Sparse High-Dimensional Representations and Large Margin Classifiers for Image Retrieval

Kinh Tieu & Paul Viola

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

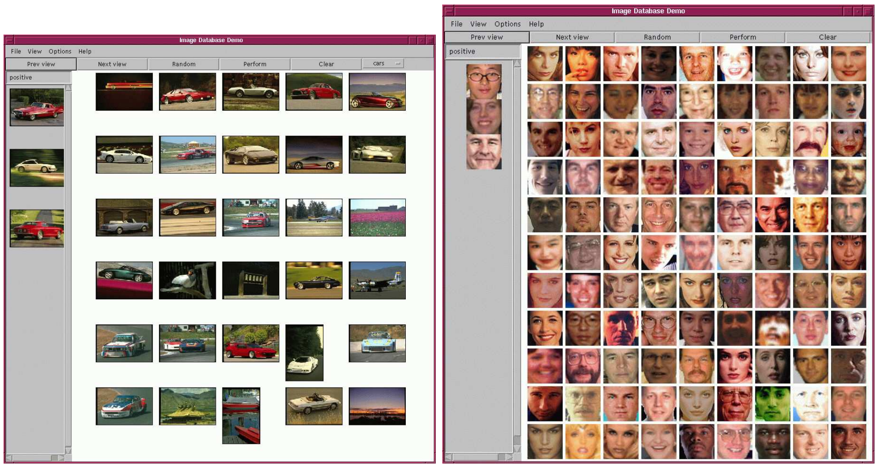

Figure 1: Retrieval results for cars and faces based on a few example images.

The Problem: There has been an explosion of content on the Web. In the very near future, images, video and virtual reality will be available on demand much as text is now. Somehow an interested user must be able to find the images or video of interest. For example a user may wish to scan a travel documentary for images of distinct locations and objects, like “Buddhist temples”, “Gothic cathedrals”, or “statues on horseback”.

Motivation: The proliferation of digital images has created a need for methods to index image databases for queries based on visual similarity. The types of images vary significantly from an image of an African gray parrot to an outdoor festival scene containing crowds of people. This, along with the large number of images, makes manual indexing not only inadequate but impractical. There is a need for an *automatic* indexing method that provides discrimination of many possible visual classes.

Previous Work: Most early image database approaches used a very small set of “simple” features: for example pixel histograms of color, vertical and/or horizontal edges. Initial approaches focused exclusively on color measures because they are pose insensitive [5]. Unfortunately these measures are also very non-selective–many images have blue pixels and vertical edges. A query with such simple features must stake out a very complex and irregular region in the feature space. Finding such a region is a complex process that requires much more data than the few examples selected by the user. As a result, instead of the system learning the query, users are required to fine-tune weights for the features based on their prior knowledge and intuition. Simple features often do not adequately separate the many possible visual classes (e.g. cars may be colored red, black, etc.).

Approach: Our approach for image database retrieval depends on two related proposals: (1) that images are best represented using a very large and selective set of features; and (2) that learning a query (image class) should quickly focus on just a few of these features.

Our system differs from others because it not only detects simple first order features, such as oriented edges and color, but also measures how these first order features are related to one another. Thus by finding patterns between image regions with particular local properties, more complex – and therefore more discriminating – features can be extracted.

The process starts out by first extracting a feature map for each type of simple feature (there are 25 simple linear features including "oriented edges", "center surround," and "bar" filters). Each feature map is rectified and down-sampled by two. The 25 feature maps are then used as input to another round of feature extraction (yielding 625 = 25 x 25 feature maps). The process is repeated again to yield 15,625 feature maps. Finally each feature map is summed to yield a single feature value. To further increase the selectivity of the set of features we extract a subset of features consisting of those features with the largest kurtosis across a sample of images.

It might seem that the introduction of tens of thousands of features could only make the query learning process infeasible. However two recent results in machine learning argue that this is not necessarily a terrible mistake: "support vector machines" and "boosting" [6, 7]. Both approaches have been shown to generalize well in high dimensional spaces because they maximize the margin between positive and negative examples. Boosting is more appropriate for our problem because it can be used to greedily select a small number of discriminating features from a large number of potential features.

We use the AdaBoost[7] learning algorithm to combine a collection of weak classifiers (expected only to correctly classify slightly more than half of the examples) to form a stronger classifier. In order for the weak learner to be boosted, it is called upon to solve a sequence of learning problems. In each subsequent problem, examples are re-weighted in order to emphasize those which were previously incorrectly classified. The final strong classifier is a weighted combination of weak classifiers. The weak learner used in the image query domain selects the single complex feature along which the positive examples are most distinct from the negative examples.

Difficulty: *Retrieval* differs from the more typical task of *classification* in that the number of potential image classes is extremely large and that the target class remains unknown until query time. Thus traditional machine learning methods for classification are difficult to apply since they often require a small number of classes and a large set of labeled data $\{x^i, y^i\}$ (where x^i is an input image and y^i is the class label). In addition, the image database system must be able to both learn the query and retrieve similar images quickly for online user interaction.

Impact: It has been observed that the structure of natural images may be based on a sparse code [8]. Our approach for image database retrieval is based on representing images with a very large set of highly-selective, complex features and interactively learning queries with a simple large margin classifier. The selectivity of the features allow effective queries to be formulated using just a small set of features and supports a "sparse" causal structure for images.

Future Work: We have shown that our system is capable of learning classes such as car and faces with very few examples (see Figure 1). We plan to apply our techniques to build other specialized detectors for particular classes. Another goal is analyze which features are most discriminative. In addition we are exploring other boosting algorithms and experimenting with methods to use other large margin classifiers such as support vector machines.

Research Support: This work supported in part by Nippon Telegraph and Telephone Corp.

References:

[1] The IBM QBIC Project Web: http://www.qbic.almaden.ibm.com

[2] The Virage Project Web: http://www.virage.com

[3] M. Kelly and T. M. Cannon and D. R. Hush Query by image example: the CANDID approach *SPIE Vol. 2420 Storage and Retrieval for Image and Video Databases III*, 238-248, 1995.

[4] A. Pentland, R. W. Picard and S. Sclaroff Photobook: Content-based Manipulation of Image Databases *Tech. Rpt.*, MIT Media Lab, 18(3):233-254, 1995.

[5] M. J. Swain and D. H. Ballard Color Indexing *Int. J. Comp. Vis.*, 7(1):11-32, 1991.

[6] C. Cortes and V. Vapnik. Support vector networks. *Mach. Learn.*, 20:1-25, 1995.

[7] Y. Freund and R. E. Schapire A decision-theoretic generalization of on-line learning and an application to boosting *J. Comp. & Sys. Sci.*, 55(1):119-139, 1997.

[8] B. A. Olshausen and D. J. Field Emergence of simple-cell receptive-field properties by learning a sparse code for natural images *Nature*, 381:607-609, 1996.