

(19) World Intellectual Property Organization
International Bureau(43) International Publication Date
11 January 2001 (11.01.2001)

PCT

(10) International Publication Number
WO 01/03307 A1(51) International Patent Classification⁷: **H03M 7/40**,
7/00, H03K 17/94, H03M 11/00, H04M 1/00, 3/00, 1/64,
11/00, G06F 13/00, 12/00, 12/02, 12/04, 12/08, G06K
9/36, 9/54, 9/60, H04L 12/28, 12/56(74) Agents: **BOSE, Romi, N. et al.**; Motorola Inc., Intellectual
Property Dept., 1303 East Algonquin Road, Schaumburg,
IL 60196 (US).

(21) International Application Number: PCT/US00/17591

(22) International Filing Date: 27 June 2000 (27.06.2000)

(25) Filing Language: English

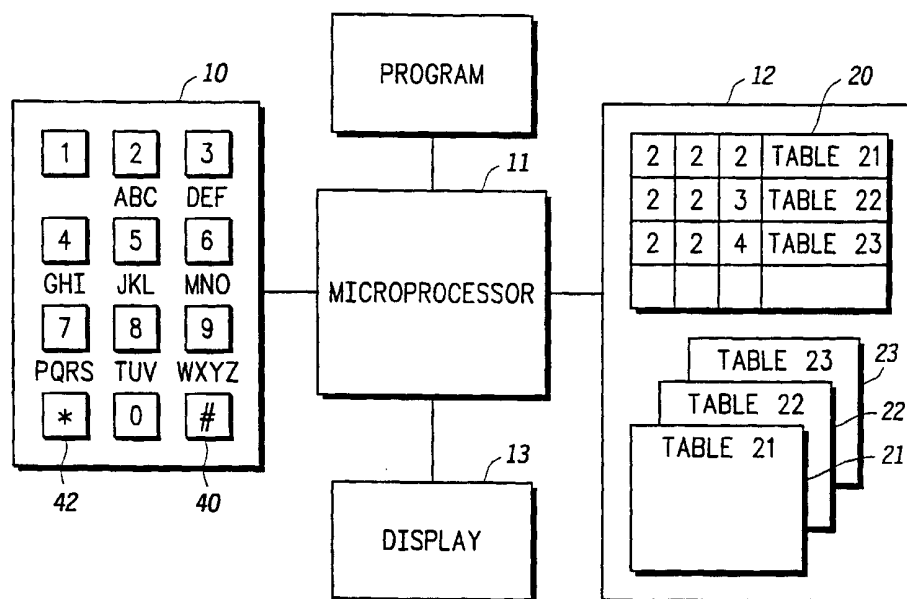
(26) Publication Language: English

(30) Priority Data:
09/350,271 6 July 1999 (06.07.1999) US(81) Designated States (*national*): AE, AG, AL, AM, AT, AU,
AZ, BA, BB, BG, BR, BY, CA, CH, CN, CR, CU, CZ, DE,
DK, DM, DZ, EE, ES, FI, GB, GD, GE, GH, GM, HR, HU,
ID, IL, IN, IS, JP, KE, KG, KP, KR, KZ, LC, LK, LR, LS,
LT, LU, LV, MA, MD, MG, MK, MN, MW, MX, MZ, NO,
NZ, PL, PT, RO, RU, SD, SE, SG, SI, SK, SL, TJ, TM, TR,
TT, TZ, UA, UG, UZ, VN, YU, ZA, ZW.(84) Designated States (*regional*): ARIPO patent (GH, GM,
KE, LS, MW, MZ, SD, SL, SZ, TZ, UG, ZW), Eurasian
patent (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European
patent (AT, BE, CH, CY, DE, DK, ES, FI, FR, GB, GR, IE,
IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG,
CI, CM, GA, GN, GW, ML, MR, NE, SN, TD, TG).(71) Applicant: **MOTOROLA INC.** [US/US]; 1303 East Al-
gonquin Road, Schaumburg, IL 60196 (US).**Published:**

— With international search report.

