

Giving order to image queries

Jonathon S. Hare^a, Patrick A. S. Sinclair^b, Paul H. Lewis^a and Kirk Martinez^a

^aSchool of Electronics and Computer Science, University of Southampton, Southampton, SO17 1BJ, UK;

^bRoom 718, BBC Henry Wood House, 3-6 Langham Place, London, W1A 1AA, UK

ABSTRACT

Users of image retrieval systems often find it frustrating that the image they are looking for is not ranked near the top of the results they are presented. This paper presents a computational approach for ranking keyworded images in order of relevance to a given keyword. Our approach uses machine learning to attempt to learn what visual features within an image are most related to the keywords, and then provide ranking based on similarity to a visual aggregate. To evaluate the technique, a Web 2.0 application has been developed to obtain a corpus of user-generated ranking information for a given image collection that can be used to evaluate the performance of the ranking algorithm.

Keywords: Image retrieval, ranking, visual features, web 2.0

1. INTRODUCTION

At the present time, the current hot-topic in the image retrieval community is the retrieval of images that are not annotated. The subject of retrieving images that have subject metadata in the form of keywords has gathered very little attention. The reason behind this is that there has been an assumption by the research community that once an image has subject metadata then the retrieval should largely be a trivial matter of finding images annotated with the terms from the query. In practice, however, this approach is not particularly satisfying for the image-searchers due to the peculiarities of the keywording schemes often used for annotating imagery.

When analyzing keywording schemes used for image annotation three specific problem areas are often encountered. Firstly, many image collections are annotated with free-form keywords, without the use of a strictly enforced controlled vocabulary. Without a controlled vocabulary, it becomes harder to find relevant imagery in response to a query. As an example, consider the problem of finding images of a “garbage can”, when an inconsistent vocabulary is used (for example “trash can”, “dustbin”, etc.). Many collections that do not strictly enforce a vocabulary also suffer from inconsistent or bad spelling of keyword terms, which again confounds the issue of finding relevant images.

The second problem area is related to the quantity and depth of keywords associated with imagery. Many image collections have a tendency to apply large numbers of keywords to their images without much thought as to whether the keywords will help or hinder the image searcher.¹ An example of this would be an image of “George Bush Sr.” that was also keyworded “man”. The problem with this is that many professional image searchers would not find images of “George Bush Sr.” helpful in a search for “man”. This issue is one that can perhaps be solved by using structured thesauri and ontologies, and then only annotating images with the most specific concepts - this would make automatic query expansion possible, if it were required by the searcher.

The third and final issue is the one which this paper aims to investigate. The problem concerns the relevance of keywords to the image, that is, how well do the keywords represent the subject and content of the image? Generally speaking, a keyword that describes an object in the background of the image is less relevant than one that describes the subject or primary object of interest within the image. Unfortunately, the standard model of annotating images with keywords does not allow this information to be modeled. This has big implications for image search as searchers would like to have the most relevant images to their query presented first, and less relevant images presented later, in order to minimize the time it takes to locate a suitable target image.

Further author information E-mail: jsh2@ecs.soton.ac.uk

Basically, keyword-based search would be better if the images could be ranked in order of decreasing relevance to the query term. The idea of ranking images with respect to their relevance is not a new idea; for example, the indexes of books tend to list relevant pages for a term in decreasing order of relevance to the term.

The first part of this paper explores how a machine learning approach can be used to learn the relevance of keywords to an image from the images' visual content. The second part describes a system for evaluating how searchers would like to see images ranked, and presents results of such an evaluation. The second part also compares how well the system compares to the user-provided ranking of images under a number of different visual features.

2. ADDING RANKING TO KEYWORDED IMAGES

Most keyword-based image retrieval strategies for annotated images will return an unordered set of images that were labeled with the corresponding query term. As far as a searcher is concerned, the images will be displayed in an essentially random order with no link between image relevance to the query and image position. In this section we propose an algebraic technique that combines visual features and keyword labels in order to provide a ranked order to returned images.

Our central hypothesis is that, at least for single-term queries, the more of the particular object representing the query the image displays, then the more relevant it is to the query. The problem, therefore, is to determine how well each image represents the keyword given a visual signature for the image.

