PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification ^{6} : G06K	A2	(11) International Publication Number: WO 97/50051 (43) International Publication Date: 31 December 1997 (31.12.97)

(21) International Application Number: PCT/IL97/00190

(22) International Filing Date: 12 June 1997 (12.06.97)

(30) Priority Data:
118642 12 June 1996 (12.06.96) IL

(71) Applicant *(for all designated States except US)*: ALIROO LTD. [IL/IL]; Trumpeldor Street 19, 44442 Kfar Sava (IL).

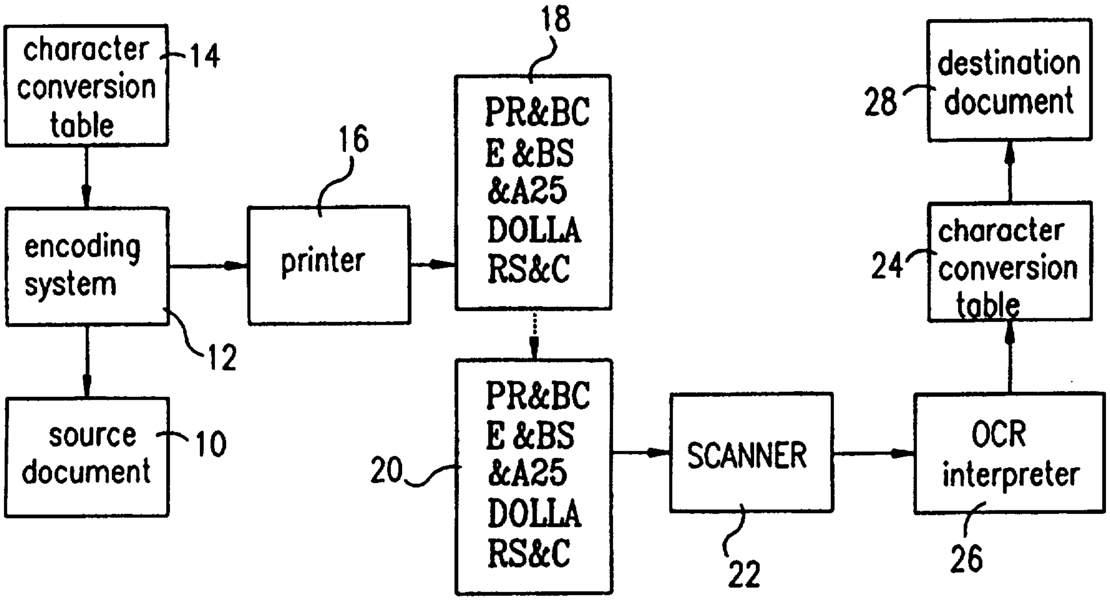
(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: POMERANTZ, Itzhak [IL/IL]; Golomb Street 18, 44357 Kfar Sava (IL). COHEN, Ram [IL/IL]; Bartenura Street 13, 62282 Tel Aviv (IL).

(74) Agents: COLB, Sanford, T. et al.; Sanford T. Colb & Co., P.O. Box 2273, 76122 Rehovot (IL).

(81) Designated States: AL, AM, AT, AT (Utility model), AU, AZ, BA, BB, BG, BR, BY, CA, CH, CN, CU, CZ, CZ (Utility model), DE, DE (Utility model), DK, DK (Utility model), EE, EE (Utility model), ES, FI, FI (Utility model), GB, GE, GH, HU, IL, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS, LT, LU, LV, MD, MG, MK, MN, MW, MX, NO, NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SK (Utility model), TJ, TM, TR, TT, UA, UG, US, UZ, VN, YU, ZW, ARIPO patent (GH, KE, LS, MW, SD, SZ, UG, ZW), Eurasian patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG).

Published
Without international search report and to be republished upon receipt of that report.

