Semantic Facets

An in-depth Analysis of a Semantic Image Retrieval System

Jonathon S. Hare
jsh2@ecs.soton.ac.uk

Paul H. Lewis
phl@ecs.soton.ac.uk

School of Electronics and Computer Science
University of Southampton
Southampton, United Kingdom

Peter G. B. Enser
p.g.b.enser@bton.ac.uk

Christine J. Sandom
c.sandom@bton.ac.uk

School of Computing, Mathematical and Information Sciences
University of Brighton
Brighton, United Kingdom

ABSTRACT

This paper introduces a faceted model of image semantics which attempts to express the richness of semantic content interpretable within an image. Using a large image data-set from a museum collection the paper shows how the facet representation can be applied. The second half of the paper describes our semantic retrieval system, and demonstrates its use with the museum image collection. A retrieval evaluation is performed using the system to investigate how the retrieval performance varies with respect to each of the facet categories. A number of factors related to the image data-set that affect the quality of retrieval are also discussed.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing;
H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Performance Evaluation*

General Terms

Performance, Experimentation, Design, Theory

Keywords

Semantic Facet Model, Semantic Image Retrieval, Image Content

1. INTRODUCTION

Over the past few years there has been a shift away from the idea that content-based retrieval is the solution to all multimedia retrieval needs [30]. A lot of the current research in multimedia retrieval is related to what is known as the semantic gap in multimedia retrieval. In essence, the semantic gap is the gap between the low-level physical features of the media and the much higher level understanding of the semantics of what is depicted by the media [30, 15]. At the present time, many of the papers on image retrieval make reference to the problem of the semantic gap in multimedia retrieval [30, 7, 15]. There is a growing awareness in the community of many of the limitations of current content-based retrieval technology and the incompatibility between queries formulated by searchers and the facilities that have been implemented so far in image retrieval systems [7, 6]. 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.

Techniques for attempting to bridge the semantic gap in image retrieval have mostly used an *auto-annotation* approach, in which keyword annotations are applied to unlabelled images [24, 17, 23, 8, 5, 2, 13]. The basic premise of these automatic annotation approaches is that a model can be learnt from a training set of images that describes how low-level image features are related to higher-level keywords. This model can then be applied to unannotated images in order to automatically generate keywords that describe their content. In essence, the process of auto-annotation is analogous to translating from one language to another [5, 2]. In fact, many of the the state-of-the-art techniques for encoding low-level image content are based around the idea of transforming or quantising the features to a vocabulary of *visual terms*, which represent a purely visual language [29, 12].

One of the problems with current auto-annotation approaches with regards to multimedia retrieval is that they can seriously harm retrieval effectiveness if the annotations they provide are wrong or they employ a limited vocabulary which fails to express the richness of semantic content interpretable within the image. In the following section a faceted representation of semantic content is presented which seeks to address the second of these problems.

CIVR'07, July 9–11, 2007, Amsterdam, The Netherlands.

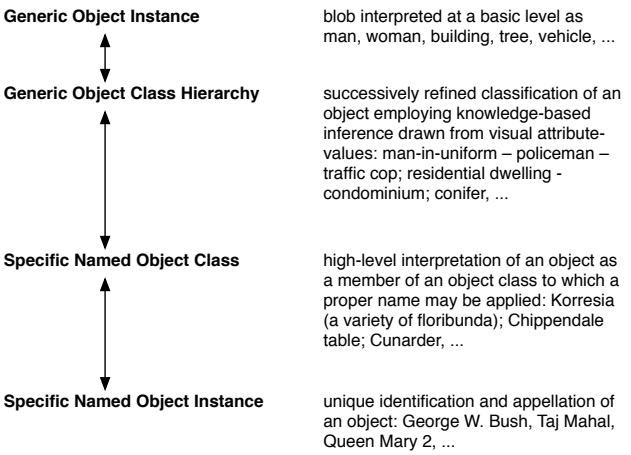


Figure 1: A generic-specific object 'continuum'.

