A Multifont Word Recognition System for Postal Address Reading

JÜRGEN SCHÜRMANN

Abstract—**This paper describes the basic design principles of a multifont word recognition system developed for postal address reading. Of the three main subsystems, image preprocessing, single character recognition, and contextual postprocessing, the last two will be considered in detail. A multiple-channel/multiple-choice approach is taken in designing the overall system. The character images produced by the image preprocessing subsystem are fed into three parallel single character recognition (SCR) channels. Each channel classifies the raster image according to one of the three character types: capital letter, small letter, or numeral. A second degree polynomial classifier is required in order to satisfy the multifont requirements of address reading. Each SCR channel outputs a rank-ordered list of potential character meanings for the character type being processed and a channel-specific figure of confidence. This figure of confidence serves a twofold purpose. First, it is used to determine the number of alternatives in the rank-ordered list, and secondly, it is used by the subsequent contextual postprocessor in calculating a word-specific discriminant function designed to discriminate between four different kinds of words: numeric, all upper case, all lower case, and lower case with upper case initial. Based on this discriminant function for every character position, only one channel output is passed on to the word recognition system. From the list of alternatives for each character position, a set of alternative words can be constructed which, with a high probability, contains the correct word. A hash-coded table look-up procedure is applied to compare the set of alternative words with the set of all legitimate words contained in a postal directory.**

Index Terms—**Character recognition, contextual postprocessing, mean-square adaptation, polynomial classifier, postal address reader.**

I. Introduction

In optical character recognition, two different lines of development can be discerned: the single-character approach and the word-recognition approach. Practically all of the existing document and page reading machines belong to the first category. The object to be recognized there is the single character image. These machines do not make an attempt to derive profit from the contextual dependencies among the characters to be read, and in many of the applications such dependencies actually do not exist. The redundancy, which is indispensable for reliable operation in a noisy environment, in this case is bound to be that contained in the character image itself. Redundancy, in this situation, characterizes the fact that not all of the possible raster images occur with equal probability. Actually, only an extremely small fraction of all thinkable raster images is likely to appear during recognition. The redundancy here is decreased by degrading printing quality as well as by increasing the number of fonts to be read. Therefore, a maximum of redundancy and a correspondingly high rate of performance is achieved by single font machines designed for and operating on perfect printing quality.

On the other hand, the word recognition approach utilizes the redundancy of character strings on a word basis. Redundancy in this situation characterizes the fact that again, only a small fraction of all conceivable character strings is likely to appear during recognition. Up to now, there have been only a few practical applications where the word recognition approach has proven to be the only promising way. The most familiar example of this kind is the problem of automatic address reading. The recognition system here is faced with the problem of reading practically unconstrained multifont print with real-life printing quality as produced by the broad spectrum of existing printing devices.

In this case, the redundancy at the raster image level is not sufficient for reliable single character recognition. Reliable operation in such an unconstrained environment can only be achieved by making use of the contextual redundancies, which fortunately, are available in address reading applications. These redundancies become evident when the correct character string of a postal place name, for example, is garbled by single recognition errors. Typically, the garbled version of the name no longer is a legitimate postal place name. Therefore, the selection of correctly recognized words and the rejection of erroneous ones basically can be accomplished by checking the result of the single character recognition subsystem up on the postal directory, thus simply sifting out the legitimate words from those recognized.

Following the word recognition approach, the recognition problem is shifted from the character recognition level to the word recognition level and, in the case of automatic address reading, to still higher logical levels if necessary. Recognizing single characters, in the concept of the word recognition system, now becomes only a subgoal, subordinate to the goal of recognizing complete words—postal place names, street names, and ultimately, the goal of directing the letter to the correct destination.

The following paper presents the description of the multifont word recognition system which, as part of the AL

Manuscript received November 24, 1977; revised August 10, 1977. This paper is an expanded version of a contribution to the 3rd International Joint Conference on Pattern Recognition, November 1976, Coronado, CA. The work was partly supported by the Federal German Postal Service (DBP) and the Federal Ministry of Research and Technology (BMFT).

The author is with AEG-TELEFUNKEN Research Institute, Ulm, Germany.

0018-9340/78/0800-0721$00.75

880-address-reader, has been developed by AEG-TELEFUNKEN for the Federal German Postal Service. A functional model of the complete address reading system has been thoroughly tested with live mail by the Post Office in the spring of 1975. The prototype machine is to be installed in Wiesbaden, Germany, in the fall of 1977.

The system is designed to handle 60 000 pieces of mail per hour. Two mechanical transports are combined with the recognition unit. Letters and postcards to be processed have a standardized format. The address can be positioned anywhere within the 60 mm reading zone beginning 15 mm from the bottom edge. Every kind of type font—typewriter, line printer, addressograph, bookprint—is admissible. Character height may vary from 1.6 to 6.0 mm. The segmentation system must be prepared to cope with every kind of occurring character spacing, including proportional spacing.

For outgoing mail, the bottom line—postcode, postal place name—is to be read, and for incoming mail the second to last line, containing street name and house number. Accordingly, the complete set of characters, small letters, capital letters, and numerals, must be recognized.

The elements of the recognition process are the individual characters. Therefore, the connected raster picture of the complete reading zone, scanned off the envelope by an integrated 512-point self-scanned photodiode array, must be broken up into pieces. These operations are executed by the picture processing and pattern preprocessing units. They operate on the input data of a 512 * 1024 black and white raster picture and produce a sequence of standardized 16 * 16 raster pictures, each containing a single character image, provided that no segmentation errors have occurred. The main features of the picture processing and pattern preprocessing units are as follows:

1) suppression of underscoring lines,
2) line-skew correction by shearing,
3) segmentation by a hierarchy of different segmentation procedures especially tailored to fixed and variable character spacing,
4) centering to the center of gravity, and
5) normalizing of varying stroke widths and sizes.

These operations are illustrated by Figs. 1 and 2. In Fig. 1, the content of the 512 * 1024 input image storage is shown as matrix raster plot. It contains one total address block and parts of envelope imprints. The two next to bottom lines are detected by the line-finding procedures and the corresponding raster images transmitted to separate line image storages of 64 picture element's height. Fig. 2 shows how the second to last line is processed further by the segmentation, centering and size normalization procedures. The resulting 16 * 16 black and white raster images are tagged with their origin, line, and character position, and are then passed on to the recognition procedures via a queueing buffer.

The subsequent multifont word recognition system can be partitioned into two major subsystems:

1) single character recognition (SCR) subsystem and
2) contextual postprocessing (CPP) subsystem.

The design philosophy applied here can be characterized by

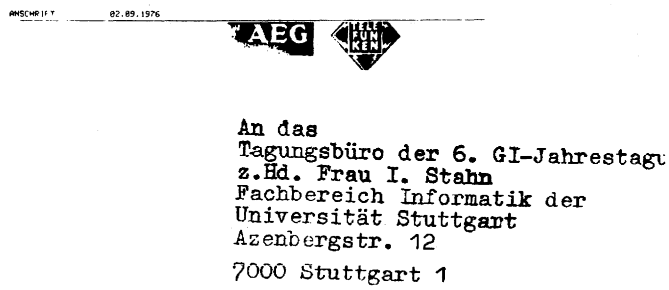


Fig. 1. Raster image of the complete reading zone. 512 * 1024 binary picture elements with 120 μm resolution in both vertical and horizontal directions.

the notion of a multiple-channel/multiple-choice system (MCMC-system). It represents a compromise solution lying between the compound approach [1]–[3], completely integrating the single character recognition process into the word recognition procedure, and the error detection and correction approach [4]–[6], which is not specifically designed for character recognition applications but can also be applied if the garbled text is produced by any kind of noisy transmission channel, not just by an imperfectly operating single character recognition system.

