2010 Second International Conferences on Advances in Multimedia

Using Context to Understand User Intentions in Image Retrieval

Christian Hartvedt
Department of Information Science and Media Studies
University of Bergen
Bergen, Norway
Christian.Hartvedt@infomedia.uib.no

***Abstract*—This paper discusses how combining existing techniques in a new way may help improve the understanding of user intentions in image retrieval. This is an especially challenging task in situations where the information need of a user is ambiguous, or if the user is unable to express it. The suggestion made here is that combining two different approaches to image retrieval with the utilization of context information and user interaction will help alleviate this problem. This paper describes how an approach that uses visual data, textual data in the form of context information, and user interaction in the image retrieval process may be developed and evaluated.**

Keywords-image retrieval; context-focus; user intention

I. INTRODUCTION

The number of searchable images available is now in the hundreds of millions, and the amount is continuously growing [1]. Two prominent sources contributing to this trend are personal albums (e.g., Flickr, photobucket, piczo, Picasa web album and photo.net) and general-purpose image collections (e.g., Google Images and Yahoo! Images). Users of most large image collections often face the problem of how to retrieve relevant images from the collection. One of the main obstacles is that the system does not necessarily understand the user's intentions behind a search.

To show how use of a context-focused approach combined with interaction can be useful to alleviate this problem in the image retrieval process, some of the challenges associated with the traditional image retrieval approaches are briefly discussed.

The two most common approaches to the process of retrieving images are Text Based Image Retrieval (TBIR) and Content Based Image Retrieval (CBIR). In TBIR, a search is performed by matching search terms submitted by the user against text-based annotations that represent the images in the collection. Two potential problems with annotations are that they may cover only one or a small subset of the possible semantic interpretations of the image content, and that they are commonly biased because of human perception and subjectivity [2, 3]. In addition, it is also a prerequisite in TBIR that users are able to formulate their information needs using text.

In CBIR, low-level features (e.g., color, shape or texture) from some visual imagery submitted to the system are matched against the features of images in the collection. The CBIR approach may be convenient in situations where textual descriptions are hard or impossible to create, but CBIR lacks the support for image retrieval based on high-level semantic concepts.

One important challenge users meet when using most systems designed for TBIR or CBIR respectively, is that the problem referred to as the semantic gap [4] largely remains. This gap represents the mismatch between semantic user requests and the capabilities of the image retrieval systems.

A central task of image retrieval systems is to aid users in the retrieval of images in an efficient manner using the depicted image content or keywords as a starting point. However, in order to be able to retrieve relevant images it is necessary that the system has a good understanding of the user's intentions behind the search. From this, the discussion on if and to what extent current techniques for creating image representations actually are suitable is important.

It is the belief of this author that neither TBIR nor CBIR alone are optimally suited to support context focused image retrieval, primarily due to challenges associated with the semantic gap. The main reason for this is that it is often difficult to understand the user's intentions from just a few keywords or a query-image alone. From this, an important question is if TBIR combined with CBIR, and the use of context information combined with user interaction may help alleviate this problem. Two central questions with regard to the evaluation of the proposed approach are to what extent it improves the quality of image retrieval result sets, and how useful it is perceived to be.

In this paper, Section II presents some work where TBIR and CBIR have been used together in various ways, and briefly discusses some problems and open issues. Section III presents a brief scenario illustrating the problem. Section IV briefly discusses how the use of TBIR and CBIR in combination, and the use of context information may contribute to a better understanding of user's intentions in image retrieval. Section V concludes the paper by giving an outline of the work and road ahead.

II. RELATED RESEARCH

In the field of image retrieval, the general usefulness of combining TBIR and CBIR has been recognized, e.g., in [5-8]. In the medical domain, an early proposal for combining image content with associated text for retrieval purposes was I^{2}Cnet, where annotations describing the images supported users in content-based queries in a health care network [9]. Two more recent approaches, the first utilizing a standard cross-language information retrieval system combined with an image retrieval system [10] and the second using the *medgift easyIR* system [7], made an attempt to enrich content-based image retrieval with text in the form of multi-lingual search terms. The approach presented in [10] combined two result sets provided by two autonomous search systems, while the approach presented in [7] generated a text-based query expansion from the annotations accompanying the top-three results from an

978-0-7695-4068-9/10 $26.00 © 2010 IEEE
DOI 10.1109/MMEDIA.2010.35

initial image query. Neither of these two approaches provided users with the possibility to explicitly specify the text to be used in the query process.

