Hierarchical long-term learning for automatic image annotation

Donn Morrison, Stéphane Marchand-Maillet, and Eric Bruno

Centre Universitaire d'Informatique
University of Geneva, Geneva, Switzerland
{donn.morrison, marchand, eric.bruno}@cui.unige.ch
http://viper.unige.ch/

Abstract. This paper introduces a hierarchical process for propagating image annotations throughout a partially labelled database. Long-term learning, where users' query and browsing patterns are retained over multiple sessions, is used to guide the propagation of keywords onto image regions based on low-level feature distances. We demonstrate how singular value decomposition (SVD), normally used with latent semantic analysis (LSA), can be used to reconstruct a noisy image-session matrix and associate images with query concepts. These associations facilitate hierarchical filtering where image regions are matched based on shared parent concepts. A simple distance-based ranking algorithm is then used to determine keywords associated with regions.

1 Introduction

The semantic gap, recognised as the major hurdle in image retrieval, can be narrowed by tracking patterns of user interaction during query [**?**, **?**, **?**, **?**, **?**]. Previous research tends to focus on fully automatic methods using low-level features such as colour, texture, and shape [**?**, **?**, **?**, **?**, **?**] or structured augmentation using ontologies [**?**]. As with patterns in web traffic analysis, users of image retrieval systems exhibit useful information via their browsing and searching habits [**?**, ?].

The inherent limitations of using only low-level features in image retrieval become apparent after a brief appraisal of the available literature. Retrieval systems cannot reliably glean high-level semantics from low-level features due to a lack of image understanding in computer vision. There are many facets to semantic meaning and images can be described in many ways [**?**]. Subjectivity and intent in photography as well as in retrieval play a critical role. Therefore, we feel it is necessary to place more focus on user interaction in image retrieval and annotation. To ignore this information can be likened to marketing goods or services without some knowledge of consumer purchase patterns. In this paper, our goal is to semantically describe the images users are searching for, thus facilitating subsequent queries. This involves the propagation of keywords across partially annotated databases using a mixture of long-term learning and low-level image features.

In a previous paper, we demonstrated the use of singular value decomposition for the reconstruction of missing values in a session-image matrix, where each session represents a query concept [**?**]. The advantage of this method was that it relied only on long-term learning via relevance feedback on a partially annotated image database. However, a fundamental limit was found during the annotation process where new annotations were selected based on the most popular concept keywords. The result was a quantised annotation where each image belonging to a concept was annotated with similar keywords. In this paper, we improve this by dividing each image into regions which can be represented by specific keywords. The two feature types have very different meaning on the semantic level, and therefore must be hierarchically fused.

The article is structured as follows: Section 2 gives a lengthy review of related work, ranging from fully automatic annotation methods to semi-supervised methods that utilise long-term learning for annotation propagation. We have omitted works dealing with annotation by ontologies, except some studies which use WordNet. Section 3 introduces our method for automatic annotation using regions of low-level features and long-term learning. Next, Section 4 details the image database we use and the experiments followed. Section 5 reviews the experimental results and Section 6 closes with a conclusion and some proposed improvements.

2 Related work

Automatic image annotation can be approached with a variety of machine learning methods, from supervised classification to probabilistic to clustering. It is common to borrow latent and generative models from text retrieval such as latent semantic analysis (LSA) [**?**] and it's probabilistic cousin, PLSA [**?**]. These two latent-space models are compared in [**?**]. The authors pose the question of whether annotation by propagation is better than annotation by inference. LSA is shown to outperform PLSA. However, they explain that some of the reasons for this may be that LSA is better at annotating images from uniformly annotated databases.

In a later paper, the authors introduce an improved probabilistic latent model, called *PLSA-words*, which models a set of documents using dual cooperative PLSA models. The intention is to increase the relevance of the captions in the latent space. The process is divided into two stages: parameter learning, where the latent models are trained, and annotation inference, where annotations are projected onto unseen images using the generated models. In the first stage, the first PLSA model is trained on a set of captions and a new latent model is trained on the visual features of the corresponding images. In the second stage, the standard PLSA technique projects a latent variable onto the new image, and annotations of an aspect are assigned if the probability exceeds a threshold [**?**].