The main idea is to pass on as much information as possible from the single character recognition level to the word recognition level. The single character recognition subsystem, therefore, is provided with the capability to output a variable number of choices depending on the reliability of the recognition result. The number of choices is dynamically controlled to ensure that, on the one hand, the correct character meaning is among the choices with a sufficient high probability and on the other hand, to keep the mean number of choices at a minimum.

From another point of view this can be considered an application of the principle of least commitment. The single character recognition subsystem is the first stage of a hierarchically organized data processing structure consisting of different levels with different decision authority. Only in the case of sufficient plausibility is the single character recognition subsystem allowed to make a final decision—to present only one choice. Otherwise, the task of resolving the remaining ambiguities must be handed over to the next higher hierarchical level.

The multiple channel organization makes allowance for the fact that three types of character images can occur: capital letters, small letters, and numerals, and that they are not intermixed arbitrarily. Therefore, each raster image to be recognized is simultaneously presented to three independent recognition channels, each responsible for one of the three character types and specifically adapted to its corresponding subset of character classes. The selection of one of the three recognition system outcomes for further processing is accomplished on a word basis by a specific word genre recognition system which has the task of deciding among one of the four types of words: numeric words and alphabetic words in three different modes of writing.

Word genre decision is the first operation of the contextual postprocessing system. After having selected one of the three recognition system channel outcomes for each single

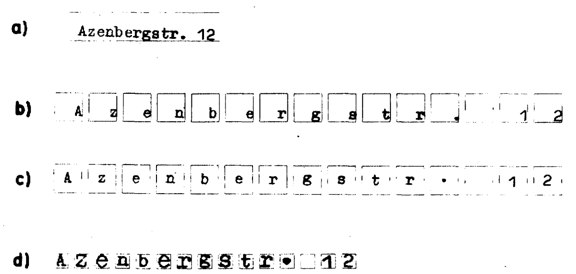


Fig. 2. Picture preprocessing operations applied to the second to last line of Fig. 1. (a) Raster image of the line after line finding and separating. (b) Raster images of the single characters after segmentation. (c) Raster images of the single characters after centering. (d) Raster images of the single characters after size normalization. In this form, the $16 * 16$ raster images are transmitted to the single character recognition subsystem.

character position of the word to be recognized, in general, a set of alternative words can be constructed by combining the multiple choice results for the single character positions. This set of alternative words with a high probability contains the correct word, which then is sifted out by means of a hash-coded table look-up procedure comparing every element of the set of alternative words with the postal directory. Special provisions are made to cope with segmentation errors which may occur during image preprocessing and may effect additional or missing characters.

The overall system design reflects many of the ideas which have been and still are currently discussed in the pattern recognition literature, especially in the area of postal address reading [7]–[10]. After all, the approach taken here is a unique solution combining some new ideas with some well-known efficient components for a successfully operating address reading system and therefore, seems to be worthy of a detailed discussion.

II. Single Character Recognition

Input datum to the SCR subsystem is the $16 * 16$ black and white raster picture of an isolated character image [see Fig. 1(d)]. It is transmitted simultaneously to the three channels of the SCR subsystem, each of which is responsible for one of the three types of character images: capital letters (CL), small letters (SL), and numerals (NU). The partitioning into three channels is a straightforward extension of the two channel organization suggested by Atrubin [7] and employed by the IBM Advanced Optical Character Reader [6]. The grouping of the characters into the three disjoint subsets is justified because, at least after the text has been completely read, a reliable decision can be made as to the types of the words read—word genre recognition. The character classes are grouped together not according to their similarity or recognizibility, but rather according to the contextual situation in which they appear. The aim is, for each decision to be made, to limit the number of alternative outcomes and thus, to increase the recognition reliability.

The three channels of the SCR subsystem are identically structured. Each of them is specifically adapted to its corresponding subset of character classes. It produces estimations as to the class membership of the raster picture presented. These are valid only if the hypothesis under which the SCR channel is operating is true.

A. Design Principles of the Single SCR Channel

For realizing the single SCR channel, the concept of the mean-square polynomial classifier has been adopted. Character recognition systems of this type have been applied for high performance document reading with remarkable success and are presently being used in numerous single and mixed font applications.

The approach [11] applied here for classifier design and classifier adaptation is closely related to those described by Yau [12] and Meisel [13].

Mathematically, the $16 * 16$ binary raster picture is described by a corresponding measurement vector

$$\boldsymbol{v} = (v_1 v_2 \cdots v_N) \tag{1}$$

with $N = 256$ binary components $v_n \in \{0, 1\}$. As there are K different classes to be discriminated, the actual class membership is indicated by a K-valued discrete integer variable $k = 1, 2, \cdots, K$, the class membership indicator. This scalar integer variable k is mapped one-to-one into a discrete K-dimensional vector variable $\boldsymbol{y}$, called the objective vector. The K discrete values, which the objective vector can assume, are the K columns of the K-dimensional identity matrix

$$\boldsymbol{I} = [\boldsymbol{y}_1 \boldsymbol{y}_2 \cdots \boldsymbol{y}_K]. \tag{2}$$

Whereas the measurement vector $\boldsymbol{v}$ constitutes the collection of measurement data on which the recognition is to be based, the objective vector $\boldsymbol{y}$, principally, is unknown to the recognition system. The task of the recognition system, essentially, therefore, is to reconstruct $\boldsymbol{y}$ approximately from the measurement data $\boldsymbol{v}$. This is brought about by means of a vector discriminant function $\boldsymbol{d}(\boldsymbol{v})$ and the optimization criterion

$$S^2 = E\{|\boldsymbol{y} - \boldsymbol{d}(\boldsymbol{v})|^2\} \stackrel{!}{=} \text{minimum}, \tag{3}$$

leading to the well-known solution

$$\boldsymbol{d}(\boldsymbol{v}) = E\{\boldsymbol{y} \mid \boldsymbol{v}\}. \tag{4}$$

The optimum mean-square discriminant function $\boldsymbol{d}(\boldsymbol{v})$ thus turns out to be the so-called regression function [14] $E\{\boldsymbol{y} \mid \boldsymbol{v}\}$. Due to the special definitions applied here for the objective vector $\boldsymbol{y}$ [see (2)] the components d_k of the regression function $\boldsymbol{d}(\boldsymbol{v})$ [see (4)] become the *a posteriori* probabilities

$$d_k(\boldsymbol{v}) = \text{prob}\,[k \mid \boldsymbol{v}]. \tag{5}$$

These are necessary for solving the recognition problem in the decision theoretic minimum risk sense.

The least mean-square approach results in a practically feasible solution when combined with the concept of the polynomial classifier. In this case the discriminant function $\boldsymbol{d}(\boldsymbol{v})$ no longer is an arbitrary vector function of $\boldsymbol{v}$ as in (3) but is designed to be a polynomial in $\boldsymbol{v}$. The components $d_k(\boldsymbol{v})$ of the discriminant function $\boldsymbol{d}(\boldsymbol{v})$ then become mean-square estimations of the *a posteriori* probabilities

$$d_k(\boldsymbol{v}) \approx \text{prob}\,[k \mid \boldsymbol{v}]. \tag{6}$$

By means of polynomial vector $\boldsymbol{x}(\boldsymbol{v})$, containing a constant term $x_0 = 1$ and a predefined number M of pattern measurements v_n and products $v_n \cdot v_m$ of them,

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ x_M \end{bmatrix} = \begin{bmatrix} 1 \\ v_1 \\ \vdots \\ v_n v_m \end{bmatrix} = x(v) \quad (7)$$

and describing the structure of the polynomial, the polynomial discriminant function $d(v)$ can be expressed in a compact form by

$$d = A^T x(v), \quad (8)$$

whereby the $(M + 1) * K$-matrix A contains all of the polynomial coefficients.

