An online system for gathering image similarity judgements

Alexei Yavlinsky
Department of Computing
Imperial College London
London SW7 2AZ
alexei.yavlinsky@imperial.ac.uk

Daniel Heesch
Department of Electrical and Electronic Engineering
Imperial College London
London SW7 2AZ
daniel.heesch@imperial.ac.uk

ABSTRACT

We describe an online application that allows users to provide similarity judgements whilst browsing a collection of 60,000 photographs. One immediate goal is to modify the initial browsing structure in response to the feedback. We thus suggest a long-term relevance feedback technique that integrates user information over multiple sessions. The principal role of the system, however, is that of a tool for acquiring a rich dataset of similarity relationships between images which we plan to make available to the community and which can be used for training and evaluation purposes. Two particular ways of how to use the data will be described in detail.

Categories and Subject Descriptors

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

General Terms

Experimentation, Human Factors

Keywords

Image Similarity, Relevance Feedback, NN^k Networks

1. INTRODUCTION

In order to evaluate the performance of an image retrieval system, or to optimise components thereof such as the distance metric, the image features, or the method of relevance feedback, it is common practice to assume that an image belongs to one particular category. The commercially available Corel collection, for example, owes its popularity among researchers as much to the quality of the images as to the convenient partitioning of the image set into categories. Many images, however, admit to multiple interpretations and a unary class membership will be too restrictive. A richer representation of the semantic content of a collection, and thus a better representation for the purpose of measuring the degree of sameness between the semantics of different images, is a set of annotations accompanying each image. However, not only does this multi-class representation render the evaluation procedure more complicated, it also fails to acknowledge that the ways in which two images can be related to each other are infinitely more varied than can be captured even by a set of terms.

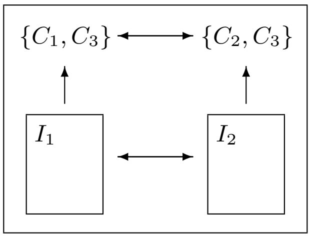

Figure 1: Two ways to establish a measure of the semantic similarity between two images I_1 and I_2: (i) annotate each image with a set of terms or concepts (upward arrows) and compare these (upper horizontal arrow), (ii) let humans compare images directly (lower horizontal arrow).

Instead, we gather information about the semantic similarity between images (as perceived by users) by letting users compare images directly (Figure 1). The similarity information thus gathered can act as an alternative, arguably more realistic ground truth for evaluating and optimising retrieval systems and its components.

Most current multimedia retrieval systems are capable of eliciting from the user some form of feedback with regard to the relevance of retrieved objects (for overviews see [6] and [3]). Almost invariably, these systems do not support long-term learning as relevance information is utilised only during the current user session. One may argue that the principal role of relevance feedback is to adjust the system to the requirements of particular users, and that, therefore, the system should revert to its original state after each session. On realistic datasets, however, the performance of content-based retrieval systems is still well below the level at which one would begin to discern differences in the similarity perception between users. The primary target should therefore be to match, in a first instance, the smallest common denominator among human users. To define the latter in a realistic way, it is critical to gather a large quantity of data from actual users about the perceived similarities between objects. Whilst users interact with the system and provide relevance judgements, the information can be used to dynamically reshape the retrieval system, based not only on the most recent feedback but the cumulative history of the judgements of all users.

This short paper is structured as follows. The next section introduces similar initiatives. Section 3 presents details of the system and Section 4 describes two potential applications of the similarity data. We conclude the paper in Section 5.

2. RELATED WORK

The view that human consensus data can be of enormous practical value to address fundamental problems in computer vision lies behind a number of similar projects. Notable examples include the LabelMe initiative at MIT[1] and the ESP game of Carnegie Mellon University[2], both of which have the objective to gather image annotations: while the former seeks term-region correspondences, the latter aims for a global bag-of-word annotation. Both enjoy considerable success, partly because researchers are encouraged to contribute annotations should they wish to use the data.

To our knowledge, the system we describe is the first of its kind that gathers user information not about the classes (annotations) of individual images but about the perceived similarity *between* images.