Extensions of PLSA have been described, for example *latent Dirichlet allocation (LDA)*, introduced in [**?**], which models documents as probabilistic mixtures of topics which are comprised of sets of words [**?**]. This model was applied to image annotation in a slightly modified version called *correspondence latent Dirichlet allocation (Corr-LDA)* [**?**]. In this study, the authors compare the algorithm with two standard hierarchical probabilistic mixture models. Three tasks are identified: modelling the joint distribution of the image and it's caption, determining the conditional distribution of words in an image, and determining the conditional distribution of words in an image region. The Corr-LDA model first generates region descriptions from the image using an LDA model. Then corresponding caption words and image regions are selected, based on how the image region was selected.

In addition to low-level image features, a semantic modality can be introduced to harness the knowledge generated by users or groups of users interacting with an image database, whether it be browsing or performing longer queries (including but not limited to relevance feedback). By observing these interactions and the associations made between relevant and non-relevant images during a query, semantic themes can start to become apparent. These themes need not be named entities such as words describing objects or concepts, but can simply be relationships between images indicating some level of semantic similarity.

This type of learning is dubbed *inter-query learning* due to the feature space spanning multiple (or all) query sessions. The converse is the traditional *intra-query learning*: the utilisation of relevance feedback examples during the current query only (after the session has ended the weights are discarded). Inter-query learning takes an approach similar to collaborative filtering; interaction (in the form of queries with relevance feedback) is required to increase density in the feature space. It is in this way that a collection can be incrementally annotated. The more interaction and querying, the more accurate the annotations become.

The Viper group produced one of the first studies which looked at inter-query learning [?]. 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 the *tf-idf* feature weightings. Images were paired based on two rules: images sharing similar features and also marked relevant have a high weight while images sharing similar features but marked both relevant and irrelevant should have a low weight (indicating a semantic disagreement). Two factors were introduced to manage the relevance feedback information. The first being a measure of the difference between the positively and negatively rated marks for each feature and the second re-weighting the positively and negatively marked features differently such that the ratio is scaled non-linearly.

Later, in [**?**], the authors focus more formally on annotation. 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.

Taking a direction toward the fusion of the two modalities, [?] 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. This refinement and fusion is facilitated by a *barycenter*. The paper also discusses the problems with asymmetrical learning, where the irrelevant images are marked irrelevant by the user for a variety of reasons, whereas relevant images are marked relevant only because they relate semantically to the query. Therefore, the authors reduce the relevance of irrelevant images during the fusion of feedback stages. Similarly, in [**?**], 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 fused with low-level features (256 colour histogram, colour moments, 64 colour coherence, Tamura coarseness and directionality) to propagate the annotations onto unseen images.

In [?], inter-query learning is used to improve the accuracy of a retrieval system and latent semantic indexing (LSI) is used in a way such that the interactions are the documents and the images correspond to the term vocabulary of the system. The authors perform a validation experiment on image databases consisting of both texture and segmentation data from the MIT and UCI repositories. Random queries were created and two sessions of relevance feedback were conducted to generate the historical information 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.

This method of exchanging RF instances and images for the documents and term vocabulary was also used in a later study where the authors use long-term learning in the PicSOM retrieval system [?]. 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. They note that the relevance feedback information provided by the users is similar to hidden annotations. Using Corel images with a ground truth set of 6 classes, MPEG7 features scalable colour, dominant colour, colour structure, colour layout, edge histogram, homogeneous texture, and region shape, the authors reported a significant increase in performance.

In [**?**], a system is proposed which shifts the document retrieval paradigm from content-based features to document similarity based on user interaction with a retrieval system. The system is built using principles from collaborative filtering (CF) which completely replace the traditional content-based technique. CF data was obtained by having users group similar images in a test environment. The CF data collected comprised 5010 similarity records 4010 of which were used as training data, and the remaining 1000 as testing data. The result of the classification experiment showed an increase in performance over a feature vector based on histograms. They concluded by stating that there exists “good inter-subject transferability of interpretation.”

