

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
23 October 2008 (23.10.2008)

PCT

(10) International Publication Number
WO 2008/127443 A1

(51) International Patent Classification:
G06K 9/20 (2006.01)

Kirvin [US/US]; c/o Bank of America Corporation, 101
S. Tryon Street, Charlotte, NC 28255 (US).

(21) International Application Number:
PCT/US2007/085949

(74) Agents: **PATEL, Aseet** et al.; Banner & Witcoff, Ltd.,
10 S. Wacker Drive, Suite 3000, Chicago, IL 60606-7407
(US).

(22) International Filing Date:
29 November 2007 (29.11.2007)

(81) Designated States (unless otherwise indicated, for every
kind of national protection available): AE, AG, AL, AM,
AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH,
CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG,
ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL,
IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK,
LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW,
MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL,
PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY,
TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA,
ZM, ZW.

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/911,666 13 April 2007 (13.04.2007) US

(71) Applicant (for all designated States except US): **BANK
OF AMERICA CORPORATION** [US/US]; 101 S. Tryon
Street, Charlotte, NC 28255 (US).

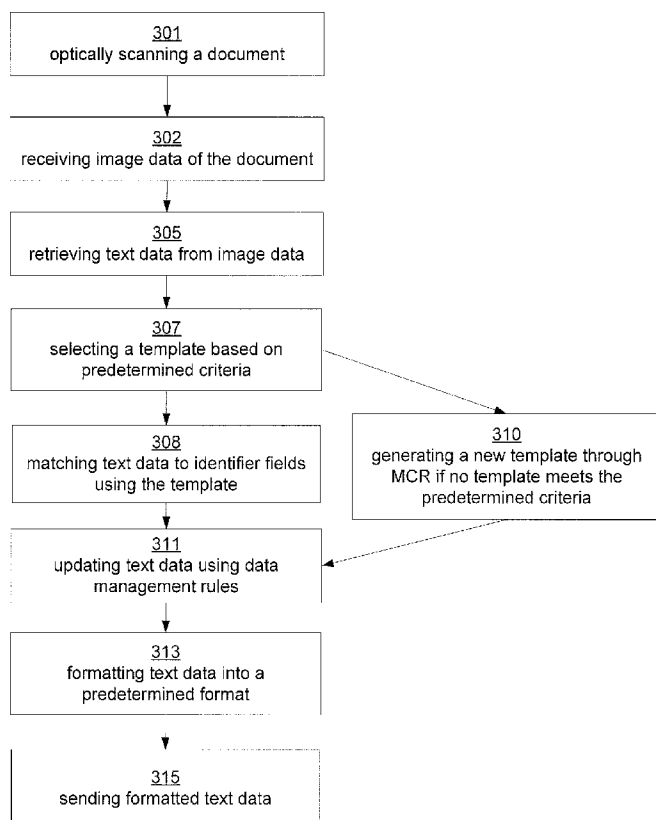
(84) Designated States (unless otherwise indicated, for every
kind of regional protection available): ARIPO (BW, GH,
GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM,

(72) Inventor; and

(75) Inventor/Applicant (for US only): **FLOYD, George**

[Continued on next page]

(54) Title: IMAGE DATA EXTRACTION AUTOMATION PROCESS



(57) Abstract: Healthcare providers, insurance providers, patients, others involved in the administration of healthcare services, and numerous others may obtain enhanced efficiency and benefit from automated and semi-automated processing of paper and/or optical documents. One aspect of the processing include a cognitive learn process for use in retrieving data from documents using templates. The templates may identify a mapping of textual data on a document with identifier fields, such as account number, patient name, insurance coverage, and others. Another aspect of the processing may include updating the textual data using data management rules. Yet another aspect of the processing may include accessing a remote data repository to enhance the accuracy of textual data retrieved using optical character recognition of image data.

FIG. 3

WO 2008/127443 A1



ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM),
European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI,
FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL,
PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM,
GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

— *as to the applicant's entitlement to claim the priority of the
earlier application (Rule 4.17(iii))*

Published:

— *with international search report*

Declarations under Rule 4.17:

— *as to applicant's entitlement to apply for and be granted a
patent (Rule 4.17(ii))*

IMAGE DATA EXTRACTION AUTOMATION PROCESS**RELATED APPLICATION**

- [01]** This application claims priority to U.S. Provisional Application No. 60/911,666 entitled "Cognitive Learn Process in Data Retrieval" which was filed on April 13, 2007, and which is herein incorporated by reference in its entirety.
- [02]** This application is related to U.S. Provisional Application No. 60/972,752, which was filed on September 15, 2007.

FIELD OF THE INVENTION

- [03]** Aspects of the disclosure relate to the automated and semi-automated processing of paper and/or optical documents. More specifically, aspects of the disclosure relate to techniques for use in retrieving data from documents.

