Framework for Improving Annotation-Based Image Retrieval Performance

Phani Kidambi, Mary Fendley, S. Narayanan
College of Engineering & Computer Science
Wright State University
3640 Colonel Glenn Highway
Dayton, OHIO, USA – 45435
{phani.kidambi, mary.fendley, s.narayanan}@wright.edu

ABSTRACT: *As the proliferation of available and useful images on the web grows, novel methods and effective techniques are needed to retrieve these images in an efficient manner. Currently major commercial search engines utilize a process known as Annotation Based Image Retrieval (ABIR) to execute search requests focused on image retrieval. The ABIR technique primarily relies on the textual information associated with an image to complete the search and retrieval process. Using the game of cricket as the domain, we describe a benchmarking study that evaluates the effectiveness of three popular search engines in executing image-based searches. Second, we present details of an empirical study aimed at quantifying the impact of inter-human variability of the annotations on the effectiveness of search engines. Both these efforts are aimed at better understanding the challenges with image search and retrieval methods that purely rely on ad hoc annotations provided by the humans. Finally, we propose a framework that utilizes generic templates to aid the human's cognitive capabilities to fill relevant for annotation needed in a specific domain in a more systematic way. The systematic annotation will not only reduce the mental task load on the human, but also would increase the precision and recall of a search engine.*

Keywords: Annotation Based Image Retrieval, Web image processing, Image search, Image retrieval performance

Received: 11 December 2010, Revised 27 January 2010, Accepted 3 February 2011

1. Introduction

Today's digital technologies allow an inexhaustible number of images to easily be uploaded to the web. As the sheer number of images increases, the user is faced with image overload, where the user has access to more images than can be viewed, with only a portion of them even being relevant [1]. Major search engines use a technique called Annotation Based Image Retrieval (ABIR) to perform queries focusing on image retrieval. Annotation is the encoding of visual stimuli as text and involves the professional judgment of an individual to interpret material and its content. The ABIR technique relies on text based information regarding the image in order to execute the search and retrieval process.

To complete a search, an ABIR-driven engine follows a set of standard steps [2]. Images are retrieved by evaluating the vector of word frequencies in their annotations and returning the images with the closest vectors. A relevancy ranking is calculated by evaluating the degree of the match between the search terms and the annotation of each individual image in terms of the order and separation of the words [3]. This means that, even though the user is searching for images, the images that are retrieved are actually determined by the *textual* annotation. This annotation usually consists of manually assigned keywords, captions, or any other text associated with the images.

While a significant body of research exists to evaluate the effectiveness of textual information retrieval processes [4, 5, 6, 7, 8], there has been very little focus on evaluating image retrieval on the Internet. The first part of this article describes a

benchmarking study which evaluates the effectiveness of three popular search engines in executing image based searches.

The domain selected to assess their performance is the game of cricket. The second phase of this study assesses the impact of variability in human annotation of web images on the image search. Finally, we propose a framework that aids the human to annotate an image in a more systematic way.

2. Background

2.1 Image Annotation

Image search and retrieval through ABIR draws its capabilities by taking advantage of natural language text to represent semantic content of images and the user's search needs [9]. Studies have shown that textual information is key to image retrieval using both video and photo image retrieval [10, 11].

To search and retrieve the images in a database (or the World Wide Web), the current commercial search engines provide some text descriptors to the images and retrieve them based on the text. Figure 1 presents an overview of the background processes in this area.

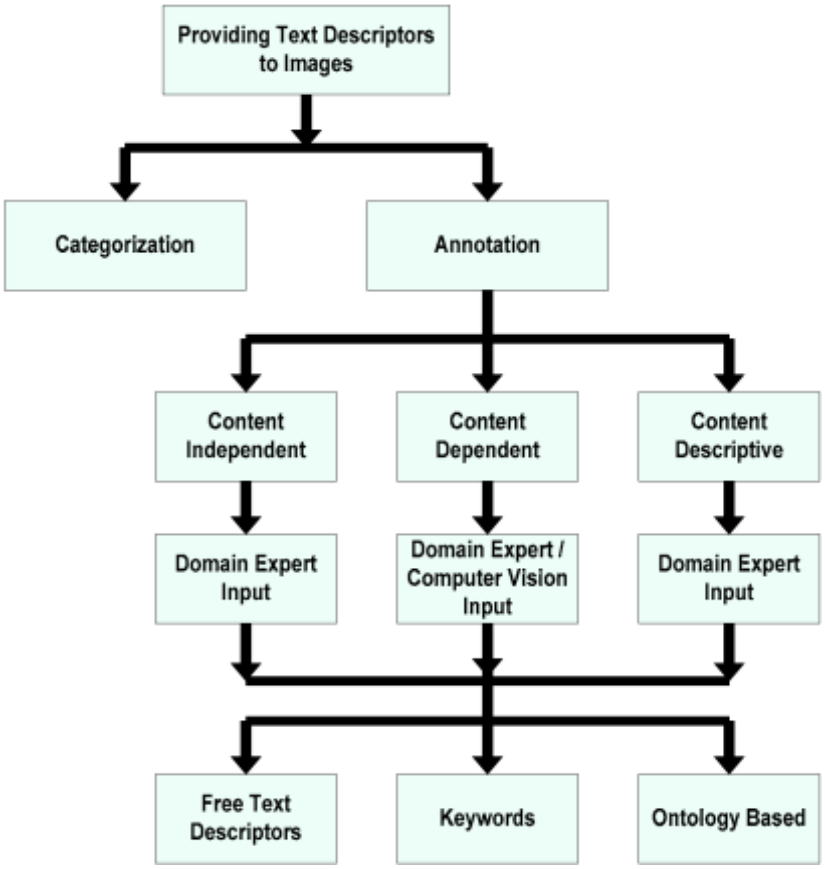

Figure 1. Overview of the background

Chen et al. [12] defined two methods of providing text to images: categorization & annotation. Categorization is the association of a predefined category of text to an image while annotation provides the image with detailed text descriptors. Bimbo [13] stated that three different types of information can be associated with an image that include

- Content-independent metadata – the metadata is related to the image but cannot be extracted with computer vision algorithms (example: date when the image was taken)
- Content-dependent metadata – the metadata can be extracted using content based image retrieval algorithms (example: color, texture & shape)

- Content-descriptive metadata – the semantics in the image which cannot be extracted using computer vision algorithms and need the expertise of a domain expert to annotate the image.

