SEMANTIC CLUSTERING OF IMAGES USING PATTERNS OF RELEVANCE FEEDBACK

Donn Morrison, Stéphane Marchand-Maillet, Eric Bruno

Computer Vision and Multimedia Laboratory
University of Geneva
Geneva, Switzerland
{donn.morrison,marchand,eric.bruno}@cui.unige.ch

ABSTRACT

User-supplied data such as browsing logs, click-through data, and relevance feedback judgements are an important source of knowledge during semantic indexing of documents such as images and video. Low-level indexing and abstraction methods are limited in the manner with which semantic data can be dealt. In this paper and in the context of this semantic data, we apply *latent semantic analysis* on two forms of user-supplied data, *real-world* and *artificially generated* relevance feedback judgements in order to examine the validity of using artificially generated interaction data for the study of semantic image clustering.

***Index Terms*—** Image clustering, relevance feedback, long-term learning, latent semantic analysis

1. INTRODUCTION

Approaches to automatic image annotation span a wide variety of methods, from latent and generative models [1, 2], to machine translation [3], to classification-based approaches [4, 5, 6]. These methods have proven a good starting point for bridging the semantic gap, but the problem still exists. This "gap" between low-level features and concepts depicted in images is *the* fundamental problem in computer vision and in order to narrow it, new methods of gathering *semantics* relating images to one another are needed. It is well known that user trends can be extracted from web server logs [7, 8] with applications in collaborative filtering, trend detection, etc., and the same can be applied in the image retrieval setting.

Inferred semantic relationships can take many forms, but the more popular can be categorised as: browsing logs, where users casually peruse a document collection with no formal information need; click-through data, where information is sought but evidence of interest (the "click") does not necessarily imply relevance [9, 10]; and relevance feedback judgements, where the user has a definite query and explicitly rates search results with respect to relevance [11, 12].

In this paper we will focus on exploiting long-term relevance feedback (RF) judgements for the semantic clustering of images. Ideally, a large amount of RF data is required. Due to difficulties in accumulating this type of user interaction, we also seek to demonstrate that, at least from an investigative perspective, artificially generated data is also very useful, if only for the validation of the machine learning models. Many studies have used artificial relevance feedback data [13, 14, 15, 16, 17], but none provide justifications on this data and no comparisons have been made between artificial and real-world data. This is an issue we wish to address in this paper.

Long-term (or *inter-query*) learning is the collection of RF data over many queries (and even possibly many users) and is ideal for building a semantic index over an image database. During a query session, images marked relevant or irrelevant with respect to the information need are recorded and used to build a semantic space where, after sufficient data is collected, similarities between otherwise unrelated images (in the low-level feature space) can be made apparent and used in later queries. Taken one step further, this data can be used to directly propagate image annotations across a database.

Over time and continued use of the retrieval system, the idea is that the semantic relationships, annotations, and indexes can become more and more accurate, reflecting the semantic knowledge held by the users. The idea itself is not new, and has been in the literature for several years [13, 12, 17, 10]. However, the problem has not been adequately explored. Questions we hope to answer in this study are:

- the viability of using artificially generated relevance feedback data to study algorithms for image clustering,
- the required level of database coverage of the RF data for sufficient image clustering (the degree of sparsity),
- and the issues surrounding performance over a traditional feature-based approach.

In the next section, we introduce some of the more relevant past works. We then describe latent semantic analysis (LSA) and how it is used for long-term learning. Next, we detail the image database and the real-world and artificially generated relevance feedback data used in the experiments. The

experiments follow, where we evaluate the model according to different parameters and make comparisons with a simple low-level feature-based approach in the context of a retrieval system. The implications of using and differences between artificially generated data and data collected from real users are then presented.

2. RELATED WORK

There are a handful of studies which use long-term learning for a variety of purposes, from image annotation to indexing and retrieval. Previously, relevance feedback was used only only within the duration of the query session (intra-query learning); once the query was finished this information was discarded. The Viper group produced one of the first studies which looked at using inter-query learning to aid future queries [11]. The authors analysed the logs of queries using the *GIFT (GNU Image Finding Tool)* demonstration system over a long period of time and used this information to update *tf-idf* feature weightings in the low-level feature index.

