A Non-Visual Photo Collection Browser based on Automatically Generated Text Descriptions

Frode Eika Sandnes
Faculty of Engineering, Oslo University College
N-0130 Oslo, Norway
e-mail: frodes@hio.no

Abstract—**This study presents a textual photo collection browser that automatically and quickly analyses large personal photo collections and produces textual reports that can be accessed by blind users using either text-to-speech or Braille output devices. The textual photo browser exploits recent advances in image collection analysis and the strategy does not rely on manual image tagging. The reports produced by the textual image browser gives the user a gist about where, when and what the photographer was doing in the form of a story. Although yet crude, the strategy can give blind users a valuable overview about the contents of large image collections and individual images which otherwise are totally inaccessible without vision.**

Keywords-image tagging; photo browsing; visual imapirment;textual image descriptions

I. INTRODUCTION

Low cost and high quality camera equipment combined with inexpensive and large memory have resulted in an explosion in the size and number of personal image collections. Most users enjoy taking photographs. The management and retrieval in such large collection is an emerging challenge [5] – even for users with perfect vision as the sheer volume of images is overwhelming. Manual tagging of individual images is tedious, laboureous and seems to be a lost battle, although some researchers are working on creative and playful ways, or games, for tagging images [3, 4]. In fact, the image tagging problem is well known in the universal design and accessibility community as the problem of missing textual, or alt-text, descriptions of images on web pages is large despite influential initiatives such as WCAG.

Images are highly visual objects that obviously are not accessible to blind individuals. From a blind users' perspective an image collection is simply a hierarchy of folders with numbered filenames. These filenames carry no useful information beyond indicating the sequence in which the photographs were taken. Date and time at which the images were taken are easily available, but carry no significant meaning to users without additional information or a context. Perhaps the key valuable clue in such collections may be the folder names as many photographers occasionally empty their cameras on their computers and create a new folder with some approximate short description of the context or setting.

Blind users may be interested in learning about the contents of image collections for a number of reasons. The prime reason is that images are social objects that we like to share, and blind users may not like to be completely excluded from the activity, yet they may want independence and not having to rely on others explaining the contents of images. Second, a blind user who is a part of a family with non-blind family members may want to know the contents of the images stored on the family computer.

Other multimedia formats such as audio and video are more accessible to blind users because of the sound. Moreover, the most common application, music collections are also often coded with text such that each audio file has a descriptive name and additional meta information coded such as the artist, album and song. This allows for non-visual browsing even without having to listen to the audio itself.

Recent advances in image analysis can also benefit non-visual interpretation of images. Google recently introduced face recognition in their Picasa photo browser. Clearly, information about which people are present on what photographs could be of interest to blind users. Knowing who is present in each picture can help establish a better understanding of the image collection structure. Moreover, if the blind user is familiar with a particular person the image will have a totally new meaning to the user. Moreover, landmark recognition in images is an ambitions, but promising, approach that may prove useful [14]. Blind users may find it interesting to know if an image contains a well known landmark – especially if the user is familiar with the landmark, has read descriptions about it, etc. Again, the image may become more meaningful to the user. Moreover, landmarks are associated with locations and locations are interesting when establishing the context and storyline for a sequence of photographs. Faces and landmarks are important as they make up a majority of the contents in images. However, both face recognition and landmark recognition are computationally expensive and not easily available. Moreover, landmark recognition requires a large database of landmarks to be useful.

In this study a similar approach is taken with very simple means based on the authors own recent results [8-11]. Image metadata are analyzed and the information are used to put an image collection into context as a comprehendible story suggesting where, when and what the photographer was doing.

II. Method

Most modern cameras store information about each photograph taken in special EXIF headers [7]. Such information includes the time and date the image was taken, optical characteristics such as shutter speed, aperture, iso film speed, the focal length of the lens and often special maker notes are provided with camera specific metadata. The EXIF format also support geo-coded images, that is, images taken together with a GPS measurement [1]. However, few cameras are equipped with GPS receivers.

It is the information stored in the EXIF headers that is the basis for the text based photo browser presented herein. There are two particular characteristics that are used for analysis, namely temporal features and exposure related features.

A. *Temporal analysis*

1) Events

Temporal features are well exploited by photo browsing applications. Applications such as Picasa primarily organize images chronologically into year and date. However, this is often done quite statically. In this study we propose to organize chronologically into events. An event is here defined as a physical event experienced by the photographers such as a two week holiday trip to Lisbon, Portugal, a three day conference trip to Oxford, England, etc. Photo collections often carry this signature, that is, long time intervals with no photographs, then a few days with bursts of photograph indicating an interesting event.

