

(19) World Intellectual Property Organization  
International Bureau



(43) International Publication Date  
7 June 2007 (07.06.2007)

PCT

(10) International Publication Number  
**WO 2007/065087 A1**

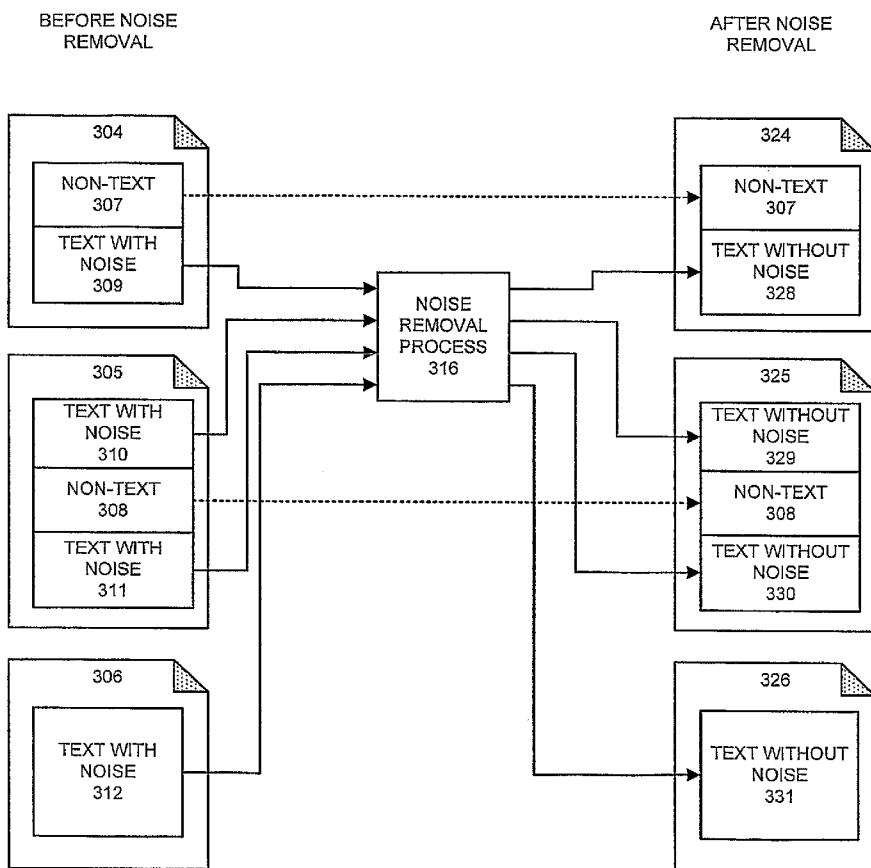
- (51) International Patent Classification:  
*G06T 5/00* (2006.01)
- (21) International Application Number:  
PCT/US2006/061294
- (22) International Filing Date:  
28 November 2006 (28.11.2006)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:  
11/291,552 30 November 2005 (30.11.2005) US
- (71) Applicant (for all designated States except US): **ADOBE SYSTEMS, INCORPORATED** [US/US]; 345 Park Avenue, San Jose, CA 95110-2704 (US).
- (72) Inventor; and
- (75) Inventor/Applicant (for US only): **NICHOLSON, Dennis, G.** [US/US]; 1 Altree Court, Atherton, CA 94207 (US).
- (74) Agent: **MEYERTONS, HOOD, KIVLIN, KOWERT & GOETZEL, P.C.**; P.O. Box 398, 700 Lavaca, Suite 800, Austin, TX 78767-0398 (US).

- (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LV, LY, MA, MD, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:  
— with international search report

[Continued on next page]

(54) Title: METHOD AND APPARATUS FOR REMOVING NOISE FROM A DIGITAL IMAGE



(57) Abstract: One embodiment of the present invention provides a system that removes noise from an image. During operation, the system first identifies blobs in the image, wherein a blob is a set of contiguous pixels which possibly represents a character or a portion of a character in the image. Next, the system analyzes the blobs to dynamically determine a "noise threshold" for the blobs. The system then removes blobs from the image which are below the noise threshold.

WO 2007/065087 A1



---

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

*For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.*

**METHOD AND APPARATUS FOR REMOVING NOISE FROM A DIGITAL IMAGE****BACKGROUND****Field of the Invention**

[0001] The present invention relates to image processing. More specifically, the present invention relates to a method and apparatus for facilitating the removal of noise from a digital image.

**Related Art**

[0002] As businesses and other organizations become more computerized, it is becoming increasingly common to store and maintain electronic versions of paper documents on computer systems. The process of storing a paper document on a computer system typically involves a "document-imaging" process, which converts a copy of the paper document into an electronic document. This document-imaging process typically begins with an imaging step, wherein document page-images are generated using a scanner, a copier, a camera, or any other imaging device. These page-images are typically analyzed and enhanced using an image-processing program before being assembled into a document container, such as a Portable Document Format (PDF) file.

[0003] Often, applications need to recognize text from the scanned page-images to facilitate subsequent document-processing operations. This is typically accomplished through an optical character recognition (OCR) process.

[0004] Unfortunately, it is very common for the performance of the OCR process to be significantly degraded by the presence of noise in scanned images. Many types of noise and noise-like artifacts arise from the printing and imaging processes. Examples of noise and noise-like artifacts may include quantization noise from the imaging light sensors, dirt on imaging device optics, ink spatters, and toner smudges.

[0005] Because of this problem, noise-removal operations are commonly applied to images prior to the OCR process. For example, a common noise-removal operation removes all blobs that are smaller than a threshold number of pixels. However, this may cause small characters such as a "period" to be removed, or may cause a particularly large noise artifact to be retained. Rarely is a fixed threshold value optimal for all character sizes. Consequently, either too much noise is left behind during the noise-removal process, or portions of a scanned image are improperly removed.

[0006] Hence, what is needed is a method and apparatus for removing noise from an image without the above-mentioned problems.

**SUMMARY**

[0007] One embodiment of the present invention provides a system that removes noise from an image. During operation, the system first identifies blobs in the image, wherein a blob is a set of contiguous pixels which possibly represents a character or a portion of a character in the image. Next, the system analyzes the blobs to dynamically determine a "noise threshold" for the blobs. The system then removes blobs from the image which are below the noise threshold.

[0008] In a variation of this embodiment, analyzing the blobs involves analyzing: the size distribution of the blobs, the number of blobs, locations of the blobs, the blob density of the image or region of the image, and colors of the blobs.

[0009] In a variation of this embodiment, the system determines the noise threshold by first identifying text regions in the image and then identifying "key characters" in the text regions, wherein a key character is a small

character or a portion of character, such as a period, an i-dot, or a comma. Next, the system computes the average size of the identified key characters, and computes the noise threshold as a fraction of this average size.

[0010] In a variation on this embodiment, prior to analyzing the blobs, the system performs an initial noise-removal operation by removing blobs from the image that are below an initial noise threshold.

