FILE RETRIEVAL DEVICE

TECHNICAL FIELD

The present invention relates to a file retrieval device in which information that is recorded in documents or the like is fetched into a computer as computer data, and the functions of a computer are utilized to consolidate and make use of that information.

Background of Art

Recent advances in computer technology have led to the development of techniques designed to transfer information that is recorded in documents, or the like, into a computer as computer data, not only to reduce the storage area of documents and the like, but also to make efficient use of information.

As a result of the above technical developments, a file retrieval device has already been proposed that is provided with an internal image scanner. This prior-art file retrieval device inputs information recorded in documents or the like, and stores it as image data. With this prior-art file retrieval device, documents are scanned with the image scanner, the image data obtained by a single scan is handled as one item of image data, retrieval keywords are input by the user for each item of image data or each group of a predetermined number of items of image data, and the image data is stored together with the retrieval keywords.

When searching for the image data of a certain document with this prior-art file retrieval device, the user inputs a keyword for the search and a retrieval means provided within the file retrieval device searches for and extracts the desired image data by searching for all image data having either an identical retrieval keyword or a partially identical retrieval keyword.

A search for a partially identical retrieval keyword

is one in which a retrieval keyword that was input as "XX 株式会社" for example when the image data was stored, would be found in response to a partially identical keyword such as "XX".

Many relevant techniques for searching on similar keywords have been proposed. If the keyword "ABC 株式会社" for example, has been input, techniques have been proposed for searching by similar retrieval keywords devised from consideration of standard representation methods in Japanese, such as "AyBeeCee 株式会社", "株式会社 ABC", or "(株) ABC". Note that 株式会社, herein, is referred to Incorporated or Limited, and is obliged by the Japanese Commercial Law to be included as a part of the name of the company, a trade name, at either a head of it or an end.

A file retrieval device with an internal character recognition device has also been proposed as an improvement on the above described file retrieval device, with the objective of compressing the volume of data to be stored.

This file retrieval device with an internal character recognition device is designed to store information from documents or the like in a small storage volume, by automatically distinguishing between text portions and photographs and other graphics portions during the acceptance of information recorded in such documents, using a character recognition device to convert the text portions into character data, accepting photographs and other graphics portions as image data, and converting the text portions into character data. With this file retrieval device too, retrieval keywords for future searches have to be input when the image data and character data are stored.

However, although retrieval keywords must be input during the process of storing the image and other data in the above described prior-art file retrieval devices, the work of inputting these retrieval keywords manually is time-consuming, hindering the rapid storage of information.

Furthermore, to enable searching by partially identical retrieval keywords and keywords with a similar representa-

tion, the user has to input even more suitable retrieval keywords for each item of image data during information storage, while considering the possibility of keywords that will be used for future searches, so that a large burden is placed on the user during information storage. Furthermore, the skill with which keywords are set determines whether a search is difficult or easy.

Even if the problem of skill in keyword setting is ignored, a prior-art file retrieval device cannot cope with a request to search by an unforeseen keyword. It is thus difficult to make flexible use of previously-stored information.

With the above described prior-art file retrieval device that converts the character data of the document for storage, the user must verify or correct the results of the conversion by the character recognition device during the acceptance of information. This verification or correction of character recognition makes the storage of information time-consuming and labor-intensive. There is also a danger that the original information may be lost if information containing an erroneously recognized character is stored without modification by some chance.

The various points described above have led to a technical problem that is it not easy to store information with a prior-art file retrieval device, and it is not possible to search for and extract desired information accurately and rapidly.

It is therefore an objective of the present invention to provide a file retrieval device that enables easy storage of information recorded in documents, and also enables easy and accurate retrieval.

SUMMARY OF THE INVENTION

In order to achieve the above objectives, a file retrieval device in accordance with the present invention comprises:

image input means for inputting characters, symbols,

and graphics as image data;

storage means for storing image data that has read in by the image input means, as image data without modification;

display means for displaying the image data in a file format wherein one item of image data is displayed as one page;

range specification means for specifying a range of a predetermined portion of image data displayed on the display means;

character recognition means for converting into character data an array of pixels within a range specified by the range specification means; and

editing means for editing as character data the character data that has been converted by the character recognition means.

