A Database for Handwritten Text Recognition Research

Jonathan J. Hull

***Abstract*—An image database for handwritten text recognition research is described. Digital images of approximately 5000 city names, 5000 state names, 10000 ZIP Codes, and 50000 alphanumeric characters are included. Each image was scanned from mail in a working post office at 300 pixels/in in 8-bit grayscale on a high-quality flat bed digitizer. The data were unconstrained for the writer, style, and method of preparation. These characteristics help overcome the limitations of earlier databases that contained only isolated characters or were prepared in a laboratory setting under prescribed circumstances. Also, the database is divided into explicit training and testing sets to facilitate the sharing of results among researchers as well as performance comparisons.**

***Index Terms*—Handwriting recognition, database, performance analysis, testing.**

I. Introduction

The recognition of handwritten characters and words is a difficult research problem where the complexity of the solution depends on the constraints that are placed on the formation of the original handwriting. A significant aspect of handwriting recognition in domains such as bank checks [8] and postal addresses [2] is that there is no control over the author, writing instrument, or writing style. For example, an arbitrary handwritten word might be produced by a felt tip pen and could include isolated, touching, or overlapping characters, cursive fragments, or fully cursive words. Also, varying degrees of neatness are possible, from very sloppy to extremely neat [5]. However, these difficulties are often offset by the constraint that the input words come from a fixed vocabulary. For example, a handwritten city name in the United States can be one of only about 30000 alternatives.

A standard database of images is needed to facilitate research in handwritten text recognition [9]. Previous databases, summarized in [7], are insufficient for a variety of reasons. Most other data sets contain only isolated characters and thus do not reflect the word images that will be encountered in practice. Also, other databases have almost exclusively been constructed from binary (bi-tonal) images, and thus the researcher is limited to the sampling rate and thresholding algorithm provided by a particular digitizer. This precludes experimentation with preprocessing or recognition in the grayscale domain, both of which may provide improvements in recognition accuracy. Another drawback to previous databases is that they are often collected in a laboratory environment in which subjects prepared samples on standard forms that were then digitized. The awareness by subjects that their handwriting would be used to develop automatic recognition algorithms could have introduced biases into the data. The desire to perform well, because of the similarity to a classroom testing environment, may yield samples that are abnormally neat. Alternatively, subjects may try to "fool the computer" by making their samples unusually sloppy.

Manuscript received April 6, 1992; revised April 1, 1993. Recommended for acceptance by Associate Editor R. Kasturi.

The author is with the Center of Excellence for Document Analysis and Recognition (CEDAR), Department of Computer Science, State University of New York at Buffalo, Buffalo, NY 14260 USA; e-mail: hull@cs.buffalo.edu.

IEEE Log Number 9215855.

Some of the criticisms of experimental pattern recognition that are related to the replication of experiments and the comparison of performance among research groups could be answered with a standard database that contains precisely defined training and testing sets [6]. Results could be reported for algorithms that had access to the same training set and were tested on precisely the same data. This would help solve problems in comparing results that are caused by minor differences in the testing data themselves [4]. Performance analysis down to the level of individual samples could also be performed. Furthermore, the sharing of test results among researchers and experimentation with combining algorithms to achieve better performance would be facilitated by the availability of a common database [1], [3]. This could be done by the exchange of plain ASCII files that identify image names and recognition results.

The rest of this correspondence describes a database for handwritten text recognition research that addresses many of the drawbacks of earlier efforts. An application domain of city name, state name, and ZIP Code recognition on handwritten addresses is assumed. This reasonably limits the decision space but still provides a challenging problem: about 30000 classes are possible for city names and 42000 for ZIP Codes. Research in preprocessing algorithms can be performed because the images were scanned at 300 pixels/in (ppi) in 8-bit grayscale on a high-quality flat bed digitizer. Also, the data were scanned from live mail in a post office. Thus, the subjects that prepared the samples had no idea that they would be used for algorithm research, thereby eliminating the subject-bias problem. The database presented here also contains explicitly defined training and testing tests. Approximately 10% of the available samples were randomly selected and placed in the test set. The remainder were retained as training data.

II. Background

The database discussed here contains city names, state names, ZIP Codes, and alphanumeric characters extracted from the digital images of handwritten addresses that were gathered as part of a research project sponsored by the United States Postal Service (USPS). The objective of this project was to develop algorithms to locate and recognize handwritten ZIP Codes.

