may not add to 1568 because a few subjects did not provide the relevant information.

determines a threshold gray-scale value such that any value higher than the threshold is deemed to be white and any value lower is deemed to be black.

Paragraph and line images were acquired from each document image by segmentation. Word images were segmented from the line image, and each character image was segmented from the word image. We used a commercial image-manipulating tool (Adobe® Photoshop®) to manually extract line, word, and character images. Examples of extracted paragraph, line, word, and character images are shown in Fig. 3.

Segmentation of the eight characters of the word "referred" is illustrated in Fig. 4. These eight characters were used as sample allographs in some of the tests conducted for individuality.

Types of Features

Features are quantitative measurements that can be obtained from a handwriting sample in order to obtain a meaningful characterization of the writing style.

These measurements can be obtained from the entire document or from each paragraph, word, or even a single character. In pattern classification terminology, measurements, or attributes, are called "features." In order to quantify the process of matching documents, each sample is mapped onto a set of features that correspond to it, called a "feature vector." For example, if measurements, f_1, f_2,..., f_d, are obtained from a sample, then these measurements form a column vector $[f_1, f_2, \ldots f_d]^t$, which is a data point in d-dimensional space (12); note that superscript t indicates vector transposition.

We distinguish between two types of features: conventional features and computational features. Conventional features are the handwriting attributes that are commonly used by the forensic document examination community. These features are obtained from the handwriting by visual and microscopic examination. Software tools such as FISH (Forensic Information System for Handwriting), developed in Germany, are used to narrow down the search. Computational features are features that have known software/hardware techniques for their extraction. The two types of features have some correspondence.

Conventional Features—Forensic document examiners use a host of qualitative and quantitative features in examining questioned documents. These features have been compiled into twenty-one discriminating elements of handwriting (7). A discriminating element is defined as "a relatively discrete element

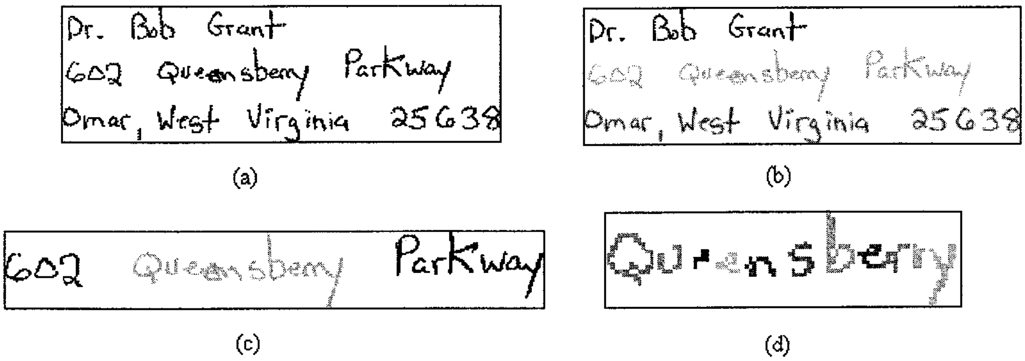


FIG. 3—*Examples of three levels of segmentation: a) paragraph (address block), b) line level, c) word, and d) character. Each distinct line, word, or character is assigned a distinct shade/color.*

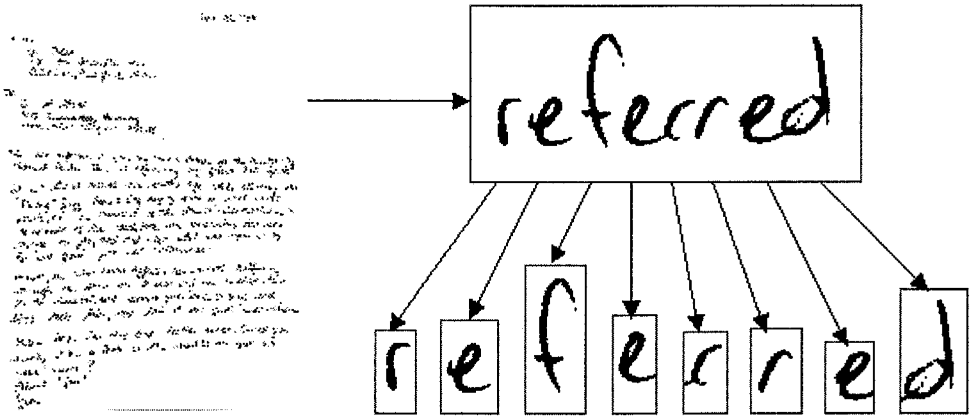


FIG. 4—*Segmented word and character images: snippets of words and characters extracted from the handwritten word "referred." The shapes of these eight characters were used to determine the writer.*

of writing or lettering that varies observably or measurably with its author and may, thereby, contribute reliably to distinguishing between the inscriptions of different persons, or to evidencing the sameness in those of common authors." The 21 features are: arrangement; class of allograph; connections; design of allographs (alphabets) and their construction; dimensions (vertical and horizontal); slant or slope; spacings, intraword and interword; abbreviations; baseline alignment; initial and terminal strokes; punctuation (presence, style, and location); embellishments; legibility or writing quality; line continuity; line quality; pen control; writing movement (arched, angular, interminable); natural variations or consistency; persistency; lateral expansion; and word proportions.

Computational Features—Computational features are those that can be determined algorithmically, e.g., by software operating on a scanned image of the handwriting. Computational features remove subjectivity from the process of feature extraction. While it could be argued that all conventional features could eventually be computational features—when the correct algorithms have been defined—the fact remains that most of the conventional features are not yet computable.

While some conventional features, like embellishments and line quality, are difficult to implement algorithmically, several of the other features are computable based on existing techniques for handwriting recognition (13,14). Handwriting recognition differs from handwriting identification in that they are two opposite processes. The objective of handwriting recognition is to filter out individual variability from handwriting and recognize the message. The objective of handwriting identification is to capture the essence of the individuality, while essentially ignoring the content of the message. The two share many aspects of automated processing, such as determining lines, strokes, etc. For instance, handwriting recognition procedures routinely compute baseline angle and slant so that a correction can be applied prior to recognition (15).

Computational features can be divided into macro- and micro-features, depending on whether they pertain globally to the entire handwritten sample, e.g., darkness, or are extracted locally, e.g., contour variations. Macro-features can be extracted at the document level (entire handwritten manuscript) or at the paragraph, line, word, and character levels. We used a set of eleven macro-features that are loosely related to the document examiner discriminating elements (Fig. 5).

Measures of Pen Pressure	1. Entropy of grey values 2. Grey-level threshold 3. Number of black pixels
Measures of Writing Movement	4. Number of interior contours 5. Number of exterior curves
Measures of Stroke Formation	6. Number of vertical slope components 7. Number of horizontal slope components 8. Number of negative slope components 9. Number of positive slope components
Slant	10. Slant
Word Proportion	11. Height

FIG. 5—*Eleven computational (macro-features) and their relationship to five conventional features.*

Micro-features are computed at the allograph, or character shape, level. They are analogous to the allograph-discriminating elements among document examiner features. The features that we used are those used in recognizing handwriting scanned from paper documents (called off-line recognition), which differ from those used in devices such as hand-held PDAs (called on-line recognition). Features corresponding to gradient, structural, and concavity (GSC) attributes, which are used in automatic character recognition for interpreting handwritten postal addresses (16,17), were used as micro-features.

Feature Extraction

Macro-Features—The macro-features can also be grouped into three broad categories: darkness features, contour features (connectivity and slope features), and averaged line-level features. Darkness features, such as entropy of gray-level values, gray-level threshold, and number of black pixels, are indicative of the pen pressure. The number of interior and exterior contours are indicative of writing movement. The number of horizontal, vertical, negative, and positive slope components are indicative of stroke formation. Brief descriptions of algorithms for computing the eleven macro-features follows (see *Ref* 10 for greater detail).

Measures of Pen Pressure

1. *Gray-level distribution (measured by its entropy)*: Entropy is an information-theoretic measure of disorder. The gray-scale histogram (frequency plot of the gray-values) of the scanned image is normalized and regarded as a probability distribution. The entropy of the probability distribution is calculated as $-\sum_i p_i \log p_i$, where p_i is the probability of the i^{th} gray value in the image. This gives an indication of the variation of gray-levels in the image. For example, an image where each gray-level is equally likely will have a very high entropy.
2. *Gray-level threshold value*: The scanned gray-scale image is converted into a pure black-and-white, or binary, image by using a thresholding algorithm. It maps the gray-level pixel values in the image that are below a particular threshold to pure black (foreground) and those above the threshold to pure white (background). The threshold value (the gray-scale value that partitions the foreground and background of the gray-level image) is determined using a gray-level histogram (18). The value of the threshold is indicative of the pen-pressure, with higher values indicating lighter pressure.
3. *Number of black pixels*: This is a count of the number of foreground pixels in the thresholded image. The number of black pixels is indicative of the pen pressure, thickness of strokes, and size of writing.