Annotation based on content-dependent metadata for an image can be generated using computer vision algorithms while for content-descriptive metadata, human annotation is required. The computer vision algorithms captures only one aspect of the image (color, texture, shape) and even a combination of these low level image features do not capture the semantic meaning (high level textual features) of the image. Liu et al. [14] coined the term 'semantic gap' to describe the gap between the low level image features and the high level textual features. The computer vision algorithms are still at an early stage, leaving the commercial search engines to focus on retrieving the images based on the text descriptors rather than the content of the image. The human can associate some text descriptor for an image in one of three ways: free flowing text, keywords from restricted vocabularies, and ontology classification [15]. Though manual annotation of images is labor intensive and time consuming, it still is the preferred way of annotating images [16]. To reduce the effort of the human, an innovative way known as the Extra Sensory Perception (ESP) game was developed to collect the text descriptors by Ahn et al. [17]. In this ESP game, two players are paired randomly, and are asked to generated keywords for a particular image in a given time interval. Whenever the keywords suggested by the two players match, the image is annotated with that keyword. Our work proposes a more methodical approach to reduce cognitive effort and increase precision and recall.

2.2 Image Retrieval

Evaluating the effectiveness of information retrieval is important but challenging. Most researchers adopt the definition of evaluation posited by Hernon et al [18] which states that evaluation is

> " the process of identifying and collecting data about specific servicesor activities, establishing criteria by which their success can be assessed,and determining both the quality of the service or activity and thedegree to which the service or activity accomplishes stated goals and objectives."

Researchers have defined several different taxonomies for evaluation processes of information retrieval. Meadow et al [19] classified information retrieval measures into two categories: evaluation of performance (descriptive of what happens during the use of the information retrieval system) and evaluation of outcome (descriptive of the results obtained). Hersh [20] also identifies two other evaluation categories: macro evaluation (investigates information retrieval system as a whole and its overall benefit) and micro evaluation (investigates different components of the system and their impact on the performance in a controlled setting). Lancaster et al. [21] went on to define three separate levels of evaluation. The first level evaluates the effectiveness of the system, the second level evaluates the cost effectiveness and the third level evaluates cost benefits of the system. Smith [22] proposed several measures for image retrieval evaluation including precision, recall, fallout and F-measure (F-measure = 2 (Precision * Recall) / (Precision + Recall)). Finally, Cooper [23] suggests Expected Search Length (ESL) as an alternative to recall and precision. ESL measures the number of unwanted documents the user can expect to examine before finding the desired number of relevant documents.

Though the research literature contains numerous studies on various ways of ascribing text to images by humans and computers, and the evaluation metrics for information and image retrieval, there is a dearth of literature benchmarking the performance of image search engines and investigating the human role in the annotation of images for search and retrieval engines.

This paper begins to fill this information gap, employing a systematic approach to evaluate search engines based on a number of independent factors including query types, number of images retrieved and the type of search engine. The research also conducts a systematic study to investigate the role of humans in the annotation of images. The remainder of this paper discusses the experiments that have been performed, the results, and their implications.

3. Benchmarking Commercial Image Search Engines

3.1 Design of Experiment

Methodology: Research studies [22, 24], have showed that the quality of images retrieved are typically a function of query formulation, type of search engine and the retrieval level. These independent factors and their levels are illustrated in the Ishikawa diagram in Figure 2.

In this experiment three search engines (Google, Yahoo, and MSN Live) are evaluated for varying query types and retrieval

levels. Details on the query types and query levels are provided below.

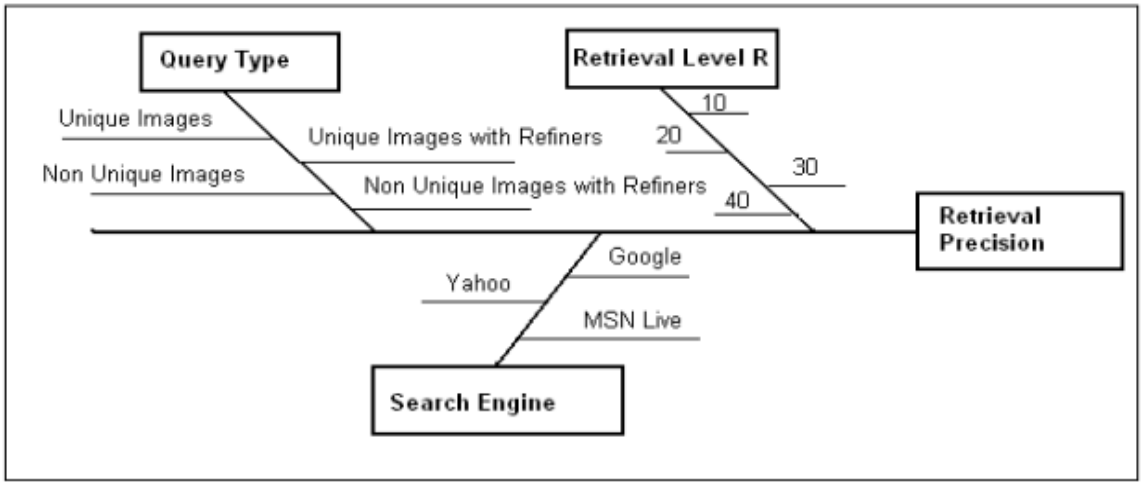

Figure 2. Ishikawa diagram of independent factors used in this study

Ensor & McGregor [25] summarized that user search requests for images fall into four different categories:

- Search for unique images – The property of uniqueness is a request for the visual representation of an entity where the desired entity (image) can be differentiated from every other occurrence of the same entity type. An example is – "find the image of Sachin Tendulkar".
- Search for unique images with refiners – Using refiners, the user can narrow down the results of the retrieved images from unique images to a certain degree of specificity. An example is – "find the image of Sachin Tendulkar in 2004".
- Search for non – unique images – The property of non – uniqueness is a request for the visual representation of an entity where the desired entity (image) cannot be differentiated from every other occurrence of the same entity type. An example is – "find the images of Indian cricketers".
- Search for non – unique images with identifiers – Using refiners, the user can narrow down the results of the retrieved images from unique images to a certain degree of specificity. An example is – "find images of Indians waving the Indian flag".

Search Engines: According to Nielsen's May 2009 ratings [26], Google's search engine accounts for 63.3% of the total searches on the internet, Yahoo's accounts for 17.3% and MSN Live accounts for 9.4%. These three search engines execute 90% of the total searches conducted on the internet. It is for this reason that they have been selected for comparison purposes in this study.

Experiment: To carry out the evaluation, a user-centered interpretative approach, based on the actual information-seeking behavior of real users [22] was employed. Since this research is focused on the domain specific evaluation of the system, a subject matter expert in the domain of the game of Cricket, was used to evaluate the existing search engines. Five queries for each query type are chosen. The queries consist of multi-word queries, related to the game of cricket, as shown in Table 1.

The queries associated with "unique images" are all internationally known cricket players from different playing eras. The queries associated with "unique images with refiners" are related to a cricket player involved in a context such as winning a world cup. The queries associated with "non-unique images" are internationally well known cricket teams and finally the queries associated with "non-unique images with refiners" are internationally known cricket teams involved in a context similar to winning a world cup.

