Automatic Feature Design for Optical Character Recognition Using an Evolutionary Search Procedure

FRED W. M. STENTIFORD

Abstract—An automatic evolutionary search is applied to the problem of feature extraction in an OCR application. A performance measure based on feature independence is used to generate features which do not appear to suffer from peaking effects [17]. Features are extracted from a training set of 30,600 machine printed 34 class alphanumeric characters derived from British mail. Classification results on the training set and a test set of 10,200 characters are reported for an increasing number of features. A 1.01 percent forced decision error rate is obtained on the test data using 316 features. The hardware implementation should be cheap and fast to operate. The performance compares favorably with current low cost OCR page readers.

Index Terms—Evolution, feature extraction, image processing, nearest neighbor classifier, optical character, pattern recognition.

INTRODUCTION

The problems of feature extraction and selection in pattern recognition are serious obstacles in the design of efficient systems [1]–[3]. The task of choosing $N$ measurements or features thought to be useful in a particular recognition problem is normally carried out by intuition. There are several techniques which may be subsequently applied for the selection of a good subset from the $N$ features [4]–[7], but there are often computational difficulties when $N$ becomes large.

Whether intuition is the best approach to initial feature selection is debatable. In the field of the recognition of handprinted characters there are examples where the choice of features has yielded very good results [8], [9]. However, it is also evident that a great deal of system complexity is frequently incorporated during the feature selection process [10]–[13]. The extraction of features is the central issue in handprint recognition [14], [15] and the success of any project generally hinges on this task.

Manuscript received July 10, 1984; revised November 19, 1984.

The author is with the British Telecom Research Laboratories, Ipswich, Suffolk, England.

0162-8828/85/0500-0349$01.00 © 1985 IEEE
An automated technique is desirable because it has significant advantages over manual methods.

a) Tedium manual involvement is reduced.

b) It is applicable to a range of pattern recognition problems.

c) Feature selection is not restricted by any lack of human insight.

d) Features may be confined to only those which are easily implementable.

e) System complexity can be more easily related to performance.

f) Features may be chosen according to a range of objective criteria.

Little work has been reported on automatic feature design primarily because of the huge number of possible features available in any one problem. A practical Kanji character recognizer produced by Toshiba [16] makes use of the K-L expansion method to select features. This approach requires a difficult computation of eigenvalues if the number of features is large, and can actually worsen class separations unless supplementary algorithms are added.

This paper describes an evolutionary technique for the automatic generation of features which have low mutual correlation. Such a set of features can theoretically be extended to achieve improved performance without suffering from a peaking effect [17].

The omni-font recognition of machine printed characters is analyzed, but the method may be applied to handprint or any recognition problem which is representable in the form of two-dimensional binary data. The results reported in this paper indicate a performance/complexity trend which promises to yield predictable improvements on relatively cheap hardware. It should be noted that some commercial devices achieve superior performances to those described here. It is not always clear, however, how these machines may be extended to cater to low quality data.

**A Simple Optical Character Recognition System**

At the outset it was decided that the OCR system should be cheap and fast to operate. This necessarily meant choosing simple hardware with very little processing associated with the recognition.

Five basic elements are distinguished in a standard system illustrated in Fig. 1. The print line on the document passes over an area illuminated by lamps, and the reflected light is then collected by a lens and imaged on a vertical line array of photodiodes. The video circuits amplify the signals from these photodiodes and quantize them into black and white grid points. The matrix is a register where the incoming grid points are assembled to form an electrical image of the part of the line which has just been scanned. The extraction and decision logic analyzes the electrical image continuously and accumulates at the appropriate time a measure of fit between the character pattern in the matrix and each of the character classes to be recognized. The final output is the identity of a recognized class or a reject indication [18].

The matrix is a long serial shift register which brings the binary patterns to be recognized into all possible translation positions. To avoid wraparound problems, white bits are inserted in the matrix shift register after each scan. For simplicity features for recognition are required to take their inputs from this matrix, and each feature (operator) therefore corresponds to a set of spatially related points. Each point or element in the operator is either black-seeking or white-seeking, fitting black or white matrix points, respectively. The whole operator fits only if all its elements fit simultaneously (Fig. 2).

The presence of an operator match may be determined from the output value of combinational logic connected to specified shift register elements. A register holds the value of all such operator outputs at one translational position of the data in the shift register. A succession of M such register values corre-
sponds to a window which is scanned over the data as it moves through the shift register. The OR function of the $M$ registers indicates the presence or otherwise of each scanning operator somewhere within the data window. The binary operator response vector obtained may be rapidly compared with reference vectors using conventional nearest neighbor classifier circuitry.

