Mind the Gap: Another look at the problem of the semantic gap in image retrieval

Jonathon S. Hare[a], Paul H. Lewis[a], Peter G. B. Enser[b] and Christine J. Sandom[b]

[a]School of Electronics and Computer Science, University of Southampton, UK;
[b]School of Computing, Mathematical and Information Sciences, University of Brighton, UK

ABSTRACT

This paper attempts to review and characterise the problem of the semantic gap in image retrieval and the attempts being made to bridge it. In particular, we draw from our own experience in user queries, automatic annotation and ontological techniques. The first section of the paper describes a characterisation of the semantic gap as a hierarchy between the raw media and full semantic understanding of the media's content. The second section discusses real users' queries with respect to the semantic gap. The final sections of the paper describe our own experience in attempting to bridge the semantic gap. In particular we discuss our work on auto-annotation and semantic-space models of image retrieval in order to bridge the gap from the bottom up, and the use of ontologies, which capture more semantics than keyword object labels alone, as a technique for bridging the gap from the top down.

Keywords: Semantic Gap, Image Retrieval, Automatic Annotation, Ontologies, Cross Language Latent Semantic Indexing

1. INTRODUCTION

At the present time, many of the papers on image retrieval make reference to the problem of the semantic gap. There is a growing awareness in the community of many of the limitations of current retrieval technology and the incompatibility between queries formulated by searchers and the facilities that have been implemented so far in image retrieval systems. Whether in papers by researchers of content based techniques who believe they may be providing a bridge to the semantics or by professional searchers frustrated by the inability of systems to accommodate their queries, the semantic gap appears as a recurring issue in their endeavours.

In a review of the early years of content-based retrieval, Smeulders *et al*[1] define the semantic gap as the "lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation". At the end of the survey the authors conclude that: "A critical point in the advancement of content-based retrieval is the semantic gap, where the meaning of an image is rarely self-evident. ...The aim of content-based retrieval systems must be to provide maximum support in bridging the semantic gap between the simplicity of available visual features and the richness of the user semantics."

Smeulders *et al* also mention another gap of relevance to content based retrieval, the sensory gap, which they define as "the gap between the object in the world and the information in a (computational) description derived from a recording of that scene". Although this is an important issue, we will confine ourselves in this paper to the problem of the semantic gap.

Our aim in this paper is to try and characterise the gap rather more clearly and explore what is and is not being done to bridge it. We begin in Section 2 by defining the gap more carefully to aid later discussion and suggest that it can be divided usefully into a series of smaller gaps between definable representations. In Section 3 we look at queries and their categorisation in order to show how an awareness of the requirements of real searchers can sharpen an understanding of the limiting effects of the gap. In sections 4 and 5 we present some of our own gap bridging work and summarise that of others. In particular, in Section 4, we describe some work on image annotation which attempts to build bridges between low level features and higher level "object" labels: i.e. tackling the gap from the bottom upwards. In Section 5 we argue that ontologies and ideas from emerging

Further author information: E-mail: {jsh02r | phl}@ecs.soton.ac.uk, {p.g.b.enser | c.sandom}@bton.ac.uk

semantic web technology can help to represent and integrate higher-level knowledge about images, potentially capturing more of the semantics than a set of “object” labels alone. In Section 6 we draw some brief conclusions and outline future work.

2. CHARACTERISING THE GAP

The semantic gap manifests itself as a computational problem in image retrieval. The representations one can compute from raw image data cannot be readily transformed to high-level representations of the semantics that the images convey and in which users typically prefer to articulate their queries. It may be useful to look at the series of representations between and including the two extremes. At the lowest level of representation are the raw media, which in this particular case refers to raw images but our analysis is quite general. Content-based retrieval algorithms typically extract feature vectors, or in MPEG 7 parlance, descriptors and these constitute the second level. They may be simple colour histograms, texture statistics or more sophisticated feature vectors developed for content based tasks and may represent parts of an image or the whole image. At a higher level there are representations of “objects” which may be prototype combinations of feature vectors or some other more explicit representation. Once identified, these objects may be given symbolic labels, ideally the names of the objects. This is a simplification as labels may be general or specific e.g. a mountain or Mount Everest. Even where it is possible, labelling all the objects in an image does not typically capture all the semantics. The relationships between the objects as depicted in the image, and the variety of connotations invoked, the implied relationship with the world at large, implied actions, and the broader context, all contribute to the rich high level full semantic representation of the image. The hierarchy of levels between the raw media and full semantics is illustrated in Figure 1.

Needless to say, this is a gross simplification. For example, the objects may have components, with their own labels. But this simple notation is sufficient to enable us to characterise the gap.

The first thing to observe is that the characteristics of the gap vary from one problem to another. There are (rather rare) situations involving simple images where it is possible to pass computationally from the raw image through descriptors to extraction of objects, labels and any required semantics fully automatically. An example might be a robot vision system that can identify parts on a conveyer belt and capture all relevant semantics to use the captured images effectively. But in general the semantic gap starts at the descriptors and goes all the way up to the semantics. In some situations it is possible to extract objects and assign labels but a gap may remain between the labels and the semantics. That is, we may be able to identify the names of the objects in an image but the meaning or significance of the image remains unknown. Our system may be capable of identifying that there are people and buildings in the image but is not able to recognise that this is a demonstration involving police and students. In some cases the required semantics in a query may be expressed directly as a set of object labels but more often the expressed semantics in the query are at a higher level than simply object label lists.

It may be instructive to see the gap in two major sections, the gap between the descriptors and object labels and the gap between the labelled objects and the full semantics.

