Context-Driven Image Annotation Using ImageNet

George E. Noel and Gilbert L. Peterson
Department of Electrical and Computer Engineering, Air Force Institute of Technology, Ohio
george.noel@afit.edu; gilbert.peterson@afit.edu

Abstract

Image annotation research has demonstrated success on test data for focused domains. Unfortunately, extending these techniques to the broader topics found in real world data often results in poor performance. This paper proposes a novel approach that leverages WordNet and ImageNet capabilities to annotate images based on local text and image features. Signatures generated from ImageNet images based on WordNet synonymous sets are compared using Earth Mover's Distance against the query image and used to rank order surrounding words by relevancy. The results demonstrate effective image annotation, producing higher accuracy and improved specificity over the ALIPR image annotation system.

Introduction

The widespread adoption of digital recording devices and social media has created a challenging environment for data mining. Unlabeled images are especially problematic with decades of research demonstrating accurate results during controlled testing (Datta et al. 2008), yet decreased performance on real world data (Pavlidis 2009). Typically, both image annotation and Content-Based Image Retrieval (CBIR) tools must choose between approaches that either sacrifice precision for generalized applicability or specialize and fail outside of their limited domains (Müller et al. 2004).

Generalized image annotation methods (Wang, Li, and Wiederhold 2001) (Zhang et al. 2002) (Chen and Wang 2002) are designed to work across a broad spectrum of images but require high intra-category clustering with adequate inter-category separation. As the search space grows, categorical separation becomes challenging. Conversely, specialized annotators often perform well within their domains (Szummer and Picard 1998) (Vailaya, Jain, and Zhang 1998), but require a-priori assumptions about the data that, for a general image set may be incorrect.

Contextual clues may offer a middle ground between generalized and specialized annotators (Wang, Li, and Wiederhold 2001) (Popescu, Moëllic, and Millet 2007). This research presents an algorithm that converts words surrounding an image into synonymous sets (synsets). Synsets related to people (e.g. doctor, philanthropist, etc) are tested using face detectors. All non-people synsets are tested against a set of specialized synset-based signatures. These signatures are generated as needed from the ImageNet (Deng et al. 2009) database using a mixture of color space and frequency features. Images are compared against the synset signatures and each synset rank-ordered based on the signature's similarity to the image, providing potential labels for the image.

Since the focus is to develop methods that work on general data, the test data consists of a wide range of Wikipedia articles (Denoyer and Gallinari 2007). The presented method effectively rank-orders highly-relevant annotations and outperforms ALIPR (Li and Wang 2008) in both word selection and word specificity.

Related Work

Recent work on CBIR can be divided between specialized annotators that limit their domain (Vailaya, Jain, and Zhang 1998) (Szummer and Picard 1998) or generalized annotators (Müller et al. 2004) (Li and Wang 2008) that sacrifice precision for broad applicability. In general annotators, an ensemble of specialized annotators could provide precision while expanding the applicable domains. The challenge is in automatically selecting appropriate specialized annotators.

Contextual clues surrounding images can be used to prompt annotator selection. Most images are embedded in context information of some kind: within a named directory structure, linked in a web page, or embedded within a word processor document. Several have attempted to link words with images, including multi-modal LDA (Barnard et al. 2003) (Blei, Ng, and Jordan 2003). More recent research used LDA to link image blobs with text into a large-scale, parallel infrastructure designed for web image annotation (Liu et al. 2008). All these techniques, however, require a significant number of images and words to effectively extract latent topics (Jeon, Lavrenko, and Manmatha 2003).

Object recognition provides an alternative to word-blob co-clustering and works with fewer data points. Successful techniques use ontology-based object recognition (Schober, Hermes, and Herzog 2004) (Wang, Liu, and Chia 2006) using image features to match to manually-crafted ontologies. Agarwal, et al. (2004) focus on specific elements of an object, such as car wheels or the front grill. Chai, et al. (2008) use edge detection to search for ellipses and quadrangles, identifying and discriminating between both cars and bicy-

cles. These techniques often require highly specialized detectors or carefully crafted ontologies, limiting their domain.

Context-Driven Image Annotation

Our context-driven image processing approach draws on the advantages of both generalized and specialized annotators in a novel way using contextual information surrounding an image. It leverages WordNet (Fellbaum 1998) synonymous sets (synsets) to select appropriate annotations. Wordnet is a lexical database that groups words into distinct concepts called synsets. ImageNet images, organized by synsets, are used to generate signatures that represent common image features associated with the words. ImageNet includes over 14 million images embedded into a hierarchical structure based on WordNet with over 21,000 synsets. Many of these images include bounding boxes and human annotated attributes that provide a valuable resource for image annotation research.