BACKGROUND

- [04]** In the healthcare industry an explanation of benefits (EOB) document is commonly provided when a payer submits funds to a payee. The EOB document can be a detailed document with a plurality of line items with sub-line items under them (and so on) itemizing the various components of a bill. In the past, funds were manually matched up to their corresponding line items in an EOB document. As one can imagine, this process was slow and tedious. In an effort to improve this process, some processing facilities used optical character recognition (OCR) technology to optically scan images of EOB documents. The data from the image file of the EOB document could be extracted and assigned to the appropriate line items and corresponding funds.
- [05]** Although optical character recognition (OCR) technology is known in the art, there is a need in the art for enhancements to OCR capabilities that permit more accurate and automated processing of EOB documents. Generally, there exists a need for enhancements to OCR and OCR-related technology and tools to enhance the processing of OCR image files.

BRIEF SUMMARY

- [06] Aspects of the present disclosure address one or more of the issues mentioned above by disclosing systems, devices, and methods for a cognitive learn process in data retrieval. The following presents a simplified summary of the disclosure in order to provide a basic understanding of some aspects. It is not intended to identify key or critical elements of the invention or to delineate the scope of the invention. The following summary merely presents some concepts of the disclosure in a simplified form as a prelude to the more detailed description provided below.
- [07] In one embodiment in accordance with various aspects of the invention, a method is disclosed for receiving image data; retrieving (i.e., extracting) text data from the image data; selecting a template based on at least one predetermined criteria; using the template to match the text data to identifier fields; and formatting and sending the text data. The predetermined criteria may be based on the source of the image data (e.g., check, explanation of benefits document, etc.) and/or based on the location of particular data on the image retrieved.
- [08] In various embodiments, the text data retrieved from the image data may be updated using data management rules. For example, letters in the text data at particular locations in the image data may be converted to numbers depending on the template and identifier fields. In another example, a remote data repository may be accessed to compare the accuracy of the text data against related documents.
- [09] Furthermore, a tangible computer-readable medium storing computer-executable instructions for causing a processor to perform various steps of methods disclosed herein. Such a processor may be located in a system performing one or more aspects of the invention. For example, a system is disclosed comprising an optical character recognition component, a data storage unit, a template selection component, a data retrieval component, and scanner component. Such a system may access a remote data repository to enhance the accuracy of character data retrieved through optical character recognition. These and other embodiments of the invention will become apparent to one of ordinary skill in the art after review of the entirety disclosed herein.

BRIEF DESCRIPTION OF THE DRAWINGS

- [10] The present disclosure is illustrated by way of example and not limited in the accompanying figures in which like reference numerals indicate similar elements and in which:
- [11] Figure 1 depicts an illustrative personal computing device with peripheral devices in accordance with various aspects of the invention;
- [12] Figure 2 shows an illustrative operating environment for establishing communication between an image acquisition device and a remote computer (e.g., banking website server computer) in accordance with various aspects of the invention; and
- [13] Figure 3 shows a flowchart illustrating a method in accordance with various aspects of the invention.

DETAILED DESCRIPTION

- [14] Healthcare providers, insurance providers, patients, others involved in the administration of healthcare services, and anyone that processes documents or electronic images of documents may benefit from one or more aspects of the embodiments disclosed herein. The features of the illustrative embodiments described herein contemplate additional other embodiments comprising one or more, or a combination thereof, of the aspects described throughout.
- [15] An example of an illustrative personal computing system 100 in which various aspects and embodiments of the invention may be implemented is shown in the simplified diagram in Figure 1. The features of such a device are well-known to those of skill in the art and need not be described at length here. The illustrative system 100 is only one example of a suitable system and is not intended to suggest any limitation as to the scope of use or functionality of the invention. Suitable computing environments for use with the invention include an image processing device 102 or system that support interaction with an input devices 122 (e.g., digital camera 128, document scanner 124, multi-function office device 126, etc.), output devices 118 (e.g., visual display 120), and communication connections 130 (e.g., Ethernet connection, IEEE 802.11, dial-up

connection, etc.). The communication connections 130 may be used to allow the image processing device 102 to communicate with other devices. With reference to Figure 1, an image processing device 102 commonly includes a memory 106 and a processor 104 (e.g., an Intel microprocessor).

- [16] Programs, comprising sets of instructions and associated data, may be stored in the memory 106, from which they can be retrieved and executed by the processing unit 103. Among the programs and program modules stored in the memory 106 are those that comprise or are associated with an operating system 110 as well as application programs 112. Application programs 112, such as a web browser application, Java runtime environment, and others, and an operating system 110 are commonly installed in an image processing device 102. The memory 106 may also include a cache 106 to enhance device performance. Computing system 100 includes forms of computer-readable media. Computer-readable media include any available media that can be accessed by the image processing device 102. Computer-readable media may comprise storage media and communication media. Storage media include volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information such as computer-readable instructions, object code, data structures, program modules, or other data. Communication media include any information delivery media and typically embody data in a modulated data signal such as a carrier wave or other transport mechanism.
- [17] In accordance with various aspects of the invention, illustrated in Figure 2 is a personal computing device 220 in communication with an image processing device 102 over a network 210 (e.g., Internet, LAN, WAN, wired network, wireless network). Figure 2 depicts a simplified, illustrative operating environment for implementing various aspects and embodiments of the invention. The illustrative operating environment in figure 2 is only one example of a suitable operating scenario and is not intended to suggest any limitation as to the scope of use or functionality of the invention.
- [18] The personal computing device 220 may be a conventional personal computer with computer software for performing manual character recognition (MCR). One skilled in the art will appreciate that in some embodiments where the personal computing device 220 is a handheld device (e.g., a personal digital assistant such as a Treo or Blackberry),

a thin-client application is contemplated for use with the invention to enable MCR. As explained below, MCR may be used to develop and maintain templates associated with the processing and performing OCR on image files corresponding to documents (step 310).

