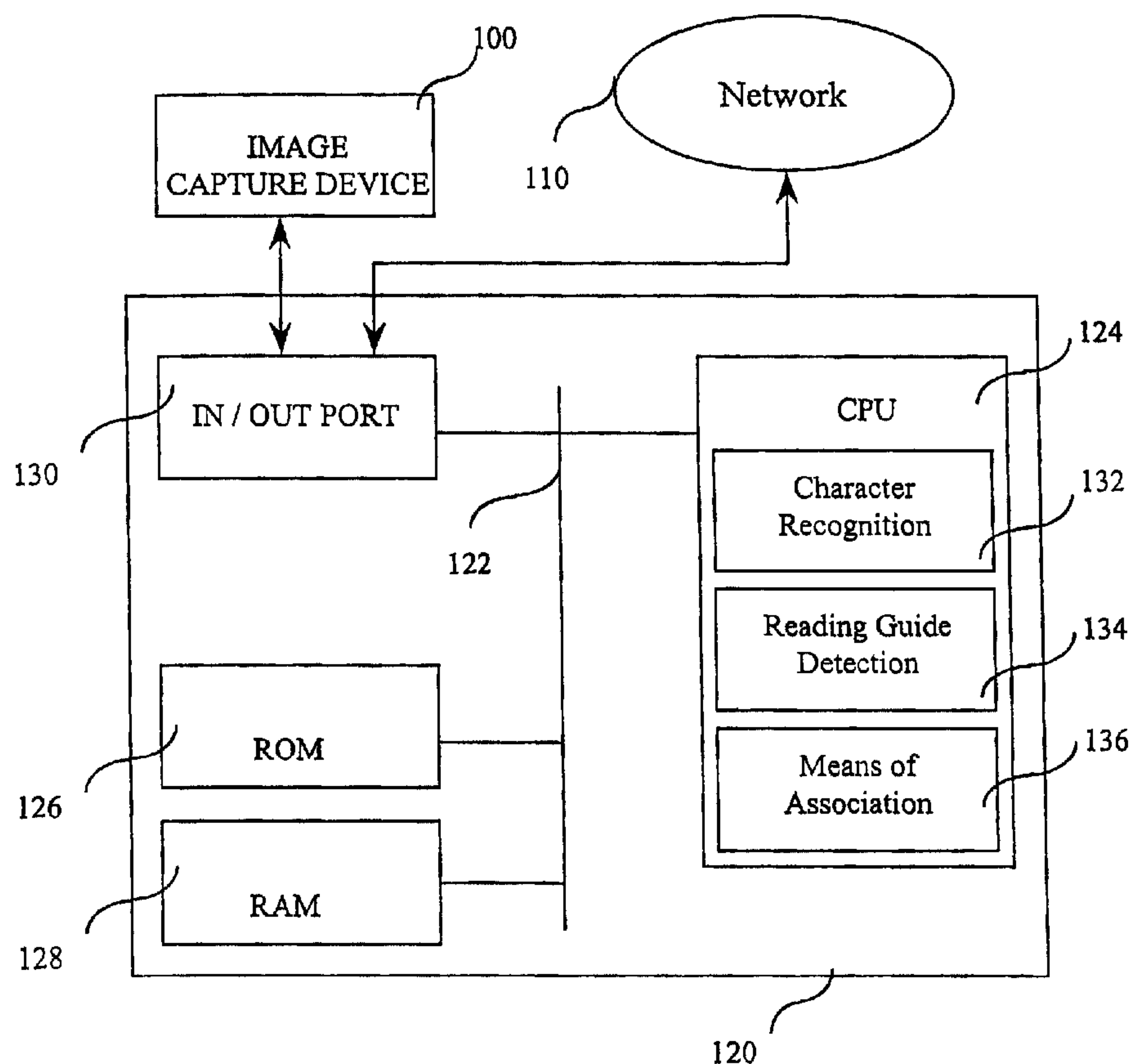




(86) Date de dépôt PCT/PCT Filing Date: 2000/12/04
(87) Date publication PCT/PCT Publication Date: 2001/06/07
(45) Date de délivrance/Issue Date: 2011/03/08
(85) Entrée phase nationale/National Entry: 2002/06/03
(86) N° demande PCT/PCT Application No.: FR 2000/003379
(87) N° publication PCT/PCT Publication No.: 2001/041050
(30) Priorité/Priority: 1999/12/02 (FR99/15306)

(51) Cl.Int./Int.Cl. *G06K 9/20* (2006.01)
(72) Inventeur/Inventor:
CHARPENTIER, DIDIER, FR
(73) Propriétaire/Owner:
ITESOFT S.A., FR
(74) Agent: KIRBY EADES GALE BAKER

(54) Titre : TECHNOLOGIE ADAPTATIVE D'ANALYSE AUTOMATIQUE DE DOCUMENT
(54) Title: ADAPTIVE TECHNOLOGY FOR AUTOMATIC DOCUMENT ANALYSIS



(57) **Abrégé/Abstract:**

The invention concerns a device for automatic document reading comprising an image sensor for imaging the document and an output for data from said document and data concerning meanings of at least part of said scanned data. It further comprises means for detecting at least a scanning guide and means associating at least a scanned information with an information of meaning using said scanning guide.



Abstract

The invention concerns a device for automatic document reading comprising an image sensor for imaging the document and an output for data from said document and data concerning meanings of at least part of said
5 scanned data. It further comprises means for detecting at least a scanning guide and means associating at least a scanned information with an information of meaning using said scanning guide.

Adaptive Technology for Automatic Document Analysis

Field of the Invention

The invention described herein concerns a process and device for automatic document reading.

5 Background of the Invention

The usual approaches to automatic document or form reading require a parameterization operation which defines the data to be read by the coordinates along two axes, X and Y, of the zones to be processed, that is the zones that might contain symbols to be recognized. Also, with each variation
10 in document format, even if the variation is slight, it is necessary to reset the parameters. This parameterization is more and more complex as the document itself increases in complexity.

When a document whose format is defined electronically, for example by being transmitted on a telecommunications network such as the Internet, is
15 printed, the positions of the zones to be processed vary according to the software program used (word processor, drawing application), the printer used, and more generally the configuration of the user's computer system. Because of these variations, the automatic reading of this document is therefore made difficult, or even impossible.

20 Document WO 98/47098 (PCT/SE98/00602) presents a method for automatic data acquisition in which an unknown form is digitized with a scanner and lines are located on this form to identify it from among all the different possible documents. However, the identification of the document does not avoid all the disadvantages described above.

Summary of the Invention

The invention described herein aims to overcome these disadvantages. For that reason, this invention proposes a device and process for automatic document reading.

5 In accordance with one aspect of the present invention there is provided an automatic document reading device comprising: an image capture device adapted for capturing an image of a document; a means of logically analyzing an image of the document, adapted for extracting logical characteristics of the document; a means of defining a document family to
10 which the document belongs from among a plurality of document families, in function of the logical characteristics, each of the plurality of document families containing documents whose contents are similar but differently localized; a means of associating to the image a document structure common to all documents in the document family; a means of segmenting the image of
15 the document into fields, in function of the document structure of the document family to which the document belongs; a means of associating the fields in the document to identifiers which depend on the document family; a means of extracting data from the fields of the document; a means of associating the data that have been read to the identifiers which are
20 associated to the fields in which the data have been read; a means of transmitting the data that have been read and the identifiers associated to the data.