Two important observations are that firstly, as we will see later, user queries are typically formulated in terms of the semantics and secondly, much of the interesting work which is attempting to bridge the semantic gap automatically is tackling the gap between descriptors and labels and not that between the labels and the full semantics.

The problem of the gap presents itself particularly because, although many image analysis researchers would like queries to be formulated in terms of the descriptors or using the query by example paradigm which can often be reduced to the problem of descriptor matching, most genuine users of image collections formulate their queries at the other side of the gap in terms of the semantics or at best in terms of labels. A number of studies have tried to characterise queries in some formal way and in the next section we review this work as a significant activity, which is taking place to understand the requirements at one side of the gap.

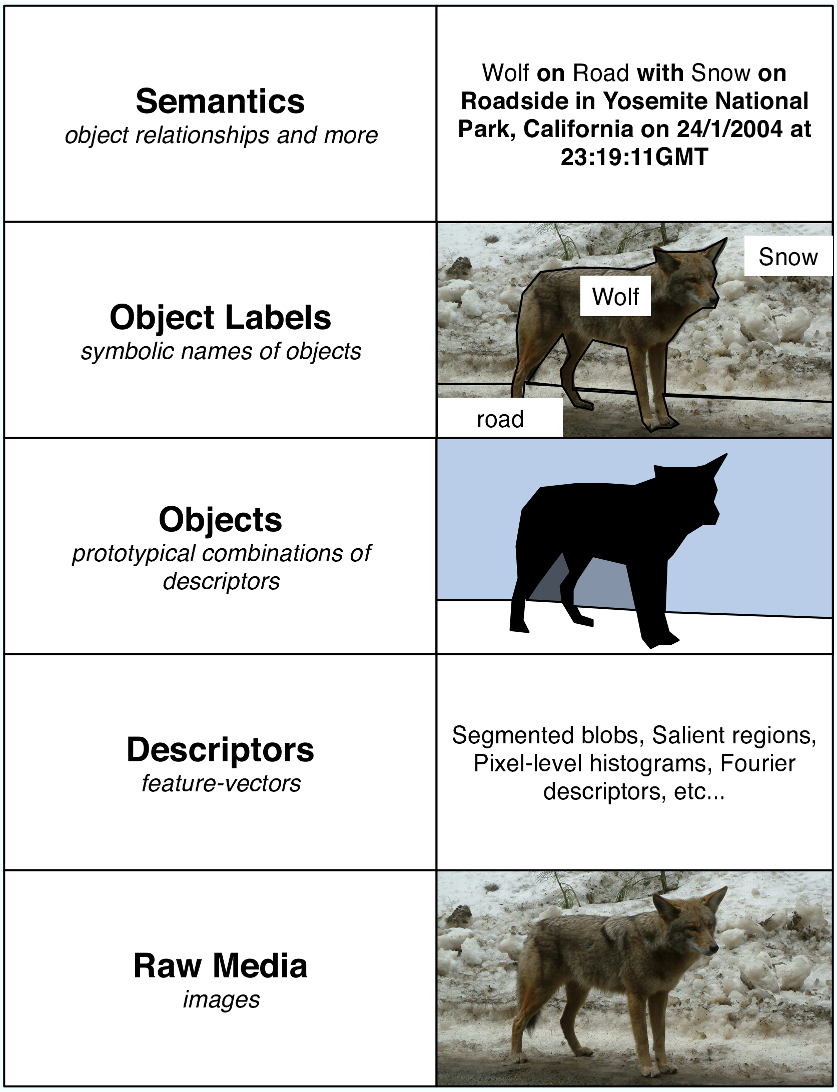
Semantics
object relationships and more
Wolf on Road with Snow on Roadside in Yosemite National Park, California on 24/1/2004 at 23:19:11GMT
Object Labels
symbolic names of objects
Snow
Wolf
road
Objects
prototypical combinations of descriptors
Descriptors
feature-vectors
Segmented blobs, Salient regions, Pixel-level histograms, Fourier descriptors, etc...
Raw Media
images

Figure 1. The Semantic Gap: Hierarchy of levels between the raw media and full semantics.

3. USERS' QUERIES SHOULD BE THE DRIVER

The hallmark of a good image retrieval system is its ability to respond to queries posed by searchers, presented in the desired way. There has been a tendency for much image retrieval research to ignore the issue of user queries and to concentrate on content-based techniques. In spite of this, some investigators have analysed and characterised image queries, providing valuable insights for retrieval system design and highlighting rather starkly the problem of the semantic gap.

One of the earliest investigations of user queries was undertaken by Enser and McGregor[2] who categorised requests in terms of unique/non-unique features, cross-classified by refinement/non-refinement whereby a request is qualified by the addition of temporal, spatial, affective, technical or other facets. Such facets generally serve to locate a query at the high-level, full semantic end of the representation spectrum

Further studies,[3,4] analysed user requests using a tool which recognised the multi-layering of semantic content in both still and moving documentary imagery. This multi-layering has been described in different ways. The art historian Panofsky, working with creative images, identified 'pre-iconographic', 'iconographic' and 'iconologic' levels of expression,[5] which Shatford's generalisation in terms of generic, specific and abstract

Title	Roomy Fridge
Date	circa 1952
Description	An English Electric 76A Refrigerator with an internal storage capacity of 7.6 cubic feet, a substantial increase on the standard model.
Subject	Domestic Life
Keywords	black & white, format landscape, Europe, Britain, England, appliance, kitchen appliance, food, drink, single, female, bending

Table 1. Metadata used for resolving the request of the query 'A photo of a 1950s fridge'.

