Using Linguistic Models for Image Retrieval

Brian Zambrano, Rahul Singh, Bibek Bhattarai

bzambran@sfsu.edu, rsingh@cs.sfsu.edu, bdb@sfsu.edu
Department of Computer Science
San Francisco State University
San Francisco, CA 94132

Abstract. This research addresses the problem of image retrieval by exploring the semantic relationships that exist between image annotations. This is done by using linguistic relationships encoded in WordNet, a comprehensive lexical repository. Additionally, we propose the use of a reflective user-interface where users can interactively query-explore semantically related images by varying a simple parameter that does not require knowledge about the underlying information structure. This facilitates query-retrieval in context of the emergent nature of semantics that complex media, such as images have. Experiments show the efficacy and promise of this approach which can play a significant role in applications varying from multimedia information management to web-based image search.

1 Introduction

Since the past decade, image retrieval has been a key area of research within the broader domain of information retrieval. Applications of image retrieval transcend multiple domains including web-based image search, object recognition, motion recognition, personal information management, satellite imagery and bio-medical imaging. The classical approaches to this problem have their origins primarily in signal and image processing and computer vision. A detailed review of the area may be obtained from [1, 14] and references therein. Ultimately, the classical formulations are based on image matching and involve selecting an appropriate image representation such as shape [12], texture [7] or color [15] along with a suitable similarity function. The query is then formulated as a set of constraints on the selected descriptors (such as the percentage of pixels in an image having a specific color(s)), or geometric constrains on image features, or an entire example image. Finally, executing the retrieval requires searching the images in the repository to find those images that satisfy the constraints or maximize the similarity value. In spite of their intuitive theoretical appeal and a large number of academic and commercial efforts [7, 6, 4, 11], the classical approach has had limited success. This can be envisaged by the lack of broadly available image retrieval systems today notwithstanding the ubiquity of image-based information in the digital world. Salient factors leading to this apparent paradox include:

- *The Signal-to-Symbol Barrier:* Queries to an image database are based on the need to satisfy a semantically meaningful information goal. Traditional retrieval formulations however, act only at the signal level. This leads to unsatisfactory performance. For instance, searching for a red colored car on a large image database using the color red typically would yield few meaningful results.

- *Data modeling for large-scale retrieval:* There is a critical paucity of formal and well validated data models for complex media-based information such as images. Most models have typically focused on the development of powerful features to describe the corresponding media and the use of similarity functions to answer queries based on these features [9]. Essentially, these are efforts to store signal-level attributes and do not facilitate bridging the signal-to-symbol barrier. A further element of complexity is introduced by the fact that semantic attributes that can be associated with images are often poorly structured. This implies that the type, number, and very presence of these attributes may cardinally vary from image to image. This is possible even in cases where the images belong to a distinct domain.
- *The emergent nature of image semantics:* The semantics associated with complex media (like video or images) is *emergent*, i.e. media is endowed with meaning by placing it in context of other similar media and through user interactions [10]. This implies that facilitating user interactions and exploratory behavior are essential components of a successful image retrieval approach.

The goal of our research has been to rethink image retrieval in context of the aforementioned issues. We propose using knowledge about the semantic content of an image to facilitate image retrieval. To do so, we utilize WordNet [16], an electronic lexical system to discern the linguistic relationships between descriptions of image content. A semi-structured data model is used to capture any available semantic or media-level information related to specific images. The model is implemented using an XML database to support large scale storage and query-retrieval. Finally, we employ a *reflective* user interface to support user-media interactions. The interface allows immediate visual feedback and response as users vary retrieval parameters. This allows retaining user context and reduces the cognitive load on the user as they engage in query-exploration.

The reader may note that our formulation presupposes the presence of semantic image descriptions or annotations. While a major departure from classical image recognition formulations, the advent of technology increasingly makes such as assumption less limiting than what may appear at the first glance. Indeed, large annotated image databases such as [5] are available today. Further, the WWW contains significant number of mixed text-image pages which forms the basis of keyword indexed image search on the web provided by facilities such as Media RSS (Yahoo), video search and image search (Google), and MSN Video (Microsoft). Finally, advances in the development of integrated annotation systems such as [13, 8] are expected to further ameliorate the process of media (and image) annotation, allowing for broader use of approaches such as the one proposed in this paper.