The context-driven method, shown in Figure 1, can be applied to any document corpus with images surrounded by text. For each document, the image caption, paragraph before, and paragraph after the image are extracted. Stopwords are removed and a list of noun synsets are generated for each word. An ensemble director is used to gain higher accuracy by using synset hierarchy to drive the decision tree, as outlined in (Opitz and Maclin 1999). The ensemble detects three categories, graphs or clip-art, people, and all other images, handling each differently. Graphs are first detected and set aside. For remaining images, it selects an annotator based on the root hyperonym of the synset. If the root synset is person related, then the image is sent to a face detector. If a face is detected, that word is marked as potentially applicable. If the image falls under any other root synset, the image is segmented using Efficient Graph-Based Segmentation (Felzenszwalb and Huttenlocher 2004) and the mean vectors calculated. A normalized histogram is generated from these mean vectors, both for the unknown image and a subset of images from the ImageNet synset. The histograms are compared using the Earth Mover's Distance (EMD) and the words rank-ordered from least EMD to most. The resulting words can then be used to annotate the images based on a calculated EMD threshold or fixed number of annotations. There currently is no way to distinguish between words related to people–either faces are detected and all 'people' words are accepted, or no faces are detected and the 'people' words are discarded.

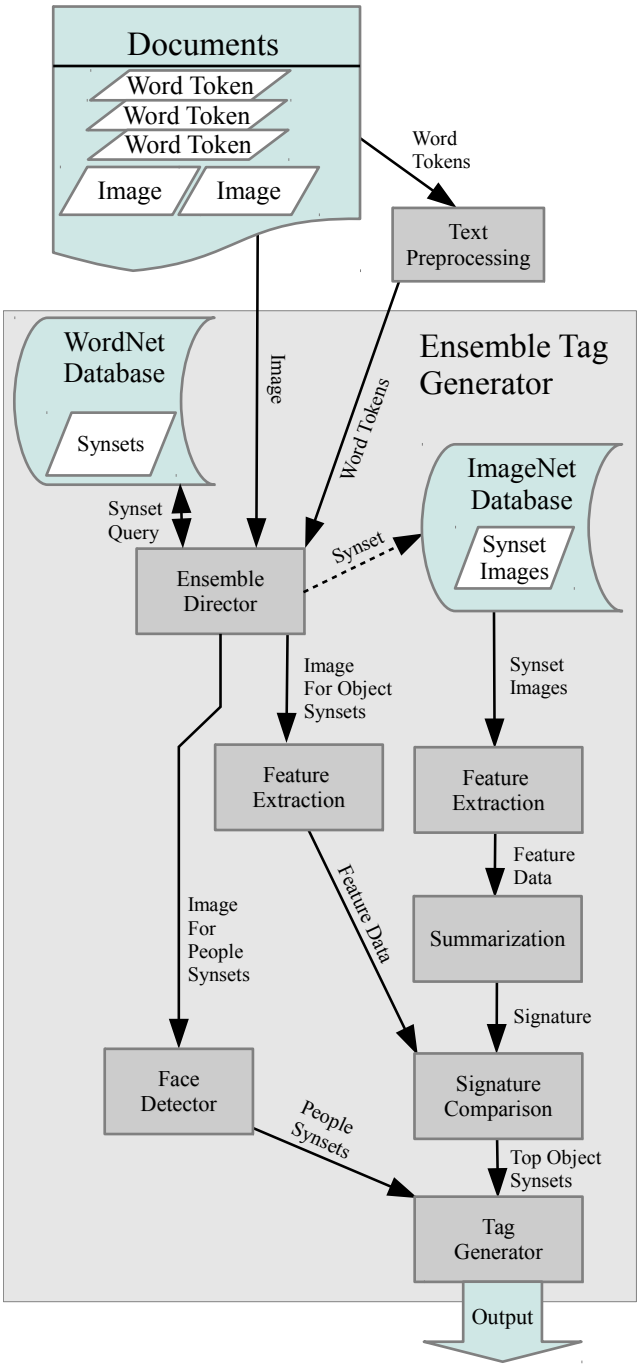

Figure 1: Context-Driven Annotation Process.