In the non-medical domain, several approaches connecting visual and textual characteristics for retrieval purposes exist. One such approach, developed for retrieving images from the World Wide Web, uses a text-based query specified by the user as a starting point for the image retrieval process. Relying on the notion that images placed near text in an HTML document are related to the image, the retrieval system retrieve all images in near proximity of words specified in the initial query. Then the user performs several user-feedback cycles in order to refine the result set [11]. A slightly different approach is found in [8]. Here, the use of keywords in a semantic network, supported by online learning through a feedback algorithm and the use of a term similarity matrix, supports the content-based image retrieval process in order to draw upon the strength of both approaches. Users specify some keywords that accompany a seed image, and these keywords support the retrieval of images annotated with corresponding keywords. The former approach relies on automatic extraction of text found in close proximity of images in a document while the latter approach relies more on manual annotation of image content. [12], presents an approach somewhat similar to that of [11]. The main difference in this new approach is that the author presents a method that combine text and images into the same semantic space using Latent Semantic Indexing (LSI) and Singular Value Decomposition [12]. The combined use of text and image content is also proposed in [13] and [14], which suggest that image descriptions can be created automatically by combining low-level image features and high-level semantic information. Another approach following this line of thought is to generate visual keywords from the analysis of the low-level image content based on learning and similarity matching [15, 16].

The roles that information needs, context information, and the users themselves may play in the image retrieval process have received little attention in many image retrieval systems. A possible reason for this may be that much of the development within the field of image retrieval has been system-driven rather than user-driven [17]. This has resulted in a general lack of focus on the system users [18]. These trends are also visible in most of the approaches to image retrieval described above in that they do not deal explicitly with many important aspects pertaining to users and user interaction. In addition, most of the previous approaches seem to assume that users have a clear information need that they are able to express, but this is not always the case. Furthermore, context focused image retrieval introduces some additional challenges. Amongst these is the problem of how the system is to understand user intentions when they are using both image content and text as query terms.

III. Scenario and Problem Formulation

Alf comes from Bergen in Norway. He is very interested in the various attractions that can be found in Bergen and spends much time exploring museums, historic buildings, artwork and statues. He also has a keen interest in sailing. One of his favorite tall ships is the "Statsraad Lehmkuhl", a Bark with three masts. During the *Tall Ships' Races* [19] in 2008, Alf was a spectator as the Statsraad Lehmkuhl and the other participating ships sailed away. He captured an image of a beautiful Bark from Germany with three masts carrying green sails. However, as the ship used its motor at the time, all the sails were down. Now, two years later, Alf is very interested to learn more about this vessel and find images of the ship with the sails up. However, as the name of the ship is not visible in the image he took, he does not know what keywords to use in his search. Alf remembers that there were other photographers present when he took the image, and decides to see if any of them have uploaded their images to Flickr. He performs a search using the keywords *"German tall ship sailing"*. The search does not result in anything useful, but Alf remembers that Svein, a fellow photo enthusiast, was there taking pictures of the ships. Alf looks up Svein's profile on Flickr. Svein does indeed have a set of images covering the race and has also taken an image of the green ship, but unfortunately he has named the image "_MG_4562". None of the annotations associated with the image reveals the name of the vessel, so the information available is not very helpful.

The scenario above describes a situation in where an information need cannot be articulated adequately as the name of the ship is unknown. The name of the ship is "Alexander von Humboldt", and if Alf had known this, he would most certainly have found some of the several beautiful images of the ship with its big green sails up on Flickr, and also the abundance of additional information about the vessel on available on the Web.

However, as Alf had an image of the desired ship, it is believed by this author that this also could have been useful as a query term. Here, the low-level features in the image and CBIR could have been used to identify the vessel, while TBIR using the words "German tall ship sailing" would have helped to indicate the actual information need.