The feature design problem is then to choose a set of operators which, when scanned in all possible positions over a set of characters, will match the characters with low correlation between the operator responses.

Kamentsky and Liu [19], [20], and Liu [21] evaluated programmed algorithms for designing scanning operators with high information measures. These algorithms chose operator elements randomly and did not ensure that outputs of the final set of operators were independent. This meant that an eventual rise in error rate with the number of measurements could be expected in an extended system [22].

Other "n-tuple" approaches [23]-[25] also employed random selection of elements to produce nonscanning operators and all, therefore, required the presupgmention and centering of each character before recognition.

A Feature Selection Criterion

The performance of a feature should be assessed not in isolation but in conjunction with the others in the working set. To be useful a feature must add materially to the information already extracted by the others and ideally there should be no correlation between feature occurrences.

Let a reference pattern $C_i$ give rise to an $N$-dimensional response vector $F_i$ in feature space where

$$F_i = \{f_{i1}, f_{i2}, \ldots, f_{iN}\} \quad i = 1, 2, \ldots, K$$

and where $f_{ij}$ is the response of the $j$th feature to the $i$th reference pattern.

In a similar way let the $j$th feature correspond to the vector response $f_j$ from the $K$ reference patterns $C_i$ where

$$f_j = \{f_{j1}, f_{j2}, \ldots, f_{jK}\} \quad j = 1, 2, \ldots, N.$$

In the task of recognizing the $C_i$ in a noisy environment, the best discrimination is achieved when the $F_i$ are all angularly separated from each other by a maximum angle. It can be shown [26] that the largest angle $\alpha$ such that the angular distance between any pair of reference vectors is at least $\alpha$ is given by

$$\alpha = \cos^{-1}\left(\frac{-1}{K - 1}\right) \quad \text{for} \quad N \geq K - 1.$$

Moreover, this maximum can be obtained when the unit reference vectors lie on the vertices of a regular $K - 1$ dimensional simplex inscribed in the $N$-dimensional unit sphere.

It is to be noted that $\alpha$ is independent of $N$ and that if the number of features is increased beyond $K - 1$, the optimum solution will not improve. In this way the most economical and optimum solution occurs precisely when $N = K - 1$ and any extension of the number of features must be accompanied by a similar increase in the number of reference patterns. In addition, as $N$ or $K$ increases, an orthogonal arrangement of the $F_i$ becomes very close to the optimum.

Consider a $K$ class OCR problem in which each character $C_k$ is represented as a two-dimensional $w \times h$ binary array

$$C_k = \{\mu_k(i, j)\} \quad i = 1, 2, 3, \ldots, w, \quad j = 1, 2, 3, \ldots, h$$

$$\mu_k(i, j) = 0 \quad \text{or} \quad 1.$$

An operator or feature $q_j$ is defined as a set of integer triples

$$q_j = \{x_{ij}, y_{ij}, q_{ij}\} \quad i = 1, 2, 3, \ldots, u_j$$

$$1 \leq x_{ij} \leq w$$

$$1 \leq y_{ij} \leq h$$

$$q_{ij} = 0 \quad \text{or} \quad 1.$$

The operator $q_j$ fits the pattern $C_k$ if there exists an offset $(x_{ij}, y_{ij})$ such that

$$\mu_k(x_{ij} + X_j, y_{ij} + Y_j) = q_{ij} \quad \text{for} \quad i = 1, 2, 3, \ldots, u_j.$$

Let

$$f_{ij} = +1 \quad \text{if the } i\text{th operator fits the reference pattern from the } j\text{th class}$$

$$f_{ij} = -1 \quad \text{otherwise}.$$

A measure of an operator’s performance is required which, when increased, guarantees an improvement in the separation or the orthogonality of a set of reference pattern responses.

It can be shown very simply that given $f_{ij}$ is a square matrix $(K = N)$, then a necessary and sufficient condition that the row vectors $F_j$ are mutually orthogonal is that the column vectors $f_j$ are mutually orthogonal (Appendix). This means that any tendency for the $f_j$ to become mutually orthogonal necessarily implies that the $F_j$ will also.