Image Signature Generation

This context-driven leverages lexical hierarchies in the WordNet and ImageNet databases. ImageNet provides an assortment of images for select synsets that consist of images from all sub-categories. For instance, the synset for 'dog' contains a wide assortment of dogs while 'great dane' only contains that particular breed. This helps when generating a signature since the signature for 'dog' will be broader than the narrower 'great dane' subset.

Not all of the synsets generated from WordNet have images associated with them. Additionally, only nouns have

associated images even though it may be relevant in some cases to annotate images with adjectives or verbs. For this reason, not every word in the surrounding paragraphs can be rank ordered. Some synsets only have a handful of images associated with them and are also ignored. During initial experimentation, synsets with less than ten images performed poorly so that number was chosen as the minimum threshold. Stemming was deemed not necessary since WordNet already recognizes alternate forms of certain words.

Before a signature is generated, the features are extracted from the image and summarized. A wide variety of color spaces have been used in past research (Smeulders et al. 2000), including the HSV and CIE LAB color space. Comparison testing demonstrated that, as expected, the RGB and HSV color space do a poor job of discriminating between disparate categories. The CIE LUV and CIE LAB color space perform the best. The lightness attribute dominated all others with its discriminative capability. Still, both the A and B frequency attribute of LAB marginally improved the signature matching.

Other features include measuring size, shape, location and texture (Datta et al. 2008). While comparison testing demonstrated size, shape and location discriminated poorly, texture information improved matching accuracy. Extracting information from texture, however, can be difficult since there is little consensus on what numerical features accurately define and distinguish between textures (Srinivasan and Shobha 2008). Based on its proven success in describing global and local features (Li and Wang 2008) (Wang, Li, and Wiederhold 2001), this research uses the Daubechies-4 fast wavelet transform (Daubechies 1992) due to its speed and localization properties, in addition to its granularity. The Daubechies wavelet provides frequency information in the horizontal, vertical, and diagonal directions. Multiple applications of the filter to the resulting approximation coefficients produces lower frequency wavelets with lower spatial resolution. Comparison testing found the best results using the first and second iteration of the Daubechies-4 fast wavelet transform and averaging the horizontal, vertical, and diagonal frequencies for each level.

Once generated, the features are reduced to representative samples to better highlight defining traits of an image set. This is accomplished through mean vectoring of the features within a region. The Efficient Graph-Based Segmentation (Felzenszwalb and Huttenlocher 2004) produced image regions whose vector means represent features common to all images of a particular synset. Efficient Graph-Based Segmentation utilizes a ratio of region size to threshold, merging nodes based on edge gradients. Consistent performance requires parameter tuning when image sizes vary. For this reason, all images are resized to a width of 400 pixels while maintaining the original aspect ratio.

By calculating the region mean vectors for each image, a representative histogram is built for each synset. Any similarity measure must detect general patterns while not being confused by the noise within the domain. The Earth Mover's Distance (EMD) provides a robust comparison algorithm that performs well in related image retrieval testing (Rubner, Tomasi, and Guibas 2000). EMD represents the

Figure 2: Efficient Graph-Based Segmentation with Mean Threshold.

amount and distance of 'earth' or histogram area that must be moved to convert one histogram to the other. This provides a numerical quantification of differences between an image's histogram and a synset's representative histogram. To eliminate noise and help highlight commonly occurring data points, a noise threshold is used to filter the signature. Second, comparison testing determined that using too many images to represent a synset averaged out key traits. Too few histogram buckets lost data while too many created a sparse signature. An image count of between ten and forty appears to produce the most accurate results, while synsets with less than ten images are ignored. A histogram bucket size of 26 best balances detail with sparsity.

People-based and Graph Images

People detection utilizes a boosted cascade of Haar-like features (Viola and Jones 2001) from the OpenCV pre-trained alternate frontal face detector. This accurately detects faces when they are dominantly displayed in the image, though has reduced accuracy with side profiles and small images. This is likely acceptable since when people do not dominate an image, they are less likely to be the focus of an image.

While future versions of this ensemble annotator may annotate clip-art and graphs, the current version does not. It uses a technique adapted from (Popescu, Moëllic, and Millet 2007) to detect graphs using standard deviation around the largest lightness histogram peak. That method, however, results in false positives for black and white images. Hence it was modified to include the average standard deviation of the top three histogram peaks. This prevents a single color from dominating and greatly improved graph detection. Graphs are then discarded as impossible to annotate.