[0011] In a variation on this embodiment, if the determined noise threshold is different from a previous noise threshold, the system repeats the noise-removal process. Furthermore, if the determined noise threshold is reduced from the previous noise threshold, the system restores previously removed blobs which are smaller than the previous noise threshold but larger than the determined noise threshold.

[0012] In a variation of this embodiment, a noise-threshold is determined independently for each identified text region.

#### **BRIEF DESCRIPTION OF THE FIGURES**

[0013] FIG. 1 illustrates a sample document in accordance with an embodiment of the present invention.

[0014] FIG. 2 illustrates several points-of-interest in the sample document in accordance with an embodiment of the present invention.

[0015] FIG. 3 illustrates a noise-removal process in accordance with an embodiment of the present invention.

[0016] FIG. 4 illustrates a computing environment in accordance with an embodiment of the present invention.

[0017] FIG. 5 illustrates a noise-removal system in accordance with an embodiment of the present invention.

[0018] FIG. 6 illustrates an optical-character-recognition (OCR) system in accordance with an embodiment of the present invention.

[0019] FIG. 7 presents a flowchart illustrating the noise-removal process in accordance with an embodiment of the present invention.

[0020] FIG. 8 presents a flowchart illustrating the OCR process which includes a refinement to the noise-removal process in accordance with an embodiment of the present invention.

#### **DETAILED DESCRIPTION**

[0021] The following description is presented to enable any person skilled in the art to make and use the invention, and is provided in the context of a particular application and its requirements. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the present invention. Thus, the present invention is not intended to be limited to the embodiments shown, but is to be accorded the widest scope consistent with the principles and features disclosed herein.

[0022] The data structures and code described in this detailed description are typically stored on a computer-readable storage medium, which may be any device or medium that can store code and/or data for use by a computer system. This includes, but is not limited to, magnetic and optical storage devices such as disk drives, magnetic tape, CDs (compact discs) and DVDs (digital versatile discs or digital video discs).

#### **Overview**

[0023] The present invention provides a technique for removing noise from an image, and can be applied to any document that may contain noise or any other extraneous artifacts that were not intended to be part of the original document. Although the embodiments of the present invention which are described below operate in conjunction with an optical character recognition (OCR) system, the present invention can easily be modified to work with any noise-removal system for digital images, including noise removal systems that are not associated with optical character recognition. For example, embodiments of the present invention may be applied to digital photography.

**[0024]** During operation, one embodiment of the present invention receives an image in digital form, which may contain both text and images. To improve the OCR process, it is beneficial to remove as much noise as possible from the image. The first step in removing this noise is to identify “blobs” in the image. A “blob” is defined as a cluster of adjacent pixels that may represent a character, part of a character, an object within the image, or a noise artifact. After identifying blobs in the image, the system analyzes the blobs to determine a “noise threshold” for the blobs. (For example, the noise threshold can specify a minimum number of pixels in a blob.) Next, the system removes all blobs which are smaller than the determined noise threshold. Note that a larger noise threshold value will remove more small artifacts, but may also remove characters (or portions of characters), such as periods or i-dots (e.g. the dot portion of the lower-case “i” character). In contrast, a smaller noise threshold will not remove characters (or portions of characters), but will not remove as many noise artifacts. Finally, after the blobs are removed, the system performs an OCR process to identify characters within the remaining blobs.

**[0025]** In one embodiment of the present invention, information gathered from blob analysis can include: blob size (the number of pixels in a blob), blob distribution (the number of blobs in a digital image, or in a section of a digital image), and blob spacing (the number of pixels between blobs). This information can be used to estimate text size. Additionally, text size can be estimated by identifying key characters in the image, and then determining the size of these key characters. Key characters are small characters (or portions of characters), such as periods, i-dots, commas, and ellipses. Key characters are useful in establishing a noise-removal threshold because they are typically the smallest characters (or portions of characters) within a font, and can thus be used to distinguish between characters (or portions of characters) and noise. Key characters may also include letters and numbers which are selected to assist in estimating text size. Typically, baseline spacing between blobs and blob height distribution are good indicators of text size, which can be determined without going through the entire OCR process.

**[0026]** In one embodiment of the present invention, the system determines the noise threshold by first identifying “text regions” in the image and then identifying key characters within the text regions. Next, the system computes the average size of the identified key characters, and computes the noise threshold as a fraction of this average size. Note that the list of key characters can be pre-defined by a user or system administrator. Also note that the key characters may be used to identify a font, which may subsequently help in discriminating characters from noise blobs during subsequent OCR operations.

**[0027]** In one embodiment of the present invention, the above-described process is iterative. In this embodiment, the system starts with an “initial noise threshold.” This initial noise threshold is used to perform an initial noise-removal process, wherein blobs which are smaller than the initial noise threshold are removed from the image. (Note that it is desirable to set this initial noise threshold to be smaller than any key characters in the smallest font of interest.) Next, the system analyzes the key characters as described above to determine a new noise threshold. If the new noise threshold is different than a previous noise threshold, the system repeats the noise-removal process. While repeating this process, if the determined noise threshold is reduced from a previous noise threshold, the system restores previously removed blobs which are smaller than the previous noise threshold but larger than the determined noise threshold.

**[0028]** In one embodiment of the present invention, the above-described iterative process continues until a “satisfactory” threshold level is reached. Note that a threshold may be considered “satisfactory” if: (1) the threshold does not change between iterations; (2) the threshold smaller than a predefined noise tolerance level; (3) the change

in threshold levels is within a predefined range; or (4) the noise-removal system has executed for a pre-specified amount of time.

**[0029]** In one embodiment of the present invention, the noise removal process is applied to the entire imaged document.

**[0030]** In another embodiment of the present invention, the noise-removal process is applied on a page by page basis. In this embodiment, the process can be applied to each page independently, or alternatively, statistical information can be carried over to each successive page to assist in the setting of an initial noise threshold for each successive page.

**[0031]** In one embodiment of the present invention, a page in the imaged document is divided into sections and each section is processed independently or in conjunction with other sections. This embodiment provides the most flexibility, and in most cases the best results, but may require more processing time.

**[0032]** In one embodiment of the present invention, regardless of how the noise-removal threshold changes, removed blobs are not restored during successive iterations of the noise-removal process. This is likely to decrease the accuracy of the noise-removal process because some blobs which are characters (or portions of characters) may be removed. However, it is also likely to increase the speed of the noise-removal process.

#### **Sample Document**

**[0033]** FIG. 1 illustrates a sample document in accordance with an embodiment of the present invention. This sample document contains: a header 101, a body 102 and a footnote 103. The body 102 of the sample document contains both text sections and an image which contains text. In addition, several artifacts 104 can be seen throughout the document. Note that this image 100 was created by scanning a document which already contained noise. However, it is also possible that the noise was introduced during the scanning process. Also note that the present invention can be applied to any digital image, and is not limited to scanned documents.

#### **Points of Interest**

**[0034]** FIG. 2 illustrates several points of interest in the sample document in accordance with an embodiment of the present invention. More specifically, FIG. 2A illustrates a section of the header 101 from the sample document illustrated in FIG. 1. Three points of interest in FIG. 2A are noise artifact 202, period 204 and i-dot 206.