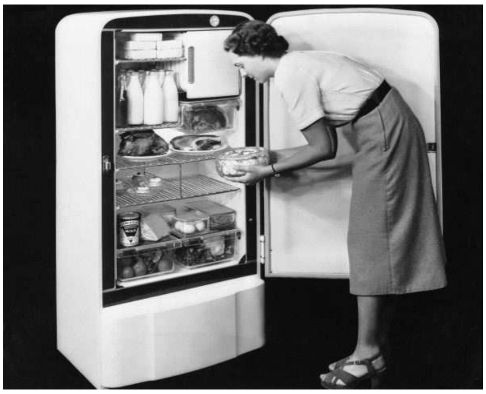

Figure 2. Roomy Fridge ©Getty Images

content, respectively, made amenable to general purpose documentary images.[6] Shatford is more particularly associated with the of-ness and about-ness of image content, the former corresponding with the denotational properties, the latter with connotational properties of visual imagery. Such an approach resonates with the perceptual and interpretive layers of meaning postulated by Jörgensen[7] and with recent classification of queries postulated by Hollink *et al.*[8]

Eakins & Graham[9] offer an alternative three level classification of queries based on primitive features, derived (sometimes known as logical) features and abstract attributes, the latter involving a significant amount of high-level reasoning about the meaning and purpose of the objects or scenes depicted. In our experiences within the realm of real user needs for visual imagery, both still and moving, the incidence of requests based on primitive features is very rare indeed.

Within the particular context of archival imagery, a large proportion of queries typically seek uniquely defined objects; e.g. 'HMS Volunteer'; 'Balshagary School (Glasgow)'; 'Marie Curie'.[2,4] A study of archival moving image requests[3] generated a similar finding, with 68% of the requests including at least one uniquely defined facet; e.g. 'Stirling Moss winning Kentish 100 Trophy at Brands Hatch, 30 August 1968'. Depiction of an event such as this, necessarily invokes the full semantic level because any event is a temporal interpretative relationship between objects. Similarly, it can be argued that the attaching of a label to a place invokes full semantics because a place has to be interpreted as a spatial relationship between objects. In all such cases, detailed textual metadata is necessary in order to represent and recover the full semantic content.

The essential nature of textual metadata is emphasised, furthermore, by the frequent occurrence of requests that address issues of identification, interpretation and significance of depicted features within still images.[10,11]

For example, a request for 'A photo of a 1950s fridge' was resolved using the metadata in Table 1.[12] The corresponding image is shown in Figure 2.

Within the metadata reference is made to a specific manufacturer and model of the depicted object, whilst enabling requests at the more generic levels of 'refrigerator' or 'fridge' and 'kitchen appliance' to be satisfied. Furthermore, the process of identification often involves context, recognition of which would seem to invoke high-level cognitive analysis supported by domain and tacit knowledge (*viz* 'Domestic Life' in the above example).

In general, contextual anchorage is an important role played by textual annotation within the image metadata. The request for a 1950s fridge is an example of query 'refinement' or qualification, moreover, which needs textual annotation for its resolution.

A yet more pressing need for supporting textual metadata occurs when the significance of some visual feature is at issue. Studies of user need have revealed that significance is an important - because frequently encountered - class of request. The problem here is that significance is a non-visible attribute, which can only be anchored to an image by means of some explanatory text. Significance frequently takes the form of the first or last occasion when some visible feature occurred in time, or the first/only/last instantiation of some physical object. Clearly, significance has no counterpart in low-level features of an image. Image retrieval operations that address significance, necessarily involve the resolution of verbalised queries by matching operations conducted with textual metadata.

When the requester's focus of interest lies with the abstract or affective content of the image, wanting images of 'suffering' or 'happiness', for example, appropriate textual cues within the metadata will help to condition our interpretation of the image.

An even more challenging scenario in this context occurs when image searchers specify features that must not be present in the retrieved image; e.g. 'George V's coronation but not procession or any royals'. Provision is sometimes made in controlled keywording schemes to indicate the absence of commonly visible features (e.g., 'no people', 'alone').

The above examples combine to indicate the scale of the challenge faced in trying to overcome the constraints innate within current automatic image indexing and retrieval techniques on their ability to recover appropriate images in response to real expressions of need.

4. IMAGE ANNOTATION AND SEMANTIC SPACES: ATTACKING THE GAP FROM BELOW

By developing systems to automatically annotate image content, we can attempt to identify symbolic labels to apply to the image, or parts of the image. Auto-annotation attempts to bridge the gap between descriptors and symbolic labels by learning which combinations of descriptors represent objects, and what the labels of the objects should be.

The first attempt at automatic annotation was perhaps the work of Mori *et al*,[13] which attempted to apply a co-occurrence model to keywords and low-level features of rectangular image regions. The current techniques for auto-annotation generally fall into two categories; those that first segment images into regions, or 'blobs' and those that take a more scene-orientated approach, using global information. The segmentation approach has recently been pursued by a number of researchers. Duygulu *et al*[14] proposed a method by which a machine translation model was applied to translate between keyword annotations and a discrete vocabulary of clustered 'blobs'. The data-set proposed by Duygulu *et al*[14] has become a popular benchmark of annotation systems in the literature. Jeon *et al*[15] improved on the results of Duygulu *et al*[14] by recasting the problem as cross-lingual information retrieval and applying the Cross-Media Relevance Model (CMRM) to the annotation task. Jeon *et al*[15] also showed that better (ranked) retrieval results could be obtained by using probabilistic annotation, rather than *hard* annotation. Lavrenko *et al*[16] used the Continuous-space Relevance Model (CRM) to build continuous probability density functions to describe the process of generating blob features. The CRM model was shown to outperform the CMRM model significantly. Metzler and Manmatha[17] propose an inference network approach to link regions and their annotations; unseen images can be annotated by propagating belief through the network to the nodes representing keywords.