3. SYSTEM OVERVIEW

This section describes the various system components. The application is served by Apache Tomcat using the Java Server Pages (JSP) API and can be accessed online.[3]

3.1 Collection

The collection comprises around 60,000 photographs that are freely available through Flickr[4] (both images and associated annotations are available upon request). The images were obtained from the Flickr group 'JPEG magazine', the aim of which is to maintain a user-submitted catalogue of unaltered digital photographs.

3.2 Features

In the following we describe the three low-level features used to construct the browsing structure to be introduced in Section 3.3.

3.2.1 HSV global colour histogram:

HSV is a cylindrical colour space with H (hue) being the angular, S (saturation) the radial and V (brightness) the height component. We choose a linear subdivision into 10 hues, 5 saturation values and 5 brightness values yielding a 205-dimensional feature vector: Since hue is singular along the achromatic axis we merge all pie-shaped three-dimensional HSV bins touching the achromatic axis. The HSV colour histogram is normalised so that the components add up to one.

3.2.2 Colour structure descriptor:

This feature is defined in the HMMD (hue, min, max, diff) colour space and is part of the MPEG-7 standard [5]. The HMMD space is derived from the HSV and RGB spaces. The hue component is the same as in the HSV space, max and min denote, respectively, the maximum and minimum among the R, G, and B values, and the diff component is the difference between max and min. The colour space is quantised non-uniformly into 184 bins with the three dimensions being hue, sum (defined as $(\max + \min)/2$) and diff.

The colour structure descriptor is obtained by sliding a 8×8 structuring window and count for each HMMD bin the number of positions for which the window contains at least one pixel from that bin. This descriptor is capable of discriminating between images that have the same global colour distribution but different local colour structures. The 184 bin values are normalised by dividing by the number of locations of the structuring window so that each of the bin values falls in the range $[0, 1]$ but the sum of the bin values can take any value up to 64.

3.2.3 Thumbnail feature:

This feature is obtained by scaling down the original image to 44 × 27 pixels and then recording the gray value of each of the pixels leaving us with a feature vector of size 1,188. It is suited to identify groups of near-identical images.

3.3 Baseline structure

Complete information about the similarity relationships between images remains practically unattainable, if only because a new user may differ from previous ones. A more practical limitation arises from the fact that while there are only N images in a collection, the number of unique pairwise relationships is of order $\mathcal{O}(N^2)$. The task would be further complicated if we wanted to weigh similarity relationships by the number of users that have voted for them, for in order to make robust estimates, we would need several votes for each link.

We alleviate the problem by initially offering users an approximation of the semantic relationships based on a number of visual image descriptors. Users are subsequently asked to refine this initial guess by providing relevance information for individual image pairs that are judged similar by the system.

We represent the approximation by NN^k networks, an associative structure proposed in [2]. NN^k networks are constructed by linking an image to all those images that are closest to it under at least one weighted linear combination of feature-specific distances. By not fixing the weights associated with different features, the networks seek to capture the range of different semantic relationships that may exist between two images. Because we do not know *a priori* which combination of features works best for a particular image (or rather for a particular semantic facet of the image), considering all feature combinations increases the chance that at least *some* neighbour turns out to be relevant.

[1] http://labelme.csail.mit.edu/
[2] http://www.espgame.org/
[3] http://www-jsp.doc.ic.ac.uk/∼agy02/similarity/
[4] http://www.flickr.com

NN^k networks have been shown to exhibit topological properties that support fast navigation between different parts of the network. For example, the average number of links between two images in a network of 100,000 images is around 4 and grows logarithmically with the collection size. This is a desirable property as we want users to be able to explore the collection effectively in search for pairs of similar images.

3.4 Interface

Like the World Wide Web, users browse the image network by following links between images. At every point of the network, we display the currently selected image (here referred to as the *focal* image) at the centre of the display surrounded by a fixed number of neighbours. Clicking on any neighbour recentres that image and retrieves the neighbours of this new focal image.

The browsing interface is shown in Figure 2. Whilst the angular component of a neighbour's position is arbitrary, the distance from the centre is determined by the number of feature combinations under which the image is the nearest neighbour of the focal image.

3.5 Relevance Feedback

To ease the operational and cognitive burden for the user, we do not consider *degrees* of similarity. The similarity relation either holds between two images or it doesn't. To gather this information from users, each of the displayed neighbours has a tick box that is initially empty. Whenever users perceive a neighbour as being similar to the focal image, they may tick the corresponding box. The event is immediately recorded by connecting to the server-side database. Users need not follow any of the selected images, so that relevance feedback and browsing coexist as independent operations. We record the number of users that have confirmed a particular link as an intuitive measure of the strength of the relationship.