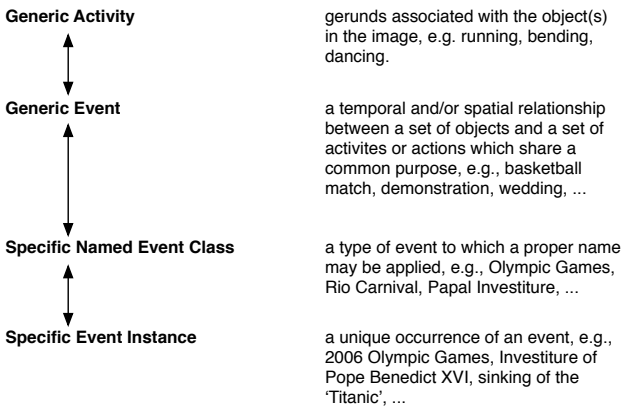


Figure 2: Specification of the activity/event facet in image indexing.

2. THE SEMANTIC FACETS OF AN IMAGE

The work reported in this paper is founded on a faceted model of semantic content which is a significant advance upon those conceptual indexing models previously encountered in the image retrieval literature which classify image semantic content broadly into generic, specific and abstract concepts [27, 1, 16]. Central to our formulation is a combination of *object, spatial, temporal* and *activity/event* facets, advocated by Shatford Layne [28] as key constituents of the 'subject' component in document cataloguing. To these are added *abstract* and *related concept* facets, together with *context* and *topic* facets, which capture the highest level, global semantic content of the image.

The *object* facets are represented in Figure 1 as a hierarchy, founded on a Generic Object Instance, which corresponds with the 'basic categories' theorised by Rosch et al. [25]. By successively finer-grained analysis of an object's attribute-value pairings hypernyms may be identified, culminating in a Specific Named Object Instance uniquely identified by proper name. In some cases such instances are associated with a Specific Named Object Class, membership of which confers a degree of specificity without unique identification of the object, as shown in Figure 1.

At any level within the object hierarchy the *related concept* facet may be encountered. This facet complements an object with those (non-visible) concepts with which the object has a semantic relationship; for example, a picture of St. Paul's Cathedral in London has a relationship with the concept Church of England. In addition, attributes other than those needed to locate an object at a particular level within the object hierarchy may be associated with the object by the addition of adjectives such as elderly, bearded, beautiful, etc.

The *temporal* facet is a continuum, discretizations of which provide a broad spectrum of generic values, both natural ('morning', 'daytime', 'winter',) and artificial ('nanosecond', 'week', 'epoch'), and specific values ('08.23', '31 January 2006', '20th century', 'Victorian', 'Jurassic',). The *spatial* facet encompasses the global scene which provides the context for the image, and an hierarchy of increasing specificity of geographical area, as in the labels Europe, Britain, England, London, Westminster, Victoria Street,

Like the object facet, the *activity/event* facet may be represented in terms of generic-specific classes and instances, as shown in Figure 2.

In combination, the facets described above provide a conceptual framework for representing semantic content in terms of either natural language or a controlled, keyword-based vocabulary, or both. We have employed a data-set comprising nearly 17,000 images drawn from the holdings of the Victoria and Albert Museum (V&A) in London, U.K., as an experimental platform for testing the effectiveness of an automated annotation technique facilitated by a faceted categorisation of the images' keyword-based subject metadata. In this data-set, the context facet is articulated as a natural language description field, and is eliminated from our keyword-based facet analysis.

2.1 The V&A data-set

The V&A data is an eclectic collection of images, from photographs of artefacts such as jewellery, ceramics, toys and statues; fashions items such as costumes, fabric and shoes; two dimensional artworks as diverse as English watercolours, advertising posters, Indian miniatures or stained glass; to manuscripts and early photography.

A vocabulary of more than 12000 keywords is used to represent some of the subject metadata, supplemented by a natural language description and a content title which is a single term descriptor of the physical form of the artefact. There is a wide variation in the number of keywords allocated to each image, ranging from zero to 136, with a mean value of 9.2. This compares with a mean value of 4.6 for the Washington data-set [31]. Our initial analysis of the distribution of keywords over the facet types in the V&A dataset is shown in Table 1.

The dominance of the Generic Object Class Hierarchy is very clear; noticeable also is the relatively frequent occurrence of the Related Concept facet, which associates the semantic content of the image with additional non-visible concepts. A further distinguishing feature of the V&A metadata is the heavy use of adjectives.