The models by Monay and Gatica-Perez,[18] Feng *et al*[19] and Jeon and Manmatha[20] use rectangular regions rather than blobs. Monay and Gatica-Perez[18] investigates Latent Space models of annotation using Latent Semantic Analysis and Probabilistic Latent Semantic Analysis, Feng *et al*[19] use a multiple Bernoulli distribution to model the relationship between the blocks and keywords, whilst Jeon and Manmatha[20] use a machine translation approach based on Maximum Entropy. Blei and Jordan[21] describe an extension to Latent Dirichlet Allocation[22] which assumes a mixture of latent factors is used to generate keywords and blob features. This approach is extended to multi-modal data in the article by Barnard *et al*.[23]

Oliva and Torralba[24, 25] explored a scene oriented approach to annotation in which they showed that basic scene annotations, such as 'buildings' and 'street' could be applied using relevant low-level global filters. Hare and Lewis[26] showed how vector-space representations of image content, created from local descriptors of salient regions within an image,[27–29] could be used for auto-annotation by propagating semantics from similar images. Yavlinsky *et al*[30] explored the possibility of using simple global features together with robust non-parametric density estimation using the technique of 'kernel smoothing'. The results shown by Yavlinsky *et al*[30] were comparable with the inference network[17] and CRM.[16] Notably, Yavlinsky *et al* showed that the Corel data-set proposed by Duygulu *et al*[14] could be annotated remarkably well by just using global colour information.

Most of the auto-annotation approaches described above perform annotations in a *hard* manner; that is, they explicitly apply some number of annotations to an image. A *hard* auto-annotator can cause problems in retrieval because it may inadvertently annotate with a similar, but wrong label; for example, labelling an image of a horse with "foal". Jeon *et al*[15] first noted that this was the case when they compared the retrieval results from a fixed-length hard annotator with a probabilistic annotator. Duygulu *et al*[14] attempt to get around this problem by creating clusters of keywords with similar meaning.

Our current approach to auto-annotation[31] is different; Instead of applying *hard* annotations, we have developed an approach in which annotation is performed implicitly in a *soft* manner. The premise behind our approach is simple; a semantic-space of documents (images) and terms (keywords) is created using a linear algebraic technique. Similar documents and/or terms within this semantic-space share similar positions within the space. For example, given sufficient training data, this allows a search for "horse" to return images of both horses and foals because the terms "horse" and "foal" share similar locations within the semantic space. The following subsections describe the approach in brief, and illustrate the performance with results using the Corel data-set proposed by Duygulu *et al.*

4.1. Building a semantic-space: Using linear algebra to associate images and terms

Latent Semantic Indexing is a technique originally developed for textual information retrieval. Berry *et al*[32] described how Latent Semantic Indexing can be used for cross-language retrieval because it ignores both syntax and explicit semantics in the documents being indexed. In particular, Berry *et al* cite the work of Landauer and Littman[33] who demonstrate a system based on LSI for performing text searching on a set of French and English documents where the queries could be in either French or English (or conceivably both), and the system would return documents in both languages which corresponded to the query. The work of Landauer and Littman negates the need for explicit translations of all the English documents into French; instead, the system was trained on a set of English documents and versions of the documents translated into French, and through a process called 'folding-in, the remaining English documents were indexed without the need for explicit translations. This idea has become known as *Cross-Language Latent Semantic Indexing* (CL-LSI).

Monay and Gatica-Perez[18] attempted to use straight LSI (without 'folding-in') with simple cross-domain vectors for auto-annotation. They first created a training matrix of cross-domain vectors and applied LSI. By querying the left-hand subspace they were able to rank an un-annotated query document against each annotation term in order to assess likely annotations to apply to the image. Our approach, described below, is different because we do not explicitly annotate images, but rather just place them in a semantic-space which can be queried by keyword.

Our idea is based on a generalisation of CL-LSI. In general any document (be it text, image, or even video) can be described by a series of observations made about its content. We refer to each of these observations as terms. In order to create a semantic-space for searching images, we first create a 'training' matrix of terms and documents that describe observations about a set of annotated training images; these observations consist of low-level descriptors and observations of which keywords occur in each of the images. This training term-document matrix then has LSI applied to it. The final stage in building the semantic-space is to 'fold-in' the corpus of un-annotated images, using purely the visual observations. The result of this process is two matrices; one representing the coordinates of the terms in the semantic space, and the other representing the coordinates of documents in the space. Similarity of terms and documents can be assessed by calculating the angle between the respective coordinate vectors.