In [18], a general framework is described which annotates the images in a collection using relevance feedback instances. As a user browses an image database using a CBIR system, providing relevance feedback as the query progresses, the system automatically annotates images using the relationships described by the user. In [14], the authors combine inter-query learning with traditional low-level image features to build semantic similarities between images for use in later retrieval sessions. The similarity model between the request and target images are refined during a standard relevance feedback process for the current session. Similarly, in [19], a statistical correlation model is built to create semantic relationships between images based on the co-occurrence frequency that images are rated relevant to a query. These relationships are also fused with low-level features to propagate the annotations onto unseen images.

Inter-query learning is used in [13] to improve the accuracy of a retrieval system with latent semantic analysis. Random queries were created and two sessions of relevance feedback were conducted to generate the long-term data to be processed by LSI. From experiments on different levels of data, they conclude that LSI is robust to a lack of data quality but is highly dependent on the sparsity of interaction data. In another study, authors use long-term learning in the PicSOM retrieval system [20]. PicSOM is based on multiple parallel tree-structured *self-organising maps (SOMs)* and uses MPEG7 content descriptors for features. The authors claim that by the use of SOMs the system automatically picks the most relevant features.

Relevance feedback is used in [17] to generate a semantic space on which a support vector machine is trained. Low-level features are used in conjunction with the long-term relevance feedback data to improve performance in the MiAlbum image retrieval system. Artificial relevance feedback data was generated by running simulated queries on a database of categorised images. The positive and negative examples were taken from the top three correct and top three incorrect results respectively.

In [15], long term user interaction with a relevance feedback system is used to make better semantic judgements on unlabelled images for the purpose of image annotation. Relationships between images which are created during relevance feedback can denote similar or dissimilar concepts. The authors also try to improve the learning of semantic features by "a moving of the feature vectors" around a group of concept points, without specifically computing the concept points. The idea is to cluster the vectors around the concept centres.

Markov random walks are employed in [10] on a large bipartite click graph of queries and documents (images) collected from popular online search engines. By following walks either backward or forward from the query on the graph, document clusters can be found for associated search keywords.

3. LATENT SEMANTIC ANALYSIS

Latent semantic analysis was born out of text retrieval and uses at its core singular value decomposition [21]. Given a sparse $m \times n$ term-document matrix A, a decomposition $A = U\Sigma V^T$ is calculated, normally through a QR decomposition, which yields U ($m \times n$), the term-concept matrix, S ($n \times n$), a diagonal matrix containing the singular values in decreasing order, and V^T ($n \times n$), the concept-document matrix.

Normally, a form of dimension reduction is then applied, often referred to as *rank lowering*, where only the top k singular values are retained, and the original matrix can be approximated by multiplying the three components

$$A_k = U_k S_k V_k^T. \tag{1}$$

This dimension reduction has the effect of causing zero valued entries in the original matrix A to become non-zero. By multiplying either the term-concept matrix U or the concept-document matrix V by the diagonal matrix S and their respective transposes, one determines directly a term-term (or document-document) similarity matrix:

$$T_{sim} = U_k S_k U_k^T, \tag{2}$$

and

$$D_{sim} = V_k S_k V_k^T. \tag{3}$$

Because latent semantic analysis traditionally work with term-document matrices in text retrieval, we shall adapt this format our relevance feedback data, as has been shown in [13, 17]. Thus, the terms become the images and the documents become the relevance feedback data. In this way, each

instance of relevance feedback can be thought of as a document containing occurrences of images as terms. Some "documents" may share terms, meaning an image has been marked relevant in more than one query.

Thus, for a database of m images and n relevance feedback sessions, the image-session matrix A would take the form:

$$\text{Image} \begin{array}{c} \\ I_1 \\ I_2 \\ I_3 \\ \vdots \\ I_m \end{array} \overset{\text{Query session}}{\left[\begin{array}{c|cccc} & J_1 & J_2 & \dots & J_n \\ \hline I_1 & 1 & -1 & \dots & 0 \\ I_2 & 0 & 0 & \dots & -1 \\ I_3 & 0 & -1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ I_m & 1 & 0 & \dots & -1 \end{array}\right]}$$

where each element

$$a_{ij} = \begin{cases} 1 & \text{where image } i \text{ is relevant to query } j \\ -1 & \text{where image } i \text{ is not relevant to query } j \\ 0 & \text{no relevance between } i \text{ and } j \end{cases}$$

Our decision to use latent semantic analysis over classical supervised classification approaches comes from the fact that LSA is very good at discovering underlying concepts in data without the need of having these categories defined initially. LSA is also naturally suited for dealing with the term-document matrix because the derived semantic classes are orthogonal. A problem does arise, however, when there are many overlapping semantic classes. Singular value decomposition is not well suited for this problem. In this paper, the semantic classes are assumed (and generated) to be non-overlapping.

