NN^k Networks and Automated Annotation for Browsing Large Image Collections from the World Wide Web

Daniel Heesch
Dept. of Electrical and Electronic Engineering
Imperial College London
London, UK
daniel.heesch@imperial.ac.uk

Alexei Yavlinsky and Stefan Rüger
Department of Computing
Imperial College London
London, UK
alexei.yavlinsky@imperial.ac.uk,
srueger@imperial.ac.uk

ABSTRACT

This paper outlines a system for searching and browsing 1.14 million images from the World Wide Web (WWW) based on their visual content. At the heart of the system lies an automatically constructed network of images that can be navigated quickly by following its edges. The browsing experience is enhanced in a number of ways including multi-dimensional scaling of the graph neighbourhood for display purposes, Markov clustering of the image network to provide summaries of its content, and automated annotation of the images to allow users to access the network through text queries.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.3.4 [**Information Storage and Retrieval**]: Systems and Software—*Information networks*

General Terms

Design, Algorithms

Keywords

NN^k networks, MDS, Markov clustering, automated image annotation

1. INTRODUCTION

A fundamental task in content-based multimedia retrieval is to infer semantic relationships between objects from their representations in terms of low-level features. Not all features are equally reliable indicators of semantic content. Determining the relative importance of different features becomes all the more challenging when objects admit to a number of different interpretations: images, videos and music pieces can all be similar to one another in different meaningful ways, and different interpretations have their own set of supporting features.

A popular technique for estimating feature importance is relevance feedback on the search results given an initial image query (see [6] for a review). This approach has a number of limitations. (i) Query by example search does not support undirected search and (ii) requires users to have a query image at their disposal. (iii) Further, if the query image is external to the collection, nearest neighbour search entails a serious computational burden at run-time for large collections. (iv) Lastly, relevance feedback systems are often initialised such that each feature is given equal importance. As such a default setting may be far from optimal and produce only few or no relevant images on which relevance feedback could be given.

Our demo has been motivated by the above limitations. In our system images are linked up in a network according to their visual similarity to each other. Because the links are precomputed, navigation through the structure can be very fast and involves users choosing neighbours in the graph that resemble more closely the target image. Crucially, the latter need only exist in the user's mind and need not be made explicit at any stage.

Unlike hierarchical browsing structures which, by using a fixed distance metric at the clustering stage, make assumptions about the relative importance of features, our networks are built with the explicit intent to impose very little structure: images are linked if they are closest under *any* instantiation of a parametrised distance metric. Thus we hope to capture the different semantic facets on an image. To test scalability of the technique and assess its suitability for realistic image collections, this demo is based on 1.14 million images downloaded from the WWW.

We describe the mechanism of network construction in Section 2.1. Section 2.2 outlines two ways how we extract structure from the resulting networks to enhance the browsing experience. Section 3 is concerned with a method for automated image annotation that allows a search to start with a text query. Section 4 concludes with some implementation details.

2. BROWSING

2.1 NN^k networks

The image networks were introduced under the term NN^k networks in [1]. The motivation behind NN^k networks is to provide a browsable representation of an image collection that captures the different kinds of similarities that may exist between images. The principal idea underlying these structures is what we call the NN^k of an image. Given some focal image q, its NN^k are all those images in a collection that are closest to it under at least one instantiation of a

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

parametrised distance metric,

$$D(p,q) = \sum_{f=1}^{k} w_f d_f(p,q),$$

where the parameters w are weights associated with feature-specific distance functions d_f. Each NN^k is a nearest neighbour (NN) of q under a different metric. Each NN^k can be associated with the average of all those weight vectors under which the image is the NN^k of q (this is denoted by $\overline{w}$ for later reference).

The NN^k idea can be used to cast a collection into a network by establishing an arc from image q to image p if p is the NN^k of q. The set of NN^k can be thought of as exemplifying the different semantic facets of the focal image that lie within the representational scope of the chosen feature set. We use a set of eight local and global colour and texture features for network construction (see Section 3 for some of these).

We see the advantage of NN^k networks in their unbiased treatment of different visual features. Each image is connected to all those images it is most similar to under different feature weightings. This guards against the semantic bias otherwise introduced by imposing a fixed set of feature weights.