The central idea behind the approach we propose is that of a mathematical factorization which is able to determine, from a collection of images and their associated visual signatures and keywords, a set of factors that relate how the image signatures and keywords are related to each other. In our approach, visual signatures are articulated as a set of 'visual terms'.²⁻⁵ The idea of using a factorization and indeed the approach described below comes from our own previous work on retrieving unannotated imagery using keywords,^{2,3} which was in-turn inspired by the text retrieval technique, Latent Semantic Indexing.

Latent Semantic Indexing (LSI)⁶ is a technique for indexing documents in a dimensionally-reduced semantic vector space. Landauer and Littman,⁷ 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. 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).

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.⁸

The technique presented here consists of creating an observation matrix (containing both visual- and keyword-terms) for the image collection and then decomposing it into separate term and document matrices. These term and document matrices can then be used to generate rankings for subsets of images retrieved using standard query-keyword matching approaches.

2.1 Decomposing the Observation Matrix

Following the reasoning of Tomasi and Kanade,⁸ 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 2.1 (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 , $\mathbf{O} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$, such that $\mathbf{U}^T\mathbf{U} = \mathbf{V}\mathbf{V}^T = \mathbf{V}^T\mathbf{V} = \mathcal{I}$, where \mathcal{I} is the identity matrix. Now partitioning the \mathbf{U} , $\mathbf{\Sigma}$ and \mathbf{V}^T matrices as follows:

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

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

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 2.2 (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 $\hat{\mathbf{T}} \stackrel{\text{def}}{=} \mathbf{U}_k$, and, $\hat{\mathbf{D}} \stackrel{\text{def}}{=} \mathbf{\Sigma}_k\mathbf{V}_k^T$, respectively. From Equation 1, we can write

$$\hat{\mathbf{O}} = \hat{\mathbf{T}}\hat{\mathbf{D}}, \quad (3)$$

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

2.2 Using the decomposition to rank images

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.

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. As we are only interested in ranking the images with keywords lexically matching the query we only need to calculate the cosines over a reduced subset of documents.

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 optimal value empirically.

2.3 Example: Images of ‘sun’

As an example of how the technique can be applied, consider the task of looking for images that represent the keyword “sun” in the well known Corel dataset.⁹ For this example, we will use a very simple 64-bin ($4 \times 4 \times 4$) RGB color histogram to represent the visual features of the image. Each bin from the histogram will represent a single visual term, and the number of pixels corresponding to that bin will be the number of occurrences of the term in the observation vector. Figure 1 illustrates how the observation matrix is created.

