(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
18 September 2014 (18.09.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/146032 A2

(51) International Patent Classification:
G06F 17/30 (2006.01)

(21) International Application Number:
PCT/US2014/030904

(22) International Filing Date:
17 March 2014 (17.03.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/799,091 15 March 2013 (15.03.2013) US

(72) Inventors; and
(71) Applicants : **SUAREZ, Sergio, David, Jr.** [US/US]; 821 N Neva, Addison, IL 60101 (US). **MESKE, Joshua, Daniel** [US/US]; 5121 N. East River Rd., Apt. 2h, Chicago, IL 60656 (US).

(74) Agent: **PASKY, Jonathan, R.**; Pasky IP Law, 320 W Ohio St Ste 300, Chicago, IL 60654 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, 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, IR, IS, JP, KE, KG, 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, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: SYSTEM FOR METHOD FOR DATA SWEEPING USING KEYWORDS

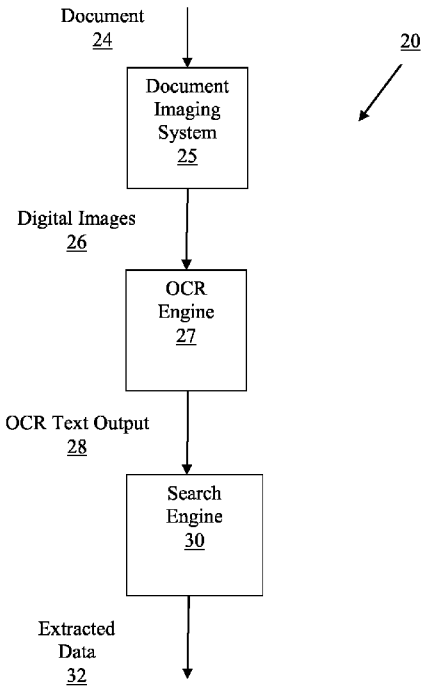

Fig. 1

(57) Abstract: A method for extracting data relating to a search term from a searchable text document is disclosed. The method includes executing on a processor instructions for extracting the data from the search term, wherein the search term is defined by at least one character pattern. The instructions include: searching the text document for the search term; performing a data sweep within an area around the search term within the text document; identifying at least one string within the area; comparing the at least one string to the at least one character pattern; and extracting data from the at least one string when the data of the at least one string matches the at least one character pattern.

WO 2014/146032 A2

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

SYSTEM AND METHOD FOR DATA SWEEPING USING KEYWORDS

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to U.S. provisional application Ser. No. 61/799,091, filed on March 15, 2013, which is hereby incorporated by reference in its entirety.

BACKGROUND OF THE DISCLOSURE

Technical Field

[0002] The present disclosure relates to processing of text documents, such as documents transcribed from an image processed by optical character recognition (OCR). More particularly, the disclosure relates to a system and method for searching for a desired search term through the text document and extracting data pertaining to the search term.

Background of Related Art

[0003] Scanning documents into digital form has become more common with the advent of improved imaging, storage and distribution techniques. By converting these documents into electronic form, institutions can reduce the cost of storage, facilitate remote access, enable simultaneous access by multiple users, and/or facilitate search and retrieval of information.

[0004] Once the content of a document is scanned, the digitally recorded image can be manipulated or otherwise processed. For example, preprocessing algorithms may be performed to de-warp, reformat, supplement with additional information, and/or compress the digitally recorded image. After performing the preprocessing algorithms, the preprocessed image may be processed with OCR software and may be indexed to facilitate electronic search. Thus,

scanning and recording of documents facilitates the creation of digital libraries that can be remotely and simultaneously accessed and searched by multiple users.

[0005] Searching of electronic documents includes identifying matches in the document for the search term to locate the term anywhere within the documents. Such searching does not allow for locating and identifying data that pertains to the search term. Accordingly, there is a need for a system and method that locates nearby data pertaining to the search term and verifies the accuracy thereof.

