ANNOTATION OF DIGITAL IMAGES USING TEXT

Filed of the Invention

The present invention relates to the annotation of digital images using text and refers particularly, though not exclusively, to the annotation of images using text recorded in a plurality of fields, each field having a predetermined title.

Background of the Invention

There are presently available a large range of portable digital image capturing devices such as, for example, digital cameras, camera attachments for desktop computers, personal computers, notebook computers, PDAs, and so forth; and camera-enabled 3G phones. These, together with the 3G mobile network, will enable a greater penetration of digital still and moving (video) images into the commercial and domestic arenas. Storing and indexing those images for subsequent retrieval is becoming a major task. Useful tools are required to enable a simple way of managing the digital images. An automatic indexing module will complement other more established and widely available tools such as database management systems for providing storage, query and retrieval functions. Present techniques are to index content using manually entered textual and graphic information. This is time consuming, labour intensive, and prone to errors due to interpretation of the content well after the capturing of the image.

The commercial success of speech recognition technology in domain-specific applications such as navigating mobile devices, voice activated command interfaces and speech-to-text transcriptions, together with the introduction of new generation of digital cameras with a built-in microphones (for example Sony CyberShot DSC-S70, Fujifilm FinePix 4700Zoom, Kodak DCS315, Ricoh RDC-i700 and many more) for speech storage, will allow speech annotation immediately after image capture. This is more effective since the information about the place, event, mood and intention are still fresh in the mind of the cameraman.

It is therefore the principal object of the invention to provide a method for the annotation of digital images by recording key information for each image as text.

A further object of the invention is to record the key information under a plurality of fields, each field having a predetermined title.

Summary of the Invention

With the above and other objects in mind, the present invention provides a method for annotating images wherein key information for each image is stored with each image as text. The key information is stored under a plurality of fields, with each field having a predetermined title. The predetermined titles form part of the key information. The predetermined titles of the fields may be determined by a user or maybe pre-set by a supplier. They may be event, place, people, and date; in any order.

The key information may be input as audio and converted to text using an automatic speech recognition engine. Alternatively, and especially for those who may have speech difficulties, the key information may be input by keyboard, keypad, touch screen or imaging.

Preferably each predetermined title is input before the key information for the field relevant for that predetermined title. After the audio input is converted to text, each of the predetermined titles may be matched to its counterpart word in the audio input. All words that are after the predetermined title and before the next occurring predetermined title or the end of the audio input, whichever occurs first, may then be extracted as a description for that field.

The automatic speech recognition engine preferably should be able to edit its vocabulary, correct frequently occurring transcription errors, incorporate new words into its vocabulary, and provide alternatives in addition to final recognition result while recognizing the audio input.

If desired, character information for an image may also recorded for that image. The character information may be information such as, for example, global positioning system coordinates, and camera-related information.

The key information and the character information are preferably stored on a database. The database may store the digital images singularly as a single image database. Alternatively, or additionally, the database may store the digital images in clusters as a multiple image database.

The key information for a plurality of digital images may be clustered by phonetically similar words that occur in a majority of the digital images in the cluster. Clustering may be achieved by using a nearest neighbour clustering algorithm. This may be based on threshold. When the descriptions for a given predetermined title are the same for the plurality of digital images, the digital images are clustered by descriptions of a different one of the predetermined titles. To prevent misplacement of a digital image, further clustering processing is conducted for fringe relocation. Data from the clusters can then be used to update the descriptions.

Words occurring in the majority of the digital images may be used in the clustering process. Also, clustering may be achieved by using a nearest-neighbour clustering algorithm based on threshold. Words that occur in the majority of the key information of all digital images in a cluster may be taken into account in the clustering process. One or more of the predetermined titles may be dependent on another of the predetermined titles during the clustering process.

Fringe relocation may be by assigning a value representing a dominant element in a field throughout all images in the cluster. A distance between the single image value for a single digital image and its cluster is then determined, as are subsequent distances between the single image value and its adjacent clusters. A normalized value of the distance and the subsequent distances is then obtained. The normalized distance and the normalized subsequent distances are compared, the single digital image is placed in the cluster where the normalized distance, or the normalized subsequent distances, are the lowest.