4. EXPERIMENTS

4.1. Database

As we have seen from previous work, exploitation of user interaction can help bridge the semantic gap by making available the underlying semantic knowledge expressed by users during image retrieval sessions. However, real-world user interaction data is often difficult and time-consuming to acquire. We seek to demonstrate that artificially generated user interaction data can help understand how algorithms work in the semantic space without the tedious data collection problems. Furthermore, the data can be specially crafted to observe certain effects, such as the emergence of new semantic relationships between queries, and how this affects automatic categorisation and annotation.

The image database used in the following experiments is a subset of the Corel collection. For purposes of data visualisation, this database was kept small with a total of 200 images from 10 categories (20 images per category)[1]. For each image, we extracted colour information to be used as the low-level features. Each image was segmented in to 9 rectangles (3x3) and the first three colour moments (mean, variance, skewness) were calculated for each segment and used to build feature vectors. Thus, for each image, there exist 81 colour features. Because of the compactness of the feature space, a simple distance measure is sufficient for determining image similarities:

$$\sum_{i}^{81} abs(x_i - y_i) \tag{4}$$

where x and y are the feature vectors for image X and Y respectively.

4.2. Relevance feedback data

We collected a pool of relevance feedback judgements on the aforementioned Corel subset. The queries were designed such that the user was shown an image and then required to locate it in the database using the GIFT (GNU Image Finding Tool) image retrieval system[2]. After each relevance feedback iteration, colour and texture weights are updated and a new set of results is displayed. Approximately five queries per for each of the 10 image categories were performed, thus, yielding around 50 relevance feedback sessions.

Next, a set of artificial relevance feedback was created based on the same image categories. The structure of this data allows us to vary the sparsity and noise for further study. For comparison with the real dataset, we constructed five queries per category. Using the 10 image categories, a complete image-session matrix was generated such that all images in each category are positively related to each other via an artificial query. Similarly, all images outside of a specified category are negatively related to those inside the category, simulating negative relevance feedback judgements (see Figure 1 (a)). This highly redundant image-session matrix is then augmented with uniform noise to closely emulate the data collected from actual relevance feedback sessions. As we will see, this will allow us to adjust levels of sparsity to determine the minimum amount of relevance feedback data for optimal image clustering.

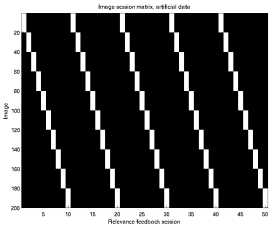
To vary the sparsity and noise of the artificial data, we generated uniform noise thresholded at a coverage percentage. For example, noise generated at a coverage percentage of 80% would randomly delete 80% of the elements in the image-session matrix, simulating the sparsity seen in the real-world data. This deletion is realised through matrix multiplication, where a matrix of uniform noise N_c thresholded at coverage c (such that each element $n_{ij} = \{0, 1\}$) is multiplied with the complete image-session matrix A:

[1] Image categories used are: *architecture, beach, bird, clouds/sky, flowers, insects, leopards, lizards, fungi, and sunset/sunrise*.

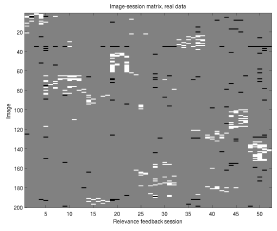
[2] `http://viper.unige.ch/demo/php/demo.php`

$$A_{sparse} = N_c * A \quad (5)$$

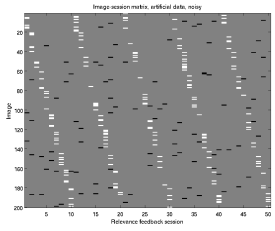
Figures 1 (b) and (c) show the image-session matrices for the real and artificially generated data, respectively. Darker areas represent negative relevance feedback while ligher areas represent positive relevance feedback, meaning the image corresponding to the row is relevant to the query, which corresponds to the column.



(a) Complete artificially generated RF data



(b) Real RF data



(c) Artificially generated RF data

Fig. 1. Image-session matrices for (a) complete artificial, (b) real and (c) sparse artificial (with 80% element deletion) relevance feedback data.