In [?] 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.

3 Proposed annotation model

The following proposed hierarchical annotation model works by selecting a general (parent) concept based on relevance feedback over past query sessions. This concept comprises a subset of images from the database, each of which may have associated keywords, depending on the amount of initial annotation. With the concept selected for a particular image, each region is matched with similar regions in the concept space based on low-level image features. A ranked list of the closest matches is used to annotate each region in the unannotated image, hence a propagation of annotation.

In this paper, similar to what has been done previously [?], we created a set of artificial queries based on concepts each of which comprise semantically similar image classes (i.e., the concept “animals” contains image categories “birds,” “insects,” “leopards,” and “lizards”). The query data is used to compose an image-session matrix, in which the rows contain the images in the database and the columns represent the relevance feedback values for each query session [?, ?]. The cells of the session columns can have the following values: -1, meaning the image is irrelevant to the query; 0, where no judgement is given; or +1, where the image is considered relevant to the query by the user.

We define a session to be a query where a typical user has performed a search using relevance feedback to locate an image belonging to a particular concept (in our study these concepts are manually defined in Table 1) based on the image database. This knowledge can be used in a hierarchical manner to filter available keywords for propagation. Figure 1 shows a flow diagram for the annotation of an unseen image i.

Table 1. Concept-category relationships

Concept	Category
landscape	beach
	sunrise/sunset
sky	beach
	cloud
	sunrise/sunset
animals	bird
	insect
	leopard
	lizard
plants	flower
	mushroom
man-made	architecture

Consider an unannotated image, i_u^C, with regions R_i, which as been found through long-term learning to belong to a concept class C. Then, the subset of annotated images within C, denoted as I_a^C, are used solely for the nearest matches on low-level features. Next, for each of the n regions r_j belonging to the image i_u^C, the top k matching histograms from the relevant concept are ranked and the most common keyword is propagated to that region. This approach assumes images share similar concepts with respect to the regions. For example, an image of an insect, in the context of our collection, has a high likelihood of also being accompanied by regions depicting leaves or plants; an image of a sunset has a high likelihood of having regions depicting water or cloud.

The ranking algorithm is simply a ranked list of k Euclidean distances between the unannotated image region and all other regions sharing the same parent concept. The

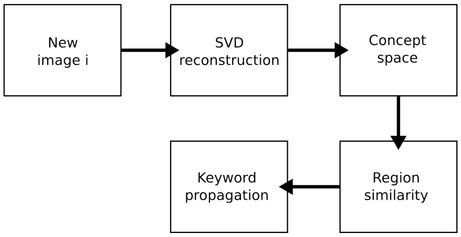


Fig. 1. Flow diagram for hierarchical annotation of an unannotated image i. The image is first added to the image-session matrix. SVD is used to decompose, filter, and reconstruct the matrix. Concepts then become apparent and are used to filter the keyword space. The nearest matching regions are ranked and the associated keywords are propagated onto the new unannotated image.

keyword with the highest vote in this ranked list is used as the new annotation. As we will see in the following section, it is possible to have no associated concept with some images due to the sparsity of the image-session matrix. In this case, we simply fall back to using the low-level feature distances to propagate region annotations. The only drawback with this is that the probability of matching irrelevant regions is increased.

The long-term learning works by storing all previous queries in a matrix A. After each query involving relevance feedback, this matrix is updated and SVD is used to associate images with concepts. In this way the annotations are never completely fixed, but can evolve with use of the retrieval system. An example artificial matrix is show in Figure 3 (a). It is highly redundant because of the large number of RF sessions generated and the low number of concepts.