- [19] The image processing device 102 may communicate over a network 210 with various devices and/or systems. For example, the image processing device 102 may communicate with a digital camera (reference 218, 128), multi-function office device (reference 216, 126), and/or document scanner 124. An image processing device 102 comprising a memory and a processing unit may be used to receive an image of a document such as, for example, a check or an EOB (explanation of benefits) data sheet. The image processing device's memory may store computer-executable instructions. In various embodiments of the invention, the image processing device 102 may be a farm of computers/servers for large-scale image processing. One skilled in the art will appreciate that the image processing functionalities may required extensive processing power that a farm of servers may be ideally equipped to handle.
- [20] The memory 106 of the image processing device 102 may also include computer-executable instructions that may be executed using a processing unit (e.g., processor 104) in the image processing device. The computer-executable instructions may be used to perform a method. The method may include the steps of retrieving templates based on, among other things, the source of the document and using automatic statistical analysis to enhance the accuracy of data and data groups retrieved from images of paper source documents. The paper source documents may be processed by OCR software installed in the memory 106 of the image processing device 102 to produce associated text data. In one embodiment in accordance with aspects of the invention, a scanner device (e.g., flatbed scanner 124) may optically scan an EOB document to generate an image file. The text data produced by OCR software (e.g., an OCR component of a system for processing a document) may be included in the image file of the paper source document or may be stored in a character data file associated with the image file. Such a scanner device 124 may include a memory 214 for storing computer-executable instructions that may be executed using a processing unit (e.g., processor 212) to perform the various method steps disclosed herein. method.

- [21] The method described above may also include using information from the documents to determine their source, and select a template based on the determined source to utilize with each document. In one embodiment, a template selection component 202 in the image processing device 102 may be used to select the appropriate template for the document. For example, if the document is determined to be a check, a template for the check may be retrieved, whereas another template for an EOB data sheet may be retrieved if the document is an EOB data sheet. In one embodiment of the invention, the method may include processing groups of documents and matching related documents for further processing. For example, the method may be used in a healthcare payment system where one set of document may include documents relating to the services received by patients including healthcare insurance coverage, costs, etc., such as, for example, EOB documents. Another set of documents may include payment documents such as, for example, checks for payments by patients for healthcare services received by the patients. Based on the information retrieved by the OCR, the EOB documents may be matched with the payment documents. The processing may be done in a parallel manner to maximize throughput.
- [22] In an embodiment in accordance with aspects of the invention, parallel processing may be achieved by creating an expandable farm of process servers, all of which may perform one or more of the process steps, working from a single queue of input files. Each process server in the farm may seek to process the oldest file in the queue utilizing First-In First-Out (FIFO) in conjunction with a file contention avoidance utility, which may ensure all files from the queue are processed in the quickest possible manner. This may be beneficial in that it may allow scalability of the overall process by expansion of the number of computers in the process farm and may avoid a delay in processing due to one or more large time consuming files preventing other files from being processed if all processes were performed in a strictly serial manner.
- [23] In an embodiment in accordance with aspects of the invention, an incoming document may not have a matching template in the system. A template may contain information that allows the process to control the extraction of individual data elements from the text file output from the OCR process. For example, an EOB document may be provided by a healthcare insurance company for the first time. A template may be

created using manual character recognition (MCR), whereby source text data may have been verified manually to find where the data comes from in the source text data. For example, a template may indicate where information appears within the source text data, and its format such as, for example, the type of service provided to a patient, its format, which may be alphabetical characters, and its location, which may be by line and column numbers or with respect to a reference point in the data sheet. Once templates are initially created using MCR, statistical patterning of documents that utilize the same template may be utilized to continuously update the template to minimize erroneous character recognition.