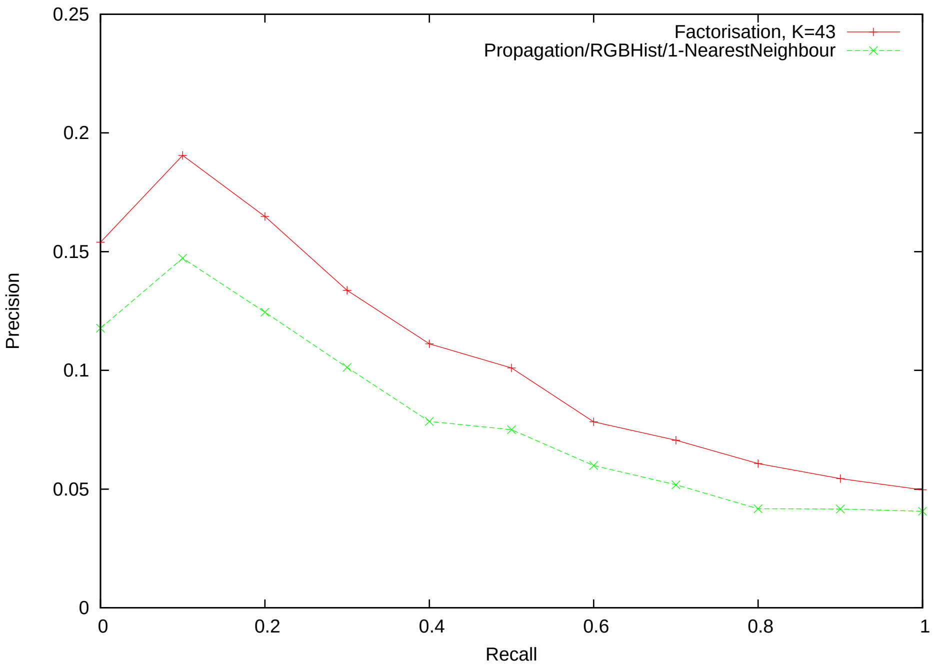

Figure 3. Average Precision-Recall plots for the Corel data-set using RGB-Histogram descriptors for both the CL-LSI and vector-space propagation algorithms.

4.2. Experiments with the Corel data-set

In order to demonstrate the approach described above, we have experimented using the training set of 4500 images and test set of 500 images described by Duygulu *et al.*[14] The visual observations have been kept simple in order to demonstrate the power of the approach; each observation term is a bin from a 64-bin global RGB histogram of the image in question. Because all of the images in the data-set have ground truth annotations, it is possible to automatically assess the performance of the retrieval. By splitting the data-sets into a training set and testing set, it is possible to attempt retrieval for each of the annotation terms and mark test images as relevant if they contained the query term in their annotations. Results from using this technique are presented against results using the 'hard' annotations from the semantic propagation technique.[26]

The overall average precision-recall curves of the CL-LSI and Vector-Space Propagation approaches are shown in Figure 3. As before, the CL-LSI approach outperforms the propagation approach. Whilst the overall averaged precision-recall curve doesn't achieve a very high recall and falls off fairly rapidly, as before, this isn't indicative of all the queries; some query terms perform much better than others. Figure 4 shows a histogram of the R-Precision for the best query terms. Figure 5 shows precision-recall curves for some queries with *good* performance.

Ideally, we would like to be able to perform a direct comparison between our CL-LSI method and the results of the statistical machine-translation model presented by Duygulu *et al*,[14] which has become a benchmark against which many auto-annotation systems have been tested. Duygulu *et al* present their precision and recall values as single points for each query, based on the number of times the query term was predicted throughout the whole test set. In order to compare results it should be fair to compare the precision of the two methods at the recall given in Duygulu2002 *et al*'s results. Table 2 summarises the results over the 15 *best* queries found by Duygulu *et al*'s[14] system (base results), corresponding to recall values greater than 0.4.

Table 2 shows that nine of of the fifteen queries had better precision for the same value of recall with the CL-LSI algorithm. This higher precision at the same recall can be interpreted as saying that more relevant images are retrieved with the CL-LSI algorithm for the same number of images retrieved as with the machine learning approach. This result even holds for Duygulu *at al*'s slightly improved *retrained* result set. This implies, somewhat surprisingly, that even by just using the rather simple RGB Histogram to form the visual observations,