Model Capabilities

This algorithm provides several capabilities over existing image annotation algorithms. First, specific concepts are often represented by a number of words and each word may have several synsets or meanings. For instance, the proper form of the word 'plane' meaning an aircraft or 'plane' meaning a tool for smoothing wood can be identified by sentence context. Accurately mapping to the intended meaning requires complex lexical analysis or a search across each possible meaning. Wordnet includes a hierarchical hyperonym/hyponym relationship between more general and specific forms of a word. This allows the words surrounding an image to provide a robust contextual prompting based on meaning instead of text. Leveraging both WordNet hierarchy with image features has been used effectively in past research (Popescu, Moëllic, and Millet 2007) to group images of placental animals within their hyperonym/hyponym synsets. They compared pixel statistics of known images within the synset and unknowns to accurately select from sets of unlabeled images. As expected, the more specific terms (e.g. Brown Swiss cow) perform better at finding similar images than more generalized terms (e.g. Bovine). Leveraging local context should provide more specific context to compare pre-categorized images against unknowns. While Leong and Mihalcea (2010) used image features from ImageNet to measure word relatedness between synsets based on their image similarity, we are not aware of any similar use of ImageNet to compare against unknown images.

Evaluation

Testing of the algorithm used the 2007 INitiative for the Evaluation of XML Retrieval (INEX) dataset (Denoyer and Gallinari 2007). The INEX dataset includes more than 600,000 documents from Wikipedia, many with embedded captioned images. Each document is labeled with overlapping topic categories, providing a robust mechanism for selecting a range of document topics. In addition, Wikipedia articles tend to be topically related to the images embedded within, helping to highlight poor results as algorithmic problems rather than inconsistent data.

To evaluate the algorithm accuracy, human annotators on Amazon Mechanical Turk (AMT) graded annotation relevancy on a scale of one to five, where 'five' indicates the annotation describes the subject completely and 'one' indicates no correlation between the annotation and image. The definition of each annotation was provided to the AMT workers to remove ambiguity on the intended word form or synset. Five AMT workers were used to score each word, with the average score considered as the standard and disagreement represented by a standard deviation value.

The results are compared with ALIPR (Li and Wang 2008) due to its availability and widespread acceptance in the research community. The top fifteen words produced by ALIPR for each image were sent to the AMT workforce and scored similarly. Unfortunately, ALIPR does not provide an indication of the synset for each word, so each worker had to assume the most relevant synset.

This test measured two things: annotation relevance and specificity. Relevance is measured as the average AMT score of the top five words [1.0-5.0]. Specificity is a measure of how precisely a word describes a concept. Describing a boat as a 'thing' is relevant but too general to be useful. It is calculated as the depth of a word within the WordNet hierarchy. While a rough measure, it provides a means for comparing average word specificity differences between two methods. Since a specific synset is not defined for ALIPR, this research utilizes an average calculation to determine best and worst-case specificity levels across potential synsets. This method is a more general form of that defined by Sakre, et al. (2009). They utilize a function of term weighting based on the number of WordNet senses in a word, the number of synonyms, the hyperonym level, and the number of children (hyponymys/troponyms).

Table 1: Results by Category.

INEX Category	Image Count	Avg Relev.	Avg Spec.
Elephants	18	1.75	9.15
Mountains	260	1.76	8.44
Aircraft	441	2.07	9.59
Dogs	215	2.00	10.10
Skyscrapers	75	2.00	8.75
Sailboats	23	2.07	8.67
Armored Vehicles	47	2.00	9.61
WWII Ships	22	1.82	8.93
Vegetables	56	2.16	9.02
Flowers	34	2.11	9.25

Table 1 lists the categories that were chosen from the INEX dataset. They were selected based on their likelihood of having topically-relevant images and for variety. It includes 6894 total documents with 1136 images. Not every document contained an image and some documents contained multiple images. Graphs were pruned, resulting in the final image numbers by category listed in Table 1.

Table 2: Method Test Results Comparison.

Method	Relevancy	Min Spec.	Avg Spec.	Max Spec.
Context-Driven	1.98	N/A	9.09	N/A
ALIPR	1.62	5.03	6.43	7.85

The image signature consisted of a mean vector containing the three CIE LAB color space elements and Daubechies-4 fast wavelet transform average of the horizontal, vertical, and diagonal frequencies. Frequency is calculated for the high and medium frequencies using one and two applications of the transform respectively. Since this results in higher dimensions, more images must be used. Thirty randomly chosen images were drawn from each synset with a minimum threshold of ten images for the smaller synsets. Twenty-six histogram bins per dimension provided a good balance between resolution and density. Any bins with only one data point were filtered to zero.