In this experiment, only the positive examples ($A(i,j) == +1$) are used in order to simplify the propagation stage (we will ignore irrelevant concepts for the moment). Next, to simulate missing relevance feedback data, the values of A are randomly dropped (set to 0) to form a new noisy matrix, A_n. Singular value decomposition (SVD) is applied to this matrix to yield:

$$A_n = U \Sigma V^T. \tag{1}$$

The diagonal matrix Σ contains the singular values. We retain only $k = 5$ concepts as Σ' to filter out unimportant concepts and reconstruct A as A_r.

$$A_r = U \Sigma' V^T \tag{2}$$

With A_n reconstructed as A_r, we now apply a thresholding measure to allow diffusion of relevance feedback examples into cells with missing data. As a result of SVD, cells previously zero will now be non-zero. These values are normalised into the same space as A_n and then empirically thresholded at 0.7, giving:

$$A_r(i,j) = \begin{cases} 1 & \text{where} \quad A_r(i,j) > 0.7 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

The result is the reconstructed matrix with values that should minimise the difference from A_n:

$$D_{nr} = \sum_{i,j} |A_n(i,j) - A_r(i,j)| \tag{4}$$

We intend to annotate the unlabelled images in the database to allow for keyword-based queries. Image similarities are specified by the user by way of relevance feedback. This alone could be sufficient for labelling, but normally the feature space is very sparse, and some diffusion is needed to propagate image labels throughout the collection.

During the matching of low-level features, we use a 64 bin histogram for each of the RGB channels segmented by normalised cuts. The Euclidean distance metric was used to find the closest matching regions using these histograms:

$$D = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}. \tag{5}$$

Other measures such as histogram intersection could easily be used here, but for this initial experiment the Euclidean measure is sufficient.

Due to the fact that the regions are dynamically sized by normalised graph cuts, each region is normalised with respect to its area. Normalised graph cuts requires the number of regions to be specified as a parameter. In our experiments we set this parameter to 4. This inflexibility can cause problems during the segmentation process because if there exist three very obvious regions, a fourth will be created by dividing one of the three. This could be alleviated with some preprocessing of the image (or manual specification) to determine an optimal number of regions. To simplify our approach we left the number of regions static.

4 Experiments

For the purposes of an initial investigation, a small, uniformly distributed subset of images was taken from the Corel collection based on 10 predefined semantic categories (recall Table 1). Twenty images from each category were taken at random so as to reduce a bias towards low-level similarity. Each image was segmented into four regions using normalised graph cuts [**?**]. Next, each region was manually annotated with a keyword which best described the majority of the region. For example, if a region exists containing a small bird on a large sky, the word 'sky' would be used as the annotation. In total, 200 images were collected. The final vocabulary comprised 23 words. Figure 2 shows the distribution of words in the vocabulary.

A pool of 100 artificial relevance feedback sessions was created by setting all images under a concept as relevant to that query. In essence, the matrix created is a ground truth matrix where all concepts are related to the categories through artificial sessions. This data simulates query sessions where users would have a concept image in mind (for example, images depicting animals), and would construct the query by selecting a number of positive and negative examples.

Figure 3 (b) shows A_n, which results from the random cell deletion on A at 80%. In our experiments, the percentage of cell deletion was varied to see how SVD handles