In order to allow users to not only confirm already existing links, but to establish new links that were not captured by the initial network structure, we record and display the history of the browsing trail. Any image from that trail can be marked as being similar to the focal image. Any image thus marked becomes a new neighbour of the focal image and will be displayed as such from then onwards. We therefore do not only record user information but respond to the feedback by continuously adjusting the network structure.

Figure 2 shows the interface after the user has followed a number of links. The history is shown at the bottom, and a number of images have been marked as similar to the focal image, which here depicts an evening scene at sea.

Note that, because image similarity is always with respect to the currently centred image, the binary similarity relationship can be established by clicking a single box. While this scheme prevents users from establishing links between two neighbours of the current focal image, we hope that the improved user experience resulting from the simplicity of unary feedback ultimately results in higher adoption rate and better coverage.

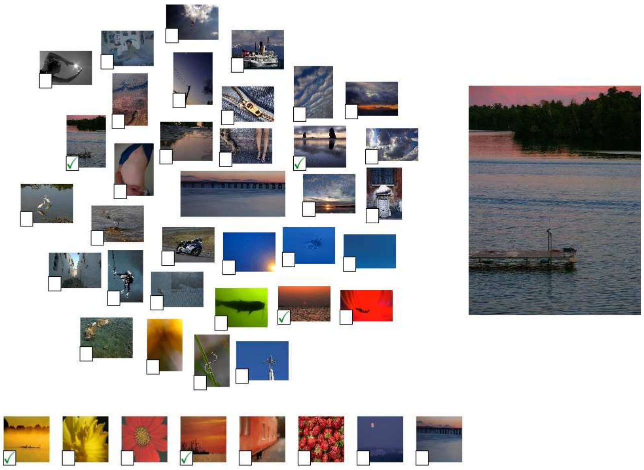

Figure 2: Main part of the browsing interface. The focal image is displayed at the centre surrounded by 33 of its neighbours. Each of the neighbours is closest to the focal image under at least one linear combination of feature-specific distances. Users confirm a similarity relationship between the focal image and one of its neighbours by ticking the corresponding checkbox. Links to previously viewed images can also be established by having access to the series of most recently selected focal images. By moving over an image, a blow-up is displayed to the right.

3.6 Issues

3.6.1 Consistency

In order to ensure a sufficient level of participation, we feel it necessary not to restrict access to the site, and to gather relevance judgements indiscriminately from all users. The information we distill from the votes is thus inevitably subject to much greater noise than if we were to set up controlled experiments locally. For example, users will naturally differ in how exhaustively they provide relevance judgements, and how similar two images need to be for them to qualify as having a semantic relationships. We aim to reduce the amount of variation in a number of ways. First, we provide a set of guidelines to help set a certain standard, e.g. we ask visitors to only confirm links if the semantic relationships is clearly perceived without prolonged deliberation. Secondly, we keep users informed about how other users tend to vote, e.g. how many images on average were marked by other users. These statistics are continuously updated and displayed alongside the browsing interface.

We are encouraged by the success of similar projects that rely on user-generated content. These include those mentioned in Section 2 but also non-academic initiatives like Wikipedia[5].

3.6.2 Negative feedback

To keep the interaction simple, we do not provide users with an opportunity to indicate explicitly the absence of a semantic relationship. Images that have not been marked may

[5] http://www.wikipedia.org

therefore belong to either of two classes: either the user did not care to indicate a positive relationship, or it was perceived as dissimilar to the focal image.

We do not try to distinguish between these two possibilities but simply add a small negative weight to each link that has not been confirmed by a user. Since users may decide to use the system without contributing any relevance information, this negative feedback only comes into effect if at least one image on the current display has been marked as similar.

4. APPLICATIONS

Below we sketch two specific uses of the similarity data we obtain. No doubt many other applications will be identified within the image retrieval and computer vision communities, to which the data will be made freely available.

4.1 Feature selection and dimensionality reduction

Once a sufficient number of NN^k network links have been confirmed by users, a new network structure will emerge that consists entirely of human similarity judgements. It is then possible to treat the weighted adjacency matrix of the network as a pairwise (dis-)similarity matrix. This can be used for performing feature selection by finding features that are correlated with human similarity judgements. We outline one method of doing so below.