Each query is run on each of the three search engines. The first forty images retrieved in each search run are evaluated for relevance by the subject matter expert using his knowledge base. Relevance is determined in a binary manner, either relevant or not relevant. In instances when the same image appears on different websites, these are evaluated as different images and each is evaluated for relevance. In instances where the same images appear in multiple places on the same website, the first

Query Types	**Queries**
Unique Images	MS Dhoni Vijay Bharadwaj Ricky Pointing Gary Sobers Abey Kuruvilla
Unique Images with refiners	Kapil Dev lifting World Cup Sreesanth + beamer + Pietersen Andy Flower + protest + black band Allan Donald + run out+ WC semifinal '99 Inzamam Ul Haq hitting a spectator + Canada
Non Unique Images	Indian Cricket Players Surrey Cricket Team Ashes (Eng vs Aus) Cricket Players Huddle Rajastan Royals + IPL
Non Unique Images with refiners	Victorious Indian Team + 20 20 WC SA chasing 438 Aus players with World Cup 2007 SL protesting against Aus + walking out of the ground Eng vs SA + WC stalled by rain + 1992

Table 1. Query Formulations for various Query types

image is evaluated for relevance and the other images are considered not relevant. Additionally, if the image retrieved is not accessible due to technical difficulties in the site domain, the image is considered to be non-relevant. In order to obtain a stable performance measurement of image search engines, all the searches are performed within a short period of time (one hour) and the relevance of the images is decided by the subject matter expert.

3.2 Results

Traditionally evaluation for information retrieval has been based on the effectiveness ratios of precision (proportion of retrieved documents that are relevant) and recall (proportion of the relevant documents that are retrieved) [22]. Since the World Wide Web is growing constantly, obtaining an exact measure of recall requires knowledge of all relevant documents in the collection. Given the sheer volume of documents this is, for all practical purposes, impossible. Because of this, recall and any measures related to recall cannot be readily used for evaluation. This necessitates that the evaluation be based on the effectiveness ratios of precision. For purposes of this evaluation precision is defined as the number of relevant images retrieved to the total number of images retrieved. The search engines were evaluated based on the precision at a retrieval length R at R=10, 20, 30 and 40.

Figures 3-6 illustrate the average precision of each of the search engines for each of the query types: unique images, unique images with refiners, non-unique images, and non-unique images with refiners.

To check the adequacy of the factors, a factorial analysis was conducted and the results were analyzed using the analysis of variance (ANOVA) method. As previously discussed, the factors that were hypothesized to have a significant effect on the average precision of the retrieved results are Query Type, Search Engine and Retrieval Level.

The response variable is the average precision at retrieval length R which is defined as the ratio of the relevant retrievals to the overall number of images retrieved.

3.2.1 Hypothesis

Null Hypothesis: H_0: There is no significant effect of Query Type, Search Engine or Retrieval level R on the precision of the retrieved results.

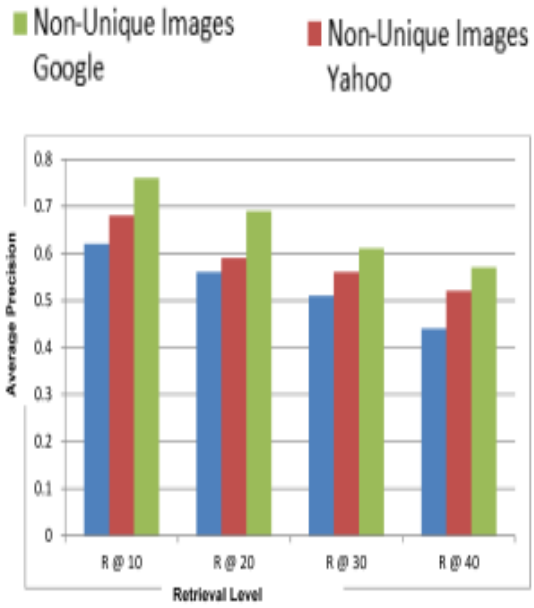

Figure 3. Unique Images - Average Precision

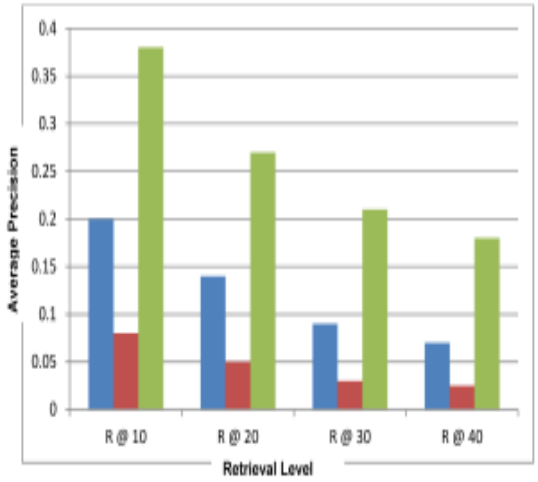

Figure 4. Unique Images with Refiners -Average Precision

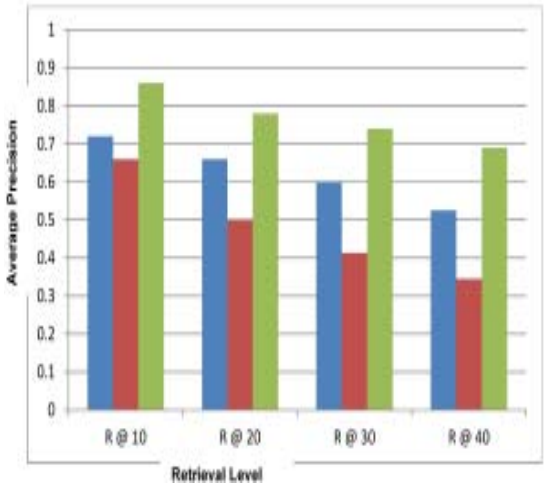

Figure 5. Non-Unique Images - Average Precision

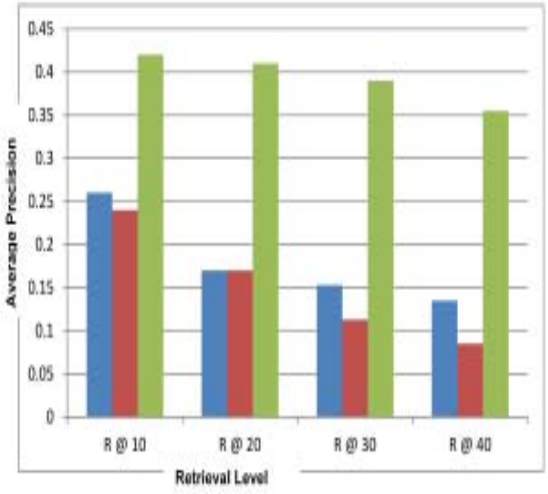

Figure 6: Non-Unique Images with Refiners -Average Precision

Alternate Hypothesis: H_1: There is significant effect of Query Type, Search Engine or Retrieval level R on the precision of the retrieved results.