NN^k networks bear structural resemblance to the hyperlinked network of the Web, but they tend to exhibit a much better connectedness: the great majority of the 1.14 million images form part of a giant component and the average number of links between any two images is less than 5.

2.2 MDS and Markov clustering

Navigation in the NN^k network involves users repeatedly selecting from among the NN^k of the current focal image. We enhance the browsing experience in two ways both of which seek to extract and to expose more of the structure present in the graph.

The first method employs multi-dimensional scaling (MDS) to arrange the graph neighbourhood of the current focal image in a visually more coherent way. The distances to which MDS is applied are between the $\overline{w}$ associated with each NN^k of the currently selected image: images are thus mapped close to each other if they are nearest neighbours of the focal image under similar feature combinations.

Secondly, we apply the recently developed Markov clustering algorithm [4] to the dual of the graph (in which vertices represent edges of the original graph). By partitioning the vertex set of the dual, an image can belong to as many clusters as it has edges in the original network. In addition to showing neighbours of the currently selected node, we display representatives of the clusters to which the neighbours belong. Thus users may find more images of the same kind by selecting clusters in addition to browsing the graph (for details see [1]).

3. AUTOMATED ANNOTATION

Once users have positioned themselves in an area of interest, they can gather more relevant images by exploring the local graph neighbourhood and associated clusters. To help with the initial positioning, we use a recently developed technique for automated image annotation that has demonstrated state-of-the-art performance using only global features [5]. With images being automatically annotated, users can access the network by issuing an initial text query.

The annotation method uses Bayes' rule to determine the probability that an image is indexed with a particular word w given that its visual representation is x, i.e. $P(w|x) \propto P(x|w)P(w)$. The likelihood function $f(w,x) = P(x|w)$ is modelled as a non-parametric density obtained through kernel-smoothing over a set of training images containing w, x_w.

$$f(w,x) = \frac{1}{Z} \sum k(x - x_w),$$

where k is a kernel function and Z a normalisation constant that depends on w and that makes $f(w,x)$ a probability density for fixed w. A new image with representation x is assigned keywords that score high $P(w|x)$. For single-word queries, images are retrieved in order of decreasing $P(w|x)$. For multi-word queries, we compute aggregate probabilities by multiplying individual conditional probabilities.

Both training images and the WWW images to be annotated are represented by global colour, texture, and frequency domain features. The images are partitioned into nine rectangular tiles for each of which we compute the mean and the variance of each of the HSV channel responses as well as Tamura's coarseness, contrast and directionality properties [3]. We also apply a Gabor filter bank [2] with 24 filters (six scales $\times$ four orientations) and compute the mean and the variance of each filter's response signal on the entire image. The result is a 129-dimensional feature vector for each image.

Our training set consist of 14,081 pre-annotated images from the Corel Photo Stock covering 253 keywords.

4. IMPLEMENTATION

The search engine is implemented within the JavaServer-Pages framework and is served using Apache Tomcat. A live version runs at `http://www.beholdsearch.com`.

The complexity of network construction is quadratic in the number of images as it relies on pair-wise distances between images. The computation can however be parallelised efficiently. On eight machines (3.2 GHz), indexing 1.14 million images takes around 8 days.

5. REFERENCES

[1] D Heesch. *The NN^k Technique for Image Searching and Browsing*. PhD thesis, Imperial College London, 2005.

[2] B Manjunath and W Ma. Texture features for browsing and retrieval of image data. *IEEE Trans Pattern Analysis and Machine Intelligence*, 18(8):837–842, 1996.

[3] H Tamura, S Mori, and T Yamawaki. Textural features corresponding to visual perception. *IEEE Trans Systems, Man and Cybernetics*, 8(6):460–472, 1978.

[4] S van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.

[5] A Yavlinsky, E Schofield, and S Rüger. Automated image annotation using global features and robust nonparametric density estimation. In *Proc Int'l Conf Video and Image Retrieval*, pages 507–517. LNCS 3568, Springer, 2005.

[6] X Zhou and T Huang. Relevance feedback in image retrieval: a comprehensive review. *ACM Multimedia Systems*, 8(6):536–544, 2003.