Measures of Writing Movement

The thresholded black-and-white images are processed to determine the connected components in the image—each connected component can be thought of as a "blob." The outlines of the blobs, or contours, are stored and manipulated. A binary image of a line of text from the handwritten source document and the corresponding contour image are shown in Fig. 6. The outlines, or contours, are stored as chaincodes (19,20). A chaincode is a series of integers in the range 0–7, each of which represents a direction of slope of the contour, e.g., 0 represents east, 1 represents north-east, 2 represents north, 3 represents north-west, etc. The chaincodes of the numeral 6 are in Fig. 7.

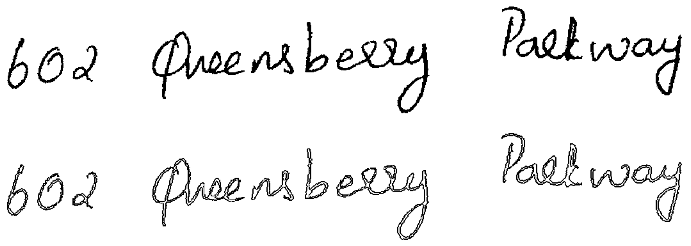
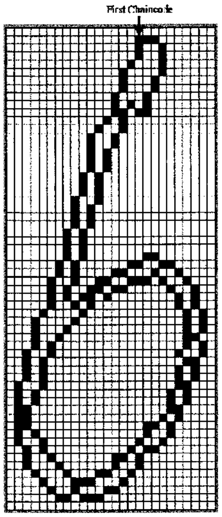


FIG. 6—*Extraction of contours of handwriting: a) thresholded image of a line of handwritten text, and b) corresponding contour image.*



Exterior contour of 6:
0667766766777666676676666766676766677
6667666766666665666555543544343332343
3232322322321322122111010707070777712
32232222322322223223232223542232322210

Interior contour of 6:
555656666757667667706777067070011112
11223232222323333323534344

FIG. 7—*Chaincode and feature representation: digitized numeral 6, and the chaincode.*

Two sets of features are extracted from the contour image as follows:

4–5. *Contour connectivity features*: The number of interior and exterior contours is extracted from the chaincode representation of the image. The average number of interior and exterior contours can be used as a measure of writing movement: cursive handwriting, for example, would have a greater number of interior contours and fewer exterior contours, while disconnected hand-printing would have a very large number of exterior contours. Examples of contour connectivity features for two samples from the database are shown in Fig. 8. Note that while the figure shows the connectivity features extracted for a line, these features can be calculated for the entire document, paragraph, line, word, or character.

Measures of Stroke Formation

6–9. *Contour slope features*: Vertical, negative, positive, horizontal slope components are indicative of the nature of stroke formation. Flattish writing would have a large number of horizontal slope components, while handwriting with a distinctive negative slope would have a large number of negative slope components. Contour slope features for two samples from the database are shown in Fig. 9, which shows the connectivity features extracted for the block of text.

Slant and Proportion

The last two macro-features, slant and height, are extracted at the line level (and averaged over the entire document, if necessary):

10. *Slant*: Vertical and near-vertical lines are extracted from the chaincode. Global slant angle is the average of all the angles of

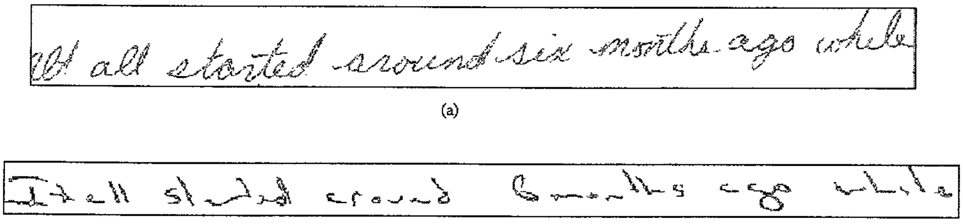

It all started around six months ago while
(a)
It all started around 6 months ago while
(b)

FIG. 8—*Macro-feature–connectivity: a) number of exterior contours = 17, number of interior contours = 49, and b) number of exterior contours = 34, number of interior contours = 7.*

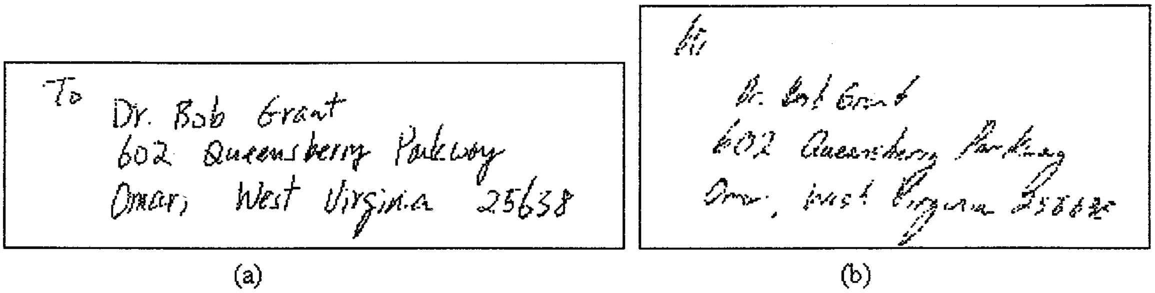


FIG. 9—*Macro-feature–contour slope: normalized number of horizontal (n_h), positive (n_p), vertical (n_v), and negative (n_n) slope components features. Illustration with two samples: a) $n_h = 0.06$, $n_v = 0.15$, $n_n = 0.68$, $n_p = 0.11$; b) $n_h = 0.04$, $n_v = 0.14$, $n_n = 0.72$, $n_p = 0.10$.*

TABLE 3—*Sample macro-features extracted from samples of three writers.*

Writer	Sample	F 1	F 2	F 3	F 4	F 5	F 6	F 7	F 8	F 9	F 10	F 11
W_1	$W_{1,1}$	0.50	188	184 K	15	14	0.31	0.13	0.28	0.28	8.8	25
	$W_{1,2}$	0.47	187	182 K	15	16	0.32	0.13	0.27	0.28	8.3	25
	$W_{1,3}$	0.52	186	181 K	16	15	0.33	0.12	0.26	0.20	10.4	23
W_2	$W_{2,1}$	0.54	198	205 K	21	23	0.20	0.12	0.43	0.25	6.5	30
	$W_{2,2}$	0.53	197	201 K	21	25	0.20	0.12	0.43	0.25	6.1	30
	$W_{2,3}$	0.57	197	200 K	21	22	0.20	0.12	0.42	0.26	7.7	30
W_3	$W_{3,1}$	0.82	191	373 K	7	20	0.29	0.10	0.29	0.32	17.2	27
	$W_{3,2}$	0.80	189	368 K	10	26	0.30	0.09	0.28	0.33	18.1	25
	$W_{3,3}$	0.85	191	390 K	10	26	0.31	0.10	0.29	0.30	14.0	29

these lines, weighted by their length in the vertical direction since the longer lines give more accurate angle values than the shorter ones.

11. *Height*: The height is calculated (for each line in the document) by considering the distance between contiguous maxima and minima in the upper contour of the chaincode. It is then averaged over the entire document.

Feature vectors composed of the eleven macro-features for three writers W_1, W_2, and W_3 with corresponding samples W_{11}, W_{12}, W_{13}, W_{21}, W_{22}, W_{23}, and W_{31}, W_{32}, W_{33} are shown in Table 3. W_1 is male, 65–84, right-handed, college educated, white, U.S. educated; writer W_2 (Sample 1 is shown in Fig. 2*b*) is female, 25–44, right-handed, college educated, API, foreign educated; and writer W_3 is female, 45–64, left-handed, college educated, white, U.S. educated. For instance, sample W_{11} had raw values as follows: entropy = 0.5, threshold = 195, # of black pixels = 184,000, # of exterior contours = 15, # of interior contours = 14, # of horizontal slope components = 0.31, # of negative slope components = 0.13, # of vertical slope components = 0.28, # of positive slope components = 0.28, slant = 8.8, and height = 25.