(72) Inventors: **GUO, Jin**; 1589 Black Hawk Drive, Sunn-
vale, CA 94087 (US). **BALAKRISHNAN, Sreeram**; 319
Missouri, San Francisco, CA 94107 (US).For two-letter codes and other abbreviations, refer to the "Guid-
ance Notes on Codes and Abbreviations" appearing at the begin-
ning of each regular issue of the PCT Gazette.

(54) Title: DISAMBIGUATION METHOD AND APPARATUS, AND DICTIONARY DATA COMPRESSION TECHNIQUES



(57) Abstract: A data input device that includes a microprocessor (11) that receives an ambiguous input from a key (10) and disambiguates the ambiguous input using a dictionary (12). The dictionary is stored in an efficient manner by being partitioned according to the first three letters of a word. A sequence of three letters in the dictionary is stored as eight bits of data, the first three letters of a word being stored using three pairs of bits, such that each pair of bits represents one of no more than four alternative letters.

DISAMBIGUATION METHOD AND APPARATUS, AND DICTIONARY DATA COMPRESSION TECHNIQUES

Field of the Invention

5

This invention relates to methods and apparatus for disambiguation of ambiguous data entry and it relates to compression of dictionary information for efficient storage in data entry devices such as handwriting recognition devices, reduced keyboard disambiguation devices, speech recognition
10 devices and the like.

Background of the Invention

In the field of data entry devices such as devices that use handwriting
15 and speech recognition and other data entry techniques, there is a need to store extensive volumes of data to assist in recognition, disambiguation or word selection and processing. In the world of mobile computing and mobile communications, memory space is very limited or is expensive and there is a need to minimize the space occupied by such data.

20 In the field of data entry, arrangements are known (for example as described in US Patent Application No. 08/754,453 of Balakrishnan, filed on 21 November 1996, assigned to the assignee of the present invention and incorporated herein by reference) in which a reduced keyboard or keypad is used for character entry where each key ambiguously represents more than
25 one character and disambiguation software is used to disambiguate a key entry to identify the probable intended key from the various ambiguous

possibilities. In such a scheme, dictionary, word or n-gram data is necessary to perform the disambiguation. Large amounts of data are required to enable satisfactory disambiguation.

Data compression techniques exist for purposes such as bulk data storage. An example is gzip compression, which is suitable for compression of alphabetical text, and is explained here by way of background. The Roman alphabet comprises 26 letters a through z, which can readily be represented as a byte of eight bits of data. Eight bits of data allow one bit for a start-of-word indicator and 128 characters (2^7). Accordingly, such a scheme has 102 unused byte values (unused in the sense of being unnecessary for coding of 26 characters). In Gzip compression, the additional 102 byte values are used to encode character pairs. By way of example if it assumed that bits 0-25 are used for a through z, value 26 can be assigned to mean "ba", value 27 to mean "ca" etc., using 102 character pairs selected as the most common character pairs in the language in question (e.g. American English).

There is a need for an improved method of storage of dictionary or other data suitable for data entry disambiguation.

20 Brief Description of the Drawings

FIG.1 is a block diagram illustrating a data entry device according to the invention, having a compressed dictionary.

FIG. 2 is a flow diagram illustrating a computer program stored in memory in the device of FIG. 1 and executed on the microprocessor of FIG. 1.

Detailed Description of the Drawings

Referring to FIG. 1, a data entry device is shown having a reduced keyboard or keypad 10 having twelve keys. Ten of the keys have the digits 0-9 marked thereon. Keys 2-9 also have the letters a-z marked on the keys or beside the keys, such that key 2 represents letter a, b, or c and key 3 represents letter d, e, or f, etc. in a manner that is standard for telephones in the United States. Other layouts for the keypad could be used, with different allocations of letters to keys, for example as described in "Probabilistic Character Disambiguation for Reduced Keyboards Using Small Text Samples" by J.L. Arnott and M. Y. Javed, *Augmentative and Alternative Communication*, Vol. 8, September 1992.

The keypad 10 is coupled to a microprocessor 11, which is coupled to a memory 12 (e.g. a read-only memory) and a display 13 in a manner known in the art. A program memory 14 is also shown coupled to the microprocessor 11 for completeness, although this can be the same memory as memory 12.

