(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 2014/145999 A2

(51) International Patent Classification:
G06K 9/03 (2006.01)

(21) International Application Number:
PCT/US2014/030867

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

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/798,223 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 AND METHOD FOR SEARCHING THROUGH TEXT TRANSCRIBED FROM AN IMAGE PROCESSED BY OPTICAL CHARACTER RECOGNITION

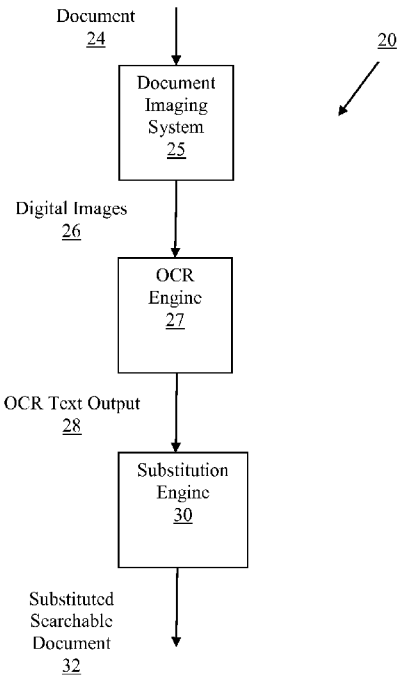

Fig. 1

(57) Abstract: A method for generating a character-by-character substitution in an optical character recognition (OCR) text output of a document including at least one character, includes: executing on a processor instructions for substituting an OCR key for the at least one character. The instructions include: identifying a class corresponding to the at least one character, wherein the class includes a character shape corresponding to at least a portion of the at least one character; substituting the OCR key including to the character shape for the at least one character; and generating a searchable substituted document including the OCR key.

WO 2014/145999 A2

Published:

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

SYSTEM AND METHOD FOR SEARCHING THROUGH TEXT TRANSCRIBED FROM AN IMAGE PROCESSED BY OPTICAL CHARACTER RECOGNITION

CROSS-REFERENCE TO RELATED APPLICATION

[0001] This application claims priority to U.S. provisional application Ser. No. 61/798,223, filed on March 15, 2013, which is hereby incorporated by reference in its entirety.

BACKGROUND

Technical Field

[0002] The present disclosure relates to processing of text transcribed from an image processed by optical character recognition (OCR). More particularly, the disclosure relates to a system and method for searching through text that has been scanned with an OCR reader using character substitutions based on character shapes.

Background of Related Art

[0003] Scanning books, magazines, and other printed material into digital form has become more common with the advent of improved imaging, storage and distribution techniques. Many institutions, such as the libraries, universities, bookstores, and private enterprises have vast collections of documents. By converting these documents into electronic form, such institutions can reduce the cost of storage, facilitate remote access, enable simultaneous access by multiple users, facilitate search and retrieval of information, and/or protect information in rare or out-of-print works from loss or destruction.

[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] Various factors may affect the accuracy of the OCR output. For example, each preprocessing algorithm performed on the digitally recorded images as well as the particular OCR software engine utilized may affect the accuracy of the OCR output. In addition, the imaging conditions and/or the original (hardcopy) document itself may also affect the accuracy of the OCR output, depending on, for example, the contents of the document (e.g., language, font, font size, page size, margins, text format such as columns, embedded images, etc.), the imaging conditions (e.g., operator, positioning of the document, camera zoom, camera focus, camera angle, and the like), etc.

[0006] OCR may be inaccurate when dealing with less than perfect text. Because of this, OCR scanners may return incorrect characters when scanning and processing documents. Accordingly there is a need for systems and method to overcome errors in the text that occur during OCR.

SUMMARY

[0007] Although OCR may produce inconsistent results with respect to specific letters, OCR scanners correctly identify get the general shapes of those letters. The present disclosure provides a system and method for substituting characters in a document obtained from an OCR scanner with a predefined shape. This bypasses the inherent inaccuracy associated with OCR and allows for searches performed on the raw OCR data to return text that matches the combination of the predefined shape of the characters rather than a direct OCR character thereby finding words/phrases that may have been missed otherwise.

[0008] According to one aspect of the present disclosure, a method for generating a character-by-character substitution in an optical character recognition (OCR) text output of a document including at least one character, including: executing on a processor instructions for substituting an OCR key for the at least one character. The instructions include: identifying a class corresponding to the at least one character, wherein the class includes a character shape corresponding to at least a portion of the at least one character; substituting the OCR key including the character shape for the at least one character; and generating a searchable substituted document including the OCR key.

[0009] According to one aspect of the above embodiment, the method further includes identifying a cardinality of the class corresponding to a frequency of occurrence of the character shape within the at least one character; and substituting the OCR key including the character shape and the cardinality for the at least one character.