For the application considered here, the discriminant polynomials are designed as second degree polynomials (quadratic discriminant functions). The large number $N = 256$ of pattern measurements in this case, prohibits the construction of complete quadratic polynomials. Therefore, additional constraints are necessary for limiting the discriminant polynomial to a manageable length. The constraints used here are heuristically inspired and concentrate the products (i.e., quadratic terms of the discriminant polynomial) in the center of the $16 * 16$ input field. Fig. 3 gives a graphical illustration of the polynomial structure applied. The background is formed by the $16 * 16$ raster field. Picture elements v_n, which are directly used as linear terms of the polynomial vector x (7), are marked by symbols $\boxdot$, whereas quadratic terms $v_n v_m$ of the polynomial are indicated by line segments connecting the corresponding picture elements. All three channels of the recognition system polynomials of similar structure are applied. The polynomials for the CL- and the SL-channels are identical [see Fig. 3(a)]. The polynomial for the NU-channel takes into account that numerals normally are less wide than high, the two leftmost and the two rightmost columns of the raster field, therefore, are omitted [see Fig. 3(b)]. The polynomial length is limited to $M = 1217$ terms for the alphabetic channels and to $M = 1215$ for the numeric channel.

Since the polynomial structure $x(v)$ is predefined by the system designer, all of the system adaptivity is concentrated in the coefficient matrix A [see (8)]. Essentially, system adaptation is accomplished by proper adjustment of all the coefficients contained in A. Inserting (8) into the optimization criterion (3) and carrying out the minimization,

$$\begin{aligned} S^2 &= E\{|y - d|^2\} \\ &= E\{|y - A^T x|^2\} \stackrel{!}{=} \text{minimum} \end{aligned} \quad (9)$$

leads to the well-known matrix equation

$$E\{xx^T\} \cdot A = E\{xy^T\}. \quad (10)$$

Constitutive parts of this matrix equation are the two moment matrices $E\{xx^T\}$ and $E\{xy^T\}$, which are statistically estimated based on a given labeled learning set by simple arithmetic averaging.

The learning set is extracted from live mail. Labeling is done partially unsupervised using a preliminary recognition system of the same kind. Only those patterns recognized with insufficient reliability are labeled manually using computer displays for presenting the raster pictures to human operators. The discrimination between reliable and unreliable recognition is based on the squared Euclidean norm of the error vector r (14), between discriminant vector d, and the estimated objective vector $\hat{y}$. By proper threshold adjustment, mislabelings among the automatically labeled patterns can be totally avoided. A certain amount of patterns mislabeled manually is tolerated by the adaptation procedure.

There are numerous methods of solving the particular least mean-square problem [15] of (9) and (10). The approach we use is the direct evaluation of (10) carried out in a sequence of two steps. The first step is the computation of the two sample moment matrices $E\{xx^T\}$ and $E\{xy^T\}$ based on the learning set. For this purpose a $(M + 1 + K)$-dimensional vector variable z is formed from the polynomial vector x and the objective vector y

$$z = \begin{pmatrix} x \\ y \end{pmatrix}. \quad (11)$$

Given the measurement vector v and the corresponding class membership indicator k, each element of the learning set is mapped into a corresponding vector z by applying (7) and (11), thus forming a learning set of vectors z. By calculating the sample moment matrix

$$M = E\{zz^T\} = \begin{pmatrix} E\{xx^T\} & E\{xy^T\} \\ E\{yx^T\} & E\{yy^T\} \end{pmatrix} \quad (12)$$

of this learning set $\{z\}$, all of the statistical data necessary for further processing are gained simultaneously.

The second step is the computation of a generalized inverse solution A from (10). The computational procedure directly operates on the matrix M (12). It is oriented towards the stepwise regression procedure and can be interrupted at every stage of the computational process, each time resulting in a valid intermediate solution.

Classifier adaptation can be repeated if new data are collected. When including the solution of (10), a kind of recursive stochastic approximation procedure is established by weighted mixing of the moment matrices M of the original learning set and of a subset of newly collected problem characters

$$M \Leftarrow (1 - \gamma)M + \gamma M \text{ problem set}. \quad (13)$$

This leads to an iterative classifier refinement procedure, whereby the set of problem characters is formed of those patterns which turn out to be difficult to recognize. For separating the problem set, again the Euclidean norm of the error vector r (14) is applied.

The discriminant polynomial $d(v)$ is optimized with respect to (9). Therefore, the error vector

$$\Delta d = y - d$$

has a tendency to assume small magnitudes in the case of easily recognizable characters and accordingly, can be used for indicating the reliability of the various individual deci-

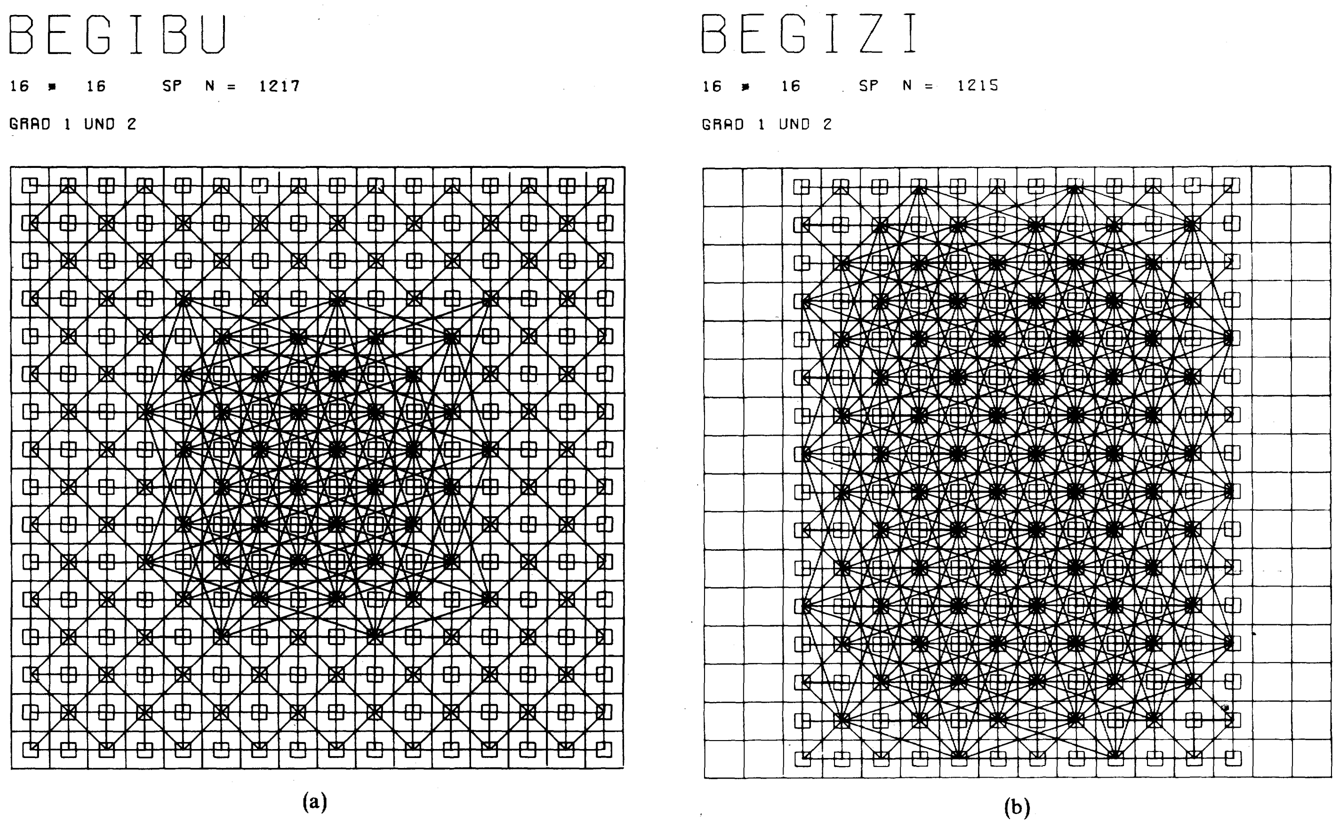


Fig. 3. Display of the polynomial structure mapped onto the raster field. (a) As applied for the alphabetic channels: CL-channel, SL-channel. (b) As applied for the numeric channel: NU channel.

