A Linear-Algebraic Technique with an Application in Semantic Image Retrieval

Jonathon S. Hare[1], Paul H. Lewis[1],
Peter G.B. Enser[2], and Christine J. Sandom[2]

[1] School of Electronics and Computer Science, University of Southampton, UK
{jsh2, phl}@ecs.soton.ac.uk
[2] School of Computing, Mathematical and Information Sciences,
University of Brighton, UK
{p.g.b.enser, c.sandom}@bton.ac.uk

Abstract. This paper presents a novel technique for learning the underlying structure that links visual observations with semantics. The technique, inspired by a text-retrieval technique known as cross-language latent semantic indexing uses linear algebra to learn the semantic structure linking image features and keywords from a training set of annotated images. This structure can then be applied to unannotated images, thus providing the ability to search the unannotated images based on keyword. This *factorisation* approach is shown to perform well, even when using only simple global image features.

1 Introduction

Automatic annotation of images has come to the fore as a means of trying to achieve the integration of content-based and text-based image retrieval. An overview of the techniques which have been used in auto-annotation has been provided by Hare et al [1], and we are currently exploring how such techniques can meet the real needs of image searchers in limited domains. This work is being undertaken within the Bridging the Semantic Gap Project, as described by Enser et al [2].

In this paper, we propose a linear algebraic method for *learning* the semantic structure between terms in an annotated training set of images. Unannotated images can then be projected into the structure. The resulting space is unique in that it allows images to be ranked on their relevance to terms that may not have been explicitly assigned to the images, even though the image is relevant to the term.

2 Using Linear-Algebra to Associate Images and Terms

Latent Semantic Indexing (LSI) [3] is a technique in text-retrieval for indexing documents in a dimensionally-reduced semantic vector space. Landauer and Littman [4], demonstrate a system based on LSI for performing text searching

H. Sundaram et al. (Eds.): CIVR 2006, LNCS 4071, pp. 31–40, 2006.

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. Landauer's system negated 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 [5] attempted to use straight LSI 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, based on a generalisation of CL-LSI, is different because we do not explicitly annotate images. The technique works by placing unannotated images in a semantic-space which can be queried by keyword.

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 n documents, it is possible to form a matrix of m observations or measurements (i.e. a term-document matrix). This $m \times n$ observation matrix, $\mathbf{O}$, essentially represents a combination of terms and documents, and can be factored into a separate term matrix, $\mathbf{T}$, and document matrix, $\mathbf{D}$:

$$\mathbf{O} = \mathbf{T}\mathbf{D} \ . \tag{1}$$

These two matrices can be seen to represent the structure of a semantic-space co-inhabited by both terms and documents. Similar documents and/or terms in this space share similar locations. The advantage of this approach is that it doesn't require *a-priori* knowledge and makes no assumptions of either the relationships between terms or documents. The primary tool in this factorisation is the Singular Value Decomposition. This factorisation approach to decomposing a measurement matrix has been used before in computer vision; for example, in factoring 3D-shape and motion from measurements of tracked 2D points using a technique known as Tomasi-Kanade Factorisation [6].

The technique presented here consists of two steps. In the first step, a fully-observed *training* observation matrix is created and decomposed into separate term and document matrices. For example, the observations may consist of both 'visual' terms and annotations from a set of training images. The second step consists of assembling an observation matrix for the documents which are to be indexed. These documents need not be fully observed; for example, they may consist of only ‘visual’ terms. Any unobserved terms are represented by zeros. The document-space of this second observation matrix is then created using the term matrix from the first stage as a basis. The idea behind this is that any term-term relationships that were uncovered in the training stage will be applied to the test data, thus giving the test data *pseudo*-values for the unobserved terms. The net result is that we are left with a new document-space which can be searched by any of the terms used in the training set, even if they were not directly observed in the test set.

2.1 Decomposing the Observation Matrix

Following the reasoning of Tomasi and Kanade [6], although modified to fit measurements of terms in documents, we first show how the observation matrix can be decomposed into separate term and document matrices.

