



- (51) International Patent Classification:  
**G06K 9/00** (2006.01) **G06K 9/20** (2006.01)
- (21) International Application Number:  
PCT/US2015/056447
- (22) International Filing Date:  
20 October 2015 (20.10.2015)
- (25) Filing Language: English
- (26) Publication Language: English
- (71) Applicant: **HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.** [US/US]; Hewlett-Packard Development Company, L.P., 11445 Compaq Center Drive West, Houston, Texas 77070 (US).
- (72) Inventors: **MACHADO, Leonardo A.**; Av. Ipiranga, 6.681 Predio 95 -5, 6 e 7 andares, 90619-900 Porto Alegre (BR). **THIELO, Marcelo R.**; Av. Ipiranga, 6.681, Predio 95 -5, 6 e 7 andares, 90619-900 Porto Alegre (BR). **TRARBACH, Lisandro L.**; Av. Ipiranga, 6.681, Predio 95 -5, 6 e 7 andares, 90619-900 Porto Alegre (BR).
- (74) Agents: **SU, Benjamin** et al.; Hewlett-Packard Company, Intellectual Property Administration, 3404 East Harmony Road, Mail Stop 35, Fort Collins, Colorado 80528 (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, 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, ST, 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: DIGITIZED DOCUMENT CLASSIFICATION

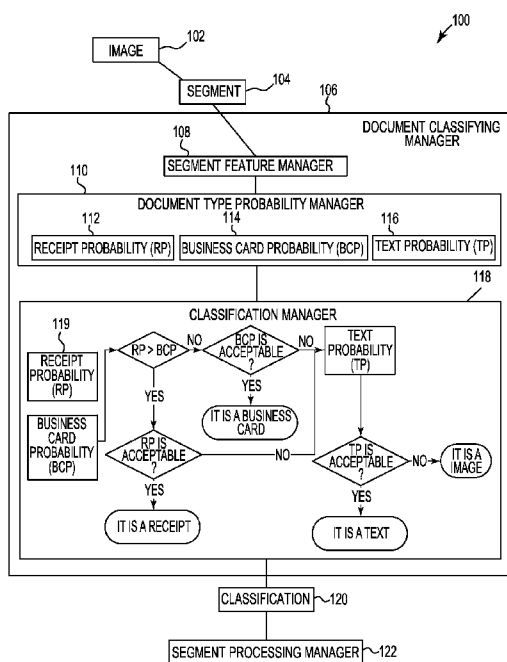


Fig. 1

(57) Abstract: An example system includes a feature engine to determine a plurality of features of a document in a segment of an image. The system also includes a probability engine to determine a probability that the document is a particular document type based on the plurality of determined features. The system also includes a classify engine to classify the segment as the particular document type or a generic document type based on at least one of the determined probabilities.

**Declarations under Rule 4.17:**

- *as to the identity of the inventor (Rule 4.17(i))*
- *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

**Published:**

- *with international search report (Art. 21(3))*

## DIGITIZED DOCUMENT CLASSIFICATION

### Background

**[0001]** In software applications that utilize and/or process digitized documents for printing, viewing or editing, the type of operation performed by the software application may depend on the type of document being utilized and/or processed.

### Brief Description of the Drawings

**[0002]** Figure 1 illustrates an example of an environment for classifying a digitized document according to the present disclosure.

**[0003]** Figure 2 illustrates a diagram of an example of a system for classifying a digitized document according to the present disclosure.

**[0004]** Figure 3 illustrates a diagram of an example of a computing device according to the present disclosure.

**[0005]** Figure 4 illustrates a diagram of an example of a system for classifying a digitized document according to the present disclosure.

### Detailed Description

**[0006]** Document classifiers may be utilized to identify digitized documents to perform different operations upon. For example, a user may want to perform optical character recognition (OCR) to extract text from receipts or business cards, or to automatically enhance a plurality of photos from a holiday trip, but not enhance images of receipts and business card. Classifiers may be utilized to efficiently discriminate between a plurality of documents that may otherwise have a same digital format (e.g., images of a receipt, images of a business card, images of a text document, images of an object, etc.).

**[0007]** Document classification of digitized documents may involve black-box approaches that operate based on a relatively large amount of time consuming user

input such as feature selection, filtering training data, filtering test data, running training to evaluate the results and iteration of the training until a percentage of hits reaches an acceptable quantity. Even after such an investment of resources, document classification may achieve the desired amount of correctness. Further, document classification may include a relatively large expenditure of computational resources, thereby increasing processing time, by utilizing computational intensive OCR, linguistic analysis, and vast trained classification database comparison to arrive at a classification.

**[0008]** In contrast to the document classification systems described above that involve additional user input, programming knowledge, training time, consumption of computation and memory resources in order to classify document, examples included herein decrease computational overhead, decrease the time it takes to make a classification, and streamlines the process of digitized document classification from the perspective of a user. For example, a system for classifying a digitized document may include determining a plurality of features of a document in a segment of an image, determining a probability that the document is a particular document type of a plurality of document types based on the number of determined features, and classifying the segment as one of the particular document type of the plurality of document types and a generic document type based on the determined probabilities.

**[0009]** Figure 1 illustrates an example environment 100 suitable for classifying a digitized document. The environment 100 is shown to include an image 102, a segment 104, a document classifying manager 106, a segment feature manager 108, a document type probability manager 110, a receipt probability (RP) 112, a business card probability (BCP) 114, a text probability (TP) 116, a classification manager 118, a classification 120, and a segment processing manager 122. The example environment may be suitable for classifying a digitized document utilizing a system (e.g., system 230 as referenced in Figure 2, system 450 as referenced in Figure 4, etc.) and a computing device (e.g., computing device 340 as references in Figure 3).

**[0010]** The environment 100 may include an image 102. The image 102 may be an image captured by an image capturing device. The image capturing device may be a digital camera, a depth camera, an infrared sensor, and/or various other image capturing devices. The image 102 may be a single image and/or a plurality of

images fused together. Further, the image 102 may be an image and/or model created from a plurality of different image capturing devices and/or imaging capturing technologies on a single device. For example, the image 102 may be a fusion of a two-dimensional (2D) image captured utilizing a red green blue (RGB) camera fused with and/or overlaid upon a three-dimensional (3D) point cloud model extracted from a depth camera and infrared camera imaging capturing device.