**[0035]** Selecting an initial noise threshold results in one of several possibilities. If the noise threshold is too fine, noise artifact 202, period 204, and i-dot 206 are removed. If the noise threshold is too coarse, noise artifact 202, period 204, and i-dot 206 remain. Because noise artifact 202 is larger than some of the legitimate blobs, such as period 204 and i-dot 206, there does not exist an initial threshold setting that removes noise artifact 202 and does not remove period 204 and i-dot 206. This example illustrates problems that other noise-removal schemes have, and which embodiments of the present invention solve by adjusting the noise threshold during the OCR process as is described in more detail below.

**[0036]** FIG. 2B illustrates a section of the footnote 103 from the sample document illustrated in FIG. 1. Two points of interest are noise artifact 208 and period 210. Choosing a noise threshold level somewhere in between the size of noise artifact 208 and period 210 removes noise artifact 208. This is easily accomplished in a single iteration, and without the refinement occurring during the (OCR) process. In this case, the noise-removal process ends without further refinement of the noise threshold, and the OCR system is able to identify the remaining blobs as valid characters in a font which is recognizable to the OCR system.

**[0037]** FIG. 2C illustrates a section of the sample document illustrated in FIG. 1. This section includes noise artifact 212, image 214, and text 216. Depending on the format of the file and the format of the section, the section

might not be subjected to the noise-removal process. The following description assumes that the section illustrated in FIG. 2C is included in the noise-removal process. If the initial noise threshold is set to a level where noise artifact 212 is removed, then many of the blobs that are part of image 214 will also be removed during the initial noise-removal process. On the other hand, if the initial noise threshold is at a level where image 214 is not altered, then noise artifact 212 will remain after the initial noise-removal stage. One embodiment of the present invention selects an initial noise threshold that results in the removal of noise artifact 212, but without affecting image 214, or alternatively, selects a threshold that neither removes noise artifact 212, nor affects image 214. In the latter situation, noise artifact 212 will be removed during subsequent iterations of the OCR process.

**[0038]** One embodiment of the present invention can process each page of a multi-page document either individually or collectively. Note that if each page is processed individually, the present invention can carry over statistical information from previously-processed pages to assist in forming the initial noise threshold for subsequent pages.

**[0039]** In another embodiment of the present invention, the items which appear in FIG. 2A, FIG. 2B, and FIG. 2C are processed collectively. Because of the numerous noise artifacts which are similar in size to legitimate blobs, most of the noise artifacts will remain during the initial noise-removal process. The remaining noise artifacts will be removed by adjusting the noise threshold during subsequent iterative operations.

#### **Process Overview**

**[0040]** FIG. 3 illustrates a noise-removal process in accordance with an embodiment of the present invention. During this process, imaged documents 304, 305, and 306 are received as inputs to noise-removal process 316. Noise-removal process 316 then produces output documents 324, 325, and 326, respectively. Note that regions containing text with noise 309, 310, 311, and 312 are transformed into regions containing text without noise 307, 329, 330, and 331 by the noise-removal process 316. Non-text regions, 307 and 308, in imaged document 304 and 305 are not processed by noise-removal process 316 and remain as non-text regions, 307 and 308, in output documents 324 and 325, respectively.

#### **Computing Environment**

**[0041]** FIG. 4 illustrates a computing environment 400 in accordance with an embodiment of the present invention. Computing environment 400 includes client 410 and laptop 420. Client 410 and laptop 420 are both coupled to network 440. Additionally both client 410 and laptop 420 have the ability to communicate with numerous devices, including printer 430, scanner 450, cellular camera phone 460, and digital camera 470.

**[0042]** Client 410 and laptop 420 can generally include any node on a network including computational capability and including a mechanism for communicating across network 440.

**[0043]** Client 410 and laptop 420 can generally include any type of computer system, including, but not limited to, a computer system based on a microprocessor, a mainframe computer, a digital signal processor, a portable computing device, a personal organizer, a device controller, and a computational engine within an appliance.

**[0044]** Printer 430 can generally include any type of printer, including, but not limited to, personal printers, network printers and multi-function printers which may include copiers, scanners, and facsimile machines.

**[0045]** Scanner 450 can generally include any type of digital scanner, including, but not limited to, standalone scanners and multi-function scanners which may include copiers, printers, and facsimile machines.

**[0046]** Devices, such as printer 430, scanner 450, cellular camera phone 460, and digital camera 470, are capable of capturing an image of a document, or creating an image that may include text. Each of these devices is capable of transmitting the image to client 410 or laptop 420. In one embodiment of the present invention, both client 410

and laptop 420 are capable of removing noise artifacts that may have occurred during the imaging process, using the removal process described herein.

#### **Noise-Removal System**

**[0047]** FIG. 5 illustrates a noise-removal system 500 in accordance with an embodiment of the present invention. Noise-removal system 500 includes blob identifier 502, blob analyzer 504, noise remover 506, noise restorer 508, OCR system 510 and memory 520. Memory 520 includes document memory 522 (which stores a copy of the images file), noise memory 524, statistical memory 526, and system settings 528.

**[0048]** Blob identifier 502 is used to identify blobs within an imaged document. Once the blobs are identified, blob analyzer 504 analyzes the blobs to determine the noise threshold. Blob analyzer 504 determines the noise threshold based on many factors, including, but not limited to, the number of blobs, the distribution of the blobs, the density of the blobs in different regions, the density of various size blobs in different regions, the position of the blobs, the alignment of the blobs, and the color of the blobs. Once the analysis of the blobs is complete, the resulting statistical information is stored in statistical memory 526. This statistical information can subsequently be used to: further refine system settings; refine the noise threshold; and to refine the noise-removal process for additional regions of the imaged document and additional imaged documents.

**[0049]** Noise remover 506 removes any blobs considered to be noise artifacts based on the noise threshold that was determined by blob analyzer 504. The removed blobs are stored in noise memory 524 in case the noise threshold is altered so that some of the removed blobs need to be restored. If this occurs, noise restorer 508 restores some or all of the removed blobs. The system can determine which blobs to restore by considering the size of the blob, or the location of the blob.

**[0050]** OCR system 510 performs the OCR process after the initial noise-removal process has completed. During this OCR process, the noise-removal is further refined by fine-tuning the noise threshold, as is described in more detail below.

**[0051]** System settings 528 contains system settings for the noise-removal process. In one embodiment of the present invention, these settings include, but are not limited to: a description of the information to be analyzed; a quality value for noise-removal process; an indicator defining when to terminate the noise-removal process; a flag which indicates whether to carry over any information to the next noise-removal task; identifiers for key characters; and an indicator which determines whether to execute the process on a file, a page, or a region of a page. Note that the quality of noise-removal process is inversely related to the speed of the noise-removal process. In addition, the quality of noise-removal process is directly related to the amount of memory available to the noise-removal process.

#### **OCR System**

**[0052]** FIG. 6 illustrates an optical character recognition (OCR) system 600 in accordance with an embodiment of the present invention. OCR system 600 includes text finder 602, text analyzer 604 and noise threshold calibrator 606. Text finder 602 identifies text regions within the imaged document. Techniques for identifying text regions within documents are well-known in the art and will not be described further herein.