IV. Foundation of the Proposed Approach

The ultimate goal of the work outlined in this paper is to improve the quality of image retrieval, especially in situations where the information need is ambiguous and/or difficult to express. It is believed that such an improvement may be achieved by improving the support given to users in the image retrieval process. A presupposition is that such an approach also may contribute to a better understanding of user intentions by the system. Two important assumptions underlying the approach suggested here is that

- Use of TBIR and CBIR in combination will enable users to specify their information need in a simpler, yet more complete manner as they may use both visual data and text.
- Combining two distinct communication channels with active use of context information and user interaction facilitates an easier query submission, possibilities for giving feedback, and provides the system with a better understanding of user intentions.

A. Combining Image Retrieval Approaches

A central challenge in the approach to image retrieval proposed here is how to create good enough image representations to be used by computers to support humans in finding relevant images. One contributing factor to this challenge is that image retrieval may be seen as having three levels [20]:

- Level one, is image retrieval using low-level image features such as colour, texture, shape or the spatial location of image elements.
- Level two, is image retrieval using derived features involving some degree of logical inference about the identity of the objects depicted in the image.
- Level three, is image retrieval using abstract attributes, involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted.

On level one, the descriptors must at least be able to identify and describe the (sequences of) symbols occurring in the image. This is commonly done by automatic extraction of image features. On level two, the goal is to identify and describe what is in the image. However, a fully automatic extraction of such high-level content features has proven very difficult [6, 21-22]. On level three, the goal is to have descriptors describing what the image is about, i.e., its meaning. This task is a far more difficult task compared to level two in that it requires abstraction from level two using high-level concepts. Creating image descriptions on level three is commonly done by humans.

TBIR is mainly aimed at image retrieval on levels two and three. Hence, if used by itself, TBIR may return images reflecting the semantic information need but that does not necessarily resemble the desired outcome visibly. On the other hand, if using a CBIR approach, this is image retrieval on level one. Here, the user may get images that are visually similar but that have little semantic similarity [8]. However, by combining TBIR and CBIR, users may utilize keywords that reflect semantic aspects of their information need combined with low-level features representing what the desired outcome should look like.

B. Utilizing Context in Image Retrieval

There exists a wide variety of definitions of the term context in the literature, e.g., [12, 23-27]. In addition, some central definitions from the field of context-aware computing are discussed in some detail in [28], where Dey also presents a specific definition that can be used prescriptively:

> *"Context is any information that can be used to characterize the situation of an entity. Here, an entity may be a person, a place, or an object that is considered relevant to the interaction between a user and an application, including the user and applications themselves*" [28]

A different definition of context is presented in [29]. Although the authors generally agree with the definition given in [28], their approach differs in an important way. In [29], the authors focus more on the users and applications themselves, and especially on what information must be available to both in order for *communication* between them to succeed.

From this brief presentation of some definitions of context, the term is here seen as being important with regard to two equally important aspects:

- Context should provide support to the process of understanding the situation of entities important to the interaction in which a user and a system participates, and
- Context should help facilitate communication between user and system.

In [28], *location*, *identity*, *time* and *activities* are presented as the primary context of an entity, and from this primary context, various forms of secondary context information may be derived. In [30], this line of thought is extended in an approach using spatial, temporal and social context to generate contextual annotations. If taking this latter approach to describing image content, then annotations specifying the identity of the photographer, the time when the image was captured, the date, GPS coordinates and so forth, can be derived automatically and assigned to the image in order to describe some aspects of the image. This kind of annotations corresponds mainly to level two described in [20], and are as such unable to provide information on what is depicted in the image.

With regard to describing image content on level three, it is believed by this author that annotations of this kind can be generated and collected through interaction and communication between the users and the system. As such, context information may be available as automatically generated annotations associated with the images, and thus already stored in the system. On the other hand, users may also contribute with valuable context information, and it is important that this information can be made available to the system. It is believed by this author that context information covering all three levels is necessary in order to understand user intentions in context-focused image retrieval, and it is vital to extend the support for collecting, maintaining and use of context information in image retrieval systems.