The variation of features (stratified across gender, age, and ethnicity) for approximately 300 writers (three samples each) is shown in Fig. 10 by mapping the normalized feature values to a color scale of eleven values. The white population has greater representation (two columns) than other ethnic groups (one column each) as an indication of a greater percentage of white people in the database (since it was based on proportional allocation). As indicated by the color map, there is consistency within different samples of a writer and considerable variation between samples of different writers.

Paragraph- and Word-Level Features

Sub-images corresponding to paragraphs, words, and characters were extracted semi-automatically, and then the features were computed from the sub-images automatically. The feature sets were slightly modified for paragraphs and words as follows:

Macro-Features: Paragraph-Level—Paragraph-level features were extracted from the destination address block that appears in the source text. Macro features 3–11 were extracted at the paragraph level. Two new features were extracted: height to width ratio (aspect ratio) and indentation (margin width).

Macro-Features: Word-Level—Macro features 3–11 were extracted at the word level if the content of the words being compared is the same. Three new features are extracted: upper zone ratio, lower zone ratio, and length. The word-level features were extracted for the word "referred" in the source text.

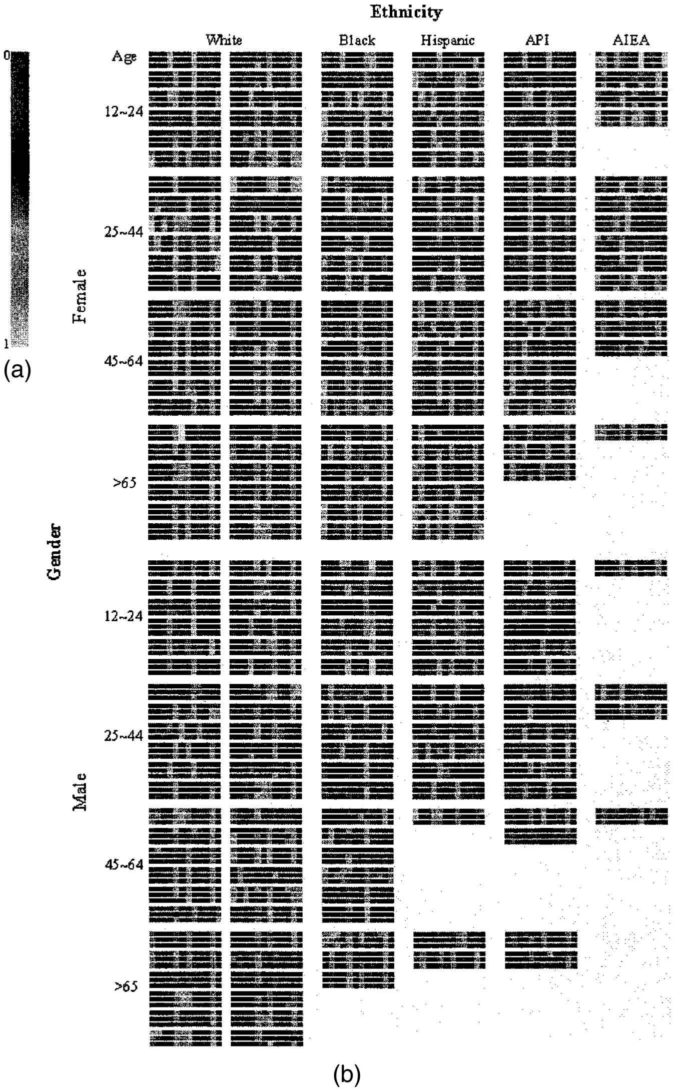


FIG. 10—*Handwriting feature map for 300 writers, each writer having three samples: a) color scale for representing normalized feature values; 0 is on top, and 1 is at the bottom of the scale, and b) feature map, where each horizontal bar represents eleven macro-features extracted from a single sample. There are three bars per writer corresponding to three exemplars. The color image can be seen at http://www.cedar.buffalo.edu/NIJ/colormap1.gif.*

The relationship between the feature sets at the word, paragraph, and document levels is shown in Table 4.

Micro-Features—The micro-features consist of 512 binary (0 or 1 value) features corresponding to gradient (192 bits), structural (192 bits), and concavity (128 bits) features. Examples of micro-features of characters are shown in Fig. 11. The first gradient feature generator computes the gradient of the image by convolving it with a 3 × 3 Sobel operator (21,22). The direction of the gradient at every edge is quantized to 12 directions. The structural feature generator takes the gradient map and looks in a neighborhood for certain combinations of gradient values. These combinations are used to compute eight distinct features that represent lines (strokes) and corners in the image. The concavity feature generator uses an eight point star operator to find coarse concavities in four directions, holes, and large-scale strokes. The image feature maps are normalized with a 4 × 4 grid, and a feature vector is generated. These features were used at the character level in our study.

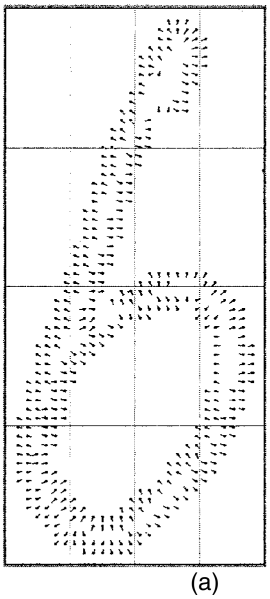
TABLE 4—*Features extracted from handwritten document at three levels of coarseness: word, paragraph, and document.*

Features	Document	Paragraph	Word
Gray-level entropy (F1)	Y		
Gray-level threshold (F2)	Y		
No. of black pixels (F3)	Y	Y	Y
No. of interior & exterior contours (F4,F5)	Y	Y	Y
No. of 4-directional slope components (F6–F9)	Y	Y	Y
Average height (F11)	Y	Y	Y
Average slant (F10)	Y	Y	Y
Aspect ratio		Y	
Margin Width		Y	
Length			Y
Upper & lower zone ratio			Y

Statistical Analysis

The analysis task is to use the handwriting samples that were collected and the features extracted from the samples to establish the discriminative power of handwriting. One possible approach to doing this is as follows. Each handwriting sample may be viewed as a point in a multi-dimensional feature space. If, in feature space, all the samples of writer W_1 are close together and all the samples of writer W_2 are close together but samples of W_1 and W_2 are far apart, then we can say that W_1 and W_2 write differently and that samples of W_1 and W_2 belong to two different classes or clusters (12,23). This is illustrated for the three-writer data in Table 3 using the two-dimensional feature space, consisting of features F1 and F2 in Fig. 12.

In order to validate individuality among n writers, we would have to determine whether the samples form n distinct clusters, where samples of the same writer belong to the same cluster and



(a)

```
Gradient    :  0000000000001100000000000110000110000000000010000011000000000001110
(192bits)      0111000001100001100000111100001011000011000000000001100000000000
               0000100000110000100011000011001110001110010000111000010000110000
Structural  :  0000000000000000010000000000011000000010001010000000000000000000
(192bits)      0011110000001001010000001000000000000011100000000001000000000100
               0000000001101100110011011110000011000000000000110000000000110000
Concavity   :  0110010011111111000000001111111001100100110110010000000001111100
(128bits)      0011011100000000000000000000000000000000000000000000000000000000
```

(b)

FIG. 11—*Micro-features of the numeral 6: a) gradient map, showing the directions of the image gradient at each pixel, and b) gradient, structural, and concavity features (512 bits).*