The significance of the Related Concept facet is well illustrated in the particular instance of the image shown in Figure 3, where a picture of an unoccupied, disambiguated bed is annotated with a variety of terms, including Ben Jonson', Twelfth Night', inn', and sleeping'. The full set of keywords and their distribution over facet types for this image

Table 1: Distribution of keywords over facets in the V&A data-set

Facet	Number of keywords	% of total vocabulary
Abstract Concept	280	2.2
Generic Activity	428	3.4
Generic Event	166	1.3
Generic Location	169	1.3
Generic Object Class Hierarchy	4238	33.5
Generic Object Instance	134	1.1
Generic Time	93	0.7
Related Concept	1246	9.9
Specific Event Instance	25	0.2
Specific Location Hierarchy	173	1.4
Specific Named Event Class	19	0.2
Specific Named Object Class	170	1.3
Specific Named Object Instance	438	3.5
Specific Time	133	1.1
Topic	95	0.8
Adjectives	2314	18.3

Figure 3: 'Great Bed of Ware'. Image Copyright ©2007, Victoria and Albert Museum, London. All rights reserved.

is shown in Table 2.

3. SEMANTIC RETRIEVAL

In this section we investigate how a semantic image retrieval system, that is able to retrieve unannotated images through textual queries, performs with respect to queries from different facet categories. The section describes our retrieval system, which is used to generate the results from a series of experiments performed with the V&A data-set, described in Section 4.

3.1 The Linear-Algebraic Semantic Space

In our current work we have been using the Linear-Algebraic Semantic Space described in [14] as the basis for investigating how current retrieval techniques work with different image collections. The Semantic Space approach is a generalisation of a text-retrieval technique called Cross Language Latent Semantic Indexing [18], which is itself an extension of Latent Semantic Indexing [4, 3].

Conceptually, the premise behind the 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

Table 2: Facet analysis of the Great Bed of Ware'

Facet	Keywords
Abstract Concept(11)	Charm, comfort, love, passion, privacy, pleasure, romance, secrecy, rest, warmth, wealth
Generic Location (1)	interior
Generic Object Class Hierarchy (66)	acanthus leaf, apple, arch, bed clothes, Bed, bed cover, bed linen, bed post, bedding, bedframe, bedmat, blanket, bolster, border, bunch of grapes, canopy, carving, caryatid, colonnade, column, counterpayne, coverlet, cushion, dome, drape, feather, flock, flower, fringe, fruit, furnishing, furnishing, grape, hanging, headboard, inlay, leaf, linen, lion's mask, marquetry, mattress, mattress, panel, pear, pillow, pillow case, pomegranate, post, rose, rush, sheet, stilted arch, strapwork, swan, tassel, tendril, testa, tester, tightener, trimming, upholstery, valance, vine, wedge, weld, wood
Generic Object Instance (3)	furniture, man, woman
Generic Time (3)	Elizabethan, Renaissance, Tudor
Related Concept (14)	Bedroom, Ben Jonson, Shakespeare, Twelfth Night, dye, hostelry, hotel, head-dress, inn, palace, public house, sleep, sleep, sleeping
Specific Named Object Class (1)	Atlantes
Specific Named Object Instance (1)	Bed of Ware
Specific Time (1)	16th century
Adjectives (34)	Architectural, carved, cinquefoil, dyed, embroidered, enormous, extravagant, famed, famous, fantastic, fantastical, flamboyant, fringed, huge, inlaid, large, luxurious, magnificent, madder, oak, ornate, ostentatious, painted, passionate, private, rich, romantic, secret, solid, trimmed, upholstered, warm, wooden, woven

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.

In general, any document (be it text, image, or even video) can be described by a series of observations, or measurements, made about its content. We refer to each of these observations as terms. Terms describing a document can be arranged in a vector of term occurrences, i.e. a vector whose i-th element contains a count of the number of times the i-th term occurs in the document. There is nothing stopping a term vector having terms from a number of different modalities. For example a term vector could contain term-occurrence information for both 'visual' terms and textual annotation terms. Given a corpus of documents, it is possible to form a matrix of observations or measurements (i.e. a term-document matrix).

Fundamentally, the Semantic Space technique works by estimating a rank-reduced factorisation of a term-document matrix of data, O, into a term matrix T and a document matrix D:

$$\mathbf{O} \approx \mathbf{TD} \ . \tag{1}$$

The two vector bases created in the decomposition form an aligned vector-space of terms and documents. The rows of the term matrix, $\mathbf{T}$, create a basis representing a position in the space of each of the observed terms. The columns of the document matrix, $\mathbf{D}$, represent positions of the observed documents in the space. Similar documents and terms share similar locations in the space.

