A Preliminary Distinctive Feature Analysis for Upper-Case, Roman, Handwritten Character Recognition

Frank C. Carman II

Follow this and additional works at: https://scholarsarchive.byu.edu/dlls

BYU ScholarsArchive Citation
Available at: https://scholarsarchive.byu.edu/dlls/vol8/iss1/14

This Article is brought to you for free and open access by the Journals at BYU ScholarsArchive. It has been accepted for inclusion in Deseret Language and Linguistic Society Symposium by an authorized editor of BYU ScholarsArchive. For more information, please contact scholarsarchive@byu.edu, ellen_amatangelo@byu.edu.
1. INTRODUCTION

Character recognition, whether by man or machine, is a subset of the general field of pattern recognition. The term "pattern" has been defined in various ways in the literature. We examine here only a few of those definitions: Jackson states that "a pattern is a collection of objects, each of which has the property that it satisfies a certain criterion, known as the pattern rule for the pattern. The objects in a pattern are said to be pattern examples." Tou simply describes a pattern as "the description of an object." As human beings we are performing pattern recognition every moment of our waking lives. The mind, through the eye, recognizes not only individual shapes and forms but classifies objects into larger categories such as humans, animals, trees, and so forth. In a similar manner, we classify sounds as linguistically meaningful or as noise or as music, etc. Again quoting Tou: "Thus, the problem of pattern recognition may be regarded as one of discriminating the input data, not between individual patterns but between populations, via the search for features or invariant attributes among members of a population." He goes on to explain that the study of pattern recognition problems is broken down into two major categories:

"1. The study of the pattern recognition capability of human beings and other living organisms.

"2. The development of theory and techniques for the design of devices capable of performing a given recognition task for a specific application."

The linguist is primarily concerned with the first of these categories; whereas the computer scientist allocates his time to the study of the second category. It is the desire of the author to merge these two viewpoints in an attempt to create a computer system for machine recognition of upper-case, Roman, handwritten characters.

One object of this paper is to probe the psychological foundation of feature selection for machine recognition of characters, and - in particular - upper-case, Roman, handwritten characters. There are numerous approaches to the problem of machine character recognition found not only in the literature but also in the marketplace. For example, the banking industry utilizes a special numerical font which is
standardized for machine reading of the pertinent data on your individualized checks. Other systems, which are readily available in the marketplace, are trained to recognize certain type fonts and are, therefore, very useful for typed data input. Many schools and universities provide computer registration that incorporates machine recognition of hand-entered registration forms. Some character recognition systems demand that the user conform to a specific printing style. In such systems the USER must be trained to perform according to the machine specifications. Other character recognition systems allow the user to train the MACHINE on the user's own personal writing style. Such machines are more versatile; yet they demand greater sophistication in the software, greatly increasing the cost of the system. Yet all of these systems are plagued with the inability to recognize, without training, an arbitrary individual writing style. This is the case for lack of a psychologically based distinctive feature analysis of human character recognition, its imitation, and implementation on a computer-based system. The human can recognize, WITHOUT prior training, the written text of an arbitrary individual. This fact alone should suggest the psychological feature invariants of the orthographic sign system that we use. By discovering these invariants we should be able to simulate human recognition on a computer-based system.

It is the intent of this paper to place machine character recognition on a solid linguistic foundation. It is the feeling of the author that such an approach, combined with the pattern recognition techniques of computer science, will bring us much closer to the day when optical character recognition becomes an inexpensive and vital part of our environment.

2. SEMIOTICS

The purpose of language is to communicate information. This is accomplished through a system of signs, be they verbal, written, or in some other form (such as American Sign Language used by the deaf). Semiotics is the science that pertains to the study of the linguistic sign. The Prague school is generally credited with the pursuit and development of semiotics. Among those most noted for its establishment as an essential part of linguistic study is Roman Jakobson. An attempt will not be made here to cover this field in depth; however, an introduction to the basic terms and concepts of semiotics will be put forth as a foundation whereon we may perform a viable distinctive feature analysis of upper-case, Roman characters.

2.1 Features

Tou states that "pattern recognition can be defined as the categorization of input data into identifiable classes via
the extraction of significant features or attributes of the data from a background of irrelevant detail." Thus, features are attributes or characteristics of an object that may be useful in classifying it. We say that a feature is distinctive, or that we have a distinctive feature, if it is psychologically necessary in the pattern classification performed by man. Thus, for example, we distinguish the difference between /p/ and /b/ through the distinctive feature VOICING. Features may be either concrete or abstract. We may distinguish objects by their color, shape, physical state, etc. We even group abstract notions such as love, hate, addition, subtraction by characteristic features.

There is much psychological evidence indicating that man utilizes binary features in his pattern recognition processes. A binary feature is one that can be represented by two values (+,-) depicting its presence or absence in the pattern being classified. In the above phonemic example, /b/ has the feature [+VOICE] whereas /p/ has the feature [-VOICE]. If a feature is required for proper identification of an object, we say that the feature is "marked". That is, it is marked if either the presence or absence of the feature must be indicated for the object's proper identification. If the feature may or may not be present (i.e., nothing is said about it in the classification scheme), then we say that that feature is "unmarked" for the classification under consideration.

