



(22) Date de dépôt/Filing Date: 2002/03/11

(41) Mise à la disp. pub./Open to Public Insp.: 2003/09/11

(51) Cl.Int.⁷/Int.Cl.⁷ G06K 9/80, G06K 9/36

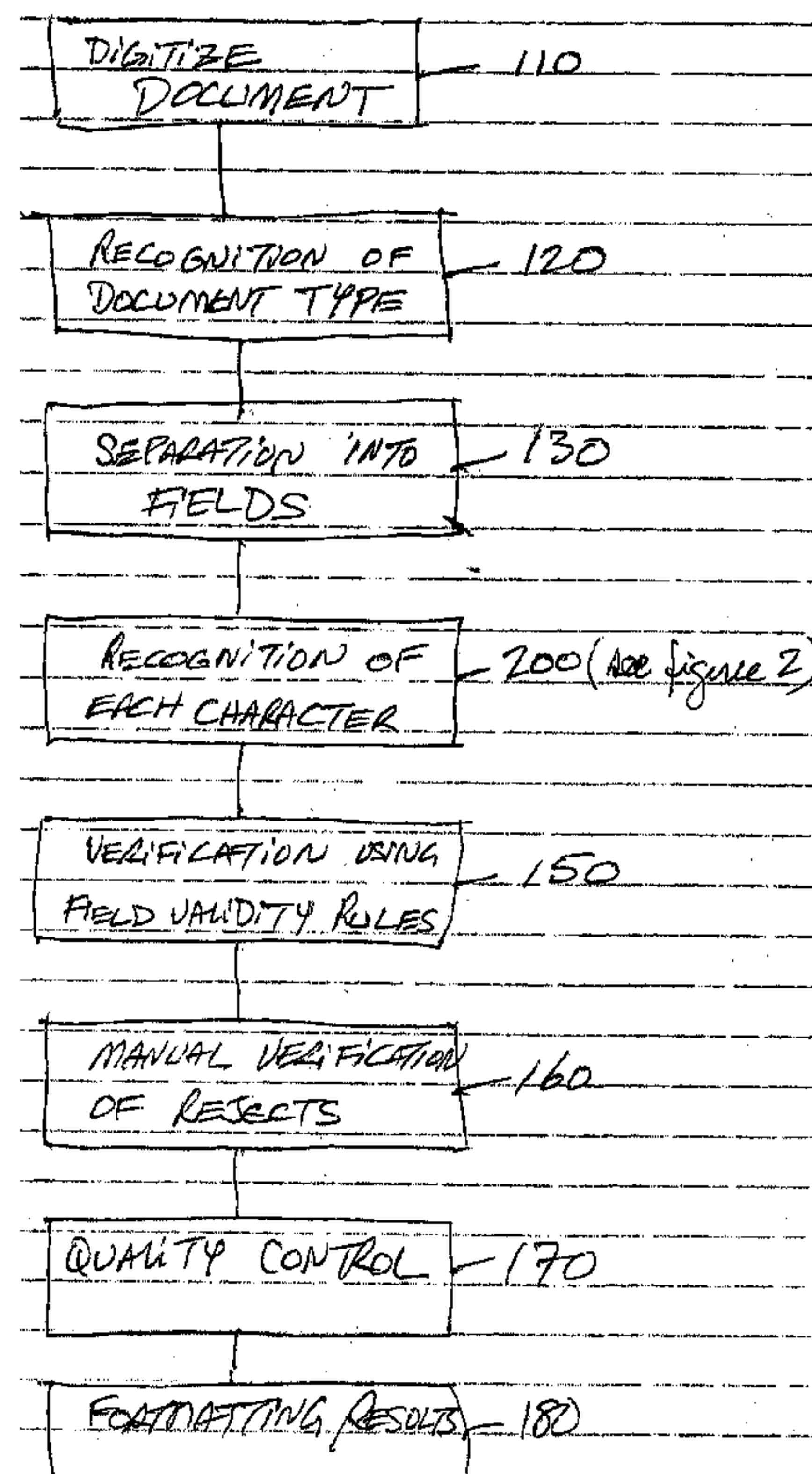
(71) Demandeur/Applicant:
NEO SYSTEMS INC., CA

(72) Inventeur/Inventor:
POLONOWSKI, JEAN-PIERRE, CA

(74) Agent: BROUILLETTE KOSIE PRINCE

(54) Titre : SYSTEME ET METHODE DE RECONNAISSANCE DE CARACTERES

(54) Title: CHARACTER RECOGNITION SYSTEM AND METHOD



CHARACTER RECOGNITION SYSTEM AND METHOD

5 BACKGROUND OF THE INVENTION

10 The present invention relates generally to character and pattern recognition machines and methods, and more particularly, to feature extraction systems for use with optical readers for reading characters which have been hand printed on forms or the like. Such forms include application forms, bank check, drafts and other similar business documents. Most of the known optical character recognition methods start with a pre-segmentation stage in which a line of text is first segmented into words and/or characters, followed by a separate character recognition stage. One novel feature of this invention is in the optimal elliptical class separation method which is used in the character recognition stage.

15 The use of word and/or character segmentation with subsequent feature extraction for image recognition is described, for example, in U.S. Pat. No. 5,438,630 issued Aug. 1, 1995 to Chen et al.

20 A template-based OCR method has been suggested which does not require pre-segmentation. However, this method has difficulty recognizing patterns because of intra-class variability of the patterns (see PCT/US93/01843 (international publication number WO93/18483), page 8).

25 While there are generally different views on the definition of the features of patterns, many studies made on the recognition of characters as well as the recognition of patterns have proved that the so-called quasi-topological features of a character or pattern such as the concavity, loop, end connectivity are very important for the recognition. To date, many different methods have been proposed for the purpose of extracting such quasi-phasic features. Some of these methods use analysis of the progressive slopes of the black pixels. U.S. Pat. No. 4,468,808 (Mori et al.)
30 classifies those analyses into three types. The first is the pattern contour tracking system developed by Grenias with IBM. Mori calls this a serial system. The second type is Mori's

preferred, the earliest patented example of which is Holt called the "Watchbird". In this type of analysis sequential rows and columns are compared. Another example of the sequential rows and column type is Holt's Center Referenced Using Red Line. Mori's third type is a parallel analysis system which Mori dismisses as either taking too long or costing too much. All systems involving the sequential analysis of the slope of black pixel groups suffer severely from smoothing end line thinning errors. Worse yet, they are very likely to produce substitution errors when the lines have voids or when unwanted lines touch.