SUMMARY

[0006] According to one aspect of the present disclosure, a method for extracting data relating to a search term from a searchable text document is disclosed. The method includes executing on a processor instructions for extracting the data from the search term, wherein the search term is defined by at least one character pattern. The instructions include: searching the text document for the search term; performing a data sweep within an area around the search term within the text document; identifying at least one string within the area; comparing the at least one string to the at least one character pattern; and extracting data from the at least one string when the data of the at least one string matches the at least one character pattern.

[0007] According to one aspect of the present disclosure, a system for extracting data relating to a search term from a searchable text document is disclosed. The system includes a computer processor that is operable to execute a computer program product tangibly embodied in a computer-readable storage medium. The computer program product being operable to cause the computer processor to: search the text document for the search term; perform a data sweep within an area around the search term within the text document; identify at least one string within the area; compare the at least one string to the at least one character pattern; and

extract data from the at least one string when the data of the at least one string matches the at least one character pattern.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The accompanying drawings, which are incorporated in and constitute a part of this specification, illustrate exemplary embodiments of the disclosure and, together with a general description of the disclosure given above, and the detailed description of the embodiments given below, serve to explain the principles of the disclosure, wherein:

[0009] Fig. 1 is a block diagram illustrating a system for generating an OCR output of a scanned document according to the present disclosure;

[0010] Fig. 2 is a flow chart illustrating a method for data sweeping the OCR output according to the present disclosure;

[0011] Fig. 3 is an illustration of the data sweeping according to the present disclosure; and

[0012] Fig. 4 is a functional diagram illustrating a computing environment and a basic computing device that can operate the data sweeping application according to the present disclosure.

DETAILED DESCRIPTION

[0013] System and methods for finding nearby data using the location of the search term and verifying accuracy of the nearby data are disclosed.

[0014] Fig. 1 is a block diagram illustrating an exemplary system 20 for extracting data 32 from an OCR text output 28 of digital images 26 resulting from scanning of a document 24. Examples of documents 24 include personal records, medical records, books, articles, magazines, and other printed material. Generally, there are errors in both the OCR text output

28 including, but not limited to, incorrect character assignment. The systems and methods according to the present disclosure may be used to process any searchable text document besides OCR text output 28, including, but not limited to, web pages, text editor documents, etc.

[0015] The document 24 may first be scanned (e.g., imaged) using a document imaging system 25 to generate one or more digital images 26 of the document 24 on which OCR may be performed by an OCR engine 27 to generate OCR text output 28. Any suitable combination of various document imaging systems 25 and OCR engines 27 (e.g., any commercially available OCR engine) may be employed. The OCR text output 28 may then be used as input to a search engine 30.

[0016] Operation of the search engine 30 is illustrated in Figs. 2 and 3. Fig. 2 shows a method 100 for sweeping data within the OCR text output 28 (e.g., searchable text document) in accordance with the present disclosure. Initially, a search term or phrase 50 (Fig. 3), such as "social security number," is input into a search field of the search engine 30 and the search engine 30 searches the OCR text output 28 for any occurrence of the search term 50 therein. Multiple search terms 50 may be input into the search engine 30. If one of the search terms 50 is found, then the search engine 30 initiates a data sweep. In particular, the search engine 30 investigates nearby strings within the OCR text output 28 to determine if they possess any data that corresponds to the search term 50.

[0017] With reference to Fig. 3, the search engine 30 sweeps an area 52 around the search term 50 of the OCR text output 28 for any strings that match predefined criteria. The area 52 may be defined by any suitable shape including, but not limited to, circle, oval, rectangle. The boundary of the area 52 may be defined using any suitable parameters including, but not limited to, pixels, lines, characters. The sweep may be commenced from within the area 52 at any point (e.g., right, left, above, below) in relation to the search term 50 and may

be performed in any direction (e.g., clockwise, counterclockwise, etc.). Sweeps may occur by quadrants or any other suitable subdivisions of the area 52. The direction and shape of the sweep may be any suitable shape (e.g., linear, arcuate, etc.) since the data may be located anywhere within the area 52. In embodiments, the area 52, direction and shape of the sweep may be based on the location of previously located data. In further embodiments, if no data is found during the data sweep either to the right or below of the search term 50, the edge of the page of the OCR text output 28 may be used as definition for the area 52.

