Semantically Relevant Image Retrieval by Combining Image and Linguistic Analysis

Tony Lam, Rahul Singh

Department of Computer Science
San Francisco State University
San Francisco, CA94132
tonster@sfsu.edu,rsingh@cs.sfsu.edu

Abstract. In this paper, we introduce a novel approach to image-based information retrieval by combining image analysis with linguistic analysis of associated annotation information. While numerous Content Based Image Retrieval (CBIR) systems exist, most of them are constrained to use images as the only source of information. In contrast, recent research, especially in the area of web-search has also used techniques that rely purely on textual information associated with an image. The proposed research adopts a conceptually different philosophy. It utilizes the information at both the image and annotation level, if it detects a strong semantic coherence between them. Otherwise, depending on the quality of information available, either of the media is selected to execute the search. Semantic similarity is defined through the use of linguistic relationships in WordNet as well as through shape, texture, and color. Our investigations lead to results that are of significance in designing multimedia information retrieval systems. These include technical details on designing cross-media retrieval strategies as well as the conclusion that combining information modalities during retrieval not only leads to more semantically relevant performance but can also help capture highly complex issues such as the emergent semantics associated with images.

1 Introduction

With the rapid proliferation of image-based information, image retrieval systems are becoming increasingly significant and have many applications in domains such as multimedia information, personal information management, geographical information systems, and bio-medical imaging to name a few. Much research has been done in this area, especially ones geared towards designing better image analysis techniques for use in Content Based Image Retrieval (CBIR) systems. As newer and better image analyzing methods are developed, *the best image-based recognition algorithm have been found to provide only partial solution to the image retrieval problem*. This is because of the fact that the image retrieval problem is more complex than analyzing pixel-level information. In recent past, researches have attempted to take advantage of the non-visual information that may be associated with images. The most common of these are related to searching for images on the Web, where the search results are typically based (at the state-of-the-art) on analysis of the textual information (such as the caption) associated with an image along with an analysis of the hyperlink structure of the page where the image is found.

Our goal is to develop retrieval strategies that go beyond examining the problem from the extremal viewpoints of image-only or text-only analysis. Beyond that, we want to understand what the query means and how meaningful the retrieved result is to the query. Since annotations are linguistic constructs and languages (in our case English) have a structure, we use WordNet [11], a psycholinguistic dictionary project developed at Princeton University. WordNet encodes a variety of semantic relationships inherent to the language. Thus, with its help, we can perform better text analysis and comparison that is based not only on keyword matching but also on the semantics of the image annotation.

In summary, we pose the retrieval problem not merely in terms of determining pixel-level similarity. This complementary formulation allows one modality to compensate for the deficiency of the other. Furthermore, the formulation also allows the retrieval strategies to be mutually supportive, especially in cases where one of them can provide results that are semantically more relevant than the other.

2 Previous Work

Numerous content-based image retrieval systems are available today; all perform excellent low level feature matching retrieval: color and texture [3][7][10][16]. Others use newer representation such as wavlets [18]. While many object recognition and region segmentation algorithms have developed from this problem space, these techniques are not perfect and will probably never be. Aside from the difficulty in developing perfect computer vision techniques, learning the semantic meaning of an image is an even harder problem.

Text-based image retrieval is more novel [14]. Many online search engines have expanded their WWW search to include subsystems made especially for image retrieval [8][9]. Most WWW images are associated with web pages and the textual content contained in them are very helpful in describing the images. Analyzing the text information as part of the image retrieval process can capture the semantic implication of the image from a linguistic level just as it can capture the same type of information from a web page. However, little research has been done to pursue capturing image semantics during the retrieval process [1][2][17].

Other text-based image retrieval systems, including meta-tagging the images, try to address the image semantic issue [19], but most are only able to address low level semantics of the image. The low level semantics of an image can be directly captured from low level feature analysis. But a more complex problem exists in capturing high level semantics, or the conceptual neighbors. Adding to the complexity is the assumption that all annotations are meaningful to the image. This brings us to the concept of emergent semantics. For example, one user may annotate a photograph of a tiger as a trip to a zoo, and another user may annotate the photograph as an animal in the jungle. Both annotations are valid given the proper user perspective and context. Combining the content and textual based approaches to image retrieval will yield more meaningful and contextual results and allow us to search using the semantics of the image in addition to low level object searching.