On two occasions (once in 1987 and again in 1988), data-gathering operations were conducted at the main post office in Buffalo, NY. Mailpieces with handwritten addresses were selected and scanned on a state-of-the-art image digitizer. Truth values were subsequently assigned by drawing bounding boxes around the city, state, and ZIP Code words and typing in their identities. A separate set of isolated ZIP Codes was also gathered and truthed.

The data described here include the original grayscale city, state, and ZIP Code images. An additional set of bi-tonal images of alphabetic and numeric characters is also included.

III. Sampling Procedures

Several two-letter acronyms were chosen to describe the sampling procedures used for subsets of the database. Each acronym is explained below together with a description of the number of images in each subset. The first letter in each acronym indicates the city in which the data were gathered (all "B" for Buffalo). The second letter specifies the sampling procedure. A description of each procedure is

TABLE I
NUMBERS OF HANDWRITTEN WORD IMAGES IN THE TRAINING AND TESTING SETS

Numbers of Training and Testing Data							
		Training Set			Testing Set		
Class	Description	Cities	States	ZIPs	Cities	States	ZIPs
BB	*biased*	363	277	269	37	31	30
BC	*carrier*	190	217	190	21	23	21
BD	*designed*	3106	2490	2201	317	252	238
BL	*local*	877	1017	875	97	114	97
BS	*test*	564	467	450	60	50	49
BR	*random*	n.a.	n.a.	3962	n.a.	n.a.	n.a.
BU	*touching*	n.a.	n.a.	1072	n.a.	n.a.	n.a.
Totals		5100	4468	9019	532	470	435

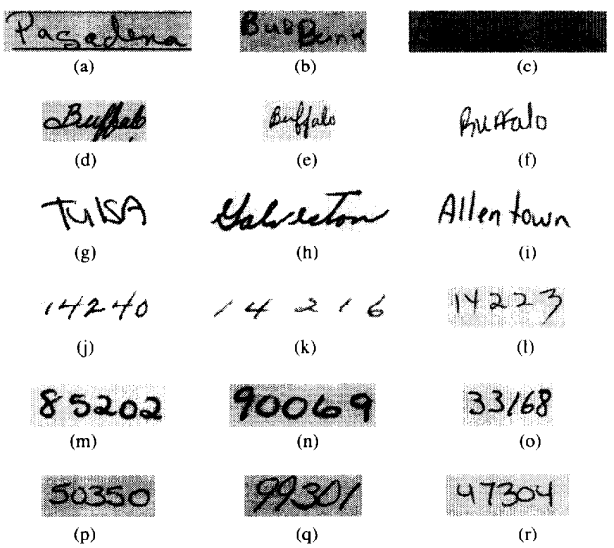

Fig. 1. Examples of images in the database.

provided below, and the correspondence between the second letter of the acronym and this description is indicated.

A. Source Images

The BB addresses (*b*iased) are composed of about 300 images that were selected because they contained a particular characteristic that might make them difficult to recognize. The characteristics included underlines in the address, digits in the ZIP Code that touch other digits, lines with different skew in the same address, ZIP Codes not on the last line, and P.O. boxes with five digits above or below a ZIP Code. These address blocks were balanced on the first digit in the ZIP Code. That is, approximately 30 of the images have a ZIP Code in which the first digit is zero, and so on.

Fig. 1(a)–(c) shows three BB city name images. The bias factors that can occur are illustrated. These include an underline (a), a patterned background (b), and an image with low contrast (c).

The BC addresses (*c*arrier route) contain 211 images that were randomly selected from a single ZIP Code in Buffalo, NY (14222). This database was gathered to support work in routing mailpieces to individual letter carriers inside a specific ZIP Code.

Fig. 1(d)–(f) shows different images of the city name "Buffalo." These illustrate some of the variations in writing style exhibited by the samples in the BC portion of the database.

The BD addresses (*d*esigned) contain 2636 images divided into two subsets. The first subset contains 10 groups of about 200 addresses each, where each group was drawn from a different ZIP Code region (first digit of the ZIP Code determines its region). The same number of addresses were sampled from each state within each region. For example, the states of Washington, Oregon, and California are in ZIP Code region nine. Thus, about 66 images were drawn from each of those states. The other subset of the BD images contains about 500 addresses drawn from 25 cities. Twenty images were gathered that contained each city name. Half of these (ten images) were primarily composed of handprinting and the other half contained a mixture of handprinting and cursive writing. Fig. 1(g)–(i) shows three city words from the BD portion of the database.

The BL addresses (*l*ocal) include 973 images that all contain the city name "Buffalo." This set was gathered for the development of algorithms that take advantage of local mail-stream characteristics to improve performance.