- [24] For example, if in a given template the first letter of the patient's name appears on line 5, column 7, but sometimes it appears on column 6 or 8, statistical patterning may indicate column 7 as the location of the first letter of the patient's name. If the first letter occasionally appears in column 6 in the same type of document, automatically-performed statistical analysis may adjust the template to allow for variances in the starting column position such that horizontal drift, which is common in OCR output, of the starting location will capture instances of where the patient name starts in column 6 as the location of the first letter of the patient's name. The statistical analysis may identify the mean and standard deviations of occurrences of the starting position. Additionally, the processes by which the statistical analysis is achieved, may control all aspects of the fully automated template development process to include vertical drift of the location, where the patient name occasionally appears on line 4 or 6 in the above example, and where related data elements may be retrieved by measuring the offset from the patient name.
- [25] In another embodiment in accordance with aspects of the invention, a method for a computing machine to process images of paper source documents is illustrated as illustrated in Figure 3. In step 301, the source documents may be captured into image files (e.g., high quality image files, low quality image files, etc). The image files may be routed in any manner including, for example, via a secure connection (such as FTP, VPN, etc.) to be received (step 302) at an image processing device 102. Image files may then be processed through one or more optical character recognition (OCR) engines to retrieve or extract text data from the image files in step 305. In step 307, an

appropriate template may be located based on predetermined criteria. In one example, information extracted from the image files may be utilized to determine whether one or more criteria have been met for the appropriate template. The template may be utilized to map the text data to data groups that can be stored in a standardized format (in step 313). Such standardized formats include predetermined formats such as ANSI defined standard formats for the healthcare industry, extensible markup language (XML), and others.

- [26] For example, when a check is imaged and processed using OCR, the bank routing information sourcing from the financial institution lockbox process may be used to find the appropriate template that maps the OCR produced text data into data groups that can be stored in unique fields within a data file (xVAL). The data file may be used to produce an ANSI, for example, or other related data file for import into a Patient Accounting System in a healthcare service system. The criteria for selecting templates may correspond to the type of document being processed and/or may be selected based on the specific source of the document such as, for example, the healthcare insurance company or the bank issuing the payment.
- [27] Referring still to Figure 3, in step 308, the templates may be used to locate data groups within the text data. For example, the template may be used to match text data to identifier fields. This may be done by creating a record in the data file (xVAL) associated with the text data. The text data may then be parsed using a “mask” on identifier key fields to evaluate each line of the text output to determine if the line is a likely start point for a top of line group or top of form. The “mask” may evaluate attributes of the multiple key data field to determine if the data group on the OCR text extract is a new account on the EOB. This process simulates the manual process of cutting out windows in a cardboard overlay where the key fields are expected to occur based on the statistically learned patterns and passing it over the EOB form, wherein when moving the cardboard template over the form, the appearance of the data in the windows resembles what one would expect to find in the key fields of patient name, number, date of service, charges, etc. The attributes of the data in the window may then look, for example, for numbers in the charges field, numbers with dashes or slashes in

the date field, alphabetic characters in the name field and numbers (and characters, based on the provider numbering scheme) in the account number field, etc.

- [28] Healthcare EOB data sheets usually have multiple lines on the paper form to identify the full detail of payment information, in addition to other information regarding the patient and his/her condition and treatment, etc. Identifier key fields may be, for example, information such as account number, name, date, amount being charged, amount paid, etc. Using this process also allows filtering out non-account data. The line group identification process may be refined by utilizing masking filters. The masking filters may be defined with character string attributes such as, for example, all numeric as is expected with account numbers and charges, or a specific number of characters as is expected with certain number like phone numbers or social security numbers. Another example of character string attributes may be containing slashes or dashes for dates. This process of refining the line group identification may account for horizontal and vertical drift which allows for variances in the text pattern created by the OCR conversion of image to text data. For example, the OCR conversion of an image may shift a character string vertically or horizontally, which the process of refining the line group identification compensates for and as a result the character strings may be recognized accurately.
- [29] The retrieved templates may be used to identify individual data fields by using the row and column offset of each data item from a point of reference in the text data. The data fields may then be used to get mapped into the data file (xVAL). For example, the reference point may be the upper left most data item in the image, which in the EOB format may be the account number. In one embodiment, a data retrieval component 204 in an image processing device 102 may be for receiving the character data (e.g., text data) and associating it with an appropriate field in the retrieved document template. The data retrieval component 204 may be further configured to update the character data using data management rules (described below.)
- [30] In step 311, the data management rules may be added to correct some errors of OCR such as, for example, converting “l” (the letter) to “1” (the number) in an all numeric field. The key fields loaded into the data file may be used to identify corresponding documents. For example, the record from the provider may be identified from the

sourced Master Patient Index (MPI) to match payment documents with the corresponding patient healthcare service documents. The MPI is the claim data and other data elements used by the provider to bill the payer, and may be stored in a remote data repository 208. The MPI ultimately returned from the payer to the provider on the EOB may be used to enhance accuracy of the OCR. For example, this identification of corresponding documents may enable improvements on the OCR results to help it more accurately match when posting transactions. The MPI may be fed through from the provider system periodically, for example, on daily basis. The information of the MPI may be stored in a data storage unit 208 in communication with the image processing device 102. The data storage unit 208 may also store document templates, image files, and/or statistical data used to enhance the accuracy of data and data groups retrieved from images of paper source documents.