A series of recent works have attempted to utilize either feature-annotation correlations [3, 18] or linguistic relations within annotations [2, 17] to facilitate image retrieval. While we share the idea of using WordNet-supported relationships with the aforementioned works, the key distinctions of our research lie in conjoining this capability with support for emergent semantics in user-image interactions as well as application of a semi-structured data model to support large scale query-retrieval.

2 A Brief Introduction to WordNet and its Utilization

The design of WordNet has been motivated by psycholinguistic theories of human lexical memory [16]. For a given word or phrase, WordNet provides lexical relationships that include among others, *synsets* (sets of synonyms), *hypernyms*, *hyponyms* (is-a relationships), and *holonyms* (part-of-a relationship). This data is available for up to four parts of speech (noun, verb, adjective and adverb). The word "bat", for instance, can either represent a noun or a verb. For each part of speech, every sense of the word will have the associated lexical information as described above. In our system, we currently focus exclusively on nouns and use the first sense of the word. Further, we use hypernyms which is the "is-a" relationship. As an example, a subset of the hypernym WordNet results for "flamingo" is shown in Figure 1. Hypernyms are returned in a hierarchy with the top level description being the most specific for the search term. Each subsequent description becomes more general. It should be noted that the proposed approach is equally extensible to other relationships that exist in WordNet.

```
flamingo -- (large pink to scarlet web-footed wading bird with down-bent bill; inhabits brackish lakes)
=> wading bird, wader-(any of many long-legged birds that wade in water in search of food)
 => aquatic bird-(wading and swimming and diving birds of either fresh or salt water)
  => bird-(warm-blooded egg-laying vertebrates characterized by feathers and forelimbs modified as wings)
```

Fig. 1. A subset of the hypernym WordNet results for "flamingo". As can be seen, a flamingo is a "wading bird, wader", which is an "aquatic bird", which is a "bird", etc. This hierarchy extends from more specific semantic concepts to less specific concept.

3 System Description

There are three major components to the system: (1) Algorithmic mechanism for computing term similarity, (2) A Semi-structured data storage and query-execution infrastructure, and (3) A front-end reflective GUI for displaying thumbnails, controlling searches, and exploration. Each of these components is described in detail below.

- **The Retrieval Strategy:** In order to calculate the similarity between two terms, we use a weighted score based on the top to bottom comparison of descriptions within the hypernym hierarchy of these terms. After the WordNet results are obtained for a query and annotation term, we step through the query term's hypernym descriptions. If any one of these descriptions is found within the hypernyms of the annotation, the match score is incremented by an amount that is in the range [1, n], where n is the total number of hypernyms of the term, with matches at the top scoring higher (n corresponding to the top of the hierarchy and 1 to the bottom). This score is then normalized by dividing it by the maximal possible match score.

  Formally, let *Q* denotes the query term and *A* the annotation term. Further, for any term *T* let T_k denote the k_{th} term in the hypernym hierarchy of *T*. Let also the predicate *depth(T)* denote the depth of the hypernym hierarchy of the term *T*. Finally, let $depth(T_k)$ indicate the position of the k_{th} term in the hypernym hierarchy of *T*. The similarity between *Q* and *A*, denoted as *S(Q,A)* can be described by formula (1), where the set-theoretic intersection of two terms T_i and V denotes their match/mismatch and takes values in {0,1}.

$$S(Q,A) = \frac{\sum_{i=0}^{depth(Q)} ((depth(Q) - depth(Q_i) \times (Q_i \cap A))}{(depth(Q) \times (depth(Q)+1))/2} \qquad \textbf{(1)}$$

Figure 2 shows results for the query/annotation pair of flamingo/penguin and penguin/flamingo. It is important to note that the direction of our search plays a key role and produces different results. Note that if the query string exactly matches a label, the corresponding image is treated as the closest match. This is done to avoid ties since, for certain label/query combinations, a relative score could be found to be 100 even though the search term did not match exactly. Using the results in Fig. 1, we see that this may happen if we searched for “wader” and we have images labeled as “wader” and “wading bird”. In Fig. 1, we can see that the hypernym hierarchies for both terms would be the same producing similarities of 100. So, if we search for “wader” and we have an image labeled “wader”, the proposed approach ensures that it is returned ahead of any images labeled “wading bird”.