Fig. 1(j)–(l) shows three ZIP Codes in the BL portion of the database. The BL images include examples of ZIP Codes that occur in the city of Buffalo. The city and state words in the BC and BL image sets provide numerous examples of writing style variation across a limited number of possibilities for the city and state name (one each) and ZIP Code (about 25 possibilities).

The BS addresses (*t*est) contain 500 images that were selected with the same technique that was used to gather the BD database. The BS data have been used to demonstrate algorithms developed on the BD image set. The examples from the BD data set shown in Fig. 1(g)–(i) and the comments about those images expressed above also apply to the BS images.

B. ZIP Codes

The BR images (*r*andom) contain 3962 grayscale images of handwritten ZIP Codes [see examples in Fig. 1(m)–(o)]. The sample was approximately balanced on the first two digits of the ZIP Code. Thus, there are about 40 samples of ZIP Codes that begin with each pair of digits. The images were randomly selected within each two-digit group.

The BU data (*u*ouching digits) contain 1072 images of ZIP Codes, in which at least two of the digits touch one another [see examples in Fig. 1(p)–(r)]. These were chosen to provide data that would stress a digit string segmentation algorithm. This sample was also approximately balanced on the first two ZIP Code digits.

IV. DATABASE DEFINITION

Complete 300 ppi 8-bit grayscale city names, state names, and ZIP Codes are in the database as well as bi-tonal alphabetic and numeric

(a) (b)

BENDAVIS, BRINKTOWN, CUBA, FANNING, INDIAN LAKE, KNOBVIEW, OAK HILL, ELDRIDGE, FT LEONARD WOOD, FORT LEONARD WOOD, CLARA, HOUSTON, LYNCH, PINEY, PRESCOTT, SIMMONS, TYRONE.

(c)

ALMARTHA, ARCOLA, ASH, BELLE, BENDAVIS, BERWICK, BIG CREEK, BILLINGSVILLE, BOLIVAR, BOONVILLE, BRINKTOWN, BUCKHORN, BURNHAM, BURNS, BUTTERFIELD, BYRON, CALHOUN, CASSVILLE, CENTER POINT, CENTERTOWN, CLARA, CLARK, CLARKS FORK, CLEAR SPRINGS, COLUMBIA, CRANE, CUBA, CULLEN, DIV OF FAMILY SERVICES, DIXIE, DUNCANS BRIDGE, EASLEY, EBENEZER, EDGEWATER BEACH, ELDRIDGE, ELKHEAD, ELSEY, FANNING, FLORIDA, FORSYTH, FORT LEONARD WOOD, FT LEONARD WOOD, GLIDEWELL, GOOCH MILL, GOSPEL RIDGE, GRANT, GUTHRIE, HALF WAY, HIGH GATE, HOLTS SUMMIT, HOUSTON, HOUSTONIA, HUNTSDALE, HURON, HUTTON VALLEY, INDIAN LAKE, KNOBVIEW, KOENIG, LAMINE, LANES PRAIRIE, LEWIS STATION, LOHMAN, LYNCH, LYNCHBURG, MADISON, MARION, MCBAINE, MCCRACKEN, MCKINLEY, MIDDLE GROVE, NEW BLOOMFIELD, NIANGUA, NORTH, OAK HILL, ONGO, OVERTON, OZARK BEACH, OZARK SPRINGS, PAYDOWN, PIERCE, PIERCE CITY, PINEY, PRESCOTT, PROSPECT, PROTEM, PULASKIFIELD, RADER, ROMANCE, ROUBIDOUX, RUCKER, SAINT ROBERT, SAINT ROBERTS, SAMPSON, SAPP, SCHOFIELD, SHADY GROVE, SIMMONS, SLAGLE, SPARTA, SPEED, SPFD, SPFLD, SPGFD, SPGFLD, SPRGFLD, SPRINGFIELD, ST ROBERT, ST ROBERTS, STOUTSVILLE, SUMMERFIELD, SUMMIT CITY, TERESITA, TYRONE, ULMAN, VAN, VANCE, WAINWRIGHT, WASOLA, WAYNESVILLE, WILLHOIT, WILLOW SPGS, WILLOW SPRINGS, WINDYVILLE, WOODLAWN, YONKERVILLE.

(d)

Fig. 2. An example of a city, state, and ZIP Code image from the test data. The lexicons that result by allowing one or two digits to vary are also shown.

characters. The bi-tonal images were derived from the corresponding gray level data by application of a thresholding algorithm. This section describes how the data were divided into training and testing sets.