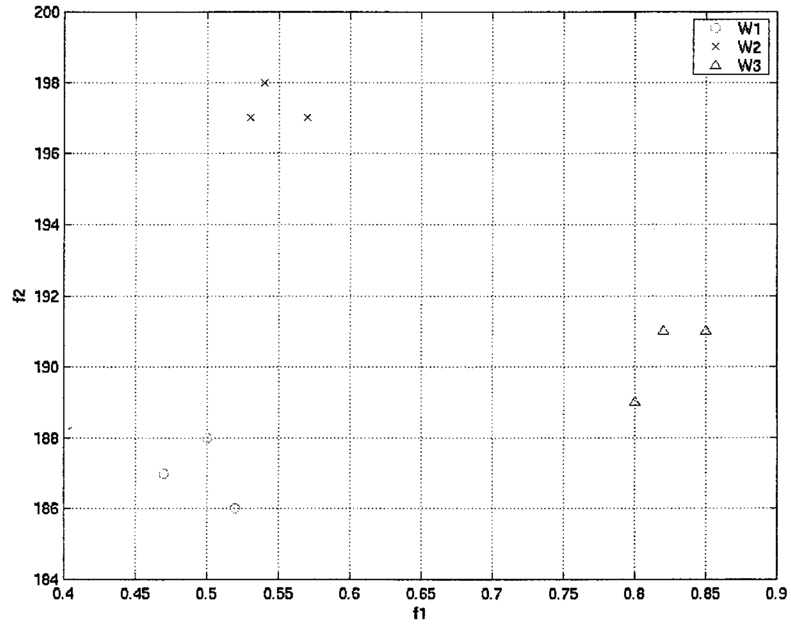


FIG. 12—*Handwriting samples of three writers in two-dimensional feature space.*

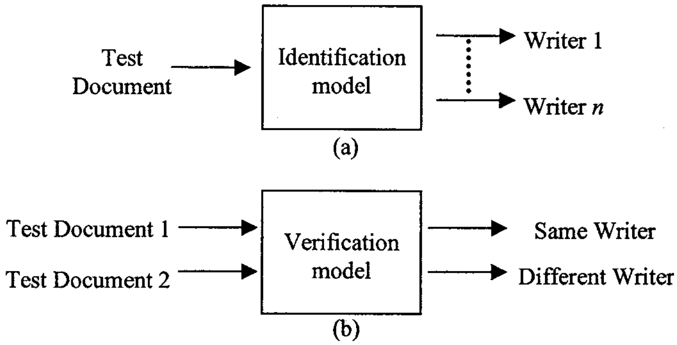


FIG. 13—*Two models for establishing the individuality of handwriting: a) the identification model, and b) the verification model.*

samples of different writers belong to different clusters. A measure of distinctness of the clusters would be a measure of confidence of individual discriminability.

The task of determining the presence of distinct clusters can be approached by using the probability of correctly classifying samples of unknown writership as the criterion of clustering. In the identification model, given a handwriting sample x whose writer is unknown and samples of handwriting of n known writers, we would like to identify the writer of x among the n writers.

In the verification model, given two handwriting samples x_1 and x_2 and samples of handwriting of n writers, we would like to determine whether x_1 and x_2 were written by the same person or by two different people among the n writers. Both models involve classification, with the identification model leading to an n-class problem (or a polychotomy of the feature space) and the verification model leading to a 2-class problem (or a dichotomy of the feature space) (Fig. 13).

Each of these models can be regarded as tasks in machine learning (24). Handwriting samples are used to learn the discrimination task. Once the task is learned, a set of samples is used to test the model for its accuracy. Both models will provide a probability of correct classification that we can use as a measure of confidence of the individuality hypothesis.

The question arises as to which model is better. The identification model has the advantage of being able to identify the writer directly. However, it is dependent on knowing all the writers in advance. The result with n writers does not generalize with $n + 1$ writers. On the other hand, the verification model provides results that have statistical inferability. The two different classification approaches would provide a measure of cross checking our results.

Both models involve a method of measuring similarity, or nearness, or distance, between two samples. For macro-features, the distance between a pair of documents with feature vectors $A = [a_1, a_2, \ldots, a_d]^t$ and $B = [b_1, b_2, \ldots, b_d]^t$ is defined by the Euclidean dis tance $\sqrt{\Sigma_{i=1}^{d} (a_i - b_i)^2}$, where d is the number of attributes. For micro-features, the distance between two characters represented by binary feature vectors A and B is calculated as:

$$d(A, B) = A^tB + \frac{\overline{A}^t\overline{B}}{2}.$$

Identification Model

Writer identification is the task of determining the writer when there are n candidates. This classification task has to be learned from a set of handwriting samples provided by each of the candidates. Given a test sample of an unknown writer, the task is to determine whether it was written by any of the n writers and, if so, to identify the writer. The writer identification procedure uses the features extracted from the test image and from the labeled prototype images to determine the writer of the test document.

Learning Algorithm—The identification model can be regarded as an n-class classification problem where writership of the samples is established based on their proximity to one another. We used the simplest learning algorithm based on storing all the samples. Classification is achieved by finding the closest match. This is known as the nearest neighbor rule (12), where the unknown input vector is classified by finding the most similar template in the prototype, or learning, set. The prototype set consisted of all the documents written by each of n writers, except for a test document that is left out from the set. So the reference set has $(3 \times n) - 1$ documents in it. The test document is assigned the class of the document nearest to it among the prototypes.

To evaluate identification accuracy, the following experiments were set up. A number of n writers was randomly selected from 1000 writers; then one document written by one of n writers was selected as a query document, and the rest of $(3 \times n) - 1$ documents was used as a reference set. This leave-one-out method was performed 1000 times for each n, and the accuracy is the number of correctly classified queries divided by 1000.

This procedure was applied with macro-features shown in Table 3 converted into normalized form obtained from the raw data by scaling the minimum and maximum values of each feature to 0 and 1, which are shown in Table 5.

Identification Accuracy—Identification accuracy was measured against the number of writers considered in three separate sets of experiments using macro-features, micro-features, and their combinations.

- *Macro-features*: Parameterizing against document, paragraph, and word levels (Fig. 14), we observed that: (i) the larger the portion of the document image we consider, the higher the accuracy, and (ii) performance decreases as the number of writers increase.
- *Micro-features*: Accuracy also improves with the number of characters considered, as shown in Fig. 15. Using character-level features of all ten characters of the word "referred" as well as "b" and "h" (Fig. 4), the correct writer was identified in 99% of the cases when all possible pairs of writers were considered. When there are five possible writers, the writer of the test document is correctly assigned with a 98% probability. We can expect the accuracy to improve when we

TABLE 5—*Normalized macro-feature data. Values are normalized to lie in (0,1) interval.*

Writer	Sample	F 1	F 2	F 3	F 4	F 5	F 6	F 7	F 8	F 9	F 10	F 11
W_1	$W_{1,1}$	0.20	0.45	0.28	0.30	0.30	0.45	0.42	0.45	0.25	0.52	0.23
	$W_{1,2}$	0.19	0.46	0.28	0.37	0.37	0.47	0.41	0.43	0.25	0.53	0.23
	$W_{1,3}$	0.19	0.46	0.30	0.33	0.33	0.50	0.43	0.40	0.23	0.56	0.19
W_2	$W_{2,1}$	0.23	0.50	0.43	0.60	0.60	0.22	0.34	0.76	0.24	0.49	0.32
	$W_{2,2}$	0.22	0.48	0.43	0.67	0.67	0.23	0.33	0.75	0.24	0.49	0.32
	$W_{2,3}$	0.22	0.48	0.43	0.57	0.57	0.23	0.36	0.74	0.22	0.51	0.32
W_3	$W_{3,1}$	0.47	0.38	0.08	0.50	0.50	0.41	0.50	0.46	0.17	0.67	0.26
	$W_{3,2}$	0.46	0.34	0.15	0.70	0.70	0.43	0.52	0.44	0.14	0.69	0.23
	$W_{3,3}$	0.50	0.38	0.15	0.70	0.70	0.46	0.45	0.46	0.17	0.62	0.30

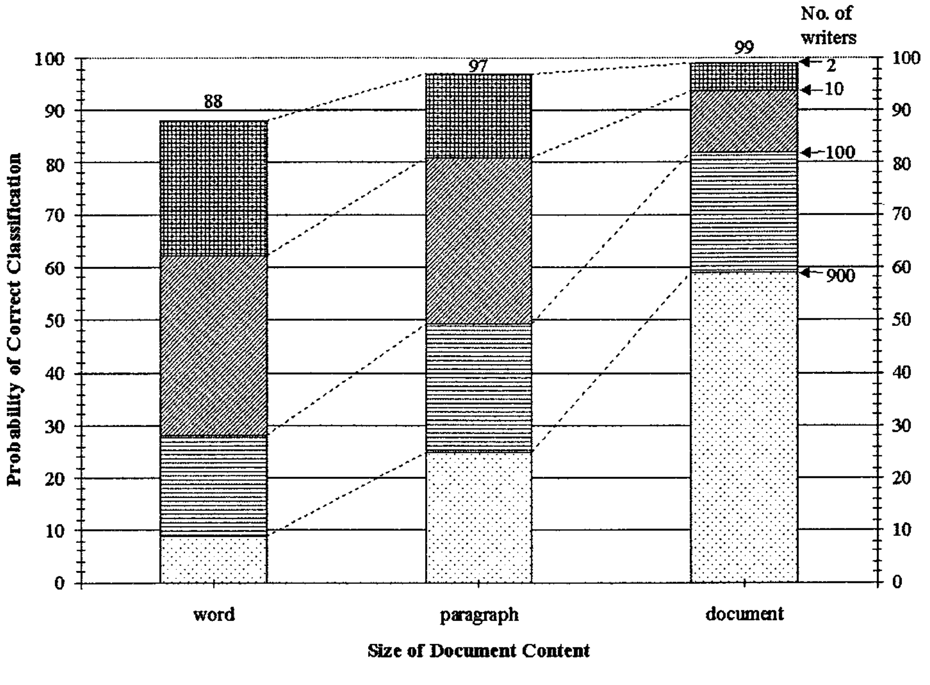


FIG. 14—*Writer identification accuracy using macro-features: shown as a function of the size of document content (document, paragraph, and word). The word level corresponds to two words ("Cohen" and "referred"); the paragraph level corresponds to the address block (Fig. 3a), which consists of 11 words; the document level corresponds to the entire document image (Fig. 2b), which consists of 156 words.*

consider: (i) more words in the document, and (ii) more discriminatory features.