[0018] The search engine 30 includes one or more parameters (e.g., character pattern) defining the search term 50 allowing the search engine 30 to correlate the strings located during the data sweep to the search term 50, e.g., strings pertaining to a social security number are defined by a series of nine numbers, potentially with spaces or dashes between the third and fourth, and fifth and sixth numbers.

[0019] As shown in Fig. 3, the sweep initially locates a first string 54 located below the search term 50. The string 54 is then analyzed by the search engine 30 to determine whether the string 54 correlates to the search term 50 by comparing the text of the string 54 to the predefined pattern. If the string 54 does not contain data that correlates to the search term 50, namely, the string 54 contains text or characters (e.g., David Wilson) that do not match the predefined pattern for search term 50 (e.g., a social security number), the search engine 30 continues to sweep the area 52.

[0020] As the data sweep continues, the search engine 30 locates a second string 56 to the right of the search term 50. The string 56 is also analyzed by the search engine 30. If the string 56 contains data that correlates to the search term 50, namely, the string 54 contains text or characters (e.g., 123-45-6789) that matches the predefined pattern for search term 50 (e.g., a social security number), the search engine 30 extracts the data from the string 56 and outputs the same as extracted data 32 (e.g., for further processing).

[0021] An example of a suitable operating environment in which the embodiments (e.g., OCR engine 27, search engine 30) of the present disclosure may be implemented is illustrated in Fig. 4. The operating environment is only one example of a suitable operating environment and is not intended to suggest any limitation as to the scope of use or functionality. Other well known computing systems, environments, and/or configurations that may be suitable for use with the embodiments, but are not limited to, personal computers, server computers, hand-held or laptop devices, multiprocessor systems, microprocessor-based systems, programmable consumer electronics, network PCs, minicomputers, mainframe computers, distributed computing environments that include any of the above systems or devices, and the like.

[0022] With reference to Fig. 4, an exemplary system for implementing the embodiments includes a computing device, such as computing device 200. In its most basic configuration, computing device 200 typically includes at least one processing unit 202 and memory 204. Depending on the exact configuration and type of computing device, memory 204 may be volatile (such as RAM), non-volatile (such as ROM, flash memory, etc.), or some combination of the two. The most basic configuration of the computing device 200 is illustrated in Fig. 4 by dashed line 206.

[0023] Additionally, device 200 may also have additional features or functionality. For example, device 200 may also comprise additional storage (removable and/or non-removable) including, but not limited to, magnetic disks, optical disks, or tape. Such additional storage is illustrated in Fig. 4 by removable storage 208 and non-removable storage 210. Computer storage media includes volatile and nonvolatile, removable and non-removable media implemented in any method or technology for storage of information, such as computer readable instructions, data structures, program modules, or other data. Memory 204, removable storage 208, and non-removable storage 210 are all examples of computer storage media.

Computer storage media includes, but is not limited to, RAM, ROM, EEPROM, flash memory or other memory technology, CD-ROM, digital versatile disks (DVD) or other optical storage, magnetic cassettes, magnetic tape, magnetic disk storage or other magnetic storage devices, or any other medium which can be used to store the desired information and which can be accessed by device 200. Any such computer storage media may be part of device 200.

[0024] Device 200 may also contain communications connection(s) 212 that allow the device to communicate with other devices. Communications connection(s) 212 is an example of communication media. Communication media typically embodies computer readable instructions, data structures, program modules, or other data in a modulated data signal such as a carrier wave or other transport mechanism and includes any information delivery media. The term "modulated data signal" means a signal that has one or more of its characteristics set or changed in such a manner as to encode information in the signal. By way of example, and not limitation, communication media includes wired media, such as a wired network or direct-wired connection, and wireless media, such as acoustic, RF, infrared, and other wireless media.

[0025] Device 200 may also have input device(s) 214 such as keyboard, mouse, pen, voice input device, touch input device, etc. Output device(s) 216 such as a display, speakers, printer, etc. may also be included. The devices 214 may help form the user interface 102 discussed above while devices 216 may display results 106 discussed above. All these devices are well know in the art and need not be discussed at length here.