A comprehensive survey of prior art handprint recognition systems is found in an article by C. Y. Suen et al. entitled "Automatic Recognition of Handprinted Characters--The State of the Art", Proceedings of the IEEE, Vol. 68, No. 4, April 1980, which is incorporated herein by reference. The preferred handprint character recognition technique of this invention uses none of the methods mentioned by Suen et al. or Mori et al.

U.S. Patent 5,504,822 (Holt) describes a character recognition system, which while using quasi-topological features, employs a different method of measuring and scoring such features, resulting in what Holt describes as a great improvement in performance of the reading machine. Briefly, the character recognition system of Holt employs measurement of the enclosure characteristics of each white pixel independently of other white pixels. Since the measurements are made in two (or more) dimensions rather than in one dimension (such as slope), the results are less sensitive to first order aberrations such as accidental voids, touching lines and small numbers of black pixels carrying noise only. In Holt's preferred embodiment, no noise processing is performed at all since all forms of noise processing are done at the expense of accuracy in recognition.

On-line handwriting recognition systems have been designed which compute feature vectors as functions of time. See, for example, T. Starner, J. Makhoul, R. Schwartz and G. Chou; "On-Line Cursive Handwriting Recognition Using Speech Recognition Methods; IEEE International Conference on Acoustics, Speech, and Signal Processing, Adelaide, Australia, Apr. 19-22, 1994, Vol. V. pp. 125-128. However, OCR applications are faced with the problem of recognizing a whole page of text which presents a two-dimensional problem for which there is no obvious way

of defining a feature vector as a function of one independent variable.

U.S. Patent 5,933,525 (Makhoul et al.) Describes an OCR system and method including a feature extraction technique of generating a signal representative of scanned text and as a function of only one independent variable (specifically, position within a line).

U.S. Patent 5,727,130 (Hung) describes the use of a fuzzy logic system. "Fuzzy Logic" was developed to enable data processors based on binary logic to provide an answer between "yes" and "no." Fuzzy logic is a logic system which has membership functions with fuzzy boundaries. Membership functions translate subjective expressions, such as "temperature is warm," into a value which typical data processors can recognize. A label such as "warm" is used to identify a range of input values whose boundaries are not points at which the label is true on one side and false on the other side. Rather, in a system which implements fuzzy logic, the boundaries of the membership functions gradually change and may overlap a boundary of an adjacent membership set. Therefore, a degree of membership is typically assigned to an input value. For example, given two membership functions over a range of temperatures, an input temperature may fall in the overlapping areas of both the functions labelled "cool" and "warm." Further processing would then be required to determine a degree of membership in each of the membership functions.

A step referred to as "fuzzification" is used to relate an input to a membership function in a system which implements fuzzy logic. The fuzzification process attaches concrete numerical values to subjective expressions such as "the temperature is warm." These numerical values attempt to provide a good approximation of human perception which is not generally limited to an environment of absolute truths. After the fuzzification step, a rule evaluation step is executed. During execution of the rule evaluation step, a technique referred to as "MIN-MAX" fuzzy inference is used to calculate numerical conclusions to linguistic rules defined by a user. Conclusions from the rule evaluation step are referred to as "fuzzy outputs" and may be true to varying degrees. Thus, competing results may be produced. A last step in the fuzzy logic process is referred to as "defuzzification." As the name implies, defuzzification is the process of combining all of the fuzzy outputs into a composite result which may be applied to a standard data processing system. For more information about fuzzy logic, refer to an article entitled

"Implementing Fuzzy Expert Rules in Hardware" by James M. Sibigtroth. The article, which is hereby incorporated by reference herein, was published in the April, 1992 issue of AI EXPERT on pages 25 through 31.

5 Fuzzy logic control systems have become increasingly popular in practical applications. Traditionally, the design of the knowledge base including membership functions and rules relies on a subjective human "rule-of-thumb" approach for decision-making. In addition, the control system is adapted (tuned) to the desired performance through trial and error. As a result, designing and adapting the fuzzy logic control system becomes a time-consuming task. To
10 overcome this drawback, neural network techniques have been used in assisting designers to generate rules and adapt the fuzzy logic control system automatically. For further discussions in this area, please refer to C. C. Hung, "Building a Neuro-Fuzzy Learning Control System," AI Expert, November, 1993; C. C. Hung, "Developing an Adaptive Fuzzy Logic System Using a Supervised Learning Method," Proceedings of the Embedded Systems Conference East, April,
15 pp. 209-220, 1994.

One of the areas being explored for utilization of the advantages of fuzzy logic systems is in optical character recognition. A fuzzy logic system is inherently well-suited for dealing with imprecise data and processing rules in parallel. However, the actual implementation of fuzzy
20 rule-based systems for this type of application often relies on a substantial amount of heuristic observation to express the knowledge of the system. In addition, it is not easy to design an optimal fuzzy system to capture the necessary features of each character.

Typically, one rule is used to recognize one character, and each character is represented as one
25 consequent of a rule. The actual implementation of fuzzy rule-based systems for this type of application often relies on a substantial amount of heuristic observation to express the membership functions for the antecedents of each rule. Each rule consists of several antecedents and consequents depending on the number of inputs and outputs, respectfully. Each antecedent in a given rule is defined as an input membership function, and each consequent is defined as an
30 output membership function.

Neural networks consist of highly interconnected processing units that can learn and globally estimate input-output functions in a parallel-distribution framework. Fuzzy logic system store and process rules that output fuzzy sets associated with input fuzzy sets in parallel. The similar parallelism properties of neural nets and fuzzy logic systems have lead to their integration in studies of the behaviour of highly complex systems. Neural systems are described in U.S. Patent 5,276,771 (Manukian et al.) and 5,251,268 (Colley et al.).

The process of designing a fuzzy rule-based system is tedious and critical for the success of the recognition. It must be done as efficiently and accurately as possible if it is to sufficiently address the OCR problem.

However, the output of Neural networks is dependent on the exact sequence of «learning» of the knowledge base. If the same knowledge base is fed twice to a neural Network with only one substitution in the learning sequence, the end result will be different in each case. This can be a major disadvantage for any OCR system.

U.S. Patent 5,727,130 (Hung) describes the use of Learning Vector Quantization ("LVQ"). LVQ, which is well-known in the art, accomplishes learning by placing input data in a finite number of known classes. The result is that this method provides the supervised effect of learning and enhances the classification accuracy of input patterns. It is also independent of the learning sequence.

The critical issue is to design more robust input membership functions that correspond to a rule. The linguistic term of a rule's antecedent, such as "input 1 is small," depends upon how accurately the input space is qualified while defining membership functions. LVQ can group similar input data into the same class by adjusting the connection weights between the inputs and their corresponding output. In other words, through supervised learning, the features of each class can be extracted from its associated inputs.