sions. Since the objective vector $\boldsymbol{y}$, principally, is unknown to the recognition device, it must be replaced by the nearest discrete value $\hat{\boldsymbol{y}}$ of $\boldsymbol{y}$ belonging to that class i for which the corresponding component d_i of $\boldsymbol{d}$ is maximal. Thus, we define a measurable error vector

$$\boldsymbol{r} = \hat{\boldsymbol{y}} - \boldsymbol{d}. \tag{14}$$

The squared Euclidean norm of it is used as a measure of reliability,

$$r^2 = |\boldsymbol{r}|^2 = \boldsymbol{r}^T\boldsymbol{r}. \tag{15}$$

According to (6), the K components d_k of the discriminant vector $\boldsymbol{d}$, produced by the mapping $\boldsymbol{v} \mapsto \boldsymbol{d}$ [see (8)] are mean-square approximations of the *a posteriori* probabilities $p(k \mid \boldsymbol{v})$ under the constraint of a predetermined structure $\boldsymbol{x}(\boldsymbol{v})$ of the discriminant polynomial (8). A conventional single character recognition system, therefore, would have to look for the maximum component d_i of $\boldsymbol{d}$ and to output the corresponding class membership indicator i as its only choice.

As indicated in Section I for interfacing the SCR subsystem with the subsequent contextual postprocessing system, here the multiple choice approach is chosen. The system is equipped with the possibility of offering up to three different choices to the subsequent procedures. The different choices are gained by rank-ordering the components d_k of the discriminant vector $\boldsymbol{d}$ according to their numerical value. The indices i, j, k of the three largest components $d_i > d_j > d_k$ of $\boldsymbol{d}$ are the best three choices the SCR channel can offer. They are accompanied with one reliability measure r^2 [see (15)] which corresponds to the first choice i.

The number of choices passed on to further processing is dynamically controlled. The goal, therefore, is to minimize the number of choices while, at the same time, to ensure that the set of alternative choices actually contains the correct meaning of the character to be recognized. For this purpose, three different thresholds $S1$, $S2$, $S3$ are applied to the reliability measure r^2 [see (15)]. The idea is illustrated by Fig. 4.

For $r^2 \leq S1$, the set of character alternatives contains only one choice. With increasing values of r^2, the number of choices is increased up to three. For $r^2 > S3$, eventually the SCR channel refuses to offer a definite choice, which is logically equivalent to offering all of the K admissible class indices.

The problem of dynamically controlling the number of choices thus reduces to the problem of properly adjusting the threshold values $S1, S2, S3$ (Fig. 4). By means of suitable empirical data, a mathematical optimization problem can be established: minimization of the mean number, a, of alternative choices with the constraint of a predefined small probability of missing the correct class membership index among the set of alternative choices.

The empirical data necessary for solving this optimization problem are contained within the classifier performance curves shown in Fig. 5, which are gained by running the recognition system after adaptation using live mail test sets. The function $\rho(r^2)$ indicates the fraction of all input patterns

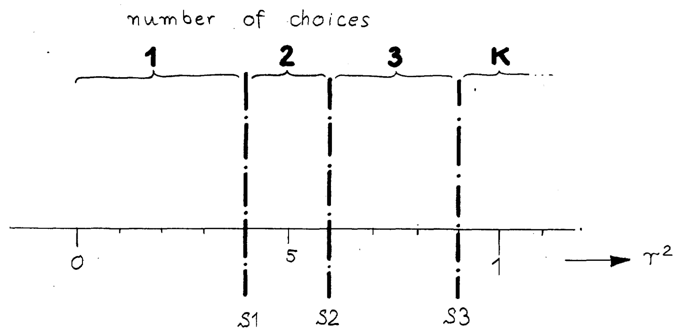


Fig. 4. Definition of reliability ranges for dynamically controlling the number of choices.

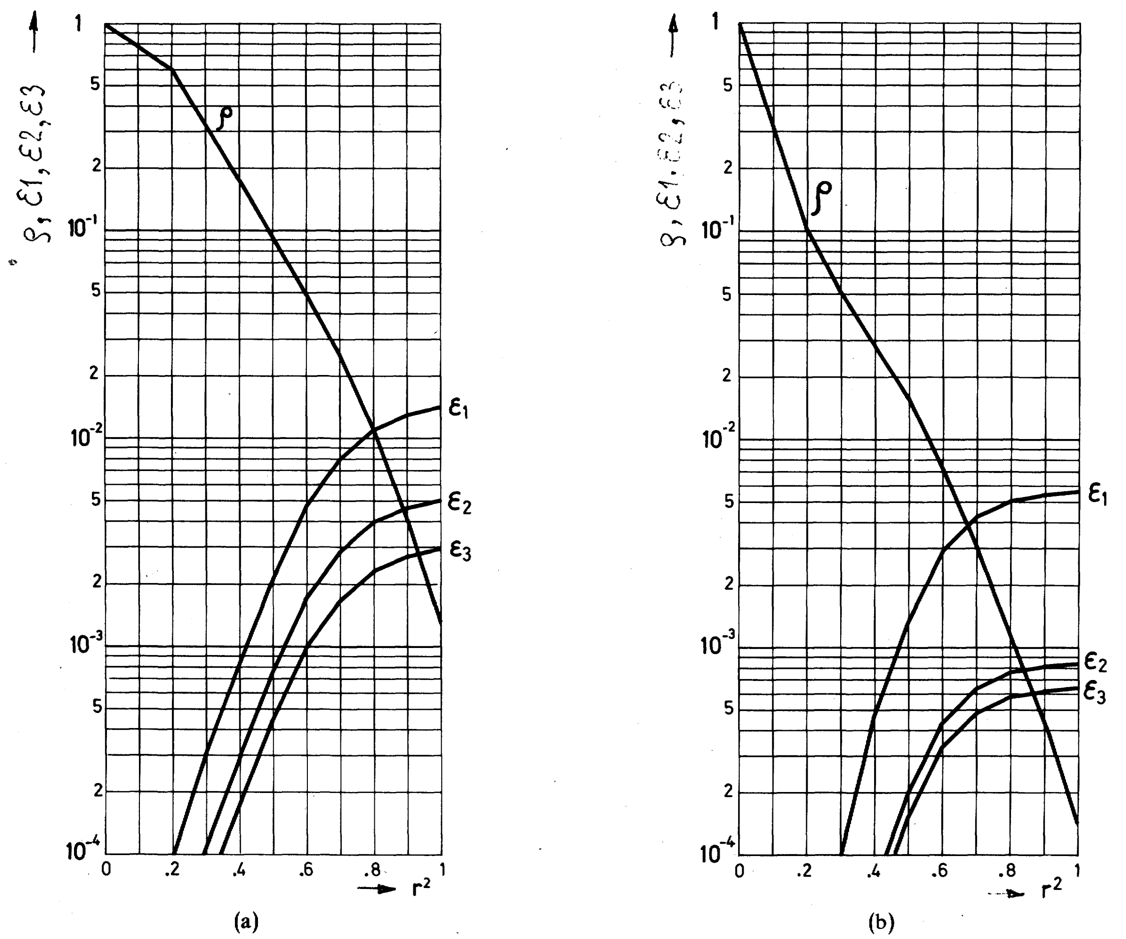


Fig. 5. Classifier performance curves. ρ = reject rate for reject threshold r^2. εI = remaining error rate for I choices. (a) For the alphabetic channel. Live mail composition: 30 percent capital letters, 70 percent small letters. (b) For the numeric channel.

resulting in a reliability measure r^2 exceeding the abcissa value. In the case of a conventional one choice single character recognition, this system fraction $\rho(r^2)$ would be the *reject rate* when operating the recognition system with a *reject threshold* r^2. The three remaining functions $\varepsilon 1(r^2)$, $\varepsilon 2(r^2)$, and $\varepsilon 3(r^2)$ are specifically defined *error rate* curves. Thereby, $\varepsilon I(r^2)$ indicates the error rate of an I-choice system depending on the reject threshold r^2, that is the probability of missing the correct meaning among the I choices, if a reject threshold r^2 is applied.