Measures of angles between the operator response vectors $f_j$ therefore have a direct bearing on the separation in feature space of the reference pattern response vectors $F_i$. Indeed

$$G = \sum_{i, j} (f_i \cdot f_j)^2 = \sum_{i, j} (F_i \cdot F_j)^2.$$

(See the Appendix.) Hence, $\Sigma_{i, j} (f_i \cdot f_j)^2$ can be used as a measure of the total performance of the reference pattern responses when $K = N$.

Now, assuming $(f_i \cdot f_i)^2 = \text{constant} \text{for all } i$

$$G = \sum_{i, j} (f_i \cdot f_j)^2$$

$$= \sum_{j \neq a} \sum_i (f_i \cdot f_j)^2 + \sum_j (f_a \cdot f_j)^2$$

$$= \sum_{j \neq a} (f_i \cdot f_j)^2 + 2 \sum_{j \neq a} (f_a \cdot f_j)^2 + (f_a \cdot f_a)^2.$$

Hence, for $G$ to improve and decrease with $f_a, f_a$ must change so as to decrease the value of $M_a$ where

$$M_a = \sum_{j \neq a} (f_a \cdot f_j)^2.$$

Suppose now that the operators in a set each have response $f_j$ and measure $M_j$. Suppose also that

$$M_m = \max\{M_j\} \quad j = 1, 2, \ldots, N.$$

A candidate operator, $a$, would replace the worst performing operator $(m\text{th})$ in the set if

$$M_a = \sum_{j \neq a} (f_a \cdot f_j)^2 < M_m.$$

A succession of such replacements would ensure an increasingly orthogonal set of the $f_j$’s and, hence, also the $F_j$’s. The operator measure $M_a$ therefore provides not only a measure of feature usefulness to be used in a search, but also an assessment for the inclusion of operators in a feature set. To some extent this overcomes some of the problems encountered when features are evaluated on their own [27].
THE EVOLUTIONARY SEARCH

The evaluation parameter $M_k$ has no known formal relation with the structure of operators and their responses over a character set. Hence, there can be no logical search which guarantees optimal answers within a practical length of time. It may be possible that the choice of certain heuristics may quicken the search under certain circumstances, but this is true only if the prior information is available concerning the total data set under study. Heuristics based on incomplete or intuitive information certainly might reduce the size of the search space but at the same time they can preclude many problem solutions. There is no way, apart from exhaustive enumeration, to prove precisely how restrictive heuristics are, for if there was a way, sufficient $a$ priori information would be available at the outset to solve the problem with a fast formal search.

A search procedure is required, therefore, which is not exhaustive and is not guided by informal heuristics. The only algorithm which completely satisfies these requirements is the random search in which each selection is equally likely to be any one of the universe of possibilities. Each random candidate is evaluated against a measure and an evolutionary sequence of steadily improving structures is obtained.

OPERATOR GENERATION

The characters used in this study were 34 classes of upper case alphanumeric published from all qualities and styles of addresses on machine printed British mail. Zeros and O's were considered to be members of the same character class, as were I's and 1's. The characters were scaled and boxed to lie within a 16 X 24 binary grid. All the black seeking elements of generated operators were restricted to lie within a 16 X 24 area without fear of barring the search from potentially good operators. The universe of possible operators therefore contains $3^{984}$ items and the number of possibilities for the selection of 100 operators, say, from this set is of course a much larger quantity.

A reference set of characters was defined initially by the selection of one character ($C_i$) from each of the classes of characters to be recognized. Also, the $f_i$ were initially set to -1.

A random arrangement of black and white seeking elements served as an operator start state. The operator was matched on each $C_i$ and a response vector $f_i$ constructed and assigned a score $M_i$. Elements were tested for redundancy by checking if this score improved after their removal. Once redundant elements had been extracted, the operator was placed in a library if there was room or if $M_i$ was an improvement of any one of the scores of the $N$ operators already present. Operator production was halted if the time from the last operator entry in the library exceeded a limit.

The replacement of an operator in the library affected the scores of all the other operators and so each $M_i$ was recomputed at this stage. The CPU requirement for this computation was tailored to increase linearly with $N$.

TRAINING AND EVALUATION

A single nearest neighbor classifier (NNC) was chosen principally for its ease of implementation. A more sophisticated classifier might have achieved an improved absolute performance but would not have shed any additional light on the problem of eliminating performance peaking effects.