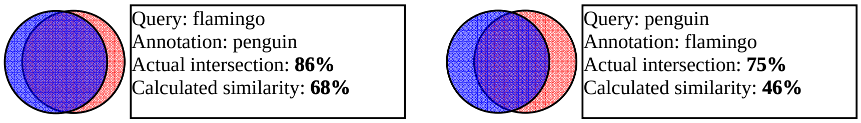
Query: flamingo
Annotation: penguin
Actual intersection: **86%**
Calculated similarity: **68%**
Query: penguin
Annotation: flamingo
Actual intersection: **75%**
Calculated similarity: **46%**

Fig. 2. Intersections of two query(left)/annotation(right) pairs. Although the actual intersections only vary by 11%, the calculated similarities vary by 24%. The difference in weighted scores suggests that a flamingo is more similar to a penguin than a penguin is to a flamingo.

- **Semi-Structured Data Storage**: Berkeley XML is a native XML database which we use for persistent data storage of search scores, image locations and image labels. This provides us with the scalability and efficiency needed for a large-scale image retrieval system. Due to the relatively high expense of a term similarity calculation (two WordNet lookups and the similarity algorithm), we store the score for a given query-label combination in our database. Subsequent searches first lookup into the database to ensure that this operation is only performed once per query-label combination. Queries into Berkeley XML are in the XQuery format which closely follows XPath syntax. Schemas for the internal representation of system data is shown in Fig. 3.
- **Graphical User Interface:** The reflective user interface provides three primary functions: creating and viewing thumbnails, labeling of images and search-exploration capabilities. Once a user has browsed to a folder containing images and created thumbnails, a single keyword can be applied to one or more selected images. A slider bar resides below the label controls and is used for specifying the relevance of search results. Below the slider bar is a text field where the user types the query term and clicks the Search button. During a search, the slider specifies a similarity value which results should be at or above. For example, when the slider bar is at 50, all results will be at least 50% similar to the search term as determined

by the algorithm described above. Thumbnail images are displayed from most to least similar. In addition to partially controlling the initial search result set, the slider bar also provides reflective feedback after the search has been performed. Moving the slider to the right (increasing the similarity) causes images with a similarity value less than the slider bar value to disappear from the displayed results. Likewise, lowering this value will cause potentially new images to appear.

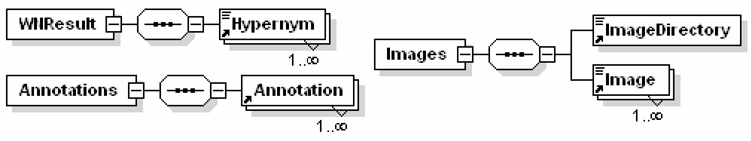

Fig. 3. Schemas for the XML database

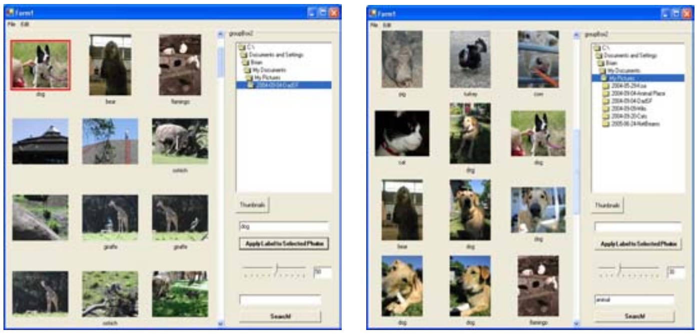
Form1
Form1
Thumbnails
Thumbnails
Apply Label to Selected Photos
Apply Label to Selected Photos
Search!
Search!

Fig. 4. The user interface. The left view shows browsing and labeling support for images. The right view shows results for the query involving the label "animal" at 30% similarity.

4 Experimental Results

Experiments were performed on a collection of 2113 images of which 388 had annotations from a group of 110 unique annotations. Various queries were performed at 10 %, 30 %, 50 %, 70 %, and 90 % similarities. For each query, the time to go through all labeled photos and calculate similarities, if needed, was recorded. Because the similarity value for each query-label pair is made only once before being stored in the database and cached in memory, we report the maximum retrieval time of 31.1s during initialization and the average of 0.28s.