- *Combination*: The micro-features are better than document-level features in that higher accuracy was obtained when more writers are considered. Combining the two sets of features yields a higher accuracy than either set alone. We combined them as follows. The macro-features were used as a filter that reduces the number of writers from 1000 to 100. Micro-features were then used to identify the writer among the 100 choices. The results of this process are displayed in the right column in Fig.15.

Verification Model

Writer verification is the task of determining whether two samples, *X* and *Y*, were written by the same writer or by two different writers. This is a 2-class categorization problem that requires a dichotomy of the feature space (Fig. 16).

We use the fact that the within-writer distance (the distance between two samples written by the same writer) will be less than the between-writer distance (the distance between two samples written by two different writers). Hence, instead of considering features, we consider distances, thereby transforming the *n*-class problem in *d*-dimensional feature space to a 2-class problem of same or different writers in multi-dimensional distance space.

When there are *n* writers contributing three documents each, the number of within-class distances is $n\cdot\binom{3}{2}$, and the number of between-class distances is $\binom{n}{2}\cdot 3\cdot 3$. Assume three writers, $\{W_1, W_2, W_3\}$ and that each writer provides three samples. If we extract two features from each sample, then each sample is a point in two-dimensional feature space (Fig. 17*a*). We then find the distance between each pair of samples, thereby transforming the $3 \times 3 = 9$ points in feature space to $3\cdot\binom{3}{2} = 9$ within-writer distances $\binom{3}{2}\cdot 3\cdot 3 = 27$ between-writer distances in feature distance space (Fig. 17*b*). The number of between-writer distances increases com-

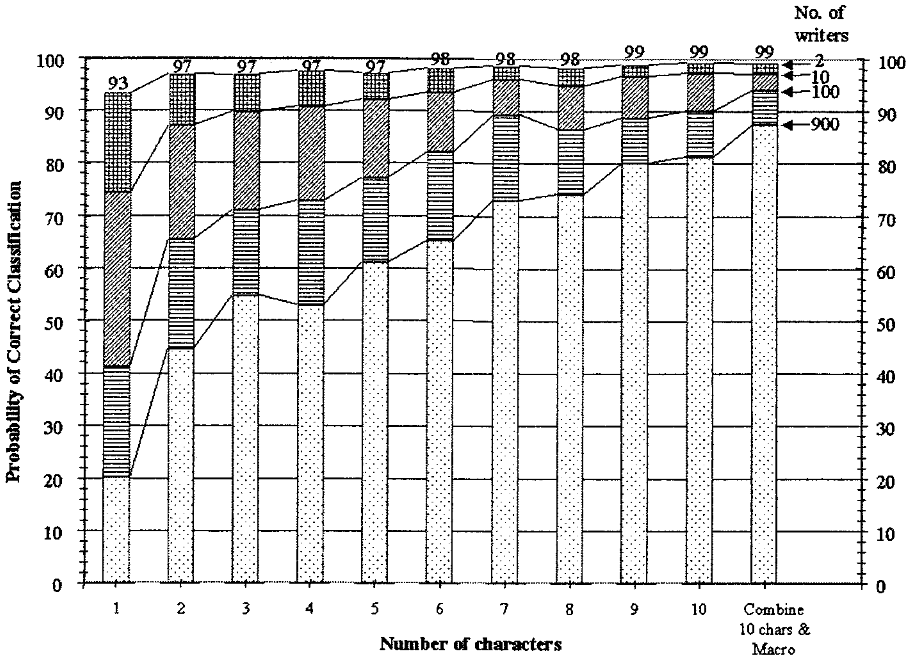


FIG. 15—*Writer identification accuracy using micro-features: shown as a function of the number of allograph shapes considered. (The number of writers is 975.) The characters were: r, e, f, e, r, r, e, d, b, h in increasing groupings considered (1 to 10). The last column shows the result of combining the micro-features of ten characters together with the macro-features of the entire document.*

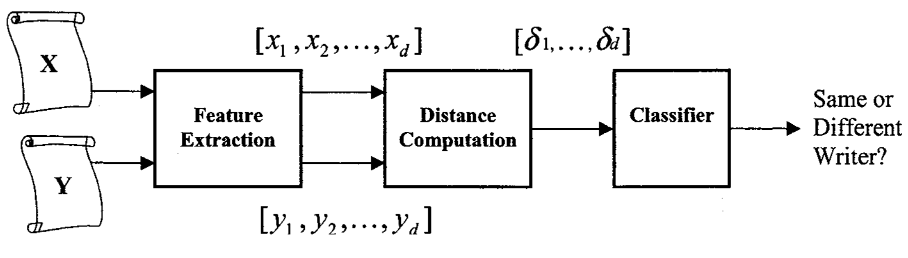

X
Y
$[x_1, x_2, \ldots, x_d]$
$[y_1, y_2, \ldots, y_d]$
$[\delta_1, \ldots, \delta_d]$
Feature Extraction
Distance Computation
Classifier
Same or Different Writer?

FIG. 16—*Verification model: feature vectors for each sample are computed as $[x_1, \ldots, x_d]$ and $[y_1, \ldots, y_d]$. Their distances along each feature, $[\delta_1, \ldots, \delta_d]$, are used by a classifier to determine whether the distance vector is classified as within- or between-writer.*

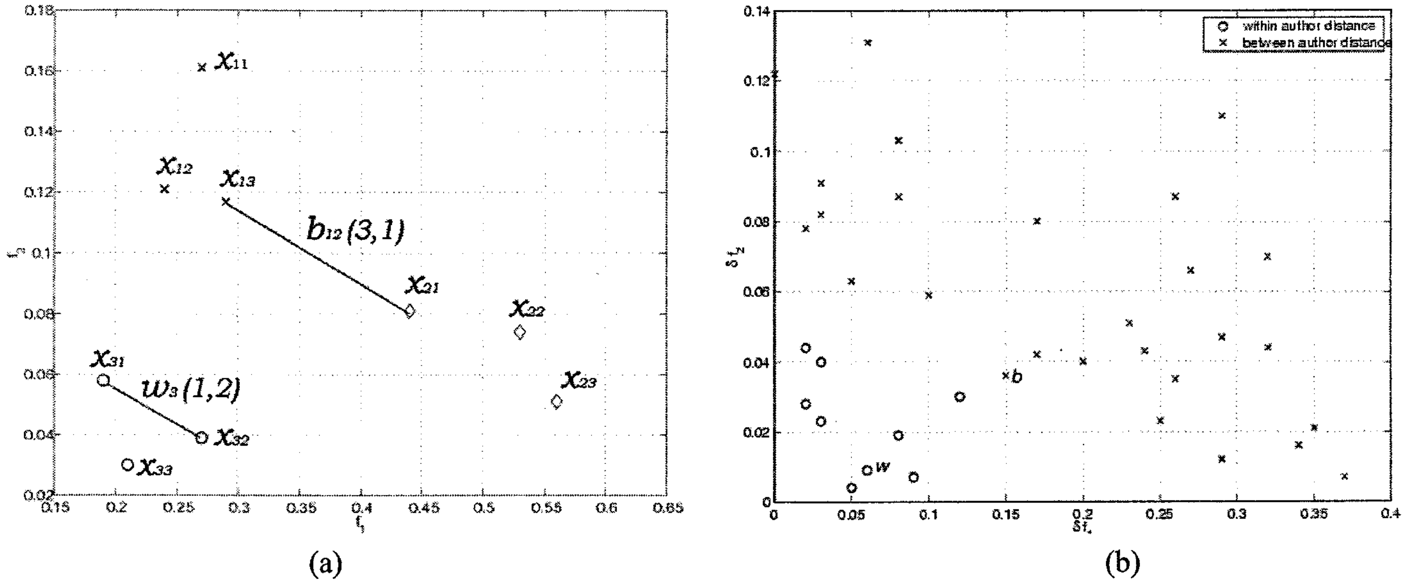


FIG. 17—*The dichotomy model: transformation from feature domain to feature-distance domain. a) Feature space: two features are extracted from each of the three samples of handwriting provided by three writers. Handwriting samples of each writer cluster together. b) Distance space: the distance between the feature vectors is mapped onto feature-distance vectors in the feature-distance space. Within-writer and between-writer distances cluster together.*

binatorially with n, the number of writers. With $n = 1000$, there are 3000 within-writer distances and 4,495,500 between-writer distances. We represented these distances as points in a d-dimensional distance-space, where each dimension represented the distance along a feature.