Lemma 1 (The rank principle for a noise-free term-document matrix). *Without noise, the observation matrix,* $\mathbf{O}$*, has a rank at most equal to the number of independent terms or documents observed.*

The rank principle expresses the simple fact that if all of the observed terms are independent, then the rank of the observation matrix would be equal to the number of terms, m. In practice, however, terms are often highly dependent on each other, and the rank is much less than m. Even terms from different modalities may be interdependent; for example a term representing the colour *red*, and the word “Red”. This fact is what we intend to exploit.

In reality, the observation term-document matrix is not at all noise free. The observation matrix, $\mathbf{O}$ can be decomposed using SVD into a $m \times r$ matrix $\mathbf{U}$, a $r \times r$ diagonal matrix $\mathbf{\Sigma}$ and a $r \times n$ matrix $\mathbf{V}^T$, $\mathbf{O} = \mathbf{U\Sigma V}^T$, such that $\mathbf{U}^T\mathbf{U} = \mathbf{VV}^T = \mathbf{V}^T\mathbf{V} = \mathcal{I}$, where $\mathcal{I}$ is the identity matrix. Now partitioning the $\mathbf{U}$, $\mathbf{\Sigma}$ and $\mathbf{V}^T$ matrices as follows:

$$\mathbf{U} = \begin{array}{c}\left[\mathbf{U}_k \middle| \mathbf{U}_N\right] \\ \underbrace{\hphantom{\mathbf{U}_k}}_{k}\underbrace{\hphantom{\mathbf{U}_N}}_{r-k}\end{array} \}m, \quad \mathbf{\Sigma} = \begin{array}{c}\left[\begin{array}{c|c} \mathbf{\Sigma}_k & 0 \\ \hline 0 & \mathbf{\Sigma}_N \end{array}\right] \\ \underbrace{\hphantom{\mathbf{\Sigma}_k}}_{k}\underbrace{\hphantom{\mathbf{\Sigma}_N}}_{r-k}\end{array} \begin{array}{l}\}k \\ \}r-k\end{array}, \quad \mathbf{V}^T = \begin{array}{c}\left[\begin{array}{c} \mathbf{V}_k^T \\ \hline \mathbf{V}_N^T \end{array}\right] \\ \underbrace{\hphantom{\mathbf{V}_N^T}}_{n}\end{array} \begin{array}{l}\}k \\ \}r-k\end{array}, \qquad (2)$$

we have, $\mathbf{U\Sigma V}^T = \mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T + \mathbf{U}_N\mathbf{\Sigma}_N\mathbf{V}_N^T$.

Assume $\mathbf{O}^*$ is the ideal, *noise-free* observation matrix, with k independent terms. The rank principle implies that the singular values of $\mathbf{O}^*$ are at most k. Since the singular values of $\mathbf{\Sigma}$ are in monotonically decreasing order, $\mathbf{\Sigma}_k$ must contain all of the singular values of $\mathbf{O}^*$. The consequence of this is that $\mathbf{U}_N\mathbf{\Sigma}_N\mathbf{V}_N^T$ must be entirely due to noise, and $\mathbf{U}_k\mathbf{\Sigma}_k\mathbf{V}_k^T$ is the best possible approximation to $\mathbf{O}^*$.

Lemma 2 (The rank principle for a noisy term-document matrix). *All of the information about the terms and documents in* $\mathbf{O}$ *is encoded in its* k *largest singular values together with the corresponding left and right eigenvectors.*

We now define the estimated noise-free term matrix, $\hat{\mathbf{T}}$, and document matrix, $\hat{\mathbf{D}}$, to be $\hat{\mathbf{T}} \overset{\text{def}}{=} \mathbf{U}_k$, and, $\hat{\mathbf{D}} \overset{\text{def}}{=} \mathbf{\Sigma}_k \mathbf{V}_k^T$, respectively. From Equation 1, we can write

$$\hat{\mathbf{O}} = \hat{\mathbf{T}}\hat{\mathbf{D}}, \tag{3}$$

where $\hat{\mathbf{O}}$ represents the estimated noise-free observation matrix.