To identify the events we simply scanned all the images in chronological order. If the difference between the timestamps of two images exceeded three days they were marked as belonging to different events, while if separated by less than three days then they were considered as belonging to the same event.

This organization allows for abstract descriptions such as “During 2010 there were 3 events – one during March 6-7, one during July 7-10 and the last during October 23-25” to be automatically composed. Such descriptions can be further enhanced with additional information such as if it was a weekday or weekend. For example “During 2010 there were 3 events – one during the weekend of March 6-7, one during the week July 7-10 and the last during the weekend October 23-25”. Whether it is weekend or weekday can greatly affect our perception and interpretation as we usually have free time in the weekend for leisure, while activities during the week are often work related. Next, this linguistic information is helpful as few people are able to determine whether a date occurred during the week or during a weekday.

Further, linguistic additions could be added such as the season (Spring, Summer, Autumn, Winter, and special holidays such as Christmas, Easter, Chinese Lunar New Year, etc. For example, “During 2010 there were 3 events – one during the Spring weekend of March 6-7, one during the Summer week July 7-10 and the last during the Autumn weekend October 23-25”.

2) Local time

People travel more than ever before and we typically take many photographs when we are visiting a new place. Assuming that we approximately know the time-zone of the location for a particular event then we can deduce the local time of the photographs. How to determine the time-zone is outlined in section 2.2. Given the local time for an event we can provide the following textual summarizations “During the first day of the Spring weekend 123 pictures were taken before noon, and 14 in the afternoon. During the second day 23 pictures were taken in the afternoon and 45 in the evening”. This information tells us that the first day was probably spent outdoors in the morning and that there was some party the second day in the evening. One could soften the statistics and give a more vague human-like description to ease comprehension and overview. For example “During the first day of the Spring weekend most pictures were taken before noon, and a few in the afternoon. During the second day a few pictures were taken in the afternoon and most in the evening”.

B. *Exposure analysis*

Exposure values stored in image files provides valuable information about the lighting conditions when the photograph was taken. Although two images may appear to be subject to the same lighting condition this may not be the case. Cameras typically adjust the shutter speed, aperture and the film speed to adjust to various lighting conditions. A measure known as exposure value has been proposed as a standard for indicating the lighting conditions of a scene and it is easily computed using the shutter speed, aperture and film speed [2, 6]. The exposure value falls in the range of 0 (total darkness) to 16 (bright sunshine) or above.

1) Outdoor/indoor

The exposure value is a very simple way of determining if an image is captured indoors or outdoors as generally an image with an exposure value of less than 10 is indoors or at night and an exposure value of 10 or more usually have to be an outside photograph taken during daytime. Outdoor classification can also be performed using just image contents, but the contents based approach is more computationally expensive and complex [12, 13].

If the exposure value is below 10 and it is during the morning of afternoon, it is a good chance that the photograph is taken indoors. Whether images taken at night are outdoors or indoors is harder to determine.

Given the capability to determine if images are taken outdoor or indoors one can create summarizing statistics as in the following example: “During first day of the Spring weekend 98 outdoor pictures and 25 indoor pictures were taken before noon.”. Alternatively, a vaguer summary can be provided, namely “During the first day of the spring weekend most pictures were taken outdoors before noon.”

Nest, the exposure value can also be used to give a rough estimate of the weather. If the exposure value is high, that is above 12 then it is likely that it is sunny and if it is 12 or below it is likely that it is cloudy. However, these factors depends on the time of day, as it is brighter in the middle of the day than in the morning or in the afternoon, and whether the scene is in the shadows or not.

2) Geographical location

It has been shown in [11] that it is possible to derive the approximate geographical coordinates for a set of images based on the exposure values for the images in that event. In simple terms this is done by fitting a sinusoidal curve with a 24 hour period on top of 24 hour windows of image data such that one is able to estimate the approximate midday and sunrise and sunset times (interested readers are referred to [11] for a detailed description of the approach). Given these estimates it is possible to compute the approximate latitude and longitude for the event. Experiments showed that a latitudinal accuracy of approximately 30 degrees and a longitudinal accuracy of approximately 15 can be achieved. This is at least enough to give a linguistic estimate of the continent where the images were taken, and in most cases it is also capable of determining the correct hemisphere. Research is underway to significantly increase this accuracy down to a couple of degrees [8]. If one is so fortunate to have geo-coded images then this information can be read directly from the EXIF information with high accuracy.