Features may be of two kinds: (1) sense determinative, or (2) sense discriminative. Sense determinative features determine the sense of or the semantic content of the sign in question. Semantic distinctive features such as those of Van Schooneveld or of the author are sense determinative. That is, their presence within the sign contributes to the semantic content or meaning of the sign. Sense discriminative features do not contribute to the meaning of the sign, their meaning being simply "mere otherness". They are used only to distinguish one sign from another. For example, the feature VOICING is used to distinguish /p/ from /b/, but in no way does it attribute a sense or meaning to either of these phonemes. It will be shown within this paper that a feature analysis of orthographic characters will yield a set of sense discriminative features, strikingly analogous to a phonemic analysis of spoken language.

2.2 The Structure of the Linguistic Sign

The linguistic sign, or "signum", consists of two inseparable components: (1) the "signans" which is the "name", or form, by which the sign is identified, and (2) the "signatum" which is the "value", or meaning, of the sign. It is important to recognize the inseparability of the signans and the signatum. Thus, whenever a specific sign is referenced, its semantic content is likewise utilized in constructing the information.
being conveyed. This inseparableness of the sign's duality implies a constancy of meaning, or semantic invariance of the sign IRRESPECTIVE of its environment.

Without going into detail or even being complete, we can illustrate the notion of invariance by examining the signatum of the preposition "to". One of the striking features of this preposition is its implicit deixis. Thus, "A... to B" indicates that to find A, one must look to B as an index. There is also implied a certain involvement or dependency between A and B. It is thus that we understand the difference between "John threw the ball at me." vs "John threw the ball to me." We say, then, that the preposition "to" has a signatum marked with the semantic features DEIXIS and INVOLVEMENT. These can be illustrated in the following:

1. "John drove to the city." The deixis is clear in this example. Involvement is evident in that the "city" is the DESIRED destination of the subject, ostensibly to carry out some action that involves being in or at the "city".

2. "John wants to play." The action desired is found through the deixis of "to". That is, the movement implied by "play" is desired by the agent of the expression. There is a deictic time motion (as opposed to the special motion of the previous example) showing the agent directed toward the desired object (or goal). "To" is used to show the involvement of the agent with the desired action.

The signatum, then, can be characterized by a set (or bundle) of distinctive features. This is common practice in specifying phonemic values. These features characterize the invariant of the linguistic sign irrespective of its environment. This invariance allows us to develop new words within our language. Without invariance, we would not be able to communicate unless each separate utterance within the language were assigned a meaning in some ad hoc manner. This is obviously not the case; we do communicate, and do so through the semantic invariance of the signs we draw upon.

3. THE FEATURE DOMAIN

3.1 Where are the Features to be Found?

We now turn our attention to the orthographic characters under consideration. In the quest for meaningful visual distinctive features, one must ask the question: Where are the features to be found? The retina of the eye consists of numerous quantized receptors. Similarly, an optical scan of a visual scene is presented to the computer in a like manner, through pixels, or quantized visual elements. Thus, the primary data provided to the brain through the eye or to a
computer through an optical scanner is a set of quantized intensity values. For the purposes of this discussion, we may consider each pixel as having only two values: either "on" or "off". We shall also ignore color, since it does not affect the perception of characters, except for its possible "prosodic" content. Thus, our receptor space - that space in which the original data is sensed and recorded - consists of a two-dimensional array of bits, each of which is in one of two states. But is this the space in which the visual distinctive features are to be found? We answer this question with a definite NO. They are not to be found in such a space any more than acoustic distinctive features are found in the time domain of a digitized voice signal. Even though the primary data is supplied to the brain as an array of pixels, the human mind does not "see" dots. What then does it see? The distinctive features are psychological phenomena and must be treated as such.

3.2 The Psychology of Visual Character Recognition

Some time ago the author took his older children on a hike through Timpanogous Cave located in Utah. Among the many peculiar characteristics of this cave are its varied stalactite and stalagmite formations. As the guide pointed out to us some of the more interesting formations, my children would often be heard to remark: "Daddy, that looks like such and such." where "such and such" might be a face, the shape of a bear, etc. In these instances the visual signal would suggest object shapes with which the children were very familiar. How often have we sought among the stars on a clear night the Big Dipper or some mythical character? This capacity to transform a set of "dots" into an abstract pattern is unique to man. Human character recognition utilizes this same process. What the eye sees suggests abstract structures whose relationships with each other form the features by which one character is distinguished from another. As a first attempt to discover the features that the human mind uses in character recognition of Roman orthography, we will posit three abstract structures that are "seen" or suggested by the pattern of dots in the receptor space. The primary distinguishing characteristic between these three structures is their "dimensionality". Thus, we distinguish between "straight" lines vs. "curved" surfaces, the former being one dimensional and the latter being two dimensional. It must be emphasized that these are abstract notions and not related to the physical world. Thus, the pattern of dots representing a "straight" line may not be found in a straight pattern in the receptor space; they may even possess a discontinuity. However, they will suggest to the mind the notion of a straight line. This is a crucial point and the base upon which human psychological perception is founded. The "curved" surface structure, possessing a two dimensional attribute, may be further characterized by the oppositions "closed" vs. "open". The notion of "within" or "inside of"
is perceived in connection with curved surfaces. Thus, with a closed surface, such as "O", one perceives an absolute boundedness or containment within the structure. On the other hand, an open surface, such as a "U" or a "C" elicits a perception of the opposition of containment, namely a freedom to enter or leave the surface at will. Thus, we have three basic structural segments that are characterized by the maximal oppositions: (1) one dimension vs. two dimensions, and (2) open (free) vs. closed (contained). The desired features, then, will be found in the topological relationships in this segment space. There is certainly a precedent set for this possibility in the acoustic realm. The human performs an effective Fourier transformation of the acoustic signal that it receives from a time domain receptor space to a frequency domain feature space. It is mapped data within this frequency domain that the hair cells of the basilar membrane quantize for feature generation. The acoustic distinctive features are then found as topological relationships within this frequency space, i.e., relationships among formant structures.