Hence, a learning vector quantization (LVQ) neural network may be used to optimize the features of each hand written character. Ming-Kuei Hu, "Visual Pattern Recognition Moment Invariant,"

IEEE Transaction on Information Theory, pp. 179-186, 1962, describes such a system which is hereby incorporated by reference herein. A LVQ network, is also disclosed in Teuvo Kohonen, "The Self-Organizing Map," Proceeding of the IEEE, Vol. 78, No. 9, pp. 1364-1479, September 1990, which is also hereby incorporated by reference herein.

5

A LVQ learning system is a two-layered network. The first layer is the input layer; the second is the competitive layer, which is organized as a two-dimensional grid. All units (a "unit" is represented as one input variable, such as x_1 , of one input pattern $(x_1, x_2, \dots)$) from the first layer to the second are fully interconnected. In the OCR example, the units of the second layer are grouped into classes, each of which pertains to one character. For purposes of training, an input pattern consists of the values of each input variable and its corresponding class (i.e., the character that it represents). A quantization unit in the competitive layer has an associated vector comprising the values of each interconnection from all the units in the input layer to itself. This vector implicitly defines an ideal form of character within a given class.

15

The LVQ learning system determines the class borders using a nearest-neighbor method. This method computes the smallest distance between the input vector $X: (x_1, x_2, \dots, x_n)$ and each quantization vector. In known systems, this computation is done in terms of Euclidean distance (straight line distance in multi-dimensional space).

20

Input vector X belongs to class $C(x)$, and quantization vector $w(I)$ belongs to class $C(w)$. If $C(x)$ and $C(w)$ belong to different classes, the $w(I)$ is pulled away from the class border to increase the classification accuracy. If $C(x)$ and $C(w)$ have the same class, the $w(I)$ closes to the center of the class. Then each input pattern is presented sequentially in the input layer and several iterations.

25 The weights of the quantization units in each class are fine-tuned to group around the center of the class. Therefore, the weight vector of the center unit within the class is represented as the optimum classification for the corresponding class. The result of the LVQ learning process is an optimized vector for each alphanumeric character.

30 U.S. Patent 5,832,474 (Lopresti et al.) also describes the use of vector quantization in a document search and retrieval system that does not require the recognition of individual characters.

However, most of the prior art character recognition systems are based on the concept of seeking to classify the greatest number of characters as possible. This means that such systems seek to attribut each character to be recognized to a class even if a certain degree of “guesswork” is necessary. As a result, such systems are far from being sufficiently accurate for many applications. This invention seeks to considerably improve the accuracy in reading characters and more particularly, hand printed or hand written characters.

SUMMARY OF THE INVENTION

The main objective of the invention is to automate as much as possible the input of data from documents. The invention is particularly useful with documents containing hand printed or hand written characters.

This invention has the following desirable characteristics and features:

1) It recognizes typed or printed, especially handprinted or hand written characters, in the various fields of a form.

2) The invention is able to pick out well known alphanumeric characters from an image which includes a great many patterns of types not previously taught to the machine: this is accomplished because the measurements of unknown shapes must have high correlation with previously taught shapes before they are even considered.

3) It is able to achieve a substitution rate (the percentage of wrongly chosen class names divided by the number of class names correctly chosen by a human judge working from only the same images) of near zero.

4) Furthermore, the character recognition system of this invention is adaptive in the sense that it can learn previously unknown patterns and automatically generate new recognition equations which are carefully crafted to be non-conflicting with previously learned patterns. Typically the

manual operator will provide a correct classification for a rejected character; non-supervised learning can also take place, e.g., learning using an automatic dictionary.

5) With respect to handprinted alphanumeric characters and symbols, one the characteristics of my invention is its ability to recognize touching and overlapping characters. This is a difficult task for any handprinted reader. This is accomplished by the use of "metaclass" training, e.g., "zero" touching another "zero" as is often seen on checks, is trained to be recognized as a new class called a "double zero". In a typical application 4 or 5 metaclasses may be used.

10 **BRIEF DESCRIPTION OF THE DRAWINGS**

These and other objects, advantages and features of the invention will be made manifest when considering the following detailed specification when taken in conjunction with the appended drawings figures wherein:

- 15
- FIG. 1 is a simplified block diagram of the general steps in the use of a character recognition system in accordance with this invention,
- FIG. 2 is a more detailed block diagram of the character recognition operation shown in figure 1;
- 20 FIG. 3 is a simplified block diagram of the construction of a recognition engine for use with this invention;
- FIG. 4 is a simplified block diagram describing the database creation process.
- FIG. 5 is a simplified block diagram describing the learning process.

25 **DESCRIPTION OF A PREFERRED EMBODIMENT**

The present invention will now be described in greater detail with reference to the drawings. The main general steps in the use of the invention are the following:

- 30
1. Digitization of the document to be read (110).
 2. Automatic recognition of the type of document (120).

3. Automatic reading of the content:

3.1 separation of the document into various fields (a field is usually a rectangular segment of the document which is destined to contain a specific information such as for example a bank account number) (130);

5 3.2 recognition of each character in the field (a detailed description is provided later) (200);

3.3 verifying the answer obtained in 3.2 by applying pre-determined field validity rules (for example the algorithm or key to validate a bank account number or a social security number) (150);

10 4. Manually verifying and inputting a field which, for any reason, the system could not automatically recognize. (160)

5. Quality control using known sampling techniques (optional) (170).

6. Formatting the data and writing it to a database (180).

15 The initial result of the recognition method is a series of possible digital characters each associated with a similarity score. Therefore, in a case where two possibilities exist for each character, a field of N characters will thus represent 2^N possibilities.

20 In order to reduce such possibilities, verification keys, (for example a bank account with its verification key). In all other cases, the only available method is to increase the quality of the bitmap image and to determine a threshold under which the result of the recognition process is considered to be unreliable.

Reading a field

25 Reading a field is the step in the character recognition process which consists in using a bitmap image of a predetermined portion of a document (generally a rectangular area) and transforming it into its equivalent digital representation (usually ASCII). A general outline of this sub-system is described in Figure 2.

30 It is preferable to read a field as a whole (210). This requires two separate operations:

1. Segmentation: this operation is used to separate the image of individual characters from each other (220). The size of the bitmap representation of a character can vary considerably from one character to the next and in the field of hand printed documents, from one writer to another. Known techniques may be used for this purpose.