Interpreting the Decomposition. The two vector bases created in the decomposition form an aligned vector-space of terms and documents. The rows of the term matrix create a basis representing a position in the space of each of the observed terms. The columns of the document matrix represent positions of the observed documents in the space. Similar documents and terms share similar locations in the space.

2.2 Using the Terms as a Basis for New Documents

Theorem 1 (Projection of partially observed measurements). *The term-matrix of a decomposed fully-observed measurement matrix can be used to project a partially observed measurement matrix into a document matrix that encapsulates estimates of the unobserved terms.*

Manipulating Equation 3 gives us a method of projecting a partially-observed observation matrix, $\mathbf{P}$ into the basis created by the term matrix, $\hat{\mathbf{T}}$. The underlying assumption is that if we were to project the original fully-observed observation matrix (i.e. $\mathbf{P} = \hat{\mathbf{O}}$), then we should get the same document basis.

$$\begin{aligned} \mathbf{P} &= \hat{\mathbf{T}}\hat{\mathbf{D}} \\ \therefore \hat{\mathbf{D}} &= \hat{\mathbf{T}}^{-1}\mathbf{P} = \hat{\mathbf{T}}^T\hat{\mathbf{T}}\hat{\mathbf{T}}^{-1}\mathbf{P} = \hat{\mathbf{T}}^T\mathbf{P} \end{aligned} \tag{4}$$

Therefore, to project a new partially observed measurement matrix into a basis created from a fully observed training matrix, we need only pre-multiply the new observation matrix by the transpose of the training term matrix. The columns of this new document matrix represent the locations in the semantic space of the documents. In order to query the document set for documents relevant to a term, we just need to rank all of the documents 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.

Thus far, we have ignored the value of k. The rank principle states that k is such that all of the semantic structure of the observation matrix, minus the noise is encoded in the singular values and eigenvectors. k is also the number of independent, un-correlated terms in the observation matrix. In practice, k will vary across data-sets, and so we have to estimate its value empirically.

3 Experimental Results

In this section, we present experiments using real images from both the Washington data-set [7] and the Corel data-set proposed in [8]. Because all of the images in these data-sets 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 keyword terms and mark test images as relevant if they contained the query term in their annotations. Results from this technique are presented against results using the *hard* annotations from the vector-space propagation technique described in [9].

3.1 Experiments with the Washington Data-Set and SIFT 'Visual' Terms

We split the Washington data-set [7] into a training set of 349 images, and a test set of 348 images. Each of the images was indexed using 'visual' terms from quantised local SIFT descriptors about interest points picked from peaks in a difference-of-Gaussian pyramid [9, 10]. The size of the visual vocabulary was fixed to 3000 terms [10].

Choosing a Good Value for *k*. In order to select a value for k, we need to try and optimise the retrieval. A good statistic of overall retrieval performance is the Mean Average Precision (MAP). Plots of the average precision versus varying values of k for four different queries in the test set are shown in Figure 1. A plot of the MAP over all possible queries in the training set, is shown in Figure 2.

Figure 1 shows that there is a very large amount of variation of average precision across different queries. This is in a large part due to biases in both the training set of images and in the test set. For example, both the training set and test set contain an approximately equal number of images of a football stadium, however, the number of *stadium* images in the training set is quite large in comparison to many of the other queries. The net effect is that the "Stadium" query is particularly well trained. Well trained queries can also result from few training images when the training image is sufficiently visually dissimilar to the other images (i.e. it contains a fairly unique combination of visual terms).

Unfortunately, Figure 2 doesn't show a peak from which to select a good value of k, instead it is asymptotic to a mean average precision of about 0.38. However, given the constraint that we want to choose k such that it is the smallest it can be whilst still giving good retrieval, we chose a value of $k = 100$ for the following experiments.