 In accordance with another aspect of the present invention there is provided an automatic document reading process comprising: an operation
25 for capturing an image of a document; an operation for logically analyzing the image of the document, during which logical characteristics of the document are extracted; an operation for defining a document family to which the document belongs among a plurality of document families, in function of the logical characteristics, each of the plurality of document families containing
30 documents whose contents are similar but differently localized; an operation for associating to the image a document structure common to all the

documents in the document family; an operation for segmenting the image of the document into fields, in function of the document structure of the document family to which the document belongs; an operation for associating fields in the document to identifiers which depend on the document family; an
5 operation for extracting data from the fields of the document; an operation for associating the data to the identifiers which are associated to the fields in which the data have been read; an operation for transmitting the data and the identifiers associated to the data.

The reading is guided and the information read is associated to a piece
10 of information whose significance is deduced from the read information thanks to this guidance. The extraction of useful information is thus guided according to the logical structure of a class of documents (documents whose contents are similar, but differently localized) rather than by the physical structure of the document.

15 The device can read documents in which the information fields are specified by a constraint rather than by an absolute position. For example, information fields may be located using the respective positions of the fields, the respective positions of the headings and the approximate positions of fields.

20 Graphic characteristics such as lines, columns, frames, boxes, hatching, colors, grayscale or combs make it possible to identify the fields of information to be read.

Written headings, such as form numbers, or box titles, make it possible to specify the significance of the information to be read.

25 Information read can be identified using key words, such as "last name", "first name", "age", "social security number", and "tax excl.", "tax incl."

The invention described herein also relates to a scanner and a computer adapted to implement the process concerned by the invention
30 described herein or to include the device concerned by this invention.

Brief Description of the Drawings

Other advantages, goals and characteristics of the invention described herein will be highlighted in the following description intended to explain them, but which should in no way be considered exhaustive, with respect to the
5 appended drawings in which:

Figure 1 is a functional diagram that represents a first way in which the device concerned by the invention described herein is implemented;

Figure 2 is a schematic representation of a sequence of operations implemented according to one aspect of the process concerned by the
10 invention described herein;

Figure 3 is a simplified flowchart that represents one way in which the process concerned by the invention described herein is carried out;

Figure 4 is a simplified flowchart that represents a second way in which the process concerned by the invention described herein is carried out;

15 Figure 5 is a simplified flowchart that represents a third way in which the process concerned by the invention described herein is carried out;

Figure 6 is a simplified flowchart that represents a fourth way in which the process concerned by the invention described herein is carried out;

Figure 7 represents a document on which the four ways of carrying out
20 the process illustrated in Figures 3 to 6 are implemented; and

Figure 8 is a simplified flowchart that represents a fifth way in which the process concerned by the invention described herein is carried out.

Detailed Description of the Drawings

Figure 1 shows an image capture device 100, an automatic document
25 reading device 120, and a computer network connection 110.

The image capture device 100 is adapted to providing an electronic signal representing the shapes and colors added to a document: for instance, a form filled out manually by a user. For example, the image capture device 100 is a known type of scanner. The connection to the computer

network 110 is of a known type. It transmits document content information once the automatic document reading device has processed the signal relayed by the capture device 100.

5 The automatic document reading device 120 includes the following interlinked items: a data and control bus 122, a central processing unit 124, a non-volatile memory (ROM) 126, a random access memory 128, and an input/output port 130, which is itself linked to the capture device 100 on the one hand, and to the network line 110 on the other. As an example, the reading device 120 can be set up by programming a typical personal
10 computer. The reading device can also be created on a printed circuit comprised of a processor or a controller, a specific circuit, and a pin grid array or in an integrated circuit. All these ways in which the process are carried out are familiar to specialists in the field of electronic circuits and are therefore not described in further detail here.

15 The central processing unit 124 is comprised of a character reading circuit 132, a detection circuit for at least one reading guide 134 and a means for associating 136 at least one piece of information read by the character reading circuit to one piece of information whose significance is deduced from that information read. The means of association implements at least one
20 reading guide for associating to a piece of read information a piece of information whose significance is deduced from the read information.

Thus, the reading device 120 transmits part of the information that was read on the document on the computer network line 110. For example, "Géraldine", "Pacaut" and, on the other hand, information whose significance
25 is deduced from the read information, which indicates that the first word corresponds to a first name and the second word to a last name. It should be noted, however, that the information whose significance is deduced from the read information may be implicit. For example, when the order in which each piece of information read on a document is transmitted on the network is
30 predefined by a communication protocol. It may also be explicit, when a predefined identifier is associated to each piece of information, or mixed when an identifier is associated to at least one piece of information.

The way in which the information transmitted on the computer network line 110 is used does not relate to the invention described herein. It should be noted, however, that all or part of the information read can be archived in a database, processed, displayed, translated, printed, sorted, added to or
5 destroyed, or, for each piece of information, a combination of these operations.

Generally speaking, the automatic document reading device 120 is adapted to implement each of the ways in which the process is carried out and the different aspects of the process concerned by the invention described
10 herein, as presented in relation to Figures 2 to 8.

Figure 2 represents schematically a sequence of operations implemented for one aspect of the process concerned by the invention described herein.

In Figure 2, it should be noted that automatic document reading
15 includes:

- a document image capture operation 210,
- a detection operation for at least one reading guide 220,
- an operation in which at least one piece of information read is associated to a piece of information whose significance is deduced from the
20 read information that implements the aforementioned reading guide 230, and
- an output operation for information read on the aforementioned document and information with whose significance is deduced from at least part of this information read 240.

The document image capture operation 210 consists in capturing an
25 image of a document and creating a piece of information representative of the aforementioned image. The operation for detecting at least one reading guide 220 consists in detecting a reading guide on the document of which an image is captured during the operation 210. The detection is performed by processing the document image information. It should be noted that this
30 reading guide does not vary with the document and supports some variations in the way the document is laid out and filled out and in the document image capture. In addition, this guide is extracted from the processing of the image

and therefore depends on the image contents. Thus, the reading guide is not a piece of information concerning a position that has pre-set Cartesian coordinates before the document is read.

5 The operation in which at least one piece of information read is associated to a piece of information whose significance is deduced from the read information that implements the reading guide 230 consists in reading a piece of information by processing the image information and associating a piece of information whose significance is deduced from read information to the information read, in function of the reading guide.

10 The output operation for information read on the aforementioned document and information whose significance is deduced from at least part of this information read 240, consists in transmitting the data read and the data whose significance is deduced from the read data (at least implicitly for the information whose significance is deduced), to prepare for processing, storing,
15 displaying, completing, printing, sorting, transmitting or destroying them.

 In the way in which the process concerned by the invention described herein is carried out, represented in Figure 3 as a simplified flowchart, can be found the operations of document image capture 210, detection of at least one reading guide 320, associating at least one piece of information read to a
20 piece of information whose significance is deduced from the read information that implements the reading guide 320, and an output for information read on the document and information whose significance is deduced from at least part of this information read 240.

 The operation for detecting a reading guide 320 includes one reading
25 operation for at least one positioning constraint for at least one field of the aforementioned document 321, a field in which a piece of information might be read, and a detection operation in which at least one positioning constraint may be complied with for at least one piece of information read 322. Each

positioning constraint mentioned here is a relative positioning constraint for different pieces of information read on the document. Such a constraint is expressed, for example, by "the name is above the address", "the total tax excl. precedes the total tax incl.", "the number representing the month and the
5 number representing the day precede the number representing the year".