3.2.2 Statistical Analysis

Data were collected for the 48 experimental trials of the 4 x 3 x 4 full factorial design that was run five times.

At the 99 % confidence level, the ANOVA results show that there is a significant effect of the main effects, Query Type**, $F_{(3,192)}$ = 55.70, p<0.0001,** Search Engine**, $F_{(2,192)}$ = 14.02, p<0.0001,** and Retrieval Level R**, $F_{(3,192)}$ = 4.46, p<0.0001,** and there are no effects due to interactions between the main effects. The ANOVA results of the overall model (taking all the main effects and interactions into consideration) are also significant at the 99% confidence level.

The results clearly show that all the main effects are significant; which takes us to the next step of further analysis of the response variable.

3.2.3 Performance Evaluation

The performance of the search engines for various queries and retrieval levels is discussed in this section. The average precision of the retrieved images for Unique Images for different levels & search engines is tabulated in Table 2 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 7

Search Engine	Average Precision			
	R @ 10	R @ 20	R @ 30	R @ 40
Google	0.76	0.69	0.61	0.57
Yahoo	0.68	0.59	0.56	0.52
Live	0.62	0.56	0.51	0.44

Table 2. Average Precision of retrieved images for Unique Images

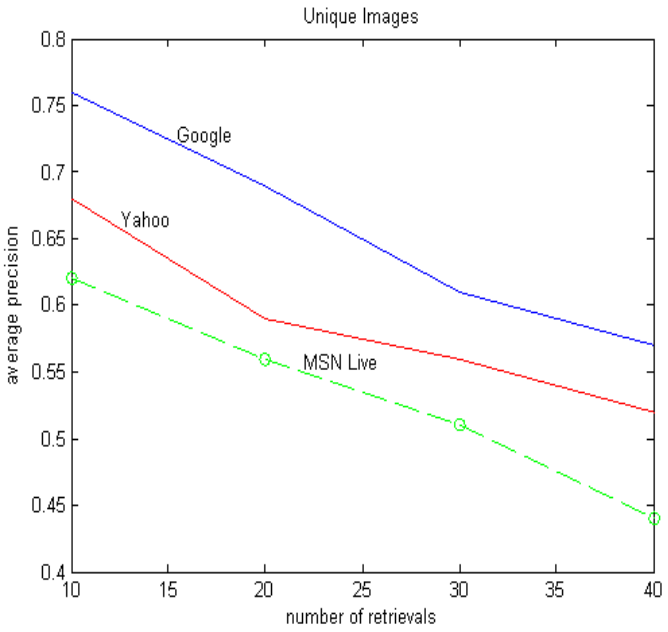

Figure 7. Average Precision of retrieved images for Unique Images

Figure 7 clearly illustrates that Google has the best average precision at any cut-off point for unique images, followed by Yahoo and MSN Live respectively; and that for each search engine the average precision tended to drop as the number of retrievals increased.

The average precision of the retrieved images for Unique Images with refiners for different levels and search engines is tabulated in Table 3 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 8.

Search Engine	Average Precision			
	R @ 10	R @ 20	R @ 30	R @ 40
Google	0.38	0.27	0.21	0.18
Yahoo	0.08	0.05	0.03	0.025
Live	0.2	0.14	0.09	0.07

Table 3. Average Precision of retrieved images for Unique Images with Refiners

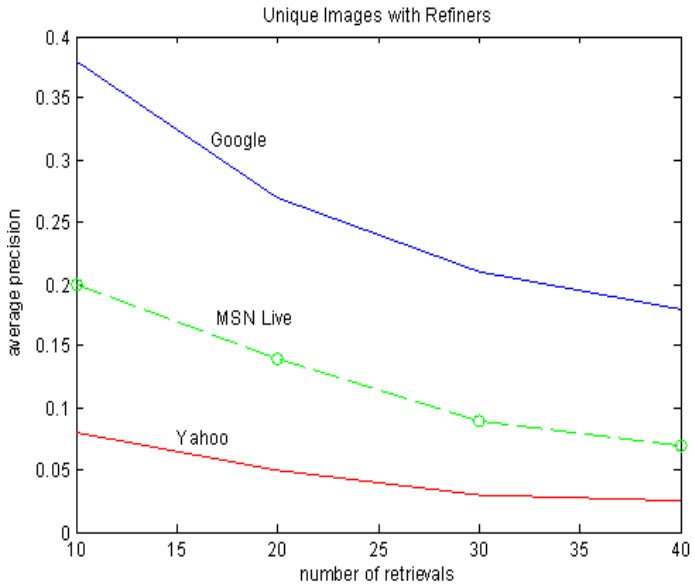

Figure 8. Average Precision of retrieved images for Unique Images with Refiners

Figure 9 illustrates that Google has the best average precision at any cut-off point for non-unique images, followed by MSN Live and Yahoo respectively; and that for each search engine the average precision tended to drop as the number of retrievals increased.

The average precision of the retrieved images for Non-Unique Images with refiners for different levels & search engines is tabulated in Table 5 and the performance of the search engines with respect to average precision is illustrated graphically in Figure 10.

Search Engine	Average Precision			
	R @ 10	R @ 20	R @ 30	R @ 40
Google	0.86	0.78	0.74	0.69
Yahoo	0.66	0.5	0.413	0.345
Live	0.72	0.66	0.6	0.525

Table 4. Average Precision of retrieved images for Non-Unique Images

Tukey's Honest Significant Difference for Search Engines (Figure 11) clearly shows that Google Image Search Engine outperforms Yahoo and MSN Live. This analysis also indicates that the performance of Yahoo and MSN Live does not differ statistically. Tukey's Honest Significant Difference for Query Types (Figure 12) clearly shows that the performance of search engines is better whenever there is no additional refiner.