After results were displayed, we manually count the total number of images returned and determine whether they are accurate to find precision and recall. It should be noted that hard boundaries may not exist when determining whether an image is correct for a given query. As an example, the word "chinchilla" as a noun has three senses. The first sense is related to its pelt: "the expensive silvery gray fur of the chinchilla". Sense three is related to a rodent as one might expect: "small rodent with soft pearly gray fur". Because we use the first sense exclusively, a query for "rodent" will not return chinchillas; however chinchillas will show up when searching for “fur” or “pelt”. This is semantically correct, but may be counterintuitive and blurs the boundaries on correctness.

Table 1 shows experimental results for four randomly selected queries at varying similarity thresholds. This dataset provides a fairly consistent picture of the system’s efficacy across a broad range of queries. The standard measures of precision and recall are used to access performances. For all the query terms, the recall is quite high across the range of similarity values. This can be attributed to our use of hypernyms as the basis of our retrieval algorithm. Many annotations were for various animals including dogs, cats, elephants and penguins. When searching for “animal”, we cast a large net and include nearly every image of an animal. Recall for “animal” would have been 100% at all levels had the first sense of “chinchilla” not been related to fur. More constrained or specific terms such as “bird” and “cat” produce more consistent recall across all levels.

Precision is more variable across all search terms. At very low similarities, we see many photos which bare little resemblance to the query term. As we deal exclusively with nouns, almost everything will be somewhat related. This accounts for the low precision at very low similarities. As the threshold value increases, loosely related images fall off. For all queries except “party”, the 100% precision included photos which labels unequal to the query. Results for “bird” included “penguin” and “flamingo”. Likewise, “cat” included “lion”. Some of the images retrieved from the query “animal” are shown in Figure 2.

Table 1. Precision and recall for different queries at various term similarity thresholds

	Similarity Threshold				
	90	70	50	30	10
			Query: cat		
Precision	100	100	40	26.7	9.9
Recall	75	100	100	100	100
			Query: bird		
Precision	100	30	30	25.7	11.1
Recall	77.8	100	100	100	100
			Query: animal		
Precision	100	100	46.9	46.9	26
Recall	85.4	85.4	92.7	92.7	92.7
			Query: friend		
Precision	none	70.8	21	21	8.5
Recall	none	81	81	81	81
			Query: party		
Precision	100	100	100	100	81.8
Recall	27	27	27	27	48.6

An interesting example that illustrates the potentially complex and distal semantic relationships such an approach can unearth is the search for the concept *friend*. As shown in table 1, this example has a large range for precision and relatively low recall. A *friend* is defined as a type of person which causes concepts like *Uncle* and *ostrich* to appear, both of which are people (the first sense of ostrich is related to a person who refuses to face reality).

A significant observation that can be made from these examples is that many of the retrieved results have significant semantic relationship to the query, even though at the image descriptor level, the corresponding images may share low similarities. This is a key asset of such an approach.

Figure 5 graphically shows how precision increases dramatically and the recall tends to decrease as the term similarity value is raised. As we constrain our results by increasing the threshold, very few unrelated images appear (high precision). At the same time, several related photos are not returned (lower recall). Lower values of similarity produce the opposite effect returning a large set of images include nearly all expected images as well as many which are extraneous.

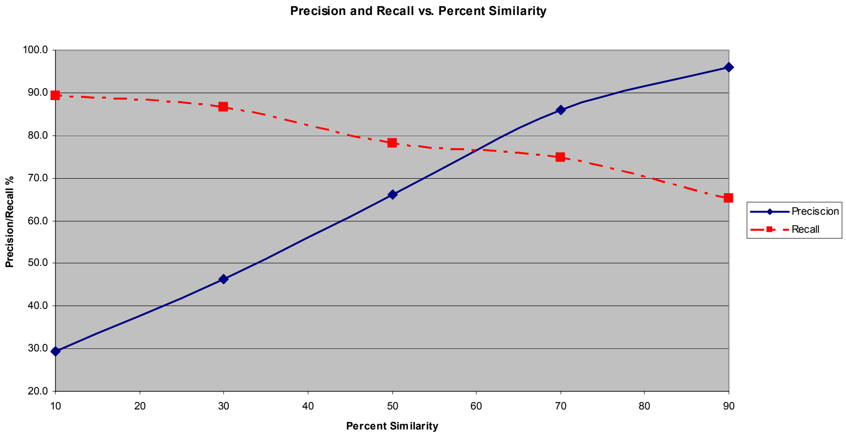

Fig. 5. A plot of precision and recall vs. query/annotation similarity averaged across 20 randomly generated queries.