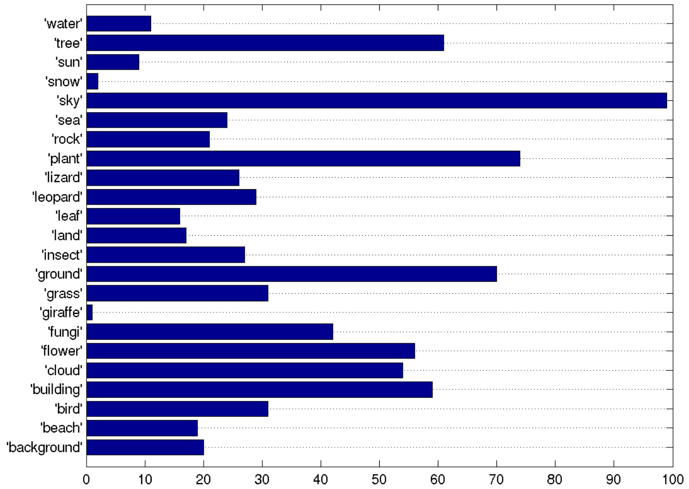


Fig. 2. Alphabetised region vocabulary distribution

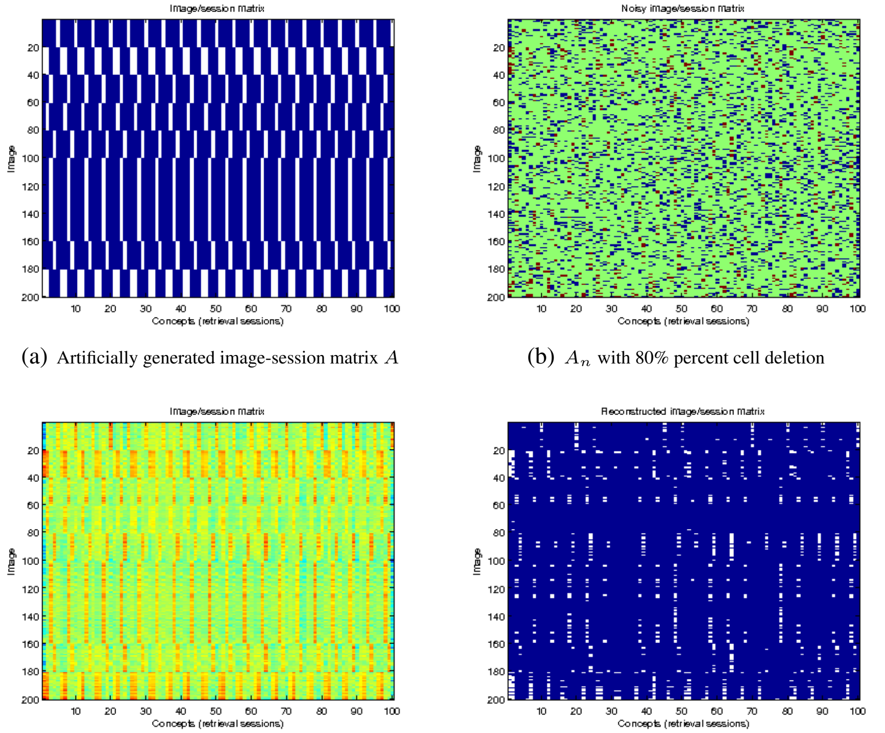
incremental missing values. Finally, Figures 3 (c) and (d) show the reconstructed image/session matrix A_r, before and after thresholding, respectively. It can be seen that one category of images (Figure 3 (d), images 61-80) suffers more corruption after reconstruction than the rest, with almost no associated concepts. This is because the category in question, “cloud”, belongs to only one concept (“sky”), while the other members of that concept belong to two concepts (“landscape” and “sky”). This causes the “cloud” category to have less influence, and thus, SVD tries to map the “cloud” concept onto these images.

To simulate a partially annotated database, we use hold-one-out cross validation to pick an unannotated image and use the remaining for distance matching.

As can be realised from Figure 3 (d), there will be missing values in the matrix that can cause some images to be unassociated with any particular concept. In this case, our algorithm falls back to simply matching the low-level feature regions.

5 Results and discussion

Figure 4 shows the prediction accuracy versus k top ranked low-level feature matches. The distribution of the vocabulary plays a part here because keywords with high distribution will eventually dilute the rankings provided their histograms are relatively close to that of the unannotated region. The accuracy (%) is calculated by strictly counting the number of predicted region keywords that match the ground truth region keywords. If this restriction is relaxed so that keywords are just associated with the image, as is the case with the Corel data set, the accuracy is significantly improved.



(a) Artificially generated image-session matrix A

(b) A_n with 80% percent cell deletion

(c) Reconstructed image-session matrix A_r

(d) Thresholded A_r showing reconstruction by SVD

Fig. 3. The various stages of the image-session matrix during singular value decomposition: (a) shows the original matrix, (b) shows the matrix A_n with entries removed simulating sparsity, (c) shows the matrix A_r reconstructed with $K = 5$ concepts, and (d) shows A_r after thresholding.