One way to capture image semantics is to combine the two aforementioned approaches [4][5][12] and pursue image retrieval from a conceptual level instead of the traditional feature level that most abovementioned research has done. This way would guarantee to outperform either approach alone. The combined approach is similarly used in hierarchical clustering of WWW image search results [6], but our proposed technique is used in the actual retrieval phase with the intention to capture the semantics of the image.

3 Image Semantics and Philosophy of the Proposed Approach

In the context of information retrieval, one predominant difference between image and textual media is how the corresponding semantics may be discerned. While textual media have relatively straightforward interpretations where each word is a contributing part of the whole meaning, image semantics is fundamentally different. Image semantics, unlike the actual image, cannot be dissected into semantically relevant blocks. It does not always possess the compositional nature of image representation and therefore cannot be captured through analysis of the image features alone.

For images, we distinguish three different levels of semantics. *Low level semantics* is the information that can be retrieved through direct interpretation of the image annotation or through content object segmentation. For the purpose of many information goals, this level of interpretation is sufficient and no additional information about the image needs to be discovered. *High level semantics* expands upon the low level semantics of an image by examining the conceptual relationships that may exist. It thus goes beyond simple image interpretation. For instance, given the image of a tiger, based on high level semantics, the image can be related to those of felines or carnivores. While such interpretation is obvious, it is hard to achieve using image-based information alone. Finally, *emergent semantics* is perceptual and contextual [15]. This level of semantics is the most difficult semantic level to capture since it is highly subjective and differential. Since emergent semantics is nearly impossible to capture with any single mode of feature-based analysis, a better approach to capture the emergent semantics is to combine information from multiple semantically correlated media (whenever available). However, and we underline this, it is important to note that in such cases some or all of the media may be noisy, have errors, or be semantically non-correlated. For image-content based strategies, causes can be noise, poor image processing or low level (signal-level) features that do not necessarily correspond to semantics of the query. For textual strategies, poor performance may be due to the lack of relevant keywords in the annotation, poor keyword selection, or inadequate capture of the underlying linguistic relationships. Therefore, retrieval by combining information from multiple media is better than focusing on a specific media alone. In addition to dealing with the above issues, an important step in combining information from different media, such as in our approach, is to determine how the annotation and the image are semantically related.

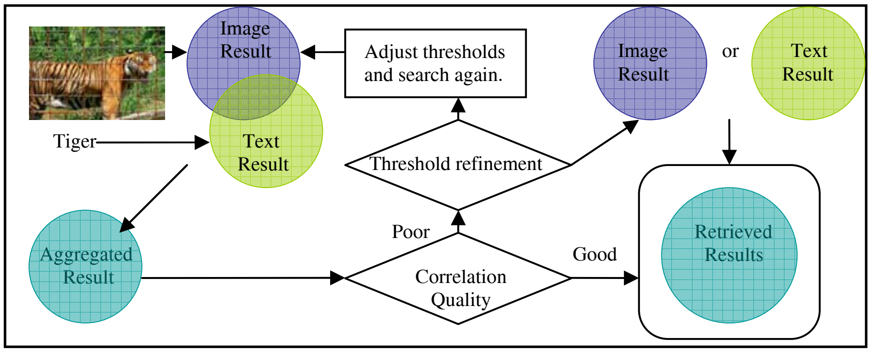

Fig. 1. Outline of the proposed approach: the correlation between the image and the annotation is analyzed. If the correlation is good, results are retrieved using information from both modalities. Otherwise the thresholds are iteratively adjusted. If the correlation continues to be poor, the more information rich modality is determined and used for retrieval.

The information at the image-level and the annotation-level may span one or more of the aforementioned semantic classes. Further, either of the modalities may suffer from any of the aforementioned sources of noise. To improve retrieval quality, we therefore need to understand the correlation between the annotation and the image and minimize the effects due to noise or imprecision in either of the modalities. Our solution philosophy is centered on the idea that if we find the annotation to image correlation to be good, the aggregated result from using both sources of information is returned. If the correlation is poor, the more semantically relevant information modality (annotation-based or image-based) is identified and used for retrieval.