The file retrieval device of this invention is characterized in having:

a similar-character-string generation portion for generating character strings similar to a predetermined character string; and

is provided with a retrieval means which prompts a user to input a character string for a search, takes as search objects the thus-input character string and a similar character string generated by the similar-character-string generation portion with respect to that input character string, and searches for these search objects within character data that has been converted by the character recognition means within a range specified by the range specification means.

The file retrieval device of this invention is characterized in that:

the similar-character-string generation portion assigns rankings corresponding to the resemblance of a character string to a predetermined character string; and

the retrieval means searches for a character string according to a ranking of candidates, where the character

string input by the user is a first candidate and similar character strings are candidates with rankings assigned by the similar-character-string generation portion.

The file retrieval device of this invention is characterized in that:

the similar-character-string generation portion generates a similar character string by referencing a file of correspondence between a predetermined character and a character similar thereto.

The file retrieval device of this invention is characterized in that:

the similar-character-string generation portion searches for a character having a shape similar to an input character, in accordance with shape-similarity rules, to generate a similar character string.

The file retrieval device of this invention is characterized in that:

the similar-character-string generation portion searches for a character that resembles an input character, in accordance with rules governing character distortion due to printing or reading, to generate a similar character string.

The file retrieval device of this invention is characterized in being further provided with:

retrieval means which converts an image in a range specified by the range specification means into character data by the character recognition means, takes the thus-converted character string as a search object, and searches for this search object within character data that has been converted by the character recognition means within a range specified by the range specification means.

The file retrieval device of this invention is characterized in that:

the range specification means specifies the same range within all image data, by specifying a range within one item of image data.

BRIEF DESCRIPTION OF THE DRAWINGS

Fig. 1 is a block diagram of an example of the configuration of a file retrieval device in accordance with this invention, showing the flow of processing thereof; Fig. 2 shows an example of a screen displayed by the display means of the file retrieval device of this invention; and Fig. 3 is an illustrative view of the search processing performed by the retrieval means of the file retrieval device of this invention.

BEST MODE FOR CARRYING OUT THE INVENTION

An embodiment of this invention will now be described, with reference to the accompanying figures.

The configuration of a file retrieval device in accordance with this embodiment of the invention is shown in Fig. 1, together with the flow of processing thereof.

The file retrieval device of this embodiment has an image input means 1, a storage means 2, a display means 3, a range specification means 4, a character recognition means 5, a retrieval means 6, and an editing means 7.

The image input means 1 could be configured in various ways, such as an image scanner 1a, a cable 1b for connection to another computer or a network, and a fax machine or photocopier that are not shown in the figure, provided that information recorded in a document or the like can be input as image data thereby.

The display means 3 of this embodiment is shown here as being independent of a display device 8 such as a monitor and is described below as a control means that controls the sending of picture data to the display device 8 for display, but the display means could equally well comprise the display device.

In a similar manner, the range specification means 4, retrieval means 6, and editing means 7 of this embodiment are shown to be independent of an input device 9, such as a keyboard or mouse, and are described below as control

means that control the range specification, search, and editing thereof, respectively, but they can equally well comprise an input means such as a keyboard.

The flow of processing within the file retrieval device of the above configuration will now be described.

The file retrieval device of this embodiment accepts all information that is recorded in documents, or the like, as image data. More specifically, a document or the like is placed on the scanning surface of the image scanner 1a, and all of the characters, diagrams, photographs, etc., recorded on that document and read in optically by the image scanner 1a are stored in an image data file 10 of the storage means 2 as image data (data recorded as an array of pixels). In this case, image data obtained by one scan of the image scanner 1a is stored as one item of image data. If there is no image scanner 1a, information that has already been turned into image data could be input from the image data file 10 through the cable 1b.