[0026] Computing device 200 typically includes at least some form of computer readable media. Computer readable media can be any available media that can be accessed by processing unit 202. By way of example, and not limitation, computer readable media may comprise computer storage media and communication media. Combinations of the any of the above should also be included within the scope of computer readable media. In embodiments, the software for executing the expression editing tool and aligning and breaking expressions is

stored on the computer readable media or in memory 204 and/or executed by the processing unit 202.

[0027] The computer device 200 may operate in a networked environment using logical connections to one or more remote computers (not shown). The remote computer may be a personal computer, a server computer system, a router, a network PC, a peer device, or other common network node, and typically includes many or all of the elements described above relative to the computer device 200. The logical connections between the computer device 200 and the remote computer may include a local area network (LAN) or a wide area network (WAN), but may also include other networks. Such networking environments are commonplace in offices, enterprise-wide computer networks, intranets, and the Internet.

[0028] When used in a LAN networking environment, the computer device 200 is connected to the LAN through a network interface or adapter. When used in a WAN networking environment, the computer device 200 typically includes a modem or other means for establishing communications over the WAN, such as the Internet. The modem, which may be internal or external, may be connected to the computer processor 202 via the communication connections 212, or other appropriate mechanism. In a networked environment, program modules or portions thereof may be stored in the remote memory storage device. By way of example, and not limitation, a remote application programs may reside on memory device connected to the remote computer system. It will be appreciated that the network connections explained are exemplary and other means of establishing a communications link between the computers may be used.

[0029] Although the illustrative embodiments of the present disclosure have been described herein with reference to the accompanying drawings, it is to be understood that the disclosure is not limited to those precise embodiments, and that various other changes and

modifications may be effected therein by one skilled in the art without departing from the scope or spirit of the disclosure.

What is claimed is:

1. A method for extracting data relating to a search term from a searchable text document, comprising:

executing on a processor instructions for extracting the data from the search term, wherein the search term is defined by at least one character pattern, the instructions comprising:

searching the text document for the search term;

performing a data sweep within an area around the search term within the text document;

identifying at least one string within the area;

comparing the at least one string to the at least one character pattern; and

extracting data from the at least one string when the data of the at least one string matches the at least one character pattern.

2. A system for extracting data relating to a search term from a searchable text document, comprising:

a computer processor that is operable to execute a computer program product tangibly embodied in a computer-readable storage medium, the computer program product being operable to cause the computer processor to:

search the text document for the search term;

perform a data sweep within an area around the search term within the text document;

identify at least one string within the area;

compare the at least one string to the at least one character pattern; and

extract data from the at least one string when the data of the at least one string matches the at least one character pattern.