In most of the experiments $N \sim K - 1$ so that corresponding to an operator library and a reference set there were produced $K$ response vectors, each of dimensionality $K - 1$. Although the response vectors were well separated in $K - 1$ space, they were not necessarily central to the clusters of co-classed character responses they represented. By virtue of the earlier search criterion it was to be expected that the $K$ response vectors would lie on the extremities of cluster boundaries so that intercluster distances were maximized.

The choice of a single nearest neighbor classifier therefore required that averaged response vectors be used as prototypes rather than those derived solely from the reference set. To this end a training set of characters was chosen and $K$ new response vectors computed from the average responses to the $K - 1$ operators.

Let the $k$th reference response vector belonging to class $p$ be $r_i^{(p)}$.

Let the response vector corresponding to the $i$th training character $C_i^{(p)}$ belonging to class $p$ be $c_i^{(p)}$.

Define

$$\text{dist} \left( F_i, F_j \right) = \sum_k | f_{ik} - f_{jk} | \quad (\text{Hamming distance}).$$

Let $S_{i}^{p}$ be the set of characters $C_i^{(p)}$ such that

$$\text{dist} \left( c_i^{(p)}, r_i^{(p)} \right) = \min_{k} \text{dist} \left( c_i^{(p)}, r_k^{(p)} \right).$$

Suppose that there are $n_l \geq 0$ characters in $S_{i}^{p}$. Then the average response vector to replace $r_i^{(p)}$ is

$$\frac{1}{n_l} \sum_{c_i^{(p)} \in S_{i}^{p}} c_i^{(p)}.$$

The training set and an unseen test set were then classified using the recomputed prototype vectors in an NNC which assigned a character $C$ with response $c$ to class $p$ where

$$\text{dist} \left( c, r_i^{(p)} \right) = \min_{k} \text{dist} \left( c, r_k^{(p)} \right) \quad \text{for some} \ l.$$  

A simple analysis of the results revealed a few prototypes which were not contributing usefully to the classification. That is, they were either not sufficiently close to pattern responses to influence their correct classification, or they were causing a disproportionate number of misclassifications. These prototypes could be eliminated.

The reference set of $K$ characters was extended by using a selection of the newly discovered errors and reduced by the elimination of redundant prototypes. Further operators were evolved on this new reference set, both extending the total number of operators and replacing some old ones. A new NNC was developed as before and a new error rate determined over the training and test sets.

The process was repeated by selecting new modes of reference characters as necessary, and, hence, increasing the number of operators and the average separation between class references (ideally $K/2$).

RESULTS

As described earlier, the iterative improvement in performance of the NNC followed an increase in both the number of features or operators and the number of character references. The character references were added according to their usefulness in reducing the training set error rate and it is interesting to note the different forms of each "3" reference, for example, in Fig. 3.

Results on a training set of 30 600 characters and a test set of 10 200 characters were obtained for a series of NNC's of increasing dimensionality. The error rates were obtained through a forced decision (no rejects) and are plotted against the number of operators in Fig. 4. It was apparent that in the experiments conducted there was a logarithmic relationship between the error rate and the NNC dimensionality. A 1.01 percent test set error rate was obtained with 316 operators and
319 reference characters. Operators possessed between five and eleven elements with an average of between six and seven.

ANALYSIS OF OPERATORS

Inspection of the structure of the operators does not easily reveal any formal way in which they may be constructed. Apart from their OCR capability the elements in each opera-
tor appear random. The magnitude of such an analysis task is illustrated in an example.

Fig. 5 shows an operator fitting 11 alphanumeric characters and not matching another 11. The three black seeking elements in the operator indicate a bottom left corner, but “N” and “S” are matched and yet do not possess this feature. The bottom left characteristic certainly prevents a match with “I,” “7,” “J,” and “T,” but three white seeking elements are strategically placed to prevent matches with “F,” “O,” “P,” and “U” even though these characters do have bottom left corners. In addition, “3,” “5,” and “M” are not matched, but not because they do not possess a bottom left feature because the black seeking elements are in fact fit. At the same time the character pairs (D, 0), (E, F) and (S, 5) are all distinguished by the same operator. In fact, 11 X 11 = 121 character dichotomies are resolved by this one operator. Although not shown in Fig. 5, the response of this operator is designed to be as
orthogonal as possible to severable hundred other operator responses.

An intuitive approach to the design of such operators is difficult because of the large number of factors which must be simultaneously handled. For this reason it is more usual for intuitively designed templates to fit single classes of character by correlation and suffer a reduced discrimination capability [28], [29].