4. Role of Humans in Image Annotation

4.1 Design of Experiment

Participants: For this study, eight participants were selected from a pool of candidates professing knowledge of the domain

Search Engine	Average Precision			
	R @ 10	R @ 20	R @ 30	R @ 40
Google	0.42	0.41	0.39	0.355
Yahoo	0.24	0.17	0.113	0.085
Live	0.26	0.17	0.153	0.135

Table 5. Average Precision of retrieved images for Non-Unique Images with Refiners

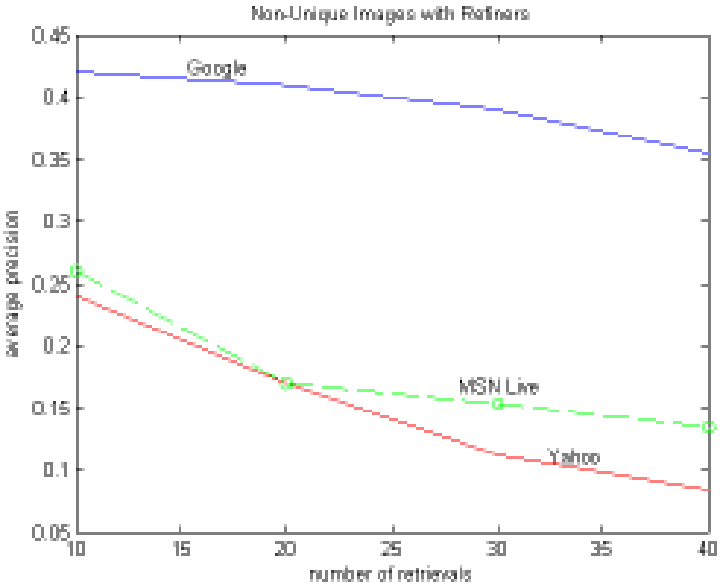

Figure 9. Average Precision of retrieved images for Non-Unique Images

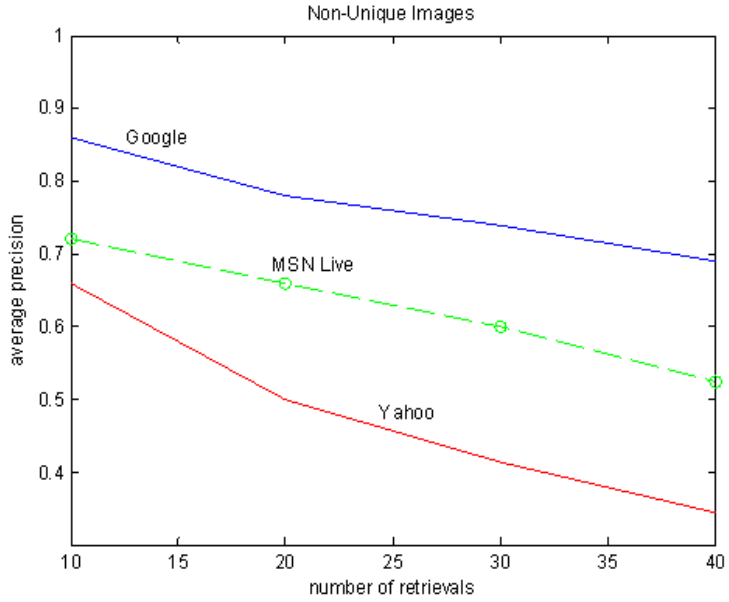

Figure 10. Average Precision of retrieved images for Non-Unique Images with Refiners

(the game of cricket). For the purposes of this experiment, a pre-annotation test was prepared to determine the level of expertise in the game of cricket. Those scoring higher than 90% were selected as the domain experts.

Apparatus and Stimuli: The participants viewed 40 images that were randomly selected from a database of 2,964 images taken over 25 years, all relating to the game of cricket. The domain experts were asked to label images on a personal computing system running at a minimum of 2.5GHz, Windows XP machine. A 17-inch LCD monitor was used to display the interface, with a mouse and keyboard used as the input devices. The experiment took place in an office type environment with ambient lighting conditions. The participants sat in an adjustable office chair, and the mouse and keyboard were placed at a comfortable position as determined by each participant.

Experimental Design: This experiment was an 8x4 full factorial design that was run twice. There were two independent variables, including search database (DE) and query type (QUERY). The eight levels of DE were DE 1, DE 2, etc., and the

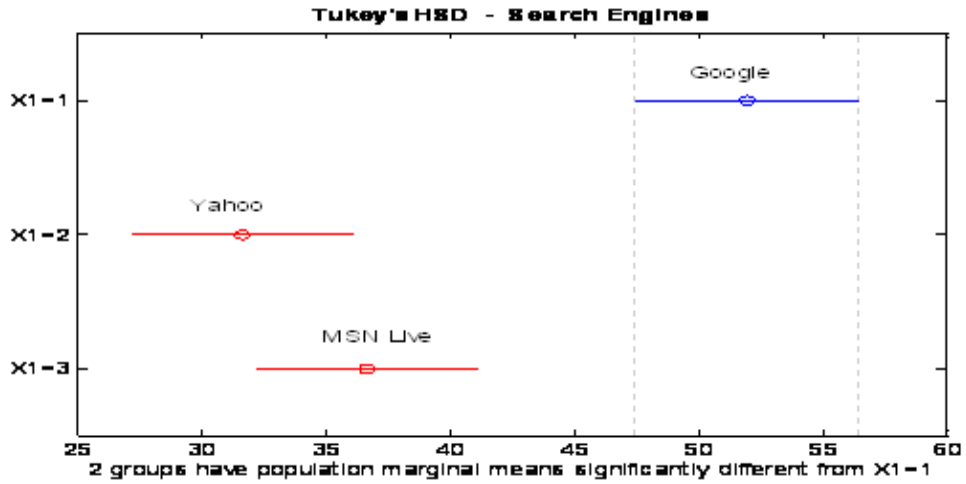

Figure 11. Tukey's Honest Significant Difference for Search Engines

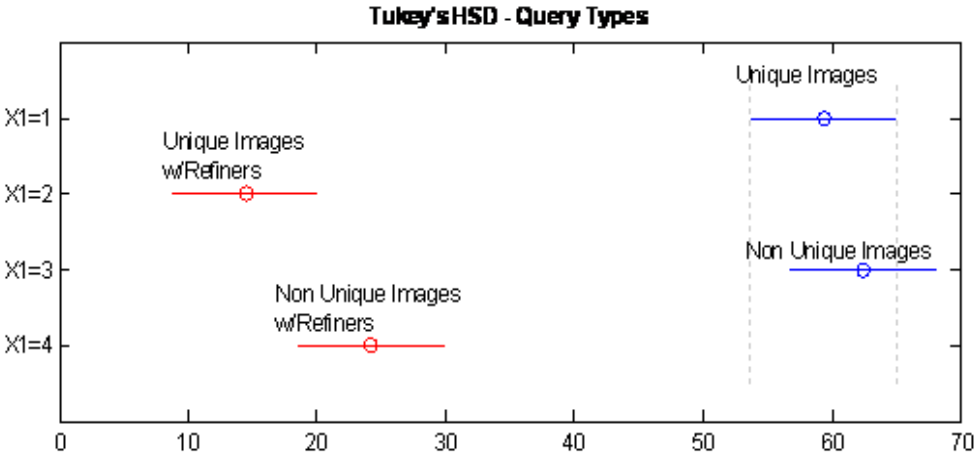

Figure 12. Tukey's Honest Significant Difference for Query Types

levels of QUERY were UNIQUE, UNIQUE with IDENTIFIER, NON UNIQUE, NON UNIQUE with IDENTIFIER. Each level of QUERY was tested with each database.