[0010] According to one aspect of the present disclosure, a system for generating a character-by-character substitution of at least one character in an optical character recognition (OCR) text output of a 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: identify a class corresponding to the at least one character, wherein the class includes a character shape corresponding to at least a portion of the at least one character; substitute a OCR key including the character shape for the at least one character; and generate a searchable substituted document including the OCR key.

BRIEF DESCRIPTION OF THE DRAWINGS

[0011] 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:

[0012] Fig. 1 is a block diagram illustrating a system for generating a character-by-character substituted OCR output of a scanned document according to the present disclosure; and

[0013] Fig. 2 is a functional diagram illustrating a computing environment and a basic computing device that can operate the OCR substitution application according to the present disclosure.

DETAILED DESCRIPTION

[0014] System and methods for searching through a text document having substituted predefined character shapes for characters obtained from OCR output are disclosed.

[0015] Fig. 1 is a block diagram illustrating an exemplary system 20 for generating a substituted searchable document 32 through character-by-character substitution within an OCR output 28 of digital images 26 resulting from scanning of a document 24. Examples of documents 24 include 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.

[0016] The document 24 may first be scanned (e.g., imaged) using a document imaging systems 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 substitution engine 30. The substitution engine 30 then assigns an OCR key to each of the characters.

[0017] The OCR engine 30 accepts as input the OCR text output 28 and assigns for each character a predetermined shape as listed in Table 1 below.

Class	Member	Cardinality
i	f, i, j, k, l, r, t, B, D, E, F, I, J, K, L, P, R, T, 1, !	1
i	n, h, u, H, N, U	2
i	m, M	3
o	a, b, d, g, o, p, q, O, Q, 6, 9, 0	1
c	e, c, C, G	1
v	v, x, y, V, Y, X	1
v	w, W	2
s	s, S, 5	1
z	z, Z	1
a	A	1

Table 1.

[0018] For each character in the OCR text output 28, the OCR engine 30 determines which class the character belongs to and the cardinality of that class, if any, and assigns an OCR key. The classes are organized by the general shape of the characters and include, but are not limited to, “i," “o,” “c,” “v,” “s,” “z,” “a.” Within certain classes the general shape may occur more than once, such as with respect to the “i" and “v” classes. Each of theses classes may include a sub-class, e.g., cardinality denoting the frequency of occurrence of the predefined class shape within a specific group of characters. In embodiments, the OCR engine 30 may ignore some or all of the spaces within the OCR text output 28. The OCR engine 30 then outputs a searchable substituted document 32.

[0019] The OCR keys assigned to each of the characters eliminate the need for much of the post processing that occurs with standard OCR post-processing algorithms, as similar strings of characters with similar shapes become equivalent. Thus, the string “Thorn” becomes “iiioiii” when the OCR key is applied according to the present disclosure, which is the

equivalent of the string "Mom." This is particularly useful when searching for a piece of text scanned via OCR, as it prevents similarly shaped strings from being ignored.

[0020] An example of a suitable operating environment in which the embodiments (e.g., OCR engine 27, substitution engine 30) of the present disclosure may be implemented is illustrated in Fig. 2. 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.

[0021] With reference to Fig. 2, 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. 2 by dashed line 206.

[0022] 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. 2 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.

[0023] 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.

[0024] 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.

[0025] 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.

[0026] 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.

[0027] 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.

[0028] 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 generating a character-by-character substitution in an optical character recognition (OCR) text output of a document comprising at least one character, comprising:

executing on a processor instructions for substituting an OCR key for the at least one character, the instructions comprising:

identifying a class corresponding to the at least one character, wherein the class comprises a character shape corresponding to at least a portion of the at least one character;

substituting the OCR key comprising to the character shape for the at least one character; and

generating a searchable substituted document comprising the OCR key.

2. The method according to claim 1, further comprising:

identifying a cardinality of the class corresponding to a frequency of occurrence of the character shape within the at least one character; and

substituting the OCR key comprising the character shape and the cardinality for the at least one character.

3. A system for generating a character-by-character substitution of at least one character in an optical character recognition (OCR) text output of a 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:

identify a class corresponding to the at least one character, wherein the class comprises a character shape corresponding to at least a portion of the at least one character;

substitute a OCR key comprising to the character shape for the at least one character; and

generate a searchable substituted document comprising the OCR key.