During the association operation 330, a piece of information whose significance is deduced from read information representative of such constraint is associated to each piece of information read that complies with the aforementioned constraint. Therefore, the piece of information "99" read
10 in the block of information "02 / 25 / 99" is associated to a piece of information that signifies "year" because it is preceded by a number, "25" that can be associated to the day and a number "02" that can be associated to the month.

In this way, documents in which information fields are specified by positioning constraints and not by absolute positioning are read. For
15 example, the respective positions of the fields, the respective positions of the headings, and the approximate positions of fields make it possible to locate them.

In the way in which the process that concerns the invention described herein is carried out, represented as a simplified flowchart in Figure 4, can be
20 found operations of document image capture 210, detection of at least one reading guide 420, associating at least one piece of information read to a piece of information whose significance is deduced from the read information that implements this reading guide 430, and of output information read on the document and information whose significance is deduced from at least part of
25 this information read 240.

The operation for detecting a reading guide 420 includes a reading operation for at least one graphic characteristic of at least one field of a document 421, a field in which a piece of information might be read, and a detection operation for a field that may present a graphic characteristic 422.
30 Each graphic characteristic mentioned here is a characteristic of shape, color, item, line, etc. For example, there is a thick horizontal line above the total in an invoice, an address is placed in a rectangular frame, the responses to

multiple choice questionnaires are in square boxes, the age of the person filling out the form appears against a yellow or hatched background, the date contains two forward slashes, etc.

During the association operation 430, to each piece of information read in a field that has one of the aforementioned graphic characteristics is associated a piece of information whose significance is deduced from the read information that is representative of the presence of that characteristic. In the examples presented above, a piece of information placed against a yellow or hatched background is associated to the significance "age", a piece of information read in a rectangular frame is associated to the significance "address" etc.

Thus, creating a document to be filled out is simple. Graphic characteristics, such as lines, columns, frames, check boxes, hatching, colors, grayscale or combs make it possible to identify the information fields to be read.

In the way in which the process that concerns the invention described herein is carried out, represented as a simplified flowchart in Figure 5, can be found operations of document image capture 210, detection of at least one reading guide 520, associating at least one piece of information read to a piece of information whose significance is deduced from the read information that implements this reading guide 530, and of output information read on the document and information whose significance is deduced from at least part of this information read 240.

The operation for detecting a reading guide 520 includes a reading operation for at least one heading of at least one field of the aforementioned document 521, a field in which a piece of information might be read, and a reading operation for a possible heading 522.

Each heading mentioned here is a sequence of at least one symbol that identifies a field. For example, the heading can be limited to:

a written number of very small dimensions (for example less than half the height of the symbols of the information read);

a series of points the number of which represents the heading;

a number of hatch marks,

a number of places for symbols of information (for example a first name on a line where fifteen symbols are pre-marked, a last name on a line in which twenty symbols are pre-marked, an address on a line with more than twenty
5 symbols, an age on a line with three places for symbols, a social security number in a sequence of exactly thirteen symbols (French standard), etc.);

one or more words (for example the word "last name" associated to the field in which the person filling out the form should write his or her last name, etc.).

10 During the association operation 530, a piece of information whose significance is deduced from read information representative of a heading is associated to each piece of information read in a field having that heading.

It is thus simple to create a document to be filled out. Headings whose significance is deduced from read information are attached to fields for the
15 information to be read.

In the way in which the process that concerns the invention described herein is carried out, represented as a simplified flowchart in Figure 6, can be found operations of document image capture 210, detection of at least one reading guide 620, associating at least one piece of information read to a
20 piece of information whose significance is deduced from the read information that implements this reading guide 630, and of output information read on the document and information whose significance is deduced from at least part of this information read 240.

The operation for detecting a reading guide 620 includes one reading
25 operation for at least one information content constraint concerning at least one field of the document 621, a field in which a piece of information might be read, and a content reading operation concerning at least one of the fields of the aforementioned document 622.

Herein, content signifies the semantics of the information. For example a number of less than three digits is the sign of an age content. A sequence of 'more than three words combined with numbers and including a 5-digit number is an address. A sequence of thirteen digits is a social security
5 number (French standard).

During the association operation 630, to each piece of information read in a field that complies with one of the aforementioned content constraints is associated a piece of information whose significance is deduced from the read information representative of that content.

10 By means of this set-up, relatively unstructured forms can be processed.

Figure 7 represents a document implementing the four ways in which the processed can be carried out, as illustrated in Figures 3 to 6. The upper part represents the image of a document and the lower part a file of
15 information and deduced significance completed on the basis of the aforementioned image.

Field 700 shows the positioning constraint of being centered within the document, with respect to the side edges (right and left) of the document. Implementing the invention in the way illustrated in Figure 3 makes it possible
20 to identify the information associated to the significance "document type".

Fields 701 to 703 show positioning constraints in accordance with the way in which the process is carried out illustrated in Figure 3. Although the totals are not at all lined up with their headings ("total excl. tax", "total tax" and "total incl. tax") the sequence in which they are in helps to determine the
25 information whose significance is deduced from read information that must be assigned to them.

In addition, fields 701 to 703 show one content constraint, since the smallest of the three totals is the Value Added Tax (VAT) total, and the greatest is the tax incl. total. Thanks to this constraint, although the totals
30 were not placed in the correct order, the way in which the invention described herein is carried out, illustrated in Figure 6, attributes the normal significance to each of them.

Field 704, which has a shaded background, corresponds to an address. In accordance with the way in which the process is carried out, illustrated in Figure 4, the deduced significance of the information read in this field is determined by this background, which constitutes a graphic guide.

5 Field 705 contains a written heading "discount", which makes it possible to identify the deduced significance of the information read in this field, in accordance with the way in which the process is carried out, as illustrated in Figure 5.

10 Field 706 contains two forward slashes and the information read therein is associated to a date in accordance with the way in which the process is carried out, as illustrated in Figure 4.

15 Finally, field 707 presents a content "with our compliments", which corresponds to a polite expression, and is therefore associated to a comment in accordance with the way in which the process is carried out, as illustrated in Figure 6.

20 It should be noted that the significance of some information read can be determined according to several ways of carrying out the process that is the object of the invention described herein. For example, a date that contains three 2-digit numbers separated by slashes can be identified by a graphic characteristic, the two slashes, or its content: three times two digits.

In the remainder of the description, the following definitions are used:

25 "family" concerns a set of documents that meet the criteria of being approximately alike. For each family, more general rules, which are less precise and more approximate than the simple plane coordinates of the zones to be read, make it possible to describe the document and the zones to be processed.

For example:

30 Half of the goods exchange declarations used by the French Customs are "cerfa" documents, and half are documents printed by the companies filing the declarations. These declarations constitute a family because the information in the header is always in approximately the same place, and the columns are in the same order on almost all the declarations.

The purchase orders used by mail-order sales companies often change for sales and marketing reasons. However, the nature of the useful information contained in these forms is always the same. In addition, this information is organized in blocks of information. These blocks can be
5 addresses, lists of items ordered, means of payment, etc.