Traditionally, evaluation for information retrieval has been based on the effectiveness ratios of precision (proportion of retrieved documents that are relevant) and recall (proportion of relevant documents that are retrieved) [22]. In this study, the two dependent variables that were examined, precision and recall are calculated as the knowledge of all relevant documents in the collection.

Procedure: The eight participants were asked to view 40 images and annotate each of these individually. The images were presented to the participants in random order and they were only allowed to see one image at a time. The participants were told that the annotations should be a descriptor of the image from a domain expert point of view and should be comprehensive. They were also told that annotations should be similar to the keywords that they use when they upload an image or a video online. The annotations were to be filled in the text box below the image. If a participant was unable to come up with a label for an image they were asked to fill "N/A" in the text area box.

Finally they were told that the annotations will be used for information retrieval (image retrieval) purposes. After the

annotation was completed, eight different search databases were built to test the queries seen in Table 6.

Query Type	Queries
Unique Images	1. Henry Olonga 2. Kevin Pietersen
Unique Images with Identifiers	1. Sachin Tendulkar + Marriage 2. Kapil Dev + World Cup
Non Unique Images	1. Pakistan Cricket Team 2. Indian Cricket Team
Non Unique Images with Identifiers	1. England Team + Ashes 2. Australian Team + World Cup

Table 6. Queries Run on the Databases

The search engine used to run the queries was Google Desktop Search Engine. This was chosen as it has been shown in previous studies to outperform other desktop search engines on many difference measures [27].

Data were collected for the 32 experimental trials to test two hypotheses:

Hypothesis 1: There is no significant effect of the Search Engine and Query Type on the Precision of the retrievals.

Hypothesis 2: There is no significant effect of the Search Engine and Query Type on the Recall of the retrievals.

Figure 13 shows an example of the annotations of images for each query type.

Figure 13. Example of an annotated image for each query type

4.2 Results

The ANOVA results were obtained for the full factorial design experiment conducted for precision (proportion of retrieved images that are relevant) and recall (proportion of relevant images that are retrieved with a search query). At a 99% confidence level, the ANOVA results show a significant main effect for QUERY ($F_{(3,32)} = 6.4$, $p < 0.0001$) for precision and QUERY ($F_{(3,32)} =$

6.72, $p < 0.0001$) for recall. The other main effect Search Engine (A) or the interaction effect is not significant at 99% confidence, for either precision or recall.

These results reject our Null Hypothesis and validate that as long as the person annotating the images in a database is a domain expert, the performance of the search engine did not change significantly. Clearly, the human's cognitive abilities have to be better tuned for the person to annotate the images in a more systematic way to improve the performance of the search engines.

5. Annotation Template

The previous sections demonstrated that the performance of the current commercial annotation based image retrieval systems could potentially be improved for image search and retrieval. Figure 14 illustrates our framework that reduces the human mental workload by providing a template which will allow the human to annotate images in a systematic way. A metadata template has been constructed which extracts text from a given annotation, and places it in specific fields of the template.

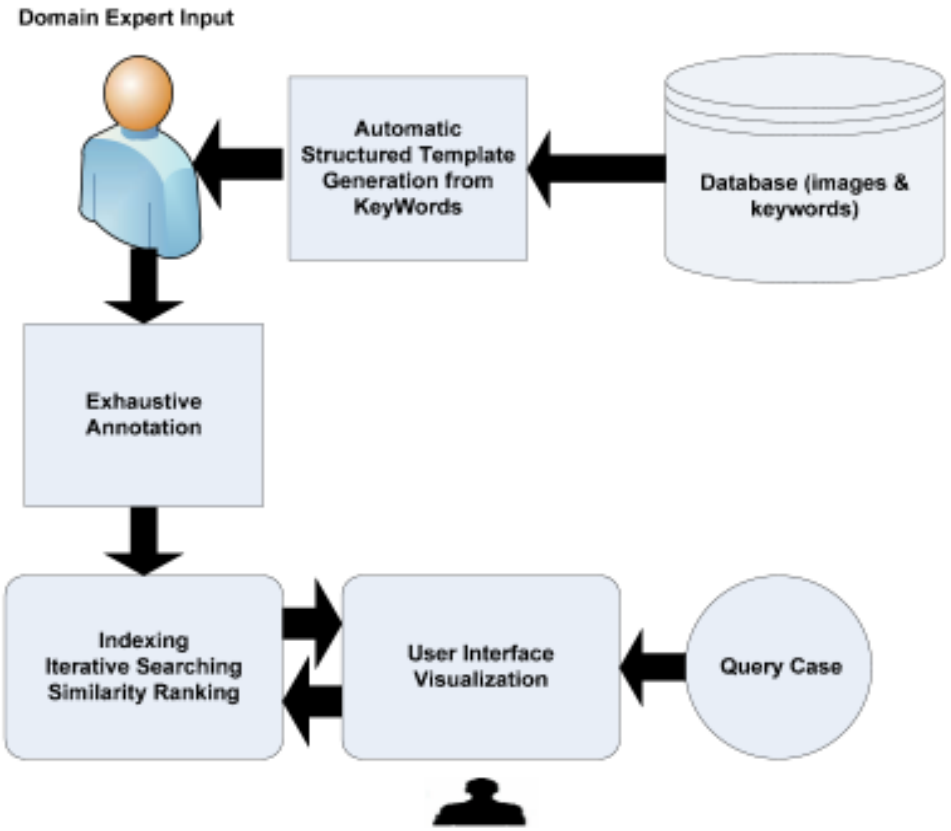

Figure 14. Human Computer Integrated System Architecture

The framework for classification of content for image retrieval by Shatford [28], is a targeted approach to understand the needs of the users who are searching for images in specific domains. The framework established by Shatford is motivated by this recognized necessity for considering classification of images from user-centered rather than the system-centered point of view. Apart from deriving taxonomy for categorization of images, the aim of this taxonomy is to allow researchers to design the user interface for image databases based on the content of images. The structure set forth by researchers is loose enough to allow the individual researcher to tailor it to their own need, but focused enough to allow the researchers to be focused on the problems. An example of the metadata template features extracted from the Shatford's model mapping to the game of cricket is illustrated in Figure 15.