Accuracy reaches a peak, just above 35%, when there are $k = 2$ top results, and declines with local maxima for $k > 2$. However, with $k = 2$ there is no majority vote, with a keyword being picked at random from the ranking if there are two suggestions. A more stable value is $k = 3$, where a majority can be found in more cases.

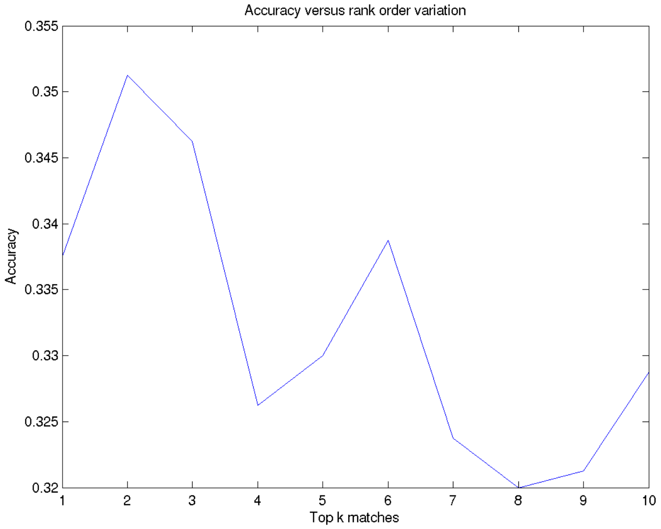


Fig. 4. Strict by-region prediction accuracy varied by k nearest regions

Some example results are pictured below in Figures 5 and 6, depicting favourable and less favourable results, respectively. It was observed that images with simple histograms were annotated more accurately (see the flower and sky images in Figure 5). Images in the animals concept were often given wrong labels for the main subject, for example, commonly mis-annotating lizards for leopards. This is partially due to the fact that the generated regions do not always directly surround semantic objects, so the colour histograms will be diluted with other areas of the image.

Further improvement could be gained by utilising WordNet to find words in similar semantic branches. In the case of the third image (bird) in Figure 5, the labels are not actually very far from the ground truth. According to WordNet, “tree” falls under the category of “plant”, which in this case is the predicted annotation.

In Figure 6, we have an example of a lizard being mistaken for a leopard. In cases of animals, especially those which exhibit patterns (scales on lizards, dots on leopards), a texture-based feature could be useful for further discrimination. The distribution of annotations in the ground truth vocabulary also has an affect on the predicted annotations. In the example of the sunset, the sun has been mis-annotated as sky. Looking back at Figure 2, we can see that the word “sun” is only associated with roughly 10 regions in

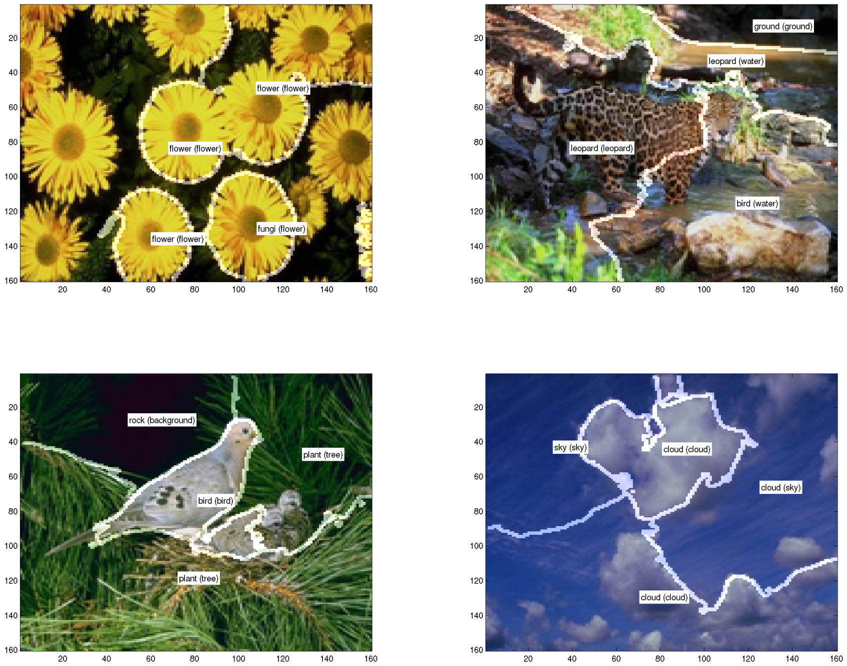


Fig. 5. Examples of well labelled regions after 80% cell deletion in the image-session matrix. Predicted labels precede ground truth labels (in parentheses).