5 Conclusions

The proposed research addresses three cardinal challenges encountered in query-retrieval of complex media such as images. These include bridging the signal-to-symbol gap, supporting emergent semantics of complex media, and formal storage models that can support the lack of structure in real-world information related to such media. Departing fundamentally from classical research in image matching, we use linguistic relationships amongst descriptions of image content as the key mechanism for information retrieval. We propose an algorithm that allows determining similarity between images using the categorical, hierarchy-based relationship model of WordNet. Further, a reflective user interface allows direct query-exploration of the images. Within it, users can explore images that are semantically proximal or distal to the query. This allows support for the emergent nature of image semantics. Finally, we use a semi-structured data model, implemented through an XML database to support queries over large datasets. Experiments verify the promise of this approach which can play a key role either independently or in conjunction with classical image processing techniques towards solving the complex challenge of image retrieval.

6 References

[1] S. Antani, R. Kasturi, and R. Jain, “A survey on the use of pattern recognition methods for abstraction, indexing and retrieval of images and video”, Pattern Recognition 35(4): 945-965, 2002

[2] Y-A. Aslandogan, C. Their, C. Yu, J. Zou, and N. Rishe, “Using Semantic Contents and WordNet in Image Retrieval”, Proc. SIGIR 1997

[3] K. Barnard and D. Forsyth, “Learning the Semantics of Words and Pictures”, ICCV, Vol. 2, pp. 408 – 415, 2001

[4] M. Flickner, et al. “Query by Image and Video Content: The QBIC System”, IEEE Computer, 28(9): 23-32,1995.

[5] http://www.gettyimages.com

[6] A. Gupta, S. Santini, and R. Jain: In Search of Information in Visual Media. Commun. ACM 40(12): 34-42, 1997

[7] A. Pentland, R. Picard, and S. Sclaroff, “Photobook: Content-Based Manipulation of Image Databases”, Int. J. of Computer Vision, Vol. 18, No. 3, 1996

[8] J. Pinzon and R. Singh, “Designing An Experiential Annotation System For Personal Multimedia Information Management”, IASTED Int. Conf. on Human-Computer Interaction, 2005 (To Appear)

[9] S. Santini and A. Gupta, “Principles of Schema Design for Multimedia Databases”, IEEE Trans. On Multimedia, Vol. 4, No. 2, 2002

[10] S. Santini, A. Gupta, and R. Jain, “Emergent Semantics Through Interaction in Image Databases”, IEEE Trans. On Knowledge and Data Engineering, Vol. 13, No. 3, 2001

[11] http://www.scimagix.com

[12] R. Singh and N. P. Papanikolopoulos, "Planar Shape Recognition by Shape Morphing", *Pattern Recognition*, Vol. 33, No. 10, pp. 1683-1699, 2000

[13] R. Singh, Z. Li, P. Kim, D. Pack, and R. Jain, “Event-Based Modeling and Processing of Digital Media”, *Proc. First ACM SIGMOD Workshop on Computer Vision Meets Databases (CVDB)*, pp. 19-26, 2004

[14] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain: Content-Based Image Retrieval at the End of the Early Years. IEEE Trans. Pattern Anal. Mach. Intell. 22(12): 1349-1380 (2000)

[15] T-F. Syeda-Mahmood, “Data and Model Driven Selection Using Color Regions”, Int. J. of Computer Vision, Vol. 21, 1997

[16] http://wordnet.princeton.edu

[17] C. Yang, M. Dong, and F. Fotouhi, “Learning the Semantics in Image Retrieval – A Natural Language Processing Approach”, Proc. CVPR Workshops, 2004

[18] R. Zhao and W. Grosky, “Narrowing the Semantic Gap – Improved Text-based Web Document Retrieval Using Visual Features”, IEEE Trans. on Multimedia, Vol4, No. 2, pp. 189 – 200, 2002