To generalize, let x_{ij} denote the feature vector of the j^{th} handwriting sample of the i^{th} writer. Distances between samples of the same class are given by $w_i(j, k) = \delta(x_{ij}, x_{ik})$ and the distances between samples of different classes are given by $b_{il}(j, k) = \delta(x_{ij}, x_{lk})$, $i \neq l$ where δ, the distance between feature vectors of two handwriting samples $X = [x_1, \ldots, x_d]$ and $Y = [y_1, \ldots, y_d]$, is given by the distance vector $\delta(X, Y) = [|x_1 - y_1|, |x_2 - y_2|, \ldots, |x_d - y_d|]$. At micro-feature level, the distance between two documents is computed quite differently. Assume that each document is represented by a set of k characters $(c_1, \ldots, c_k)$. For each character, the GSC feature generator extracts a 512-dimensional binary feature vector. Using the similarity measure given in Eq 1, the distance is computed for each pair of k characters. Hence, the distance vector between sets of k characters is given by $\delta(X, Y) = [d(x_{c1}, y_{c1}), \ldots, d(x_{ck}, y_{ck})]$.

Most statistical experiments require the assumption that observed data be statistically independent. Distance data points are not statistically independent, since knowing two distances for a given person, the third distance is bounded by the triangle inequality for metrics. A solution is to choose randomly a smaller sample from a large sample. We partition 3000 within-writer distance data into disjoint subsets of 500. Similarly, we randomly select several subsets of 500 in size from the between-writer distance data set. These subsets are used for training, validating, and testing purposes.

The accuracy of performing the dichotomy by using a given set of features can be measured by the probability of misclassification: *Type-I error* is defined as the probability of misclassifying two handwriting samples as written by two different writers when they actually were written by the same writer; *Type-II error* is defined as the probability of misclassifying two handwriting samples as written by the same writer when they actually were written by two different writers. Our goal was to minimize the misclassification error. Type-I and Type-II errors for the within- and between-writer distributions are illustrated in Fig. 18.

Learning Algorithm—There are several methods available for statistical classification. When the number of classes is few, which is true in the verification model since there are only two classes, a machine-learning technique that is accurate and yet easy to implement is based on artificial neural networks (ANNs). We used an ANN to classify the between- and within-writer distances while minimizing misclassification errors. ANNs have several desirable properties: (i) they are a sound statistical procedure (23), (ii) they are a practical software implementation of the Bayesian (optimal) procedure (25), (iii) they make no presumptions about the nature of the data (unlike other classifiers), and (iv) they let us tap into the full multivariate nature of the data and enable us to use a non-linear discrimination criterion. We used a 3-layered (Fig. 19) network: an input layer with eight units and a hidden layer with five units.

Verification Accuracy—Verification accuracy was determined with varying amounts of information available in the handwritten samples. The results, corresponding to the macro-features of the entire document, a paragraph (address block) and a word ("referred"), are shown in Fig. 20. Micro-feature results with ten characters are shown in Fig. 21. Details of the methods used to perform the testing at the document, paragraph, word, and character levels are as follows:

(i) *Document Level*: In order to ensure independence in the data and to avoid testing on the training data, we divided the writers up into four groups of 250 each. Within- and between-writer distances were then computed within these groups. We used one group for training, one for validation, and one each for two test sets. We trained the ANN using 750 within-writer distances

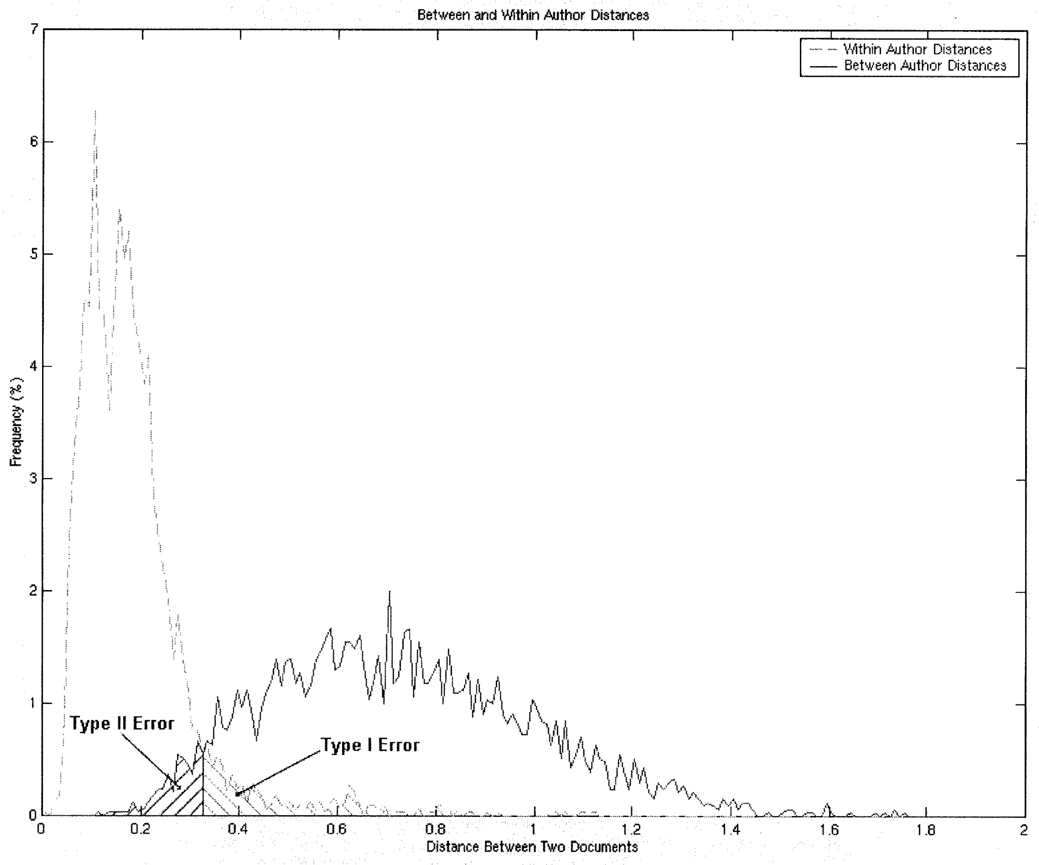


FIG. 18—*Error probabilities in distance space: Type-I and Type-II errors for within- and between-writer distributions with only one measured feature.*

and 750 between-writer distances (of 250 individuals). We then tested it on two separate (previously unseen) test sets, each with 750 within-writer distances and 750 between-writer distances. The training set classified 95% of the data correctly (with Type-I error = 6.3% and Type-II error = 3.8%). The two test sets classified 96% (with Type-I error = 4.5% and Type-II error = 3.6%) and 94% (with Type-I error = 7.5% and Type-II error = 4.4%) of the data correctly.