The supplier invoices received by a company are all different because each supplier prints its invoice in a particular way. However, the zones to be processed concern, at least, the total tax excl., the total of the Value Added Tax, the total tax incl., the date and the item number.

10 A "form" is a group of one or more pages. Each page can be broken down into one or more blocks.

Each "block" contains graphic elements (horizontal lines, vertical lines, frames, logos, images, etc.), textual elements (headings and identification labels, various text, etc.) as well as fields and zones to be read.

15 Each "element" is defined by characteristics whose degree of tolerance must represent all the possible variations within the family.

Certain documents or families of documents contain specific objects that make it possible to locate the data to be extracted in a general and global manner.

20 For example, a textual heading is often followed by the value of the field, on a line, in a frame, in a check box or series of check boxes, or with no graphic guide. For example, in a frame, the heading, if there is one, is often in the upper left-hand corner. The response fields on questionnaires are often rectangular graphic objects. Column headings are often at the beginning of
25 columns, at the top.

Thus, the analysis of the image extracts all the elements that make it possible to identify the document (graphic objects, lines, frames, headings, logos, etc.). Identifying the page makes it possible to associate to the image to be processed the template of the document, which describes the page
30 structure.

The elements defined in the page structure are compared to those found or searched for in the image. This search is performed within the limits of their characteristics (tolerances concerning the positions, relative variations, similar character strings, etc.), and results in a segmentation hypothesis.

This segmentation hypothesis makes it possible to locate with precision the fields to be processed and to extract the images from these fields for the subsequent steps.

The logical definition of the field (last name, type, consistency check, etc.) and of the physical field found on the page are performed on the basis of the absolute physical position of the position with respect to another field and/or its association with a heading (company, etc.).

The utilization of specific objects facilitates the parameterization of documents because the automatic analysis of the page localizes all the objects that meet the characteristics being searched for, verifies the constraints expressed and extracts the fields to be read with no complex segmentation of the page.

It should be noted that the adaptive process developed for processing variable formats offers exceptional robustness and flexibility for all the shapes due to slips and other printing anomalies found in document processing.

Like the human eye, adaptive technology uses "natural" or intuitive graphic characteristics, content, headings, field (or zone) titles, or relative positions to identify the information contained in each field.

The analysis and the segmentation of the page is based on logical rules, (e.g., total columns on the bottom of the page, definition of the type of column depending on its order, heading associated to the field, etc.), independently of the precise physical position as represented by planar coordinates.

The utilization of adaptive technology makes it possible to define a single template for a family of documents, and therefore all the physical variations of the same logical description. This template associates the fields and their identifiers to characteristics that are not too constraining for document designers, but define the family of documents being considered.

Thus, the process and device of the invention described herein process documents that are similar but different from a physical point of view, as well as standard, computer generated forms.

The utilization of adaptive reading, in accordance with the invention described herein, makes it possible to define a single document template for an entire family of physical variations with the same logical description. Automatic page analysis technology makes it possible to extract "at a glance" all the fields to be processed without even having learned the type of page used.

The following document types are among the classes of documents to which the invention described herein can be applied:

- purchase orders (mail-order sales, etc.);
- goods exchange declarations (customs);
- Intrastat and Extrastat documents;
- tax forms;
- invoices;
- questionnaires;
- single social security contribution declarations;
- financial instrument forms;
- bank checks;
- social security statements;
- legal correspondence and other documents, etc.

In Figure 8, subsequent to an operation 802 for identifying the document class of the document analyzed, an operation 804 defining the graphical and logical data localization rules is performed. After that, a document reading step 806 is performed. Document reading consists in
5 digitizing the document and accessing the properties of the electronic image of the document and the objects that represent the image, for example graphic objects such as icons or parts of images, or elements, as understood in SGML and XML languages, which define certain documents as they are generated.

10 Then the primitives are extracted from the data read during an operation 808. The primitives are primary elements of the document that have elementary properties. For example, if the document is represented by an electronic image, a primitive can be the pixel (or image element) or a connected component, for example a collection of adjacent pixels having the
15 same structural properties, such as the same color. During an operation 810, the page read is identified by implementing the primitives extracted. Then, during an operation 812, the data to be read are extracted from the image of the document, for example by optical character recognition. The guiding rules are used on the data read, including, according to the case, the relative
20 positions, the graphic guides, the heading or the content of the reading zones during an operation 814.

During an operation 816, the information read on the document is associated to information whose significance is deduced from the read information, both of which are related to the reading guides used, by means of
25 a look-up table, for example. Finally, during an operation 818, the information read and the information whose significance is deduced from it are transmitted to a device for processing, storing, and displaying them, etc. as described, for example, with regard to Figures 1 and 2.

An operation to manage iterations, failures, and ambiguities 820
30 enables the process to improve through automatic learning.

Claims

1. Automatic document reading device comprising:
 - an image capture device adapted for capturing an image of a document;
 - 5 a means of logically analyzing an image of the document, adapted for extracting logical characteristics of the document;
 - a means of defining a document family to which the document belongs from among a plurality of document families, in function of the logical characteristics, each of the plurality of document families containing
 - 10 documents whose contents are similar but differently localized;
 - a means of associating to the image a document structure common to all documents in the document family;
 - a means of segmenting the image of the document into fields, in function of the document structure of the document family to which the
 - 15 document belongs;
 - a means of associating the fields in the document to identifiers which depend on the document family;
 - a means of extracting data from the fields of the document;
 - a means of associating the data that have been read to the identifiers
 - 20 which are associated to the fields in which the data have been read;
 - a means of transmitting the data that have been read and the identifiers associated to the data.
2. The automatic document reading device of claim 1, further comprising:
 - a means of saving at least one positioning constraint for at least one
 - 25 field of the document, a field in which a piece of information might be read;
 - the means of logically analyzing is adapted for detecting a compliance with at least one positioning constraint of at least one of the fields; and
 - the means associating the fields is adapted for associating to the data read in a field which complies with one of the constraints, an identifier which is
 - 30 representative of that constraint.

3. The automatic document reading device of either of claims 1 or 2, further comprising:

a means of saving at least one graphic characteristic of at least one field of the document, a field in which a piece of information might be read;

5 the means of logically analyzing is adapted for detecting at least one field of the document that has one of the graphic characteristics;

the means of associating the fields is adapted for associating to each field that has one of the graphic characteristics, an identifier which is representative of a presence of that graphic characteristic.

10 4. The automatic document reading device of any one of claims 1 to 3, further comprising:

a means of saving at least one heading from at least one field of the document, a field in which a piece of information might be read;

15 the means of logically analyzing is adapted for reading at least one of the headings; and

the means of associating is adapted for associating to each field that contains the heading, an identifier that is representative of that heading.

5. The automatic document reading device of any one of claims 1 to 4, further comprising:

20 a means of saving at least one information content constraint for at least one field of the document, a field in which a piece of information might be read;

the means of logically analyzing is adapted for reading the content of at least one field of the document; and

25 the means of associating is adapted for associating to each field that complies with the content constraint, an identifier that is representative of that content.