A point to be aware of in this case is that, when information is accepted with the file retrieval device of this embodiment, there is no need to input retrieval keywords for future searching of that image data during information storage. This makes it possible for the user to store a large volume of documents mechanically and rapidly. Note that the search for the thus-stored information is described below.

The display means 3 extracts image data from the image data file 10 and displays it on the display device 8. The display provided by the display means 3 is in a file-like format such that one item of image data is shown as a page to which are affixed headers defined by a predetermined classification scheme, as shown by way of example in Fig. 2. This file-format display makes it possible for the user to open any desired portion of the image data rapidly, by clicking on a header with a mouse or the like.

The display means 3 is also preferably provided with functions such as "Turn the pages rapidly", "Expand,

compress, rotate, or browse", "Mark", and "Add comment".

The description now turns to the method used to search for certain information, or rather, certain image data.

With a search done by the file retrieval device of this embodiment, the portion of the image data to be searched is initially specified by the range specification means 4. In practice, the user uses an input means such as a mouse while viewing the display device 8, to specify a search range field 11 on the image data, such as is shown in Fig. 2. Thus if, for example, the title is written at a predetermined location in a fixed-format document such as a form and a keyword for the search is comprised within the title, enclosing that portion alone by the search range field 11 makes it possible to perform an efficient search with a small search volume.

If the desired keyword is not written at a specified portion (location), a search can be performed over all the portions of the image data, by using the range specification means 4 to enclose all of the image data within the search range field 11.

Note that the range specification means 4 in accordance with this embodiment can specify the same range in all of the items of image data, by specifying a range within one item of image data. Use of this function with the above described example of a form makes it possible to search the title portions of all of the forms by enclosing the title of one form within the search range field 11. This function is particularly effective when searching an image data file 10 that contains only image data for fixed-format documents.

The search range specified in this manner is stored in a range specification file 12 of the storage means 2.

The array of pixels in the portion specified as a range by the range specification means 4 is then converted into character data by the character recognition means 5. The character recognition means 5 references the range specification file 12, extracts image data from the image data

file 10, and converts the array of pixels within the specified search range into character data while referencing a dictionary file 13.

The thus converted character data is stored in a character data file 14. These items of converted character data form a collection of character strings for a search.

The retrieval means 6 then prompts the user to input the character string for a search, then searches for the input character string and a character string that resembles this input character string from the collection of character strings in the character data file 14.

The flow of the search performed by the retrieval means 6 is shown in Fig. 3. The retrieval means 6 of this embodiment is characterized in that it searches not only for the input character string, but also for character strings that resemble the input character string. This is described below by way of a specific example.

If, for example, a search is made for image data comprising the character string 中間決算 in Sino-Japanese characters, it is possible that the character string 中間決算 has been identified incorrectly by the character recognition means 5 for storage in the character data file 14. For example, it is possible that the character 中 has been erroneously identified as 牛, 年, or 甲, etc; the character 間 has been erroneously identified as 問, 関, or 門, etc; the character 決 has been erroneously identified as 法, 沫, or 洟, etc; or the character 算 has been erroneously identified as 筧 or 筧, etc.

It is therefore possible that the character string 中間決算 is stored as a combination of the above characters in the character data file 14. Such a erroneously identified character string will not be retrieved by the character string 中間決算.

To counter this, the retrieval means 6 of this invention has a similar-character-string generation portion 15 that generates character strings which are similar to the input character string. If, for example, the character 中

is input, the similar-character-string generation portion 15 selects 牛, 年, or 甲, etc., which each resemble that character, as components in the construction of similar character strings.

In this case, there are three methods that can be used as methods of selecting these similar characters.

With the first similar-character selection method, a file of correspondence between a predetermined character and a character similar thereto is prepared beforehand, and this file is referenced for the selection of similar characters. For example, if the character 中 is input when the characters 牛, 年, and 甲, etc., have been previously stored in the file as characters that could be mistaken for the character 中, those characters 牛, 年, and 甲, etc., are selected. If the character recognition means 5 can be used in common, this equivalence file is also used in common thereby.