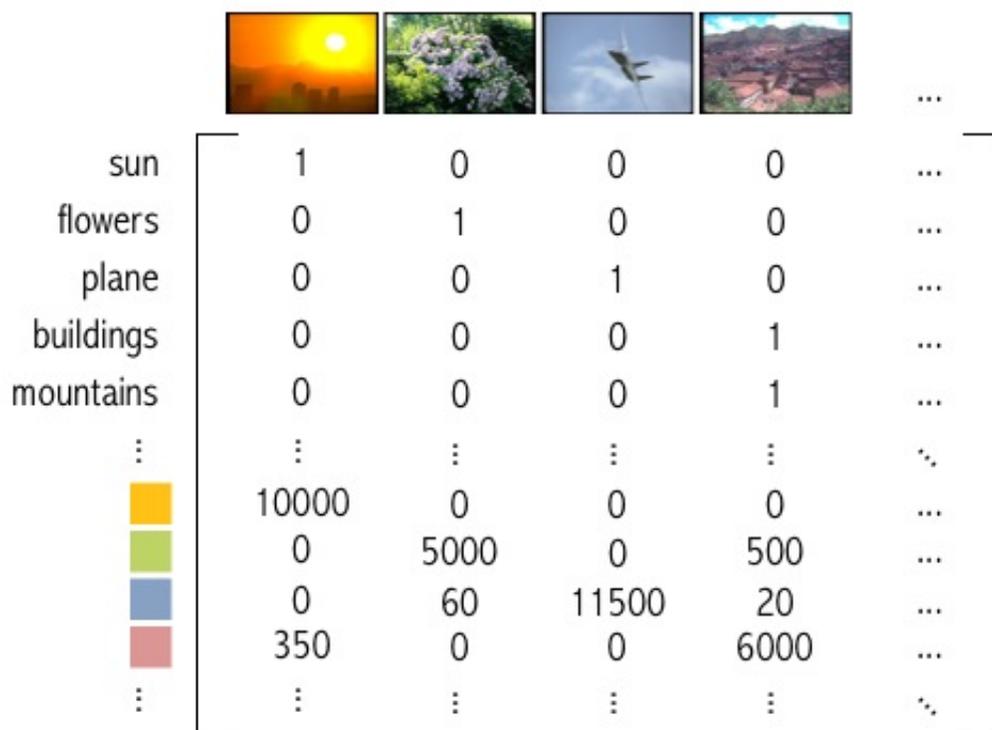


Figure 1. Creation of an observation matrix from the Corel dataset using color histogram visual terms.

In the 5000 image Corel dataset, there are 111 images labeled with the keyword “sun”. Of these, there are at least a couple of images where the keyword is not directly relevant to the image. By applying our factorization technique to the image collection we are able to essentially learn which visual features are most closely associated with each keyword by looking at the spatial proximity of the features and keywords in the resultant vector space. Figure 2 attempts to illustrate the topology of such a vector space.

In the case of the Corel images and color histogram features, the keyword ‘sun’ occurs very near to the set of red and yellow colors that often occur in images depicting the sun. The relevant images are also similarly co-located in the space, however they are arranged in such a way that the images with the closest visual similarity to the prototypical colors are closer than those images containing less of those colors.