Overall Retrieval Effectiveness. The overall retrieval effectiveness of the technique is characterised in Figure 3. As can be seen, the factorisation approach outperforms the propagation approach at all values of recall.

The precision-recall curves in Figure 3 don't truly reflect the whole performance of the approach because certain queries are better performing than others. Figure 5 illustrates this by showing the average precision for each of the queries, sorted by decreasing precision. For clarity, only queries yielding an average precision of above 0.5 are shown.

Example: Querying for "Bridge". We now take an example query using the term "Bridge" to investigate the performances of the approaches in more detail.

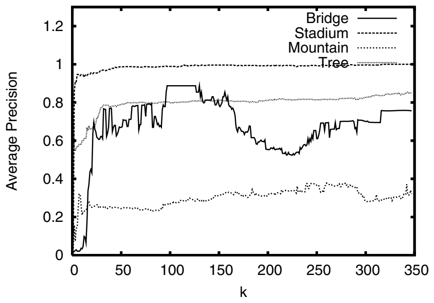

Fig. 1. The effect of k on average precision for four different queries

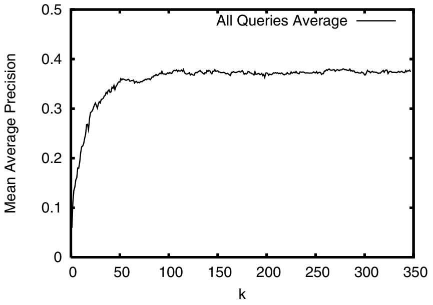

Fig. 2. The effect of k on the Mean-Average Precision over all queries

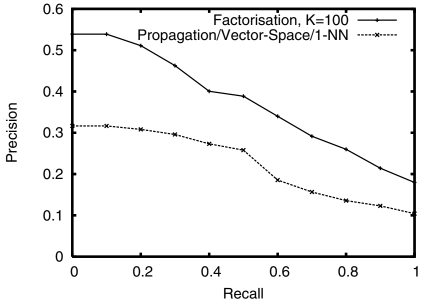

Fig. 3. Average precision-recall curves for the different algorithms over all queries

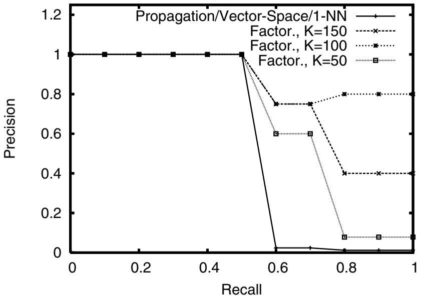

Fig. 4. Precision-Recall curves for querying with the keyword "Bridge"

There are ten occurrences of the annotation keyword "Bridge" in the Washington data-set. Of these ten occurrences, four images are in the test set and six in the training set. One of the training images has been labelled with "Bridge", although it doesn't actually appear to contain a bridge. This mislabelling of images corresponds to noise, and the algorithms need to be robust to noise within the data-set. The training images are shown in Figure 6. Figure 4 illustrates the effect on precision over different recall values using both the Factorisation algorithm and the vector-space propagation algorithm. Three different values of k for the factorisation algorithm are shown in the figure. The precision recall curves show that both of the algorithms exhibit perfect precision up to recall values of 0.5, but then tend to drop off.

Figure 7 shows the test images containing the "Bridge" keyword, along with the rank-position of the images using the factorisation and propagation techniques. The images were retrieved in the same order by the two algorithms, however, the positions at which they occur varies greatly. The factorisation approach retrieved all four relevant images within the top five images, whilst the propagation approach didn't achieve full recall until 332 images had been retrieved.