**[0011]** The image 102 may include a digital representation of a portion of a physical object, a single physical object, and/or a plurality of physical objects. The objects may include any physical object including documents. Documents may include a piece of written, printed, or electronic matter that provides information and/or serves as a record. For example, a document may be a receipt, a business card, a text document, and/or an object image document. A receipt may include a record of payment and/or another exchange for which a record is common. A business card may include a relatively small (e.g., sized for insertion into a wallet, pocket, etc.) card printed with contact information such as a name, a professional occupation, a company position, a business address, a phone number, an email address, etc. A text document may include a text-dense document such as a newspaper article, a book page, a word processing document, a patent application, etc.). An object image document may include a photograph, an illustration, a picture in a magazine, etc. The object image document may be an image of an object without any text present and/or captured.

**[0012]** The segment 104 may be a segment of the image 102. A segment 104 may be a portion of the image 102 partitioned from the other portions of the image 102. For example, an image 102 may be partitioned into multiple segments 104, each including a set of pixels (e.g., super pixels). The segment 104 of the image 102 may be formed (e.g., the individual pixels may be grouped into a super pixel) based on boundaries (e.g., lines, curves, etc.) of the objects contained within the segment 104. The segment 104 may be a portion of the image 102 corresponding to the object. That is, the segment 104 may be substantially limited to including only pixels that include the object. In furtherance of the above mentioned examples, the segment 104 may be a digitized document (e.g., the portion of the image 102 substantially consisting of the pixels of the image corresponding to the physical document). The segment 104 may be a separate computer file and or digital representation extracted from the image 102 and/or may be a highlighted or

otherwise indicated portion of the image 102. Further, where the image 102 is a fusion of a plurality of image captures, the segment 104 may include portions of the plurality of images captures fused into a single segment 104.

**[0013]** The segment 104 may be received by the document classifying manager 106. The document classifying manager 106 may receive the segment 104 in substantially real time (e.g., within a fraction of a second from image 102 capture). The document classifying manager 106 include hardware and/or a software application that may utilize a processing resource to execute instructions to classify the segment 104 and/or the digitized document included therein.

**[0014]** The document classifying manager 106 may include a segment feature manager 108. The segment feature manager 108 may determine a feature and/or a plurality of features of the object appearing in the segment 104. For example, the segment feature manager 108 may determine a feature and/or a plurality of features of a document appearing in the segment 104.

**[0015]** As used herein, a feature may include a physical property of an object appearing in the segment 104. For example, a feature may include physical properties such as a physical dimension and/or a plurality of physical dimensions of the object (e.g., document) appearing in the segment 104. A physical dimension of the object may be communicated to and/or received by the document classifying manager 106 as a part of and/or associated with the segment 104. Alternatively, a physical dimension and/or a plurality of physical dimensions of the object may be determined by the segment feature manager 108 utilizing a relationship (e.g., known, received from an image capturing device, included with the segment 104, etc.) between pixel dimensions within the segment and device dot-per-inch (DPI) resolution for the image capturing device that captures the image 102 and/or for the segment 104. The segment feature manager 108 may determine a physical dimension and/or a plurality of physical dimensions of the object appearing in the segment 104 by converting the pixel dimensions of the object to corresponding physical dimensions utilizing the device DPI factor. A physical dimension may be determined in any units of measurement (e.g., inches, centimeters, millimeters, etc.). A physical dimension may include a dimension corresponding to a height, length, width, radius, circumference, etc. of the physical object.

**[0016]** Another example of the feature and/or plurality of features may include a physical property such as an ink coverage for the segment 104. Determining an

ink coverage for the segment 104 may include determining a portion and/or percentage of the segment 104 that is covered by ink. Such a determination may be based on an assumption that a background (e.g., the portion of the document that does not include printing, but is still part of the document) of an object (e.g., document) is un-patterned and/or is light or otherwise in some high-contrast to the printing thereupon. Determining the ink coverage may also include determining a disposition and/or uniformity of distribution of ink upon the object.

**[0017]** Yet another example of the feature and/or plurality of features may include physical properties such as a quantity and a format of an identified candidate text box and/or a plurality of identified candidate text boxes. Determining the quantity and the format of a candidate text box may include identifying a plurality of candidate text boxes. Identifying a candidate text box may include recognizing portions of the segment 104 that are candidates for corresponding to a block or grouping of text. That is, identifying a candidate text box may include recognizing text in the segment 104 and characterizing the text as a candidate for incorporation into a text box. Text portions of a segment 104 may be recognized based on contrast of the text with a background. Determining the quantity of text boxes may include adding the amount of text boxes appearing in the segment 104 to arrive at the quantity. Determining a format of the identified candidate box may include recognizing the format of the candidate box within the segment 104. For example, determining the format may include recognizing the layout of text and/or text box candidate/s across the segment 104. A specific example may include recognizing the parallel and aligned item description and price columns of a receipt based on the disposition of a text box associated with each description/price and/or item description column/price column. Determining the quantity and the format of the candidate text boxes may reduce the incidences of false positives in text box detection by factoring in the quantity and format of identified candidate text in the recognition of text boxes.

**[0018]** Another example of the feature and/or plurality of features may include a physical property such as the edge energy detected inside each identified candidate text box of the segment 104. Determining an edge energy may include applying an edge detector to detect a quantity of edge energy within an identified candidate text boxes. The edge energy may be a value representing a quantity of high contrast points within an identified candidate text box. Since text generally has

a high edge energy owed to the contrast of the text with the background that exists in order to make the text visible and readable to the human eye. That is, text generally appears as black characters against a white background yielding a relatively high edge energy. Whereas, a blue object on a green background appearing in a segment would have relatively low edge energy compared to the black and white text edge energy level. An identified candidate text box edge energy may be compared to a threshold edge energy indicative of a minimum level of edge energy corresponding to a legitimate text box (e.g., non-false positive text box). An identified candidate text box with an edge energy below the threshold edge energy level may be discarded from consideration as a legitimate text box. Moreover, determining an edge energy may include analyzing a contour and/or a regularity of each edge recognized within the identified candidate text box. A contour and/or regularity of an edge may indicate if an identified candidate text box is text. For example, if an edge is relatively smooth then it may not be an edge of a text.

**[0019]** The document classifying manager 106 may also include a document type probability manager 110. The document type probability manager 110 may receive data related to the features of the object in the segment 104 from the segment feature manager 108. The document type probability manager 110 may determine a receipt probability (RP) 112, a business card probability (BCP) 114, and a text probability (TP) 116 based on the features of the object appearing in the segment 104.

**[0020]** Determining an RP 112 may include determining a probability that an object (e.g., a document) appearing in the segment 104 is a receipt. As described above, the RP may be determined based on the features of the object.