- [31] This process of identifying corresponding documents may utilize, for example, the MPI, which may contain frequently updated source claims data and related codes, etc. Additionally, identifying corresponding documents may involve a “fuzzy match” process that allows selection of the best matching record from the MPI by weighting the accuracy of the match from all key fields. The fuzzy match process may allow for near matches within a threshold determined acceptable by match rates as controlled in the templates. For example, account numbers are often long and contain many numbers, therefore, when one of the numbers is erroneously recognized, the erroneous number may not have a perfect match, but may be close to some numbers, and the “closeness” of the erroneous number may be quantified as to determine the best match. For a number ending in 678, for example, if the 7 is interpreted as a 2, and the recognized number therefore ends in 628 instead of 678, the best match may be the number ending in 678 based on the match between the remaining digits of the numbers. This may be done using statistical analysis, which may determine the percentage of accuracy between the recognized character and the record from the MPI to select the character string with the highest match percentage and change it if there is no perfect match. Identifying corresponding documents may also pull other information from the MPI which are not on the source document, for example, the EOB. This other information may be, for example, the payer, insurance plan codes, etc.

- [32] Other data management rules may be applied to the text data (in step 311). For example, certain character strings may contain characters that may not be necessary, such as hyphens in account numbers.
- [33] Manual verification may also be utilized to provide further verification of the identification process. A post export process may use the manually-verified output to the original and intermediate results from the OCR text data to determine the parameters in the templates. This process may determine significant statistical relationships of the MCR data to the original and intermediate results of the OCR data to measure where data elements fall on the form, for key fields and, once mapped, the other data elements; measure the patterns of field length, formatting, etc. to facilitate extraction and conversion; and measure how the compared to MPI data is effected. In one example, the manually verified data and corresponding OCR text dumps are maintained in a data store from which statistical correlation used to automate template creation is performed. The post export process may also utilize a template update cycle based on volumes from ongoing verified production data. Statistical patterns used in the templates may be adjusted as the underlying data adjusts, and the templates may be updated accordingly.
- [34] Although not required, one of ordinary skill in the art will appreciate that various aspects described herein may be embodied as a method, a data processing system, or as a computer-readable medium storing computer-executable instructions. For example, a computer-readable medium storing instructions to cause a processor in a computing device to perform steps of a method in accordance with aspects of the disclosure is contemplated.
- [35] In addition, various signals representing data or events as described herein may be transferred between a source and a destination in the form of electromagnetic waves traveling through signal-conducting media such as metal wires, optical fibers, and/or wireless transmission media (*e.g.*, air and/or space).
- [36] Aspects of the invention have been described in terms of illustrative embodiments thereof. The features of the embodiments described below contemplate other embodiments comprising one or more, or a combination thereof, of the aspects

described throughout. Numerous other embodiments, modifications and variations within the scope and spirit of the appended claims will occur to persons of ordinary skill in the art from a review of this disclosure. For example, one of ordinary skill in the art will appreciate that the steps illustrated in the illustrative figures may be performed in other than the recited order, and that one or more steps illustrated may be optional in accordance with aspects of the disclosure.

I claim:

1. A method, comprising:

receiving image data;

retrieving text data from the image data;

selecting a template from among a plurality of templates based on at least one predetermined criteria, where the template identifies a mapping of the text data to identifier fields and is based at least in part on patterns obtained through manual verification;

matching the text data to the identifier fields using the template; and

formatting the text data using a predetermined format.

2. The method of claim 1, where the predetermined criteria is based on at least a source of the image data.

3. The method of claim 2, where the source of the image data is an explanation of benefits document.

4. The method of claim 1, where the predetermined criteria is based on at least a location of the text data on an image corresponding to the image data.

5. The method of claim 4, where the image data includes data corresponding to an image of an explanation of benefits document, and where the identifier fields comprise payee name, insurance coverage, and amount being charged.

6. The method of claim 1, where the retrieving text data uses optical character recognition, the method further comprising:

updating the text data using data management rules.

7. The method of claim 6, where the updating the text data includes converting a letter in the text data to a number if the text data matches to an identifier field consisting of all numeric data.

8. The method of claim 6, where the identifier fields comprise account number and payee name, and where the updating the text data includes accessing a remote data repository to compare accuracy of the text data against related documents.

9. The method of claim 8, where the related documents are patient healthcare service documents.

10. The method of claim 1, further comprising:

generating a new template through manual character recognition if none of the plurality of templates meets the predetermined criteria to be selected.

11. A computer-readable medium storing computer-executable instructions for causing a processor to perform a method comprising:

receiving image data;

retrieving text data from the image data using optical character recognition;

selecting a template from among a plurality of templates based on a predetermined criteria, where the template identifies a mapping of the text data to identifier fields;

matching the text data to the identifier fields using the template;

updating the text data using data management rules;

formatting the text data into a predetermined format; and

sending formatted text data.

12. The computer-readable medium of claim 11, where the predetermined criteria is based on at least a source of the image data being an explanation of benefits document.

13. The computer-readable medium of claim 12, where the identifier fields comprise account number, and where the updating the text data includes accessing a remote data repository to compare accuracy of the text data against healthcare documents.