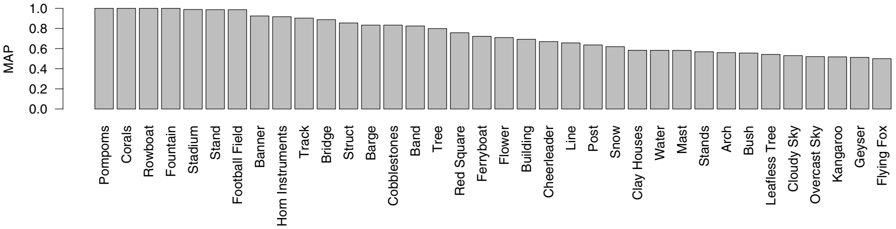

Fig. 5. Average precision of all queries with precision > 0.5, sorted by decreasing precision

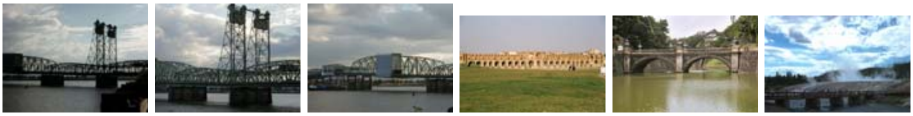

Fig. 6. Training images containing the "Bridge" keyword

3.2 The Corel Data-Set

In the previous subsection we proposed using SIFT visual terms to model the image content. However, this is not the only option; the observation matrix could conceivably contain observations of any type of feature. In order to demonstrate the power of the factorisation technique, we use a much simpler feature; a 64-bin global RGB histogram. We use the training set of 4500 images and test set of 500 images from the Corel data-set described in [8].

Following the methodology for optimising k based on MAP described previously, we set $k = 43$. Overall averaged precision-recall curves of the factorisation and propagation approaches are shown in Figure 8. As before, the factorisation approach outperforms the propagation approach. Whilst the overall averaged precision-recall curve doesn't achieve a very high recall and falls off fairly rapidly, this isn't indicative of all the queries; some query terms perform much better than others. Figure 9 shows precision-recall curves for some queries with *good* performance.

Ideally, we would like to be able to perform a direct comparison between our factorisation method and the results of the statistical machine-translation (MT)

			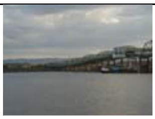	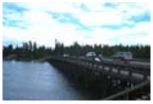
Factorisation (k=100)	1	2	3	5
Vector-space Prop. (1NN)	1	2	125	332

Fig. 7. Test Images and the rank-order in which they were retrieved by the two algorithms

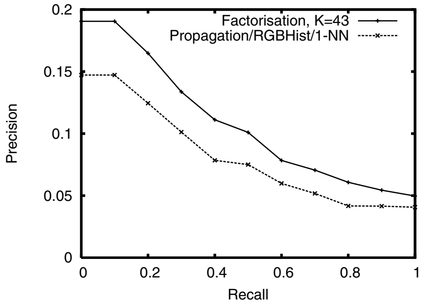

Fig. 8. Average Precision-Recall plots for the Corel data-set using RGB-Histogram descriptors for both the factorisation and propagation algorithms

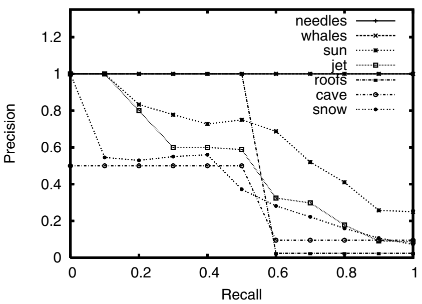

Fig. 9. Precision-Recall curves for the top seven Corel queries using factorisation ($k = 43$)

model presented by Duygulu et al [8], 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 the MT results. Table 1 summarises the results over the 15 *best* queries found by the MT system (base results), corresponding to recall values greater than 0.4.