In operation, the microprocessor 11 receives an ambiguous input from a key 2-9 and calculates in a manner described below a word start n-gram probability from accumulated probabilities for words in the memory 12 beginning with the ambiguous input. The microprocessor disambiguates the ambiguous input using the word start n-gram probability so calculated.

Memory 12 is illustrated as having a look-up table 20 and a series (up to 512) of dictionary tables of which three are shown, labeled 21, 22 and 23. These will be described in greater detail.

Look-up table 20 has four columns and 512 rows (or vice-versa). The first three elements of a row are three digits that are possible key entries from keypad 10. The fourth element of a row is a number identifying a dictionary table from among the dictionary tables 21, 22 and 23 etc.

5 The dictionary tables 21, 22, 23 etc. together form a complete dictionary of words including common proper nouns and acronyms partitioned into sub-dictionaries (collectively referred to as "words"). The dictionary is partitioned according to the first three letters of each dictionary entry (word). Thus, all words beginning with a/b/c followed by a/b/c, followed
10 by a/b/c appear in table 21 and all words beginning with a/b/c followed by a/b/c, followed by d/e/f appear in table 22, etc. Of course there may be combinations of letters that have no dictionary entries, in which case there will be fewer than 512 tables (for example there are no English language words beginning with w/x/y/z followed by w/x/y/z, followed by w/x/y/z and
15 there may be no table for this combination, unless acronyms such as "www" are included).

By partitioning the dictionary in this manner, it is possible to more compactly store representations of words in the various tables. In each table, the first three characters of a word are represented by three pairs of
20 bits. Thus, for example, in table 21, the first three letters of any word in the table are exclusively "a" or "b" or "c". In table 21, "a" is represented as 00, "b" is represented as 01, and "c" is represented as 00. An entry of "11" would be an error and would give an error indication or some other action such as no result.

25 In the tables 21-23 etc., the first three letters of a word are represented by a single byte of eight bits. The first bit is used as a word-

start indicator (as in gzip compression), the second bit is used to indicate that the first letter is capitalized and the remaining six bits represent the three letters, where the actual interpretation of the bits into ultimate letters depends on the table being referenced. There now follows some examples

5 for entries for tables 21 and 22. These examples give a miscellaneous selection of words from each table and are not necessarily complete. In these examples, the first three letters of each word are shown encoded as a byte and the remaining letters are shown unencoded. Any letters of a word in excess of three are coded using ordinary gzip compression already

10 described.

Table 21 - words beginning with {a/b/c, a/b/c, a/b/c}

aback	10 10 01 10 ck
abbot	10 10 01 01 ott
15 babble	10 01 10 01 ble
back	10 01 10 00 k
cab	10 00 10 01
cable	10 00 10 01 le

20 Table 2 - words beginning with {a/b/c, a/b/c, d/e/f}

bad	10 01 10 10
Badminton	11 01 10 10 minton
baffle	10 01 10 00 fle
caddie	10 00 10 10 die
25 cafe	10 00 10 00 e
caffeine	10 00 10 00 feine

Note that the first digit of each word is "1" indicating start-of-word. This may appear redundant, but it should be noted that the letters to the left of each table entry will be encoded in further bytes, the first of which will begin "0". Note also that the second digit of the word "Badminton" is "1" indicating that the word is capitalized. In each entry, the second, third and fourth pairs of bits represent the first, second and third letters of the encoded word. It is not possible to decode a word from a table without reference to the table in which the word is stored. For each pair of bits in the second, third and fourth pairs of a byte, the selected correspondence between the bit-pair and the selection from among the ambiguous letters represented thereby is not critical, so long as the rules for determining that correspondence are fixed. Thus, in table 22, in the third position, "10" represents "d", "01" represents "e", "00" represents "f" and "11" would be an error state. Similar correspondences are selected for other letters g-z.