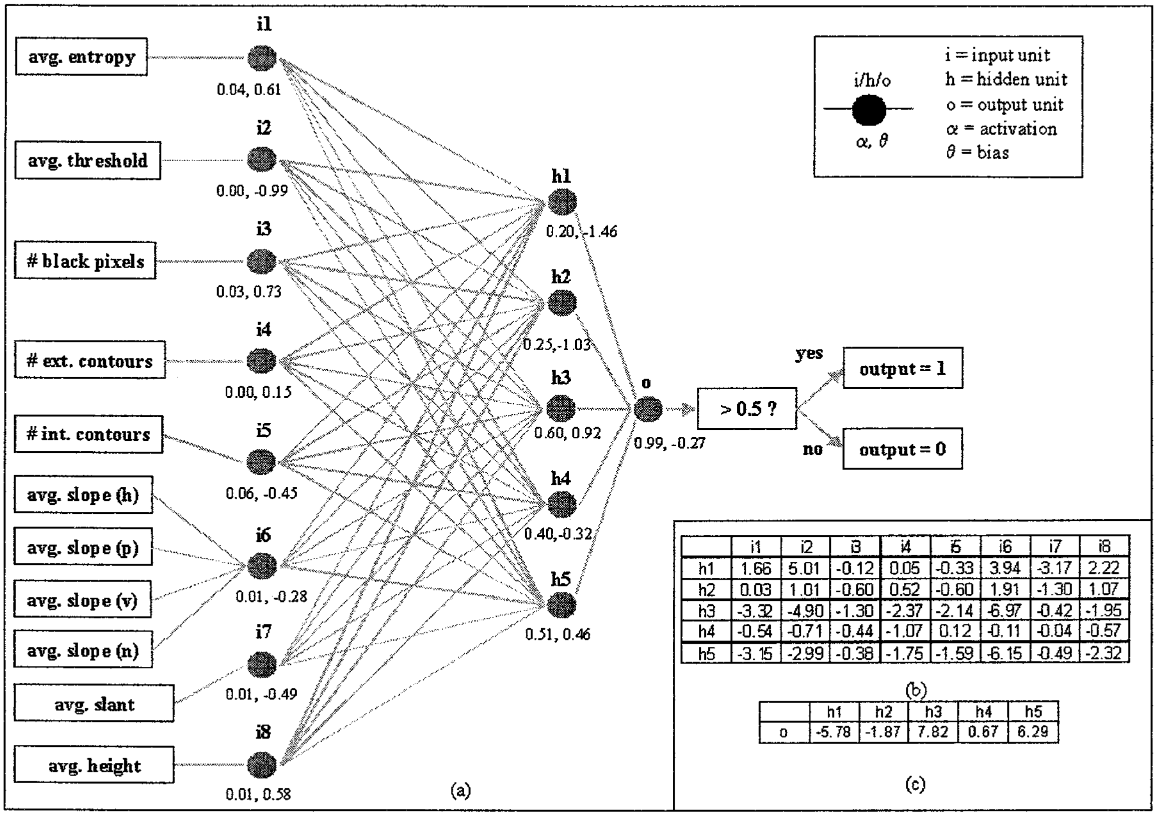
(ii) *Paragraph Level*: Using macro-features for the address block, we trained the ANN using 711 within-writer distances and 711 between-writer distances (of 237 individuals). We then tested it on two separate (previously unseen) test sets each with 711 within-writer distances and 711 between-writer distances. The training set classified 90% of the data correctly (with Type-I error = 11.8% and Type-II error = 7.5%). The two test sets classified 89% (with Type-I error = 14.2% and Type-II error = 7.6%) and 87% (with Type-I error = 16.9% and Type-II error = 9.6%) of the data correctly.

(iii) *Word Level*: Using macro-features for the word "referred," we trained the ANN using 834 within-author distances and 836 between-writer distances. We then tested it on two separate (previously unseen) test sets, each with 834 within-writer distances and 836 between-writer distances. The training set classified 82.3% of the data correctly (with Type-I error = 18% and Type-II error = 17.3%). The two test sets classified 83.1% (with Type-I error = 14.5% and Type-II error = 19.3%) and 82.7% (with Type-I error = 14.4% and Type-II error = 20.2%) of the data correctly.

(iv) *Character Level*: Based on micro-features of ten characters *r*, *e*, *f*, *e*, *r*, *r*, *e*, *d*, *b*, *h*, we trained the ANN using 723 within-author distances and 723 between-writer distances (of 964 individuals). We then tested it on two separate (previously unseen) test sets each with 723 within-writer distances and 723 between-writer distances. The training set classified 91.2% of the data correctly (with Type-I error = 9.8% and Type-II error = 7.7%). The two test sets classified 91.1% (with Type-I error = 12.4% and Type-II error = 5.3%) and 91.8% (with Type-I error = 10.0% and type-II error = 6.5%) of the data correctly. The same experiments with different number of characters were performed and, as shown in Fig. 21, we observe that higher accuracy is achieved with more characters considered.

Comparison of the Two Models

The discriminative power of handwriting using the features extracted was established by using two different approaches, both based on classificatory models: (i) the approach of identifying the writer from a set of possible writers, and (ii) the approach of determining whether two documents were written by the same writer. Writer identification accuracy was close to 98% for two writers. In the verification approach, the features were mapped onto the fea-



	i1	i2	i3	i4	i5	i6	i7	i8
h1	1.66	5.01	-0.12	0.05	-0.33	3.94	-3.17	2.22
h2	0.03	1.01	-0.60	0.52	-0.60	1.91	-1.30	1.07
h3	-3.32	-4.90	-1.30	-2.37	-2.14	-6.97	-0.42	-1.95
h4	-0.54	-0.71	-0.44	-1.07	0.12	-0.11	-0.04	-0.57
h5	-3.15	-2.99	-0.38	-1.75	-1.59	-6.15	-0.49	-2.32

	h1	h2	h3	h4	h5
o	-5.78	-1.87	7.82	0.67	6.29

FIG. 19—*Artificial neural network used to classify within- and between-writer distances: a) Fully connected, feed-forward, back-propagation, 8-5-1 neural network. The feature distance vector is presented at the input layer. The neural network then classifies it as a within- or between-writer distance. A 1 at the output implies different writers, and a 0 implies the same writer. The sigmoid function on each unit is defined by the activation (α) and bias (θ) values. b) Weights on edges connecting input units to hidden units. c) Weights on edges connecting hidden units to output unit.*

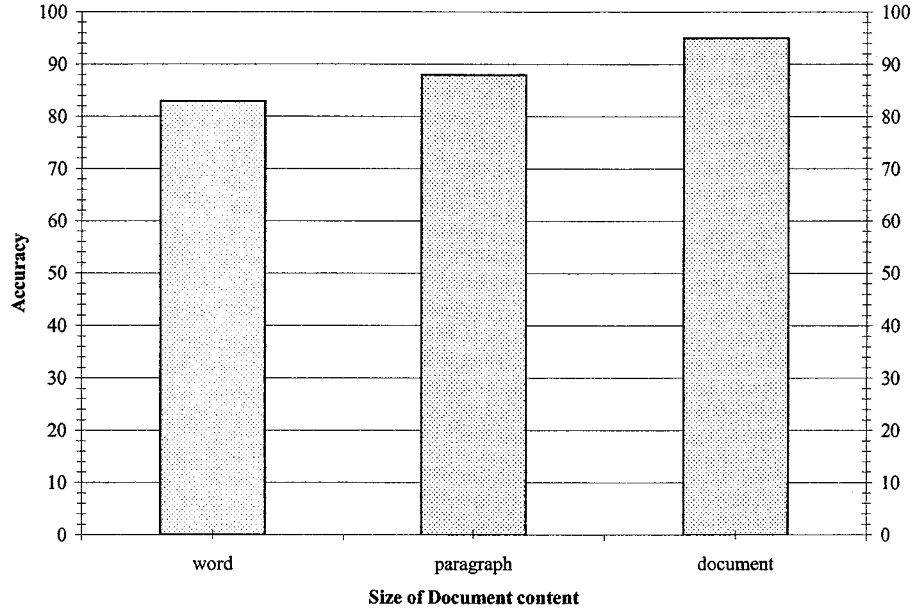


FIG. 20—*Verification analysis using macro-features: performance at word "referred", paragraph (address block), and document levels.*