(54) Title: TEXT COMMUNICATION VIA OPTICAL CHARACTER RECOGNITION


character conversion table
14
encoding system
12
source document
10
16
printer
18
PR&BC E &BS &A25 DOLLA RS&C
20
PR&BC E &BS &A25 DOLLA RS&C
SCANNER
22
OCR interpreter
26
character conversion table
24
destination document
28

(57) Abstract

Apparatus for encoding written text including an encoder for encoding the text into an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus and apparatus for providing a hard copy output of the encoded form. An encoding method and a method and apparatus for decoding are also disclosed.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AL	Albania
AM	Armenia
AT	Austria
AU	Australia
AZ	Azerbaijan
BA	Bosnia and Herzegovina
BB	Barbados
BE	Belgium
BF	Burkina Faso
BG	Bulgaria
BJ	Benin
BR	Brazil
BY	Belarus
CA	Canada
CF	Central African Republic
CG	Congo
CH	Switzerland
CI	Côte d'Ivoire
CM	Cameroon
CN	China
CU	Cuba
CZ	Czech Republic
DE	Germany
DK	Denmark
EE	Estonia
ES	Spain
FI	Finland
FR	France
GA	Gabon
GB	United Kingdom
GE	Georgia
GH	Ghana
GN	Guinea
GR	Greece
HU	Hungary
IE	Ireland
IL	Israel
IS	Iceland
IT	Italy
JP	Japan
KE	Kenya
KG	Kyrgyzstan
KP	Democratic People's Republic of Korea
KR	Republic of Korea
KZ	Kazakstan
LC	Saint Lucia
LI	Liechtenstein
LK	Sri Lanka
LR	Liberia
LS	Lesotho
LT	Lithuania
LU	Luxembourg
LV	Latvia
MC	Monaco
MD	Republic of Moldova
MG	Madagascar
MK	The former Yugoslav Republic of Macedonia
ML	Mali
MN	Mongolia
MR	Mauritania
MW	Malawi
MX	Mexico
NE	Niger
NL	Netherlands
NO	Norway
NZ	New Zealand
PL	Poland
PT	Portugal
RO	Romania
RU	Russian Federation
SD	Sudan
SE	Sweden
SG	Singapore
SI	Slovenia
SK	Slovakia
SN	Senegal
SZ	Swaziland
TD	Chad
TG	Togo
TJ	Tajikistan
TM	Turkmenistan
TR	Turkey
TT	Trinidad and Tobago
UA	Ukraine
UG	Uganda
US	United States of America
UZ	Uzbekistan
VN	Viet Nam
YU	Yugoslavia
ZW	Zimbabwe

TEXT COMMUNICATION VIA OPTICAL CHARACTER RECOGNITION

The present invention relates to apparatus and techniques for scrambling of written text.

Apparatus and techniques have been developed for scrambling written text to prevent it from being read by unauthorized persons.

Examples of such apparatus and techniques are described in Israel patent application 115227 of the present applicant/assignee.

A difficulty has been appreciated by the present inventor in scanning scrambled written text using conventional optical character recognition (OCR) techniques. These techniques often employ statistical models of letter combinations in written language as a secondary tool for resolving ambiguities in the scanned text. Scrambled text does not follow most such statistical models.

The present invention seeks to provide apparatus and techniques for scrambling written text such as to avoid ambiguities which could make optical character recognition difficult.

In this specification and claims the terms “scrambling” and “encrypting”, and the terms “descrambling” and “decrypting” are respectively synonymous.

There is thus provided in accordance with a preferred embodiment of the present invention apparatus for encoding written text including an encoder for encoding the text into an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus and apparatus for providing a hard copy output of the encoded form.

Preferably the apparatus also includes a scanner for receiving a hard copy written text and supplying an electronic output thereof to the encoder.

Additionally in accordance with a preferred embodiment of the present invention the apparatus also includes a symbol conversion table employed by the encoder for converting symbols which are not readily distinguished from each other by OCR apparatus to symbols which are readily distinguished from each other by OCR apparatus.

Preferably, the symbols which are readily distinguished from each other by OCR apparatus consist of alias symbol sequences.

In accordance with a preferred embodiment of the present invention, the encoder also carries out scrambling of the text. For the purposes of the present specification and claims, scrambling is deemed to include encryption.

Preferably, the encoded form includes a set of symbols which includes a number of different symbols which is less than the number of different symbols contained in the non-encoded written text.

In accordance with a preferred embodiment of the present invention, the encoded form includes a set of different symbols which is a subset of a set of different symbols contained in the non-encoded written text.