A weighing rule may be applied for the plurality of fields in determining the subsequent distances.

The matching of the predetermined titles is by keyword spotting; or may be by using the automatic speech recognition results and searching for the predetermined titles.

The present invention also provides a computer useable medium comprising a computer program code that is configured to cause a processor to execute one or more functions as described above.

Description of the Drawings

In order that the invention may be readily understood and put into practical effect, a preferred embodiment of the invention will now be described by way of non-limitative example only, and with reference to accompanying illustrative drawings, in which:

Figure 1 is a flow chart for a speech annotation process;
Figure 2 is an illustration of an annotation speech structure; and
Figure 3 is an illustration of the field segmentation process.

Description of the Preferred Embodiment

The indexing process assumes that the image captured, and the results of the speech annotation process, are stored in a memory device in any multi-media format (e.g. MPEG, AVI and so forth). Figure 1 shows the indexing processing after the digital still/moving images, and the associated speech annotations, are downloaded to the host. The first module 10 is to separate the multi-media content (image plus audio) into the image 12 and audio fields 14. The image content can either be
(a) stored directly into the database 16; or
(b) further processed at 18 to derive contextual information using spatial attributes.

The further processing is to:

(a) increase the signal-to-noise ratio;

(b) remove artifacts caused during recording (i.e. microphone and ambient); and

(c) be converted to a suitable format, if necessary.

CBIR is Content-Based Image Retrieval. Known image-based processes will not be described as they do not form part of the present invention.

As depicted in Figure 1, the audio or speech content is first preprocessed at 20 to enhance the quality of the speech signals before being transcribed into text at the Automatic Speech Recognition ("ASR") module 22. The ASR 22 can be any commercial off-the-shelf engine 28 that preferably has the flexibility for word editing and training in its vocabulary structure. This is a useful feature as some words of local flavor or the name of places and persons may not reside in the indigenous vocabulary of the ASR. The ability to add these words can further improve the performance of the speech recognition engine 28.

To further improve the ASR module 22, the engine 28 should preferably support the following additional customization functions:

(a) enhancing the speech model to correct frequently occurring transcription errors, as well as incorporating new words into the vocabulary;

(b) providing alternatives in addition to the final recognition result, while recognizing an audio input.

The pre-determined field titles can be emphasized using the first function (a) in order to achieve high recognition rates for the field words. Likewise, the name of family members can also be trained into the ASR engine as, for home photographs, family members will quite often be in the photograph. The field titles, and commonly used field descriptions (e.g. names of family members) are stored in a field-based dictionary 36. In addition, due to the well-known uncertainties in the speech recognition process, the second function (b) can be especially helpful to determine the original content of the speech by providing more detailed information about the process.

A pre-determined syntax for the input structure of the speech is preferably used. Structured speech has been used to control many speech-activated devices such as cell phones, and other handheld devices. A relatively high recognition accuracy of these implementations is achieved by restricting the vocabulary of commands and indexed words. Here, the key information of images, either description or raw content, can be extracted from the speech annotation. These extracted terms will be used as index descriptions of the image. Because of the subjective nature of index creation for images of different categories (e.g. scenery, family portrait, interior design, urban setting, countryside, and so forth), the user may be given the flexibility to define content sub-categories or titles that best suit their indexing needs. Alternatively, these may be pre-set by the supplier. For example, in a digital camera intended for general domestic use, they may be set by the manufacturer as Event, Place, People and Date as these would be the most commonly used, and most relevant, titles or sub-categories for everyday general domestic use. These sub-categories or titles are for fields in the speech structure.

The basic speech structure is shown in Figure 2. The field titles can vary with the categories of the photos/video. For example, as is explained above, in a home photo scenario, it may be useful to categorize the photos into the following fields: ***Event***, ***Place***, ***People*** and ***Date***. Following each field is the list of the elements or description of the field. Querying the elements of these fields will enable retrieving the content of the "album".