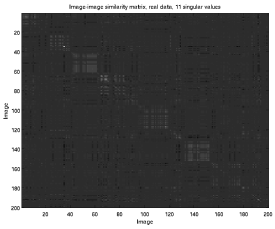
4.3. System model

The semantic clustering is designed to operate in an image retrieval system where relevance feedback can be applied. The general overview is as follows:

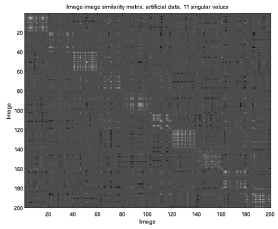
1. Relevance feedback data is stored from query sessions
2. A semantic similarity index is generated using latent semantic analysis
3. These new image similarities are used to enhance future retrieval results

We conducted the experiments as follows. We first did empirical studies on the parameters of our latent semantic analysis model over both RF datasets. Using these tuned parameters, we then ran tests on the artificial RF data to determine the minimum coverage needed for sufficient clustering in the semantic space. Average precision was used for each of these measures. These results are compared with colour-moment based similarity.

Image similarity in the semantic space is evaluated using the method described in Section 3 Eq. 2. Figure 2 shows the similarity matrices for real and artificial relevance feedback data for SVD retaining 10 singular values. These are for example only, the values chosen were empirically set (see Figure 3 (a) and (b)).



(a) Similarity matrix for real data



(b) Similarity matrix for artificially generated data

Fig. 2. Example similarity matrices based on SVD retaining 10 singular values for (a) real and (b) artificially generated data

Figure 3 (a) shows the average precision on the real data while varying the number of singular values for SVD. Average precision is also given for the colour features and a randomly generated similarity matrix for comparison. The average precision for the colour features is based on a query returning 20 results. The data points do not change as no parameters for colour similarity were altered during the experiment. Figure 3 (b) shows the average precision on the artificial data while varying the number of singular values for SVD. The artificial data leads to much less stable performance curves yet both appear to peak near 10 singular values and then descend slightly. This is due to the low-rank approximation of the original matrix by retaining the largest k singular values. Retaining too many singular values inhibits the propagation of data, essentially leaving the concept space too large. We know from the data that there are indeed 10 distinct concepts. In [17], it was shown that optimal rank in the approximated image-session matrix is closely related to the number of semantic classes in the data. In this way the latent models act to cluster the images into these semantic groups, yielding the highest average precision when an optimal value has been reached.

Figure 4 shows the average precision while varying the coverage of the artificial data over the database. We begin with a complete image-session matrix, where each image can be linked to any other semantically relevant image, iteratively and randomly deleting elements from the image-session matrix. The purpose is to show the minimum percentage of relevance feedback coverage required to have a reasonably useful precision for future queries. Interestingly, the system shows reasonable performance with just 30% relevance feedback coverage (70% simulated element deletion).

4.4. Discussion

The average precision for the real data is visibly more stable than for the artificial data. The peak performance for LSA is between 10 and 15 singular values for both datasets, indicating that this is the optimal parameterisation for the low-rank approximation and propagation of non-zero values for this data[3]. Overall, the artificial data, while modeled on the real-world data, performs relatively well. There is some instability, but both plots appear very similar, peaking near 30% average precision and gradually descending to around 24% as more singular values are retained.

On the amount of coverage required for sufficient semantic clustering, Figure 4 shows us that for LSA, we do not have much to gain by having more than 30% of the images in the database associated to one another via the RF data, and at more than 50% coverage there is virtually no improvement. These findings are supported by an earlier work, where the authors note that as little as 25% of the images are required to be annotated with RF data [13].

With respect to performance over the use of low-level features, the semantic data is ideally suited for clustering the images into semantic categories because the relevance feedback judgements come from users who can understand these relationships. Low-level feature distances have a very limited knowledge of the higher-level concepts inherent in the data and are thus more suited to measuring distances in these low-level spaces. The Corel subset used in this study was selected to show unambiguous semantic similarity while also having relatively similar inter-class colour characteristics.

5. CONCLUSION

In this paper we have demonstrated the usefulness of latent semantic analysis for generating a similarity index over an image database. With continued use of the retrieval system, relationships between images become stronger and benefit future queries. We discussed the differences between real-world and artificially generated interaction data and have shown that for the purposes of algorithm and parameter selection and validation, artificially generated data is a suitable candidate when real-world data is difficult to acquire. This validation has never been considered in previous studies on long-term learning where artificial data was used.