Further in accordance with a preferred embodiment of the present invention the apparatus includes an error control code generator for generating error control codes for said encoded text, and an inserter for inserting said error control codes into said encoded text.

Still further in accordance with a preferred embodiment of the present invention the apparatus includes a divider for dividing said encoded text into blocks and for inserting end-of-block markers between said blocks.

There is also provided in accordance with a preferred embodiment of the present invention apparatus for decoding encoded text utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus, the apparatus including:

a text receiver for receiving a text in an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus; and a decoder for decoding the text in the encoded form and converting it into a decoded form wherein the subsets are not excluded and wherein the symbols which are not readily distinguished from each other by OCR apparatus appear.

Preferably, the receiver comprises a scanner which provides an electronic output to the decoder.

The apparatus for decoding preferably also includes apparatus for providing a hard copy output of the decoded form.

The apparatus for decoding also preferably includes a symbol conversion table employed by the decoder for converting certain symbols which are readily distinguished from each other by OCR apparatus to symbols which are not readily distinguished from each other by OCR apparatus.

Preferably, the symbols which are readily distinguished from each other by OCR apparatus comprises alias symbol sequences.

In accordance with a preferred embodiment of the present invention, the decoder also carries out unscrambling of the text.

Preferably, the encoded form includes a set of symbols which includes a number of different symbols which is less than the number of different symbols contained in the non-encoded written text.

The encoded form preferably includes a set of different symbols which is a subset of a set of different symbols contained in the non-encoded written text.

There is also provided in accordance with a preferred embodiment of the present invention a method for encoding written text including:

encoding the text into an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus; and

providing a hard copy output of the encoded form.

Further in accordance with a preferred embodiment of the present invention the method for encoding written text includes a step of embedding error correction characters into said text.

Still further in accordance with a preferred embodiment of the present invention the method for encoding written text includes a step of embedding end-of-block identifier characters at regular intervals in said text.

Preferably, the method also includes receiving a hard copy written text and supplying an electronic output thereof for encoding.

Further in accordance with a preferred embodiment of the present invention the text receiver further includes an end of block counter adapted to identify end-of block markers embedded in said text and to compare a number of characters between each said marker thus determined with a predetermined expected number of characters.

Still further in accordance with a preferred embodiment of the present invention the text receiver further includes an error correction character identifier which is operable to identify error correction characters embedded in said text and associated with given characters of said text.

Additionally in accordance with a preferred embodiment of the present invention the text receiver further includes an error identifier operable to compare any error correction characters identified by the error correction character identifier with associated characters of the text to identify whether a character recognition error has occurred.

Moreover in accordance with a preferred embodiment of the present invention the text receiver further comprises an error correction character identifier, and means for replacing textual characters on the basis of statistically likely errors until an error correction character identified by said error correction character identifier indicates that no error is present.

Further in accordance with a preferred embodiment of the present invention the text receiver further includes a means for replacing textual characters on the basis of statistically likely errors until said error identification means indicates that no error is present.

The method also preferably includes causing the encoder to employ a symbol conversion table for converting symbols which are not readily distinguished from each other by OCR apparatus to symbols which are readily distinguished from each other by OCR apparatus.

In accordance with a preferred embodiment of the present invention, the symbols which are readily distinguished from each other by OCR apparatus comprises alias symbol sequences.

The method also preferably includes scrambling of the text.

There is also provided in accordance with a preferred embodiment of the present invention a method for decoding encoded text utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus, the method including:

receiving a text in an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus; and

decoding the text in the encoded form and converting it into a decoded form wherein the subsets are not excluded and wherein the symbols which are not readily distinguished from each other by OCR apparatus appear.

The decoding method preferably also includes scanning the written text to provide an electronic output to the decoder and providing a hard copy output of the decoded form.

The decoding method also preferably includes employing a conversion table for converting certain symbols which are readily distinguished from each other by

OCR apparatus to symbols which are not readily distinguished from each other by OCR apparatus.

The decoding method additionally preferably includes unscrambling of the text.