To partition the transcribed text into the appropriate field content, a field word-detecting algorithm is used to ascertain the location of the fields within the text. There may be two ways to do this. The first is using the ASR results 24. This algorithm sequentially searches through the words in the text and their alternatives, and then matches the selected word with the list of keywords, comprising the fields titles words entry. As is described above, the prior word level training of the ASR module enables the fields titles words to be detected with relatively high confidence. Thus the interval between the detected field titles (i.e. F_n and F_{n+1}) will determine the sub field content of the F_n field.

Alternatively, the field segmentation can also be regarded as a form of keyword spotting ("KWS") 26 and may be carried out based on signal-level processing methods. Because the set of pre-defined field words or titles is preferably relatively small (e.g. four), it is appropriate to create a template for each word. The minimum number of fields is two with no theoretical maximum, although practicality such as available memory, and processing speed will provide the maximum limit in each case. Additional filler templates may also be needed to absorb all other words. Templates may be represented by a sequence of feature vectors. Upon establishment, the speech annotation is compared with the templates through, for example, dynamic-time warping to determine which part of the speech signal is most similar to which template. Thus the beginning and ending points of the field words can be approximately determined. This facilitates the recognition of content in the field.

Figure 3 shows an illustration of the field segmentation process. The transcribed annotation is shown at the top. The four field titles in this illustration are ***Event***, ***Place***, ***People*** and ***Date***. The field segmentation process will yield the locations of the uttered field words within the text, either through ASR 24 results or through signal level processing 26. The ensuing text (within in the shaded rectangular sub field box) before the next field word (for example "wedding ceremony") will contain the description of the respective field. Because of the sequential way of detecting field words, there are no restrictions on how these field words need to be organized for each audio input. For example, a speech annotation in the form of "***Date...People...Place...Event...***" will, after segmentation, lead to the same field content as that of "***Event...Place...People...Date...***". This allows flexibility in defining the sub-categories or title in a speech structure, and performing annotations for images of different categories.

In the proposed speech structure of the annotation, the content of the ensuing field title, before the next field title or the end of the annotation, whichever occurs first is the description of the field category. These are extracted at 38. For example, in the previous example described above in relation to Figure 3, *"wedding ceremony"* are the textual elements describing the ***Event*** field. These elements can be stored directly as the field meta data of the accompanying digital image, or may passed through a parser to

extract higher level information. When the field segmentation process 26 is implemented correctly, every segment generated can be re-fed into the ASR engine 28 for the ASR engine 28 to recognize as belonging to the corresponding field. This may improve recognition, as well as the resulting extraction performance..

Digital images may be associated with information stored in any character format representation (for example, ASCII, ISO-8859-1 or UNICODE). Examples of such information are:

- GPS (Global Positioning System) coordinates for longitude/latitude/altitude where the photograph is taken; and/or
- camera-related information such aperture, zoom information, speed, use of flash, landscape/portrait mode, and so forth.

This information describes each image, and is processed by an adequate character-based extraction process 30 and stored in the database 16 to be used for visualization and/or retrieval purposes.

The extracted fields are stored in the Single Image Database ("SID") 32. The SID 32 stores the speech-based and character-based extracted fields for each image. The storage is for each image separately from all other images. The database 32 allows effective and efficient storage of index information pertaining to the image. The database 32 facilitates the retrieval of relevant images, as well as providing the required information for report generation 34 purposes.

The accuracy of the field element extraction process 38 is dependent on the recognition performance of the ASR engine 28. To improve the accuracy, clustering 40 may be used. Here, the collection of images is partitioned using a predefined structure of at least one field content. The clustering process is then used to group together images that are similar. Images that are similar according to the clustering fields may also have strong similarities in the non-clustering fields. The cluster can then be indexed using representative elements of the image fields. For fields corresponding to text extracted from speech, it is possible to represent the cluster fields by a group of phonetically similar words that occur in a majority of the images of the cluster. For character-based

fields processes such as interval generation, or sets, intersection of field attributes may be used to represent the major features of the images of the cluster.