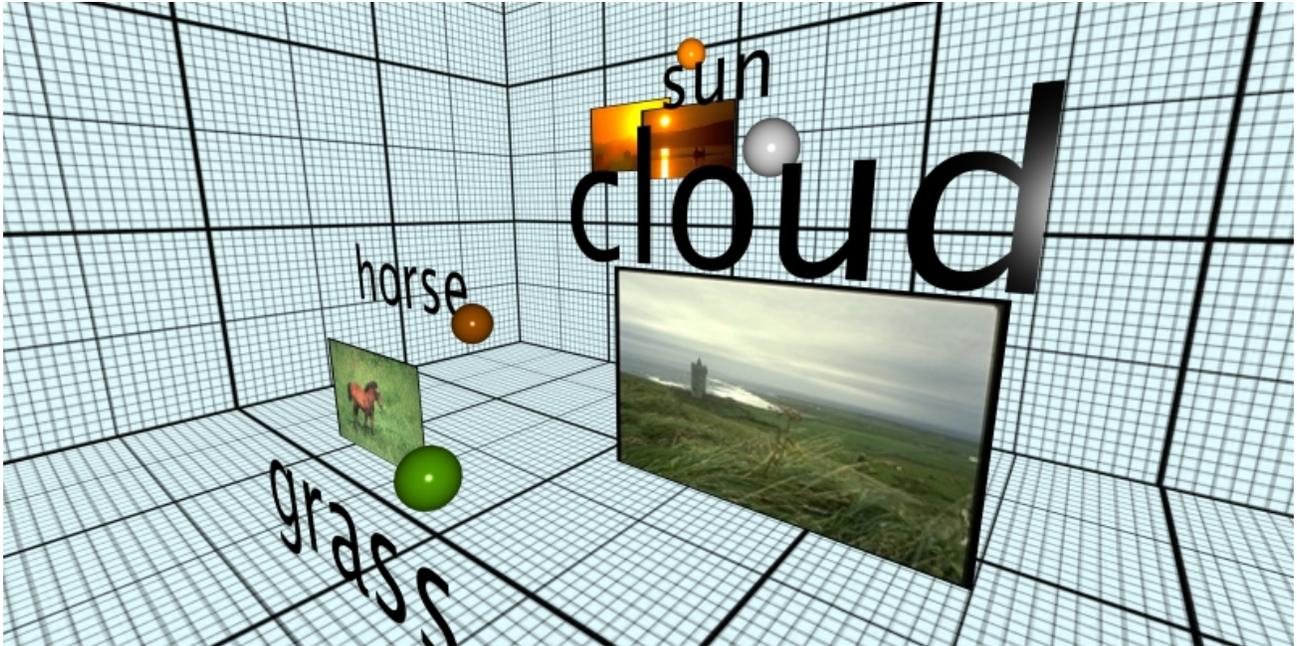


Figure 2. Example illustration showing the topology of the vector-space created through the factorization process. The colored spheres represent color-based image features.

Table 1. Example images of ‘sun’ from the Corel dataset after ranking by the factorization technique.

Image						
Rank	1	26	49	73	97	100

Table 1 shows some of the images retrieved together with their associated rank position for a query of ‘sun’. The author’s opine that there is a definite correlation between the relevance of the images to the query term and the respective rank of the images. Section 3 explores this in more detail through a process of user-evaluation.

3. USER EVALUATION OF IMAGE-KEYWORD RELEVANCE

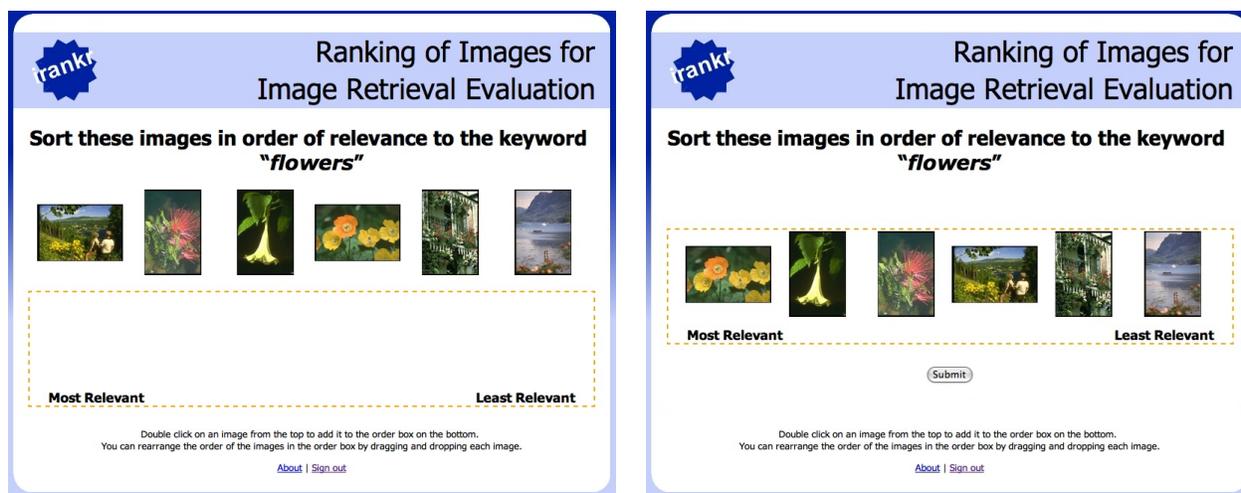
The challenge in evaluating image ranking mechanisms for keyword-based search is in obtaining the ground-truth; an image ranking must be determined against which search systems can be evaluated. This is a problem requiring human intelligence, that is asking a group of users to determine which images they find most relevant to satisfy a given query. Separate user trials on each algorithm being evaluated were considered, for example asking users to comment on the relevance of the ranking of images returned by each system, but it was felt that this approach would not have been scalable. For each new algorithm being tested, a new user trial would have to be conducted. A base system, to which each algorithm could be compared, would also be required and it is not clear which traditional ranking mechanism would be suitable to employ.

Instead, our approach involves a user-based experiment to determine suitable image rankings that can be used to evaluate the image ranking mechanism described in Section 2. Users are asked to rank various subsets of images. The subsets are the result of a keyword search, and the ranking is order of relevance for the respective keyword. The aim is to create a corpus of rankings, which can also potentially be used to evaluate the ranking performance of other image retrieval systems.