These data are sufficient for calculating for every arbitrary combination of threshold values $S1$, $S2$, and $S3$ (see Fig. 4) the mean number, a, of alternative choices as well as the resulting error rate, ε, that is the probability of missing the correct character meaning among the set of alternative choices [16]. By minimizing the mean number, a, of alternative choices for given error rate ε, the diagram $a(\varepsilon)$, Fig. 6, can be computed. To every point on this curve, a different combination of threshold values $S1$, $S2$, and $S3$ belongs. By applying the corresponding threshold values, every point on the $a(\varepsilon)$ curve of Fig. 6 can be chosen as operating point of the SCR subsystem.

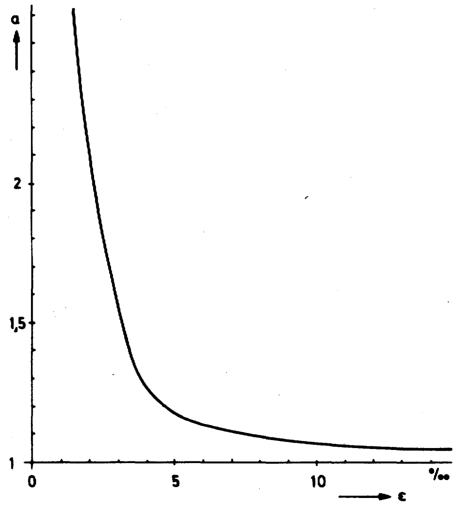


Fig. 6. Mean number a of choices as a function of the residual error rate ε. Valid for live mail composition of 30 percent capital and 70 percent small letters.

TABLE I
EMPIRICAL PROBABILITIES FOR THE iTH CHOICE OUT OF I OF BEING CORRECT

i \ I	1	2	3
1	0.9980	0.9761	0.8403
2		0.0154	0.1027
3			0.0235

The alphabetic channels of the postal address-reader discussed here are tuned to $\varepsilon = 4 \cdot 10^{-3}$ which leads to a mean number $a = 1.25$ of character alternatives. For this operating point the probability can be computed [16], that the ith choice out of I is correct, Table I. These probabilities play an important role for the dynamic control of the word recognition procedure.

B. The Complete Single Character Recognition Subsystem

The three channels CL, SL, and NU are combined, thus constituting the complete SCR subsystem (see Fig. 7). The discriminant polynomials actually used have different lengths for the alphabetic and numeric channels, 1024 terms for both the CL and the SL channel and 512 terms for the NU-channel. The alphabetic channels are designed to discriminate between $K = 32$ classes each. The numeric channel is designed for $K = 16$ different classes. Thus, a total of approximately 80 000 polynomial coefficients must be stored and processed during recognition. Proper quantization leads to the modest storage requirements of only 5 bits/coefficient without loss of classifier performance. The clear-cut mathematical structure of the polynomial discriminant function approach is particularly well-suited for microprocessor implementation. A specially prepared microprocessor system is used for realizing the single character recognition subsystem. The operating speed is 1000 chars./s. The complete subsystem is housed in one 19-in wide chassis with 27 printed circuit boards of $7 * 10$ inch squared (double Europe format).

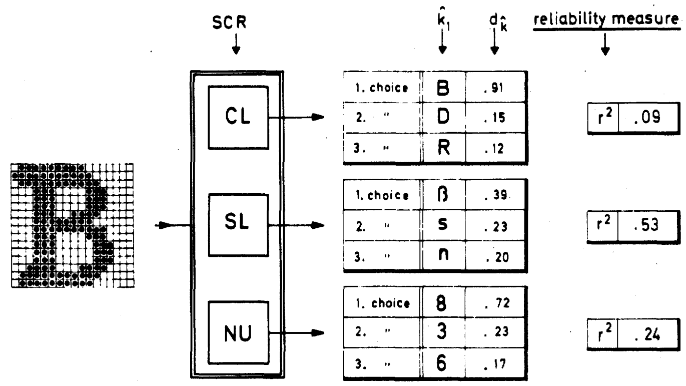


Fig. 7. Three-channel three-choice SCR subsystem operating on a sample raster picture.

The subsystem output consists of three identically arranged data messages containing the three first choices k_1, k_2, k_3 and the reliability measure r^2 for each of the three channels

$$[k_1, k_2, k_3, r^2]_i \qquad i = \text{CL, SL, NU}. \tag{16}$$

Fig. 7 gives a practical illustration with realistic values. Depending on the actual value which the three reliability measurements r^2 have, a variable number of different choices is to be processed further. Thus, a stochastically varying number of alternative choices is transmitted to the contextual postprocessing system.

III. CONTEXTUAL POSTPROCESSING

There are three independent problems to be solved by the contectual postprocessing (CPP) subsystem:

1) to tie together the character recognition results belonging to one word of text input (subfield formation),

2) to select the correct channel output from the three-channel SCR (word genre decision), and

3) to recognize the word actually picked up from the envelope based on the SCR selected output data (word recognition).

A. Subfield Formation

Principally, there are two different criteria for word border detection, blank space detection, and detection of special symbols such as ( ).-/. Whereas these special symbols are recognized by the SCR subsystem, the blank space detection must be carried out by the segmentation module which has access to geometrical information concerning the gap width between adjacent characters as they appear in the connected raster picture of the scanned line of text (see Fig. 1). To cope with spaced typewriter material and the different varieties of bookprint, a clustering procedure is applied to the set of gap width measurements collected from the line being tested.

For this purpose the sequence of gap width measurements w_i, which are produced during segmentation by the segmentation module, is analyzed in their natural order of appearance. The first available width measurement is taken as the center c of a first cluster. If the relative distance $(w - c)/c$

between cluster center and a new width measurement w exceeds a certain threshold, a new cluster is established with the cluster center w. Otherwise, the width measurement w is associated with that cluster nearest in the minimum relative distance sense. Then the cluster center is updated to be the average of all the width measurements w_i already members of the cluster.

In this way, gap widths which are significantly larger than the majority of gap widths can be detected. We call those gaps which actually divide the line picture into meaningful words "syntactic gaps." A syntactic gap is found if a wider gap is observed in a sequence of small gaps or even if a sequence of wide gaps follows a sequence of smaller ones. Syntactic gaps define word borders as do the special symbols described above. The string of segmented and preprocessed raster pictures surrounded by word border symbols or syntactic gaps are considered to belong to one word of text.

B. Word Genre Decision

There are two types of words to be discriminated:

1) numeric words as postal *Zip* codes, postal district numbers, house numbers, post box numbers, and

2) alphabetic words as city names, street names, etc.

The latter may appear in three different modes of writing:

1) capital letters only, as common for high speed printing (capital mode),

2) first letter capital, followed by small letters, as common for typewriting (mixed mode), and

3) small letters only (small mode).

Both decision systems conceptually are organized in the decision theoretic sense. This will be explained using the example of the number/name decision problem. The measurements on which the decision is based are the $J * 3$ reliability measures belonging to the J character positions of the word just considered which are produced by the three SCR channels. They form a sequence of J three-dimensional measurement vectors.

$$\boldsymbol{m}_1, \boldsymbol{m}_2 \cdots \boldsymbol{m}_J. \tag{17}$$

This approach has some resemblance to the BOND procedure described by Rosenbaum [6]. The main difference is that here the reliability measures r^2_{CL}, r^2_{SL}, and r^2_{NU} are used instead of the first choice decisions. The subsequent processing is quite similar to Rosenbaum's.