**[0021]** For example, the RP 112 may be based on a physical dimensions and/or a plurality of physical characteristics of the object (e.g., document) appearing in the segment 104. For example, a physical dimension of the object may be compared to a corresponding physical dimension standard for receipts. In an example, a width of an object appearing in a segment 104 may be compared to a standard width for receipts (e.g., 3 inches, 3 1/8 inches, 2 1/4 inches, etc.). While the length of the object may be collected and compared, it may be of less weight in the comparison since a receipt may have various different lengths depending on the quantity of items sold. The closer the physical dimensions of the object appearing in the segment 104 are to the standards, the higher the RP 112 may be.

**[0022]** The RP 112 may be based on the ink coverage of the object appearing in the segment 104. The ink coverage may be compared to corresponding ink coverage standards for receipts. For example, a receipt may have a substantially uniform type of ink distribution. Therefore, the closer the ink coverage is to the ink coverage standard for receipts the higher the RP 112 may be.

**[0023]** The RP 112 may be based on the quantity and format of identified candidate text boxes. The quantity and format of the identified candidate text boxes of the object may be compared to standard identified candidate text box quantities and/or formats. For example, receipts may have a substantially standard quantity and format of text boxes. As an example, receipts may have substantially parallel and evenly spaced columns having equal quantities of items in both columns. The columns may correspond to a column of item descriptions and a column of corresponding prices of the items. Therefore, a quantity and a format of identified candidate text boxes may be able to be compared to these standard text box quantities and formats. Therefore, the closer the quantity and/or format of identified candidate text boxes to the standard identified candidate text box quantities and/or formats for receipts, the higher the RP 112.

**[0024]** The RP 112 may also be based on the edge energy within the identified candidate text boxes. The edge energy for identified candidate text boxes may be compared to a standard candidate text box edge energy for receipts. This comparison may reduce the amount of false positive text boxes and therefore increase the accuracy of the quantity and format of identified candidate text boxes discussed above by disqualifying candidate boxes from classification as a text box. The closer the edge energy of the identified candidate text boxes to a standard text box edge energy for receipts, the higher the RP 112.

**[0025]** Determining a BCP 114 may include determining a probability that an object (e.g., a document) appearing in the segment 104 is a business card. As described above, the BCP 114 may be determined based on the features of the object.

**[0026]** For example, the BCP 114 may be based on a physical dimension and/or a plurality of physical dimensions of the object (e.g., document) appearing in the segment 104. For example, the physical dimensions of the object may be compared to corresponding physical dimension standards for business cards. In an example, a length and a width of an object appearing in a segment 104 may be

compared to a standard length and width for business cards (e.g., 3.5 inches x 2 inches, etc.). The closer the physical dimensions of the object appearing in the segment 104 are to the standards, the higher the BCP 114 may be.

**[0027]** The BCP 114 may be based on the ink coverage of the object appearing in the segment 104. The ink coverage may be compared to corresponding ink coverage standards for business cards. For example, a business card may have a substantially non-uniform distribution but a predictable format for the distribution. Therefore, the closer the ink coverage is to the ink coverage standard for business cards the higher the BCP 114 may be.

**[0028]** The BCP 114 may be based on the quantity and format of identified candidate text boxes. The quantity and format of the identified candidate text boxes of the object may be compared to standard identified candidate text box quantities and/or formats. For example, business cards may have a substantially standard quantity and format of text boxes. As an example, business card may prominently feature a centered logo or off-center logo, a quantity of lines of text of a first size and/or font (e.g., business name, professional title, name, etc.), and a quantity of lines with a second size and or font spaced closer together (e.g., contact information, etc.) appearing below the quantity of lines of the first size. Therefore, a quantity and a format of identified candidate text boxes may be able to be compared to such standard text box quantities and formats for business cards. Therefore, the closer the quantity and/or format of identified candidate text boxes to the standard identified candidate text box quantities and/or formats for business cards, the higher the BCP 114.

**[0029]** The BCP 114 may also be based on the edge energy within the identified candidate text boxes. The edge energy for identified candidate text boxes may be compared to a standard candidate text box edge energy for business cards. This comparison may reduce the amount of false positive text boxes and therefore increase the accuracy of the quantity and format of identified candidate text boxes discussed above by disqualifying candidate boxes with relatively low edge energy from classification as a text box. The closer the edge energy of the identified candidate text boxes to a standard text box edge energy for business cards, the higher the BCP 114.

**[0030]** The document type probability manager 110 may determine the RP 112, BCP 114, TP 116 at substantially the same time and/or substantially serially in

an automatic fashion. However, the document type probability manager 110 may instead determine the RP 112 and BCP 114 before progressing to determine TP 116. If the probabilities are substantially high enough (e.g., exceed a threshold) the document type probability manager 110 may acknowledge that the object appearing in the segment 104 is highly likely to be one of the receipt or business card document type. If neither the RP 112 nor the BCP 114 values are high enough (e.g., exceed a threshold) the TP 116 may be determined. Computational resources may be conserved in this manner, since not all documents will have the TP 116 determined.

**[0031]** Determining a TP 116 may include determining a probability that an object (e.g., a document) appearing in the segment 104 is a text document. As described above, the TP 116 may be determined based on the features of the object.

**[0032]** For example, the TP 116 may be based on a physical dimension and or a plurality of physical dimensions of the object (e.g., document) appearing in the segment 104. For example, the physical dimensions of the object may be compared to corresponding physical dimension standards for text documents. In an example, a length and a width of an object appearing in a segment 104 may be compared to a standard length and width for a text document such as a page of a book (e.g., 6 inches x 9 inches, etc.). The closer the physical dimensions of the object appearing in the segment 104 are to the standards for a text document, the higher the TP 116 may be.

**[0033]** The TP 116 may be based on the ink coverage of the object appearing in the segment 104. The ink coverage may be compared to corresponding ink coverage standards for text documents. For example, a text document such as a newspaper article may have a substantially non-uniform distribution but a predictable format for the distribution. Therefore, the closer the ink coverage is to the ink coverage standard for a text document the higher the TP 116 may be.

**[0034]** The TP 116 may be based on the quantity and format of identified candidate text boxes. The quantity and format of the identified candidate text boxes of the object may be compared to standard identified candidate text box quantities and/or formats of a text document. For example, newspaper articles may have a substantially standard quantity and format of text boxes. As an example, a newspaper article may feature a title line, a credits line, the text of the article substantially dominating a page (sometimes in a two column format), and an image.

Therefore, a quantity and a format of identified candidate text boxes may be able to be compared to such standard text box quantities and formats for text documents. Therefore, the closer the quantity and/or format of identified candidate text boxes to the standard identified candidate text box quantities and/or formats for text documents, the higher the TP 116.