6. Automatic document reading process comprising:
- an operation for capturing an image of a document;
 - an operation for logically analyzing the image of the document, during which logical characteristics of the document are extracted;
 - 5 an operation for defining a document family to which the document belongs among a plurality of document families, in function of the logical characteristics, each of the plurality of document families containing documents whose contents are similar but differently localized;
 - an operation for associating to the image a document structure
 - 10 common to all the documents in the document family;
 - an operation for segmenting the image of the document into fields, in function of the document structure of the document family to which the document belongs;
 - an operation for associating fields in the document to identifiers which
 - 15 depend on the document family;
 - an operation for extracting data from the fields of the document;
 - an operation for associating the data to the identifiers which are associated to the fields in which the data have been read;
 - an operation for transmitting the data and the identifiers associated to
 - 20 the data.
7. The automatic document reading process of claim 6, wherein the operation for logically analyzing comprises:
- an operation for reading at least one positioning constraint for at least one field of the document, a field in which a piece of information might be
 - 25 read;
 - an operation for detecting compliance with at least one positioning constraint of at least one of the fields; and
 - during the operation for associating fields, to each field which complies with one of the constraints, is associated an identifier which is representative
 - 30 of that constraint.

8. The automatic document reading process of claims 6 or 7, wherein the operation for logically analyzing further comprises:

an operation for reading at least one graphic characteristic of at least one field of the document, a field in which a piece of information might be read;

an operation for detecting at least one field of the document that has one of the graphic characteristics; and

during the operation for associating fields, to each field which complies with one of the graphic characteristics, is associated an identifier which is representative of that graphic characteristic.

9. The automatic document reading process of any one of claims 6 to 8, wherein the operation for logically analyzing comprises:

an operation for reading at least one heading from at least one field of the document, a field in which a piece of information might be read;

an operation for detecting reading one of the headings; and

during the operation for associating fields, to each field which contains the heading is associated an identifier which is representative of that heading.

10. The automatic document reading process of any one of claims 6 to 9, wherein the operation for logically analyzing comprises:

an operation for reading at least one information content constraint for at least one field of the document, a field in which a piece of information might be read;

an operation for detecting the content of at least one field of the document; and

during the operation for associating fields, to each field which complies with the content constraint is associated an identifier which is representative of that content.