the database whereas “sky” is the most common word in the vocabulary. This will have a direct effect on the predictions because the difference in the distribution is so great.

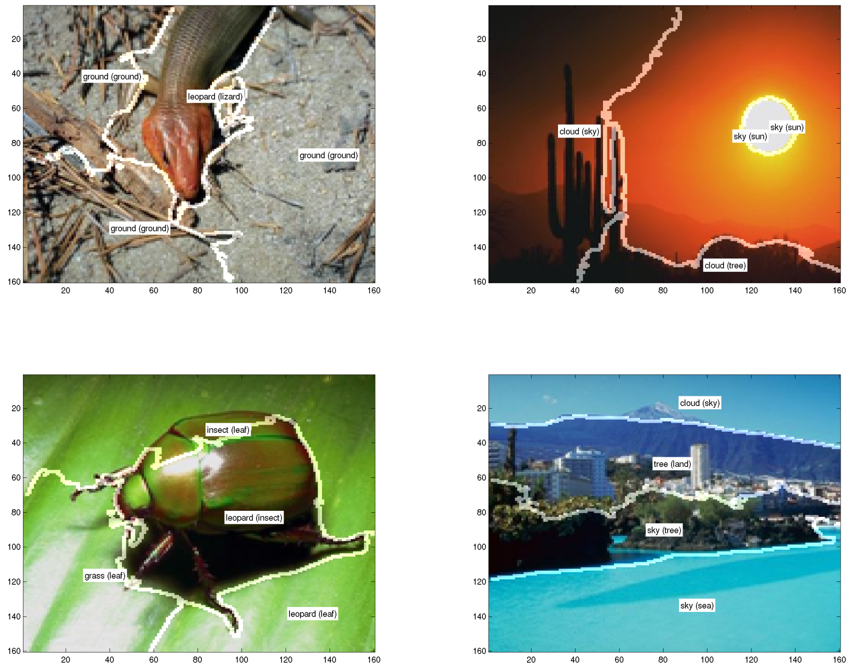


Fig. 6. Examples of poorly labelled regions after 80% cell deletion in the image-session matrix. Predicted labels precede ground truth labels (in parentheses).

In our previous study, the reliance on only relevance feedback information demonstrated the need to segment the images into regions which could more closely model specific keywords with low-level features [?]. These experiments show that image regioning provides a much finer grained approach after concept selection from relevance feedback.

Figure 6 shows examples where incorrect keywords were propagated due to the oversimplistic nature of the colour histograms used in the distance measure. Improvement could be found by adding more discriminant features such as texture and shape, although shape features would require regions to be better suited to object shape.

Due to the redundancy in the artificial data, we expect to see a large drop in performance when performing the same experiments on natural data. The natural data will normally have a much sparser image-session matrix, and thus many more images will not have category information.

6 Conclusion

From the foundation of an earlier study, this paper has demonstrated a hierarchical annotation system that combines relevance feedback and low-level colour-based features. The relevance feedback is crucial for determining the concept to which an image belongs and provides a narrowing of the secondary distance-based feature space. Because of the semantic gap, user interaction – which can be seen as a sparse approximation of image semantics, is very important for automatic image annotation. In this study, the two sets of features are complimentary. The low-level features are used to find similar image regions within the same concept space as specified by the relevance feedback information, thus allowing a much more accurate propagation.

In the longer term, we hope to add more low-level features to the distance measure, compare the distance measure with a classification approach, and use a larger image database to verify these initial findings. We also expect to begin gathering real-world data to use in place of the artificially generated relevance feedback instances.