A. Handwritten Words

The handwritten words (cities, states, and ZIP Codes) were divided into explicit training and testing sets. This was done by randomly assigning approximately 10% of the ZIP Code images as well as the city and state images from the corresponding address to the test data. The remainder of the data were retained for training purposes. The breakdown into training and test data by class is shown in Table I.

Multiple-word city or state names are split into separate files but not split across the training and test sets. Each city and state word image in the test data is provided with three lexicons that simulate recognition results for the corresponding ZIP Code. Often, only some of the digits in a ZIP Code can be recognized with high confidence and the other digits must be rejected. The lexicons provided here simulate cases where two, three, and four of the digits in each ZIP Code were confidently recognized. This was done by randomly choosing one, two, and three positions in each ZIP Code and allowing the corresponding digits to vary, thus simulating their rejection. The positions were randomly chosen in each ZIP Code to simulate an image-to-image variation in performance.

The three simulated partial ZIP Code recognition results were matched against the USPS list of legal ZIP Code–city–state combinations. The city and state names with corresponding ZIP Codes that matched each of the three conditions were placed in the lexicons. It should be noted that the ZIP Code–city–state list is provided with the database. Researchers can thus experiment with various strategies for simulating ZIP Code recognition.

Fig. 2 shows an example of lexicon generation and how a word recognition algorithm could take advantage of partial information from recognition of the ZIP Code. If the fourth digit of the ZIP Code shown in Fig. 2(b) were rejected, a lookup in the USPS database would provide the 17 city names shown in Fig. 2(c). Instead, if the digits in positions three and four were rejected, the 125 city names shown in Fig. 2(d) would be returned. In either case, a word recognition algorithm could use these lexicons to narrow its focus of attention. That is, the decision space would be reduced from approximately 30000 possible city names to either 17 or 125.

B. Alphabetic and Numeric Character Data

The bi-tonal images of alphabetic and numeric characters in the database are divided into two groups. One set contains mixed alphabetics and numerics and the other set contains numeric characters only. The mixed alphabetics and numerics were extracted from entire address blocks and are suitable for general character recognition algorithm development. The set of numeric characters only were specifically segmented from ZIP Codes and are suitable for digit recognition algorithm development and testing.

The mixed alphabetic and numeric data were segmented from the BD and BL address blocks. This was done to satisfy a need for training data for character recognition algorithm development. The truthing of these data was performed by a procedure that extracted connected components from an address image and displayed them in isolation to an operator. A truth value was assigned to a component if its truth was obvious in isolation. If the truth was not obvious, the operator used the coordinates of the component to locate it in the original address block. The surrounding context of the component was then used to assign the truth value. It should be noted that this procedure caused each of the mixed alphabetic and numeric characters to be composed of just one component. A test set was chosen from these data by randomly choosing 10% of the available images. The breakdown into training and testing sets is shown in Table II. Altogether, there are 24947 characters in the training set and 2890 in the testing set.

The set of numeric characters (digits) only were extracted from two sources. The digits in the training set were extracted from the BR ZIP Codes after they were truthed. The BR ZIP Codes were segmented into isolated digits and the truth value for each position in the ZIP Code was mapped onto the isolated digit that was output by the segmentation algorithm at the corresponding position. Each digit image was then manually inspected. Any images that resulted

TABLE II
MIXED ALPHABETIC AND NUMERIC CHARACTER SET DISTRIBUTION

Data Set	Composition												
	A	B	C	D	E	F	G	H	I	J	K	L	M
	1237	595	588	388	490	287	143	274	490	68	160	563	588
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	1022	905	516	3	749	834	441	268	201	249	112	259	24
Training	a	b	c	d	e	f	g	h	i	j	k	l	m
	527	84	211	249	736	120	93	205	803	1	94	684	184
	n	o	p	q	r	s	t	u	v	w	x	y	z
	469	833	129	5	460	490	428	333	127	102	173	136	15
		0	1	2	3	4	5	6	7	8	9		
		811	1160	778	467	607	342	469	429	346	393		
	A	B	C	D	E	F	G	H	I	J	K	L	M
	162	69	51	49	57	32	19	25	56	15	19	87	59
	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
	123	102	59	2	97	77	54	31	27	30	19	39	7
Testing	a	b	c	d	e	f	g	h	i	j	k	l	m
	68	8	17	26	107	16	14	23	70	2	14	69	21
	n	o	p	q	r	s	t	u	v	w	x	y	z
	60	84	7	2	47	29	46	26	20	15	10	15	0
		0	1	2	3	4	5	6	7	8	9		
		102	136	103	68	63	41	47	48	46	53		