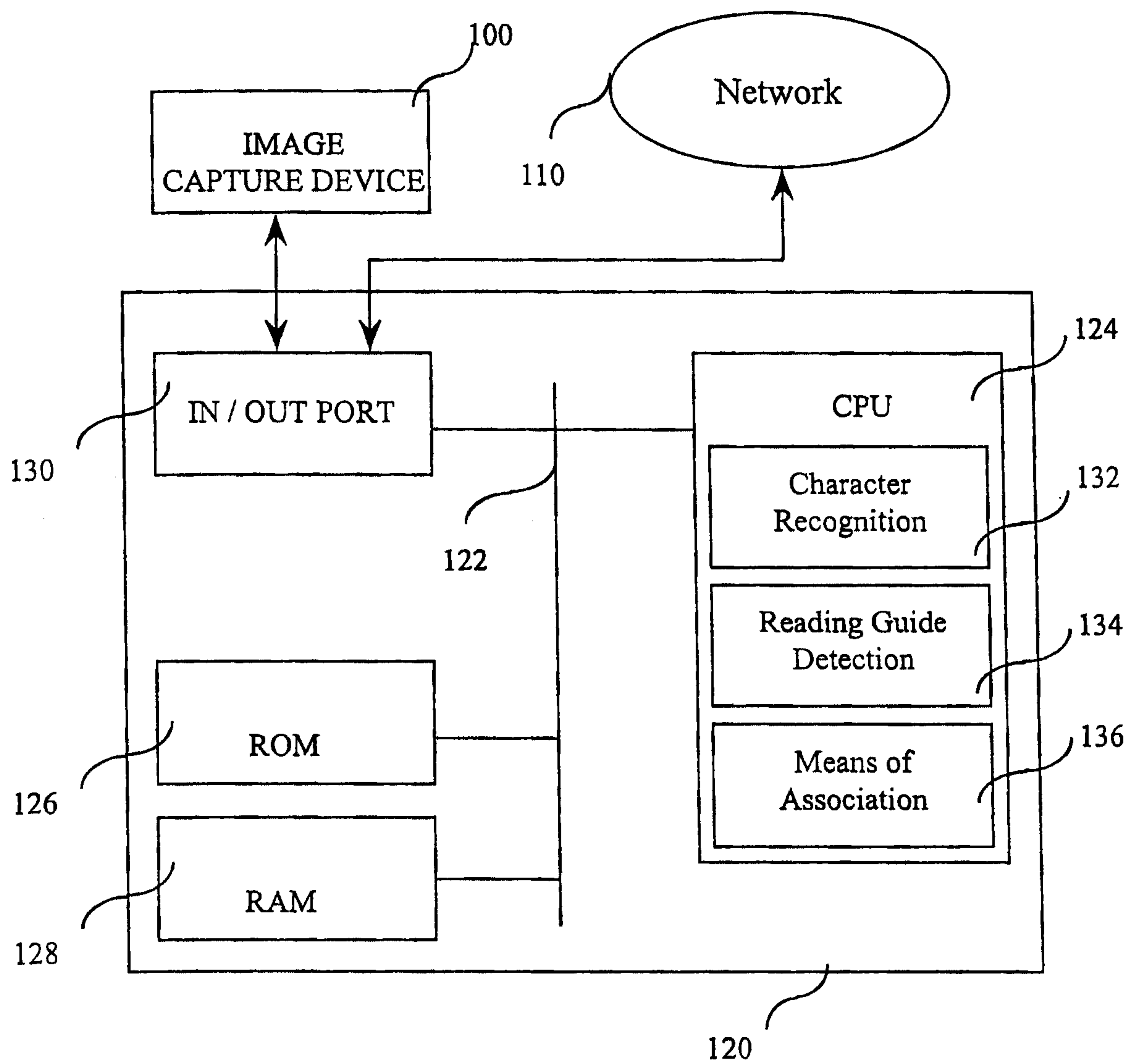


Fig. 1

2/8

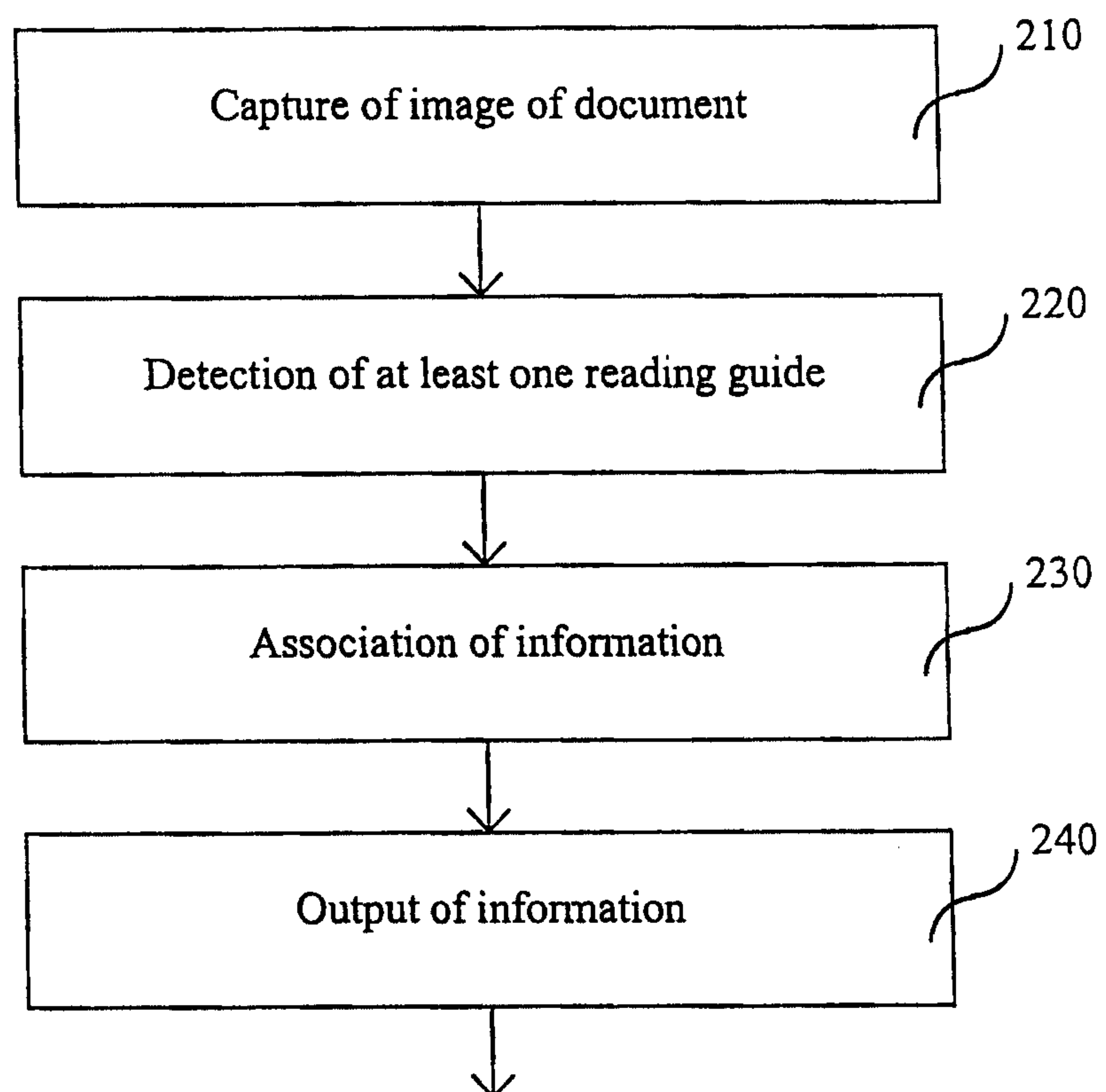


Fig. 2

3/8

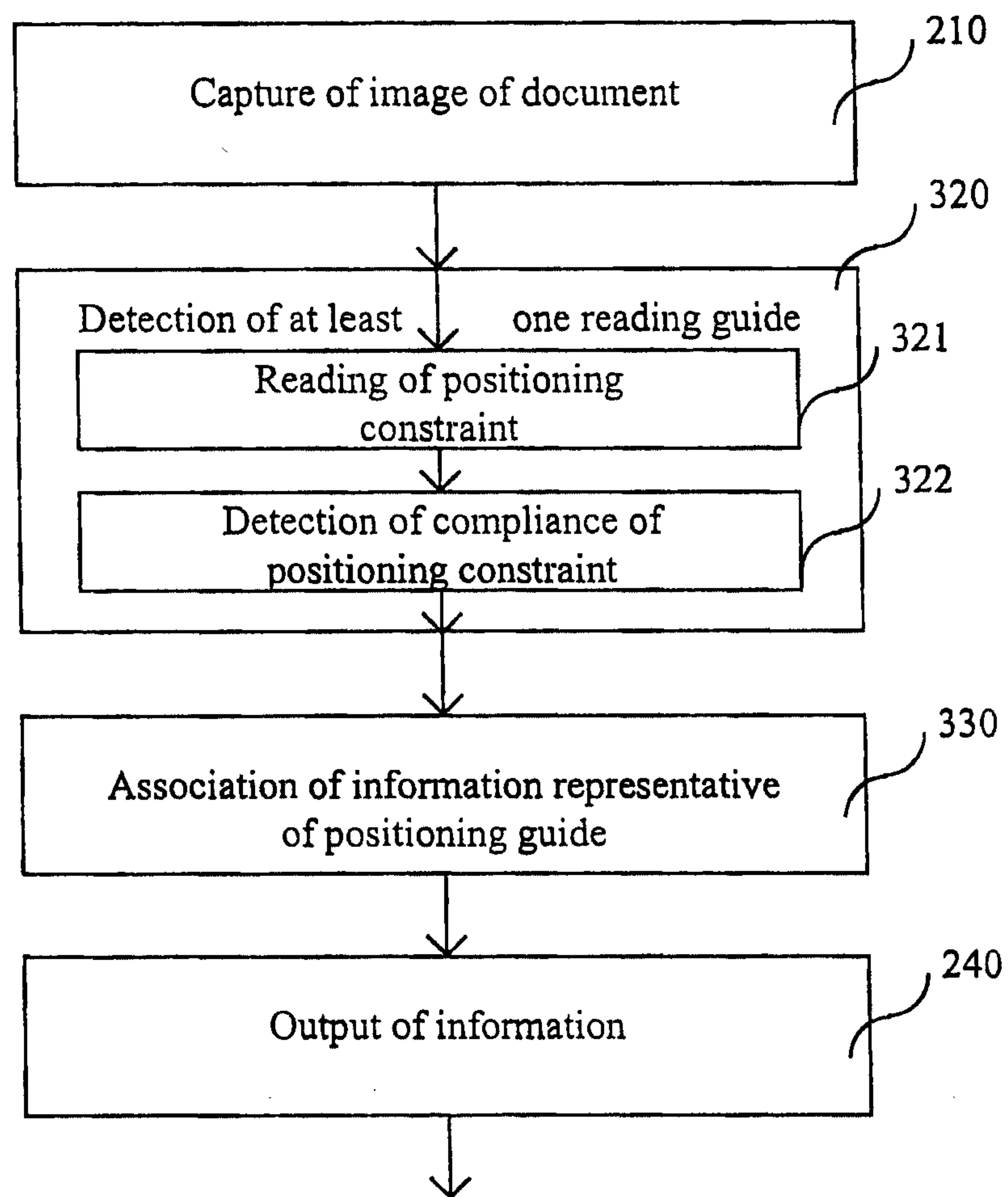


Fig. 3

4/8

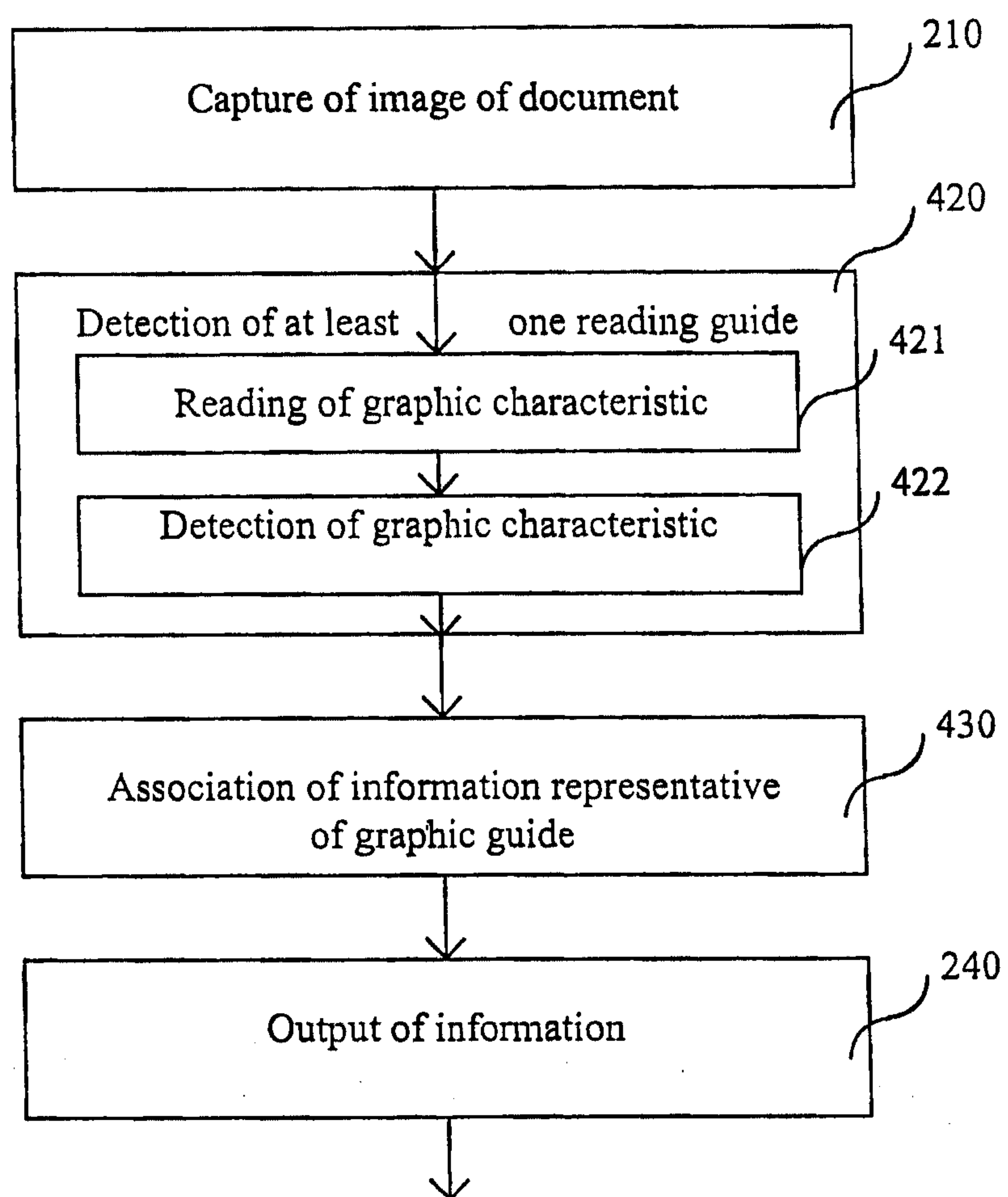


Fig. 4

5/8

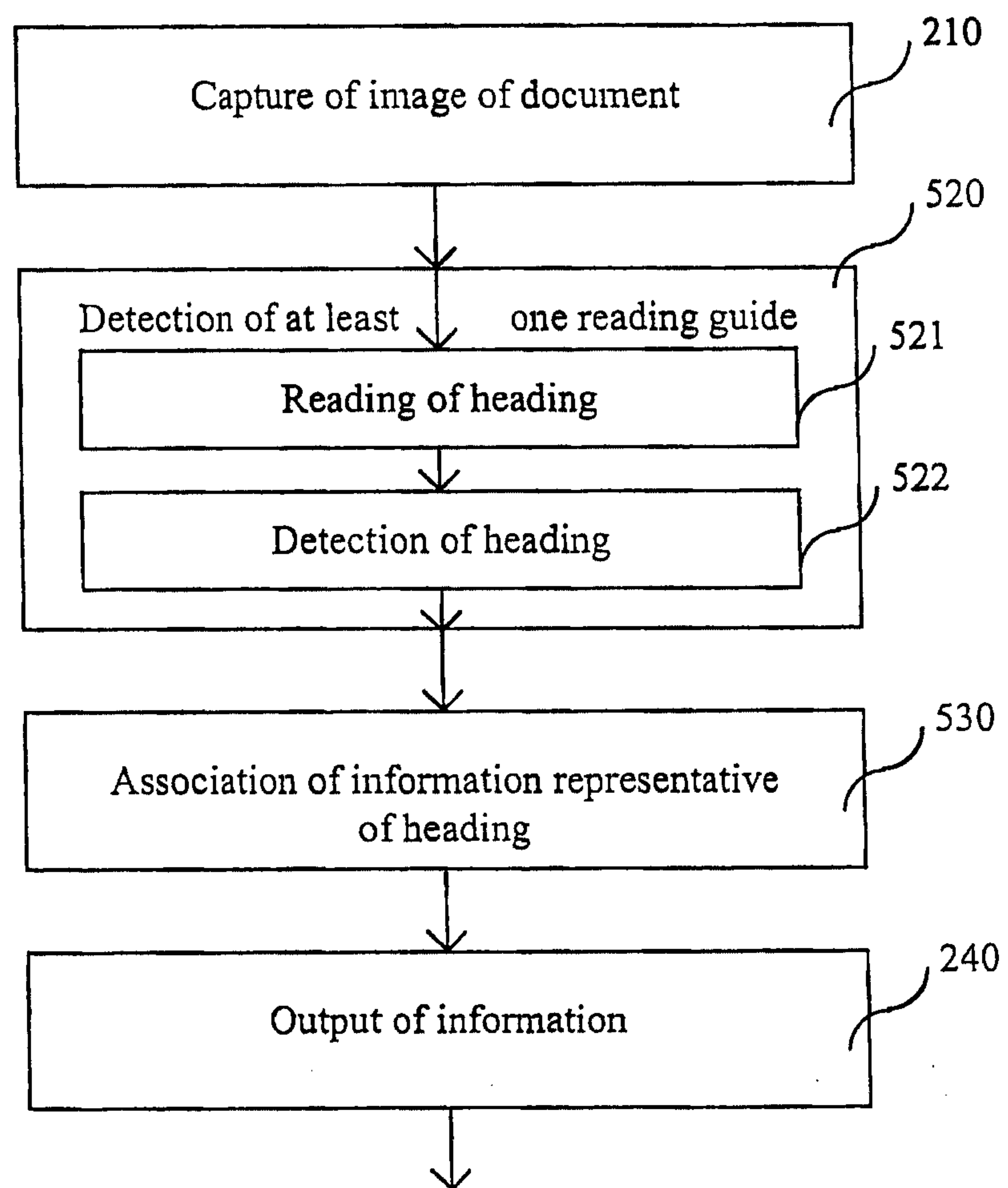


Fig. 5

6/8