V. CONCLUSION AND FUTURE WORK

In this paper, it has been discussed one way in which existing techniques may be used to improve the quality of image retrieval and the understanding of user intentions. An approach that combines two different approaches to image retrieval, together with active use of context information and interaction has been proposed.

An important aspect of the research work outlined here is to design and evaluate a system that, in addition to combine the use of TBIR with CBIR, uses context information and user interaction in the process. An underlying hypothesis is that such an approach will make it easier for an image retrieval system to understand the user's intentions behind a search. Furthermore, it is also believed that such an approach may also help generate context annotations, as having users interacting with the system in the image retrieval process may provide valuable information on various aspects of image content and image context. This kind of information may be of help to other users.

To test the hypothesis, the first step is to develop a prototype that utilizes both TBIR and CBIR in the retrieval process, and in addition supports user interaction. When performing a search using the prototype, a TBIR search and a CBIR search are run in parallel. Relevance scores from the two queries are collected and combined into one score, which is used to rank results. As users submit both text and image data, the system may use these data to infer the information need and present users with available information corresponding to the search criterion. Users may then choose from the available information to refine the result set further. An underlying notion is that interaction and communication where system and user both contribute with their best effort may help in this process.

The evaluation of the approach will be done through experiments with human participants. The experiment will include concrete image retrieval tasks that the participants use the prototype to solve. Aspects pertaining to system performance and the quality of results will be assessed using quantitative measurements in terms of precision measures, while aspects pertaining to usability will be investigated using a questionnaire and an interview.

References

[1] R. Datta, D. Joshi, J. Li, and J. Z. Wang, “Image Retrieval: Ideas, Influences, and Trends of the New Age,” ACM Computing Surveys, vol. 40, 2008, pp. 1-60, doi:http://doi.acm.org/10.1145/1348246.1348248.

[2] Y. Rui, T. S. Huang, and S. Mehrotra, “Relevance feedback techniques in interactive content-based image retrieval,” in SPIE/IS&T Conf. on Storage and Retrieval for Image and Video Databases San Jose, CA, pp. 25--36 1998, doi:10.1117/12.298455.

[3] O. Marques and B. Furht, Content-Based Image and Video Retrieval, 1st ed.: Kluwer Academic Publishers 2002.

[4] W. Arnold, M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, “Content-Based Image Retrieval at the End of the Early Years,” IEEE Transactions on Pattern Analysis and Machine Intelligence vol. 22, 2000, pp. 1349–1380 doi:http://doi.ieeecomputersociety.org/10.1109/34.895972

[5] G. Kowalski and M. T. Maybury, “Information Storage and Retrieval Systems - Theory and Implementation,” Kluwer Academic Publ., 2000.

[6] G. Lu, Multimedia database management systems. Boston: Artech House, 1999.

[7] H. Müller, P. Ruch, and A. Geissbuhler, “Enriching content-based image retrieval with multi-lingual search terms,” Swiss Medical Informatics, vol. 54, 2005, pp. 6-11.

[8] X. S. Zhou and T. S. Huang, “Unifying keywords and visual contents in image retrieval,” Ieee Multimedia, vol. 9, Apr-Jun 2002, pp. 23-32, doi:10.1109/93.998050.

[9] S. C. Orphanoudakis, C. E. Chronaki and D. Vamvaka, “I Cnet: Content--based similarity search in geographically distributed repositories of medical images,” Computerized Medical Imaging and Graphics, 20(4), pp. 193-207, 1996.

[10] G. J. F. Jones, D. Groves, A. Khasin, A. Lam-Adesina, B. Mellebeek, and A. Way. “Dublin City University at CLEF 2004: Experiments with the ImageCLEF St Andrew's Collection,” Proc. CLEF 2004: Workshop on Cross-Language Information Retrieval and Evaluation, 2004.

[11] S. Sclaroff, M. L. Cascia, and S. Sethi. “Unifying textual and visual cues for content-based image retrieval on the world wide web,” CVIU, 75(1-2), 1999, pp. 86 - 98.

[12] T. Westerveld, Image Retrieval: “Content versus Context,” in RIAO 2000 Conference Proceedings, Paris, 2000, pp. 276–284.