The data in Table 2 provides the overall improvement of

Table 3: Sample Results.

Word (in order)	Definition	ALIPR words
cockpit	Area pilot sits	man-made
bomber	Aircraft that drops bombs	indoor
panel	Electrical device	photo
control	Mechanism	old
throttle	Controls throttle valve	people
throttle	Regulates fuel to engine	decoration
engine	Motor	decoy
range	Series of mountains	snow
ocean	Large body of water	ice
limited	Public transportation	winter

the context-driven algorithm over ALIPR. Since ALIPR did not provide the synonymous set for a particular word, we calculated three values. The minimum specificity takes the least specific synset for each word and produces the worst-case scenario. The average of all possible synset specificity values is second while the third calculates the greatest possible specificity, or best possible case. This calculation was unnecessary for the context-driven approach, so those fields in Table 2 are marked as N/A. The data in Table 1 breaks this performance down into the various document categories. The third column provides the average relevance measure for each category while the fourth column indicates specificity.

Some categories performed better than others. Documents about elephants had very difficult images for even humans to analyze, often in the dark, covered in tapestries or lights, or consisting of charcoal drawings. This category had the second highest standard deviations among AMT worker scores, indicating confusion on their part. Vegetables had the highest, demonstrating the difficulty that humans had in determining what kind of label a leafy plant receives. While the elephants were hard for the algorithm to annotate, it had little problems annotating vegetables, likely due to the consistent color and texture among common varieties. The skyscraper category tended to use very generalized words such as tower, structure, and building, accounting for the low specificity. Images of 'WWII Ships' tended to be black and white, often grainy and taken from a distance or with high background noise.

Table 3 illustrates performance of the context-driven algorithm on an image of a B-1B bomber aircraft. Words drawn from the area surrounding the image helped to populate the table and the algorithm rank ordered them based on their image features. The first column provides the ordered list of words generated by this context-driven algorithm, along

Table 4: Sample Results.

Word (in order)	Definition	ALIPR words
dog	Domestic dog	people
human	family Hominidae	man-made
street	Thoroughfare	sport
sign	A public display	car
street	Thoroughfare (variant 2)	cloth
control	Operates a machine	plane
sign	Advertising board	guard
retriever	Dog variant	parade
people	Group of humans	sky
blind	A protective covering	race
dog	Supports for fireplace logs	motorcycle

with a very brief definition in the second column. The definitions had to be abbreviated due to space limitations, but they illustrate the difference between variants of a word. The third column provides the ordered top words generated by the ALIPR algorithm. As anticipated, bomber was near the top while completely unrelated words, such as 'ocean' and 'range' were pushed to the bottom. The word 'cockpit' did appear high on the list, likely due to color and texture similarities between the aircraft image and an actual cockpit.

Table 4 provides example results from one of the more challenging images within the 'dog' category. The algorithm matched color and texture features common to dogs, humans, and a thoroughfare (street) from the ImageNet database and rank-ordered them highest. The alternative definition for the word 'dog', meaning metal supports for logs in a fireplace, was ranked low due to incompatible features.

Conclusion & Future Work

Automated image annotation research has demonstrated the difficulty of achieving precise annotations within a generalized image set. This paper presented a method for annotating images, taking advantage of local image context to drive specialized image annotation algorithms using image features. By using an ensemble image annotation method to separate graphs, people-based images, and all other images, specialized annotators can be applied for each domain. The broad category of images that are not graphs and do not contain people are annotated using a signature-based histogram comparison with Earth Mover's Distance. Each signature is generated using ImageNet images from a particular synset

cued by surrounding text. This algorithm was tested on a relatively diverse assortment of real world data taken from the INEX 2007 data set. Using this context-driven method outperforms ALIPR, one of the more popular image annotation tools within the field.

While the findings of this research were promising, several questions remain. Any useful system would need an acceptably-low false positive rate to avoid flooding the user with spurious results. Currently there is no way of accurately determining where to make the cut for which words apply to the image and which do not. While this research made an attempt to experiment on data representing the chaotic nature of real world data, it still is not as unstructured and unpredictable as those found across the Internet. Further research is required to determine how well this method maps to more complex data.