But one might ask: Can not many characters be created from a single continuous line? How can features be derived from a single line? In answer to this query, one must remember that it is not the physical continuity of a line that we are interested in. Rather, it is the PSYCHOLOGICAL continuity or discontinuity that is sought for. It is the psychological segmentation that occurs within the mind that generates the topology necessary for feature extraction. Therefore, to approach the level of human character recognition through a machine, we must first map the receptor space onto the abstract segment space perceived by the human mind. Creating such a machine is the crux and pivotal point of the entire recognition problem.

4. THE "GRAPHHEME"

The term "grapheme" has been used in the literature to designate an orthographic character, the supposed smallest unit in the perceptual process. We will show below that a character is not the smallest perceptual unit in the human recognition system. The more basic units, of which the character is composed, we shall term "graphemes". Henceforth, then, we shall depart from the customary usage of the word "grapheme", using it to reference a structural component of a character instead. These structural components, or graphemes, consist of single segments or segment pairs, where the segments are those described in the previous section. A single segment may belong to two graphemes. For example, the two line segments in the letter "T" form a grapheme by virtue of their topological relationship with each other. A character composed of four segments, on the other hand, might consist of six graphemes, since four segments taken two at a time results in six possible pairs. This is the case with the
letter "E", the mutual relationship of each pair of line segments forming separate graphemes. Each grapheme is a linguistic sign. The signans of this sign consists of the form that the grapheme takes and the signatum consists of the distinctive features that identify the grapheme as unique from other graphemes.

As a spoken word consists of a set of phonemes, so the written character consists of a set of graphemes. There is one major difference in this analogy. The phonemes are ordered sequentially in time and are perceived according to this sequential relationship. However, the graphemes within a character constitute an unordered set and are perceived simultaneously, without regard to time sequencing.

4.2 Types of Graphemes

As phonemes are classified into groups such as vowels, stops, and fricatives, graphemes are similarly classified according to the segments from which they are formed. Since we are dealing with three basic segments, we have nine possible grapheme types as follows:

1. "I" graphemes - line alone
2. "C" graphemes - open curve alone
3. "O" graphemes - closed curve alone
4. "X" graphemes - line with line
5. "S" graphemes - open curve with open curve
6. "8" graphemes - closed curve with closed curve
7. "P" graphemes - line with open curve
8. "Q" graphemes - line with closed curve
9. "g" graphemes - open curve with closed curve

The name of the grapheme type is meant to be iconic, suggesting the segment pairs that form the grapheme. Each of these classifications consists of a set of individual graphemes which are characterized by topological distinctive features. The feature distinctions between single segment graphemes are those of dimensionality and the open/closed opposition, whereas those between the segments of a pair grapheme are relational features. It is these distinctive features that form the sense discriminative network for the character recognition of man. We shall describe the feature distinctions in the pair graphemes below. Since the "8" and the "g" graphemes are not represented in the upper-case, Roman characters, we shall omit them from our descriptions. It is also unnecessary to discuss further the structure of single segment graphemes. Table 3 at the end of this section contains a summary of the graphemic character representations.

4.3 The "X" Graphemes

The X-graphemes are relationships between pairs of straight
lines. The first notable feature that one encounters when examining these graphemes is the opposition intersection/no-intersection. To determine the intersection feature, we bound or enclose the character by the smallest possible rectangular box. We then extend the line segments under question in both directions until they meet the box boundaries on both ends of each segment. A pair of line segments composing a grapheme are then said to intersect if their extended representations intersect within the character bounding box; otherwise they do not intersect.

For graphemes that intersect, the place of intersection on each line segment becomes the next distinctive feature. The intersection point is either at the end of the line or in the center of the line. If the intersection point of the extended line is beyond the end-point of the original line segment, we say that the intersection is at the end of the line. The region of the original line that can be considered as the "center" of the line for classifying the intersection as at the "center" must yet be determined by statistical studies of actual handwritten character samples.

![Decision Tree for X-Grapheme Feature Selection](image)

**INTERSECTION(1):**
- **ee** = end-end
- **cc** = center-center
- **ce** = center-end
- **ec** = end-center

**CONNECTIVITY BIT REPRESENTATION:**
- **ee** = end-end
- **cc** = center-center
- **ce** = center-end
- **ec** = end-center

**NO-INTERSECTION(0):**
- **ae** = adjacent ends
- **bc** = both centers
- **de** = diagonal ends
- **nc** = not connected

**Fig. 1:** The decision tree for X-grapheme feature selection

Those X-graphemes that are labeled as non-intersecting may yet be distinguished by a connectivity feature. That is, they are either disjoint (unconnected) or they are connected.
by means of a single segment. If they are thus connected, we
classify this connection into three areas: (1) the adjacent
ends are connected, (2) the centers are connected, or (3) the
opposite (diagonal) ends are connected. Fig. 1 illustrates
the feature options as a decision tree for X-graphemes.
It will be noted from Fig. 1 that there are eight possible