3.1.1 Application to retrieval of unannotated images

Assume that we have two collections of images; a training set with keyword annotations and a test set without. The content of each image can be represented by a vector of 'visual-term' occurrences. A cross-modality term-document matrix, $\mathbf{O}_{train}$ can be created for the training set of images by combining the visual-term occurrence vector with the keyword-term occurrence vector for each image. This can then be factorised according to Equation 1 into a term matrix $\mathbf{T}_{train}$ and a document matrix $\mathbf{D}_{train}$.

In order to make the unannotated test images searchable, we can project them into the semantic space described by $\mathbf{T}_{train}$ (and $\mathbf{D}_{train}$). Firstly, a cross-modality term-document matrix, $\mathbf{O}_{test}$ must be created for the test set of images by setting the number of occurrences of each (unknown) keyword to 0. It can be shown that it is possible to create a document matrix, $\mathbf{D}_{test}$ for the test documents as follows:

$$\mathbf{D}_{test} = \mathbf{T}_{train}^{\mathrm{T}} \mathbf{O}_{test} \ . \tag{2}$$

In order to query the test set for images relevant to a term, we just need to rank all of the images based on their position in the space with respect to the position of the query term in the space. The cosine similarity is a suitable measure for this task.

3.2 Visual features

In order for the above retrieval strategy to be applied, each image needs to be described as a set of discrete visual terms that can be counted. In most of our current work we have been using a relatively crude, but remarkably effective visual term description based on quantised local descriptors of salient interest regions.

3.2.1 Salient Regions.

In previous work, it has been shown that content-based retrieval based on salient interest points and regions performs much better than global image descriptors [11, 26]. For our algorithm, we select salient regions using the method described by Lowe [20], where scale-space peaks are detected in a multi-scale difference-of-Gaussian pyramid. Peaks in a difference-of-Gaussian pyramid have been shown to provide the most stable interest regions when compared to a range of other interest point detectors [11, 21].

3.2.2 Local Feature Descriptors.

There are a large number of different types of feature descriptors that have been suggested for describing the local image content within a salient region; For example colour moments and Gabor texture descriptors [26]. The choice of local descriptor is in many respects dependent on the actual application of the retrieval system; for example some applications may require colour, others may not. In the current implementation of the algorithm, Lowe's SIFT (Scale Invariant Feature Transform) descriptor [20] is used. The SIFT descriptor was shown to be superior to other descriptors found in the literature [22], such as the response of steerable filters or orthogonal filters, for general matching and retrieval scenarios. The performance of the SIFT descriptor is enhanced because it was designed to be invariant to small shifts in the position of the salient region, as might happen in the presence of imaging noise.

3.2.3 Creating Visual Terms.

One immediately obvious problem with taking local descriptors to represent words is that, depending on the descriptor, there is a possibility that two very similar image patches will have slightly different descriptors, and thus there is a possibility of having a massive vocabulary of words to describe the image. A standard way to get around this problem is to apply vector quantisation to the descriptors to quantise them into a known set of descriptors. This known set of descriptors then forms the vocabulary of 'visual' terms that describe the image. This process is essentially the equivalent of the stemming, where the vocabulary consists of all the possible stems. The next problem is that of how to design a vector quantiser. Sivic and Zisserman [29] selected a set of video frames from which to train their vector quantiser, and used the k-means clustering algorithm to find clusters of local descriptors within the training set of frames. The centroids of these clusters become the 'visual' words representing the entire possible vocabulary. The vector quantiser then proceeded by assigning local descriptors to the closest cluster.

In this work, a similar approach was used. A sample set of images from the data-set was chosen at random, and feature vectors were generated about each salient region in all the training images. Clustering of these feature descriptors was then performed using the batch k-means clustering algorithm with random start points in order to build a vocabulary of 'visual' words. Each image in the entire data-set then had its feature vectors quantised by assigning the feature vector to the closest cluster.

One remarkable thing to note about this approach is that the visual term vocabulary tends to generalise well. This means it is not always necessary to generate a completely new vocabulary for individual image collections [10]. For our

experiments we have used a 3000 term vocabulary generated from a random sample of 100,000 interest regions from images in the Washington Ground Truth Image Data-set [31].

4. EXPERIMENTS AND DISCUSSION

In order to perform experiments with the V&A data-set we first reduced the number of keywords by separating the textual keywords into the facet categories and retaining the keywords that occurred relatively commonly. In total 119 commonly occurring keywords were manually selected from the 15 facet categories, with an average of just under 8 keywords per facet. Two facet categories had only three keywords assigned, however the majority had 9 or 10 keywords. No facet had more than 10 keywords assigned to it. The keyword reduction was performed because it was impractical to analyse the total vocabulary of around 12100 keywords, most of which occurred only one or two times.