4 Algorithm and System Development

Our approach of image retrieval combines text and image analysis to capture the semantics of the image collection. Image collections are added and annotated by the user. The annotation can range from a simple one word descriptor or a verbose paragraph. Retrieval is initialized by a text search, and a representative image is picked from result. Our system will search the database using both image features and text query and/or image annotation using the user-selected minimal similarity thresholds. Each analysis is discussed in greater detail in the following sub-sections.

4.1 Image Analysis

Semantically similar images seldom have high visual similarity due to different color and texture features. We adapted the image segmentation technique used in the Blobworld image retrieval system [3]. Blobworld segments an image using Expectation-Maximization (EM) algorithm based on color and texture.

In our system, an image is size-normalized and converted to the L*a*b* color space. Each pixel is then analyzed and clustered together to form regions with similar characteristics based on ten iterations of EM. Only regions with size of at least 2% of the entire image are deemed significant and have their characteristics stored in the database. Each region's simplified color histogram, contrast, anisotropy, and size are stored as a feature vector in the database.

Upon choosing the image that will be used for image similarity comparison, the user selects the region(s) of interest. Each selected region is compared with regions of other images in the database. The feature vector for each region includes color, texture, and size. Each image comparison is indexed after the initial search for efficiency purposes. All images are preprocessed offline since processing time for each image is between three and five minutes.

We compute the image similarity score μ_i for the chosen image with selected regions r_i in the follow manner:

1. For each region r_j in the database image with feature vector v_j, Euclidean distance between v_i and v_j, and similarity between the two regions is calculated:

$$d_{ij}^2 = (v_i - v_j)^T \Sigma(v_i - v_j) \qquad \mu_{ij} = e^{-\frac{d_{ij}}{2}} \tag{1}$$

2. The score μ_i is:

$$\mu_i = \max{}_j \; \mu_{ij} \tag{2}$$

The matrix Σ is block diagonal with the block corresponding to the texture and size features being the identity matrix with diagonal values 20, 50 and 5. The block corresponding to the color feature is the matrix *A* for *i* and *j* corresponding to the color bin for the color histogram of the region:

$$A_{ij} = \{1.0 \;\; if \;\; i = j, \;\; 0.5 \;\; if \;\; d_{ij} < 1.8, \;\; 0.0 \;\; otherwise \tag{3}$$

4.2 Text Analysis

Due to inherit subjectivity of information retrieval, low level keyword matching is not enough to completely capture the image semantics. We expand the text information associated with the image with psycholinguistic data available through WordNet and TFIDF to capture the semantics of the image annotation. TFIDF is popular text analysis technique that employs a term weighting scheme that adjusts the frequency weight of terms in a document by taking the product of the term frequency and the inverse document frequency (document frequency refers to the number of documents the term appears in).

We extended our prior research [20] for text analysis of the image annotations, including also synonym and holonym (x is a holonym of y if y is part of x) hierarchies in addition to hypernyms. This technique allows image semantics beyond the low level to emerge. WordNet stores each lexical unit as a synset (synonym set) that consists of terms that semantically mean the same thing. Each term may have several sense associations. For example, "tiger" can mean a big cat or a fierce person. For each term in the annotation, the follow is computed:

1. A TFIDF value is computed for each term where the IDF consists of a background set. We queried Google using random keywords and retrieved the top 50 documents for each query and built a term frequency matrix based on those results.
2. A hierarchy of lexical relationships including synonym, hypernym and holonyms. As the level of the hierarchy becomes more general, the weight assigned to the level decreases. The weight assigned to each term in the hierarchy is adjusted by the total length of the tree. For any term *T* let T_k denote the k_{th} term in the hierarchy. Also, let *depth(T)* denote the depth of the hierarchy for term *T*. The weight W_k for T_k is computed:

$$W_k = \frac{depth(T) - k}{depth(T)} \tag{4}$$

The similarity score between the query and the image annotation is computed from how closely the two hierarchies match. Comparisons are made with hierarchies of the same type (noun hypernym tree with noun hypernym) and the maximum score is assigned as the text score for the image. This score is then multiplied by the TFIDF value of term in the annotation. For queries with multiple keywords, the search is performed with each keyword having equal weight. The total score for each image is the sum of the text score for each keyword in the query. For a set of terms *T*, let T_i denote i_{th} term in *T* and T_{ij} denote j_{th} term in the lexical hierarchy for T_i. Let *W* denote the weight and n be the number of terms in the query. For query *Q* and each image annotation *A* in the database, the similarity score t_i between query Q_i and annotation A_k for matching Q_{ij} and A_{kl}, and the total text score *t* is:

$$t_i = \max(W_{Q_{ij}} \times W_{A_{kl}} \times tfidf_{A_k}) \qquad t = \frac{1}{n}\sum_{i=1}^{n} t_i \tag{5}$$