The second similar-character selection method uses character-shape rules that determine a character by factors such as character outline and line density, to select characters that resemble the shape of the input character. If, for example, the character 中 is input, the characters 牛, 年, and 甲, etc., which resemble that character are selected by the shape-similarity rules. If the character recognition means 5 can be used in common, these character-identification rules are also used in common thereby.

The third similar-character selection method selects characters that resemble the input character according to character-distortion rules that give a number of examples of the ways in which characters can be distorted by the printing or reading process. For example, the numeral "1" could be mistaken for the alphabetic characters "i" or "l" ("ell"), or the parenthesis "(", depending on the printing or reading circumstances, so that "i", "l" ("ell"), and "(" are selected as similar characters if "1" is input.

Thus, if the character string 中間決算 in Sino-Japanese

characters is input as the search object character string, the retrieval means 6 of this device uses 牛間決算, 中問決算, 中間法算, etc., as search objects in addition to 中間決算. These search object character strings are checked one-by-one against the character strings in the character data file 14, to search for identical character strings.

In this case, the retrieval means 6 preferably has a list of the possible erroneous identifications of predetermined characters, as probability values, and assigns rankings to the similar character strings. This ensures that the device first of all searches for a character string that exactly matches the search object character string, then searches for other character strings, starting with the erroneous identification that is most likely to occur, and displays the results in a manner that represents the possibility of error.

The search results are displayed as image data n1, n2, n3, etc., comprising the relevant character string, as shown in Fig. 3, with the portion of the relevant character string emphasized. These retrieved results are stored in a retrieved data file 16 (see Fig. 1) of the storage means 2.

In accordance with this search method, character recognition is performed only on portions of the image data comprising a retrieval keyword, and the character string input for the search and a character string that is similar thereto are used as search objects, without any check that the results of the character recognition are correct, to search image data that comprises that character string. This makes it possible to 1) reduce the volume of character recognition processing, 2) reduce the labor involved in checking the results of the character recognition, and 3) enable retrieval of the image data comprising the character string being searched for, without omissions.

Note that the above description assumed that the retrieval means 6 searches for similar character strings that resemble the entire character string being searched

for, but the present invention is not limited thereto and the retrieval means 6 can equally well be devised to generate a search object and similar character strings relating to a fragment of a character string that has been input for a search.

In other words, if the keyword for a search is assumed to be "ABC 株式会社", for example, it is possible to retrieve "ABC 株式会社" by an input character string that is merely "AB", "ABC", "A****会社", or "A".

Furthermore, if similar character strings are used in a search where the keyword is 中間決算, it is possible to specify the 中 in 中間決算 and search on similar character strings such as 牛間決算, 年間決算, and 甲間決算, which comprise the 牛, 年, and 甲 characters that resemble 中. Of course the above character specification can be extended so that any two or three characters are substituted into the similar character string, as specified by the user.

This file retrieval device also makes it possible for the user to search by another method that differs from the above-described method of inputting a search object character string. This search method concentrates on a predetermined character string within predetermined image data, and retrieves image data having a character string that is identical to this character string. This alternative search method is described below.

This search method is exactly the same as that of the previously described search in which a character string is input, as far as the creation of the collection of character strings 14 for the search. The character strings for the search are then converted into character data by the range specification means 4 and the character recognition means 5. In this case, the character strings identified by the character recognition means 5 are left unmodified, so that even if there are errors in the identification, they are used unmodified as search object character strings.

If, for example, the character recognition means 5 has erroneously identified 中間決算 as 牛間決算 when a search is

to be made on the character string 中間決算, 牛間決算 is used without modification as the character string that is the search object. This is because it is highly probable that 中間決算 has been erroneously identified by the same character recognition means 5 as 牛間決算 in the collection of character strings 14 for the search, so that a search on 牛間決算 will extract the desired image data. The display and storage of the search results are exactly the same as in the above search by the input of a character string.

The above describes the search performed by this file retrieval device. This file retrieval device uses the character recognition means 5 to convert a predetermined array of pixels within image data into character data, and this data can then be edited by means such as copying it into a document within a word-processor.