A nearest neighbour clustering algorithm based on threshold may achieve the clustering of digital images. The general nearest neighbour clustering algorithm is described below. To find the nearest clusters in step 4 of the algorithm, a clustering criterion $D_{min}(D_i,D_j) = \min(\text{similarity}(x_{il},y_{jm}))$ (with $x_{il} \in D_i$, $y_{jm} \in D_j$) is compared to a threshold T dependant on the application.

1. initialize $D_i = \{x_i\}$; $i = 1,...,n$
2. do
3. if any nearest clusters D_i and D_j with $D_{min}(D_i,D_j)$ less than T
4. then Merge D_i and D_j
5. until no merging
6. return the clusters

For example, in the domain of home photo albums using the annotation structure mentioned above, all photos taken during holidays in USA spanning several days or weeks might share the same information in the field ***Event***. For example, this may be "USA Holiday". Meanwhile, for images taken on the same day, the clustering rule below may be used so that an excessively large number of images are not in the one cluster. This enables the clusters to be of a manageable size. It is preferred that the maximum number of images in a cluster be limited to a maximum number, the maximum number may be predetermined, or may be determined by the processing capabilities of the host. In this clustering rule, the similarity between images is the difference of the Date attribute computed in seconds, and the threshold used is $T_{seconds}$.

For each group of photos

- Sort the photo according to the time in which it is taken.
- Apply the Threshold-based nearest neighbour algorithm on the Date attribute of the photographs with a threshold $T_{seconds}$.

Photo 1: ***Event*** USA Holiday ***Place*** New York City ***People*** Jill ***Date*** 14 April 1992 10 am
Photo 2: ***Event*** Holiday in USA ***Place*** New York City ***People*** Jack ***Date*** 14 April 1992 10.30 am
} TS_1

Photo 3: ***Event*** Holiday in USA ***Place*** New York Cited ***People*** Jack ***Date*** 14 April 1992 10.34 am
Photo 4: ***Event*** USA Holiday ***Place*** New York City, Battery Park ***People*** Jill ***Date*** 14 April 1992 12.50 pm } TS_2
Photo 5: ***Event*** USA Holiday ***Place*** New York City, Battery Park ***People*** Jill ***Date*** 14 April 1992 12.56 pm
Photo 6: ***Event*** USA Holiday ***Place*** New York City, Battery Park ***People*** Jack ***Date*** 14 April 1992 2.01 pm
Photo 7: ***Event*** USA Holiday ***Place*** Ferry Ride to Ellis Island ***People*** Jack ***Date*** 14 April 1992 2.45 pm
Photo 8: ***Event*** USA Holiday ***Place*** Ferry Ride to Ellis Island ***People*** Jack, Jill ***Date*** 14 April 1992 2.55 pm } TS_3
Photo 9: ***Event*** USA Holiday ***Place*** Ferry Ride ***People*** Jack, Jill ***Date*** 14 April 1992 3.05 pm

Table 1

The result of the time clustering is composed of the three clusters TS1, TS2 and TS3 of Table 1 which shows the results of the grouping by time with T = 60 minutes. The fields that describe a cluster have the same names as those coming from the images. From each field of the images in a cluster that is dependent from the clustering criterion, the value for the corresponding field of the cluster is computed. For the fields that are not dependant from the clustering criterion, the corresponding field of the cluster is not computed. A dependant attribute from a clustering criterion is an attribute for which its value is expected to be similar across the images of the cluster.

For example, taking into account:

1. phonetic similarity between words (for example, City close to Cited) within each field;
2. words that occur in the majority of the images in a cluster;
3. the Event and Place attribute are dependent on the Date clustering;

the following fields for the three clusters defined in Table 1 are obtained:

- TS_1: **Event** Holiday in USA, **Place** New York City, **Date** 14 April 1992, 10:00-10:34 am.
- TS_2 : **Event** USA Holiday , **Place** New York City, Battery Park, **Date** 14 April 1992, 12:50 pm – 12:56 pm
- TS_3 : **Event** USA Holidays, **Place** Ferry Ride to Ellis Island, **Date** 14 April 1992, 2:01 - 3:05 pm.