We found that only fraction of the images in the database need to be judged with respect to a query in order for a semantic clustering to take place. This may, to some extent, help to allay fears of the "cold-start" problem associated with long-term learning where the retrieval system will not be "usable" until it has been "used" for a sufficient period of time. Deployed efficiently in application areas with high user traffic such as internet search engines, the cold-start problem may not be noticeable.

Human activity tends to follow a power-law distribution, which we have not taken into account. For a more effective study, we propose to generate the artificial data according to a power-law distribution, with the majority of RF judgements being from several classes, and the remainder making up the tail of the distribution.

In the future, we look to validate these experiments on larger sets of both real and artificial data. We propose to collect data from many users of an image retrieval system on a large catalogue of images over a larger number of semantic categories. This will both further validate the relationship between the optimal rank approximation of SVD and the perceived concepts in the semantic space, as well as help uncover scalability issues relating to LSA and large matrices.

[3]Some results from the real dataset may have higher precision because more positive examples were needed to find the desired image (forming a power-law distribution) during the query. Contrast this to the artificial data which was generated using uniformly distributed element deletions.

6. ACKNOWLEDGMENTS

This research was funded by the Swiss National Science Foundataion (NSF) through IM2 (Interactive Multimedia Information Management).

7. REFERENCES

[1] F. Monay and D. Gatica-Perez, "On image auto-annotation with latent space models," in *Proc. ACM Int. Conf. on Multimedia (ACM MM), Berkeley, 2003.*, 2003.

[2] J. Tang, J. S. Hare, and P. H. Lewis, "Image auto-annotation using a statistical model with salient regions," in *In Proceedings of IEEE International Conference on Multimedia & Expo (ICME), Hilton Toronto, Toronto, Ontario, Canada*, 2006.

[3] P. Duygulu, K. Barnard, J. de Freitas, and D. Forsyth, "Object recognition as machine translation :learning a lexicon for a fixed image vocabulary," in *In Proceedings of ECCV 2002.*, 2002.

[4] Beitao Li and Kingshy Goh, "Confidence-based dynamic ensemble for image annotation and semantics discovery," in *MULTIMEDIA '03: Proceedings of the eleventh ACM international conference on Multimedia*, New York, NY, USA, 2003, pp. 195–206, ACM Press.

[5] King-Shy Goh, Edward Y. Chang, and Beitao Li, "Using one-class and two-class svms for multiclass image annotation," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 10, pp. 1333–1346, 2005.

[6] Gustavo Carneiro, Antoni B. Chan, Pedro J. Moreno, and Nuno Vasconcelos, "Supervised learning of semantic classes for image annotation and retrieval," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 3, pp. 394–410, 2007.

[7] Osmar R. Zaïane, Man Xin, and Jiawei Han, "Discovering web access patterns and trends by applying OLAP and data mining technology on web logs," in *Advances in Digital Libraries*, 1998, pp. 19–29.

[8] O. Nasraoui, C. Cardona, C. Rojas, and F. Gonzalez, "Mining evolving user profiles in noisy web clickstream data with a scalable immune system clustering algorithm," 2003.

[9] Gui-Rong Xue, Hua-Jun Zeng, Zheng Chen, Yong Yu, Wei-Ying Ma, WenSi Xi, and WeiGuo Fan, "Optimizing web search using web click-through data," in *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*, New York, NY, USA, 2004, pp. 118–126, ACM.

[10] Nick Craswell and Martin Szummers, "Random walks on the click graph," in *In Proceedings of SIGIR 2007*, 2007.

[11] Henning Müller, Wolfgang Müller, David McG. Squire, Stéphane Marchand-Maillet, and Thierry Pun, "Long-term learning from user behavior in content-based image retrieval," Tech. Rep., Université de Genève, 2000.

[12] Stéphane Marchand-Maillet and Eric Bruno, "Exploiting user interaction for semantic content-based image retrieval," Tech. Rep., Computer Vision and Multimedia Laboratory, Computing Centre, University of Geneva, 2003.

[13] D. Heisterkamp, "Building a latent-semantic index of an image database from patterns of relevance feedback," 2002.

[14] J. Fournier and M. Cord, "Long-term similarity learning in content-based image retrieval," 2002.

[15] M. Cord and P. H. Gosselin, "Image retrieval using long-term semantic learning," in *IEEE International Conference on Image Processing*, 2006.