Sheet 1/3

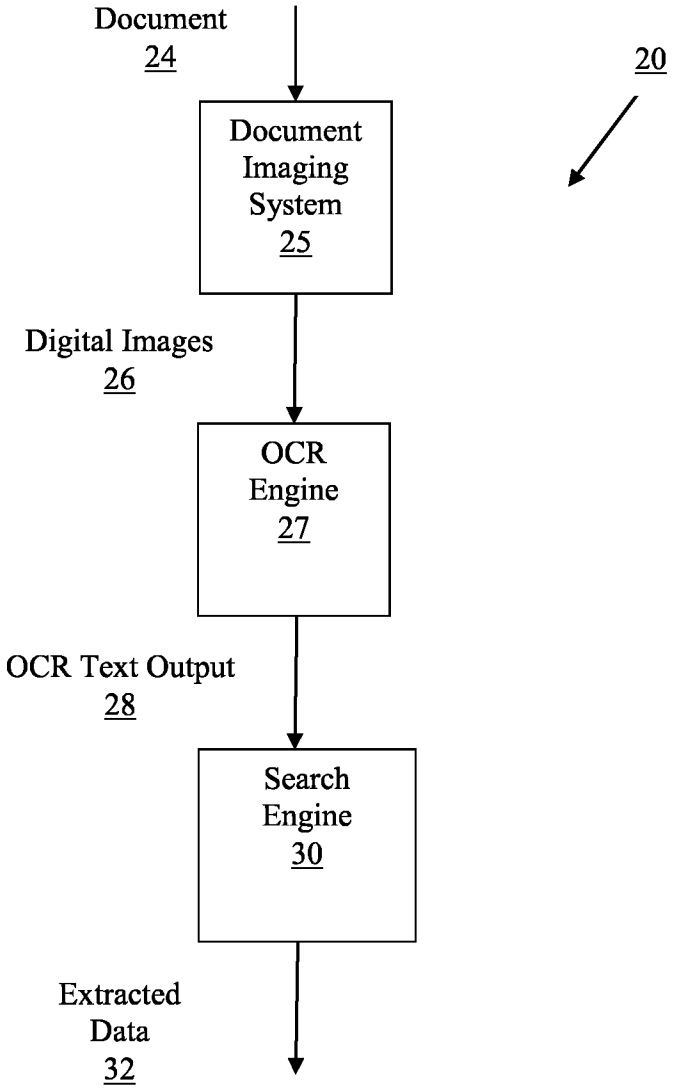

Fig. 1

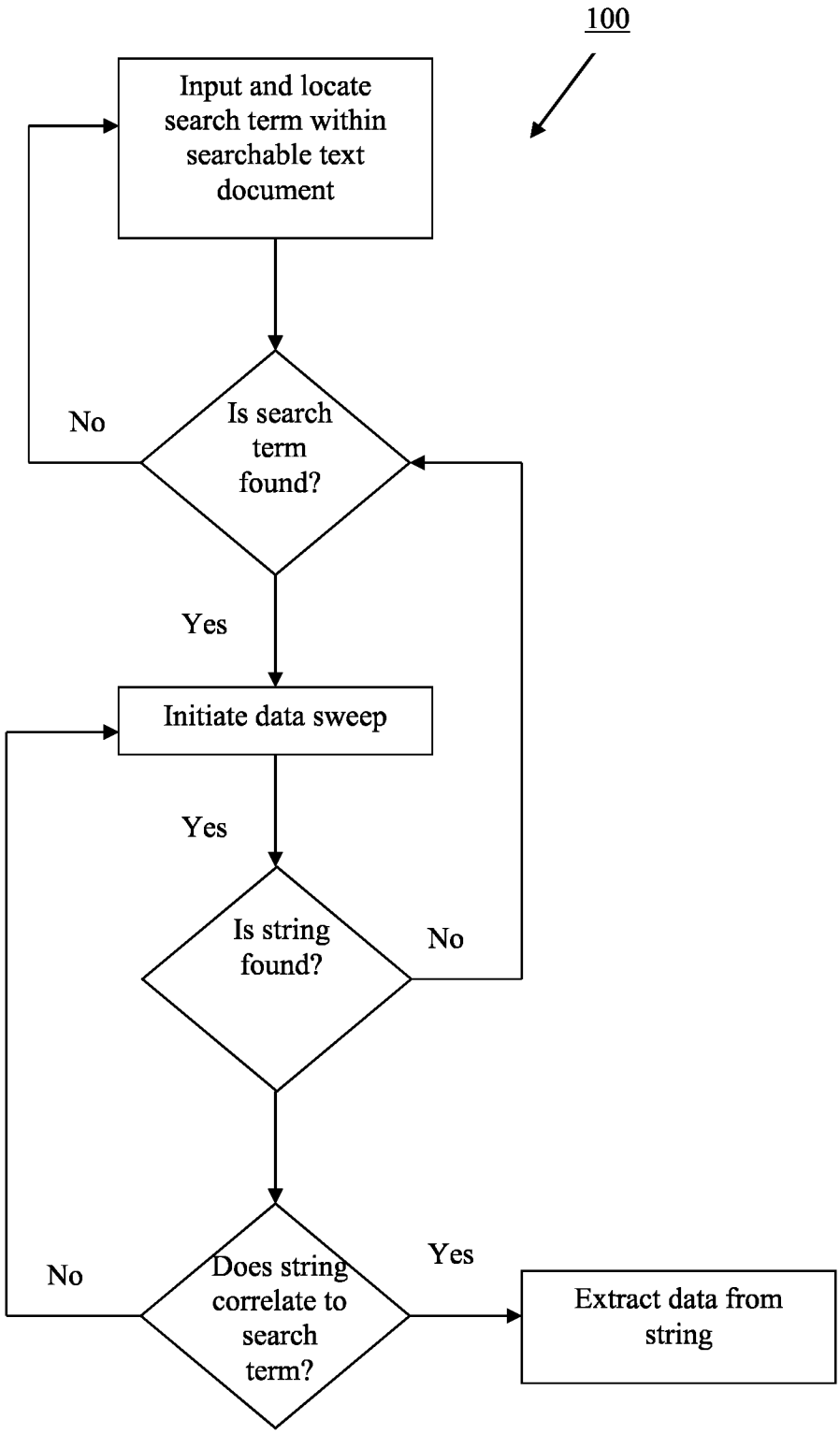
100
Input and locate search term within searchable text document
Is search term found?
No
Yes
Initiate data sweep
Yes
Is string found?
No
Does string correlate to search term?
No
Yes
Extract data from string