The initial plan was to use a paper-based evaluation, where users would manually rank the subsets of images printed out on cards. For each user, their preferred ranking would have been recorded manually. Although easy to implement, this approach would have been too time-consuming and expensive to evaluate large image collections. Inspired by systems such as the ESP game¹⁰ and Amazon’s Mechanical Turk,¹¹ we decided to build an evaluation platform, in the form of a dynamic web application, that would allow us to collect image ranking information for any image collection.

3.1 Experiment design

The web application that has been developed,¹² illustrated in Figure 3, is available for anyone to use. Users sign up, providing details such as a nickname. Users must provide a valid email address to activate their account, after which they are able to access the ranking page.



(a) (b)
Figure 3. Screen shots of the ranking interface from the web application.

The experimental results described below are based on the Corel dataset, although this has been restricted to a subset of images matching a selection of keywords. Currently, only the keywords ‘flowers’, ‘sun’, ‘plane’, ‘mountain’, ‘buildings’ and ‘train’ are being used. Each time the ranking page is visited, one of the keywords is chosen at random and a random subset of images matching that keyword is displayed in arbitrary order, as shown in Figure 3(a).

The image subset size was chosen to be six to ensure that users can perform each ranking task quickly. If too many images were displayed at once it would make the task overly complex for the user. Six images can also be conveniently displayed on the screen at once and still be visible to the user, so they do not waste any time scrolling up and down the page.

To ensure that users perform the ranking task, i.e. so they do not spam the system with the initial random order, the system requires that they move the images from the top to the area on the bottom. This is achieved by the user double clicking on an image to move it. The user is able to adjust the ranking by dragging and dropping the images in the box on the bottom.

Once the user has ranked all of the images, they are able to submit their selection to the system to obtain a new set of images to rank (Figure 3(b)).

As described above, the aim of this web application is to obtain a corpus of ranking information for a given image collection. This corpus contains ranked image subsets, with each subset matching a keyword-based query. The system offers the ability for administrators to collect this information in a structured format, from which a statistical analysis can be performed to evaluate image retrieval systems.

Table 2. Average correlation between machine & user rankings.

Query Term	Color Histogram	Local Color Histogram	Blobs
buildings	0.09	0.06	0.07
flowers	0.36	0.37	0.23
plane	0.19	0.19	0.26
sun	0.14	0.24	0.14
train	0.13	0.25	0.19
mountain	0.25	0.31	0.12

3.2 Initial Experiments & Results

In order to investigate how well the machine technique proposed in Section 2 performs against the user-rankings from the experimental system described in this section, we performed an experiment to attempt to generate quantitative results. The results presented here are preliminary and will be described in the final paper in greater detail.

3.2.1 Image features

For the purposes of our initial evaluation we have used three different visual feature morphologies for creating visual terms. The first is simply a 64-bin RGB color histogram as described earlier in Section 2. The second is a “local color histogram” which encodes both the pixel color and approximate location. The local color histogram is calculated by splitting the image into 16 blocks (in a 4×4 grid), and calculating a 64-bin RGB histogram for each block. Each histogram bin for each block becomes a visual term, with a total visual vocabulary size of 1024 terms. The final visual feature used is the blob-like feature created by Duygulu et al.⁹ These features were created by segmenting the images, and then calculating a number of shape, color and texture statistics for the segmented shapes, which were then arranged into feature-vectors. A visual-term representation was created by clustering the feature-vectors into a vocabulary of 500 terms using K-Means and then vector-quantizing the features using the vocabulary.

3.2.2 Measuring correlation between machine and user rankings

A suitable measure for comparing the correlation between the ranked lists of images provided by the users and the machine is Kendall’s tau rank correlation coefficient.¹³ Kendall’s τ takes on values between 1 and -1 denoting perfect agreement and perfect disagreement respectively. A value of 0 indicates an absence of any association between the lists.

3.2.3 Preliminary results

Results from 387 individual ranking tasks have been collected using the iRankr system. In total, 29 individual users performed the tasks, and each keyword had an approximately equal number of tasks associated with it.