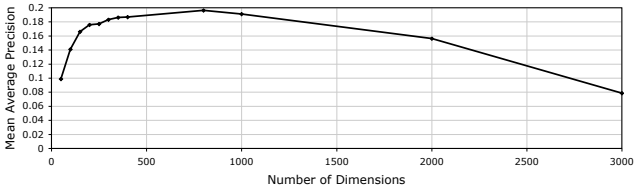
The data-set was trimmed to contain only the images that contained the selected keywords, and was then split randomly into two halves. The random split was performed in such a way as to attempt to keep approximately equal numbers of images representing each keyword in both the training and test sets.

4.1 Building the Semantic Space and choosing the number of dimensions

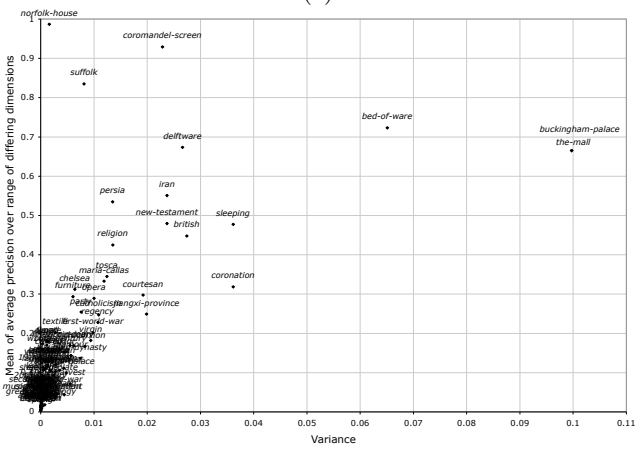
The linear-algebraic technique for building a semantic space described in Section 3.1 relies on a rank-reduced factorisation. The parameter of this factorisation is the target rank, or the number of dimensions of the resultant semantic space. As with previous experiments with the semantic space using other data-sets [14], in order to find the optimal number of dimensions we attempt to optimise the mean average precision (MAP) of retrieval of images using each of the keywords as queries. Figure 4(a) shows how the number of dimensions of the semantic space affects retrieval performance. Of course, as with the other data-sets with which we have previously experimented, the mean average precision doesn't give us the complete picture because the average precisions of each individual keyword query are affected by the number of dimensions in different ways. Figure 4(b) attempts to illustrate this by showing the variance of average precision against the mean value of the average precision over the range of dimensionality, from 50 to 3000 dimensions.

Figure 4(b) illustrates that some queries are much more sensitive to the selection of the number of dimensions than others; the "Buckingham Palace" and "The Mall" queries exhibit a strong sensitivity to the number of dimensions, whilst the "Norfolk House" query is virtually insensitive to the number of dimensions and performs well over the entire range (see Section 4.2 for a more detailed analysis of this query). The large number of queries in the lower-left hand corner of Figure 4(b) indicates that there are many queries that are insensitive to the number of dimensions and do not perform well.

Analysis of the data helps us choose a value for the number of dimensions that trades-off retrieval effectiveness and computational cost (more dimensions requires more work). For the remainder of the results discussed in this paper, 200 dimensions were used in the semantic space, which gives a reasonable trade-off between computational time and retrieval performance.



(a)



(b)

Figure 4: (a)Plot showing variation in retrieval performance as the number of dimensions is increased; (b)Plot showing sensitivity of individual queries to choice of dimensionality.

4.2 General Performance

As with most, if not all, semantic retrieval and automatic annotation systems, the performance of our system varies from query-to-query. The reasons for this are numerous; for example, the choice of visual feature has a large effect — it is rather difficult to accurately learn the colour of an object from grey-level features. Biases and errors in the training data also have an effect on how certain semantic concepts are learnt. An example of these errors is the 'Room' keyword, which is used rather inconsistently. Interestingly, when using the retrieval system to search for images of the 'Room' concept, the first two resulting images depict the interior of a room in a stately home, but were not originally annotated with the 'Room' keyword.

Approximately 8% of all the queries achieved an average precision (AP) of over 0.5, and 52% achieved an AP of at least 0.1. In order to show the range in retrieval performance, some sample queries are shown in Table 3 together with their MAP scores. These queries represent almost the full range of MAP scores achieved by the system.