The present invention will be understood and appreciated more fully from the following detailed description, taken in conjunction with the drawings in which:

Figs. 1A, 1B, 1C, 1D, 1E and 1F are illustrations of six examples of subsets of symbols which are not readily distinguished from each other by OCR apparatus;

Fig. 2 is a functional block diagram of apparatus for encoding and decoding written text in accordance with a preferred embodiment of the present invention;

Fig. 3 is an illustration of mapping of the individual symbols of the subset of Fig. 1A into sequences of individual symbols which are readily distinguished from each other by OCR apparatus;

Figs. 4A and 4B respectively illustrate a portion of written text, including some of the symbols of the subset of Fig. 1A, in unencoded and encoded form respectively;

Figs. 5A and 5B respectively illustrate a portion of encrypted written text, including some of the symbols of the subset of Fig. 1A, in unencoded and encoded form respectively;

Fig. 6 is a functional block diagram of a method of embedding error correction abilities into the text in accordance with an embodiment of the invention; and

Fig. 7 is a functional block diagram of a method of reading text produced in accordance with the method of fig. 6.

Reference is now made to Figs. 1A, 1B, 1C, 1D, 1E and 1F, which are illustrations of six examples of subsets of symbols which are not readily distinguished from each other by optical character recognition (OCR) apparatus. The subset of symbols of Fig. 1A, includes left and right facing bracket symbols, the number 1, an exclamation mark, the small case "l" and a capital I, which can readily be confused with each other by optical character recognition apparatus and techniques.

Similarly the symbols in each of the subsets of Figs. 1B - 1F may be confused with each other. It is appreciated that Figs. 1A - 1F are merely exemplary of such subsets and that additional subsets having similar characteristics may exist.

Reference is now made to Fig. 2, which is a functional block diagram of apparatus for encoding and decoding written text in accordance with a preferred embodiment of the present invention. A source document 10 typically comprises alphanumeric text which includes a multiplicity of symbols including at least one subset of symbols which may readily be confused with each other by optical character recognition apparatus and techniques. Examples of such subsets are illustrated in Figs. 1A - 1F.

An encoding system 12 encodes the source document 10. The encoding system may be any suitable encoding system producing a scrambled or encrypted text and is preferably an encoding system as described in applicant/assignees pending Israel Patent Application 115277, the disclosure of which is hereby incorporated by reference. This document describes a method of encrypting all or only selected portions of a document and the reader is referred in particular to pages 5 to 9 thereof. Alternatively, the encoding system may not provide scrambling or encryption but only provide symbol conversion as described hereinbelow. An encoding system of the type described is available from the applicants as Aliroo Private Suite Version 2.13.

In accordance with a preferred embodiment of the present invention, the encoding system 12 is operative to convert alphanumeric characters or other symbols from the source document into a non-ambiguous text, which is normally, but need not necessarily be, scrambled, by employing a symbol conversion table 14, such as that illustrated, for example, in Fig. 3. Symbol conversion table 14 is employed for converting alphanumeric characters or other symbols which may be ambiguous, i.e. which may readily be confused with each other by optical character recognition apparatus and techniques, into unambiguous sequences of symbols.

In accordance with a preferred embodiment of the invention, the unambiguous symbols are sequences of an infrequently occurring symbol such as an "&", hereinafter termed an "alias character" and another symbol, hereinafter collectively termed an "alias character sequence". In such a case, if the alias character actually occurs in the source document, it also is converted to a sequence of the alias character and another symbol.

For example, in the example of Fig. 3, the symbol & in the source document could be converted to "&R".

The encoding system 12 preferably outputs to a printer 16 which produces an encoded printout 18. In the illustrated embodiment, encoded printout 18,is shown to be scrambled. Alternatively this need not be the case. In accordance with a preferred embodiment of the present invention, the encoded printout 18 is formed of a set of symbols which includes a number of different symbols which is less than the number of different symbols contained in the non-encoded written text.

The encoded printout 20, or a copy thereof, may then be scanned by an OCR device, with or without having been previously transmitted by fax or any other unprotected communications medium. The OCR device typically includes a conventional scanner 22 which outputs to an OCR interpreter 26 which employs a symbol conversion table 24, which corresponds to symbol conversion table 14 and produces a destination document 28, which may be a hard copy or virtual document.

OCR interpreter 26 and conversion table 24 are operative to replace all of the alias character sequences and to verify the error correction code in the encoded printout 18 by the original characters with a high degree of accuracy in the destination document.