Using the three different image feature morphologies, we constructed three separate factorizations and used them to rank images for the sample query keywords selected for the initial user evaluation. These rankings were then compared to the results from each individual ranking task using Kendall’s τ , and the results averaged for each query term. The averaged correlations are shown in Table 2.

It is difficult to say anything conclusive about these preliminary results. However, there is a strong indication that for most queries there is at least some correlation between the user rankings and machine rankings. From our previous experience in retrieving un-annotated images using a similar approach,³ we hypothesize that there will be a fairly large amount of variation in correlation over different queries. The reason for this is that the particular feature morphologies chosen are likely to model certain keyword terms well, and others less well. In particular, none of the feature morphologies chosen are particularly good at representing and distinguishing buildings. Some example rankings can be seen in Figures 4, 5 and 6.

User	
Colour Histogram	
Local Colour Histogram	
Blobs	

Figure 4. Example orderings for “Buildings”

User	
Colour Histogram	
Local Colour Histogram	
Blobs	

Figure 5. Example orderings for “Mountain”

User	
Colour Histogram	
Local Colour Histogram	
Blobs	

Figure 6. Example orderings for “Flowers”

4. CONCLUSIONS AND FUTURE WORK

This paper has presented a computational machine learning approach for ranking keyworded images in order of relevance to a given keyword. Whilst the technique appears promising for sample sets of images, such as the images keyworded “sun” in Table 1, it is important to perform a formal evaluation of the system. This is challenging as the ground truth, i.e. a ranking for a collection of images, cannot be determined computationally and ideally requires human intelligence to resolve. To overcome this problem, we have developed a Web 2.0 approach to obtain a user-generated corpus of ranking information for image collections that can be used to evaluate the performance of the machine learning technique. While the initial results are inconclusive, there is a strong indication that for many of the queries there is a correlation between the user rankings and the machine rankings.

In terms of future work, various techniques to improve the engagement of users with the web application are being considered. For example, the inclusion of a high-score table, displaying the number of rankings performed by the most active users, might encourage users to use the system and provide more results. Other forms of reward, such as announcing a prize for the most active user, could also be employed. Alternatively, the system could be exposed through the Amazon Mechanical Turk service where users would be paid for individual ranking tasks.

We would also like to investigate by how much different users rankings vary for a given query, however this would require a much larger amount of data to be collected.

ACKNOWLEDGMENTS

The third author is grateful for the support of EU funding under the OpenKnowledge and HealthAgents grants numbered IST-FP6-027253 and IST-FP6-027213.

REFERENCES

1. V. Bullen and S. Corr, “Is a picture worth 1000 words?,” in *Semantic Image Retrieval - The User Perspective*, (Brighton, UK), March 2007.
2. J. S. Hare, *Saliency for Image Description and Retrieval*. PhD thesis, University of Southampton, 2005.
3. 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 *Image and Video Retrieval, 5th International Conference, CIVR 2006, Tempe, AZ, USA, July 2006, Proceedings*, H. Sundaram, M. Naphade, J. R. Smith, and Y. Rui, eds., *Lecture Notes in Computer Science* **4071**, pp. 31–40, Springer, 2006.
4. 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.
5. M. Westmacott and P. H. Lewis, “An inverted index for image retrieval using colour pair feature terms,” in *Proceedings of the SPIE Image and Video Communications and Processing Conference*, pp. 881–889, January 2003.
6. 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), pp. 391–407, 1990.
7. 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.
8. C. Tomasi and T. Kanade, “Shape and motion from image streams under orthography: a factorization method,” *International Journal of Computer Vision* **9**, pp. 137–154, November 1992.
9. 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.
10. L. von Ahn and L. Dabbish, “Labeling images with a computer game,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pp. 319–326, ACM Press, (Vienna, Austria), April 2004.

11. Amazon Inc., "Amazon's Mechanical Turk." <http://www.mturk.com/mturk/welcome>, 2007.
12. P. Sinclair and J. S. Hare, "iRankr." <http://multimedia.ecs.soton.ac.uk/irankr>, 2007.
13. M. Kendall, *Rank Correlation Methods*, Charles Griffin & Company Limited, 1990.