**TABLE 1**

**SAMPLE X-GRAPHEME REPRESENTATIONS**

<table>
<thead>
<tr>
<th>Segment #</th>
<th>Connectivity</th>
<th>Grapheme Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>I(a) I(b) Int</td>
<td>X(1,2:3) or X(1,2:7)</td>
</tr>
<tr>
<td>1</td>
<td>2 1 1 1</td>
<td>X(1,2:7)</td>
</tr>
<tr>
<td>1</td>
<td>3 1 0 1</td>
<td>X(1,3:5)</td>
</tr>
<tr>
<td>2</td>
<td>3 1 0 1</td>
<td>X(2,3:5)</td>
</tr>
<tr>
<td>1</td>
<td>2 1 1 1</td>
<td>X(1,2:7)</td>
</tr>
<tr>
<td>1</td>
<td>3 1 0 1</td>
<td>X(1,3:5)</td>
</tr>
<tr>
<td>2</td>
<td>4 1 1 1</td>
<td>X(1,4:7)</td>
</tr>
<tr>
<td>3</td>
<td>4 0 0 0</td>
<td>X(2,3:0)</td>
</tr>
<tr>
<td>2</td>
<td>4 0 0 0</td>
<td>X(2,4:0)</td>
</tr>
<tr>
<td>1</td>
<td>2 1 1 1</td>
<td>X(1,2:7)</td>
</tr>
<tr>
<td>1</td>
<td>3 1 0 1</td>
<td>X(1,3:5)</td>
</tr>
<tr>
<td>2</td>
<td>3 0 0 0</td>
<td>X(2,3:0)</td>
</tr>
<tr>
<td>1</td>
<td>2 0 1 0</td>
<td>X(1,2:2)</td>
</tr>
<tr>
<td>1</td>
<td>2 1 1 1</td>
<td>X(1,2:7)</td>
</tr>
<tr>
<td>1</td>
<td>4 1 1 1</td>
<td>X(1,4:0)</td>
</tr>
<tr>
<td>1</td>
<td>2 1 1 1</td>
<td>X(1,2:7)</td>
</tr>
<tr>
<td>2</td>
<td>3 1 1 1</td>
<td>X(2,3:7)</td>
</tr>
<tr>
<td>3</td>
<td>4 1 1 1</td>
<td>X(3,4:7)</td>
</tr>
<tr>
<td>1</td>
<td>2 0 0 1</td>
<td>X(1,2:1)</td>
</tr>
<tr>
<td>1</td>
<td>2 1 0 1</td>
<td>X(1,2:5)</td>
</tr>
<tr>
<td>1</td>
<td>2 1 0 0</td>
<td>X(1,2:4)</td>
</tr>
<tr>
<td>1</td>
<td>2 0 0 1</td>
<td>X(1,2:1)</td>
</tr>
</tbody>
</table>

X-graphemes. Thus, three bits are sufficient to represent
these possibilities. Table 1 illustrates these graphemes
with some specific examples. It will be noted that the
graphemic composition for the letter "A" as shown in Table 1 may have one of two possible graphemes as the relationship between segments 1 and 2. This does not mean that there is a variability in the compositional invariant of the letter A. Such would be a contradiction in terms. Rather, the feature "intersection/no-intersection" is not marked in the feature specification for the grapheme. That is, nothing is said concerning its value. Therefore, it may be either 0 or 1. This allows for individual variations in the letter such as would be found among $\bar{A}$, $\bar{A}$, and $\bar{A}$.

4.4 The "S" Graphemes

The S-graphemes are relationships between pairs of open curves. Although several topological relationships can be created between such pairs, only two of these relationships are found in the upper-case, Roman characters. Thus, we restrict our classification of these graphemes to a single distinctive feature which we term "uni-directional". The letter "B" possesses this feature with the value 1. That is, the open curves - taken as "arrows" - point in the same direction. If they pointed in opposite directions, the feature value would be 0, as is found in the letter "S".

4.5 The "P" Graphemes

The P-graphemes are relationships between segment pairs consisting of a line and an open curve. We always treat the line segment as the first segment in this relationship. Using the methodology of minimal pair comparisons, we find four distinctive features that characterize this class of graphemes. The first feature embodies the oppositions extended(1) vs. compacted(0). The pair relationship is extended if the two segments are connected "in line" with each other, whereas they are compacted if they are "side-by-side". These relationships may be illustrated as follows:

\[
\begin{align*}
\text{J} \quad \text{and} \quad \text{P} & \quad \text{are extended;} \\
\text{Y} \quad \text{and} \quad \text{P} & \quad \text{are compacted.}
\end{align*}
\]

The next evident feature that arises out of our comparisons is designated through the oppositions parallel(1) vs. perpendicular(0). If the segments are "parallel", then the line segment is parallel to the tangent of the open curve at the point of connection. Similarly, the segments are "perpendicular" if the line segment is perpendicular to this same tangent at the connection point. These are illustrated in the following:

\[
\begin{align*}
\text{J} \quad \text{and} \quad \text{Y} & \quad \text{are parallel;} \\
\text{P} \quad \text{and} \quad \text{P} & \quad \text{are perpendicular.}
\end{align*}
\]
The last feature, applied to each segment individually, relates the general size of the segment to the size of the boxed character. Thus, for example, the open curve in the letter "D" is designated as "whole size"(l) whereas the same open curve in the letter "P" is designated as "half size"(0). Table 2 illustrates these features with examples of upper-case characters.