Operation of the apparatus of Fig. 2 in a non-encrypting and non-scrambling mode is illustrated in Figs. 4A and 4B. Figs. 4A and 4B respectively illustrate a portion of written text, including some of the symbols of the subset of Fig. 1A, in unencoded and encoded form respectively. It is seen that only ambiguous characters 30, 32, 34 and 36 are converted, the remaining characters are unchanged.

Operation of the apparatus of Fig. 2 in an encrypting or scrambling mode is illustrated in Figs. 5A and 5B. Figs. 5A and 5B respectively illustrate a portion of encrypted written text, including some of the symbols of the subset of Fig. 1A, in unencoded and encoded form respectively. It is seen that all or most of the characters are changed, but that no ambiguous characters 38, 40 and 42 remain in the encoded form.

Fig. 6 is a functional block diagram of a method of embedding error correction abilities into the text in accordance with an embodiment of the invention.

In general, as discussed above, optical character readers are known not to be able to read text with 100% accuracy. Even with replacement of the more ambiguous characters as above, total accuracy is not guaranteed. Statistical models of letter combinations cannot be used for operating on the scrambled text but for descrambling to be possible reading by the OCR has to be 100% accurate. Therefore, as shown in figure 6, error correction codes may be inserted into the written text. Many forms of error correction code are well-known to the skilled man and can easily be applied to the present circumstances with minimal experimentation. In the present embodiment the encrypted text is divided into blocks of equal length, for example 19 characters and an XOR operation is carried out between all the bytes of the block bit by bit. The resultant byte is then added to the block as the twentieth character.

In the version of the embodiment shown an end-of-block marker is added to the block as a twenty-first character. A suitable character for use as an end of block character is any character that is distinctive in appearance (to an OCR) and is not likely to appear with much frequency in the body of the text. A capital "X" is an appropriate choice.

Figure 7 shows the decryption process for the version shown in figure 6. In this process the characters as input from the OCR are first of all divided into blocks using the end of block markers. Then the characters in each block are counted. If the block length is found to be too long (a split error, one character has been misread as two) then pairs of characters are successively replaced by single characters until the error code is matched. If, on the other hand the length is found to be too short (a merge error, two characters have been read as one) then single characters are successively replaced by pairs of characters until the error code is matched. If the length is correct and the error code is still not matched by the characters (a recognition error, a character has been misread as another character) then single characters are successively replaced by single characters until the error code is matched.

The successive replacement of characters is not carried out at random. Rather a set of statistics of the most likely errors is used to replace characters by their likely alternatives. This applies both to recognition errors and to split/merge errors.

This is followed by a stage of removing the error control characters and end of block markers, at which point decryption can proceed in exactly the same way as described above.

The single byte error correction system described above will fail in the case of there being two or more errors in the block that compensate each other to match the error code. Furthermore even if multiple errors do not compensate each other and the error is consequently detected the correction algorithm of successively replacing characters will not supply the right correction. The problem can be overcome however by using a multiple byte error correction character.

It is pointed out that the error correction character is as much subject to the correction algorithm as any other character.

It will be appreciated by persons skilled in the art that the present invention is not limited by what has been particularly shown and described hereinabove. Rather the scope of the present invention is defined only by the claims which are:

CLAIMS

1. Apparatus for encoding written text comprising:

an encoder for encoding the text into an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus; and

apparatus for providing a hard copy output of the encoded form.

2. Apparatus according to claim 1 and also comprising a scanner for receiving a hard copy written text and supplying an electronic output thereof to said encoder.

3. Apparatus according to claim 1 and also comprising a symbol conversion table employed by said encoder for converting symbols which are not readily distinguished from each other by OCR apparatus to symbols which are readily distinguished from each other by OCR apparatus.

4 Apparatus according to claim 1 and wherein said encoder also carries out scrambling of said text.

5. Apparatus according to claim 1 and wherein said encoded form includes a set of symbols which includes a number of different symbols which is less than the number of different symbols contained in the non-encoded written text.