For applying Bayes' decision rule, the *a priori* probabilities p_{AL}, p_{NU}, and the conditional probabilities $p(\boldsymbol{m}_1, \boldsymbol{m}_2 \cdots \boldsymbol{m}_J \mid \mathrm{AL})$, $p(\boldsymbol{m}_1, \boldsymbol{m}_2 \cdots \boldsymbol{m}_{\mathrm{U}} \mid \mathrm{NU})$ for two classes AL (alphabetic words), and NU (numeric words) have to be provided. The evaluation of the conditional probabilities is simplified by the assumption of statistical independence between different character positions. This leads to

$$\begin{aligned} p(\boldsymbol{m}_1, \boldsymbol{m}_2 \cdots \boldsymbol{m}_J \mid \mathrm{AL}) &= \prod_{j=1}^{J} p(\boldsymbol{m}_j \mid \mathrm{AL}), \\ p(\boldsymbol{m}_1, \boldsymbol{m}_2 \cdots \boldsymbol{m}_J \mid \mathrm{NU}) &= \prod_{j=1}^{J} p(\boldsymbol{m}_j \mid \mathrm{NU}). \end{aligned} \tag{18}$$

The procedure results in comparing the log likelihood ratio

$$L = \sum_{j=1}^{J} \ln \frac{p(\boldsymbol{m}_j \mid \mathrm{AL})}{p(\boldsymbol{m}_j \mid \mathrm{NU})} \tag{19}$$

with a threshold value, which is given by the *a priori* probabilities of names and numbers. The single position log likelihood function

$$l(\boldsymbol{m}) = \ln \frac{p(\boldsymbol{m} \mid \mathrm{AL})}{p(\boldsymbol{m} \mid \mathrm{NU})} \tag{20}$$

necessary for calculating L according to (19) is empirically defined by statistically analyzing the behavior of the SCR subsystem when confronted with live mail.

The log likelihood function $l(\boldsymbol{m})$ here is a real-valued function of a three-dimensional vector variable $\boldsymbol{m}$ with nonbinary components. The problem of storing this function $l(\boldsymbol{m})$ can be solved in a number of different ways. One of them is the table look-up approach with suitable quantized input variables. Another approach which proved to be economically superior is to use a polynomial approximation for $l(\boldsymbol{m})$. Therefore, for both the number/name discrimination and the discrimination between small and capital letters, second degree polynomial approximations are applied.

C. Word Recognition

At this stage of information processing within our multiple-channel/multiple-choice system, the ambiguities generated by the multiple channel approach are resolved. The ambiguities caused by the multiple choice approach, however, are still existing. Namely, the set of data transmitted from the word genre decision system to the word recognition system consists of a string of J sets of character alternatives $\{\hat{k}\}$, where J is the wordlength of the word presented to the recognition system:

$$\{\hat{k}\}_1 \{\hat{k}\}_2 \cdots \{\hat{k}\}_J. \tag{21}$$

This string of positional character alternatives can be transformed into an equivalent set $\hat{\Omega}$ of alternative words

$$\hat{\Omega} = \{\omega_1, \omega_2 \cdots\} \tag{22}$$

by combinatorically arranging the character alternatives to word alternatives. The data contained in the string of positional character alternatives [see (21)], or in another kind of presentation, in the set $\hat{\Omega}$ of alternative words constitute the complete set of measurements available for the special recognition system which has to decide on which word most likely had been picked up from the piece of mail.

Conceptually, we are faced here again with a statistical recognition problem. The measurement data are the string of character alternatives or the equivalent set $\hat{\Omega}$. The classes are the legitimate words contained in the postal directory. The number of classes is the total number of directory entries, for the case considered here, approximately 16 000. To decide which of the legitimate words most likely would have been the cause of those findings, we have made here, in form of the set of alternative words, contained in $\hat{\Omega}$ [see (22)], a best match procedure is applied, matching the measurement data against the postal directory.

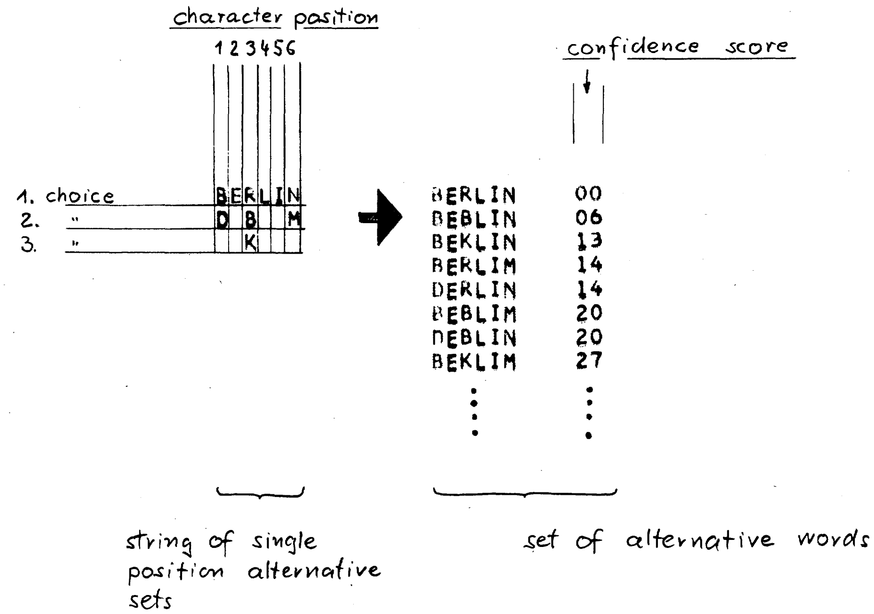


Fig. 8. Implicit and explicit representation of the set $\hat{\Omega}$ of alternative words. On the right, the elements $\hat{\omega}_i$ are ordered in decreasing reliability by the confidence score CM.

The basic idea underlying this best match procedure is that the correct word ω should be contained in the set $\hat{\Omega}$ of alternative words with a high probability and that, at the same time, the competing candidate words, also contained in $\hat{\Omega}$, would be almost unlikely to become legitimate words again. This leads to a word match of all the elements $\hat{\omega}$ contained in $\hat{\Omega}$ against all dictionary entries contained in Ω. In other words, the two sets $\hat{\Omega}$ and Ω have to be intersected.

There are some difficulties preventing this simple idea from being applied without modifications. The main obstacles are the occurrence of occasional segmentation and recognition errors, and the large volume of the postal directory prohibiting any kind of exhaustive matching procedures.

To keep the computing times limited, a dynamic system organization is applied. Only those operations which are necessary in a given situation shall actually be executed. The first measure to limit the requirements in computing time is to rank-order the set $\hat{\Omega}$ of alternative words $\hat{\omega}$ according to their probabilities of being correct.

This is accomplished by using the conditional probabilities of Table I which indicate, for known position of the single character within the set of character alternatives, the probability p_j for that specific character of being correct. With the assumption of positional independence, the probability for the complete word of being correct can be computed

$$\text{prob}\,(\hat{\omega} = \omega) = \prod_{j=1}^{J} p_j, \tag{23}$$

where J indicates the wordlength. Taking the logarithms of the p_j, the multiplication can be transformed into a summation. This way, an additive figure of confidence CM is achieved which can be used for rank-ordering the set $\hat{\Omega}$ of alternative words (Fig. 8).