ADVANTAGES AND DISADVANTAGES OF THE APPROACH

The main advantage over other feature extraction procedures is shown by the promising yield of sets of features possessing a high degree of independence. The computational requirement during evolution increases linearly with the number of pattern references \( K \). About 100 suitable operators can be produced in a few hours on a Ferranti ARGUS 700 minicomputer.

The method involves no manual intervention save that of identifying different modes of characters in the training data.

In the OCR problem reported in this paper, the evolutionary search technique enables a very large space of features to be considered. The simple structure of the features means that such a system can be implemented cheaply and operated at TV video speeds.

It is fundamental to this approach to feature extraction that as few restrictions as possible be placed on the search to avoid the exclusion of potential problem solutions. This necessarily requires an evolutionary strategy in which randomly chosen items are evaluated to determine their usefulness. In effect, this procedure abandons the use of any formal or heuristic strategy for the choice of good operators, and means that it is impossible to predict how the numbers of operators and references will grow in any particular problem. There is also no firm guarantee that the results obtained are optimal, although the convergence of earlier results can provide evidence for this. For the same reasons there is no logical indication of the existence of operators having the required properties during a search.

CONCLUSIONS AND FUTURE WORK

An automatic feature extraction system has been described which was applied to an OCR problem. A recognition system was designed which was capable of extension as unrecognizable forms of pattern were encountered without loss of earlier performance. This was achieved by maintaining a high degree of independence between features and no performance peaking effects were observed during the generation of a 316 feature NNC.

The evolutionary search proved successful in a large space of possibilities, and there was no evidence to indicate that there would be difficulty in finding more operators as required during the extension of the system. The work could be applied to handprint recognition or other visual recognition tasks.

In the OCR problem it was evident that high performance features existed which defied an intuitive explanation of their structure. Operators possessed between five and eleven elements and averaged between six and seven elements. The simple structures of the operators and the NNC mean that such a system should be cheap to implement and would operate at video speeds. The 1.01 percent test set error rate on omni-font data compares favorably with the performance of low cost commercial page readers which achieve very low error rates only on 6–8 high quality character fonts [301], [311]. Further work with more sophisticated classifiers could achieve an improved performance.

The flexibility of the OCR system may be increased by incorporating tilted and unscaled modes of characters into the training set. This, if successful, would remove some of the need to de-skew and scale characters before recognition. Work is continuing to establish the feasibility of extending the recognition mechanism to reject intercharacter contingencies and, hence, contribute towards the location of the characters. Recognition would then be independent of the text format.

An indication of the application of the methods to handprinted alphanumeric characters is given by a preliminary result on a 34 class training data set derived from 97 writers. 225 operators were generated which yielded a 4.03 percent error rate. An equivalent performance on machine printed characters is obtained with about 25 fewer operators (Fig. 4). The increased variability of handprint does account for the need for extra operators. However, the difference is not great and further work may show that the handprint performance improves in a similar fashion to machine print as more operators are added.

Since the technique requires no a priori knowledge of the form of the patterns to be recognized, it could have a more general application in the computer vision field. It could avoid some of the human effort involved in the extraction of detailed information in the solution of specific visual processing problems.

APPENDIX

\[
G = \sum_{i,j} (f_i \cdot f_j)^2 \\
= \sum_{i,j} \left( \sum_k f_{ik} f_{jk} \right)^2 \\
= \sum_{i,j,k,m} f_{ik} f_{jk} f_{im} f_{jm} \\
= \sum_{i,j,k,m} f_{ik} f_{im} f_{jk} f_{jm} \\
= \sum_{i,j,k,m} f_{ik} f_{jk} f_{im} f_{jm}.
\]

Interchanging dummy variables \( i, k, j, \) and \( m \)

\[
= \sum_{i,j,k,m} f_{ik} f_{jk} f_{im} f_{jm} \\
= \sum_{i,j} \left( \sum_k f_{ik} f_{jk} \right)^2 \\
= \sum_{i,j} (F_i \cdot F_j)^2.
\]

Hence, the mutual orthogonality of the \( f_j \) is a necessary and sufficient condition for the mutual orthogonality of the \( F_j \).

ACKNOWLEDGMENT

Acknowledgment is made to the Director of Research of British Telecom for permission to publish this paper and to the British Post Office for the use of their data. The author also wishes to acknowledge the Plessey Company Limited, the Science Research Council, and his supervisor Prof. D. W. Lewin for their support during the early stages of this work.

REFERENCES


[31] "Workless station," Dest Corp., San Jose, CA.