4.3 Determining Image-Annotation Correlation

After retrieving results based on text and image query, we proceed to analyze their correlation. We establish the correlation score by looking at the images in the intersection of the text search result and the image search result with total similarity score above user provided threshold which is preset at 0.7. Let *T* be the result set returned by text analysis and *I* be the result set returned by image analysis and *S* be the set of images in the intersection above the threshold. Each image result S_i has a text similarity score t_i and an image score μ_i. Let *n* be the cardinality of *S*, we compute the correlation score *c* by:

$$c = \frac{1}{n}\sum_{i=1}^{n}(t_i \times \mu_i) \tag{6}$$

If we determine that a good correlation ($c > 0.6$) exists between the text query and the image, we can simply return the intersection ranked by the total score, followed images with text or image scores above the corresponding text or image threshold. The minimum text and image thresholds are established by taking the min(t_i) and min(μ_i) from *S*. For poor correlation ($c \leq 0.6$), we need to determine which retrieval technique would yield more accurate overall results. We compute the text similarity score $\bar{t}$ and image similarity score $\bar{\mu}$ by taking the mean of the respective similarity score of images in *S*.

We then adjust the search text and image search threshold with the text and image similarity scores by multiplying each threshold by the ratio of the average text and image similarity scores. Let *T(q,t)* denote the text results retrieved with the query *q* with minimum text threshold *t* and *I(p, μ)* denote the image results retrieved with image *p* with minimum image threshold μ . Also, let *S* denote the new aggregated result set after the threshold adjustments:

$$a = \frac{\bar{t}}{\bar{\mu}} \quad t^* = \begin{cases} \bar{t} & \mathit{if}\ a \geq 1 \\ \bar{t} \times a & \mathit{otherwise} \end{cases} \quad \mu^* = \begin{cases} \bar{\mu} & \mathit{if}\ a \leq 1 \\ \bar{\mu} \times a^{-1} & \mathit{otherwise} \end{cases} \tag{7}$$

$$S = T(q, t^*) \bigcap I(p, \mu^*) \tag{8}$$

If the overall correlation *c* does not improve beyond the preset minimum of 0.6 within three search iterations, the more relevant text or image only result is returned.

5 Experimental Evaluation

We conducted our experiments on a collection of 1564 images where each image is annotated with a brief description of what is in the image. Annotations varied in length, context and semantic level. The experiments investigate the efficacy of our solution philosophy, provide insights about the nature of the data, compare our approach with more-classical single media retrieval, and analyze parameters characterizing the retrieval performance.

To demonstrate how the system works we consider two queries with different levels of semantic complexity. The first of these is the query “tiger” with the similarity thresholds set at 0.50. Retrieval using text-only information and semantic relations from Wordnet is shown in Fig. 2 and can be observed to be highly relevant. However, an image of a woman is included in the top results. This stems from the fact that the image’s annotation is “model” and the word “model,” in the sense of a person, who is role model, is highly similar to “tiger,” which also has the common root of “person” (fierce person). This example illustrates the fact that even with highly relevant annotation; text analysis alone cannot adequately capture the semantics of the query.

Our combined approach returned images that are much more coherent than either approach. The image of the model is demoted to a lower rank due to the lack of image similarity with the query image. Other images of tiger are promoted to the top to reflect the positive reinforcement of having the text and image scores combined. Those irrelevant images returned as top matches for the image result are filtered away because they have no textual similarity with the text. The calculated correlation in this search was 0.68. In this example, there is good correlation between the text query and the query image and our technique is able to capture the correct semantics of the query and retrieved images similar to the tiger semantics.