**[0053]** Once the text regions have been identified, text analyzer 604 analyzes the text regions both to determine the size of the text and to identify key characters within the text.

**[0054]** The key characters are then analyzed by text analyzer 604 to determine their size (for example, in number of pixels). As was mentioned above, key characters are small characters (or portions of characters) which are used to distinguish characters from noise. The key characters may vary from font to font, but they generally include periods, i-dots, commas, ellipses and other characters (or portions of characters) which are smaller than the other



characters in a font. Information ascertained from text analyzer 604 (including for example a noise threshold) is stored in statistical memory 526. This information can be used during subsequent iterations or for other noise-removal tasks. As described below, the output of text analyzer 604 is used to refine the noise threshold.

**[0055]** Noise threshold calibrator 606 adjusts the noise threshold. If the noise threshold is determined to have changed outside of tolerances specified in system settings 528, then an additional noise-removal operation is performed. This additional noise-removal operation may involve restoring some or all previously removed blobs if the noise threshold has decreased.

#### **Noise-Removal Process**

**[0056]** FIG. 7 presents a flowchart illustrating the noise-removal process in accordance with an embodiment of the present invention. The process begins by identifying blobs in the image (step 702). Next, the system analyzes the blobs (step 704) and stores the resulting information in statistical memory 526. This resulting information may include: average blob size, blob density information, blob color, and any other information useful for setting a noise-removal threshold.

**[0057]** Using the results of step 704, the system determines the noise threshold (step 706). In one embodiment of the present invention, the noise threshold is a fraction of the average blob size for key characters in the image. The noise threshold may also differ over various sections of a given page. For example, in one embodiment, if the top 33% of the page has an average key-character blob size of 15 pixels, the noise threshold for the upper 33% of the page may be set to 5 pixels. However, if the lower 66% of the page has an average key-character blob size of 45 pixels, the noise threshold for the lower 66% of the page may be set to 15 pixels. Other factors besides blob size can be used to determine the noise threshold. For example, if system settings 528 indicate that the image is strictly monochrome, but during the imaging process a red artifact was introduced into the document, the system may identify the red artifact as noise and remove it from the image.

**[0058]** After the noise threshold has been established, the system removes all blobs containing fewer pixels than the noise threshold (step 708). Finally, the system initiates the OCR process (step 710).

#### **OCR Process**

**[0059]** The previous section describes a fast and flexible noise-removal process that can be adjusted to match the needs of the user. In one embodiment of the present invention, the process is completed at step 712.

**[0060]** In another embodiment of the present invention, the noise-removal process is refined during the OCR process. This refinement produces a higher quality result than the previously described embodiment.

**[0061]** FIG. 8 presents a flowchart illustrating the OCR process which includes a refinement to the noise-removal process in accordance with an embodiment of the present invention. During this process, the system first identifies text regions within the imaged document (step 802). Note that in one embodiment, text which is part of a figure or image is typically ignored in this step and all future steps of the OCR process.

**[0062]** Once the text regions are identified, the text regions are analyzed (step 804). This analysis may involve, for example, estimating text size, identifying key characters, and identifying the font used for the text. Next, the system re-evaluates the noise threshold for the image based on results of the analysis (step 806).

**[0063]** The system then determines whether the noise threshold has changed (step 808). If not, the system completes the OCR process (step 820). If the noise threshold has changed, the system determines if the new noise threshold is smaller than the previous noise threshold (step 810). If so, previously removed blobs, which were larger than the new noise threshold but smaller than the initial noise threshold, are restored.

**[0064]** In one embodiment of the present invention, a pre-specified "tolerance level" is also used to determine whether a removed blob should be restored. For example, given a tolerance-level of two pixels, if the old noise threshold was twelve pixels and the new noise threshold is eight pixels, a blob of nine pixels will not be restored. The tolerance level can be indicated by system settings 528.

**[0065]** After removed blobs have been restored (or if the new noise threshold is not lower than the previous noise threshold in step 810), the system repeats the noise-removal process using the new noise threshold (step 814).

**[0066]** Next, the system determines if the new noise threshold is satisfactory (step 816). This determination can be based upon: whether any blobs have been removed; how many blobs have been removed; how many times the noise threshold has been adjusted; or in which direction the noise threshold has been adjusted. If the new noise threshold is satisfactory, the system completes the OCR process (step 820). On the other hand, if the new noise threshold is not satisfactory, the noise-removal process is repeated by returning to step 802.

**[0067]** In one embodiment of the present invention, the system refines the noise-removal process by adjusting system settings 528 before returning to step 802. This can involve adjusting: the information to be analyzed; the quality of noise-removal process; when to terminate the noise-removal process; whether to carry over any information to the next noise-removal task; which characters are key characters; and whether to execute the process on a document, a page, or a region of a page.

**[0068]** The foregoing descriptions of embodiments of the present invention have been presented for purposes of illustration and description only. They are not intended to be exhaustive or to limit the present invention to the forms disclosed. Accordingly, many modifications and variations will be apparent to practitioners skilled in the art. Additionally, the above disclosure is not intended to limit the present invention. The scope of the present invention is defined by the appended claims.

**What Is Claimed Is:**

1. A method for removing noise from an image, comprising:
  - receiving the image;
  - identifying blobs in the image, wherein a blob is a set of contiguous pixels which possibly represents a character or part of a character in the image;
  - analyzing the blobs to determine a noise threshold, wherein blobs smaller than the noise threshold are likely to be noise; and
  - removing blobs which are smaller than the noise threshold from the image.
2. The method of claim 1, wherein determining the noise threshold involves:
  - identifying a text region in the image;
  - identifying key characters within the text regions;
  - computing the average size of the identified key characters; and
  - computing the noise threshold as a fraction of the average key character size.
3. The method of claim 2, wherein if the determined noise threshold is different from a previous noise threshold, the method for removing noise from the image is repeated.
4. The method of claim 3, wherein if the determined noise threshold is reduced from the previous noise threshold, the method further comprises restoring previously removed blobs which are smaller than the previous noise threshold but are larger than the determined noise threshold.
5. The method of claim 1, wherein prior to analyzing the blobs, the method further comprises performing an initial noise-removal operation by removing blobs from the image that are below an initial noise threshold.
6. The method of claim 1, wherein analyzing the blobs involves analyzing one or more of:
  - a size distribution of the blobs;
  - a number of blobs;
  - locations of the blobs;
  - a density of the blobs within the image or region of the image; and
  - colors of the blobs.
7. The method of claim 1, wherein a noise threshold is determined independently for each of a plurality of identified text regions.
8. A computer-readable storage medium storing instructions that when executed by a computer cause the computer to perform a method for removing noise from an image, the method comprising:
  - receiving the image;
  - identifying blobs in the image, wherein a blob is a set of contiguous pixels which possibly represents a character or part of a character in the image;
  - analyzing the blobs to determine a noise threshold, wherein blobs smaller than the noise threshold are likely to be noise; and
  - removing blobs which are smaller than the noise threshold from the image.