### Table 2

**Sample P-Grapheme Representations**

$I(a) = \text{line } a $  \hspace{1cm}  $C(b) = \text{open curve } b$

<table>
<thead>
<tr>
<th>Segment #</th>
<th>Extended(l)</th>
<th>Parallel(l)</th>
<th>Whole size(l)</th>
<th>Grapheme Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>I(a) C(b)</td>
<td></td>
<td></td>
<td>I(a) C(b)</td>
<td></td>
</tr>
<tr>
<td>B{l 1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>P(1,2:2)</td>
</tr>
<tr>
<td></td>
<td>3</td>
<td>0</td>
<td>0</td>
<td>P(1,3:2)</td>
</tr>
<tr>
<td>D{l 1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>P(1,2:3)</td>
</tr>
<tr>
<td>P{l 1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>P(1,2:2)</td>
</tr>
<tr>
<td>F{l 1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>P(1,2:12)</td>
</tr>
<tr>
<td>G{l 1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>P(1,2:1)</td>
</tr>
</tbody>
</table>

### Table 3

**Graphemic Representation of Upper-Case Roman Characters**

\[
\begin{align*}
A\{l & : X(1,2:3), X(1,2:7), X(1,3:5), X(2,3:5) \\
B\{l & : P(1,2:2), P(1,3:2), S(2,3:1) \\
C & : C \\
D\{l & : P(1,2:3) \\
E\{l & : X(1,2:7), X(1,3:5), X(1,4:7), X(2,3:0), X(3,4:0), X(2,4:0) \\
F\{l & : X(1,2:7), X(1,3:5), X(2,3:0) \\
G\{l & : P(1,2:1) \\
H\{l & : X(1,2:2) \\
\end{align*}
\]
4.6 The "Q" Graphemes

The Q-graphemes are used to distinguish relationships between a line and a closed curve. There is only one basic distinctive feature that need be considered with these graphemes. It is analogous to the parallel/perpendicular oppositions of the P-graphemes and is referenced by the oppositions tangent(1)/perpendicular(0). The segments are tangent if the line coincides with the tangent to the closed curve at the point of connection. On the other hand, they are perpendicular if the line intersects the closed curve such that it is perpendicular to the tangent of the curve at the point of intersection. There is only one letter in the upper-case
characters that contains this grapheme, namely, the letter "Q". Its graphemic notation is simply Q(1,2:0), where the first segment of the pair is the line.

5. THE GRAPHIC SYLLABLE

We have developed a graphemic theory of character perception much along the lines of phonemic theory of verbal perception. One might naturally ask at this stage whether or not the analogical relationship with phonology can be carried to a greater extent. More specifically, we might ask questions such as: Is there a graphic "syllable"? In phonology we treat the syllable as a structural unit above that of the phoneme. The syllable is a psychological reality wherein the mind automatically creates a "natural" segmentation of the verbal signal. That is, we impose a structural hierarchy upon the signal that we hear. A given discourse is treated as a set of sentences, which are composed of phrases, which - in turn - are composed of words, which are then built up of syllables, which finally are structured from phonemes. This, obviously, is a simplified view of language perception; but it serves to illustrate the function of psychologically imposed structure in the communicative process. Is there, then, a comparable "graphic" syllable? There is evidence in the languages of man that would demand an affirmative answer to this question. The pictographic nature of Chinese is an illustration of this phenomenon. The Chinese character is recognized - not as an arbitrary set of lines with peculiar relationships - but as consisting of structural components, the core of which is known as a "radical". There are 214 radicals, or primitive structural units, that are used in building some 15,000 Chinese characters. These radicals may be looked upon as constituting graphic syllables. Entire Chinese dictionaries are developed around the order of complexity of the governing radical within the character.

Turning our attention to the Roman alphabet, we see evidences of a similar psychological hierarchy in the structure of the upper-case characters. For example, if one were to separate the letter "A" into two constituents, which of the following possibilities would seem most "natural"?

1. /\  
2. \  
3. \ -

If the reader selected option (3), he will not find himself alone. There seems to be a natural division of the letter "A" according to this pattern. Then, could one not suggest that the letter "A" consists of two graphic "syllables", namely "\" and "-"? Carrying this a step further, could not the featural invariant of the letter "A" be dependent
upon the relationship between these two syllables? This being the case, a free variation of the syllable "/" (e.g., realizations such as "\" and "/\") would not affect the featural invariant between this syllable and its counterpart "—".

The above inquiries are meant simply to stimulate thinking along the lines of a psychologically imposed syllabic structure in the character recognition processes of man. However, this phenomenon has not been taken into account in the preliminary computer implementation of the recognition system described in this paper.

6. ORIENTATIONAL FEATURES

The reader may have noticed in Table 3 that some characters have identical graphemic representations. Thus, we are forced to conclude the existence of other featural entities that allow man to uniquely distinguish one character from another. We have treated in the graphemic analysis only those relationships between the three basic segment types, irrespective of the observer. For example, the letter "b" is identical to the letter "d", which is also identical to the letter "p" from a graphemic point of view. We find, however, that the "point of view" of the observer plays an important role in character recognition. Hence, we may conclude that characters are further distinguished by orientational distinctive features. These features operate upon the character as a whole and not upon individual graphemes of which the character is constructed. The application of such features is illustrated by the following examples:

1. "L" rotated 45 degrees yields "V".
2. "N" rotated 90 degrees yields "Z".
3. "M" rotated 180 degrees yields "W".