It is informative to briefly describe some features of the best and worst queries; The "Norfolk House" query corresponds to specific photographic images taken inside a room in a stately home. The "Norfolk House" images fall into two groups that contain many visual similarities, and are visually disparate from the other images in the collection. The training set contains exemplars from both of these groups.

The "Russia" query corresponds to a small number (8 in the test set) of images of very different objects, from posters depicting the 1905 revolution, to jewellery, to a vase. There is little visual similarity between any of the images, hence it

(a)

(b)

Figure 5: Query for "Russia". (a) Sample of the training images; (b) Top 4 unannotated images retrieved by the semantic space. Images Copyright ©2007, Victoria and Albert Museum, London. All rights reserved.

is very difficult for the system to learn how to accurately retrieve images relevant to the term "Russia". It should however be noted that if one looks at the first 20 unannotated images retrieved by the system there are many examples of jewellery (which apparently are not Russian), a French poster in the same *bold-style* as the ones in the "Russia" training images, and even a couple of examples of vases! Figure 5 illustrates a query for "Russia".

Table 3: Some good, and bad queries.

Query	Average Precision
norfolk house	1
buckingham palace	0.8274
the mall	0.8274
bed of ware	0.7355
new testament	0.5784
jesus christ	0.1201
musical instrument	0.0307
russia	0.0028

Facets.

It is also informative to investigate how well the system retrieves images with respect to each of the facet categories described previously. Table 4 shows the mean average precision for each facet, formed by averaging the APs of the queries relevant to the particular facet. Figure 6 shows a histogram of each query, grouped by facet. From Figure 6, it can be seen that within each facet group there is a large variation in retrieval performance, with some queries performing much better than others. Looking at the averages in Figure 4, it is clear that the best retrieval comes from the specific location hierarchy and specific named object class/instance facets. Intuitively, this seems reasonable as it is fair to assume that the visual features that describe images within these facets are themselves quite specific and distinctive.

The retrieval performance of each facet is tightly coupled with the way the images in the collection are annotated. For

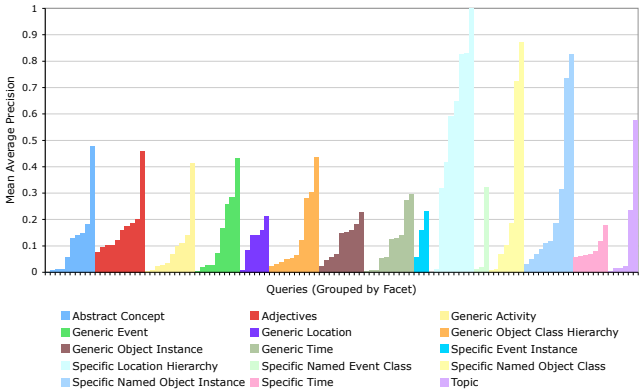


Figure 6: Histogram of average precision scores for all queries, grouped by facet.

example, the MAP of the specific location hierarchy facet is somewhat inflated by the way the images were labelled with locations that are inferred from the image content. The 'Iran' and 'Persia' keywords all correspond to images that depict textiles with distinctive patterns (i.e. Persian Rugs, bags, etc.) that were produced in the Middle East.

4.3 Effect of training set size on retrieval

When investigating the performance of the retrieval system, it is interesting to look at what factors make a particular query work well compared to other queries that work less well. One such factor that is likely to be important is the size of the training set used for learning each particular semantic concept. Figure 7 illustrates this by showing the number of training images for learning each keyword against the average precision for a query with that keyword. The figure shows three salient features; firstly there is a large cluster of points near the origin. These points represent queries that had few training examples and perform badly. Secondly, there is a trend for average precision to increase as the number of training examples increases — this is quite intuitive; the more examples you have of a concept, the easier it is to determine what the salient features that describe the concept are.

Table 4: Mean Average Precision per facet.