**[0035]** The TP 116 may also be based on the edge energy within the identified candidate text boxes. The edge energy for identified candidate text boxes may be compared to a standard candidate text box edge energy for text documents. This comparison may reduce the amount of false positive text boxes and therefore increase the accuracy of the quantity and format of identified candidate text boxes discussed above by disqualifying candidate boxes with relatively low edge energy from classification as a text box. The closer the edge energy of the identified candidate text boxes to a standard text box edge energy for text documents, the higher the TP 116.

**[0036]** The document classifying manager 106 may include a classification manager 118. The classification manager 118 may receive the RP 112, the BCP 114, and the TP 116 as inputs from the document type probability manager 110. The classification manager 118 may utilize the RP 112, the BCP 114, and the TP 116 in achieving a classification 120 of the segment 104. The classification manager 118 may utilize the cascade binary classifier decision tree 119 illustrated in Figure 1.

**[0037]** As illustrated in the decision tree 119, the classification manager 118 may compare the RP 112 to the BCP 114. The classification manager 118 may determine that the RP 112 is greater than the BCP 114 as a result of this comparison. If RP 112 is greater than BCP 114 then the classification manager 118 may determine whether the RP 112 is acceptable. An RP 112 may be acceptable when the RP 112 meets or exceeds a RP threshold acceptability value. Therefore, the classification manager 118 may compare the RP 112 to the RP threshold value. In this example, if the RP 112 meets or exceeds the RP threshold value then the segment 104 to which the RP 112 corresponds may receive a classification 120 as a receipt. Conversely, if the RP 112 does not meet or exceed the RP threshold value then the segment 104 to which the RP 112 corresponds will not receive a classification 120 as a receipt and its TP 116 will be analyzed.

**[0038]** Alternatively, the classification manager 118 may determine that the RP 112 is less than the BCP 114 as a result of the comparison of the two. If the RP

112 is less than the BCP 114, the classification manager 118 may determine whether the BCP 114 is acceptable. A BCP 114 may be acceptable when the BCP 114 meets or exceeds a BCP threshold acceptability value. Therefore, the classification manager 118 may compare the BCP 114 to the BCP threshold value. In this example, if the BCP 114 meets or exceeds the BCP threshold value then the segment 104 to which the BCP 114 corresponds may receive a classification 120 as a business card. Conversely, if the BCP 114 does not meet or exceed the BCP threshold value then the segment 104 to which the BCP 114 corresponds will not receive a classification 120 as a business card and its TP 116 will be analyzed.

**[0039]** As described above, if the RP 112 and/or the BCP 114 are determined to be unacceptable based on their comparison with a corresponding acceptability threshold, then a TP 116 may be analyzed. In such examples, the classification manager 118 may determine whether the TP 116 is acceptable. A TP 116 may be acceptable when the TP 116 meets or exceeds a TP threshold acceptability value. Therefore, the classification manager 118 may compare the TP 116 to the TP threshold value. In this example, if the TP 116 meets or exceeds the TP threshold value then the segment 104 to which the TP 116 corresponds may receive a classification 120 as a text document. Conversely, if the TP 116 does not meet or exceed the TP threshold value then the segment 104 to which the TP 116 corresponds may receive a classification 120 as a generic document type (e.g., an image document type).

**[0040]** As described above, the classification 120 of the segment 104 as one of a receipt, a business card, a text document, and a generic image document may be output by the document classifying manager 106. The classification 120 of a particular segment 104 as one of these document types may be received by a segment processing manager 122. In some examples, the segment processing manager 122 may be part of and/or incorporated within document classifying manager 106.

**[0041]** A segment processing manager 122 may process the segment 104 in a variety of ways determined by the classification 120 associated with that segment 104. For example, the segment processing manager 122 may organize the storage of the segment 104 based on the classification 120. In an example, organizing the storage of the segment 104 may include saving the segment 104 in a particular location (e.g., database, folder, file, etc.), as a particular file type, with a particular

name, with a particular identifier, and/or associated with a particular software program determined by the classification 120. For example, a segment 104 with a receipt classification 120 may be saved into a folder for receipts and saved associated (e.g., a compatible file type, as a resource for use by, etc.) with expense management software.

**[0042]** Moreover, the segment processing manager 122 may organize segments 104 according to places identified and people that appear within the segment 104.

**[0043]** A segment processing manager 122 may extract particular data from the segment 104 based on its classification 120. Up to this point, the environment 100 has not included the utilization of any optical character recognition tools. However, the extraction and further use of the extracted data may involve the use of optical character recognition. For example, for a segment 104 with a classification 120 as a business card, the segment processing manager 122 may automatically integrate a contact detail and/or a plurality of contact details extracted from the business card appearing in the segment 104. The segment processing manager 122 may populate the extracted contact details into a contact management database. Further, the segment processing manager 122 may transform extracted contact details into interact-able data where a user could call and/or email a desired contact with a selection (e.g., a click/touch of the extracted data).

**[0044]** In another example, for a segment 104 with a classification 120 as a receipt, the segment processing manager 122 may automatically integrate an expense detail into an expense control database. In this manner, expenses may be recognized and stored according to type, value, date, etc.

**[0045]** In yet another example, the segment processing manager 122 may extract and organize portions of a text document. For example, specific words, URLs, emails, sentences, quantity of words, etc. may be extracted from a segment 104 with a text document classification and stored according to the classification 120.

**[0046]** Figure 2 illustrates a diagram of an example system 230 for classifying a digitized document. The system 230 may include a database 232, a document classifying manager 234 and/or an engine and/or a plurality of engines (e.g., feature engine 236, probability engine 238, classify engine 240, etc.). The document

classifying manager 234 may include additional or fewer engines than are illustrated to perform the various functions as will be described in further detail.

**[0047]** An engine and/or a plurality of engines (e.g., feature engine 236, probability engine 238, classify engine 240, etc.) may include a combination of hardware and programming (e.g., instructions executable by the hardware), but at least hardware, that is configured to perform functions described herein (e.g., determining a plurality of features of a document in a segment of an image, determining a probability that the document is a particular document type based on the plurality of determined physical features, and classifying the segment as the particular document type or a generic document type based on the determined probability, etc.). The programming may include program instructions (e.g., software, firmware, etc.) stored in a memory resource (e.g., computer readable medium, machine readable medium, etc.) as well as hard-wired program (e.g., logic).

**[0048]** The feature engine 236 may include hardware and/or a combination of hardware and programming, but at least hardware to determine a features and/or a plurality of features of a document in a segment of an image. A feature may include a physical property such as a physical dimension of the document. The physical dimension of the document may be determined based on a pixel dimension and a dot-per-inch resolution of an image capturing device that captured the image from which the segment being processed was segmented.