Given such estimates it is possible to provide descriptions as provided in the following example: "The spring weekend of 2010 was probably photographed in North America (34 degrees North, 116 degrees West)". This information makes the report much more interesting to the user.

C. Focal length

Many digital cameras come equipped with zoom lenses allowing the focal length to be adjusted. This information also usually encoded into the EXIF headers. Typically, we use a small focal length, that is, a wider angle lens, to capture large objects in the view or if the objects are very close. Long focal lengths represent zoom magnification and are usually used to capture very small objects, or objects far away. Unfortunately, few cameras store the autofocus distance measurements as also these would be immensely useful.

Focal length information can also be summarized in a useful manner. For example, "During the first day of the Summer week 34 of the images had a long focal length (55mm or more) and 22 images had a short focal length (7.5 mm or smaller)" or in a more abstract form "During the first day of the summer week most of the images were taken with zoom."

D. Manual information

People are unlikely to tag individual images in their collections. However, many users tend to organize their images in folders with meaningful names "USA conf 2010". These manually added textual descriptions can be added to the textual reports. For example, "During 2010 there were 3 events – one during the weekend of March 6-7 labeled 'USA conf 2010', one during the week July 7-10 labeled 'Australia conf' and the last during the weekend October 23-25 labeled 'Mountain trip.'"

III. Experimental evaluations

The textual image collection browser was run on a subset of the author's personal image collection including 3,046 unique images taken over a period of four years at various locations around the world. The textual image collection browser was implemented in java and run on a Dell personal computer with an AMD Athlon Dual core processor and 4 Gb RAM running Windows Vista Personal edition. Drew Noakes' freely available (EXIF) metadata-extractor library (available at http://drewnoakes.com/) was used for extracting EXIF information from the images. It took about 3 minutes to generate the image collection report. The Java code was not optimized and significant speedups could be achieved with simple tuning.

Figure 1 shows a short extract of the output produced by the textual image collection browser.

```
This collection comprises 23 events.

The 6-day event entitled Asia-Taiwan-Tainan-
iceer  contains 615 images. It occurred during
the winter week, starting February 28, 2005 in
AUSTRALIA AND SOURTHEAST PACIFIC (77 degrees
Sourth, 154 degrees East
On the first day (Monday) 120 images were taken.
A majority of these photographs were taken
indoors during a mostly cloudy day
On the second day (Tuesday) 80 images were
taken. A majority of these photographs were
taken indoors during a mostly cloudy day
On the third day (Wednesday) 88 images were
taken.
Most of the photographs were taken at night
On the fourth day (Thursday) 114 images were
taken. A majority of these photographs were
taken indoors during a mostly cloudy day
On the fifth day (Friday) 155 images were taken.
A majority of these photographs were taken
indoors during a mostly sunny day
On the sixth day (Saturday) 53 images were
taken. A majority of these photographs were
taken indoors during a mostly sunny day

The 9-day event entitled Americas-USA-
SuthernIllonois contains 768 images. It
occurred during the summer week, starting June
23, 2005 in NORTH AND SOUTH AMERICA (61 degrees
North, 101 degrees West
On the first day (Thursday) 37 images were
taken. A majority of these photographs were
taken indoors during a mostly sunny day
On the second day (Friday) 67 images were taken.
A majority of these photographs were taken
indoors during a mostly sunny day
On the third day (Saturday) 91 images were
taken. A majority of these photographs were
taken indoors during a mostly cloudy day
On the fourth day (Sunday) 96 images were taken.
A majority of these photographs were taken
indoors during a mostly sunny day
On the fifth day (Monday) 60 images were taken.
A majority of these photographs were taken
indoors during a mostly sunny day
On the sixth day (Tuesday) 91 images were taken.
```

Figure 1. An extract of automatically generated textual image collection descriptions.

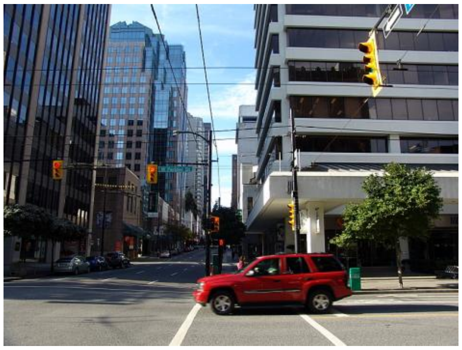

Image DSC03458.JPG, showing a cloudy outside motive, is taken in the early afternoon (13:17) Sunday October 14 during the autumn of 2007. The image occurs on the fifth day of a 6 day event entitled "Americas-Canada-Victoria " that takes place in Americas (80 degrees Sourth, 112 degrees West). This image is automatically annotated.