From these examples, it can be seen that a three-letter word such as "cab" or "bad" can be encoded in a single byte. A four-letter word such as "back" can be encoded with just two bytes, where the second byte is a gzip representation of "k" (e.g. decimal 10 represented in binary, where k is the eleventh letter of the alphabet and a-z are represented as decimal 0-25). A five letter word can also be represented in just two bytes if the fourth and fifth letter pair is one of the 102 most-common two-letter combinations representable as decimal 26-127.

One-letter and two-letter words are treated as special cases and supported with an extra one-letter table (8 rows each for key 2-9, and 3 to 4 columns for letters under the key) and two-letter tables (8x8x5, for up to 5 words under each two-key combinations).

This level of dictionary compression is an improvement over the prior art, in which at least two bytes were necessary to encode three letters and at least three bytes were necessary to encode five letters. In fact, the average length of a word in the English language is just five letters, so compression of the first three letters gives a substantial saving in space in dictionary storage. Simulations have shown that a 20% compression gain can be achieved using this method. There is an overhead cost in terms of the need to store table 20, but this is a relatively small table and the measured gain of 20% is achieved after including table 20.

Operation of the device of FIG. 1 will now be described with reference to the flow diagram of FIG. 2.

In operation, the user strikes a key of the keypad 10 to enter a letter (step 100). The microprocessor 11 performs a look-up operation in table 20 to determine which of the tables 21, 22, 23 etc.. are valid for the key or keys so far entered. In the case where only one key has so far been entered, there will be 64 valid tables. The microprocessor 11 identifies these tables in step 102. Following identification of the tables, the microprocessor performs a look-up in the tables (step 104) to identify a complete set of possible words that include the word being entered by the user. In step 106, the microprocessor 11 uses language model data to identify one or more of the possible words that is the most likely candidate for the key sequence being entered.

In step 106, the most likely entered key is preferably identified by the order of storage of the words in the dictionary. It is preferred that the words in the dictionary are stored in order of decreasing frequency of usage in the language in question, with the most common words being stored at lower

order addresses and least common words being stored at higher order addresses (or vice-versa).

In this manner, where there are three ambiguous possibilities for a given letter, (e.g. "a", "b" and "c") the letter identified first in the words of the dictionary (in word address order) is selected as the most likely desired
5 result. This letter is present in the display 13 to the user. The second and third most likely candidates (and the forth if relevant) can also be displayed.

This method is particularly useful for entry of Chinese characters. The characters are stored in frequency order as stroke sequences.
10 (Incidentally, a character can be stored and represented in more than one possible stroke order.) When a stroke is entered (i.e. a classification of stroke or an ambiguous selection of one of a plurality of strokes in a set) at a position in a sequence of entered strokes, the first character in the dictionary that is formed using the entered stroke (at that position in a sequence of
15 strokes) appears foremost as the most likely desired result. Other characters that are formed using the same stroke at the same position in the same sequence appear lower down the order.

Thus, in a single device, the same search software is used for English and Chinese, and the dictionary is selected by the user according to the
20 language desired. The dictionary of words is a dictionary of Roman alphabet words and is interchangeable in the device with a dictionary of ideographic characters (e.g. Chinese or Japanese characters). The two dictionaries are searched by a common search engine (search software) using the technique described.

25 There are alternative ways of providing language model data. One alternative is to use n-gram data as described in the paper by Arnott

described above. Another alternative is to provide an absolute score for each word in the dictionary and to display an entire word in the display 13, where the first choice is the word that has the highest score (i.e. is the most common of the words as used in the language in question).

- 5 The dictionary is preferably exclusively a dictionary of words, but may include acronyms and common misspellings of words and may include word stems. Also it may include common letter combinations that are not word stems but are found in non-dictionary words (e.g. found in proper nouns). Accordingly, in the preferred embodiment, the n-gram probabilities
- 10 calculated are all or substantially all word-start n-grams.