Other examples, outside of the upper-case set, will be noted in the next section of this paper. There appear to be eight possible orientations for a given letter demanding three bits, or orientational distinctive features, to accommodate all possible permutations of the observer's point of view.

These observations suffice for rotational features, but they do not include a parity feature. We can not obtain the letter "d" from the letter "b" by rotation, and yet they have the same graphemic description. It is therefore necessary to posit a parity feature. Even though it is not needed for upper-case characters, this feature is requisite to complete the set of orientational distinctive features used by man.
7. PERMUTING THE FEATURES BEYOND UPPER-CASE ROMAN CHARACTERS

If the distinctive features discussed in this paper are truly representative of the psychological featural analysis performed subconsciously by man in character recognition, then permutations of these features, not found in the upper-case characters, should result in "characters" that are recognized by related alphabets (i.e. those that are diachronically related, possessing, subsequently, a similar or identical feature structure) or that would appear to be "natural" additions to our own alphabet if required. Many of these "new" characters will be found in the lower-case, Roman character set as well as the Greek and Russian alphabets. Several such characters are also found as "invented" symbols in mathematics as well as the various sciences. Table 4 illustrates a few of these.

**TABLE 4**

SAMPLE (NON UPPER-CASE ROMAN) CHARACTERS POSSESSING GRAPHEMIC AND ORIENTATIONAL FEATURES

<table>
<thead>
<tr>
<th>Lower-case Roman</th>
<th>Greek</th>
<th>Russian</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>( \Gamma )</td>
<td>( \mathcal{B} )</td>
<td>( \mathcal{L} )</td>
</tr>
<tr>
<td>b</td>
<td>( \Delta )</td>
<td>( \Gamma )</td>
<td>( \Pi )</td>
</tr>
<tr>
<td>d</td>
<td>( \Theta )</td>
<td>( \Pi )</td>
<td>( \Gamma )</td>
</tr>
<tr>
<td>g</td>
<td>( \Lambda )</td>
<td>( \Pi )</td>
<td>( \Gamma )</td>
</tr>
<tr>
<td>h</td>
<td>( \Xi )</td>
<td>( \Pi )</td>
<td>( \Gamma )</td>
</tr>
<tr>
<td>m</td>
<td>( \pi )</td>
<td>( \phi )</td>
<td>( \psi )</td>
</tr>
<tr>
<td>n</td>
<td>( \Sigma )</td>
<td>( \chi )</td>
<td>( \psi )</td>
</tr>
<tr>
<td>p</td>
<td>( \gamma )</td>
<td>( \Psi )</td>
<td>( \psi )</td>
</tr>
<tr>
<td>q</td>
<td>( \phi )</td>
<td>( \Theta )</td>
<td>( \psi )</td>
</tr>
<tr>
<td>u</td>
<td>( \Omega )</td>
<td>( \Psi )</td>
<td>( \psi )</td>
</tr>
<tr>
<td>y</td>
<td>( \mu )</td>
<td>( \Theta )</td>
<td>( \psi )</td>
</tr>
</tbody>
</table>

13.15
8. GRAPHICAL PROSODICS - GRAPHEMICS VS. GRAPHEMICS

Prosodics, in the traditional sense, has referred to the functions of stress, pitch, duration, and intonation in the spoken language. Each of these attributes carries information in the communication process. This information may denote a mood, such as fear or anger; or it may change the entire sense of a statement as found in sarcastic expressions. Likewise, there exists a graphical prosodics with similar functions. We recognize this in the individual writing styles that we encounter. So strongly does the prosodic information in handwritten text carry information about the author of the text, that entire studies have been devoted to personality analysis through handwriting traits. Similarly, we utilize different type styles for different occasions, each conveying their own special mood. One would not think of using a formal business style on a wedding invitation. The art of calligraphy perhaps epitomizes the extent to which graphical prosodics can be carried. Whereas we associate psychological invariance with the study of graphemics, we now introduce the study of graphical prosodics under the name of "graphetics". Our only motive for introducing this subject is to contrast it with graphemics. Submerged in a multitude of prosodic environments, the sign's distinctive feature description remains invariant. This is an important point and the foundation upon which we build our machine recognition system. It is also the psychological foundation upon which man recognizes textual material, be it handwritten or typed.

9. THE MACHINE RECOGNITION SYSTEM

We now come to that aspect of our inquiry that is the goal of the above extensive analysis - the actual machine implementation of a character recognition system. Recognition systems are generally divided into two parts: (1) the training stage, and (2) the testing stage (or recognition stage). Although we do not have a formal "training stage" in our system, we have by no means neglected this aspect of computer character recognition. One must remember the purpose of the training stage - namely, to discover those features upon which the pattern classes may be distinguished. Training, then, is a discovery procedure. We have here applied an alternative discovery procedure - a linguistic psychoanalysis. Such an approach has been taken because of its applicability to this area of pattern recognition. It is felt that a linguistic inquiry is not only appropriate as a discovery procedure, but is far superior to a mathematical approach that might be taken by a computer system. In saying this, the author realizes full well that the opposite would be true in other circumstances. The linguist would never claim the ability to ferret out those distinctive features in biomedical applications of pattern recognition. However, he is
fully competent to bring his expertise to bear in linguistic and linguistically related investigations.