Manual note: the image is taken in the morning in the city center of Vancouver, Canada. The photograph shows a car driving through an intersection.

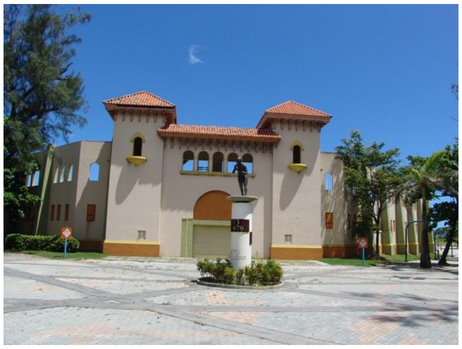

Image DSC06932.JPG, showing a sunny outside motive, is taken at noon (12:07) Saturday July 22 during the summer of 2006. The image occurs on the second day of a 8 day event entitled "Americas-PuertoRico-SanJuan " that takes place in Americas (67 degrees North, 75 degrees West). This image is automatically annotated.

Manual note: the image is taken at noon in the Old Town of San Juan, Puerto Rico showing a gate.

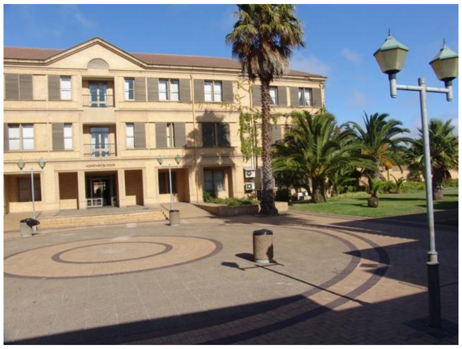

Image DSC08919.JPG, showing a sunny outside motive, is taken in the late morning (11:33) Friday February 27 during the winter of 2009. The image occurs on the fifth day of a 6 day event entitled "Africa-SouthAfrica-CapeTown " that takes place in Europa/Africa (78 degrees Sourth, 21 degrees East). This image is automatically annotated.

Manual note: The image is taken in the morning in Cape Town showing one of the administration buildings in one of the campuses of Cape Peninsula University of Technology.

Figure 2. Automatically tagged example images (arbitrarily selected)..

The extract shows that the data is presented in a semi hierarchical manner. First a statement regarding the entire collection is provided. Next, an overview of each event is provided followed by details on a day-based level of detail.

This extract represents a relatively large collection of images and the listing is relatively verbose and repetitive. A more compact form could be used, but then the data would be more cryptic. The linguistic representations are easy to interpret. Moreover, it should be possible to present this data in a semantically structured way such that users with screen readers could jump directly to the desired items. It is also possible that one could have made the day descriptions optional and instead present an overall trend for the event regarding the time of the day the photographs are taken and whether this is outside or inside and the overall weather.

The strategy can also produce a description for individual photograph. General information is then collected from the event as a whole and added to the individual information about the image. The examples in Figure 2 illustrate this.

The images in Figure 2 were chosen arbitrarily and the results show that the descriptions are more or less correct. The local time estimate in the first image is slightly incorrect as it is claimed that this image is taken in the early afternoon, when it was actually taken a few hours before lunch. Moreover, the weather predictor tags this as a “cloudy outside motive” while this day had clear skies. The reason is that this image is dark as the motive depicts a street shadowed by tall buildings. The second and third images are quite correctly explained. However, the simple linguistic image classifier could with simple means have resolved that this is a motive taken in Africa by eliminating Europe due to the image occurring on the southern hemisphere.

Note that the three images are taken in three totally different time-zones and the example illustrates the important of adjusting the presentation of time into local time as humans more naturally relate to local time.

A. *Implications*

Besides allowing non-visual insight into the contents of large image collections the strategy could also be used to help improve the accessibility of web pages. First, the strategy could be run on an image collection and each individual image could be labeled with the information extracted for its respective event and day and the information written back to each of the image files into the maker note of the EXIF header. Here a unique label such as for example “auto-alt” could be used to identify the automatic image tag. Web publishing tools could then automatically extract the automatically generated description and insert this into the alt-attribute of the image tags on the web pages. This would not be as good as a manual alt-text, but it would be more informative than no alternative image text at all.