14. The computer-readable medium of claim 11, where the updating the text data includes converting a number in the text data to a letter if the text data matches to an identifier field consisting of all non-numeric data.

15. A system, comprising:

an optical character recognition component configured to perform optical character recognition on an image file to generate character data;

a data storage unit storing document templates;

a template selection component configured to select one of a plurality of templates; and

a data retrieval component configured to receive the character data and associating the character data with an appropriate field in the document template.

16. The system of claim 15, where the image file corresponds to an electronic image of an explanation of benefits paper document.

17. The system of claim 16, further comprising:

a scanner component for optically scanning a document corresponding to the image file.

18. The system of claim 15, where the data retrieval component is further configured to update the character data using data management rules.

19. The system of claim 18, where the updating includes converting a letter in the character data to a number if the character data matches to a field in the document template configured to correspond to all numeric characters.

20. The system of claim 15, where the data retrieval component is further configured to access a remote data repository to compare accuracy of the character data against related documents.

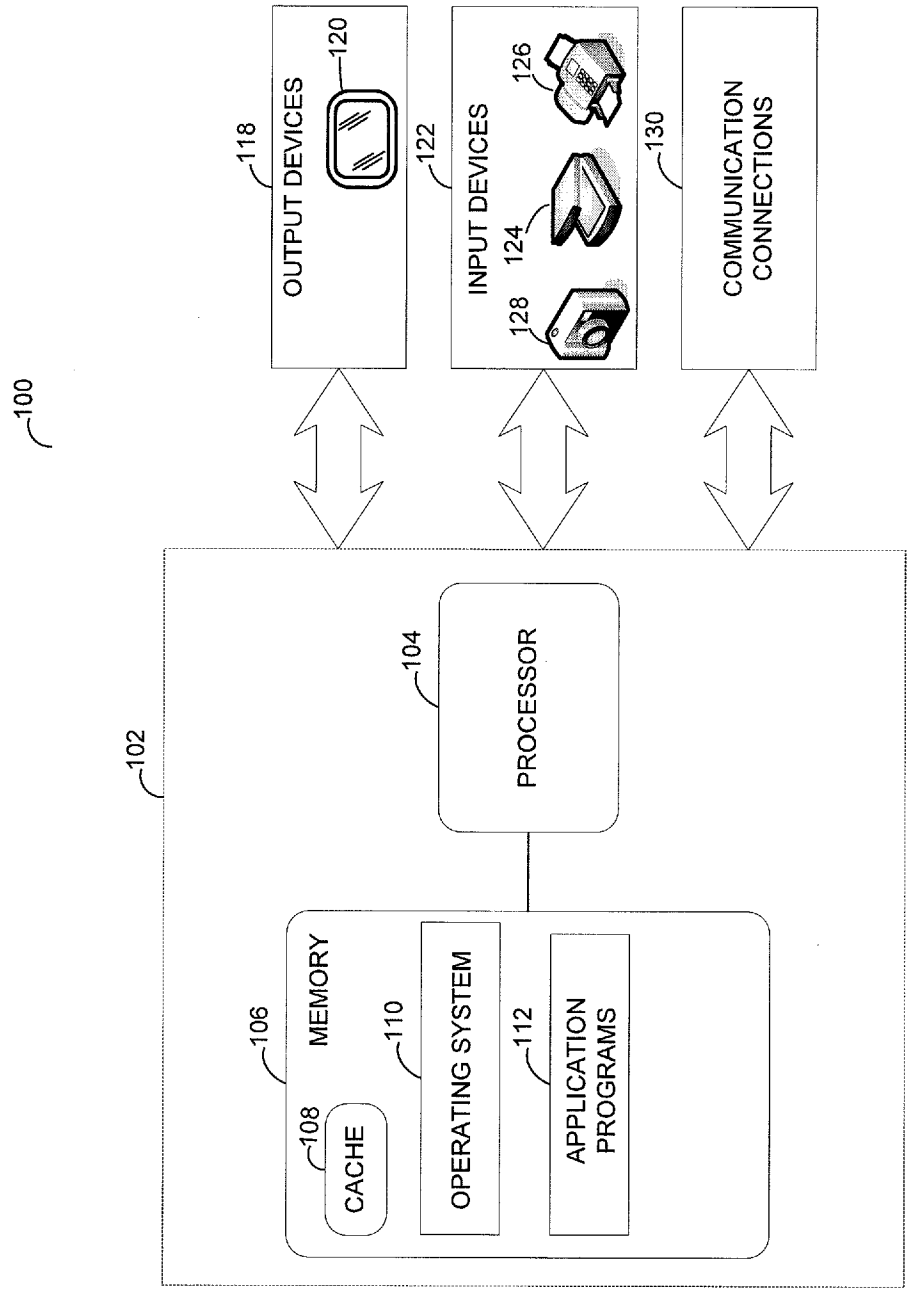


FIG. 1

2/3

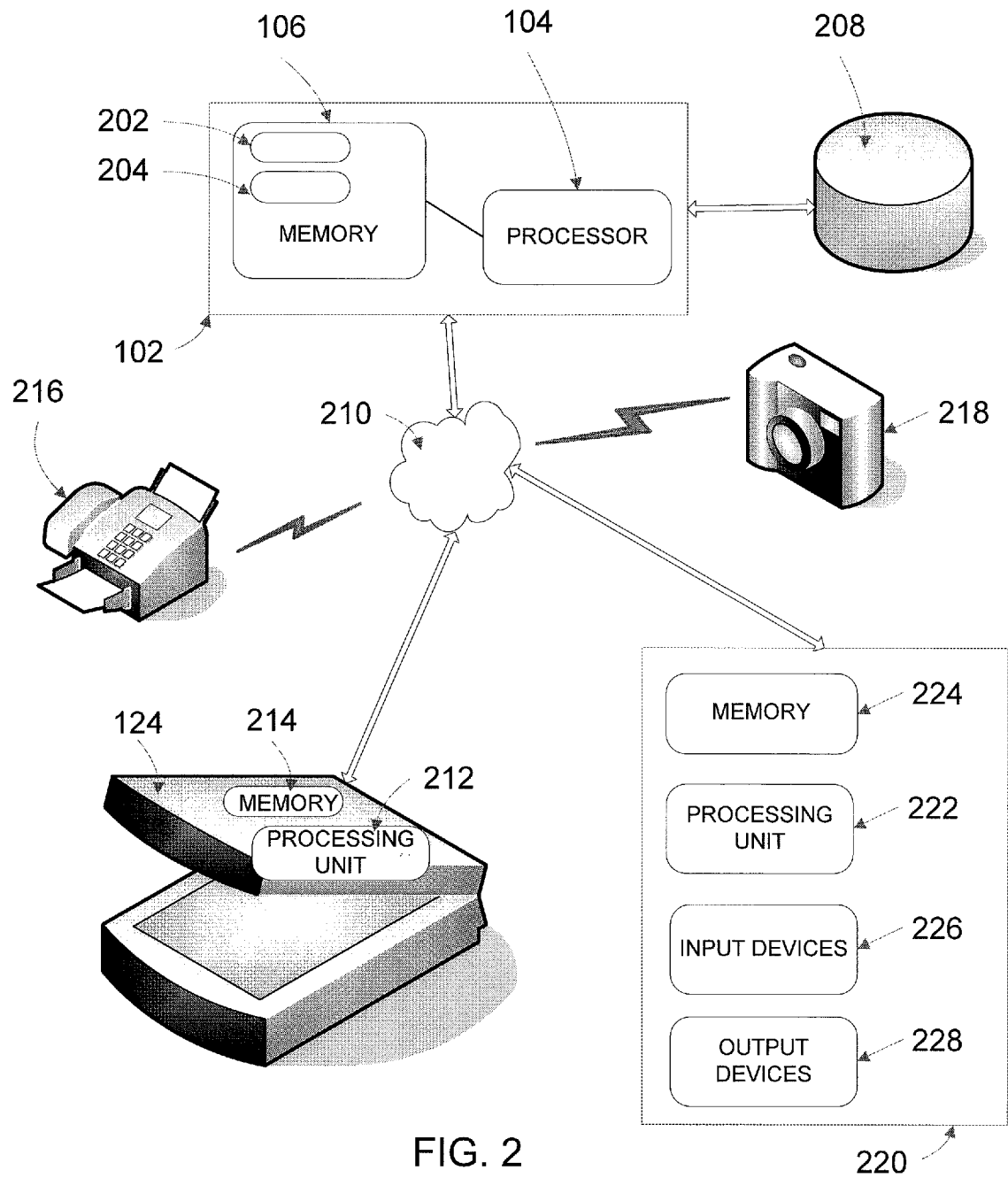


FIG. 2

3/3

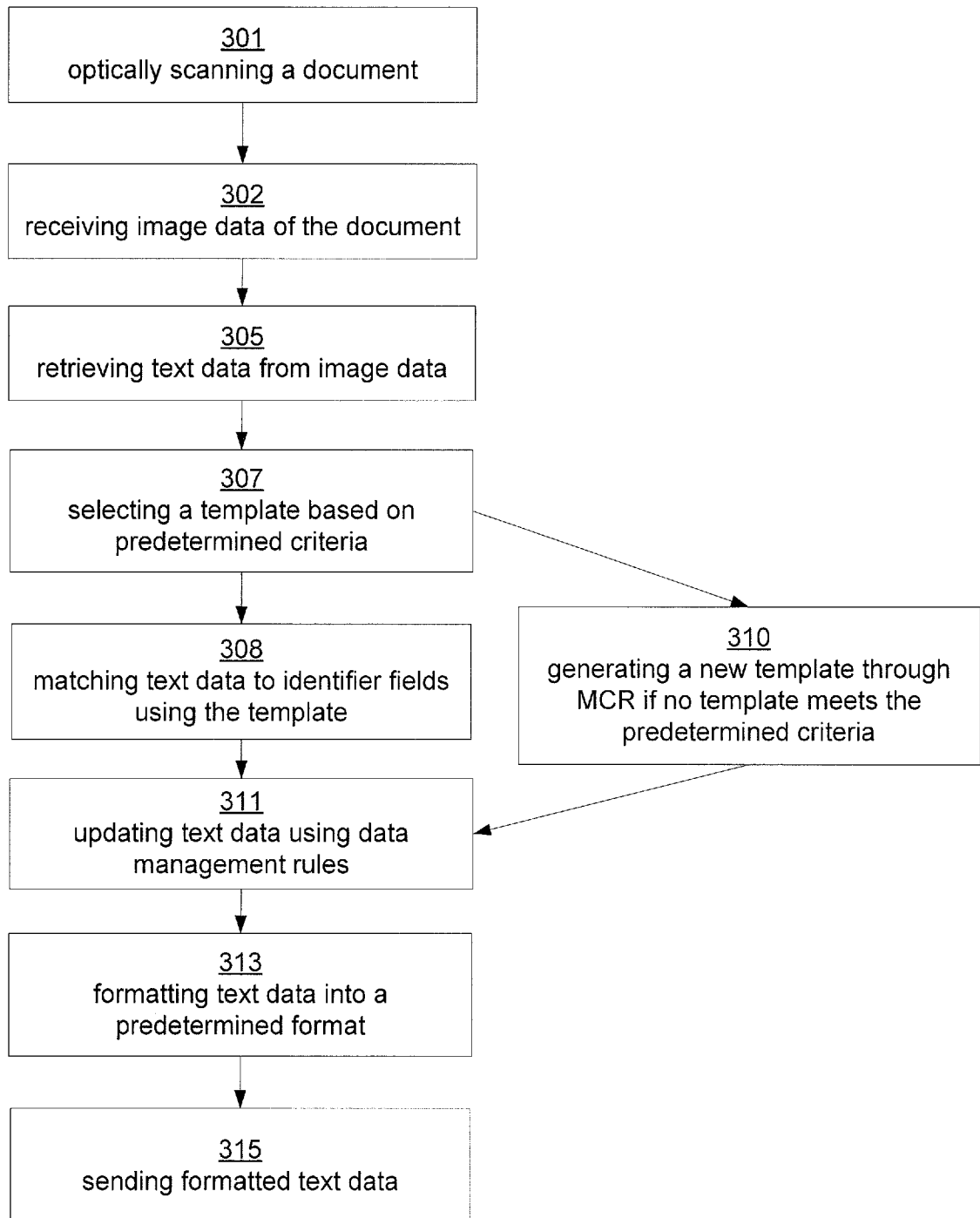


FIG. 3

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2007/085949

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06K9/20

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)
G06K

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 practical, search terms used)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y	US 7 149 347 B1 (WNEK JANUSZ [US]) 12 December 2006 (2006-12-12) column 7, line 45 - column 11, line 53	1-5, 10, 12, 15-17 6-9, 11, 13, 14, 18-20
A	column 1 - column 4, line 3 ----- WO 03/038735 A (CALL TELL LLC [US]) 8 May 2003 (2003-05-08) abstract; claim 1 paragraph [0005] paragraph [0009] paragraph [0032] - paragraph [0034] ----- -/--	1-20

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"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)

"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

"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

"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

"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.

"&" document member of the same patent family