However, a clustering scheme based on similarities may cause inconsistent clusters. For example, TS_3 in Table 1. In this case, Photo 6 should belong to cluster TS_2, rather than

TS_3. To prevent such misplacement of images, a Fringe Relocation or Cluster Adaptation algorithm may be used. This may be implemented in any of a number of ways including by the following steps:

1. Upon the completion of the initial clustering, each dependent field of a cluster is assigned a value to represent its dominant element throughout all images in this cluster. Meanwhile, each image in a cluster is labeled with a field-specific distance metric to describe its disparity relative to the dominant value in each field.
2. For the image at the fringe of a cluster, or the image with the largest distance value in this cluster, also calculate the distance values for all fields between this image and all other clusters adjacent to that to which it presently belongs, still through the dominant representatives. Optionally, a proper weighting rule can be applied to discriminate the various significance of different fields. It is then possible to obtain a normalized distance for each cluster and to identify the minimum candidate among the current cluster and its neighbors.

$$index\,of\,closest\,cluster = \arg\min_{i}[d_i(FringePhoto)]$$

where *i* indicates the index of the n+1 clusters in the above comparison (n being the adjacent clusters of the initial cluster of the FringeImage), and d_i(FringeImage) represents the normalized distance between the fringe image and each cluster i.

If the *index of closest cluster* corresponds the current cluster for the fringe image, no fringe relocation needs to be performed and hence proceed directly to step 4. Otherwise, re-categorize the fringe image into the cluster corresponding the *index of closest cluster*. Update the dominant elements in each field of both clusters involved in the relocation and recalculate related distance values.

3. Identify the image with the second largest value of distance. Repeat Step 3 until no more fringe images need to be relocated.
4. Select another cluster and repeat the process until all clusters are processed.

The number of neighbouring clusters may be greater than two for clustering fields greater than two.

In the previous example, as the first photo of TS_3, Photo 6 is relocated to TS_2 after Fringe Relocation. Although it is consistent with TS_3 in terms of the ***People*** field, its similarity with the ***Place*** field in TS_2 is much stronger, according to the weighting rule.

The data from the cluster fields may also be used to update the image field database at 42. This is because the information related to the clusters is considered trustworthy because it abstracts information coming from several images. For instance, in the Place field of the photograph 3 of Table 1, "cited" is a wrongly extracted word. If phonetic similarity determines that Cited and City are close to each other, and that Cited is not a word of the Event Field of the cluster containing the photo 3, then the event field of photo 3 of Table 1 should be updated to contain City not Cited.

The data concerning the cluster composition and the cluster fields are stored in the multiple image database 44. As for the SID 32, the multiple image database 44 should allow efficient and effective retrieval of clusters that are relevant to queries formulated by a user or an external system.

Besides being used it for image retrieval purposes, the database can be used as a tool to complement report generation 34. This would usually be done after a surveying operation. The single 32 and multiple 44 image databases may provide descriptions of single or multiple images at various granularity of content. This, together with manual intervention, may help expedite the report generation process.

The present invention also extends to a computer useable medium having a computer program code that is configured to cause a processor to execute one or more functions described above.

Whilst there has been described in the foregoing description a preferred embodiment of the present invention, it will be understood by those skilled in the art that many variations or modifications in details of design or construction or operation may be made without departing from the present invention.

The present invention extends to all features disclosed either individually, or in all possible permutation and combinations.

Claims

1. A method for annotating images wherein key information for each image is stored with each image as text, the key information being stored under a plurality of fields, each field having a predetermined title, the predetermined titles forming part of the key information.