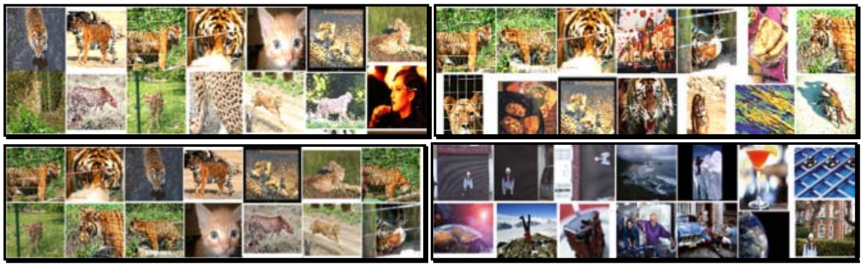

Fig. 2. *From top to bottom, left to right:* Top text result for the query "tiger" with text similarity threshold set at 0.60 (the image of the female "model" lower right), top image results for the image query with similarity threshold set at 0.75, top combined search result for the query "tiger" and the tiger image with overall threshold set at 0.70, and search results for the query "London vacation".

In the second example, we use the information goal "London vacation". The reader may note that semantics underlying this information goal are highly subjective and contextual. From the text result, an image of a person standing in from of a closed store is selected as representative because the person should be in most of the London vacation images being looked for. The result for this search is shown in Figure 2. The poor coherence in the result demonstrates the difficulty in capturing such semantics from the image. However, our approach has better recall than those obtained with text or image-based information alone. Additional experiments were conducted and data is summarized in Table 1 and Figure 3. Our combined approach has good precision performance with a high similarity threshold set at 0.90. As we relaxed the threshold requirement, there is a noticeable drop in precision for some queries such as "roses" and "vacation." This phenomenon is due to the highly contextual nature of those queries. Vacation is a concept that is difficult to capture with annotation and nearly impossible with image features; roses come in all different colors and they are not usually enough of a featured object in an image to have an explicit annotation. Since our approach analyzes result from text and image, our performance is dependant on the quality of those retrievals. In almost all cases, the quality of the combined result surpasses both techniques.

Table 1. Precision and recall for selected queries using text only, image only, and the proposed method at 0.70 and 0.90 similarity thresholds

	Query	**Retrieval Technique**			
		Text	Image	0.70	0.90
Precision	tiger	27%	31%	35%	50%
	wine	67%	10%	75%	75%
	cake	60%	8%	68%	100%
	vacation	100%	6%	22%	100%
	strawberry	65%	33%	83%	83%
	roses	17%	7%	15%	100%
Recall	tiger	30%	85%	85%	85%
	wine	23%	20%	10%	10%
	cake	88%	11%	88%	18%
	vacation	16%	16%	16%	5%
	strawberry	100%	40%	100%	100%
	roses	80%	20%	80%	80%

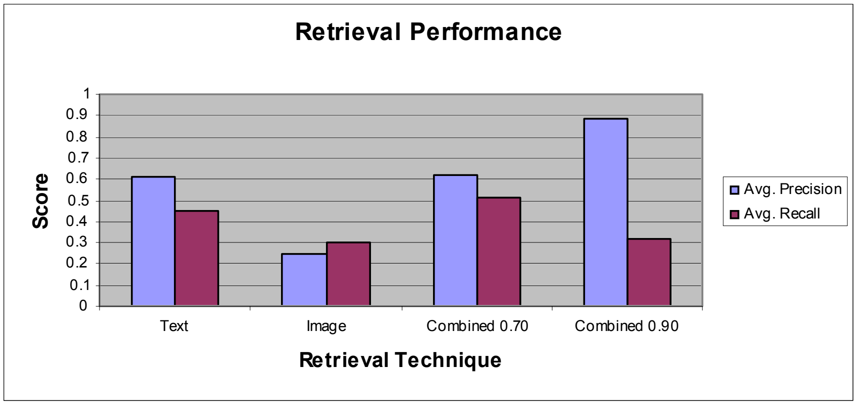

Fig. 3. Average retrieval performance for ten random queries using different retrieval strategies. Text and image similarity thresholds are set at 0.6 and 0.75 respectively. For our technique, the overall similarity thresholds are set at 0.7 and 0.9.

5 Conclusion

In this paper we address the problem of image retrieval by combining visual information along with the linguistic relationships that exists in the textual annotations associated with images. This formulation is distinct from pure image-based retrieval, where techniques have been stymied due to the signal-to-symbol gap. It is also distinct from pure text-based approaches, which are prone to textual ambiguities and by disregarding image-based information can not detect inconsistencies between the image-based and text-based descriptions, when they occur. The proposed approach combining these two perspectives and is able to retrieve information that is semantically more meaningful and robust, than what is possible using image-only or text-only strategies. Furthermore, the proposed approach is able to handle complex characteristics, such as emergent semantics, that are associated with images. While preliminary, our investigations provide intriguing evidence in favor of the proposed strategy for using multiple-media in problems of complex information retrieval.