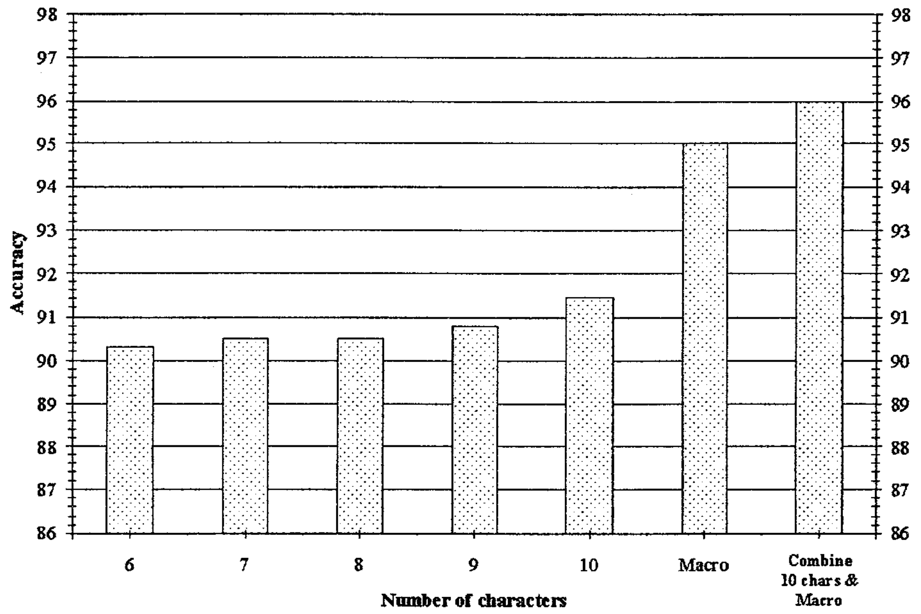


FIG. 21—*Verification analysis using micro-feature: performance at the character level (r, e, f, e, r, r, e, d, b, h). The right-most column shows performance combining the micro-features of the characters with the macro-features of the entire document.*

ture distance domain, and the individuality problem was tackled as a 2-class problem of classifying within- and between-author distances. Verification accuracy was about 96%.

The verification model has a slightly lower accuracy, as can be expected due to its mapping into a space of distances before performing classification. It was seen that performance deteriorated with a decrease in document content for both models. The verification model cannot be parameterized corresponding to the number of writers considered, unlike the identification model. However, repeated application of the verification model, considering one writer at a time, will yield a method of identification. Such a use of the verification model will have a reject option built in.

The principal advantage of the verification model over the identification model is its statistical generality. The identification model is easy to set up for establishing individuality as long as a substantial number of instances for every class is observable. When the number of classes is too large, e.g., the U.S. population, most parametric or non-parametric multiple classification techniques are of no use to validate the individuality of classes, and the problem is seemingly insurmountable.

In the verification model, one need not observe all classes, yet it allows for inferential classification of patterns. It is a method for measuring the reliability classification about the entire set of classes based on samples obtained from a small sample of classes.

Summary and Conclusion

A study was conducted for the purpose of establishing the individuality of handwriting. The work was motivated by U.S. high court rulings that require expert testimony be backed by scientific methodology. Since handwriting had not been subjected to such a study, we decided to undertake this endeavor.

A database was built representing the handwriting of 1500 individuals from the general U.S. population. The sample population was made representative of the U.S. population by stratification and proportional allocation. The population was stratified across different genders, age groups, and ethnicities. Each individual provided three handwritten samples, produced by copying-out a source document that was designed to capture many attributes of the English language: document structure; positional variations of alphabets, numerals, and punctuation; and interesting alphabet and numeral combinations. Features were extracted at a global level of the document, from the entire document, from a paragraph of the document, and from a word of the document. Finer features were extracted at the character level from each sample.

Individual discriminability was established by using a machine-learning approach where some samples are used to learn writer characteristics, and other samples are used to test the learned models. Based on a few macro-features that capture global attributes from a handwritten document and micro-features at the character level from a few characters, we were able to establish with a 98% confidence that the writer can be identified. Taking an approach that the results are statistically inferable over the entire population of the U.S., we were able to validate handwriting individuality with a 96% confidence. By considering finer features, we should be able to make this conclusion with a near 100% confidence.

An assumption here is that we have a representative sample of handwriting. For instance, it would not be possible to establish the individuality of handwriting based on a single stroke of handwriting.

Our work has employed handwriting features similar to, but not exactly the same as, those used by document analysts in the field. However, the objective analysis that was done should provide the basis for the conclusion of individuality when the human analyst is measuring the finer features by hand.

There are many important extensions of the work that could be done. Some of these are to study the handwriting of similarly trained individuals, to study temporal variations of handwriting over periods of time, etc.

Acknowledgments

This work was funded by National Institute of Justice (NIJ) grant 1999-IJ-CX-K010. We are grateful to Dr. Richard Rau who introduced us to the domain of forensic science and provided guidance. We would like to thank Eugenia Smith for assistance in collecting handwriting samples and Kristen Pfaff for editorial assistance.

References

1. U.S. Court of Appeals, *Frye v. United States*. 54 App. D.C. 46, 47, 293 F. 1013, 1014, 1923.
2. U.S. Supreme Court ruling, *Daubert v. Merrel Dow Pharmaceuticals*. 509 U.S. 579, 1993.
3. U.S. District Court ruling, *United States v. Starzecpyze*l. 880 F. Supp. 1027 (S.D.N.Y.), 1995.
4. U.S. Supreme Court ruling, *General Electric Co. v. Joiner*. (96–199) 78 F.3d 524.
5. U.S. Supreme Court ruling, *Kumho Tire Co. v. Carmichael*. (97–1709) 131 F.3d 1433.
6. U.S. 11th Circuit Court of Appeals, *United States vs. Paul*. (97–9302), 1999.
7. Huber RA, Headrick AM. Handwriting identification: facts and fundamentals. Boca Raton: CRC Press, 1999.
8. Osborn AS. Questioned document. 2nd ed. Albany, NY: Boyd Printing, 1929.
9. Lohr, SL. Sampling: design and analysis. Pacific Grove, CA: Duxbury Press, 1999.
10. Srihari SN, Cha S-H, Arora H, Lee S. Handwriting identification: research to study validity of individuality of handwriting & develop computer-assisted procedures for comparing handwriting. Buffalo (NY): University at Buffalo, State University of New York; 2001 TR No.: CEDAR-TR-01-1.
11. Gilbert AN, Wysocki CJ. Hand preference and age in the United States. Neuropsychologia 1992; 30:601–8.
12. Duda RO, Hart PE. Pattern classification and scene analysis. NY: Wiley, 1973.
13. Srihari SN. Feature extraction for locating address blocks on mail pieces. In: Simon JC, ed. From pixels to features. Amsterdam: North Holland, 1989;261–73.
14. Srihari SN. Recognition of handwritten and machine-printed text for postal address interpretation. Pattern Recognition Letters 1993;14: 291–303.
15. Govindaraju V, Shekhawat A, Srihari SN. Interpretation of handwritten addresses in U.S. mail stream. In: Proceedings of the 2nd International Conference on Document Analysis and Recognition; 20–22 Oct. 1993; Tsukuba Science City, Japan: International Association for Pattern Recognition, 1993.
16. Srikantan G, Lam SW, Srihari SN. Gradient-based contour encoding for character recognition. Pattern Recognition 1996;29:1147–60.
17. Srikantan G, Lee DS, Favata JT. Comparison of normalization methods for character recognition. In: Proceedings of the 3rd International Conference on Document Analysis and Recognition; 14–16 August 1995; Montreal: International Association for Pattern Recognition, 1995.
18. Otsu N. A threshold selection method from gray-scale histograms. IEEE Trans System, Man and Cybernetics 1979;9:62–6.
19. Freeman H. On the encoding of arbitrary geometric configurations. IRE Trans Electronic Computers 1961;18:312–24.
20. Kim G, Govindaraju V. A lexicon-driven approach to handwritten word recognition for real-time applications. Trans on Pattern Analysis and Machine Intelligence 1997;19:366–79.
21. Favata JT, Srikantan G, Srihari SN. Handprinted character/digit recognition using a multiple feature/resolution philosophy. In: Proceedings of the Fourth International Workshop on the Frontiers of Handwriting Recognition; 7–9 December 1994; Taipei: NA, 1994.
22. Gonzalez RC, Woods RE. Digital image processing. 3rd ed. Reading, MA: Addison-Wesley, 1992.
23. Mirkin B. Mathematical classification and clustering. Dordrecht: Kluwer Academic Pub, 1996.
24. Mitchell TM. Machine learning. Boston: McGraw-Hill, 1997.
25. Lee DS, Srihari SN, Gaborski R. Bayesian and neural network pattern recognition: a theoretical connection and empirical results with handwritten characters. In: Sethi IK, Jain AK, ed. Artificial neural networks and statistical pattern recognition. Amsterdam: North Holland, 1991: 89–108.

Additional information and reprint requests:
Sargur N. Srihari
CEDAR
520 Lee Entrance, Suite 202
Amherst, NY 14228-2567
E-mail: Srihari@cedar.buffalo.edu