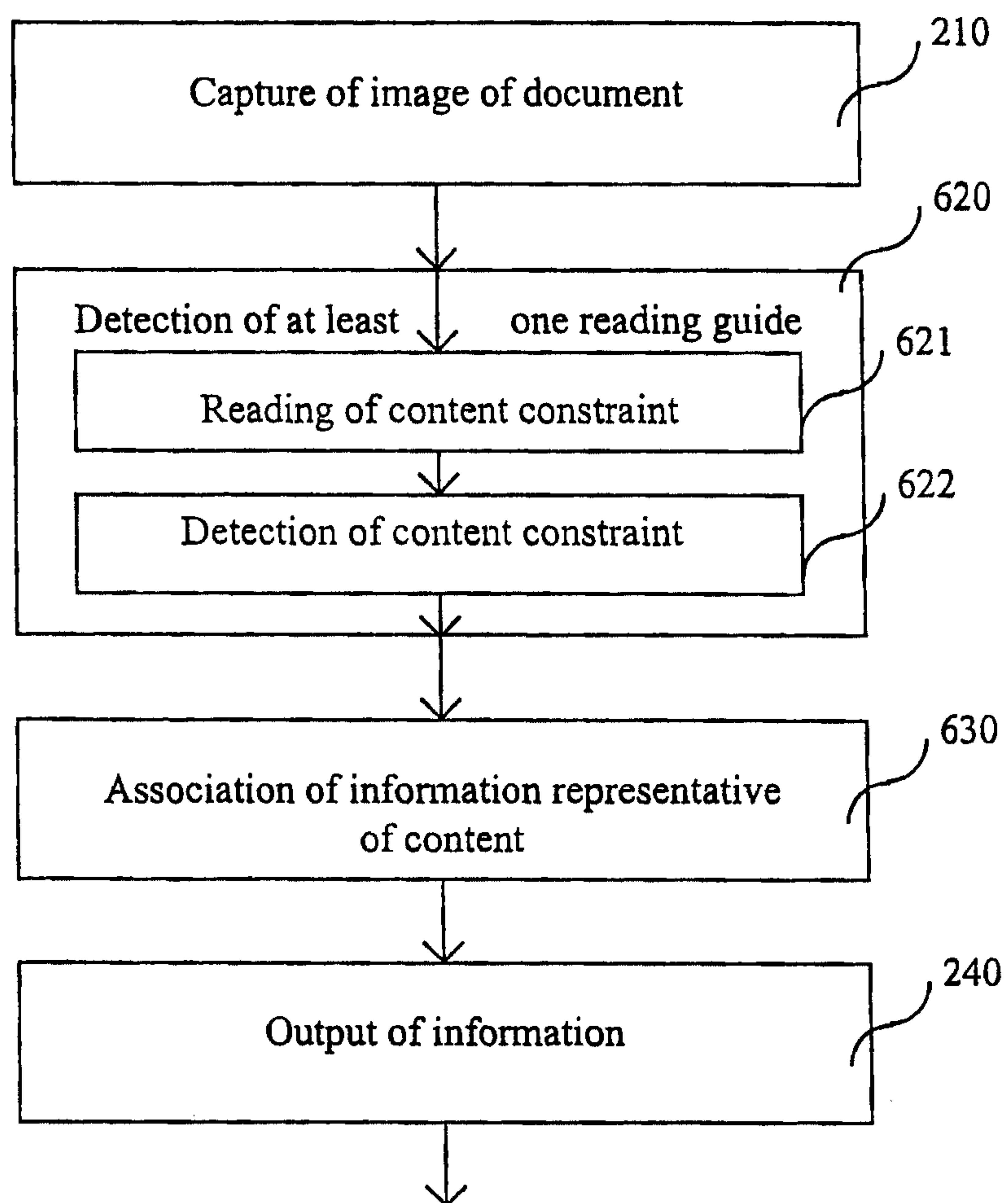


Fig. 6

7/8

704

E.D.F.
220, rue Volt
32200 AMPERES

ITESOFT
32, rue des Bruyères
23210 Cannes

706

7/14/99

700

Invoice N° 9652

705

discount 15 %

120.6 FF. 701

Net Total 20.6 FF. 702

V.A.T. (20.6 %) 100.00 FF. 703

Gross Total

707