IV. Limitations and future work

The current strategy assumes that the camera has a correctly set clock where its relative position to UTC is known. Moreover, it is assumed that the camera stores exposure attributes as described herein. A study of a large set of current compact and digital SLR cameras revealed that they all store this information – this includes modern mobile phones

One ultimate danger of automatic analysis is that there is a false positive rate, that is, the algorithms occasionally produce erroneous results, emphasizes insignificant details and overlook details that would be perceived to be important to a non-blind user. However, also humans err and occasionally miss important details, or misinterpret when visually inspecting visual objects.

The current approach is limited in terms of analysis and the information extracted. This is because only image-meta information was considered. Future work should also consider the image contents and incorporate some of the existing image analysis research results within face recognition, object recognition, landmark recognition, etc, to provide more rich descriptions of the images. More work should also be put into extracting high level summaries from the low level image analysis results.

V. Conclusions

This study has presented a textual image browser. The browser summarizes the contents of image collections in the form of stories comprising chronologically ordered events suggesting, when the event to place, where the event to place and what was done. Although still very limited, the approach extracts useful information about highly visual image object into a textual domain accessible by blind users. The current approach does not actually look at image contents and are therefore computationally highly effective. The current approach also considers individual images and their relationship to their respective events. This could for instance be used to automatically populate alt attributes of web-page images. However, further work is needed to derive information about the contents of individual photographs. Another important future research question is what information blind users find useful in context of image collections.

References

[1] Ahern, S., Naaman, M., Nair, R., and Hui-I Yang, J., "World explorer: visualizing aggregate data from unstructured text in geo-referenced collections," in the proceedings of 7th ACM/IEEE-CS joint conference on Digital libraries, 2007, pp. 1-10.

[2] ANSI, "[ANSI PH2.7-1986. American National Standard for Photography - Photographic Exposure Guide," American National Standards Institute, New York 1986.

[3] Diakopoulos, N. and Chiu, P., "Photoplay: A collocated collaborative photo tagging game on a horizontal display.," in the proceedings of UIST '07, 2007, pp. 53-54.

[4] Golder, S. and Huberman, B. A., "The structure of collaborative tagging systems," Journal of Information Sciences, 32, 2 (2006), pp. 198-208.

[5] Jang, C.-J., Lee, J.-Y., Lee, J.-W., and Cho, H.-G., "Smart Management System for Digital Photographs using Temporal and Spatial Features with EXIF metadata," in the proceedings of 2nd International Conference on Digital Information Management, 2007, pp. 110-115.

[6] Ray, S. F., "Camera Exposure Determination," in The Manual of Photography: Photographic and Digital Imaging, R. E. Jacobson, S. F. Ray, G. G. Atteridge, and N. R. Axford, Eds.: Focal Press, 2000.

[7] Romero, N. L., Chornet, V. V. G. C. G., Cobos, J. S., Carot, A. A. S. C., Centellas, F. C., and Mendez, M. C., "Recovery of descriptive information in images from digital libraries by means of EXIF metadata," Library Hi Tech, 26, 2 (2008), pp. 302-315.

[8] Sandnes, F. E., "Determining the Geographical Location of Image Scenes based on Object Shadow Lengths," Journal of Signal Processing Systems for Signal, Image, and Video Technology, (2010) in press.

[9] Sandnes, F. E., "Sorting holiday photos without a GPS: What can we expect from contents-based geo-spatial image tagging?," Lecture Notes on Computer Science, 5879, (2009), pp. 256-267.

[10] Sandnes, F. E., "Unsupervised and Fast Continent Classification of Digital Image Collections using Time," in the proceedings of ICSSE 2010, 2010, pp. 516-520.

[11] Sandnes, F. E., "Where was that photo taken? Deriving geographical information from image collections based on temporal exposure attributes," Multimedia Systems, 16, 4-5 (2010), pp. 309-318.

[12] Serrano, N., Savakis, A., and Luo, A., "A computationally efficient approach to indoor/outdoor scene classification," in the proceedings of 16th International Conference on Pattern Recognition, 2002, pp. 146-149.

[13] Szummer, M. and Picard, R. W., "Indoor-outdoor image classification," in the proceedings of IEEE International Workshop on Content-Based Access of Image and Video Database, 1998, pp. 42-51.

[14] Zheng, Y.-T., Ming, Z., Yang, S., Adam, H., Buddemeier, U., Bissacco, A., Brucher, F., Chua, T.-S., and Neven, H., "Tour the world: Building a web-scale landmark recognition engine," in the proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2009), 2009, pp. 1085 - 1092.