6. Apparatus according to claim 3 and wherein said encoder also carries out scrambling of said text.

7. Apparatus for decoding encoded text utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus, the apparatus comprising:

a text receiver for receiving a text in an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus; and

a decoder for decoding the text in said encoded form and converting it into a decoded form wherein said subsets are not excluded and wherein said symbols which are not readily distinguished from each other by OCR apparatus are represented by symbols which are readily distinguished from each other.

8. Apparatus according to claim 7 and wherein said symbols which are readily distinguished from each other by OCR apparatus comprises alias symbol sequences.

9. Apparatus according to claim 7 and wherein said decoder also carries out unscrambling of said text.

10. A method for encoding written text comprising:

encoding the text into an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus; and

providing a hard copy output of the encoded form.

11. A method according to claim 10 and also comprising causing said encoder to employ a symbol conversion table for converting symbols which are not readily distinguished from each other by OCR apparatus to symbols which are readily distinguished from each other by OCR apparatus.

12. A method according to claim 10 and wherein encoding also includes scrambling of said text.

13. A method for decoding encoded text utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus, the method comprising:

receiving a text in an encoded form utilizing a set of symbols which excludes subsets of symbols which are not readily distinguished from each other by OCR apparatus; and

decoding the text in said encoded form and converting it into a decoded form wherein said subsets are not excluded and wherein said symbols which are not readily distinguished from each other by OCR apparatus are represented by symbols which are readily distinguished from each other.

14. A method according to claim 13 and also including scanning the written text to provide an electronic output to said decoder.

15. A method according to claim 13 and also comprising providing a hard copy output of the decoded form.

16. A method according to claim 13 and also comprising employing a conversion table for converting certain symbols which are readily distinguished from each other by OCR apparatus to symbols which are not readily distinguished from each other by OCR apparatus.

17. A method according to claim 13 and wherein said symbols which are readily distinguished from each other by OCR apparatus comprises alias symbol sequences.

18. A method according to claim 14 and wherein said symbols which are readily distinguished from each other by OCR apparatus comprises alias symbol sequences.

19. A method according to claim 12 and wherein said decoding also includes unscrambling of said text.

20. A method according to claim 14 and wherein said decoding also carries out unscrambling of said text.

21. Apparatus according to claim 1 further comprising an error control code generator for generating error control codes for said encoded text, and an inserter for inserting said error control codes into said encoded text.

22. Apparatus according to claim 1 further comprising a divider for dividing said encoded text into blocks and for inserting end-of-block markers between said blocks.

23 Apparatus according to claim 7, wherein said text receiver further comprises an end of block counter adapted to identify end-of block markers embedded in said text and to compare a number of characters between each said marker thus determined with a predetermined expected number of characters.

24. Apparatus according to claim 7 wherein said text receiver further comprises an error correction character identifier which is operable to identify error correction characters embedded in said text and associated with given characters of said text.

25 Apparatus according to claim 24 wherein said text receiver further comprises an error identifier operable to compare any error correction characters identified by the error correction character identifier with associated characters of the text to identify whether a character recognition error has occurred.

26. Apparatus according to claim 23 wherein said text receiver further comprises an error correction character identifier, and means for replacing textual characters on the basis of statistically likely errors until an error correction character identified by said error correction character identifier indicates that no error is present.

27. Apparatus according to claim 25 wherein said text receiver further comprises a means for replacing textual characters on the basis of statistically likely errors until said error identification means indicates that no error is present.

28 A method according to claim 10 further comprising a step of embedding error correction characters into said text.

29 A method according to claim 10 further comprising a step of embedding end-of-block identifier characters at regular intervals in said text.

30 A method according to claim 12 further comprising the steps of identifying end-of-block and error control characters in said encoded and scrambled text, dividing the text into blocks in accordance with the positions of said end-of-block

characters, comparing an actual number of characters in each block with an expected number of characters in each block, determining whether characters in each block conform to an error control character associated with that block, and in the case of non-conformity with either said expected number of characters or said error control character, replacing characters on the basis of statistically likely errors until conformity is achieved.