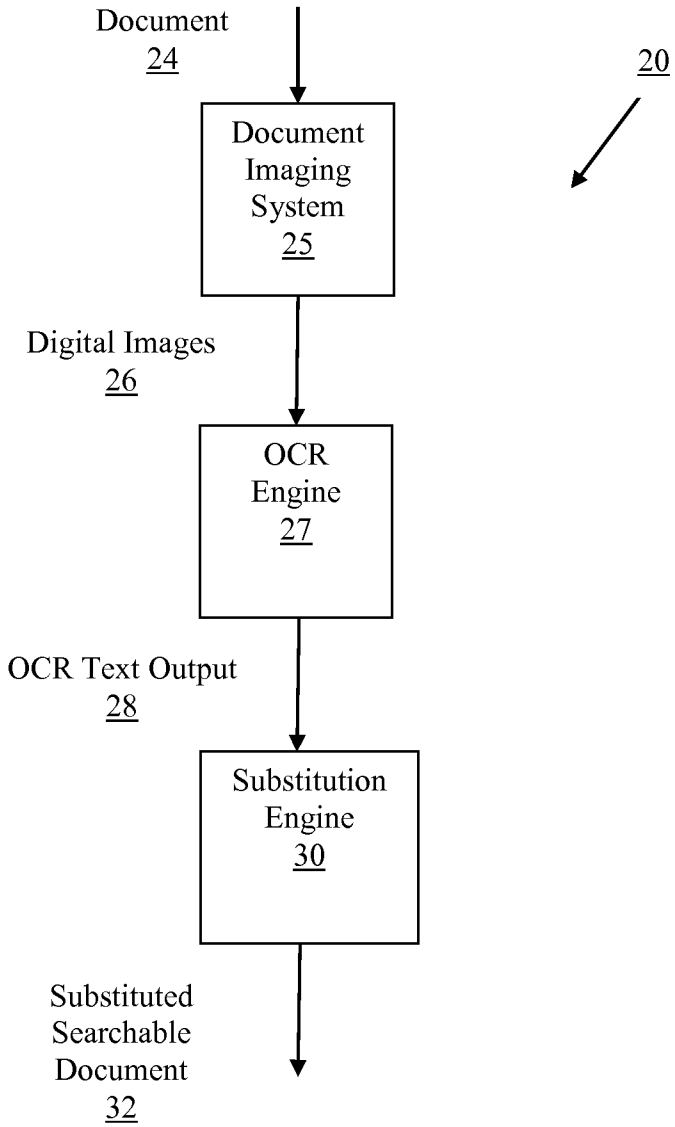
Document
24
20
Document Imaging System
25
Digital Images
26
OCR Engine
27
OCR Text Output
28
Substitution Engine
30
Substituted Searchable Document
32

Fig. 1

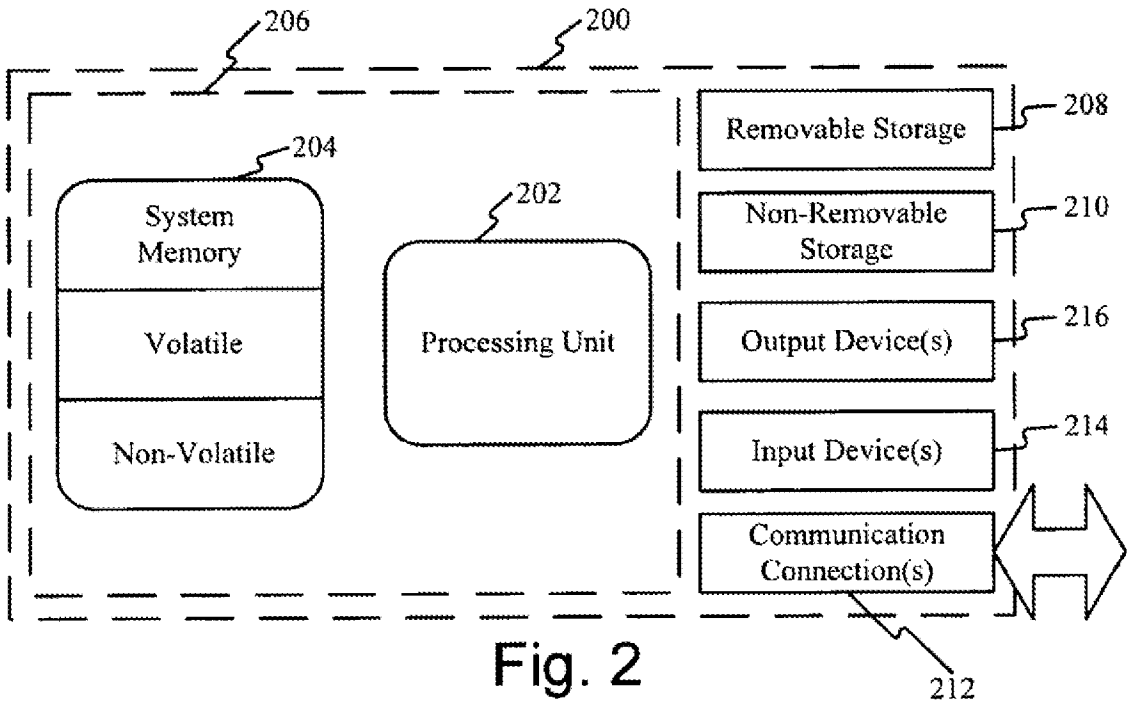
206
200
204
System Memory
Volatile
Non-Volatile
202
Processing Unit
Removable Storage
208
Non-Removable Storage
210
Output Device(s)
216
Input Device(s)
214
Communication Connection(s)
212

Fig. 2