With our compliments

Type of document	Invoice N° 9652
Address	ITESOFT, 32 rue des Bruyères 23210 CANNES
Date	7/14/99
Discount	15 %
Net Total	100.00
V.A.T. Total	20.6
Gross Total	120.6
Comments	With our compliments

Fig. 7

8/8

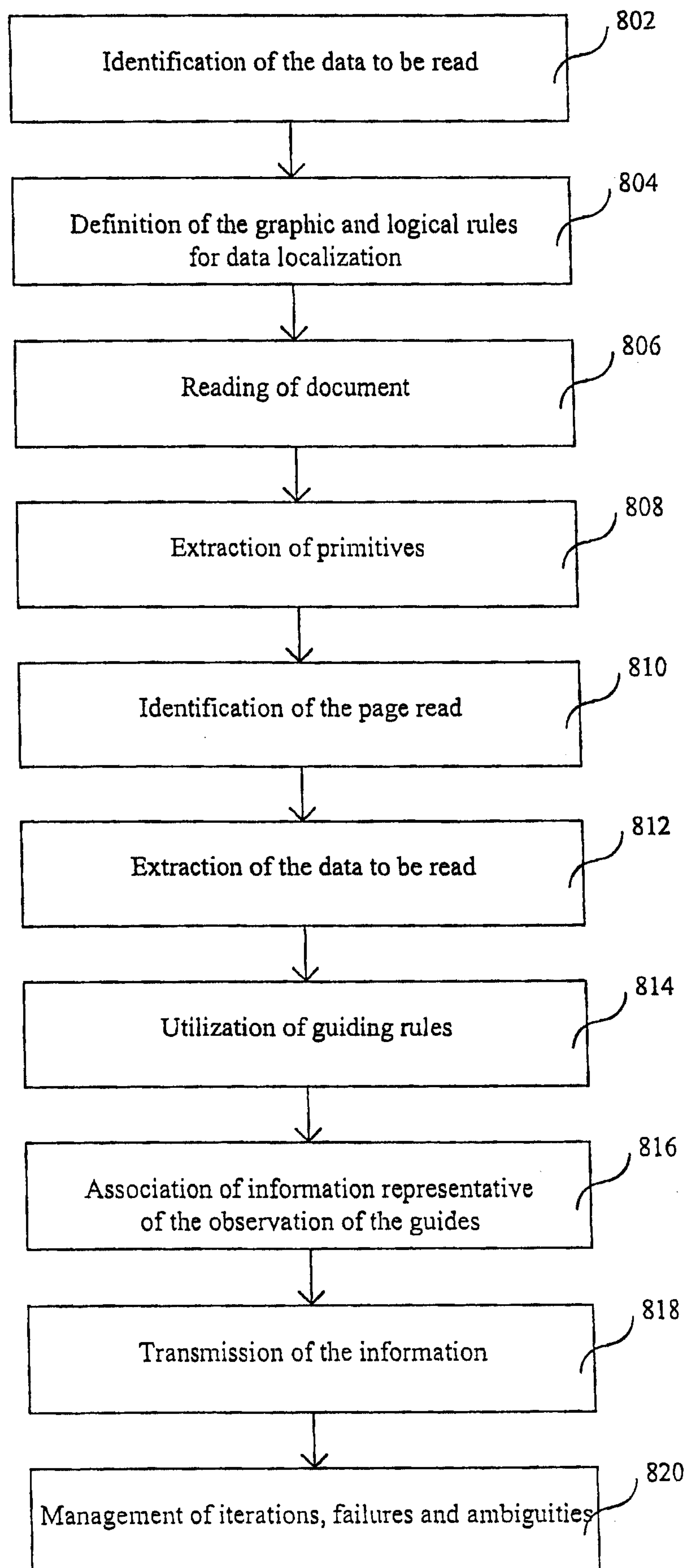


Fig. 8