**[0049]** The probability engine 238 may include hardware and/or a combination of hardware and programming, but at least hardware to determine a probability that the document is a particular document type. The particular document type may be a document type of a plurality of document types. The probability that the document is a particular document type may be based on the determined feature and/or plurality of features of the document. For example, the particular document types may include a receipt, a business card, and a text document. The probability that the document in the segment of the image is a particular one of these types of documents may be based at least in part on a comparison of the physical dimensions of the document to a standard dimension respectively associated with each of the document types (e.g., a standard dimension associated with a receipt, a standard dimension associated with a business card, a standard dimension associated with a text document, etc.).

**[0050]** The classify engine 240 may include hardware and/or a combination of hardware and programming, but at least hardware to classify the segment as the particular document type or a generic document type based on the determined probability. In an example, the segment may be classified as one of the particular document type of the plurality of document types and a generic document type based on a set of determined probabilities. The set of determined probabilities may include a plurality of probabilities where each probability is a probability that the document is a corresponding document type (e.g., a receipt probability that the document is a receipt, a business card probability that the document is a business card, a text probability that a document is a text document, etc.) Classifying the segment may include comparing the probability that the document in the segment is a particular type of document to a respective threshold probability. For example, this may include comparing the probability that the document is a business card with a threshold probability for business cards, past which the document may be characterized as a business card. The generic document type may be a default or catch-all document type such as an image document type.

**[0051]** Figure 3 illustrates a diagram of an example of a computing device 350 according to the present disclosure. The computing device 350 may utilize software, hardware, firmware, and/or logic to perform functions described herein.

**[0052]** The computing device 350 may be any combination of hardware and program instructions to share information. The hardware, for example, may include a processing resource 352 and/or a memory resource 354 (e.g., non-transitory computer-readable medium (CRM), machine readable medium (MRM), database, etc.). A processing resource 352, as used herein, may include any quantity of processors capable of executing instructions stored by a memory resource 354. Processing resource 352 may be implemented in a single device or distributed across multiple devices. The program instructions (e.g., computer readable instructions (CRI)) may include instructions stored on the memory resource 354 and executable by the processing resource 352 to implement a desired function (e.g., determine a plurality of physical features of a document in a segment of an image, determine a probability that the document is a particular document type based on the plurality of determined physical properties, classify the segment as the particular document type based on a comparison of the determined probability to a probability

that the document is a different document type and a comparison of the determined probability to a corresponding probability threshold, etc.)

**[0053]** The memory resource 354 may be in communication with the processing resource 352 via a communication link (e.g., a path) 356. The communication link 356 may be local or remote to a machine (e.g., a computing device) associated with the processing resource 352. Examples of a local communication link 356 may include an electronic bus internal to a machine (e.g., a computing device) where the memory resource 354 is one of volatile, non-volatile, fixed, and/or removable storage medium in communication with the processing resource 352 via the electronic bus.

**[0054]** A module and/or a plurality of modules (e.g., feature module 358, probability module 360, classify module 362, etc.) may include CRI that when executed by the processing resource 352 may perform functions. The module and/or a plurality of modules (e.g., feature module 358, probability module 360, classify module 362, etc.) may be sub-modules of other modules. For example, the probability module 360 and classify module 362 may be sub-modules and/or contained within the same computing device. In another example, the module and/or a plurality of modules (e.g., feature module 358, probability module 360, classify module 362, etc.) may comprise individual modules at separate and distinct locations (e.g., CRM, etc.).

**[0055]** Each of the modules (e.g., feature module 358, probability module 360, classify module 362, etc.) may include instructions that when executed by the processing resource 352 may function as a corresponding engine as described herein. For example, the feature module 358, probability module 360, and classify module 362 may include instructions that when executed by the processing resource 352 may function as the feature engine 236, the probability engine 238, and the classify engine 240, respectively.

**[0056]** The feature module 358 may include CRI that when executed by the processing resource 352 may determine a plurality of features of a document in a segment of an image. The segment of the image may be one of a plurality of segments from a single image, each segment corresponding to one of a plurality of documents in the image. The plurality of features may include a physical property such as a physical dimension, a global and/or per document region density of ink

coverage, a format and a quantity of identified text boxes, and an edge energy of a document in a segment of an image.

**[0057]** The probability module 360 may include CRI that when executed by the processing resource 352 may determine a probability that the document is a particular type of document based on the determined feature and/or plurality of determined features. For example, the determined probability may be a probability that the document is the particular document. In another example, the determined probability may be a set of document type probabilities (e.g., a probability that the document is a business card type, a probability that the document is a receipt type, a probability that the document is text type) each one based on the determined feature and/or plurality of determined features. The probability that the document is a text document may be determined after a determination that the probability that the document is a business card or a receipt has been determined to be unacceptable.

**[0058]** The classify module 362 may include CRI that when executed by the processing resource 352 may classify the segment as the particular document type based on the determined probability. For example, the segment may be classified as a particular document type of a plurality of document types (e.g., a business card type, a receipt type, a text type, etc.) based on corresponding determined document type probabilities (e.g., a probability that the document is a business card type, a probability that the document is a receipt type, a probability that the document is text type). For example, the segment may be classified as a business card, a receipt, a text document, or an image based on the corresponding determined document type probabilities (e.g., a probability that the document is a business card type, a probability that the document is a receipt type, a probability that the document is text type). The classification may proceed as a cascade binary classifier decision tree. If the segment fails to be classified as one of a receipt, a business card, and a text document, then the segment may be classified as an image document. Classifying the segment utilizing the cascade binary classifier decision tree may include comparison of the determined probability that the document is a particular document type to a probability that the document is a different document type. For example, the probability that the document is a receipt may be compared to the probability that the document is a business card. Further, classifying the segment utilizing the cascade binary classifier decision tree may include comparison of the determined probability that the document is a particular document type to a corresponding

probability threshold. To continue the above previous example, if the probability that the document is a receipt is greater than the probability that the document is a business card, then the probability that document is a receipt may be compared to a receipt probability threshold. The threshold may be a probability beyond which the document may be classified as a receipt and below which the probability that the document is a text document is determined and compared to a text probability threshold.

**[0059]** In some examples, the classify module 362 may include CRI that when executed by the processing resource 352 may automatically integrate a contact detail extracted from the document into a contact detail database based on a classification of the segment as the business card. In another example, the classify module 362 may include CRI that when executed by the processing resource 352 may automatically integrate an expense detail into an expense control database based on a classification of the segment as the receipt.