Acknowledgment

The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

References

Barnard, K.; Duygula, P.; Forsyth, D.; de Freitas, N.; Blei, D. M.; and Jordan, M. I. 2003. Matching words and pictures. *Journal of Machine Learning Research* 3:1107–1135.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022.

Chen, Y., and Wang, J. Z. 2002. A region-based fuzzy feature matching approach to content-based image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24:1252–1267.

Chia, A. Y.; Rajan, D.; Leung, M. K.; and Rahardja, S. 2008. Category-level detection based on object structures. In *16th European Signal Processing Conference*, 1407–1421.

Datta, R.; Joshi, D.; Li, J.; and Wang, J. Z. 2008. Image retrieval: Ideas, influences, and trends of the new age. *ACM Computing Surveys* 40(2):1–60.

Daubechies, I. 1992. Ten lectures on wavelets. In *Society for Industrial and Applied Mathematics*, 284–289.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 785–792.

Denoyer, L., and Gallinari, P. 2007. The wikipedia XML corpus. In *Special Interest Group on Information Retrieval*, 12–19.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Felzenszwalb, P. F., and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *International Journal of Computer Vision* 59:167–181.

Jeon, J.; Lavrenko, V.; and Manmatha, R. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Special Interest Group on Information Retrieval*, 119–126.

Leong, C. W.; Mihalcea, R.; and Hassan, S. 2010. Text mining for automatic image tagging. In *International Conference on Computational Linguistics*, 647–655.

Li, J., and Wang, J. Z. 2008. Real-time computerized annotation of pictures. *IEEE Transactions on Pattern Analysis and Machine Itelligence* 30:985–1002.

Liu, J.; Hu, R.; Wang, M.; Wang, Y.; and Chang, E. Y. 2008. Web-scale image annotation. In *Proceedings of the 9th Pacific Rim Conference on Multimedia: Advances in Multimedia Information*, 663–674.

Müller, H.; Michoux, N.; Bandon, D.; and Geissbuhler, A. 2004. A review of content-based image retrieval systems in medical applications–clinical benefits and future directions. *International Journal of Medical Informatics* 73:1–23.

Opitz, D., and Maclin, R. 1999. Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research* 11:169–198.

Pavlidis, T. 2009. Why meaningful automatic tagging of images is very hard. In *Proceedings of the 2009 IEEE International Conference on Multimedia and Expo*, 1432–1435.

Popescu, A.; Moëllic, P.-A.; and Millet, C. 2007. Semretriev: An ontology driven image retrieval system. In *Proceedings of the 6th ACM International Conference on Image and Video Retrieval*, 113–116.

Rubner, Y.; Tomasi, C.; and Guibas, L. J. 2000. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision* 40:99–121.

Sakre, M. M.; Kouta, M. M.; and Allam, A. M. N. 2009. Weighting query terms using wordnet ontology. *International Journal of Computer Science and Network Security* 9:349–358.

Schober, J.; Hermes, T.; and Herzog, O. 2004. Content-based image retrieval by ontology-based object recognition. In *Workshop on Applications of Description Logics*, 61–67.

Smeulders, A. W.; Worring, M.; Santini, S.; Gupta, A.; and Jain, R. 2000. Content-based image retrieval at the end of the early years. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(12):1349–1380.

Srinivasan, G., and Shobha, G. 2008. Statistical texture analysis. In *Proceedings of World Academy of Science, Engineering and Technology*, 1264–1269.

Szummer, M., and Picard, R. 1998. Indoor-outdoor image classification. In *IEEE International Workshop on Content-Based Access of Images and Video Databases*, 42–51.

Vailaya, A.; Jain, A.; and Zhang, H. 1998. On image classification: City images vs. landscapes. *Pattern Recognition* 31:1921–1935.

Viola, P., and Jones, M. J. 2001. Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 1511–1518.

Wang, J. Z.; Li, J.; and Wiederhold, G. 2001. SIMPLIcity: Semantics-sensitive integrated matching for picture libraries. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23:947–963.

Wang, H.; Liu, S.; and Chia, L.-T. 2006. Does ontology help in image retrieval?: A comparison between keyword, text ontology and multi-modality ontology approaches. In *Proceedings of the 14th annual ACM International Conference on Multimedia*, 109–112.

Zhang, Q.; Goldman, S. A.; Yu, W.; and Fritts, J. E. 2002. Content-based image retrieval using multiple-instance learning. In *International Workship on Machine Learning*, 682–689.