To generate the automatic structured template generation from keywords, domain knowledge was utilized to construct the

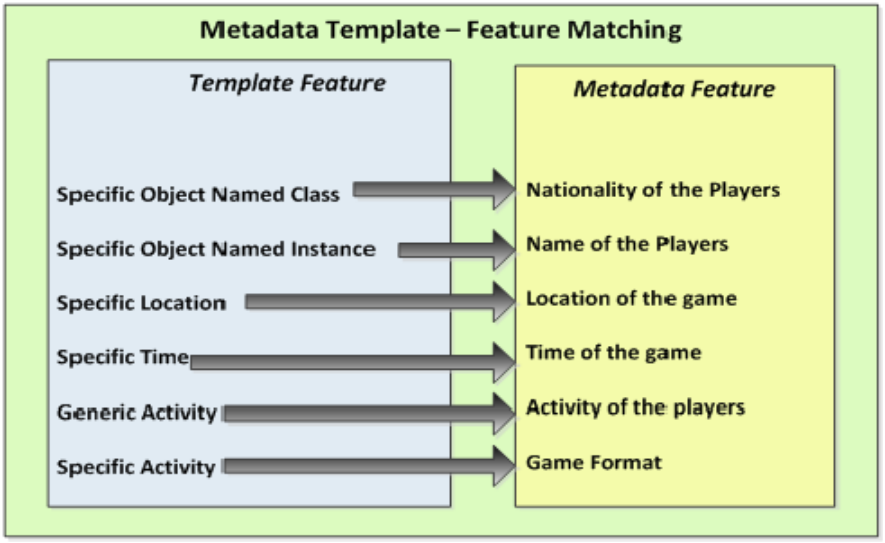

Figure 15. Functional flow of automatic structured template generation from keywords

template features. From the metadata available for the image, stop-words like is, the, and etc. were removed. The remaining words were compared with the arrays of each template feature, and if matched with a particular array, were placed in that template field. The template features used from Shatford's model are discussed in the following paragraphs.

Specific Object Named Class: The specific object named class is related to the nationality of the players for the domain of cricket. An array was constructed with each element of the array containing a country name. The country information was downloaded from http://www.cricinfo.com/ci/content/page/417881.html. If a word from the metadata of the template matches with the specific object named class array, the word is placed in the specific object named class field.

Specific Object Named Instance: The specific object named class is related to the name of the players for the domain of cricket. If the first character of the words in the metadata is an upper case letter and the word does not fit in the Specific Object Named Class, Specific Location, Specific Time or Specific Activity, it will be automatically placed in the Specific Object Named Instance field.

Specific Location: The specific location class is related to the location of the playground for the domain of cricket. An array was constructed with each element of the array containing a playground name. The playground information was downloaded from http://www.cricinfo.com/ci/content/current/ground/index.html. If a word from the metadata of the template matches with the specific location array, the word is placed in the specific location field.

Specific Time: The specific time class is related to the time of the game for the domain of cricket. An array was constructed with each element of the array containing a time which includes months, years, and days. If a word from the metadata of the template matches with the specific time array, the word is placed in the specific time field.

Generic Activity: The generic activity time class is related to the activity of the players for the domain of cricket. An array was constructed with each element of the array containing activities of the players that include "batting", "bowling", etc. If a word from the metadata of the template matches with the generic activity array, the word is placed in the generic activity field.

Specific Format (Activity): The specific format (activity) class is related to format of the game for the domain of cricket. An array was constructed with each element of the array containing formats of the games that include "Test", "One-Day", "World Cup" etc. If a word from the metadata of the template matches with the specific format (activity) array, the word is placed in the specific format (activity) field.

Any other information that does not fill in any of the metadata fields are filled in the Other Information field. If any metadata template feature field is empty, "N/A" is filled into the field. An example of the metadata template generation is showcased in Table 7.

Image Metadata:	"Kapil Dev and Mohinder Amarnath are all smiles after winning the World Cup. India beat West Indies in the final. June 25, 1983"
Remaining Metadata after Remove Stop-Words:	Kapil Dev Mohinder Amarnath smiles winning World Cup India West Indies final June 25 1983
Specific Object Named Class field:	India West Indies
Remaining Metadata:	Kapil Dev Mohinder Amarnath smiles winning World Cup final June 25 1983
Specific Location:	"N/A"
Remaining Metadata:	Kapil Dev Mohinder Amarnath smiles winning World Cup final June 25 1983
Specific Time:	June 25 1983
Remaining Metadata:	Kapil Dev Mohinder Amarnath smiles winning World Cup final
Generic Activity:	smiles winning
Remaining Metadata:	Kapil Dev Mohinder Amarnath World Cup final
Specific Format (Activity):	World Cup final
Remaining Metadata:	Kapil Dev Mohinder Amarnath
Specific Object Named Instance:	Kapil Dev Mohinder Amarnath
Remaining Metadata:	----------

Table 7. Example of Metadata Template Generation

Once the template is created from the given annotation, this template is sent to the subject matter expert who accepts or rejects the annotations provided in the template, adds any other information that is pertinent, and finally saves it. An exhaustive annotation html file is generated as the last step to be stored in the human integrated database.

This approach of using the domain expert to fill in the information in the template helps him to annotate images more systematically with minimal effort. An example of the metadata template generated during this step is showcased in Figure 16.

We hypothesize that the performance of the search engine in metrics of precision and recall will be enhanced with the addition of the template. This template will incorporate human expertise to capitalize on the strengths of manual annotation and that we have shown with using domain experts while avoiding the out-of-the-loop performance problems that occur with a completely automated system [29, 30].