**[0060]** Figure 4 illustrates a diagram of an example system 450 for classifying a digitized document. The system 450 may include a database 452, a document classifying manager 454 and/or an engine and/or a plurality of engines (e.g., feature engine 456, probability engine 458, classify engine 460, extract engine 462, etc.). The document classifying manager 454 may include additional or fewer engines than are illustrated to perform the various functions as will be described in further detail.

**[0061]** The engine and/or plurality of engines (e.g., feature engine 456, probability engine 458, classify engine 460, extract engine 462, etc.) may include a combination of hardware and programming (e.g., instructions executable by the hardware), but at least hardware, that is configured to perform functions described herein (e.g., determine a physical dimension, an ink coverage, a format and a quantity of identified text boxes, and an edge energy of a document in a segment of an image, determine a probability that the document is one of a business card, a receipt, and a text document based on the determined physical dimension, the determined ink coverage, the determined format and the determined quantity of identified text boxes, and the determined edge energy of the document, classify the segment as one of a business card, a receipt, a text document, and an image based on the determined probabilities, and extract data from the document in the segment based on the classification of the segment etc.). The programming may include program instructions (e.g., software, firmware, etc.) stored in a memory resource

(e.g., computer readable medium, machine readable medium, etc.) as well as hard-wired program (e.g., logic).

**[0062]** The feature engine 456 may include hardware and/or a combination of hardware and programming, but at least hardware to determine a physical feature and/or a plurality of physical features of a document in a segment of an image. Determining the physical feature and/or a plurality of physical features may include determining a physical property such as a physical dimension, an ink coverage, a format and a number of identified text boxes, and an edge energy of a document in a segment of an image. Determining the ink coverage may include quantifying a percentage of the document that is covered by ink. Determining the edge energy may include quantifying high contrast points within each identified text box. Determining the format and quantity of identified text boxes may include discarding an identified box from the determination when the box is associated with a quantity of high contrast points below a threshold.

**[0063]** The probability engine 458 may include hardware and/or a combination of hardware and programming, but at least hardware to determine a probability that the document is a particular document type of a plurality of document types based on the determined features. For example, the probability engine 458 may include hardware and/or a combination of hardware and programming, but at least hardware to determine a probability that the document is one of a business card, a receipt, and a text document based on the determined physical dimension, ink coverage, format and quantity of identified text boxes, and edge energy of the document.

**[0064]** The classify engine 460 may include hardware and/or a combination of hardware and programming, but at least hardware to classify the segment as one of a business card, a receipt, a text document, and an image based on the determined probabilities. Classifying the segment may include comparing the determined format and the quantity of identified text boxes of the document to a corresponding standard format and a corresponding standard quantity of text boxes (e.g., a standard text box format and a standard quantity of text boxes associated with each of the business card, the receipt, and the text document).

**[0065]** The extract engine 462 may include hardware and/or a combination of hardware and programming, but at least hardware to extract data from the document in the segment based on the classification of the segment. The type of data

extracted and the handling of that data may be determined by the classification of the segment whence it came.

**[0066]** As used herein, “logic” is an alternative or additional processing resource to perform a particular action and/or function, etc., described herein, which includes hardware, e.g., various forms of transistor logic, application specific integrated circuits (ASICs), etc., as opposed to computer executable instructions, e.g., software firmware, etc., stored in memory and executable by a processor. Further, as used herein, “a” or “a plurality of” something may refer to one or more such things. For example, both “a widget” and “a plurality of widgets” may refer to one or more widgets.

**[0067]** The figures herein follow a numbering convention in which the first digit or digits correspond to the drawing figure number and the remaining digits identify an element or component in the drawing. As will be appreciated, elements shown in the various embodiments herein may be added, exchanged, and/or eliminated so as to provide a plurality of additional embodiments of the present disclosure. In addition, as will be appreciated, the proportion and the relative scale of the elements provided in the figures are intended to illustrate certain embodiments of the present disclosure, and should not be taken in a limiting sense.

**[0068]** The above specification, examples and data provide a description of the method and applications, and use of the system and method of the present disclosure. Since many examples may be made without departing from the spirit and scope of the system and method of the present disclosure, this specification merely sets forth some of the many possible embodiment configurations and implementations.

What is claimed is:

1. A system for classifying a digitized document comprising:
  - a feature engine to determine a plurality of features of a document in a segment of an image;
  - a probability engine to determine a probability that the document is a particular document type based on the plurality of determined features; and
  - a classify engine to classify the segment as the particular document type or a generic document type based the determined probability.
2. The system of claim 1, wherein the plurality of features includes a physical dimension of the document.
3. The system of claim 2, wherein the probability engine is to determine the probability that the document is the particular document type based at least in part on a comparison of the physical dimension to a standard dimension associated with each of a plurality of document types.
4. The system of claim 2, wherein the feature engine is to determine the physical dimension based on a pixel dimension and a dot-per-inch resolution of a capturing device that captured the image.
5. The system of claim 1, wherein the classify engine is to classify the segment by comparing the probability that the document is the particular document type to a respective threshold probability and designating the document as the particular document type when the probability for the particular type of document exceeds the threshold probability
6. A non-transitory computer readable medium storing instructions executable by a processing resource to cause a computer to:
  - determine a plurality of physical properties of a document in a segment of an image;
  - determine a probability that the document is a particular document type based on the plurality of determined physical properties; and

classify the segment as the particular document type based on a comparison of the determined probability to a probability that the document is a different document type and a comparison of the determined probability to a corresponding probability threshold.

7. The medium of claim 6, wherein the segment corresponds to one of a plurality of documents in the image.

8. The medium of claim 6, including organizing a storage of the segment based on the classification.

9. The medium of claim 6, including automatically integrating a contact detail extracted from the document into a contact detail database based on a classification of the segment as a business card.

10. The medium of claim 6, including automatically integrating an expense detail into an expense control database based on a classification of the segment as a receipt.

11. A system for classifying a digitized document comprising:

a feature engine to determine a physical dimension, an ink coverage, a format and a quantity  $r$  of identified text boxes, and an edge energy of a document in a segment of an image;

a probability engine to determine a probability that the document is one of a business card, a receipt, and a text document based on the determined physical dimension, the determined ink coverage, the determined format and the determined quantity of identified text boxes, and the determined edge energy of the document;

a classify engine to classify the segment as one of a business card, a receipt, a text document, and an image document based on the determined probabilities; and

an extract engine to extract data from the document in the segment based on the classification of the segment.