2. A method as claimed in claim 1, wherein the predetermined titles of the fields are determined by a user.

3. A method as claimed in claim 1, wherein the predetermined titles are event, place, people, and date.

4. A method as claimed in claim 2, wherein the predetermined titles are event, place, people and date.

5. A method as claimed in claim 1, wherein the key information is input as audio and converted to text using an automatic speech recognition engine.

6. A method as claimed in claim 1, wherein the key information is input by one or more selected from the group consisting of: keyboard, keypad, touch screen, and imaging.

7. A method as claimed in claim 5, wherein each predetermined title is input at the commencement of the key information for the field relevant for that predetermined title.

8. A method as claimed in claim 5, wherein the automatic speech recognition engine is able to perform one or more functions selected from the group consisting of:

a. edit its vocabulary;
b. correct frequently occurring transcription errors;
c. incorporating new words into it vocabulary; and

d. providing alternatives in addition to final recognition result while recognizing the audio input.

9. A method as claimed in claim 7, wherein after the audio input is converted to text, each of the predetermined titles is matched to its counterpart word in the audio input, and all words that are after the predetermined title and before the next occurring predetermined title or the end of the audio input, whichever occurs first, are extracted as a description for that field.

10. A method as claimed in claim 1, wherein character information for an image is also recorded for that image.

11. A method as claimed in claim 10, wherein the character information includes one or more selected from the group consisting of: global positioning system coordinates, and camera-related information.

12. A method as claimed in claim 10, wherein the key information and the character information are stored in a database.

13. A method as claimed in claim 1, wherein the key information for a plurality of digital images is clustered by phonetically similar words that occur in a majority of the digital images in the cluster.

14. A method as claimed in claim 13, wherein clustering is achieved by using a nearest neighbour clustering algorithm based on threshold.

15. A method as claimed in claim 14, wherein words that occur in the majority of all digital images in a cluster are used in the clustering process.

16. A method as claimed in claim 9, wherein the key information for a plurality of digital images is clustered by phonetically similar words that occur in a majority of the digital images in the cluster.

17. A method as claimed in claim 16, wherein clustering is achieved by using a nearest neighbour clustering, algorithm based on threshold; words that occur in the majority of the key information of all digital images in a cluster being taken into account in the clustering process.

18. A method as claimed in claim 15, wherein one or more of the predetermined titles may be dependent on another of the predetermined titles during the clustering process.

19. A method as claimed in claim 17, wherein one or more of the predetermined titles may be dependent on another of the predetermined titles during the clustering process.

20. A method as claimed in claim 16, wherein when the description for one or more given predetermined titles are the same for the plurality of digital images, the digital images are clustered by descriptions of a different predetermined title.

21. A method as claimed in claim 20, wherein to prevent misplacement of a digital image further clustering processing is conducted for fringe relocation.

22. A method as claimed in claim 21, wherein data from the clusters after fringe relocation is used to update the descriptions.

23. A method as claimed in claim 12, wherein the database stores the digital images singularly in a single image database.

24. A method as claimed in claim 12, wherein the database stores the digital images in clusters in a multiple image database.

25. A method as claimed in claim 24, wherein there is provided a further database being a multiple image database for storing digital images in clusters.

26. A method as claimed in claim 21, wherein fringe relocation is by the steps of:

(a) assigning a value representing a dominant element in a field throughout all images in the cluster;

(b) determining a distance between the single image value for a single digital image and its cluster;

(c) determining subsequent distances between the single image value and its adjacent clusters;

(d) obtaining a normalized value of the distance and the subsequent distances;

(e) comparing the normalized distance and the normalized subsequent distances; and

(f) placing the single digital image in the cluster where the normalized distance, or the normalized subsequent distances, are the lowest.

27. A method as claimed in claim 26, wherein a weighing rule is applied for the plurality of fields in step (c).

28. A method as claimed in claim 9, wherein the matching of the predetermined titles is by keyword spotting.

29. A method as claimed in claim 9, wherein the matching of the predetermined titles is by using the automatic speech recognition results and searching for the predetermined titles.

30. A computer useable medium comprising a computer program code that is configured to cause a processor to execute one or more functions as claimed in claim 1.