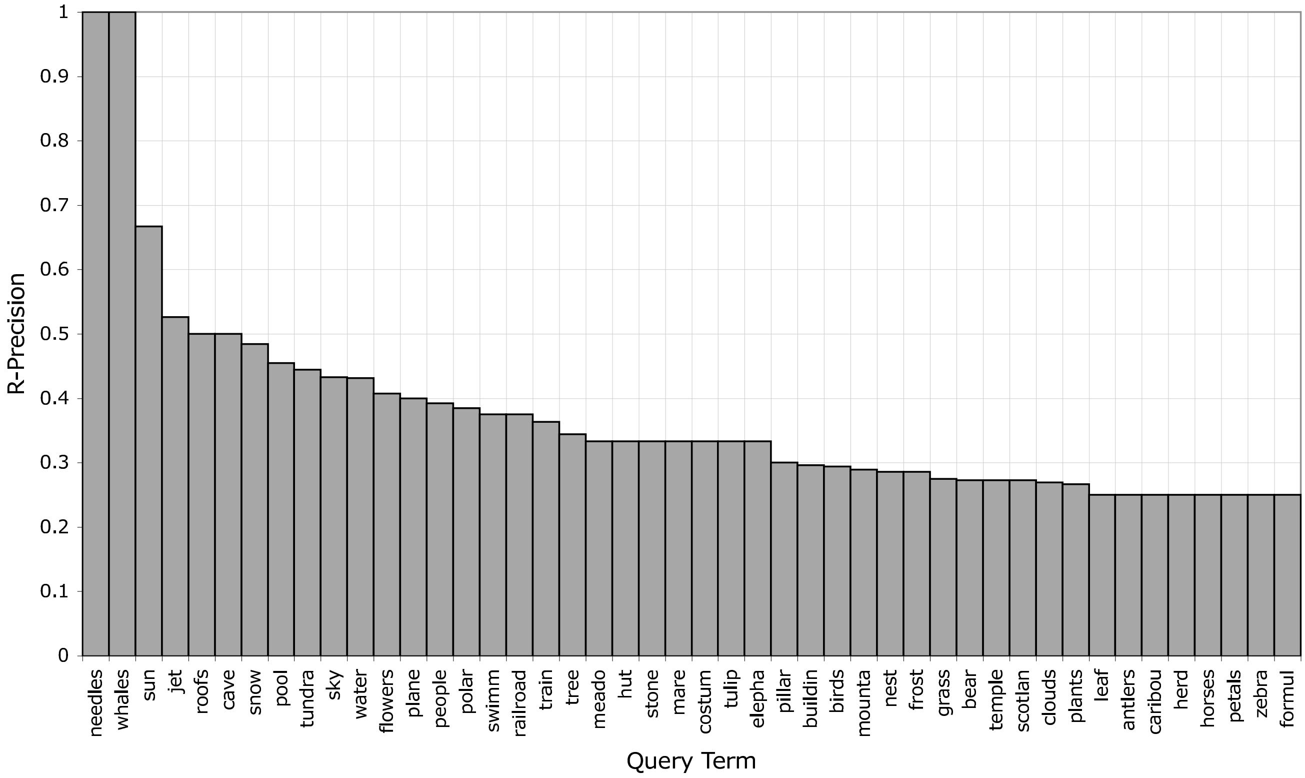

Figure 4. R-Precision of all queries with an R-Precision of 0.25 or above, in decreasing order.

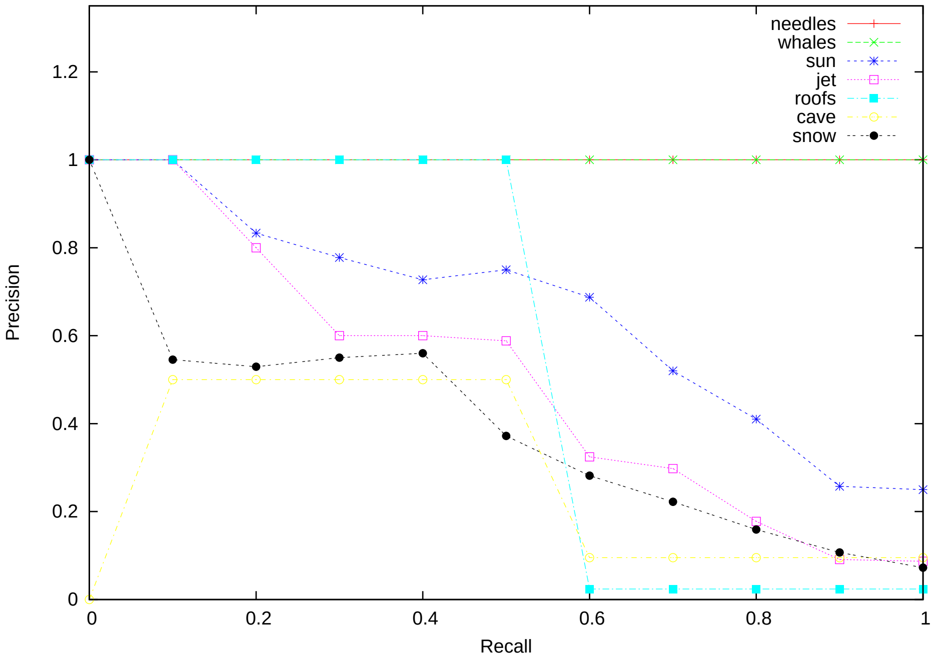

Figure 5. Precision-Recall curves for the top seven Corel queries.

Query Word	Recall	Precision	
		Machine Translation Base Results, th=0	CL-LSI, RGB Histogram, K=43
petals	0.50	1.00	0.13
sky	0.83	0.34	0.35
flowers	0.67	0.21	0.26
horses	0.58	0.27	0.24
foals	0.56	0.29	0.17
mare	0.78	0.23	0.19
tree	0.77	0.20	0.24
people	0.74	0.22	0.29
water	0.74	0.24	0.34
sun	0.70	0.28	0.52
bear	0.59	0.20	0.11
stone	0.48	0.18	0.22
buildings	0.48	0.17	0.25
snow	0.48	0.17	0.54

Table 2. Comparison of precision values for equal values of recall between Duygulu *et al*'s machine translation model and the CL-LSI approach.

the CL-LSI approach performs better than the machine translation approach for a number, of queries. This, however does say something about the relative simplicity of the Corel dataset.[30] Because not all of the top performing results (c.f. Figure 4) from the CL-LSI approach are reflected in the *best* results from the machine translation approach, it follows that the CL-LSI approach may actually perform better on a majority of *good* queries compared to the machine translation model. Of course, whilst the CL-LSI approach may outperform the machine translation approach in terms of raw retrieval performance, it doesn't have the capability of applying keywords to individual segmented image regions that the translation model does.

5. ONTOLOGIES: ATTACKING THE GAP FROM ABOVE

Although automatic image annotation techniques can take us some way across the semantic gap and may enable us to reach the label representation of Section 2, above, as we have shown in Section 3, even a very full set of image labels falls far short of the richness required to represent the full semantics required to describe most images. How might such semantics be represented? The artificial intelligence community has developed many knowledge representation schemes over the years, but recently, the use of ontologies is seen as an increasingly popular way of representing high-level knowledge about application domains. Part of the reason for this increasing interest is the role which ontologies are playing in the emerging semantic web technologies aimed at making web based information understandable by software systems as well as by humans. An ontology is a *shared conceptualisation of a domain* and typically consists of a comprehensive set of concept classes, relationships between them, and instance information showing how the classes are populated in the application domain.

Once knowledge from documents is represented richly in this way several new capabilities are facilitated. First and foremost at least some of the semantics is made explicit and allows queries to be formulated in terms of concepts and their relationship. It is possible to reason over the knowledge domain via the ontology using reasoning software. The ontology can provide a platform for interoperability between systems and a versatile vehicle for browsing and navigating around the document collection.