9. The computer-readable storage medium of claim 8, wherein determining the noise threshold involves:
  - identifying a text region in the image;
  - identifying key characters within the text regions;
  - computing the average size of the identified key characters; and
  - computing the noise threshold as a fraction of the average key character size.
10. The computer-readable storage medium of claim 9, wherein if the determined noise threshold is different from a previous noise threshold, the method for removing noise from the image is repeated.
11. The computer-readable storage medium of claim 10, wherein if the determined noise threshold is reduced from the previous noise threshold, the method further comprises restoring previously removed blobs which are smaller than the previous noise threshold but are larger than the determined noise threshold.
12. The computer-readable storage medium of claim 8, wherein prior to analyzing the blobs, the method further comprises performing an initial noise-removal operation by removing blobs from the image that are below an initial noise threshold.
13. The computer-readable storage medium of claim 8, wherein analyzing the blobs involves analyzing one or more of:
  - a size distribution of the blobs;
  - a number of blobs;
  - locations of the blobs;
  - a density of the blobs within the image or region of the image; and
  - colors of the blobs.
14. The computer-readable storage medium of claim 1, wherein a noise threshold is determined independently for each of a plurality of identified text regions.
15. An apparatus that removes noise from an image, comprising:
  - a blob-identification mechanism configured to identify blobs in the image, wherein a blob is a set of contiguous pixels which possibly represents a character or part of a character in the image;
  - a threshold-determination mechanism configured to analyze the blobs to determine a noise threshold, wherein blobs smaller than the noise threshold are likely to be noise; and
  - a blob-removal mechanism configured to remove blobs which are smaller than the noise threshold from the image.
16. The apparatus of claim 15, wherein the threshold determination mechanism is configured to:
  - identify a text region in the image;
  - identify key characters within the text regions;
  - compute the average size of the identified key characters; and to
  - compute the noise threshold as a fraction of the average key character size.

17. The apparatus of claim 16, wherein the apparatus is configured to repeat the noise-removal process if the determined noise threshold is different from a previous noise threshold.
18. The apparatus of claim 17, further comprising a blob-restoring mechanism, wherein if the determined noise threshold is reduced from the previous noise threshold, the blob-restoring mechanism is configured to restore previously removed blobs which are smaller than the previous noise threshold but are larger than the determined noise threshold.
19. The apparatus of claim 15, further comprising an initial-noise removal mechanism configured to perform an initial noise-removal operation by removing blobs from the image that are below an initial noise threshold.
20. The apparatus of claim 15, wherein while analyzing the blobs, the threshold-determination mechanism is configured to analyze one or more of:
- a size distribution of the blobs;
  - a number of blobs;
  - locations of the blobs;
  - a density of the blobs within the image or region of the image; and
  - colors of the blobs.
21. The apparatus of claim 15, wherein the threshold-determination mechanism is configured to determine a noise threshold independently for each of a plurality of identified text regions.

100

101

Here is a sample header. Notice how small the text is, especially the i-dot and periods.

102

This is a sample document which contains some noise that can be generated during the scanning process or may be caused by other factors.

Here is some more sample text.

104

Look. Here is a picture that is part of the sample document.



*Park, Vaughan & Fleming* LLP

ATTORNEYS AT LAW

104

104

104

104

103

104

104

Here is a sample footnote. It is a little bigger than the header. It is possible for the text to be even smaller in some documents.