[13] W. I. Grosky, R. Agrawal, and F. Fotouhi. “Mind the Gaps - Finding Appropriate Dimensional Representation for Semantic Retrieval of Multimedia Assets,” In I. Kompatsiaris & P. Hobson (Eds.), Semantic Multimedia and Ontologies, pp. 229-252, London: Springer, 2008.

[14] R. Möller, and B. Neumann, B. “Ontology-Based Reasoning Techniques for Multimedia Interpretation and Retrieval,” In I. Kompatsiaris & P. Hobson (Eds.), Semantic Multimedia and Ontologies. London: Springer, 2008.

[15] L. Joo-Hvee, “Learnable visual keywords for image classification,” in Proc. ACM conference on Digital libraries Berkeley, California, United States: ACM, 1999.

[16] S. Dasiopoulou, C. Saathoff, P. Mylonas, Y. Avrithis, Y. Kompatsiaris, S. Staab, and M. G. Strinztis, “Introducing Context and Reasoning in Visual Content Analysis: An Ontology-Based Framework," in Semantic Multimedia and Ontologies, Y. Kompatsiaris, Ed. London: Springer, 2008, pp. 99-122.

[17] C. Jörgensen, Image Retrieval Theory and Research. Laham, Maryland: Scarecrow Press, Inc, 2003.

[18] M. S. Lew, N. Sebe, D. Chabane, and J. Ramesh, “Content-based Multimedia Information Retrieval: State of the Art and Challenges,” ACM Transactions on Multimedia Computing, Communications, and Applications, 2006, pp. 1-19, doi:10.1145/1126004.1126005.

[19] http://www.tallshipsracesbergen.no/

[20] J. P. Eakins and M. E. Graham,"Content Based Image Retrieval: A report to the JISC Technology Applications Program,” Inst. For Image Data Research, University of Northumbria., Newcastle 1999.

[21] A. Jaimes and S. F. Chang, “Concepts and Techniques for Indexing Visual Semantics" in Image Databases: Search and Retrieval of Digital Imagery, V. Castelli and L. D. Bergman, Eds. New York: John Wiley & Sons, Inc, 2002, pp. 497-565.

[22] Y.-G. Jiang, C.-W. Ngo, and J. Yang, “Towards optimal bag-offeatures for object categorization and semantic video retrieval,” in Proc. 6th ACM international conference on Image and video retrieval Amsterdam, The Netherlands: ACM, 2007.

[23] T. M. Strat, “Employing Contextual Information in Computer Vision,” in DARPA93, pp. 217-229 1993, doi:http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.57.4277

[24] P. Brézillon and J.-C. Pomerol, “Contextual Knowledge Sharing and Cooperation in Intelligent Assistant Systems,” Le Travail Humain, vol. 62, 1999, pp. 223-246.

[25] A. K. Dey, “Understanding and Using Context,” Personal and Ubiquitous Computing, vol. 5, 2001, pp. 4-7, doi:10.1007/s007790170019.

[26] P. Dourish, “What We Talk About When We Talk About Context,” Personal and Ubiquitous Computing, vol. 8, 2004, pp. 19-30, doi:10.1007/s00779-003-0253-8.

[27] A. Mani and H. Sundram, “Modeling user context with applications to media retrieval,” Multimedia Systems, vol. 12, 2007, pp. 339-353, doi:10.1007/s00530-006-0054-9.

[28] A. K. Dey, G. D. Abowd, P. J. Brown, N. Davies, M. Smith, and P. Steggles, “Towards a better understanding of context and contextawareness,” Handheld and Ubiquitous Computing, Proceedings, vol. 1707, 1999, pp. 304-307, doi:http://dx.doi.org/10.1007/3-540-48157-5.

[29] D. Elgesem and J. Nordbotten, “The Role of Context in Image Interpretation,” in Context-based Information Retrieval, CIR'07 at CONTEXT'07, Roskilde University, Denmark, 2007.

[30] D. Marc, K. Simon, G. Nathan, and S. Risto, “From context to content: leveraging context to infer media metadata,” in Proc. 12th annual ACM international conference on Multimedia, New York, NY, USA, 2004, doi:http://doi.acm.org/10.1145/1027527.1027572.