Although most published work on the use of ontologies has been concerned with textual information, there is increasing interest and research into the use of ontologies with multimedia collections. Some early work on semantic description of images using ontologies as a tool for annotating and searching images more intelligently was described by Schreiber *et al*.[34] More recently his team have extended the approach[35] and also shown how spatial information could be included in the annotations semi-automatically.[36] Jaimes, Tseng and Smith described a semi-automatic approach to the construction of ontologies for semantic description of videos, using associated text in the construction[37] and several authors have described efforts to move the MPEG-7 description of multimedia information closer to ontology languages such as RDF and OWL.[38, 39] Currently, the aceMedia Project[40] is developing a knowledge infrastructure for multimedia analysis, which incorporates a visual description ontology and a multimedia structure ontology.

It is useful to consider ontologies for semantic description of multimedia in two parts, one describing the multimedia content i.e. capturing knowledge about objects and their relationships in the image for example and the other part capturing wider contextual knowledge about the multimedia object, how it was formed, by whom it was created etc.

In the MIAKT project[41, 42] we integrated image annotation tools for region delineation, feature extraction and image analysis with an ontology to capture the semantics associated with the various imaging modalities associated with the breast screening process. The aim of the project was to demonstrate enhanced support at the semantic level for decision making which needs to draw on low level features and their descriptions as well as the related case notes. It also provides a platform for reasoning about new cases on the basis of the semantically integrated set of (multimedia) case histories. By contrast, in the Sculpteur project[43] we mapped museum multimedia object metadata (as opposed to image content) to an ontology based on the CIDOC Conceptual Reference Model in order to provide semantic level navigation and retrieval which could be combined with content based techniques which were also developed in the project.

6. CONCLUSIONS AND FUTURE WORK

In Section 3 we saw how the majority of queries by searchers are presented at the semantic level and in Section 4 we explored image annotation which attempts to bridge part of the gap from below; from the descriptors to the object labels. The use of ontologies as a way of capturing the semantics of multimedia data was explored briefly in Section 5 and if annotations (labels) can be linked automatically into ontology based representations of the semantics, a tentative bridge across the semantic gap begins to emerge. However, current descriptors are inadequate and current annotations and ontologies are far from rich. But on the positive side, multimedia retrieval research is tackling the semantic issue. Eventually approaches to annotation will be coupled with software to discover spatial and other relations between objects in images and more of the semantics will be integrated into the ontological representation automatically to provide a richer platform for the support of semantic level query mechanisms.

In the ‘Bridging the Semantic Gap’ project, funded in the UK by the Arts and Humanities Research Council, we are exploring how well test-bed ontologies, combined with content-based techniques and annotation can meet the real needs of image searchers in limited domains.

ACKNOWLEDGMENTS

The ‘Bridging the semantic gap in visual information retrieval’ project is funded by the Arts and Humanities Research Council (MRG-AN6770/APN17429), whose support together with that of our various contributors, is gratefully acknowledged.

REFERENCES

1. A. W. M. 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), pp. 1349–1380, 2000.
2. P. G. B. Enser and C. G. McGregor, “Analysis of visual information retrieval queries,” in *British Library Research and Development Report*, (6104), p. 55, British Library, London, 1992.
3. C. J. Sandom and P. G. B. Enser, “Virami - visual information retrieval for archival moving imagery.,” in *Library and Information Commission Report 129*, p. 159, *Re*:source: The Council for Museums, Archives and Libraries, 2002.
4. L. H. Armitage and P. G. B. Enser, “Analysis of user need in image archives,” *Journal of Information Sciences* **23**(4), pp. 287–299, 1997.
5. E. Panofsky, *Studies in Iconology: Humanistic Themes in the Art of the Renaissance*, Harper & Rowe, 1962.

6. S. Shatford, "Analyzing the subject of a picture: A theoretical approach," *Cataloguing & Classification Quarterly* **5**(3), pp. 39–61, 1986.
7. C. Jörgensen, *Image Retrieval: Theory and Research : Theory and Research*, Scarecrow Press, Lanham, MA and Oxford, July 2003.
8. L. Hollink, A. T. Schreiber, B. J. Wielinga, and M. Worring, "Classification of user image descriptions," *Int. J. Hum.-Comput. Stud.* **61**(5), pp. 601–626, 2004.
9. J. Eakins and M. Graham, "Content-based image retrieval," Tech. Rep. JTAP-039, JISC, 2000.
10. P. G. B. Enser, C. J. Sandom, and P. H. Lewis, "Surveying the reality of semantic image retrieval," in *8th International Conference on Visual Information Systems, VISUAL2005*, (Amsterdam, Netherlands), July 2005.
11. P. G. B. Enser, C. J. Sandom, and P. H. Lewis, "Automatic annotation of images from the practitioner perspective.," in Leow *et al.*,[44] pp. 497–506.
12. Edina, "Education image gallery." `http://edina.ac.uk/eig`.
13. Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *Proceedings of the First International Workshop on Multimedia Intelligent Storage and Retrieval Management (MISRM'99)*, 1999.
14. P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, pp. 97–112, Springer-Verlag, (London, UK), 2002.
15. J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 119–126, ACM Press, (New York, NY, USA), 2003.
16. V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Advances in Neural Information Processing Systems 16*, S. Thrun, L. Saul, and B. Schölkopf, eds., MIT Press, Cambridge, MA, 2004.
17. D. Metzler and R. Manmatha, "An inference network approach to image retrieval.," in Enser *et al.*,[45] pp. 42–50.
18. F. Monay and D. Gatica-Perez, "On image auto-annotation with latent space models," in *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, pp. 275–278, ACM Press, 2003.
19. S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation.," in *CVPR (2)*, pp. 1002–1009, 2004.
20. J. Jeon and R. Manmatha, "Using maximum entropy for automatic image annotation.," in Enser *et al.*,[45] pp. 24–32.
21. D. M. Blei and M. I. Jordan, "Modeling annotated data," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pp. 127–134, ACM Press, (New York, NY, USA), 2003.
22. D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.* **3**, pp. 993–1022, 2003.
23. K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.* **3**, pp. 1107–1135, 2003.
24. A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," *Int. J. Comput. Vision* **42**(3), pp. 145–175, 2001.
25. A. Oliva and A. B. Torralba, "Scene-centered description from spatial envelope properties," in *BMCV '02: Proceedings of the Second International Workshop on Biologically Motivated Computer Vision*, pp. 263–272, Springer-Verlag, (London, UK), 2002.
26. J. S. Hare and P. H. Lewis, "Saliency-based models of image content and their application to auto-annotation by semantic propagation," in *Proceedings of the Second European Semantic Web Conference (ESWC2005)*, (Heraklion, Crete), May 2005.
27. J. S. Hare and P. H. Lewis, "Salient regions for query by image content.," in Enser *et al.*,[45] pp. 317–325.