Facet	MAP
Abstract Concept	0.12
Generic Activity	0.9
Generic Event	0.14
Generic Location	0.13
Generic Object Class Hierarchy	0.14
Generic Object Instance	0.12
Generic Time	0.11
Specific Event Instance	0.15
Specific Location Hierarchy	0.52
Specific Named Event Class	0.12
Specific Named Object Class	0.28
Specific Named Object Instance	0.25
Specific Time	0.9
Topic	0.15
Adjectives	0.17

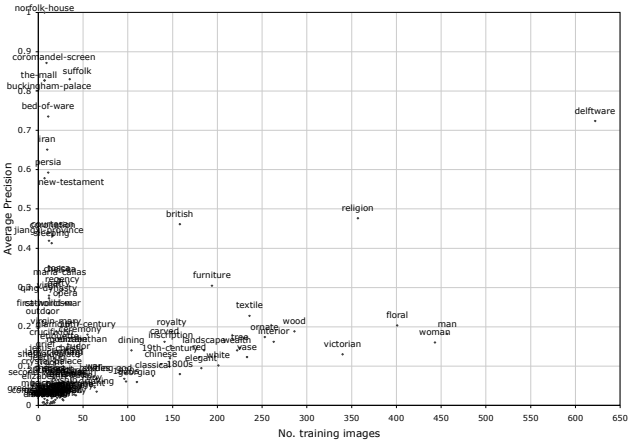


Figure 7: Plot of training-set size versus average precision for each query.

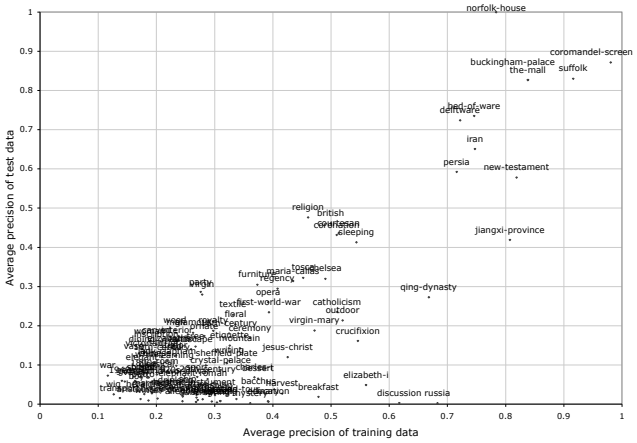


Figure 8: Plot of average precision of training data versus average precision of test data.

Thirdly, there is a distribution of queries with low numbers of training examples, but high retrieval scores. Essentially, these queries are ones that exhibit relatively low intra-class variability, but large inter-class variability; that is is to say that all of the images corresponding to the concept are very visually similar, but also visually distinctive from the other non-relevant images.

4.4 Correlation between training and test set performance

Recall from Section 3.1 that both the training documents are projected into the semantic space during the learning stage of the algorithm. Normally, the training documents are discarded or ignored, and only the test documents are retrieved. However, it is equally possible to retrieve the training documents from the space. It is interesting to investigate whether there is any correlation between the retrieval performance when retrieving training documents and retrieving test documents. Such a correlation would enable the likely retrieval performance of individual queries to be estimated from the training data alone, which could be very useful in real situations where the annotations of the test data is completely unknown (and thus it is impossible to assess performance on the test data).

Figure 8 shows the average precision of the training data plotted against the average precision of the test data for each query. In general, it appears that there is a trend for queries in the training set with high AP to also have a high AP in the test set. There are however a number of outliers; taking the 'Russia' query as an example again, we can see that when retrieving images from the training set, the AP is about 0.7, but with the training set is almost 0. This implies that the retrieval system was able to learn a good representation of 'Russia' from the training data, but that representation did not generalise at all well to the test data. This can be explained by the relatively few training images and the diversity in content of the 'Russia' training images.

5. CONCLUSIONS AND FUTURE WORK

This paper has introduced the idea of a faceted model of semantic content which is a significant advance over previously proposed conceptual indexing models. The facet model has been applied to a large image collection in order to analyse how keyword indexing has been used.

Using the same data-set, an investigation of how well our semantic retrieval technique performs with respect to queries from each of the facet categories has been performed. This investigation has shown that in the data-set used, the best retrieval results come from queries in the specific *object* and *spatial* facets, whilst the other facets, in particular the temporal facets perform less well. The poor precision of the temporal facet is in particular likely to result from the difficulty of determining the ages of a wide variety of objects from purely visual features.

In the future we hope to apply the facet model to different media collections and compare the per-facet performance of different retrieval approaches against our own.

6. ACKNOWLEDGEMENTS

The work reported in this paper has formed part of the 'Bridging the semantic gap in information retrieval' project is funded by the Arts and Humanities Research Council (MRG-AN6770/APN17429), whose support is gratefully acknowledged. We are also grateful to the Victoria and Albert museum for providing the image data used in this work, and to the EPSRC and the Motorola UK Research Laboratory for their support of the initial development of the Semantic Space retrieval technique.