References

1. Y. Aslandogan, C. Their, C. Yu, J. Zou and N. Rishe, “Using Semantic Contents and WordNet in Image Retrieval”, in Proceedings of *ACM SIGIR* Conference, Philadelphia, PA, July 1997.
2. K. Barnard and D. Forsyth, "Learning the Semantics of Words and Pictures", International Conference on Computer Vision, vol 2, pp. 408-415, 2001.

3. C. Carson, S. Belonge, H. Greenspan, and J. Malik, “Blobworld: Image segmentation using Expectation-Maximization and its application to image querying,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, SUB.
4. F. Chen, U. Gargi, L. Niles, and H. Schütze, “Multi-modal browsing of images in web documents,” *Proc. SPIE Document Recognition and Retrieval*, 1999.
5. M. La Cascia, S. Sethi and S. Sclaroff, “Combining Textual and Visual Cues for Content-based Image Retrieval on the World Wide Web,” *IEEE Workshop on Content-based Access of Image and Video Libraries*.
6. C. Deng, X. He, Z. Li, W. Ma and J. Wen, “Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information”, *In Proceedings of the 12th annual ACM international conference on Multimedia*, pp. 952 – 959, 2004.
7. Y. Deng, B. Manjunath, C. Kenney, M. Moore and H. Shin, “An Efficient Color Representation for Image Retrieval”, *IEEE Transactions on Image Processing*, Vol.10, No.1, pp.140-147, 2001.
8. Flickr, http://www.flickr.com/.
9. Google search engine, http://www.google.com/.
10. C. Jacobs, A. Finkelstein, D. Salesin, "Fast Multiresolution Image Querying", *in Proceedings of Computer Graphics, Annual Conference Series*, pp. 277-286, 1995.
11. G. Miller, R. Beckwith, C. Fellbaum, D. Gross and K. Miller, “Introduction to WordNet: An on-line lexical database”, *International Journal of Lexicography*, Vol. 3, No. 4, pp. 235—312, 1990.
12. S. Paek, C. L. Sable, V. Hatzivassiloglou, A. Jaimes, B. H. Schiffman, S.-F. Chang and K. R. McKeown, "Integration of Visual and Text based Approaches for the Content Labeling and 21 Classification of Photographs," *ACM SIGIR'99 Workshop on Multimedia Indexing and Retrieval*, 1999.
13. K. Rodden, W. Basalaj, D. Sinclair, and K. R. Wood. “Does organisation by similarity assist image browsing?” *In Proceedings of Human Factors in Computing Systems*, 2001.
14. C. Sable and V. Hatzivassiloglou, “Text-based approaches for the categorization of images”, *In Proceedings of the 3rd European Conference on Research and Advanced Technology for Digital Libraries*, pp. 19-38, 1999.
15. S. Santini, A. Gupta and R. Jain, “Emergent Semantics Through Interaction in Image Databases”, Knowledge and Data Engineering, Vol. 13, No. 3, pp. 337 351, 2001.
16. S. Sclaroff, L. Taycher and M. La Cascia, “ImageRover: A Content-Based Image Browser for the World Wide Web,” *IEEE Workshop on Content-based Access of Image and Video Libraries*, TR97-005 06/97.
17. J. Wang, J. Li, G. Wiederhold, “Semantics-Sensitive Integrated Matching for Picture Libraries”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 23, No. 9, pp. 947-963, 2001.
18. J. Wang, G. Wiederhold, O. Firschein, and S. Wei, "Content-based image indexing and searching using daubechies' wavelets," *International Journal of Digital Libraries*, Vol.1, No. 4, pp. 311-328, 1998.
19. K. Yee, K. Swearingen, K. Li and M. Heart, “Faceted Metadata for Image Search and Browsing”, *In Proceedings of the Conference on Human Factors in Computing Systems*, pp. 401-408, 2003.
20. B. Zambrano, R. Singh and B. Bhattarai, "Using Linguistic Models for Image Retrieval", *Proc. International Symposium on Visual Computing, Lecture Notes in Computer Science, Springer Verlag*, 2005.