As shown in Fig. 1, the editing means 7 of this device enables the user to specify a predetermined range by the range specification means 4 while referring to image data displayed on the display means 3, and convert that into character data by the character recognition means 5. This character data is stored in an editing data file 17 and can then be used in the editing of text in a word-processor or the like. Furthermore, a predetermined range can be cut out by the range specification means 4, this can be stored unchanged in the editing data file 17, and it can be submitted for inclusion in text in a word-processor or the like.

This makes it possible to utilize information on various different documents stored as image data, so that new information can be created from the existing information if it should become necessary.

As is clear from the above description, the file retrieval device of this invention can store information obtained by an image input means from various documents, as image data without modification, and it is not necessary to affix retrieval keywords thereto or perform character recognition during the information storage process, as it

is in the prior art. Thus document information can be stored mechanically and a large volume of information can be stored rapidly.

The file retrieval device of this invention uses a range specification means to specify a range to be searched, then uses a character recognition means to subject an array of pixels within that range to character recognition. The results of the character recognition are used as a character string for a search, without being checked for accuracy or corrected. A character string that is input for a search by the retrieval means and a character string that resembles it are used as search object character strings, and character strings that are equivalent thereto are retrieved from the character string subjected to the above described search. This makes it possible to reduce the volume of character recognition processing, reduce the labor involved in checking the results of the character recognition, and also enable retrieval of the image data comprising the character string being searched for, without omissions.

If a search is performed by specifying a character string within the image data, to search for image data that comprises the same character string, character recognition errors caused by the printing or reading circumstances are used without modification, enabling the desired image data to be extracted accurately with a small volume of processing.

Furthermore, a predetermined portion of the image data can be cut out by this file retrieval device if necessary, that portion can be converted by the character recognition means so that it is not handled as image data alone, and it can be utilized in a simple manner by the editing means.

INDUSTRIAL APPLICABILITY

The file retrieval device of this invention can be applied as a database device for image data.

CLAIMS

1. A file retrieval device comprising:

image input means for inputting characters, symbols, and graphics as image data;

storage means for storing image data that has read in by said image input means, as image data without modification;

display means for displaying said image data in a file format wherein one item of image data is displayed as one page;

range specification means for specifying a range of a predetermined portion of image data displayed on said display means;

character recognition means for converting into character data an array of pixels within a range specified by said range specification means; and

editing means for editing as character data the character data that has been converted by said character recognition means.

2. The file retrieval device as defined in claim 1, having a similar-character-string generation portion for generating character strings similar to a predetermined character string; and being provided with a retrieval means which prompts a user to input a character string for a search, takes as search objects the thus-input character string and a similar character string generated by said similar-character-string generation portion with respect to that input character string, and searches for said search objects within character data that has been converted by said character recognition means within a range specified by said range specification means.

3. The file retrieval device as defined in claim 2, wherein said similar-character-string generation portion assigns rankings corresponding to the resemblance of a

character string to a predetermined character string; and

said retrieval means searches for a character string according to a ranking of candidates, where said character string input by said user is a first candidate and similar character strings are candidates with rankings assigned by the similar-character-string generation portion.

4. The file retrieval device as defined in claim 2, wherein said similar-character-string generation portion generates a similar character string by referencing a file of correspondence between a predetermined character and a character similar thereto.

5. The file retrieval device as defined in claim 2, wherein said similar-character-string generation portion searches for a character having a shape similar to an input character, in accordance with shape-similarity rules, to generate a similar character string.

6. The file retrieval device as defined in claim 2, wherein said similar-character-string generation portion searches for a character that resembles an input character, in accordance with rules governing character distortion due to printing or reading, to generate a similar character string.

7. The file retrieval device as defined in claim 1, being further provided with retrieval means which converts an image in a range specified by said range specification means into character data by said character recognition means, takes the thus-converted character string as a search object, and searches for said search object within character data that has been converted by said character recognition means within a range specified by said range specification means.

8. The file retrieval device as defined in claim 7, wherein said range specification means specifies the same range within all image data, by specifying a range within one item of image data.