]1!|[

FIG. 1A

S5

FIG. 1B

0Q0

FIG. 1C

Z2

FIG. 1D

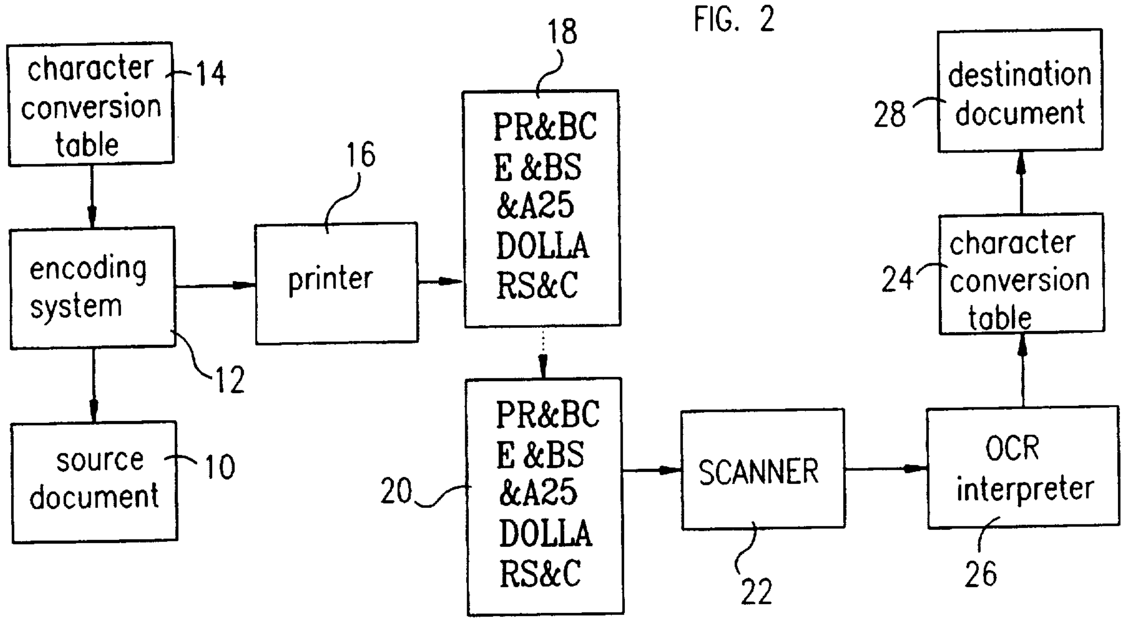
.,'

FIG. 1E

C(

FIG. 1F

FIG. 2


character conversion table
14
encoding system
12
source document
10
printer
16
18
PR&BC E &BS &A25 DOLLA RS&C
20
PR&BC E &BS &A25 DOLLA RS&C
SCANNER
22
OCR interpreter
26
character conversion table
24
destination document
28

1 → &A

| → &B

! → &C

] → &D

| → &E

[→ &F

FIG. 3

FIG. 4A

FIG. 4B

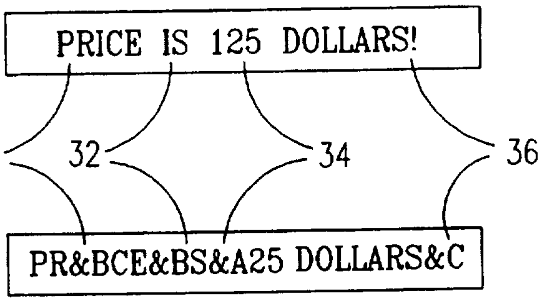
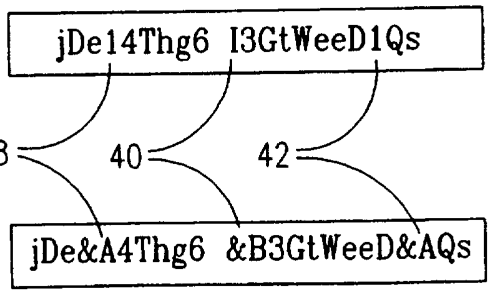