7. REFERENCES

[1] L. H. Armitage and P. G. B. Enser. Analysis of user need in image archives. *Journal of Information Sciences*, 23(4):287–299, 1997.

[2] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan. Matching words and pictures. *J. Mach. Learn. Res.*, 3:1107–1135, 2003.

[3] M. W. Berry, S. T. Dumais, and G. W. O'Brien. Using linear algebra for intelligent information retrieval. Technical Report UT-CS-94-270, University of Tennessee, 1994.

[4] S. C. Deerwester, S. T. Dumais, T. K. Landauer, G. W. Furnas, and R. A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 41(6):391–407, 1990.

[5] 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*, pages 97–112, London, UK, 2002. Springer-Verlag.

[6] P. G. B. Enser, C. J. Sandom, and P. H. Lewis. Automatic annotation of images from the practitioner perspective. In Leow et al. [19], pages 497–506.

[7] P. G. B. Enser, C. J. Sandom, and P. H. Lewis. Surveying the reality of semantic image retrieval. In S. Bres and R. Laurini, editors, *VISUAL 2005*, volume 3736 of *LNCS*, pages 177–188, Amsterdam, Netherlands, July 2005. Springer.

[8] S. L. Feng, R. Manmatha, and V. Lavrenko. Multiple bernoulli relevance models for image and video annotation. In *CVPR (2)*, pages 1002–1009, 2004.

[9] W. I. Grosky and R. Zhao. Negotiating the semantic gap: From feature maps to semantic landscapes. *Lecture Notes in Computer Science*, 2234:33, 2001.

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

[11] J. S. Hare and P. H. Lewis. Salient regions for query by image content. In P. G. B. Enser, Y. Kompatsiaris, N. E. O'Connor, A. F. Smeaton, and A. W. M. Smeulders, editors, *Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. Proceedings*, volume 3115 of *Lecture Notes in Computer Science*, pages 317–325. Springer, 2004.

[12] J. S. Hare and P. H. Lewis. On image retrieval using salient regions with vector-spaces and latent semantics. In Leow et al. [19], pages 540–549.

[13] 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.

[14] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom. A Linear-Algebraic Technique with an Application in Semantic Image Retrieval. In H. Sundaram, M. Naphade, J. R. Smith, and Y. Rui, editors, *Image and Video Retrieval, 5th International Conference, CIVR 2006, Tempe, AZ, USA, July 2006, Proceedings*, volume 4071 of *Lecture Notes in Computer Science*, pages 31–40. Springer, 2006.

[15] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom. Mind the gap. In E. Y. Chang, A. Hanjalic, and N. Sebe, editors, *Multimedia Content Analysis, Management, and Retrieval 2006*, volume 6073, pages 607309–1–607309–12, San Jose, California, USA, January 2006. SPIE.

[16] A. Jaimes and S. F. Chang. A conceptual framework for indexing visual information at multiple levels. In *IS&T/SPIE Internet Imaging*, volume 3964, San Jose, California, USA, January 2000.

[17] 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*, pages 119–126, New York, NY, USA, 2003. ACM Press.

[18] 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*, pages 31–38, UW Centre for the New OED and Text Research, Waterloo, Ontario, Canada, October 1990.

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

[20] D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, January 2004.

[21] K. Mikolajczyk. *Detection of local features invariant to affine transformations*. PhD thesis, Institut National Polytechnique de Grenoble, France, 2002.

[22] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. In *International Conference on Computer Vision & Pattern Recognition*, volume 2, pages 257–263, June 2003.

[23] 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*, pages 275–278. ACM Press, 2003.

[24] 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.

[25] E. Rosch, C. Mervis, W. Gray, D. Johnson, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 8(3):382–439, 1976.

[26] N. Sebe, Q. Tian, E. Loupias, M. Lew, and T. Huang. Evaluation of salient point techniques. *Image and Vision Computing*, 21:1087–1095, 2003.

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

[28] S. Shatford Layne. Some issues in the indexing of images. *Journal of the American Society for Information Science*, 45(8):583–588, 1994.

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

[30] 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):1349–1380, 2000.

[31] University of Washington. Ground truth image database. `http://www.cs.washington.edu/research/imagedatabase/groundtruth/`, Accessed 6/11/2003.