5

2. Character recognition: this operation seeks to associate a given digital character to each character image defined during the segmentation step. A similarity score is given to each character so recognised. This operation presumes that each segmented image is an image of a single character. A few steps are required in the interpretation of a bitmap representation of a character:

10

- 2.1 The construction of a vector using a vector quantization technique (230). This is a mathematic tool having a number of components each describing a characteristic of the shapes to be analysed. These vectorial representations must allow the "recognition" of all known characters in the universe of valid characters which may be inputted in the field to be read;

15

- 2.2 The classification or separation of such vectors into a number of classes each corresponding to one shape which is sought to be recognised. Each such shape is usually one of the characters in the universe of valid characters in the application. However, it may also be useful to include vectors relating to frequently used invalid characters or marks. For example alphabetic characters in a numeric field or a minus sign. This character recognition method and system should preferably be optimized for certain specific uses. For example, if the field to be recognize is a numeric field, the class separation will be optimized to produce a digit from zero to nine and to exclude all others. In such an example, there are ten distinct outputs each being associated with its own similarity score (240). Then, the different outputs and their associated scores are sorted by decreasing similarity scores. In those cases where one output does not have a similarity score which is high enough to determine the "winning" character, further processing is required.

20

25

30

Recognition engine

As shown in figure 3, the character recognition system of the invention requires the use of one or more recognition engine(s).

5

The construction of a recognition engine requires the following steps:

1. The choice of a vector quantization method (310).
2. The choice of a vector classification/separation method (320).
- 10 3. The determination of the valid digital characters or “vocabulary” (usually ASCII) (330).
4. The creation of a vector base. This is an iterative process. The objective is to add to a pre-existing base of vectors each assigned to a given class (each character which needs to be recognized is considered to be a different class). This is done as follows (see Figure 4):
 - 15 4.1 real documents are read (410) and those characters for which the similarity score is under a pre-determined value are noted;
 - 4.2 an operator, can use all of the unrecognised “characters” and determine manually to which class they belong (460). A vector representing each such new shape is added to the vector database in relation to its manually assigned class (470).
 - 20 4.3 There follows a series of iterative adjustments of network coefficient to optimize the performance of the recognition engine (See Figure 5). This is a relatively long process and requires powerful processing equipment.

It can thus be seen that the recognition engine so created cannot provide other results than the
 25 classes for which it was built.

A number of different recognition engines may be used for a given task. For each type of field (numeric, alphabetical or other), a recognition method is specifically assigned. A recognition method may comprise a number of recognition engines each being composed of a particular
 30 vector and a class separation method as well as a “voting system” allowing the combination of the results of the various recognition engines using simple averages, weighted averages or veto

rules.

Reliability issues

- 5 Very often, it is impossible to be 100% certain that the character which was read is in reality the digital representation which was assigned to such character by the recognition method. Therefore, it is necessary to establish a measure of confidence in the accuracy of the recognition method. The confusion rate consists in the number of characters which were thought to have been recognized but were in fact wrongly recognized divided by the total number of characters
- 10 read. The rejection rate is the number of characters which the recognition method has failed to recognize over the total number of characters read. The read rate consists in the total number of characters that were accurately read over the total number of characters read. Therefore, the read rate plus the confusion rate plus the rejection rate should equal 100%.
- 15 This invention seeks to reduce the confusion rate to as close to zero as possible. Indeed, in many applications it is preferable to consider a character to be unrecognisable even if it is one of the ASCII characters that the system is seeking to recognize than to assign a wrong ASCII value to the character which was read. This is especially the case in financial applications.
- 20 This being said, the read rate has to be high enough for the recognition system to be worthwhile. Therefore, the ideal system is the one in which the confusion rate is zero and the read rate is as close as possible to perfect. It is extremely difficult to achieve such a result because of the various difficulties encountered during the recognition process. These include:
- poor quality images including poor contrast images caused by the use of a low quality
 - 25 writing instrument or a color that is not easy to digitize;
 - an image that is poorly recognized by the digitizing sub-system or the presence of a background image;
 - a poor definition of the zone in which the characters are to be written;
 - printed characters that extend outside the area reserved for the field which can include
 - 30 characters which are too large, character that are patched together or open characters;
 - errors may also result from using characters which are outside the desired “vocabulary”

(for example alphabetic character in a numeric field).

- poor class separation performance may also result from the quality or quantity of the vector examples in the vector database or the inability of the recognition engine to generalize. Indeed, hand printed documents are by definition nearly never identical from one person to the next or from one expression to the next even when written by the same person.

Vector Separation

The present invention seeks to increase the reliability of the vector separation method by the use of elliptical separation and preferably optimal elliptical separation. Technically, known LVQ learning systems determine the class borders using known nearest-neighbour methods. One such method is called optimal linear separation. It can be described summarily as follows :

- i. each class vector has important dimensions (from a hundred to three hundred and fifty components);
- ii. for each pair of classes it is possible to find an hyper plan allowing to separate them. In the case of N classes, they are separated two by two by $N(N-1)/2$ hyperplans. The equation of each hyperplan is simple: $\sum (\alpha_i x_i) = 0$. Therefore, for all members of class A, $\sum(\alpha_i x_i) > 0$ and for all members of class B, $\sum(\alpha_i x_i) < 0$. By the use of a simple algorithm, the various coefficients α_i converge toward the most efficient value. This known system can be useful when used with characters which are very close to those in the database. This is the case for example of typed characters.

However, it has a number a deficiencies, the most important of which is the difficulty, in real life, to find a hyperplan to separate very complex objects such as handprinted characters. Because a plan is by definition open ended, it is difficult to reject characters which are relatively distanced from the characters which are sought to be read (commas, exclamation marks, question marks, etc...)

In order to diminish such difficulties and obtain more precise results, it has been suggested to use optimal linear separation by segments. However, the results are not adequate.

Multi-layer perceptron is a well known application of neural networks. This method can be excellent if great care is used in the training phase. However, because no theoretical base exists to improve the result in a structured way, one must rely on trial and error processes which are extremely costly. As a result if the multi-layer perceptron system is "taught" the same data twice, two different results will be obtained.

10 **Elliptical Separation**

The instant invention seeks to increase the quality of the result at a lower cost. In order to reach such results, the following concepts have been adopted:

- 15 i. "my friends friends are my friends;
- ii. either a shape is very close or it is declared "unknown";
- iii. use of a nearest neighbour method to regroup various class members into "centers";
- iv. quantifying the population of each class "center";
- 20 v. the use of closed surfaces which greatly increases the exclusion of non- characters (spots, deletions, alterations, etc...).
- vi. low sensitivity to small classification errors;
- vii. availability of reliable mathematical base to insure optimization of read rate, computing time and the number of "centers".

25 It has been discovered that by using elliptical separation class members can be regrouped more efficiently.

30 Reading hand printed characters are particularly difficult because it is impossible to ever have a database of vectors representing every possible variation in hand written characters. However, the use of the invention considerably reduces the negative impact resulting from the use of

“incomplete” vector databases. However, by searching all the various ellipsoids and retaining all of those which may relate to the proposed character and by assigning a similarity score, it is possible to “learn” where the character belongs and make the appropriate adjustments to the various ellipsoids.