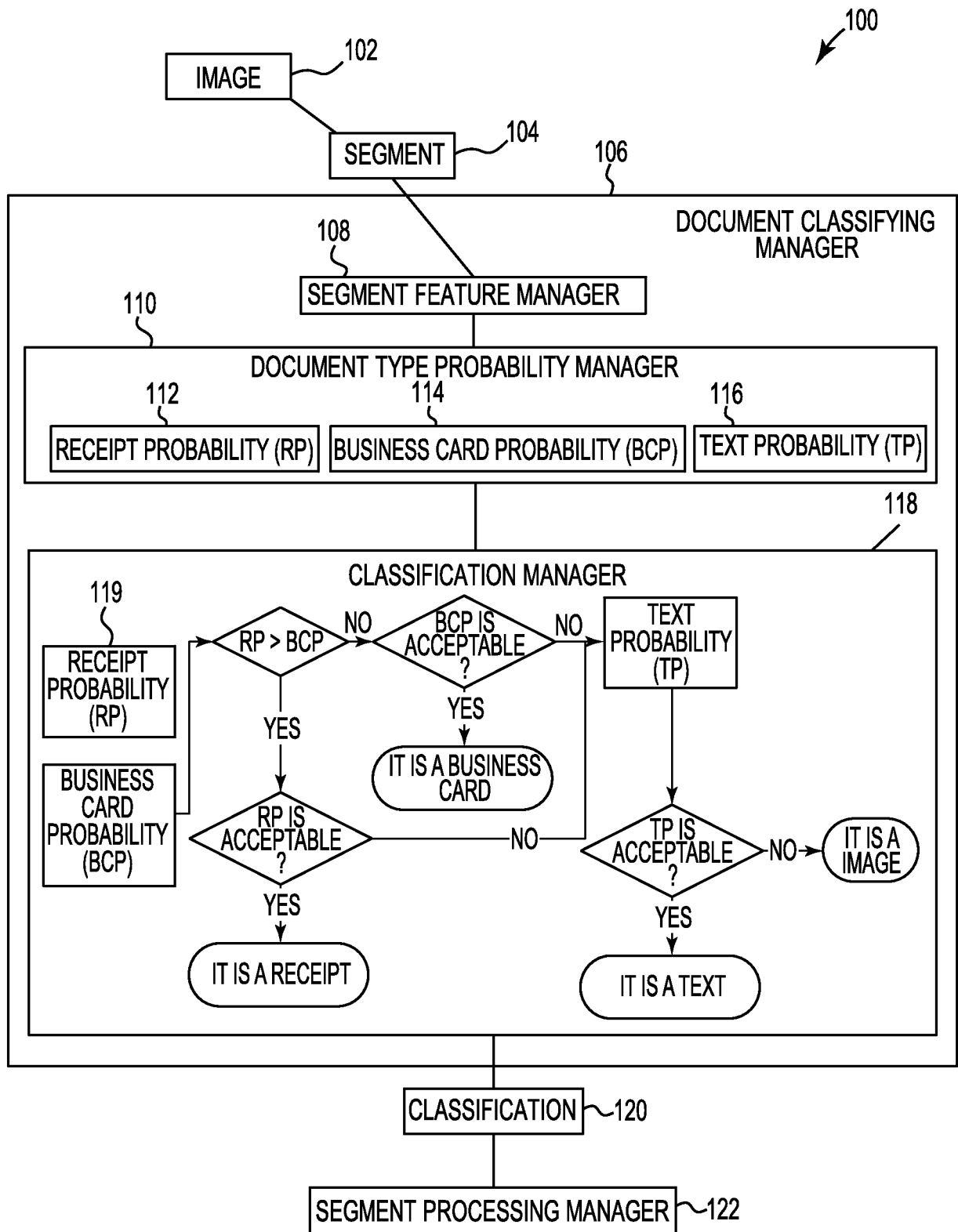
12. The system of claim 11, wherein the feature engine is to determine the ink coverage by quantifying a percentage of the document that is covered by ink.

13. The system of claim 11, wherein the classify engine is to classify the segment by comparing the determined format and the determined quantity of identified text boxes of the document to a standard format and a standard quantity of text boxes associated with each of the business card, the receipt, and the text document.

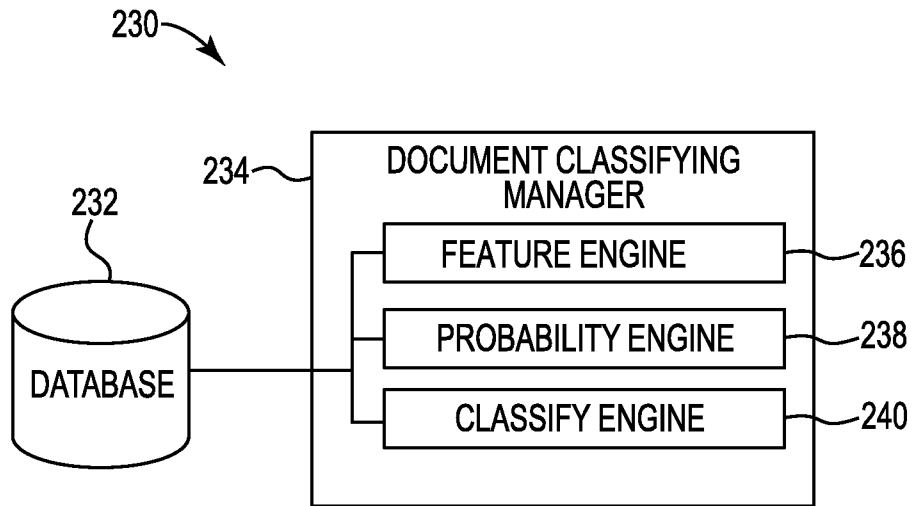
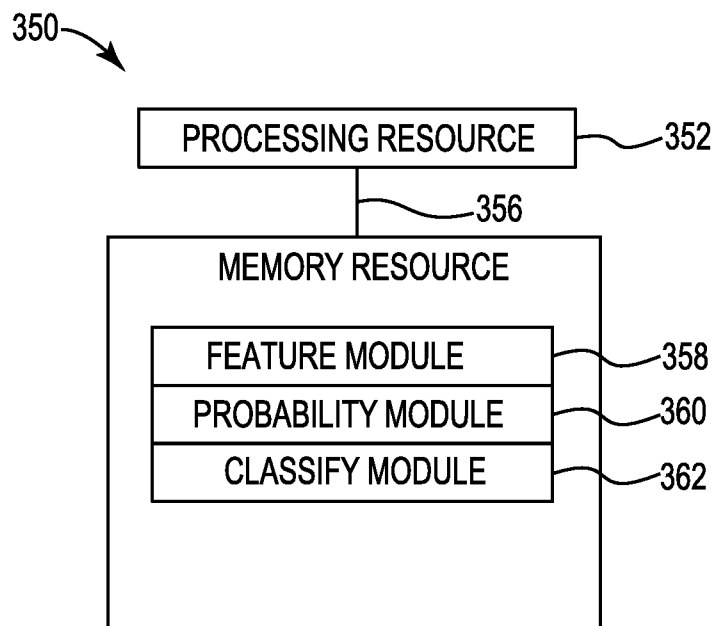
14. The system of claim 11, wherein the feature engine is to determine the edge energy of the document by quantifying high contrast points within each of the identified text boxes.