TABLE III
BI-TONAL DIGIT IMAGES IN THE DATABASE

	Class										
Data Set	0	1	2	3	4	5	6	7	8	9	Total
Training—BR	2866	2544	2047	1731	1676	1459	1722	1616	1453	1354	18468
Test—All BS	434	345	296	260	234	193	281	241	216	211	2711
Test—Good BS	355	289	224	208	183	117	245	221	191	180	2213

from poor segmentations or had been assigned incorrect truth values were removed. The digits in the test set were generated from the BS ZIP Codes. A segmentation algorithm was applied to these images, and every single digit that was output by this technique was saved to a separate file. This resulted in a test data set of 2711 digit images. Each of these digits is provided with its associated truth value by corresponding the position of the segmented digit to the truth value at that position in the ZIP Code. The objective of this test set is to provide as realistic as possible a simulation of the data that would be encountered by an isolated digit recognition algorithm when it is applied to the results of segmenting handwritten ZIP Codes.

A carefully chosen subset of the BS digits is also provided. This set contains 2213 digits chosen from the 2711. Each of the 2213 were judged to be well segmented by three research groups. Thus, any obvious segmentation artifacts should have been removed. Table III provides a histogram of the number of images in each digit class in both the training and test sets.

V. SUMMARY AND CONCLUSION

The image database discussed in this correspondence addresses several important aspects not covered by most other data sets. Only images scanned from mailpieces in a working post office are included. This overcomes the subject-bias problems of other databases that were scanned in laboratory settings. Also, the data were scanned at 300 pixels/in in 8-bit grayscale. This allows for experimentation with preprocessing and grayscale recognition techniques.

The context that can be provided by a restricted domain was also addressed in the design of the database. The inclusion of cities, states, and ZIP Codes from the same address block allows for partial results from ZIP Code recognition to provide constraints for the recognition of city and state names. This is a realistic simulation of how handwritten word recognition can be applied, in practice, and can thus show the effect that improvements in recognition of digits in a ZIP Code can have on the recognition of city and state names.

APPENDIX
DATABASE SPECIFICATION AND AVAILABILITY

Medium

The *data* are provided on an ISO-9660 format CDROM. The CDROM is readable on either PC's or workstations equipped with the appropriate hardware and driver software.

Format

The image files are compressed with a public domain difference coding algorithm. The bi-tonal images are stored as 1-bit/pixel. Complete C language source is provided for the difference coding algorithm and a routine to convert the 1-bit/pixel data to 1-byte/pixel format.

Imager Digitizer

All the images were scanned on an Eikonix EC850 4096x4096 CCD digitizer.

File Naming

Each file is assigned a unique name that will allow for individual researchers to compare performance on specific images.

Storage Requirements

The number of megabytes of storage for each portion of the database are now listed.

Megabytes for Each Portion of the Database

Database Component	Training	Testing
Grayscale cities	152	16
Grayscale states	83	9
Grayscale ZIP Codes	193	9
Mixed bi-tonal alphabetics and numerics	52	6
Bi-tonal numerics only	38	10

Overall, approximately 600 Mbytes of the CDROM are used. This includes the storage needed for formatting information.

Availability

The database described in this correspondence is available from the Center of Excellence for Document Analysis and Recognition (CEDAR) at the State University of New York at Buffalo.

ACKNOWLEDGMENT

Prof. S. N. Srihari, Director of CEDAR, was the Principal Investigator for the handwritten ZIP Code recognition project. Dr. J. Tan of Arthur D. Little, Inc., helped shape the composition of the database. Dr. E. Cohen assisted in the supervision of the data capture process. J. Giattino directed the formatting of the CDROM.

REFERENCES

[1] R. Bradford and T. Nartker, "Error correlation in contemporary OCR systems," in *Proc. 1st Int. Conf. Document Anal. Recogn.*, Saint-Malo, France, Sept. 30–Oct. 2, 1991, pp. 516–524.

[2] E. Cohen, J. J. Hull, and S. N. Srihari, "Understanding handwritten text in a structured environment: Determining ZIP Codes from addresses," *Int. J. Pattern Recogn. Artif. Intell.*, vol. 5, no. 1 & 2, pp. 221–264, 1991.