28. J. S. Hare and P. H. Lewis, "On image retrieval using salient regions with vector-spaces and latent semantics.," in Leow *et al.*,[44] pp. 540–549.

29. J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, pp. 1470–1477, October 2003.

30. A. Yavlinsky, E. Schofield, and S. Rüger, "Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation," in *Proceedings of the 4th International Conference on Image and Video Retrieval*, D. Polani, B. Browning, A. Bonarini, and K. Yoshida, eds., *Lecture Notes in Computer Science* **3568**, pp. 507–517, Springer-Verlag, (Singapore), July 2005.

31. J. S. Hare, *Saliency for Image Description and Retrieval.* PhD thesis, University of Southampton, 2005.

32. M. W. Berry, S. T. Dumais, and G. W. O'Brien, "Using linear algebra for intelligent information retrieval," Tech. Rep. UT-CS-94-270, University of Tennessee, 1994.

33. T. K. Landauer and M. L. Littman, "Fully automatic cross-language document retrieval using latent semantic indexing," in *Proceedings of the Sixth Annual Conference of the UW Centre for the New Oxford English Dictionary and Text Research*, pp. 31–38, (UW Centre for the New OED and Text Research, Waterloo, Ontario, Canada), October 1990.

34. A. T. G. Schreiber, B. Dubbeldam, J. Wielemaker, and B. Wielinga, "Ontology-based photo annotation," *IEEE Intelligent Systems* **16**(3), pp. 66–74, 2001.

35. L. Hollink, A. T. Schreiber, B. Wielemaker, and B. Wielinga, "Semantic annotation of image collections," in *In Proceedings of the KCAP'03 Workshop on Knowledge Markup and Semantic Annotation*, (Florida, USA), October 2003.

36. L. Hollink, G. Nguyen, A. T. G. Schreiber, J. Wielemaker, B. J. Wielinga, and M. Worring, "Adding spatial semantics to image annotations," in *4th International Workshop on Knowledge Markup and Semantic Annotation at ISWC'04*, 2004.

37. A. Jaimes, B. L. Tseng, and J. R. Smith, "Modal keywords, ontologies, and reasoning for video understanding.," in *CIVR*, E. M. Bakker, T. S. Huang, M. S. Lew, N. Sebe, and X. S. Zhou, eds., *Lecture Notes in Computer Science* **2728**, pp. 248–259, Springer, 2003.

38. J. Hunter, "Adding multimedia to the semantic web: Building an mpeg-7 ontology.," in *SWWS*, I. F. Cruz, S. Decker, J. Euzenat, and D. L. McGuinness, eds., pp. 261–283, 2001.

39. C. Tsinaraki, P. Polydoros, N. Moumoutzis, and S. Christodoulakis, "Coupling owl with mpeg-7 and tv-anytime for domain-specific multimedia information integration and retrieval," in *Proceedings of RIAO 2004*, (Avignon, France), April 2004.

40. I. Kompatsiaris, Y. Avrithis, P. Hobson, and M. Strinzis, "Integrating knowledge, semantics and content for user-centred intelligent media services: the acemedia project," in *Proceedings of Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS '04)*, (Lisboa, Portugal), April 2004.

41. B. Hu, S. Dasmahapatra, P. Lewis, and N. Shadbolt, "Ontology-based medical image annotation with description logics," in *Proceedings of The 15th IEEE International Conference on Tools with Artificial Intelligence*, pp. 77–82, IEEE Computer Society Press, 2003.

42. D. Dupplaw, S. Dasmahapatra, B. Hu, P. H. Lewis, and N. Shadbolt, "Multimedia Distributed Knowledge Management in MIAKT," in *Knowledge Markup and Semantic Annotation, 3rd International Semantic Web Conference*, S. Handshuh and T. Declerck, eds., pp. 81–90, (Hiroshima, Japan), 2004.

43. M. J. Addis, K. Martinez, P. H. Lewis, J. Stevenson, and F. Giorgini, "New Ways to Search, Navigate and Use Multimedia Museum Collections over the Web," in *Proceedings of Museums and the Web 2005*, J. Trant and D. Bearman, eds., (Vancouver, Canada), 2005.

44. W. K. Leow, M. S. Lew, T.-S. Chua, W.-Y. Ma, L. Chaisorn, and E. M. Bakker, eds., *Image and Video Retrieval, 4th International Conference, CIVR 2005, Singapore, July 20-22, 2005, Proceedings*, *Lecture Notes in Computer Science* **3568**, Springer, 2005.

45. P. G. B. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, and A. W. M. Smeulders, eds., *Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. Proceedings*, *Lecture Notes in Computer Science* **3115**, Springer, 2004.