Consider embedding images from the network into a multidimensional Euclidean space by applying metric Multidimensional Scaling to the pairwise image dissimilarities calculated in the above manner. Metric Multidimensional Scaling (MDS), also known as Principal Coordinates Analysis, takes a complete set of inter-point dissimilarities – which it assumes are distances – and creates a configuration of points in real vector space [4]. The Euclidean distances between them approximately reproduce the original inter-point distances. We can assume that the resulting image vectors are important image features that we are unable to observe directly, and look for correlations of these features with the low-level image features that we are able to extract.

One method for doing so is the Canonical Correlation Analysis (CCA). Consider two vector variables, $\mathbf{x} \in \mathbb{R}^m$ and $\mathbf{y} \in \mathbb{R}^n$ — in our case one corresponding to the Euclidean coordinates of images induced by MDS and the other to our extracted image features. CCA can be defined as the problem of finding two basis vectors, $\mathbf{w}_x$ for $\mathbf{x}$ and $\mathbf{w}_y$ for $\mathbf{y}$, such that the correlations between the projections of the variables onto these basis vectors are mutually maximised. Consider the linear projections $x = \mathbf{x}^T\mathbf{w}_x$ and $y = \mathbf{y}^T\mathbf{w}_y$. The exact function to be maximised is:

$$\rho = \frac{\mathbf{w}_x{}^T\mathbf{C}_{xy}\mathbf{w}_y}{\sqrt{\mathbf{w}_x{}^T\mathbf{C}_{xx}\mathbf{w}_x\mathbf{w}_y{}^T\mathbf{C}_{yy}\mathbf{w}_{\ y}}}, \qquad (1)$$

where $\mathbf{C}_{xy}$ is the covariance matrix of x and y. The basis vectors can be obtained by solving a set of eigenvalue equations (see [1]). By investigating the covariance between x and y we can identify the dimensions of our low-level feature space that maximally correlate with human similarity judgements. This can be used for visual feature selection and feature dimensionality reduction.

4.2 Measuring image annotation quality

As pointed out in the introduction, the semantics of many images is not easily captured by a small number of terms. In fact, some aspects of an image escape verbal description altogether. Having at our disposal a user-generated similarity network based on semantic relationships of *any* kind, we can proceed to a quantitative evaluation of the quality of annotations accompanying this particular and other collections.

The central idea is to compare the user-generated network with those that can be constructed on the basis of image annotations alone. The latter could be achieved by, for example, linking two images if they share annotation terms, where the number of such terms could be used as the edge weight. From both networks, certain topological properties can be extracted, such as the average distance between nodes, or the distribution of the number of neighbours. These, in turn, would provide the basis for a comparison between networks, and thus, essentially, shed light on the extent of distorsion/information loss resulting from a representation of the image by a set of classes (Figure 1).

5. CONCLUSIONS

We have proposed an online, community-oriented image browsing application that gathers image similarity judgements from users. There are two novel applications of this system. Firstly, user feedback enables us to continuously improve the underlying image network structure on which the browsing mechanism is based. Secondly, through wide community participation, we hope to gather a sufficient quantity of manual similarity judgements of images. Using these data, we will study how the human notion of similarity correlates with that derived from either low-level image feature analysis or manually obtained image annotations. We look forward to sharing the data with the research community.

6. REFERENCES

[1] D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis an overview with application to learning methods. Technical Report CSD-TR-03-02, Royal Holloway University of London, 2003.

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

[3] D. Heesch and S. Rüger. Interaction models and relevance feedback in content-based image retrieval. In Y.-J. Zhang, editor, *Semantic-Based Visual Information Retrieval.* Idea-Group, 2006.

[4] J. Kruskal and M. Wish. *Multidimensional Scaling,.* Beverly Hills and London: Sage Publications, 1978.

[5] B. Manjunath, J.-R. Ohm, V. Vasudevan, and A. Yamada. Color and texture descriptors. *IEEE Trans Circuits and Systems for Video Technology*, 11(6):703–715, 2001.

[6] Y. Rui, T. Huang, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Trans Circuits, Systems and Video Technology*, 8(5):644–655, 1998.