FIG. 5A

FIG. 5B



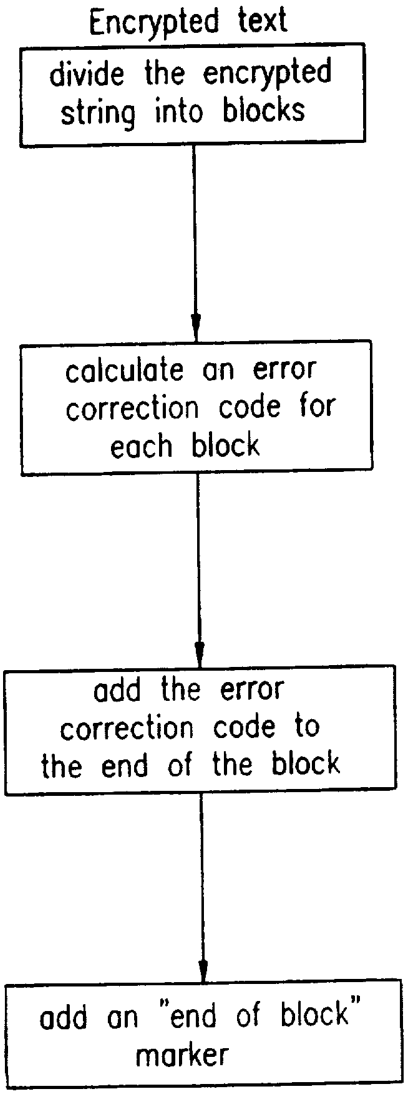

Encrypted text
divide the encrypted string into blocks
calculate an error correction code for each block
add the error correction code to the end of the block
add an "end of block" marker

FIG. 6

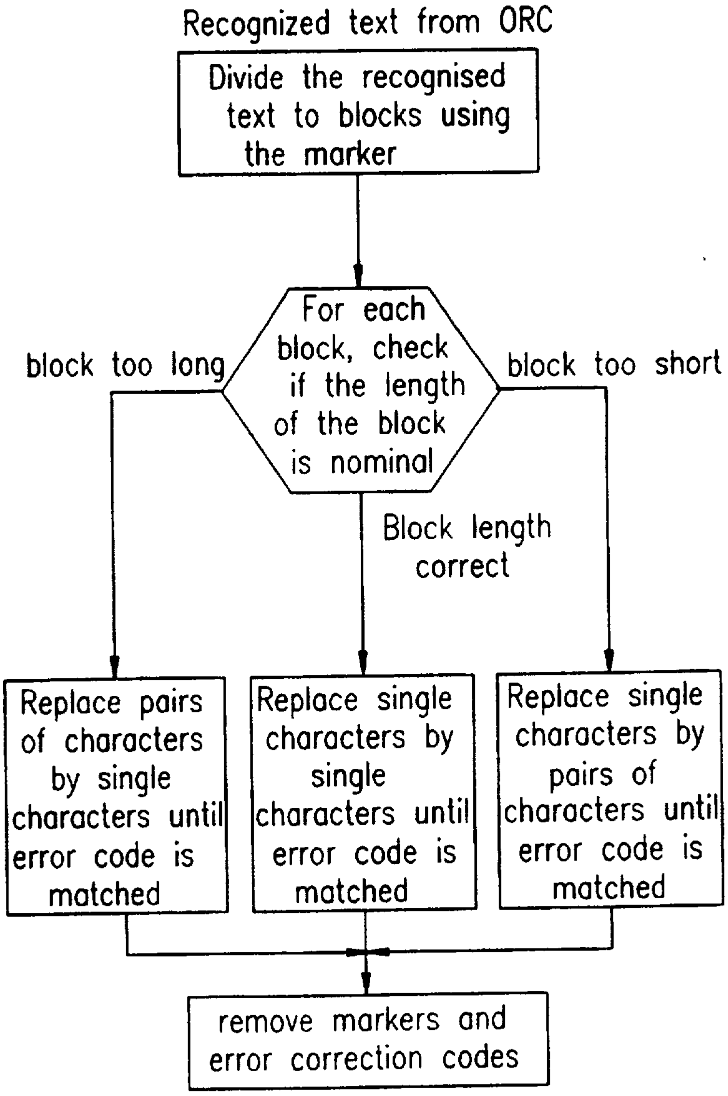

Recognized text from ORC
Divide the recognised text to blocks using the marker
For each block, check if the length of the block is nominal
block too long
block too short
Block length correct
Replace pairs of characters by single characters until error code is matched
Replace single characters by single characters until error code is matched
Replace single characters by pairs of characters until error code is matched
remove markers and error correction codes

FIG. 7