The dictionary look-up is started with the most promising element $\hat{\omega}_1$ of $\hat{\Omega}$. The matching procedure is terminated if a matching entry is found. When following this approach, the speed of dictionary look-up becomes a crucial point of system design. For fast dictionary access the hash-code technique is applied. Since hash-code dictionary look-up, principally, is applicable only if no segmentation and recognition errors have occurred, the hash-code procedure must be modified essentially. The two main extensions are

1) twofold hash-code access via the left- and the right-half of the word,

2) application of a hierarchy of different distance matches.

The approach is based on the fact that the single character recognition subsystem, by suitably tuning the threshold values for the multiple choice control system, is operated with a low residual error rate ε (Fig. 6 and Table I). Consequently, there is a high probability that at least one half of the word contains no recognition error at all, whereby recognition error means that the correct character meaning is not among the set of alternative choices. Therefore, if no segmentation error has occurred, it must be possible to find the matching entry of the dictionary by sequentially using the left- and second-half of the word for hash-code access.

Since the segmentation error rate as well is low, these considerations apply for the case of occasional segmentation errors, too. With a sufficient high probability, one half of the word is error free and can be used for hash-code dictionary access.

Words containing no segmentation errors at all are much more likely than those with segmentation errors. Therefore, the first stage of the word recognition procedure operates with the assumption of no segmentation error. Only if this assumption proves to be wrong in the sense that no matching entry could be found, the set $\hat{\Omega}$ of alternative words is

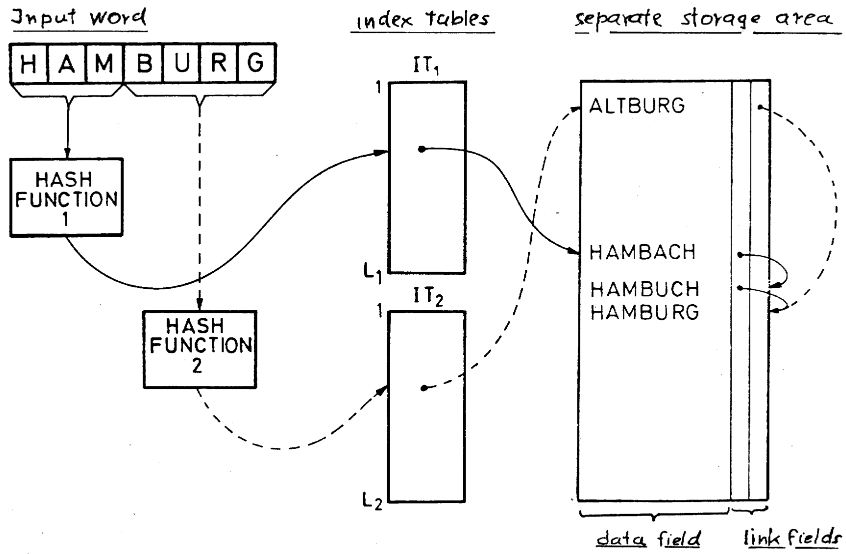


Fig. 9. Organization of the twofold hash-code dictionary access.

transmitted to specially tailored matching procedures which are designed to cope with additional or missing characters.

In order to explain the principles applied, the hash-code dictionary access in the first place shall be discussed for the segmentation error-free case.

To enable the twofold hash-code access to be executed, two index tables are provided, one for access using the left-half and one for access using the right-half of the word. From the character string of the word half used, in combination with the wordlength J, by means of the division method [17], the so-called hashing index is computed pointing to a certain cell of the corresponding index table. This cell is empty if the dictionary contains no entry of wordlength J and the corresponding combination of characters in the word half used; otherwise, an address is found pointing to the subset of those dictionary entries having the same wordlength and identical sequence of characters in the word half used. The dictionary is stored here in a separate storage area, with each entry consisting of a data field, containing the legitimate word and accompanying additional postal information, and two linkfields, linking together all those dictionary entries which, with respect to the hash-code applied, belong to the same dictionary subset (see Fig. 9). Due to statistical properties of German postal place names, the hash-code access via the left-half of the word more directly leads to the wanted entry than the hash-code access via the right-half of the word. The mean number of accesses is 3.4 for the left-half of the word and 14.7 for the right-half. If possible, therefore, the left-half access is preferred.

As indicated above, the word match procedure is organized sequentially. The first step makes use of the assumption of segmentation error-free recognition and the additional, and actually most likely assumption, that the correct word ω is contained within the set $\hat{\Omega}$ of alternative words [see (22)]. This first step of the matching procedure, therefore, applies an identity comparison between the input words, taken from the set $\hat{\Omega}$, and those dictionary entries yielded by the hash-code access. Only in the case of failure are three different distance matches activated, which tolerate nonmatching single characters up to a certain threshold depending on the wordlength J of the input word. The three variations of this distance match procedure arise from changing the wordlength J by -1, 0, and $+1$.

In the case of unaltered wordlength $J \Leftarrow J + 0$ errors in that half of the word, not used for hash-code access, are tolerated, which may be caused by recognition errors or by the so-called crowding segmentation error [6], [16].

In the case of an increased wordlength $J \Leftarrow J + 1$, the assumption is made that one splitting segmentation error [6], [16] has occurred. The input word is compared twice with the dictionary entries, the first time the two strings of characters are synchronized on the left-hand side and the second time on the right-hand side (Fig. 10). Again, a certain number of nonmatching character positions is tolerated, depending on the wordlength J.

The case of a decreased wordlength $J \Leftarrow J - 1$ is prepared for catenation segmentation errors [6], [16]. The procedure is a straightforward inversion of that just described and needs no further explanation.

Since the whole word recognition system is organized dynamically, the processing times strongly depend on the distortions that have been acting on the image of the word to be recognized (bad printing quality, insufficient printing contrast) and the susceptibility of printing to segmentation errors. A detailed description of the contextual postprocessing system accompanied by additional experimental data is given in a paper by Doster [16].

IV. Conclusions

The word recognition system described here has proven to be remarkably tolerant of segmentation and classification errors. Overall performance measurements—including segmentation, pattern preprocessing, SCR, and CPP, but excluding the preceding picture processing operations carried out on live mail—have resulted in a segmentation error rate of approximately 1.3 percent and a forced recognition error rate of approximately 1.4 percent, both with

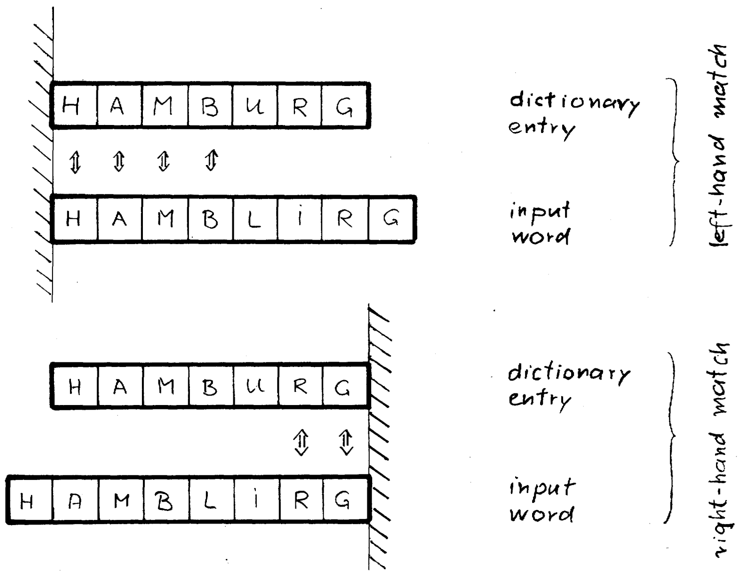


Fig. 10. Word matching in the case of a splitting segmentation error.

respect to single alphabetic characters. Operating with a dictionary of about 16000 entries, the word recognition rate was 98 percent with 1 percent word-error rate and 1 percent word-reject rate. In the case of the postal address reader, the postcode serves as additional error-detecting information. This allows the 1 percent error-rate almost completely to be converted into a reject rate of the same value.