- Returning to FIG. 2, in step 108, a determination is made as to whether a space key has been pressed. The space key can be any key otherwise not assigned (for example key 40 of FIG. 1). The space key indicates that a complete word has been entered or that the currently
- 15 displayed word is the desired result. If the space key is pressed, the desired result is output at step 110. In step 110, the selected or first displayed word or acronym is transferred into an output buffer or text display area of the display 13. Following step 110, the entry of a complete word is completed and the program returns step 112 for commencement of the next word. If a
- 20 space is not yet entered, the program determines at step 114 whether some other key has been pressed, for example for correcting or editing. If, for example, key 42 has been pressed, an editing mode can be entered at step 116. In the editing mode, previously entered digits can be counseled or corrected and other functions can be carried out. Following editing, the
- 25 program returns at step 118 for the next key entry. If, following step 114, no space key and no other key has been pressed, the process returns to step

100 for entry of the next key from keys 2-9. Steps 100-106 are repeated for a second digit and a third digit, i.e. second letter and a third letter. After a third letter has been entered, step 102 can be skipped, because after entry of a third letter there will be only one remaining table that needs to be
5 searched for the desired word. After entry of the third key, further disambiguation relies on the dictionary bytes after the first byte, i.e. dictionary bytes that represent the remainder of each dictionary word beyond the first three letters where the remainder of the word is encoded using gzip compression.

10 Accordingly, a method and apparatus for data entry has been described in which entered data is disambiguated using a dictionary that is partitioned and where the partitioning of the dictionary is dependant on at least the first letter of each dictionary word and preferably the first three letters of each dictionary word and wherein the compressed dictionary has
15 pairs of bits, where a pair of bits represents one of no more than four alternative letters of the alternative letters being represented by a pair of bits being letters that are ambiguously represented by a digit in a reduced key pad that has multiple keys, where each key represents more than one character. The invention can be applied to stroke-based character entry,
20 such as described in U.S. Patent Application 09/220,308 of Guo filed on December 23, 1998, which is assigned to the assignee of the present invention and is hereby incorporated by reference. For these purposes, the term "letter" encompasses a stroke of an ideographic character, and the term "word" encompasses an ideographic character made up of strokes.