[3] T. K. Ho, J. J. Hull, and S. N. Srihari, "Combination of decisions by multiple classifiers," in *Structured Document Image Analysis*, H. S. Baird, H. Bunke, and K. Yamamoto, Eds. New York: Springer-Verlag, 1992, pp. 188–202.

[4] J. J. Hull, A. Commike, and T. K. Ho, "Multiple algorithms for handwritten character recognition," in *Proc. Int. Workshop Frontiers Handwriting Recogn.*, Montreal, Canada, Apr. 2–3, 1990, pp. 117–130.

[5] J. J. Hull, T. K. Ho, J. Favata, V. Govindaraju, and S. N. Srihari, "Combination of segmentation-based and wholistic handwritten work recognition algorithms," in *Proc. From Pixels to Features III: Int. Workshop Frontiers Handwriting Recogn.*, Bonas, France, Sept. 23–27, 1991, pp. 229–240.

[6] G. Nagy, "Candide's practical principles of experimental pattern recognition," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-5, pp. 199–200, Mar. 1983.

[7] ——, "At the Frontiers of OCR," *Proc. IEEE*, vol. 7, pp. 1093–1100, July 1992.

[8] J. C. Simon and O. Baret, "Cursive word recognition," in *Proc. From Pixels to Features III: Int. Workshop Frontiers Handwriting Recogn.*, Bonas, France, Sept. 23–27, 1991, pp. 1–20.

[9] C. Y. Suen, presented at the Int. Workshop Frontiers Handwriting Recogn., Montreal, Canada, Apr. 2–3, 1990.

Feature-Preserving Clustering of 2-D Data for Two-Class Problems Using Analytical Formulas: An Automatic and Fast Approach

Ja-Chen Lin and Wen-Hsiang Tsai

***Abstract*—We propose in this correspondence a new method to perform two-class clustering of 2-D data in a quick and automatic way by preserving certain features of the input data. The method is analytical, deterministic, unsupervised, automatic, and noniterative. The computation time is of order n if the data size is n, and hence much faster than any other method which requires the computation of an n-by-n dissimilarity matrix. Furthermore, the proposed method does not have the trouble of guessing initial values. This new approach is thus more suitable for fast automatic hierarchical clustering or any other fields requiring fast automatic two-class clustering of 2-D data. The method can be extended to cluster data in higher dimensional space. A 3-D example is included.**

***Index Terms*— Two-class clustering, cluster representatives, feature-preserving, analytical formulas, decision boundary, automatic fast clustering, k-means, hierarchical methods.**

I. INTRODUCTION

Two-class clustering problems are frequently encountered in real applications. For example, block truncation coding for image compression [1], divisive clustering for hierarchical clustering [2], binary decision tree construction, etc. It is therefore desired to develop a fast automatic method that can be employed to partition an input set H of n patterns into two classes. Unfortunately, most of the clustering tools developed so far, such as the k-means method [3], the divisive method using a dissimilarity matrix [4], etc., are iterative and thus unsuitable for performing fast automatic two-class clustering.

It is desirable to avoid iterative computation by using mathematical formulas to express the decision boundary, which separates the two classes, in terms of the input patterns directly. One way of achieving this goal based on the moment-preserving principle is explained below. When the n input patterns are one-dimensional, say, forming a set $H = \{x_i\}_{i=1}^{n}$, the partition of H into two disjoint clusters H_A and H_B is an easy job. We may assume that every pattern in cluster H_A resembles (in some sense) a single point x_A, and similarly, every pattern in cluster H_B resembles another single point x_B. The two points x_A and x_B are called cluster representatives. Assume further that the fractions of the numbers of patterns in H_A and H_B are p_A and p_B, respectively. It is clear that

$$p_A + p_B = 1. \tag{1}$$

By preserving the first three moments, i.e., by requiring that

$$p_A x_A^k + p_B x_B^k = \overline{x^k} = \frac{1}{n}\sum_{i=1}^{n} x_i^k \quad \text{for } k = 1, 2, \text{ and } 3, \tag{2}$$

and by the natural requirement (1), we can solve Eqs. (1) and (2) to obtain the four unknowns $\{x_A, x_B, p_A, p_B\}$. The solution can be

Manuscript received March 9, 1992; revised January 5, 1993. This work was supported by the National Science Council, Republic of China, under Contract NSC 81-0408-E-009-589. Recommended for acceptance by Associate Editor R. P. W. Duin.

The authors are with the Department of Computer and Information Science, National Chiao Tung University, Hsinchu, Taiwan 30050, Republic of China.

IEEE Log Number 9214426.