The advantages of the system design presented here seem to be that, by introducing the multiple choice approach, a maximum of relevant information is transmitted from the single character recognition level to the word recognition level. By dynamically controlling the number of alternative decisions, care is also taken that no excess information is passed on. The second advantage is the employment of hash-code techniques, known for their fast access to large data bases, in the contextual postprocessing subsystem.

Although there is some resemblance in overall system design philosophy with other postal address reading machines as described in the literature [6], [8], it is particularly the multiple choice approach, which discriminates our design from those cited. Here the capability of a statistical single character recognition system to weigh its own decisions according to their reliability is effectively utilized. Therefore, no sophisticated treatment of unidentified (reject) characters and frequent character perturbations is necessary. Instead, the single character recognition subsystem presents a selected set of alternative choices which actually reflects the ambiguities of the present situation. The probability that the set of alternative choices contains the correct meaning of the character can be measured by observing the system during operation. This probability is used to control the subsequent operations.

The employment of the modified hash-code procedures for fast dictionary access constitutes another significant difference to the address reading machines cited. The approach related most to our's is that of Rosenbaum [6] in as far as there, too, an attempt is made to find access to dictionary subsets which should be as small as possible and at the same time should contain the correct entry with a high probability. The twofold hash-code access applied here seems to be a more straightforward technique than Rosenbaum's vector-fetch methodology.

Within the postal address reading machine described here, both the single character recognition subsystem as well as the contextual postprocessing subsystem are open to variation with respect to print quality, distribution of type fonts, and dictionary content. Classifier redesign can easily be realized by running some of the classifier adaptation algorithms again, implemented outside the address reader on an ordinary large scale computer. Dictionary redesign requires recalculation of the hash-code index tables and can be carried out by the process computer that is part of the address reader itself.

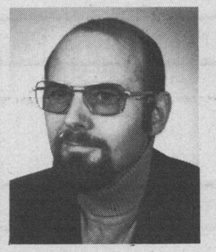
Acknowledgment

The author gratefully acknowledges the cooperation of his colleagues Drs. W. Doster, R. Ott, and K. Poike, who have been involved in the address reader system design.

References

[1] K. Abend, "Compound decision procedures for unknown distributions and for dependent states of nature," in *Pattern Recognition*, L. N. Kanal, Ed. Washington, DC: Thompson, 1968, pp. 207–249.
[2] J. Raviv, "Decision making in Markov chains applied to the problem of pattern recognition," *IEEE Trans. Inform. Theory*, vol. IT-3, pp. 536–551, Oct. 1967.
[3] M. Shimura, "Recognition machines with parametric and nonparametric learning methods using contextual information," *Pattern Recognition*, vol. 5, pp. 149–168, 1973.
[4] E. M. Riseman and A. R. Hanson, "A contextual postprocessing system for error correcting using binary *n*-grams," *IEEE Trans. Comput.*, vol. C-23, pp. 480–493, May 1974.
[5] A. R. Hanson, E. M. Riseman, and E. Fisher, "Context in word recognition," *Pattern Recognition*, vol. 8, pp. 35–45, 1976.
[6] W. S. Rosenbaum and J. J. Hilliard, "Multifont OCR postprocessing system," *IBM J. Res. Develop.*, vol. 19, pp. 398–421, July 1975.
[7] A. J. Atrubin and A. Hamburgen, "Two channel character recognition of single-channel input data," in *Proc. 1972 Int. Conf. on Cybernetics and Society*, Washington, pp. 527–530.
[8] P. M. Hahn and N. C. Randall, "A best-match file search procedure

for postal directories," in *Proc. 1972 Int. Conf. on Cybernetics and Society*, Washington, pp. 512–518.
[9] P. J. Rasch, "Contextual recognition of mail piece address information," in *Proc. 1972 Int. Conf. on Cybernetics and Society*, Washington, pp. 615–621.
[10] A. J. Atrubin, R. S. Fursa, and D. L. Malaby, "Mail piece address-simulation," in *Proc. 1972 Int. Conf. on Cybernetics and Society*, Washington, pp. 522–526.
[11] J. Schürmann, "Über eine vektorielle Unterscheidungsfunktion zur Unterscheidung von k Klassen," *TELEFUNKEN-Zeitung 40*, pp. 15–25, 1967.
[12] S. S. Yau and J. M. Garnett, "Least-mean-square approach to pattern classification," in *Frontiers of Pattern Recognition*, S. Watanabe, Ed. New York: Academic, 1972, pp. 575–578.
[13] W. S. Meisel, *Computer Oriented Approaches to Pattern Recognition.* New York: Academic, 1972.
[14] C. R. Rao, *Linear Statistical Inference and its Applications.* New York: Wiley, 1965.
[15] C. L. Lanson and R. J. Hanson, *Solving Least Square Problems.* Englewood Cliffs, NJ: Prentice-Hall, 1974.
[16] W. Doster, "Contextual postprocessing system for cooperation with a multiple-choice character-recognition system," *IEEE Trans. Comput.*, vol. C-26, pp. 1090–1101, Nov. 1977.
[17] W. D. Maurer and T. G. Lewis, "Hash table methods," *Comput. Surveys*, vol. 7, no. 1, pp. 5–19, 1975.



Jürgen Schürmann was born in Frankfurt/Oder, Germany, in 1934. He received the Dipl.-Ing. degree in 1960 and the Dr.-Ing. degree in 1968 from the Technical University of Berlin, Berlin, Germany.

In 1961 he joined AEG-TELEFUNKEN. Since 1963 he has been with the AEG-TELEFUNKEN Research Institute, Ulm, Germany, where he is engaged in digital signal processing with special emphasis on pattern recognition and radar signal processing. Since 1974 he has been teaching Pattern Recognition at the Technical University of Darmstadt, Germany.

System Functions for an Optical/Digital Processor

DAVID CASASENT, SENIOR MEMBER, IEEE, AND PETER D. RAPP, MEMBER, IEEE

Abstract—**The software and digital sections of a hybrid optical/digital processor are described with emphasis on the required operations, their realization, and the hierarchy of software and hardware used. Several applications of this interface are also included.**

Index Terms—**Hybrid processor, image processing, modular optical/digital interface, optical computing, optical/digital processor, radar signal processing, register transfer modules.**

I. Introduction

THE PHILOSOPHY of combining an optical and digital processor into a hybrid system in which the best advantages of each system are realized is well documented [1]–[4]. This architecture efficiently realizes the best advantages of both processing disciplines: the high speed and parallel processing of an optical system and the programability and flexibility of a digital system. In this paper we present a detailed description of the software and hardware for our optical/digital interface. The digital section of the hybrid processor described is used only to analyze the contents of the Fourier transform or correlation planes of the optical section of the system. The role of this interface and the discussion of its fabrication can best be understood after a brief review of the salient operations of an optical processor and the output format to be expected (Section II). Since other more lengthy reviews [5], [6] are available, this section will be intentionally brief.

The general operations required of the digital section of the system are presented as a tree in Section III followed by several examples of these operations in Section IV. Implementation of the specific functions in hardware and software are then discussed in Section V.

Manuscript received October 7, 1976; revised July 22, 1977. This work was supported by the Office of Naval Research under Contracts NR048-600 and NR350-011.

D. Casasent is with the Department of Electrical Engineering, Carnegie-Mellon University, Pittsburgh, PA 15313.

P. D. Rapp was with the Department of Electrical Engineering, Carnegie-Mellon University, Pittsburgh, PA 15313. He is now with the Department of Systems Engineering, Electronics Systems Division, Harris Corporation, Melbourne, FL 32901.

II. Optical Processors

The hybrid processor block diagram of Fig. 1 and the general remarks to follow will suffice to explain the operations possible in a coherent optical processor. The main purpose of this section is to describe the output plane

0018-9340/78/0800-0732$00.75