5

One way to assign a score to each particular member of an ellipsoid is to attribute a score of zero if the given character is located at the perimeter of the ellipsoid while attributing a score of one if such character is located at its “centre”. The appropriate scoring equation is : $1 - \exp(-\sum(\alpha_i x_i^2))$.

10 The equation for an hyper ellipses is the following : $\sum(\alpha_i x_i^2) = 0$. Optimization is used in order to find the optimal α_i coefficient.

Learning Process

15 As described above, once it has been determined that optimal elliptical separation will be used (320) and that the valid character vocabulary has been chosen (330), it is necessary to build a database of vectors from known specimens of the character vocabulary. Figure 4 describes this process.

20 Each character or shape is read from a learning base (410). A digitized (pixel) representation of the character or shape is then created (420). This pixel representation is then passed through the chosen (310) vector quantization method (430). A similarity score is then assigned for the character in respect of all classes (440). If the similarity score exceeds a predetermined threshold, it means the character is already known and can thus be ignored (455). If not, it is
25 manually examined to determine if it belongs to the desired vocabulary (460). If not, it is rejected (465). If it belongs to the vocabulary, it is added to the database and assigned to the proper class (470).

30 The learning process is completed as described in Figure 5. For each pattern in the database, the shortest distance between the vector and the closest vector of another class is measured (510).

This distance is then used to define, for each class, a sphere which will comprise only vectors

which are members of the same class (520). Therefore, the closed surface formed by this sphere separates all the members of the class contained in the sphere from all other classes. The number of same class vectors contained in the sphere is determined and the database is sorted using such number (from the largest to the smallest) (530).

5

Each N-dimension sphere (starting by the one containing the largest number of elements) (540) is then deformed into a N-dimension ellipsoid, N being the size of the vector until the optimization process is complete (550). This is a very long iterative process.

10 An algorithm similar to those used for hyperplan separation is used to optimize the separation between the various members contained in the ellipsoid from all others. The ellipsoid is deformed until members of other classes are reached. The result is the best vector available because more members of the same class are now found in the ellipsoid.

15 However, such iterations consume a large quantity of processing time and power. Therefore, it is preferable to end the iterations when a predetermined percentage of the perfect separation solution is reached. For example, in the case where the characters are numbers, the predetermined percentage is preferably 99.5%.

20 **Character Recognition**

Fig. 1 illustrates various steps in the use of the character recognition method. The first step (110) is to read a document. Then, in those cases where many different types of documents may be read, to determine the document type (120). Then, the document is separated into different fields (130). For each field, characters are separated from each other (220). Various known technics are used to separate individual pixel clusters from each other. The next step (230) consist in vector quantization of each pixel cluster. Then, as shown in figure 2, each vector is assigned a similarity or confidence score in respect of each possible class output (240). The best score (s) are verified to determine if they reach a pre-determined threshold (250). If none of the scores reaches the threshold the character is rejected as being unrecognizable and may be sent to manual verification (160). If there is a single best score that is clearly above the threshold, the character

which was read is assigned to the corresponding class (260). If two or more scores are “close” to each other, the result is ambiguous and further processing or manual verification (160) is required. If a clear answer is not found, the character is rejected.

- 5 While there has been shown and described the preferred embodiments of the invention, it will be appreciated that numerous modifications and adaptations of the invention will be readily apparent to those skilled in the art. It is intended to encompass all such modifications and adaptations as may come within the spirit and scope of the claims appended hereto.

What is claimed is:

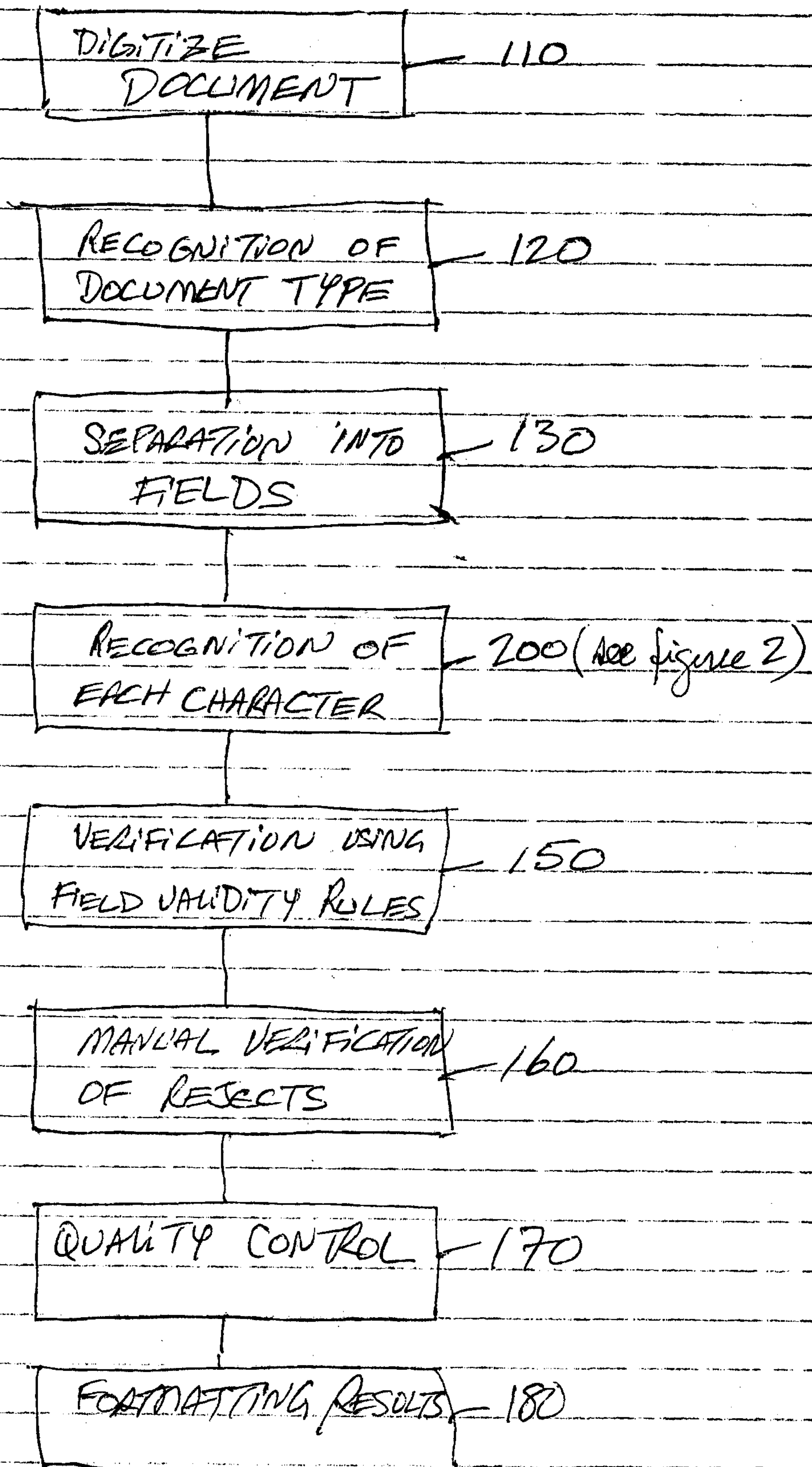
1. A system for the recognition of characters comprising:
 - a) means for generating a first bitmap image of all the characters to be recognized,
 - 5 b) means for segmenting said first bitmap image into as many second bitmap images as there are individual characters to be recognized,
 - c) means for transforming each said second bitmap image into a quantized vector,
 - d) means for automatically identifying the character represented by each said second bitmap image,
 - 10 wherein said identification means comprise elliptical class separation means.
2. The system for the recognition of characters of claim 1 wherein the identification means comprise optimal elliptical class separation means.
- 15 3. The system for the recognition of class of claim 2 wherein the characters are hand printed.
4. The system for the recognition of characters of claim 2 wherein the characters are hand written.
- 20 5. A system for the recognition of shapes comprising:
 - a) means for generating a first bitmap image of all the shapes to be recognized,
 - b) means for segmenting said first bitmap image into as many second bitmap images as there are individual shapes to be recognized,
 - 25 c) means for transforming each said second bitmap image into a quantized vector,
 - d) means for automatically identifying the shape represented by each said second bitmap image,
 - wherein said identification means comprise elliptical class separation means.
- 30 6. The system for the recognition of shapes of claim 1 wherein the identification means comprise optimal elliptical class separation means.