Acknowledgements

This study was funded by the Swiss NCCR Interactive Multi-modal Information Management (IM2).

References

1. Müller, H., Müller, W., Squire, D.M., Marchand-Maillet, S., Pun, T.: Long-term learning from user behavior in content-based image retrieval. Technical report, Université de Genève (2000)
2. Heisterkamp, D.: Building a latent-semantic index of an image database from patterns of relevance feedback. (2002)
3. Fournier, J., Cord, M.: Long-term similarity learning in content-based image retrieval (2002)
4. Koskela, M., Laaksonen, J.: Using long-term learning to improve efficiency of content-based image retrieval (2003)
5. Cord, M., Gosselin, P.H.: Image retrieval using long-term semantic learning. In: IEEE International Conference on Image Processing. (2006)
6. Wang, J.Z., Li, J.: Learning-based linguistic indexing of pictures with 2-d mhmms. In: MULTIMEDIA '02: Proceedings of the tenth ACM international conference on Multimedia, New York, NY, USA, ACM Press (2002) 436–445
7. Kosinov, S., Marchand-Maillet, S.: Multimedia autoannotation via hierarchical semantic ensembles. In: Proceedings of the Int. Workshop on Learning for Adaptable Visual Systems (LAVS 2004), Cambridge, UK (2004)
8. Kosinov, S., Marchand-Maillet, S.: Hierarchical ensemble learning for multimedia categorization and autoannotation. In: Proceedings of the 2004 IEEE Signal Processing Society Workshop (MLSP 2004), São Luís, Brazil (2004) 645–654
9. Goh, K.S., Chang, E.Y., Li, B.: Using one-class and two-class svms for multiclass image annotation. IEEE Transactions on Knowledge and Data Engineering **17**(10) (2005) 1333–1346

10. Tang, J., Hare, J.S., Lewis, P.H.: 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)
11. Srikanth, M., Varner, J., Bowden, M., Moldovan, D.: Exploiting ontologies for automatic image annotation. In: SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, New York, NY, USA, ACM Press (2005) 552–558
12. Baldi, P., Frasconi, P., Smyth, P.: Modeling the Internet and the Web: Probabilistic Methods and Algorithms. John Wiley & Sons, West Sussex, England (2003)
13. Smeulders, A.W.M., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Trans. Pattern Anal. Mach. Intell. **22**(12) (2000) 1349–1380
14. Morrison, D., Marchand-Maillet, S., Bruno, E.: Automatic image annotation with relevance feedback and latent semantic analysis. In: Proceedings 5th International Workshop on Adaptive Multimedia Retrieval, Paris, France (July 5-6 2007)
15. 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**(6) (1990) 391–407
16. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. IEEE Trans. on PAMI **25** (2000)
17. Monay, F., Gatica-Perez, D.: On image auto-annotation with latent space models. In: Proc. ACM Int. Conf. on Multimedia (ACM MM), Berkeley, 2003. (2003)
18. Monay, F., Gatica-Perez, D.: Plsa-based image auto-annotation: constraining the latent space. In: MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia, New York, NY, USA, ACM Press (2004) 348–351
19. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research **3** (2003) 993–1022
20. Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., Jordan, M.: Matching words and pictures. Machine Learning Research **3** (2003) 1107–1135
21. Blei, D.M., Jordan, M.I.: Modeling annotated data. In: SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval, New York, NY, USA, ACM Press (2003) 127–134
22. Wenyin, L., Dumais, S., Sun, Y., Zhang, H., Czerwinski, M., Field, B.: Semi-automatic image annotation. (2001)
23. Li, M., Chen, Z., Zhang, H.: Statistical correlation analysis in image retrieval (2002)
24. Kanade, T., Uchihashi, S.: User-powered "content-free" approach to image retrieval. In: Proceedings of International Symposium on Digital Libraries and Knowledge Communities in Networked Information Society 2004 (DLKC04). (2004) 24–32
25. Shi, J., Malik, J.: Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(8) (2000) 888–905