Table 1 shows that nine of of the fifteen queries had better precision for the same value of recall with the Factorisation algorithm. This higher precision at

Table 1. Comparison of precision values for equal values of recall between the machine translation model [8] and the factorisation approach

Query Word	Recall	Precision	
		Machine Translation Base Results, th=0	Factorisation, 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

the same recall can be interpreted as saying that more relevant images are retrieved with the factorisation algorithm for the same number of images retrieved as with the machine learning approach. This result even holds for Duygulu et 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, the factorisation 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 [11]. Because not all of the top performing results from the factorisation approach are reflected in the *best* results from the machine translation approach, it follows that the factorisation approach may actually perform better on a majority of *good* queries compared to the machine translation model.

4 Conclusions and Future Work

This paper presented a novel approach to building a semantic space for image retrieval using a linear algebraic factorisation. Performance of the technique is good, even when using a simple global image feature such as the RGB histogram. The approach is exciting because it models the semantic gap between image descriptors and keywords in a flexible way. The factorisation technique does not produce equal performance for all queries. The reasons for this are most likely two-fold; firstly, the visual features used to represent the image may not have been sufficient to represent the keyword. Secondly, the training data may not have been sufficient to learn a good representation for the term. In terms of the Corel data-set using RGB histogram features, the factorisation approach works particularly well with annotations that can be described globally across the image by colour alone. For example, searching for 'sun' returns images with many warm yellow tones, and searching for 'snow' returns images with lots of white colours.

More experimentation needs to be performed to investigate the performance of the factorisation approach. In particular, it would be interesting to use the image descriptors created by [8] to build our observation matrix, and then to directly compare retrieval results with other automatic annotation approaches. It would also be interesting to investigate the scalability of the approach.

In the 'Bridging the semantic gap' project, we aim to test this approach more extensively with picture librarians, in an attempt to establish its ability as a system offering the potential of semantic search.

Acknowledgements

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 EPSRC and Motorola UK Research Laboratory for their support of parts of this work.

References

[1] Hare, J.S., Lewis, P.H., Enser, P.G.B., Sandom, C.J.: Mind the gap. In Chang, E.Y., Hanjalic, A., Sebe, N., eds.: Multimedia Content Analysis, Management, and Retrieval 2006. Volume 6073., San Jose, California, USA, SPIE (2006) 607309–1–607309–12

[2] Enser, P.G.B., Sandom, C.J., Lewis, P.H.: Surveying the reality of semantic image retrieval. In Bres, S., Laurini, R., eds.: VISUAL 2005. Volume 3736 of LNCS., Amsterdam, Netherlands, Springer (2005) 177–188

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

[4] Landauer, T.K., Littman, M.L.: 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, Waterloo, Ontario, Canada (1990) 31–38

[5] Monay, F., Gatica-Perez, D.: On image auto-annotation with latent space models. In: MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia, ACM Press (2003) 275–278

[6] Tomasi, C., Kanade, T.: Shape and motion from image streams under orthography: a factorization method. IJCV **9** (1992) 137–154

[7] University of Washington: Ground truth image database. `http://www.cs.washington.edu/research/imagedatabase/groundtruth/` (2004)

[8] Duygulu, P., Barnard, K., de Freitas, J.F.G., Forsyth, D.A.: 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, London, UK, Springer-Verlag (2002) 97–112

[9] Hare, J.S., Lewis, P.H.: 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 (2005)

[10] Hare, J.S., Lewis, P.H.: On image retrieval using salient regions with vector-spaces and latent semantics. In Leow, W.K., Lew, M.S., Chua, T.S., Ma, W.Y., Chaisorn, L., Bakker, E.M., eds.: Image and Video Retrieval. Volume 3568 of LNCS., Singapore, Springer (2005) 540–549

[11] Yavlinsky, A., Schofield, E., Rüger, S.: Automated Image Annotation Using Global Features and Robust Nonparametric Density Estimation. In Leow, W.K., Lew, M.S., Chua, T.S., Ma, W.Y., Chaisorn, L., Bakker, E.M., eds.: Image and Video Retrieval. Volume 3568 of LNCS., Singapore, Springer (2005) 507–517