[16] P.-H. Gosselin and M. Cord, "Semantic kernel learning for interactive image retrieval," in *IEEE International Conference on Image Processing*, Genoa, Italy, sept. 2005, IEEE.

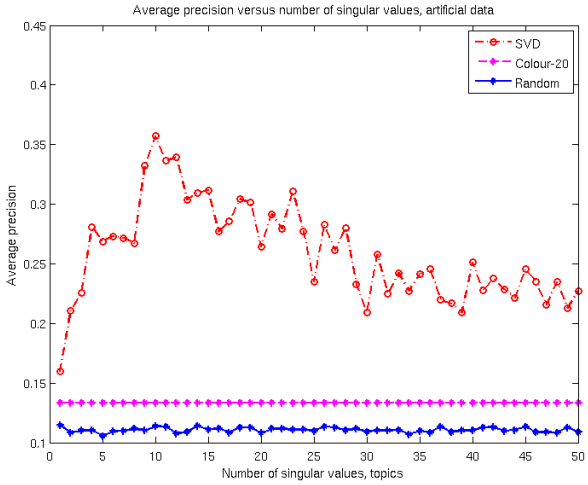
[17] Xiaofei He, King O, Wei-Ying Ma, and Zhang Hong-Jiang Li, Mingjing, "Learning a semantic space from user's relevance feedback for image retrieval," *IEEE Transactions on Circuits and Systems for Video Technology*, 2003.

[18] L. Wenyin, S. Dumais, Y. Sun, H. Zhang, M. Czerwinski, and B. Field, "Semi-automatic image annotation," 2001.

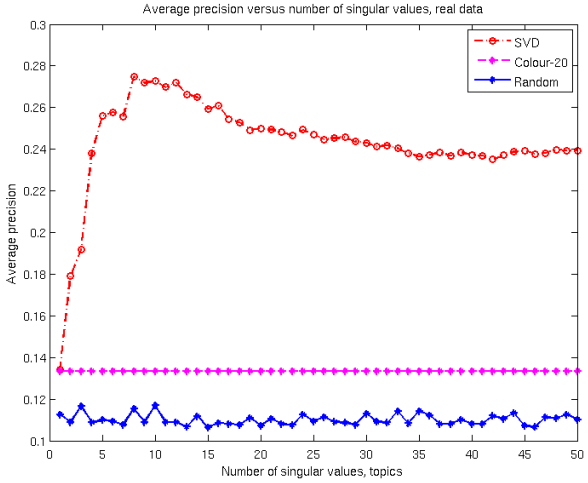
[19] M. Li, Z. Chen, and H. Zhang, "Statistical correlation analysis in image retrieval," 2002.

[20] Markus Koskela and Jorma Laaksonen, "Using long-term learning to improve efficiency of content-based image retrieval," 2003.

[21] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman, "Indexing by latent semantic analysis," *Journal of the American Society of Information Science*, vol. 41, no. 6, pp. 391–407, 1990.



(a) Average precision on artificial data



(b) Average precision on real data

Fig. 3. Average precision while varying the number of singular values retained on real and artificial relevance feedback data. The Colour-20 plot shows the average precision for colour-only retrieval returning 20 results and remains constant for comparison because no parameters were changed.

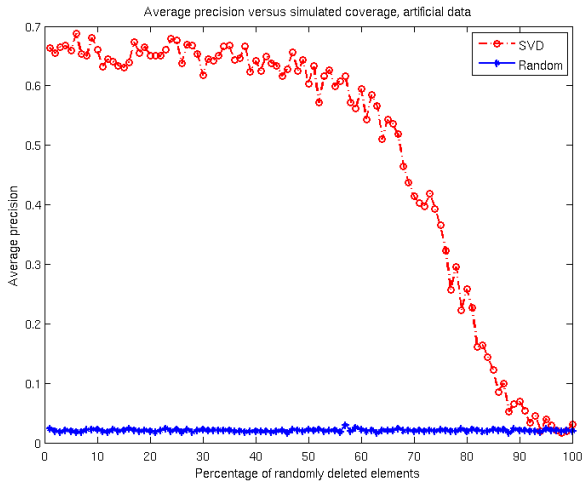


Fig. 4. Average precision while varying the coverage of the artificial data

Fig. 5. Example queries using the different algorithms - query images appear in the left columns. The top row shows query results for the artificial RF data at 80% sparsity and the bottom row using only colour features.