Fig. 2

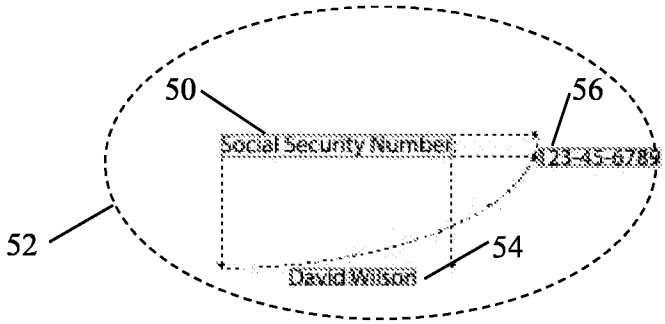
50
56
Social Security Number
123-45-6789
52
54
David Wilson

Fig. 3

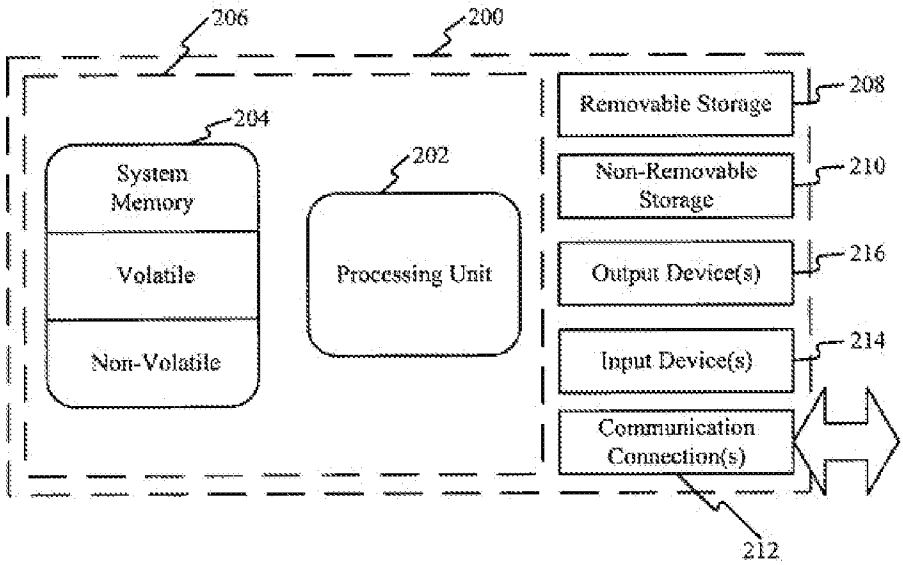
206
200
208
Removable Storage
204
System Memory
202
210
Non-Removable Storage
Volatile
Processing Unit
216
Output Device(s)
214
Non-Volatile
Input Device(s)
Communication Connection(s)
212

Fig. 4