15. The system of claim 14, wherein the feature engine is to determine the format and the quantity of identified text boxes by discarding an identified box from the determination when the box is associated with a quantity of high contrast points below a threshold.

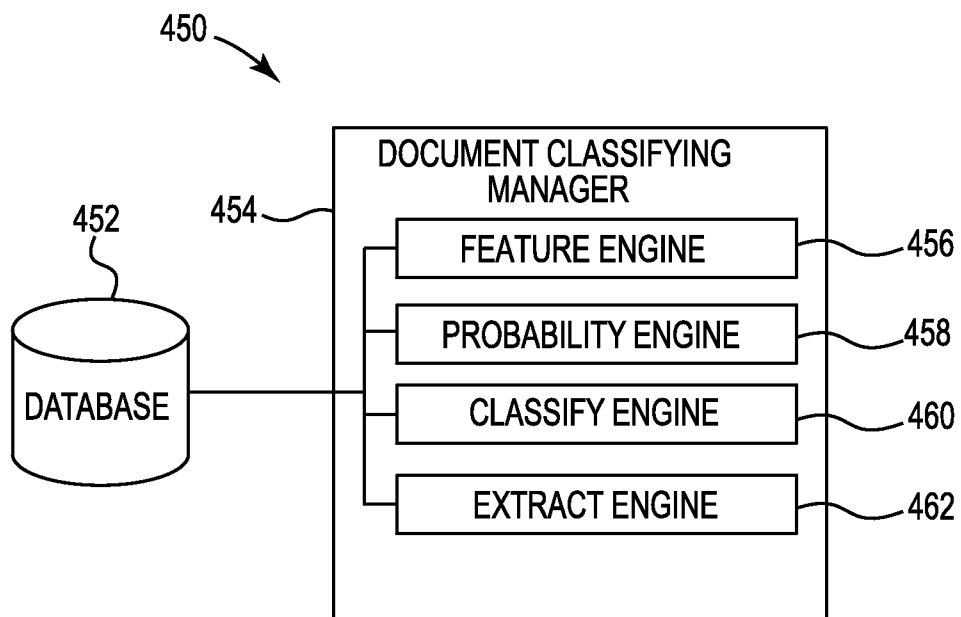
1/3

**Fig. 1**

2/3

**Fig. 2****Fig. 3**

3/3

**Fig. 4**

## INTERNATIONAL SEARCH REPORT

International application No.  
**PCT/US2015/056447****A. CLASSIFICATION OF SUBJECT MATTER****G06K 9/00(2006.01)i, G06K 9/20(2006.01)i**

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 9/00; G06K 9/68; G06F 17/30; G06F 17/60; G06F 17/21; G06K 9/62; G06K 9/34; G06K 9/20

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

Korean utility models and applications for utility models

Japanese utility models and applications for utility models

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

eKOMPASS(KIPO internal) &amp; Keywords: image, classify, segment, probability, classify

**C. DOCUMENTS CONSIDERED TO BE RELEVANT**

| Category* | Citation of document, with indication, where appropriate, of the relevant passages                                                               | Relevant to claim No. |
|-----------|--------------------------------------------------------------------------------------------------------------------------------------------------|-----------------------|
| X         | WO 2014-018001 A1 (HEWLETT-PACKARD DEVELOPMENT COMPANY, L.P.)<br>30 January 2014<br>See abstract, paragraphs [19],[22]-[29] and claims 13,15,17. | 1-4,6-8               |
| Y         |                                                                                                                                                  | 5,9-15                |
| Y         | US 2008-0152238 A1 (PRATEEK SARKAR) 26 June 2008<br>See abstract, paragraphs [18]-[23], claim 13 and figure 1.                                   | 5                     |
| Y         | US 2009-0067729 A1 (ADAM TURKELSON et al.) 12 March 2009<br>See abstract, paragraphs [23],[85],[86], claims 2-9 and figures 1-6B.                | 9-15                  |
| A         | US 2002-0085755 A1 (SU-YOUNG CHI et al.) 04 July 2002<br>See abstract, paragraphs [24]-[29] and figures 3-4B.                                    | 1-15                  |
| A         | US 2003-0050790 A1 (WILLIAM CHANG et al.) 13 March 2003<br>See abstract, paragraphs [29]-[37] and figures 2-5.                                   | 1-15                  |



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 application or patent 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

"&amp;" document member of the same patent family

Date of the actual completion of the international search

20 July 2016 (20.07.2016)

Date of mailing of the international search report

**20 July 2016 (20.07.2016)**

Name and mailing address of the ISA/KR

International Application Division

Korean Intellectual Property Office

189 Cheongsa-ro, Seo-gu, Daejeon, 35208, Republic of Korea

Facsimile No. +82-42-481-8578

Authorized officer

KANG, Sung Chul

Telephone No. +82-42-481-8405



**INTERNATIONAL SEARCH REPORT**

Information on patent family members

International application No.

**PCT/US2015/056447**

| Patent document<br>cited in search report | Publication<br>date | Patent family<br>member(s)                                                           | Publication<br>date                                                |
|-------------------------------------------|---------------------|--------------------------------------------------------------------------------------|--------------------------------------------------------------------|
| WO 2014-018001 A1                         | 30/01/2014          | CN 104487966 A<br>EP 2875446 A1<br>US 2015-0178563 A1                                | 01/04/2015<br>27/05/2015<br>25/06/2015                             |
| US 2008-0152238 A1                        | 26/06/2008          | EP 1936536 A2<br>EP 1936536 A3<br>EP 1936536 B1<br>JP 2008-159056 A<br>US 8000538 B2 | 25/06/2008<br>09/05/2012<br>24/06/2015<br>10/07/2008<br>16/08/2011 |
| US 2009-0067729 A1                        | 12/03/2009          | US 8503797 B2                                                                        | 06/08/2013                                                         |
| US 2002-0085755 A1                        | 04/07/2002          | KR 10-0411894 B1<br>KR 10-2002-0055454 A                                             | 24/12/2003<br>09/07/2002                                           |
| US 2003-0050790 A1                        | 13/03/2003          | None                                                                                 |                                                                    |