9.1 The Segmental Search

Inherent in the recognition system must be the capability to identify the three basic graphemic segments. This is not an easy task since the pattern of dots received by the retina (or, in our case, the optical scanner) only suggests the abstract pattern which we are attempting to identify. Therefore, the first task of the computer system is to recognize those abstract segments in the same manner as would the human mind recognize them.

9.2 A Heuristic Search

Once the fundamental segments have been located within a character, their topological, or graphemic, relationships need to be ascertained. This can be accomplished by identifying those features that distinguish a segment pair as a legitimate grapheme. Once the graphemes are located, it is a simple task to look up the character's identification in a table that maintains the entire featural description for each character.

The process of recognition can be greatly helped by utilizing a heuristic approach. We know what we are looking for. Therefore, rather than searching randomly for unexpected graphemes, we can search heuristically for expected graphemes. This will not only speed up the recognition process, but it will also help to eliminate "noise", i.e., spurious strokes on the page resulting, generally, from stylistic variations.

9.3 The Treatment of Ambiguities

It is recognized that legitimate ambiguities may arise. Such is often the case in "sloppy" handwriting. What is the computer system to do in such cases? Each grapheme will be determined by its "closeness" in featural description to the specified grapheme's binary description. For example, an intersection of two line segments may be found as occurring in the "center" of one of the segments. This is a binary evaluation. The raw data used to make this judgment may indicate that the intersection was so classified by a "small" margin. When such is the case, and an ambiguity in recognition exists, alternative featural classifications are taken in an attempt to disambiguate the character analysis.

When this approach fails to eliminate the ambiguity, a context sensitive examination may be performed. Although not implemented in the system described in this paper, it would check a "spelling" dictionary in a search for a legitimate word, alternating the ambiguous character's identifi-
cation during the search. The human performs a similar search, subconsciously, extending the domain beyond a simple spelling check to include the sense or contextual meaning of possible alternatives for the ambiguous word.

10. A PRELIMINARY IMPLEMENTATION

10.1 The Main Control Algorithm

The main control algorithm allows the user to examine up to 1248 character patterns, 48 patterns for each of the 26 characters in the Roman alphabet. These characters were obtained by digitizing handwritten samples from eight individuals. Each individual contributed six sample patterns for each of the 26 letters of the alphabet.

The program begins in a query mode, requesting from the user the following information:

A. To which output device (scope or plotter) should the letters be sent?

B. Do you wish to examine all letters of each character selected?

C. Do you wish all characters to be examined?

D. If the answer to C was "NO", the system asks, on a character by character basis, if the user wishes the character to be examined.

E. If the answer to B was "NO", the system asks for the upper and lower bounds on the contiguous set of samples (1 - 48 possible) to be examined for the given character.

F. Do you wish each character examined to be sent to the output device (scope or plotter)? If the answer to this question is "NO", only those characters that were not properly identified by the recognition system are sent to the output device.

After the above query, the system proceeds to examine the requested character patterns, one at a time. This examination consists of a recognition attempt. If the character was not recognized correctly, its dot matrix plus the line segment abstractions used in the recognition attempt are sent to the output device. The system proceeds thus until all requested character patterns have been examined for identification.
10.2 Segment Recognition

We found in the foregoing discussion that there were three psychological segments whose mutual relationships with each other yield the distinctive features in the character recognition of man. These segments are (1) the "straight line", (2) the "open curve", and (3) the "closed curve". These are abstract entities, related to the set of source data pixels through the abstraction mechanism of the human mind. It is this mechanism that must be modeled in segment identification. For this preliminary recognition implementation, we only utilized a "straight line" recognizing mechanism, the "open curve" and "closed curve" segment recognition being postponed until subsequent implementations. Thus, we were limited in the feature recognition capability possible in the current system.

The "straight line" algorithm consisted of strobing in from the left, right, top, or bottom of a boxed character, looking for a left vertical, right vertical, top horizontal, or bottom horizontal line, respectively. The strobe generally started at the center of a box edge, thus limiting our line identification to the information gained from a single row or column. The sought-for lines were allowed to vary plus or minus 45 degrees from the vertical (or horizontal). The desired straight line was determined by examining a 91 degree arc (-45 degrees to +45 degrees), one degree at a time, for the longest contiguous sequence of pixels contained between two parallel bounding lines whose separation was a function of the sought-for line's thickness. The line was then determined by a linear least squares fit, using those pixels that were contained in the desired line but not also contained in an intersecting line.

10.3 Feature Determination

The topological relationships between "straight lines" were based on those developed in this paper. However, liberties were taken to vary from these strict featural relationships in order to accommodate a more simplified system as a first implementational approximation to human character recognition modeling. Also, with the limitation to a single segment type abstraction, we were forced to examine other than strict graphemic relationships to produce a sufficiently large set of attributes for distinguishing between characters. These attributes include distance relationships between line end- and mid-points, slope relationships between lines, and slope bounds for given lines, etc. The Boolean test conditions for each of the ten letters considered are found in the program code of the CHANLZ subroutine.
11. DISCUSSION OF RESULTS

Figures 2a and 2b show samples of characters that were recognized correctly. The straight line abstractions by the system are also superimposed on these characters. In general, the line recognition algorithm was successful. However, there are certain weaknesses that it possesses which can be discovered from a close inspection of the characters that were NOT recognized correctly.