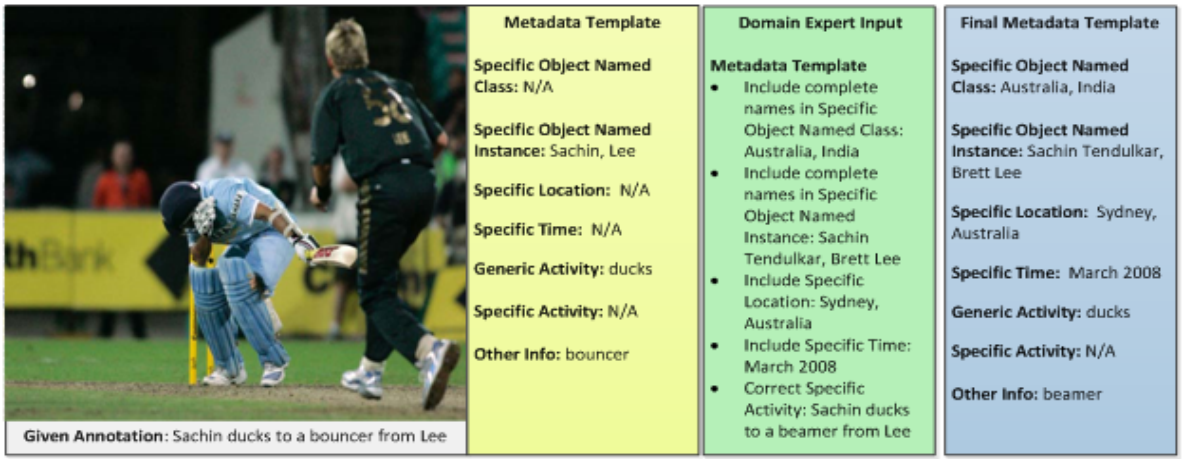

Figure 16. Human Input to Metadata Template

6. Conclusion and Future Work

The results of the benchmarking research suggest that overall, commercial search engines continue to have significant difficulties effectively executing image retrieval tasks. The Google search engine performs significantly better than Yahoo or MSN Live in any query type. The results also indicate that the precision of the search engines tended to drop with the increase in the number of retrievals. This performance reduction was noted across-the-board, that is, irrespective of the search engines and the query types. The performance of the search engines also dropped dramatically when the queries had refiners (unique or non-unique).

The role of human annotation results also show that there is no significant difference in the performance of the search engines when a domain expert annotates the images. This result is important as it supports that we do not need a panel of annotators to label an image, as long as the annotator is proficient in the image domain. The results also indicate that the performance of the current commercial search engines can be improved by a disciplined annotation approach by domain experts. With the number of images available on the internet growing exponentially, the human is incapable of annotating all these images in a systematic manner. Computer Vision algorithms can potentially be used to alleviate the load of the human operator for annotating the images. These algorithms can annotate images for content dependent metadata, but as they are still in their infancy, they fail when annotations requiring content descriptive metadata are required.

Clearly, the human's cognitive abilities have to be better tuned for the person to annotate the images in a more systematic way to lead to improve the performance of search engines. We have proposed a semi-automated annotation framework that provides a systematic metadata template to improve upon the negative aspects associated with manual annotation. A systematic evaluation of the framework is one of the future goals of this study. As computer vision algorithms are refined, we will also be evaluating their use in filling in content dependent metadata in our template. Our goal is to bring the human into the loop to validate not only the metadata template, but also the template that is created from the visual features of the image by the Computer Vision algorithms to create a highly comprehensive image search database.

References

[1] Kidambi, P., Narayanan, S. (2008). A human computer integrated approach for content based image retrieval. *Recent Advances in Computer Engineering. Proceedings of the 12th WSEAS International Conference on Computers*, 691-696.

[2] Yates, B., Neto, R., (1999). Modern Information Retrieval. *ACM Press.*

[3] Witten, I.H., Moffat, A., and Bell, T. (1999). Managing Gigabytes: Compressing and Indexing documents and images. *Morgan Kaufmann Publishers.*

[4] Kuralenok, I. E., Nekrestyanov, I. S. (2002). Evaluation of Text Retrieval Systems. *Programming and Computer Software, 28(4)*, 226-242.

[5] Text Retrieval Conference (TREC) (1992). National Institute of Standards and Technology (NIST) and U.S. Department of Defense, *http://trec.nist.gov/*

[6] Vorhees, E.M., Harman, D.K. (2005). TREC: Experiment and Evaluation in Information Retrieval, *The M.I.T. Press, ISBN 0-262-22073-3*

[7] Buckley, C., Vorhees, E.M. (2004). Retrieval Evaluation with Incomplete Information. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, 25-32.

[8] Sanderson, M., Zobel, J. (2005). Information retrieval system evaluation: effort, sensitivity, and reliability. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, 162-169.

[9] Inoue, M. (2004). On the need for annotation-based information retrieval. *Information Retrieval in Context, SIGIR IRiX Workshop*, 44-49.

[10] Choi, Y., Rasmussen, E. M. (2002). *Users'* relevance criteria in image retrieval in American history. *Information Processing & Management, 38(5)*, 695–726.

[11] Hughes, A., Wilkens, T., Wildemuth, B., and Marchionini, G. (2003). Text or pictures? An eyetracking study of how people view digital video surrogates. *Proceedings of the International Conference on Image and Video Retrieval*, 271–280.

[12] Chen, Y., Wang, J.Z. (2004). Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research ,5*, 913–939.

[13] del Bimbo, A. (1999). Visual Information Retrieval. *Morgan Kaufmann, Los Altos, CA*.

[14] Liu, Y., Zhang, D., Lu, G., Ma, W. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition, 40(1)*, 262-282.

[15] Hyvo¨ nen, E., Styrman, A., Saarela, S. (2002). Ontology-based image retrieval. *Proceedings of XML Finland Conference*, 27–51.

[16] Hanbury, A. (2008). A survey of methods for image annotation. *Journal of Visual Languages & Computing (19)*, 617-627.

[17] Ahn, L.V., Dabbish, L. (2004). Labeling images with a computer game. *Proceedings of ACM CHI*, 319–326.

[18] Hernon et al. (1990). Evaluation and Library Decision Making. *Alex Publishing*.

[19] Meadow et al, (1999). Text Information Retrieval Systems, Library and Information Science series. *Elsevier publications.*

[20] William Hersh (1995). Information Retrieval – A Health Care perspective. *Springer publications.*

[21] Lancaster et al. (1993). Information Retrieval Today. *Information Resource Press*.

[22] Smith, J. R. (1998). Image Retrieval Evaluation. *IEEE Workshop on Content-Based Access of Image and Video Libraries, 21*, 112-113.

[23] Cooper W.S. (1968). Expected Search Length – A single measure of retrieval effectiveness based on weak ordering action of retrieval systems. *Journal of the American Society for Information Science, 19*, 30-41.

[24] Cakir, E., Bahceci, H., Bitirim, Y. (2008). An Evaluation of Major Image Search Engines on Various Query Topics. *The Third International Conference on Internet Monitoring and Protection, IEEE Computer Society*, 161-165.

[25] Enser, P.G.B., BMcGregor, C. (1993). Analysis of Visual Information Retrieval Queries. *British Library Research, and Development Report 6104*.

[26] Nielsen Search Rankings (2009). *http://www.nielsen-online.com/pr/pr_090616.pdf*

[27] Lu etal. (2007). Performance Evaulation of Desktop Search Engines. *IEEE International Conference on Information Reuse and Integration*, 110-115.

[28] Shatford, S. (1986). Analyzing the subject of a picture – A theoretical approach. *Cataloguing & Classification Quarterly, 5(3)*, 39-61.

[29] Endsley, M., & Kiris, E. (1995). The out-of-the-loop performance problem and level of control in automation. *Human Factors, 37*, 381-394.

[30] Thackray, R., & Touchtone, R. (1989). Detection efficiency on an air-traffic control monitoring task with and without computer aiding. *Aviation Space and Environmental Medicine. 60*, 744-748.