25

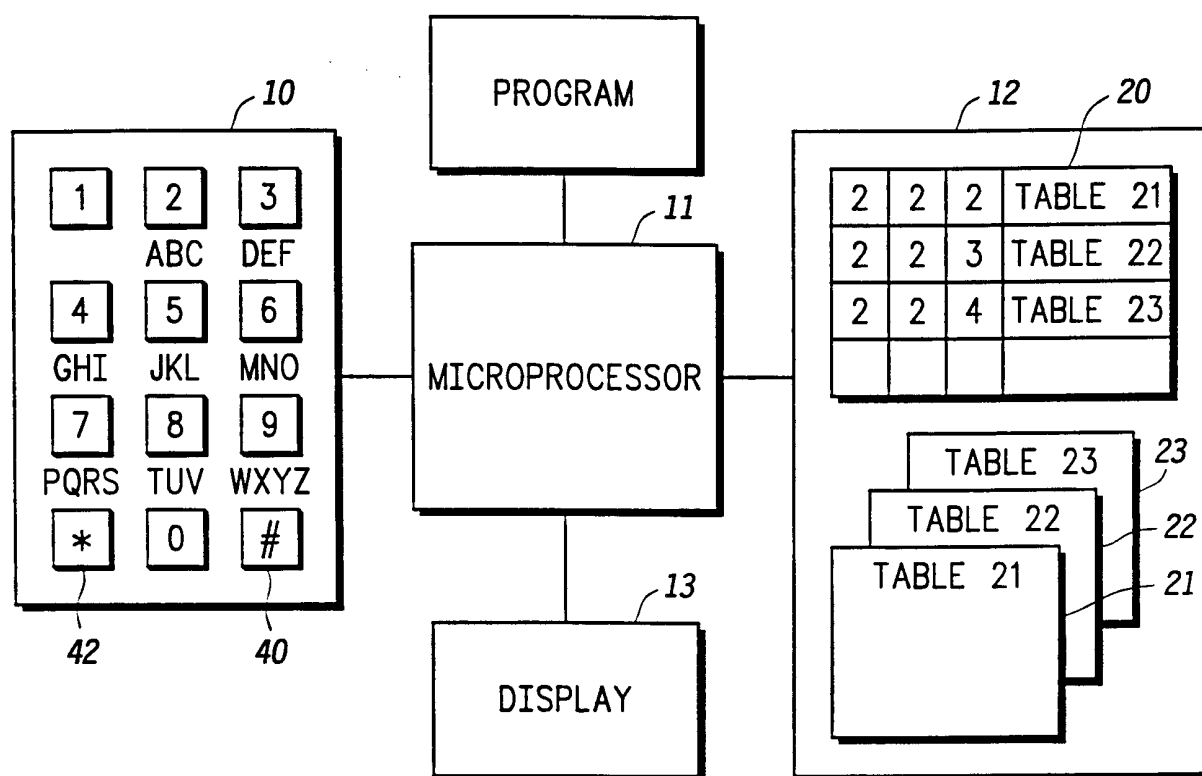
Claims

1. A method of compression of data comprising:
storing a sequence of three letters as eight bits of data, wherein the
5 first three letters of a word are stored using three pairs of bits, such that
each pair of bits represents one of no more than four alternative letters,
wherein the data is partitioned according to the first three letters of a word.
2. The method of claim 1, wherein, of each set of eight bits representing
10 three letters, one bit is reserved for a word start flag.
3. A data input device comprising:
a plurality of inputs, each representative of a letter from an ambiguous
set of letters;
15 a dictionary (21, 22, 23) of words, wherein the dictionary is partitioned
according to the first three letters of a word, and wherein a sequence of
three letters in the dictionary is stored as eight bits of data, the first three
letters of a word being stored using three pairs of bits, such that each pair of
bits represents one of no more than four alternative letters.
20
4. The device of claim 3, further comprising a one-letter table of one-
letter words.
5. The device of claim 3, further comprising a two-letter table of two-
25 letter words.

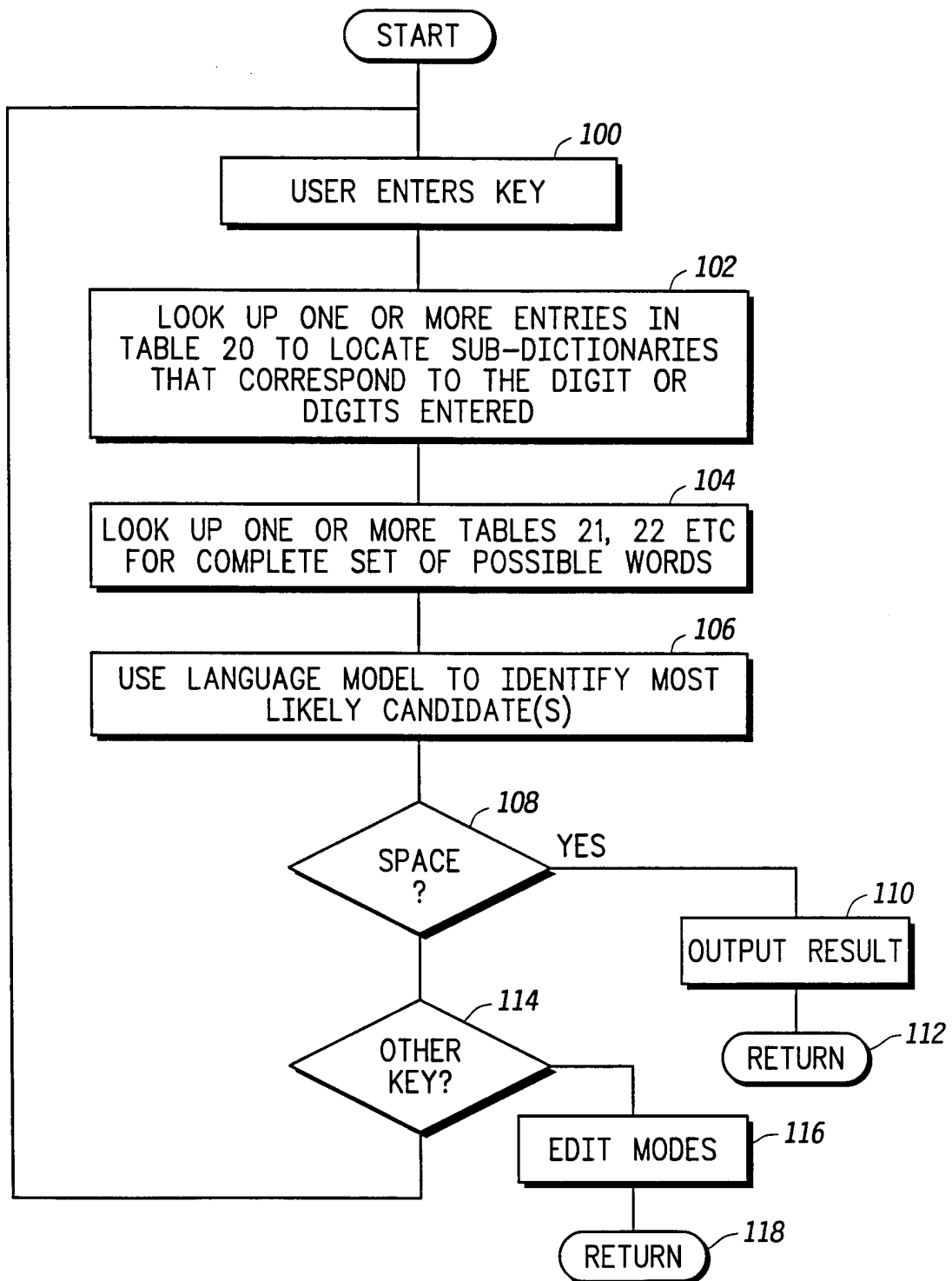
6. The device of claim 3, wherein the dictionary (21, 22, 23) of words is a dictionary of Roman alphabet words and is interchangeable in the device with a dictionary of ideographic characters searchable by a common search engine.