It is instructive to note those characters which yielded the greatest numbers of errors. Time did not allow us to incorporate a topological recognition algorithm for each of the 26 letters of the alphabet. Yet an analysis of those considered was very revealing. The system was able to recognize with 100% accuracy the letters I, L, T, V, and Z if they alone constituted the set being examined. When J was added to this set, one error appeared. Adding F increased the number of incorrect recognition attempts by 5. The system, then, performed with a 98.2% correct recognition on this expanded set of characters. However, adding the characters K, X, and Y resulted in a large increase of incorrect recognition attempts (28 total errors) yielding only a 94.2% correct recognition attempt for this final set of ten characters. These additional errors provided a means of scrutinizing the line recognition algorithm, supplying us with an analysis which will aid in improving this algorithm for future implementations.

Figures 3a, 3b, and 3c show the letters that were incorrectly recognized and the results of the line algorithm applied to these characters in the recognition process. These errors can be grouped into the following categories:


4. Diagonal was found as a "vertical": K-25, K-27, K-28. These could have been grouped in (3). However, it was evident that they resulted from the wide-angle line search and are thus listed separately.
Fig. 2a: Correctly recognized characters with "line" abstractions.
5. Ambiguity due to "thick" lines: K-31, K-36, X-33, X-34, X-35.

6. Sloppy character: Y-12. This could also be included in (3). However, from a human recognition standpoint, it was determined to isolate its primary deficiency.

7. The errors in "T" were not errors in the T-algorithm but resulted from the K and Y algorithms being applied (with an improper recognition) prior to the T algorithm. Without K, X, and Y, these errors did not exist.

The line recognition mechanism used in this preliminary implementation could be improved for future systems. As was noted earlier, the line routine scans a 91 degree angle for the longest line within this arc. If the system were able to specify the scan angle, several errors could be eliminated by looking for specific lines, as determined heuristically, reducing the chances of locking onto a "line" which intersects the desired line at an acute angle. The line algorithm could be further improved by allowing the system to determine the average line thickness (used in the search process) over a specified scan range instead of scanning the entire length (or width) of the boxed character. This would prevent the system from missing the desired line whose thickness is greater than the maximum allowed thickness, determined primarily from another (thinner) line in the same scan range. Determining the existence of a line that one is looking for would become more efficient if the line search algorithm were to scan for the desired line from several levels instead of just a single level. A heuristic could then be used to determine the proper line sought for (or its absence).

The actual character could receive some preliminary massaging prior to character identification. The object of such an effort would be to (1) eliminate "noise" due to spurious, isolated, set pixels, (2) smooth lines by closing gaps consisting of single pixel discontinuities, and (3) "shrink" the character, artificially reducing the line width. Each of these operations would improve the chances of identifying a given character correctly.

Additional attributes would be necessary to include the remainder of the upper-case characters into the recognition system. The set of letters \{A(some), H,M,N,U,W\} could be approached by looking for a pair of "parallel" vertical lines that are connected in a unique manner. A connectivity algorithm would be used to determine where the parallel lines were connected, thus completing the identification of this subset of the alphabet. Some of the "A" characters could be recognized in the same manner as the "V" was recognized, since such "V-shaped A's" are simply inversions of an actual
Fig. 32: Incorrectly recognized characters with "line" abstractions.
Fig. 3b: Incorrectly recognized characters with "line" abstractions.

13.25
Fig. 3c: Incorrectly recognized characters with "line" abstractions.
The remaining letters, most of which contain closed or open curves, will need yet another algorithm to detect these curves. Without modeling the curve recognition system of the human mind, a simple row strobe could be used to look for enclosed surfaces, using a heuristic to determine the type and number of closed or open surfaces encountered.

The attributes formed from the relationships between lines are only approximations to the graphemic distinctive features discussed in this paper. The location of line intersections was not computed for the sake of programming simplicity and run-time speed. These were approximated on the basis of the distances between line end-points. However, to be accurate and avoid some of the pitfalls of this approach, a rigorous methodology should be used in which the actual intersections would be calculated in order to accurately identify graphemic features. In addition to correctly looking for line-line distinctive features, we need to develop the open-curve and closed-curve segment recognition mechanisms. These algorithms will certainly be more sophisticated than the simple line recognition algorithm.

It is felt that the accuracy of character recognition achieved with the ten characters examined certainly supports the distinctive feature analysis provided in this paper. This becomes more evident when we consider that only approximations to the actual topological distinctive features were implemented. Throughout the error analysis, it could be seen that an actual implementation of the graphemic features (e.g. intersection locations, etc.) would have averted several recognition errors that resulted from approximations only to the true topological features. It is hoped that future research along these lines will yet vindicate this approach in its entirety.

12. REFERENCES


4. Dodwell, P. C., "Perceptual Processing; Stimulus Equivalence and Pattern Recognition"

5. Fischer, George L., "Optical Character Recognition"

6. Jackson, Philip C., "Introduction to Artificial Intelligence"
8. ibid, 1980, "The Framework of Language"
9. ibid, 1979, "The Sound Shape of Language"
12. Lyons, John, 1977, "Semantics"
14. ibid, 1977, "Semiotic and Significs -- The Correspondence between Charles S. Peirce and Lady Welby"
16. ibid, 1964, "A Prague School Reader in Linguistics"
18. Tou, Julius T., Gonzalez, Raphael C., 1974, "Pattern Recognition"
20. Watanabe, S., 1969, "Methodologies of Pattern Recognition"
22. ibid, 1977, "A Semantic Analysis of Word Order -- Position of the Adjective in French"
23. Weaver, J. A., 1972, "Reading Machines"