Date of the actual completion of the international search

22 February 2008

Date of mailing of the international search report

05/03/2008

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040, Tx. 31 651 epo nl,
Fax: (+31-70) 340-3016

Authorized officer

Müller, Martin

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2007/085949

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 5 642 435 A (LORIS KEITH [US]) 24 June 1997 (1997-06-24) abstract; figure 7 column 1, line 37 - column 2, line 10 column 5, line 27 - column 6, line 54 -----	6-9, 11, 13, 14, 18-20
A	LIEBOWITZ TAYLOR S ET AL: "EXTRACTION OF DATA FROM PREPRINTED FORMS" MACHINE VISION AND APPLICATIONS, SPRINGER VERLAG, DE, vol. 5, 1992, pages 211-222, XP000828275 ISSN: 0932-8092 the whole document -----	1-20
A	EP 0 407 935 A (HITACHI LTD [JP]; HITACHI SOFTWARE ENG [JP]) 16 January 1991 (1991-01-16) abstract column 2, line 17 - line 33 column 3, line 8 - line 67 -----	1-20

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2007/085949

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 7149347	B1	12-12-2006	NONE
WO 03038735	A	08-05-2003	EP 1449157 A1 25-08-2004
		US 2003084402 A1	01-05-2003
		US 2007206223 A1	06-09-2007
		US 7296221 B1	13-11-2007
US 5642435	A	24-06-1997	NONE
EP 0407935	A	16-01-1991	DE 69033312 D1 11-11-1999
		JP 3179570 A	05-08-1991
		JP 2000155795 A	06-06-2000
		US 5228100 A	13-07-1993