FIG. 1

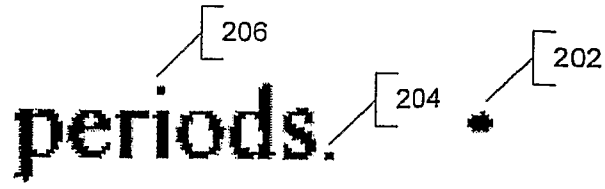


FIG. 2A

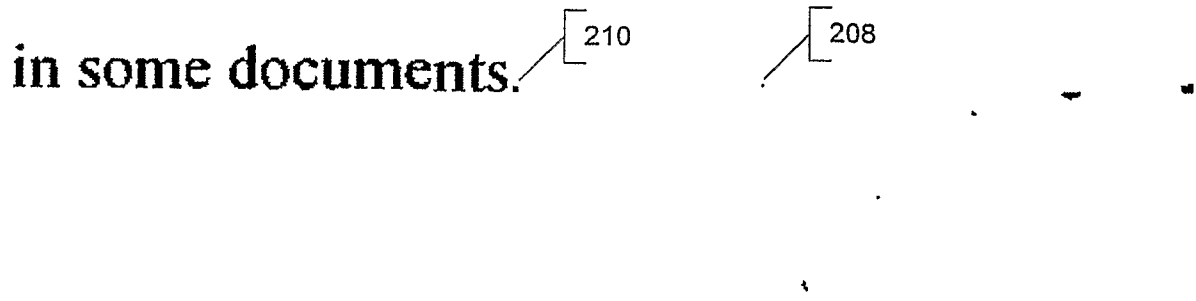


FIG. 2B

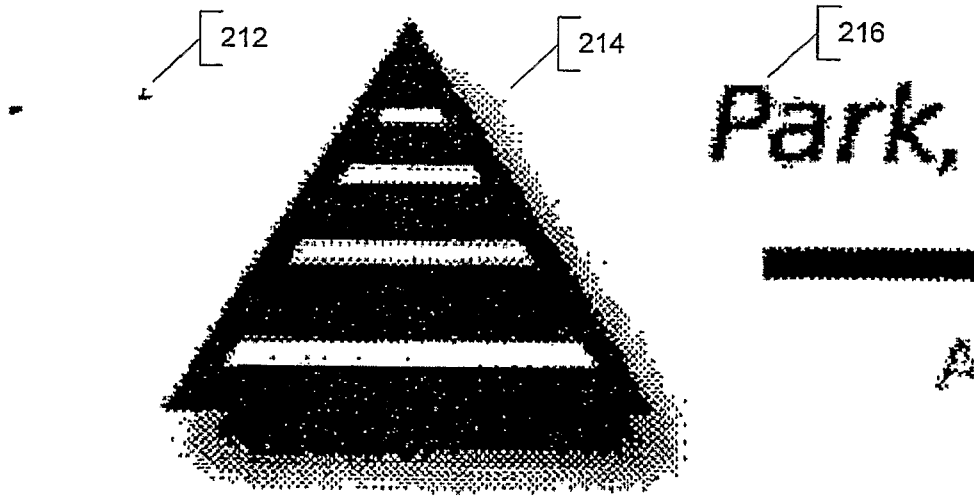


FIG. 2C

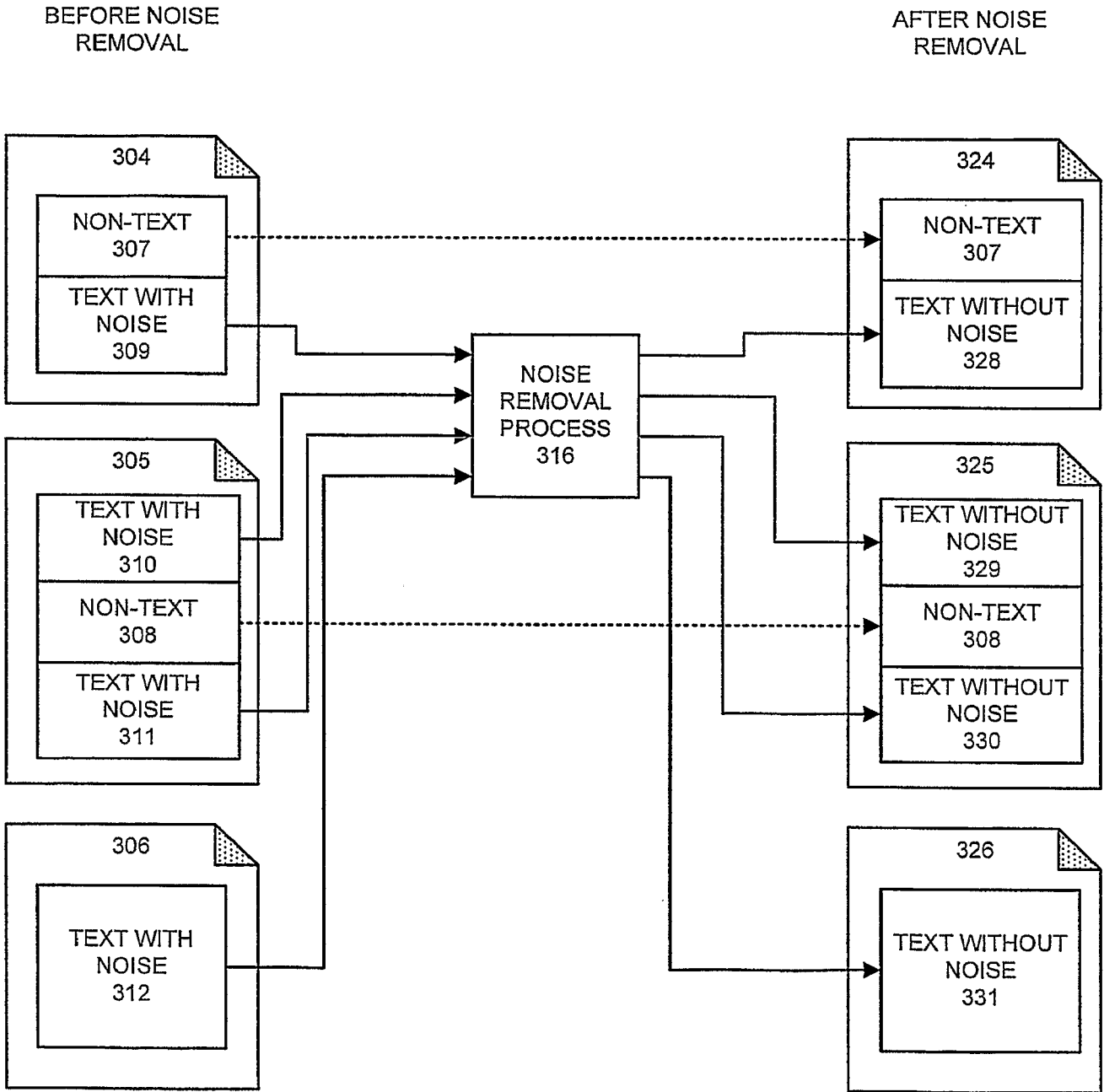


FIG. 3



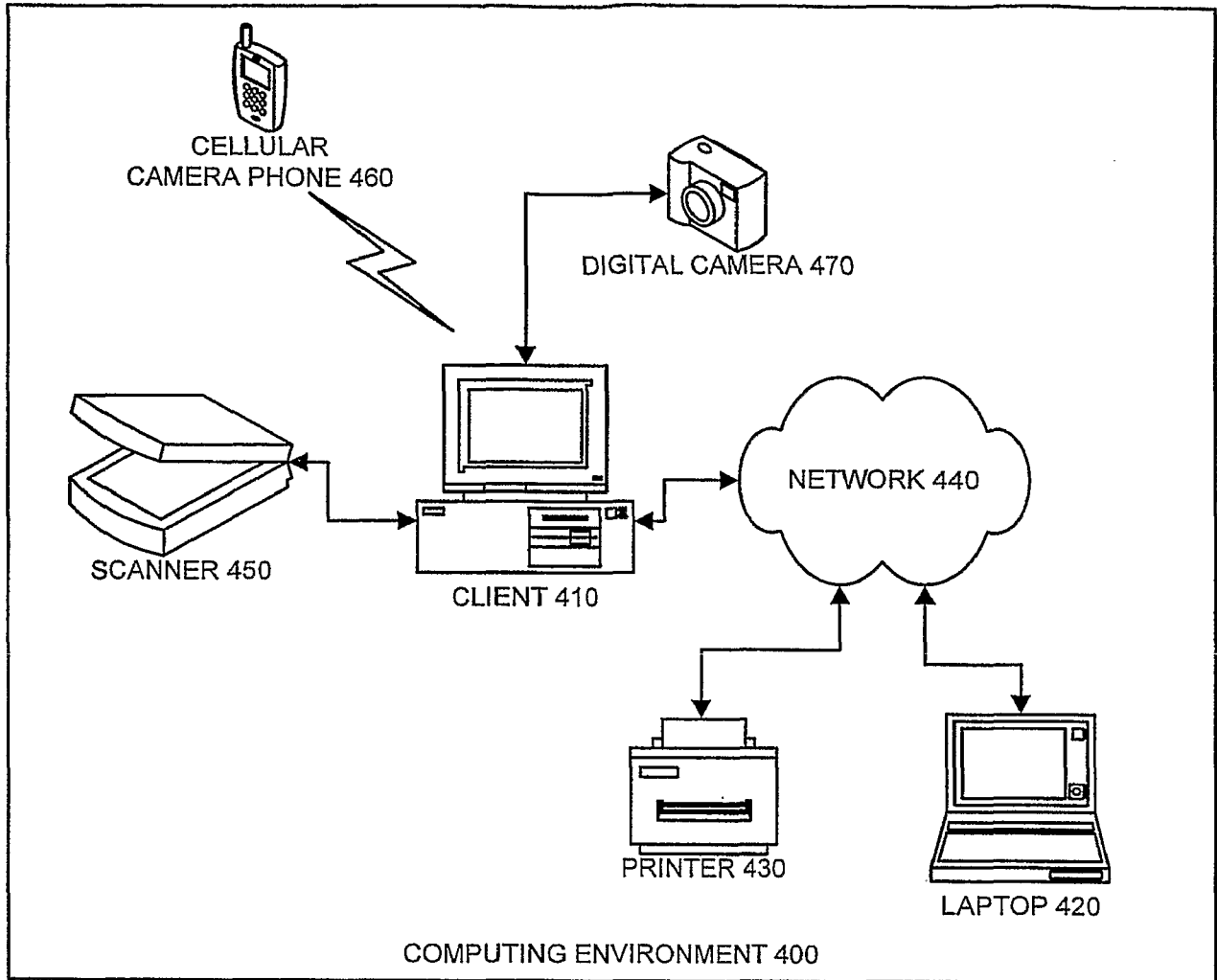


FIG. 4

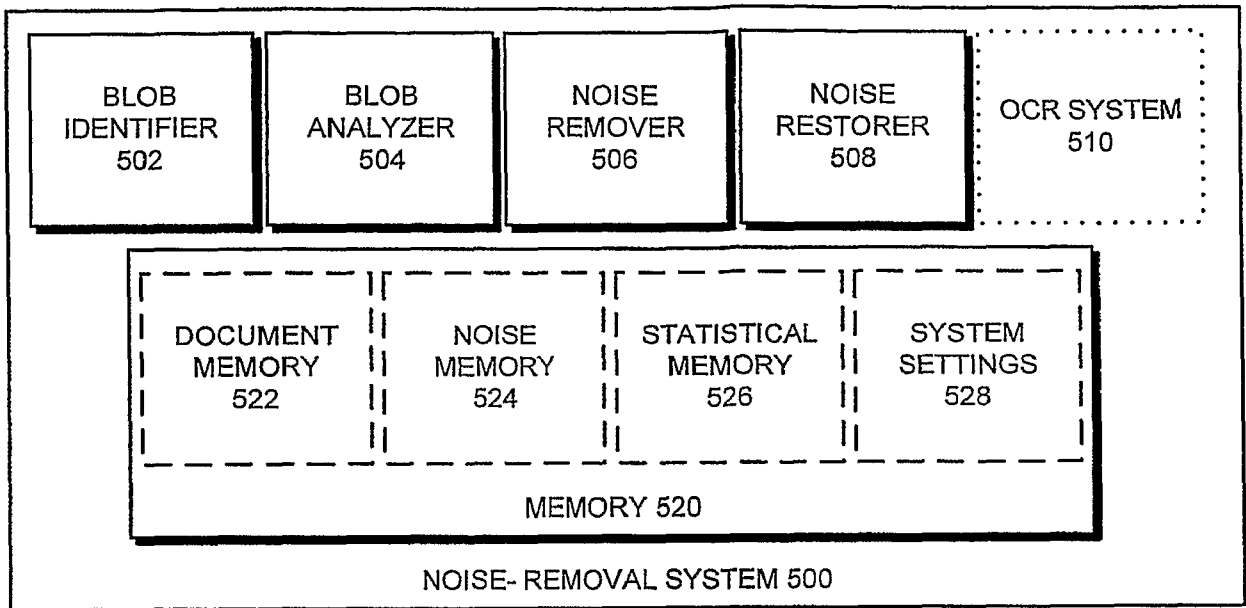


FIG. 5

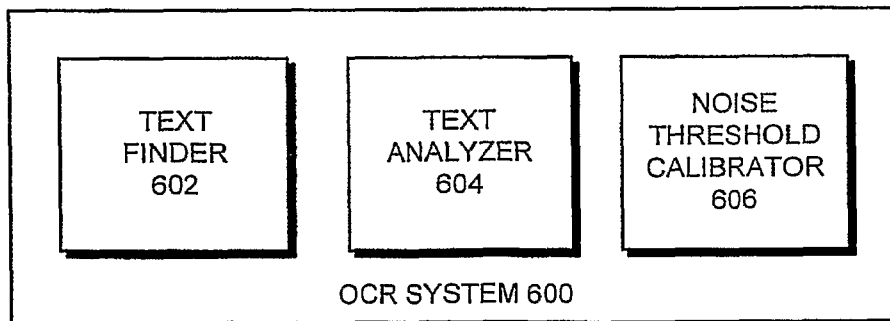


FIG. 6

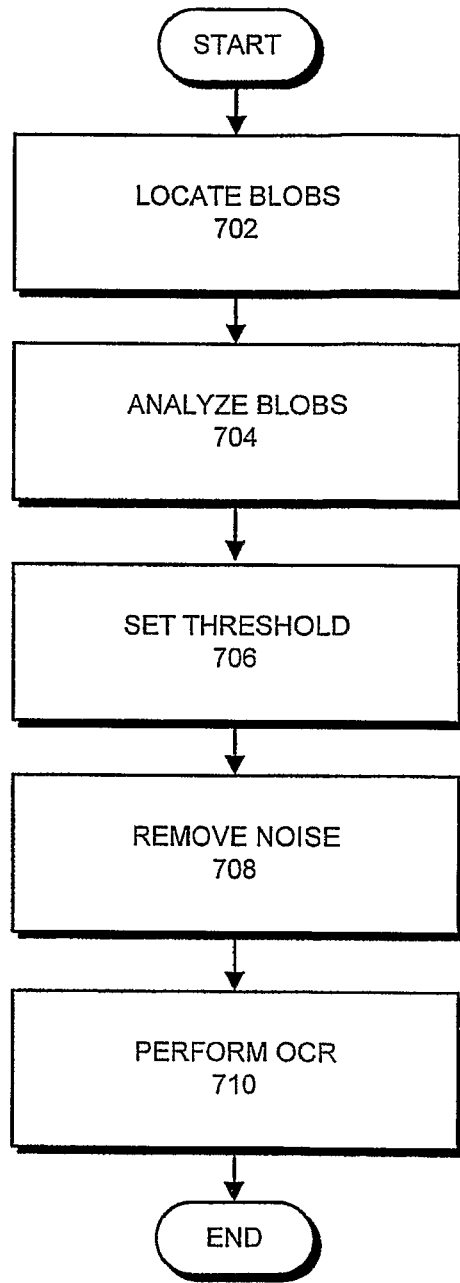


FIG. 7

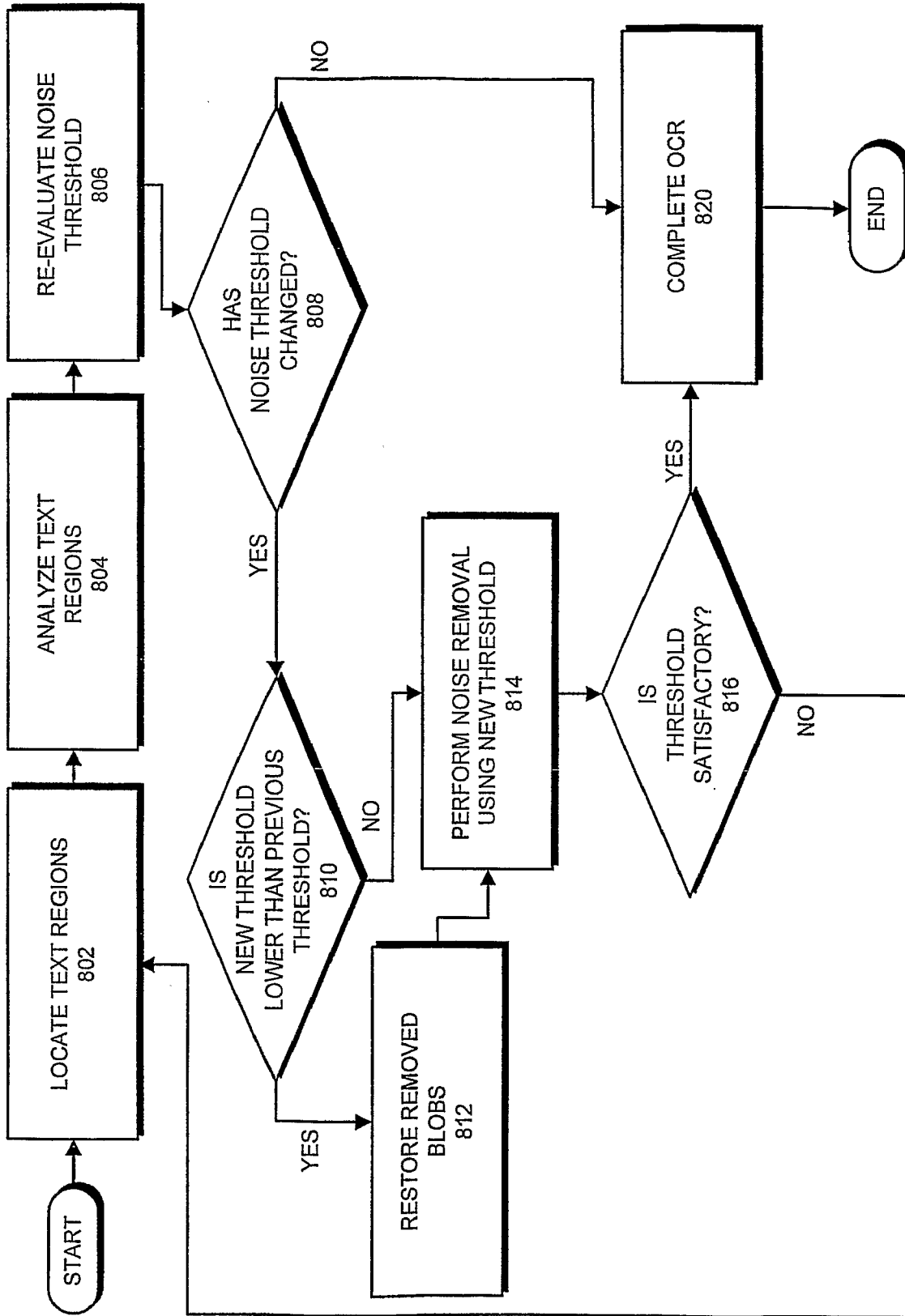


FIG. 8

# INTERNATIONAL SEARCH REPORT

International application No  
PCT/US2006/061294

<b>A. CLASSIFICATION OF SUBJECT MATTER</b> INV. G06T5/00				
According to International Patent Classification (IPC) or to both national classification and IPC				
<b>B. FIELDS SEARCHED</b>				
Minimum documentation searched (classification system followed by classification symbols) G06T				
Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched				
Electronic data base consulted during the international search (name of data base and, where practical, search terms used) EPO-Internal, WPI Data, INSPEC, IBM-TDB				
<b>C. DOCUMENTS CONSIDERED TO BE RELEVANT</b>				
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.		
P, X  P, A	US 2006/008151 A1 (LIN SIMING [US] ET AL) 12 January 2006 (2006-01-12) abstract paragraphs [0004], [0092]	1, 5, 8, 12, 15, 19  2-4, 6, 7, 9-11, 13, 14, 16-18, 20, 21		
E	US 2006/269111 A1 (STOECKER WILLIAM V [US] ET AL) 30 November 2006 (2006-11-30) abstract paragraph [0103]	1, 5, 8, 12, 15, 19		
E	US 2007/040062 A1 (LAU DANIEL L [US] ET AL) 22 February 2007 (2007-02-22) abstract paragraphs [0074], [0075]	1, 5, 8, 12, 15, 19		
----- -/-- -----				
<input checked="" type="checkbox"/> Further documents are listed in the continuation of Box C.				
<input checked="" type="checkbox"/> See patent family annex.				
* Special categories of cited documents :				