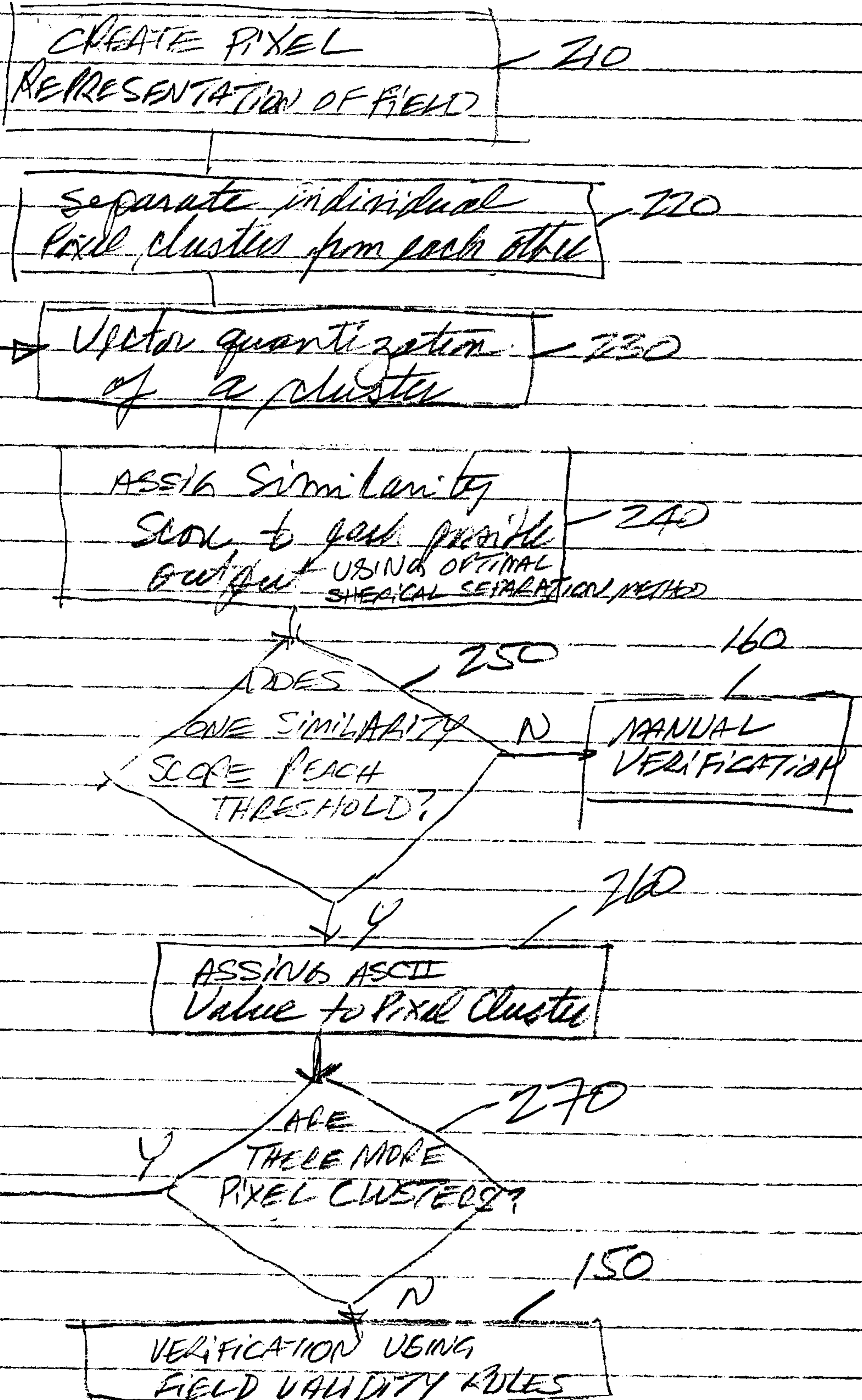
7. The system for the recognition of class of claim 6 wherein the shapes are hand printed characters.
- 5 8. The system for the recognition of shapes of claim 6 wherein the shapes are hand written characters.
9. A method for the recognition of characters on a form comprising the following steps:
 - 10 a) defining the region of the form on which the characters to be read are supposed to be located;
 - b) generating a first bitmap image of the entire said region, said bitmap image comprising a plurality of pixels;
 - c) segmenting said first bitmap image into as many second bitmap images as there are individual characters in said region;
 - 15 d) transforming each said second bitmap image into a quantized vector;
 - e) comparing each said quantized vector to pre-existing quantized vectors grouped by centers using an elliptical separation method;
 - f) assigning a similarity score between the quantized vector being compared and each pre-existing vector center;
 - 20 g) eliminating those pre-existing vectors centers for which the similarity score is lower than a pre-determined threshold;
 - h) assigning to each second bitmap image the digital character corresponding to the highest similarity score;
 - 25
10. A method as described in claim 9 wherein said separation method is an optimal elliptical separation method.
11. A method as described in claim 10 wherein said pre-existing quantized vectors are arranged in a plurality of same class clusters.
- 30 12. A method for the recognition of shapes on a form comprising the following steps:

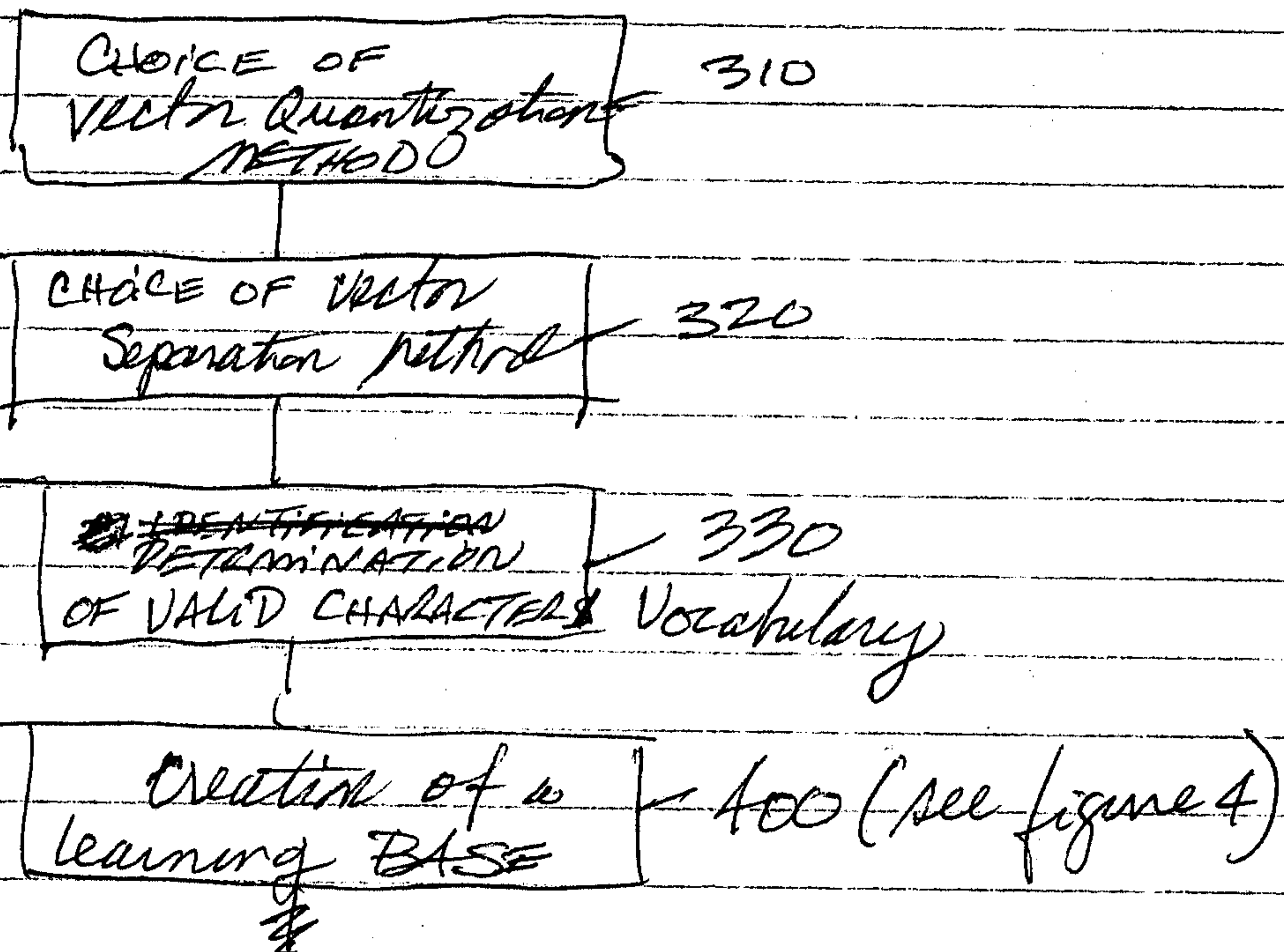
- a) defining the region of the form on which the shapes to be read are supposed to be located;
- b) generating a first bitmap image of the entire said region, said bitmap image comprising a plurality of pixels;
- 5 c) segmenting said first bitmap image into as many second bitmap images as there are individual shapes in said region;
- d) transforming each said second bitmap image into a quantized vector;
- e) comparing each said quantized vector to pre-existing quantized vectors grouped by centers using an elliptical separation method;
- 10 f) assigning a similarity score between the quantized vector being compared and each pre-existing vector center;
- g) eliminating those pre-existing vectors centers for which the similarity score is lower than a pre-determined threshold;
- h) assigning to each second bitmap image the digital shape corresponding to the
15 highest similarity score;

13. A method as described in claim 12 wherein said separation method is an optimal elliptical separation method.

14. A method as described in claim 13 wherein said pre-existing quantized vectors are arranged in a plurality of same class clusters.