5

1/2

*FIG. 1*

2/2

*FIG. 2*

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/17591

A. CLASSIFICATION OF SUBJECT MATTER

IPC(7) :Please See Extra Sheet.

US CL :Please See Extra Sheet.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 341/67, 23, 28, 106; 379/367, 74, 52, 93; 345/326; 382/232, 305; 370/395; 711/221; 375/240.23

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

EAST

search terms: memory, compress, compression, compressed, lookup, table

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,786,776 A (KISAICHI et al) 28 July 1998, entire document	1-6
A	US 4,817,129 A (RISKIN) 28 March 1989, ALL	1-6
A	US 5,200,988 A (RISKIN) 06 April 1993, ALL	1-6
A	US 5,253,053 A (CHU et al) 12 October 1993, ALL	1-6
A	US 5,488,366 A (WU) 30 January 1996, ALL	1-6
A	US 5,101,487 A (ZALENSKI) 31 March 1992, ALL	1-6
A	US 5,589,829 A (ASTLE) 31 December 1996, ALL	1-6



Further documents are listed in the continuation of Box C.



See patent family annex.

* Special categories of cited documents:		*T*	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
A	document defining the general state of the art which is not considered to be of particular relevance	*X*	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
E	earlier document published on or after the international filing date	*Y*	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
L	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	*&*	document member of the same patent family
O	document referring to an oral disclosure, use, exhibition or other means		
P	document published prior to the international filing date but later than the priority date claimed		

Date of the actual completion of the international search 11 AUGUST 2000	Date of mailing of the international search report 29 AUG 2000
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231 Facsimile No. (703) 305-3230	Authorized officer ALBERT WONG Telephone No. (703) 305-8884

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/17591

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 5,128,672 A (KAEHLER) 07 July 1992, ALL	1-6

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US00/17591

A. CLASSIFICATION OF SUBJECT MATTER:

IPC (7):

H03M 7/40, 7/00; H03K 17/94; H03M 11/00; H04M 1/00, 3/00, 1/64, 11/00; G06F 13/00, 12/00, 12/02, 12/04, 12/08;
G06K 9/36, 9/54, 9/60; H04L 12/28, 12/56

A. CLASSIFICATION OF SUBJECT MATTER:

US CL :

341/67, 23, 28, 106; 379/367, 74, 52, 93; 345/326; 382/232, 305; 370/395; 711/221; 375/240.23