<table style="width: 100%; border: none;"> <tr> <td style="width: 50%; border: none; vertical-align: top;">                     *A* document defining the general state of the art which is not considered to be of particular relevance                      *E* earlier document but published on or after the international filing date                      *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)                      *O* document referring to an oral disclosure, use, exhibition or other means                      *P* document published prior to the international filing date but later than the priority date claimed                 </td> <td style="width: 50%; border: none; vertical-align: top;">                     *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                 </td> </tr> </table>			*A* document defining the general state of the art which is not considered to be of particular relevance *E* earlier document but published on or after the international filing date *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) *O* document referring to an oral disclosure, use, exhibition or other means *P* document published prior to the international filing date but later than the priority date claimed	*T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art. *&* document member of the same patent family
*A* document defining the general state of the art which is not considered to be of particular relevance *E* earlier document but published on or after the international filing date *L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified) *O* document referring to an oral disclosure, use, exhibition or other means *P* document published prior to the international filing date but later than the priority date claimed	*T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention *X* document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone *Y* document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art. *&* document member of the same patent family			
Date of the actual completion of the international search <p style="text-align: center; font-size: 1.2em;">5 April 2007</p>		Date of mailing of the international search report <p style="text-align: center; font-size: 1.2em;">17/04/2007</p>		
Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016		Authorized officer <p style="text-align: center; font-size: 1.2em;">Herter, Jochen</p>		

## INTERNATIONAL SEARCH REPORT

International application No

PCT/US2006/061294

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 457 754 A (HAN CHIA Y [US] ET AL) 10 October 1995 (1995-10-10) abstract	1,5,8, 12,15,19
A	column 11, line 53 - column 12, line 32	2-4,6,7, 9-11,13, 14, 16-18, 20,21
X	----- WO 01/26038 A1 (PENN STATE RES FOUND [US]; SURROMED INC [US]; NATAN MICHAEL J [US]; WA) 12 April 2001 (2001-04-12) abstract	1,5,8, 12,15,19
A	page 22, line 24 - page 23, line 23	2-4,6,7, 9-11,13, 14, 16-18, 20,21
A	----- US 6 728 401 B1 (HARDEBERG JON Y [US]) 27 April 2004 (2004-04-27) the whole document	1-21
A	----- JP 63 250787 A (FUJI ELECTRIC CO LTD) 18 October 1988 (1988-10-18) abstract	1-21
	-----	

# INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2006/061294

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2006008151	A1	12-01-2006	NONE
US 2006269111	A1	30-11-2006	NONE
US 2007040062	A1	22-02-2007	NONE
US 5457754	A	10-10-1995	NONE
WO 0126038	A1	12-04-2001	AU 7746200 A 10-05-2001 AU 7844800 A 10-05-2001 CA 2386047 A1 12-04-2001 CA 2386165 A1 12-04-2001 EP 1222609 A1 17-07-2002 EP 1227929 A1 07-08-2002 JP 2003529128 T 30-09-2003 JP 2003511675 T 25-03-2003 MX PA02002787 A 26-02-2003 WO 0125002 A1 12-04-2001
US 6728401	B1	27-04-2004	NONE
JP 63250787	A	18-10-1988	NONE