FIGURE 1

FIGURE 2

FIGURE 3

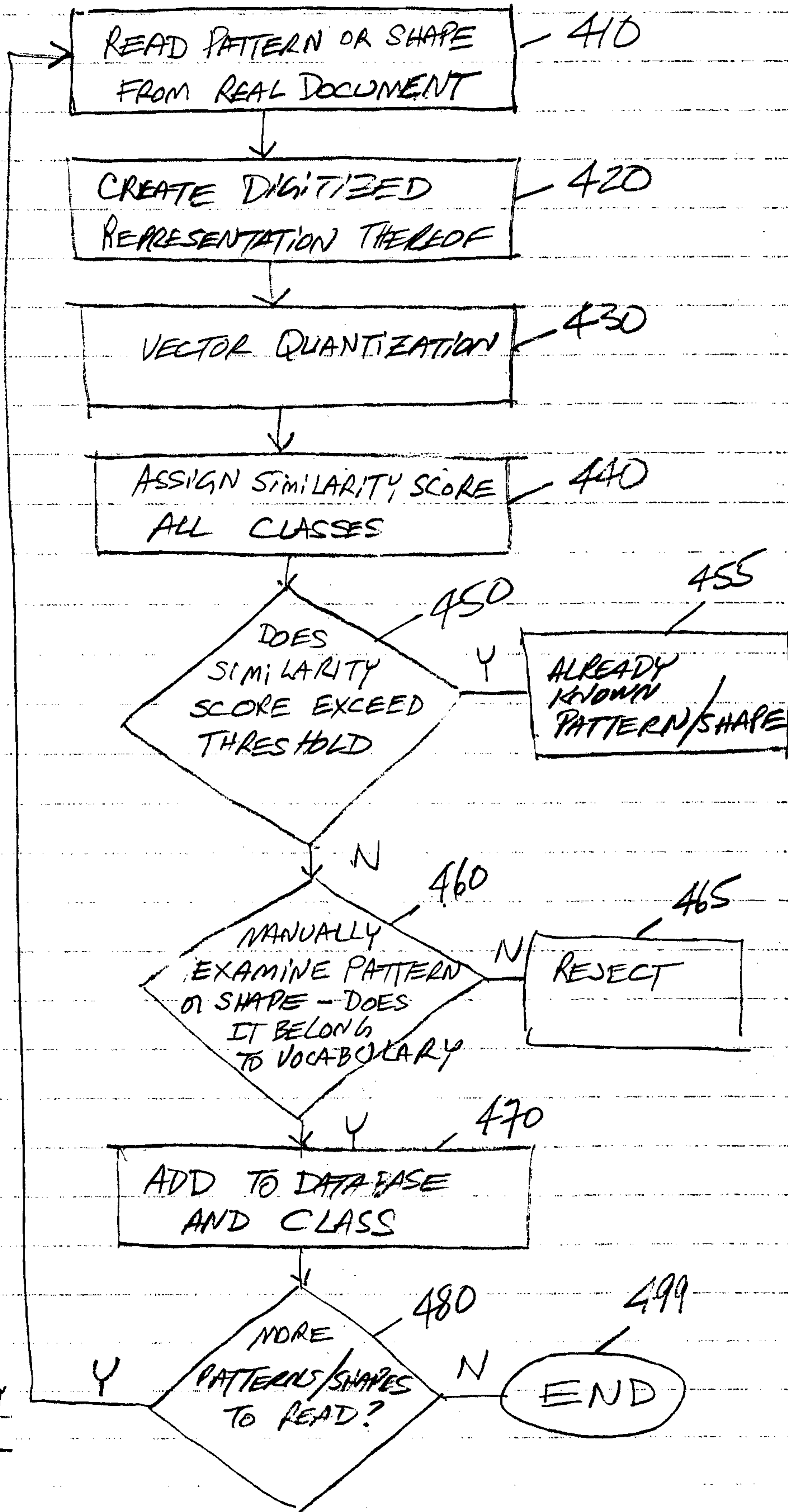
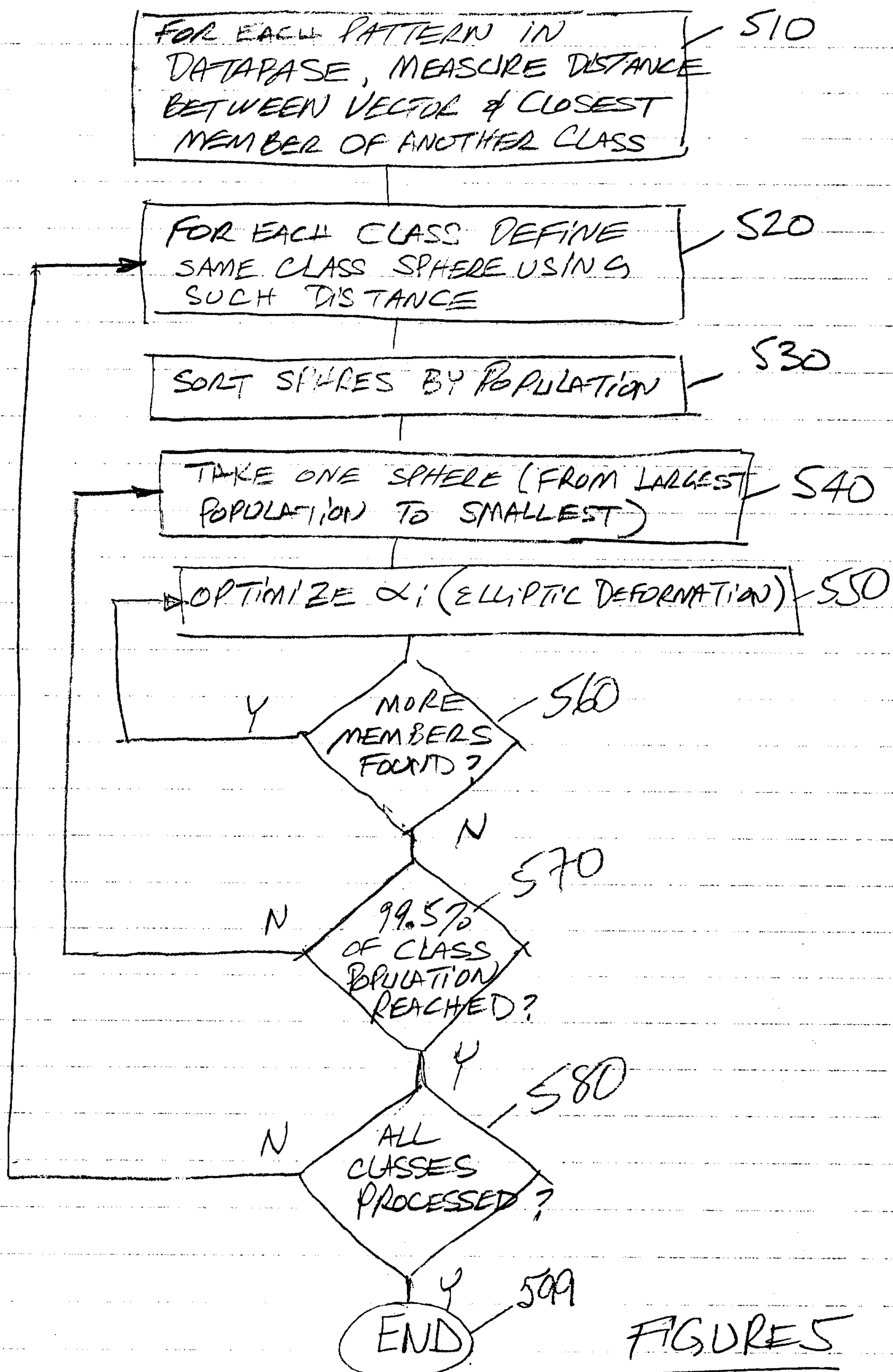


FIGURE 4



DIGITIZE
DOCUMENT

110

RECOGNITION OF
DOCUMENT TYPE

120

SEPARATION INTO
FIELDS

130

RECOGNITION OF
EACH CHARACTER

200 (See figure 2)

VERIFICATION USING
FIELD VALIDITY RULES

150

MANUAL VERIFICATION
OF REJECTS

160

QUALITY CONTROL

170

FORMATTING RESULTS

180