Image and Video Retrieval using a Visual Encyclopedia 9807-NTT03

Proposal for 1998-1999 Funding

Paul A. Viola

Abstract

We propose a collaborative research project with NTT on image and video database search. We plan to study and create systems that can scan images and video to locate items of interest. For example, such a system should be able to scan a travel documentary for images of distinct locations and objects, like ``Buddhist temples'', ``gothic cathedrals'', or ``statues on horseback''. We believe that by leveraging our existing work in this area, we can play a key role in setting the standard for research in visual information retrieval. At the same time, this research provides an excellent opportunity for transition to practical applications. We believe that the complementary skills of MIT and NTT are well suited for pursuing this dual path of developing practical applications of image indexing in conjunction with fundamental progress in associated science and engineering.

Image and Video Search

As confidence in the Web as a marketplace for information increases, there has been an associated 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. Indexing into image and video content will require very different techniques than those currently associated with text, however. This follows both from the huge volume of information associated with imagery, and from the need to develop the visual equivalent of textual ``keywords'' to support indexing based on image content.

Currently there are two prototypical schemes for searching text on the web: Altavista and Yahoo. Altavista provides a flat interface to the web; all documents which are similar to the query are returned. Yahoo presents an hierarchical organization of the web; documents are clustered by topic and arranged so that similar documents are nearby in the hierarchy. When it works well, Yahoo provides an important service: from a few key words documents from an entire subject area can be retrieved. Though this class of documents might have little surface similarity, by retrieving a single document from the class the entire class can be identified.

Previous approaches to image database search have taken the "altavista" approach, wherein the image database lacks any organization. In response to a query, all images which are similar to the query are returned. We propose to construct an organized visual encyclopedia in which images are clustered and organized into a hierarchy. Like Yahoo our system should be able to take an under-specified query and return an entire class of images.

We argue that an organized visual encyclopedia will address one of the greatest weaknesses of current image database systems: visual experience and memory. Without this memory it will be impossible for a computer retrieval system to match the performance of a human being. Given two pictures of dogs, a small black dog and a large spotted dog, the human observer has very little difficulty determining that these are dogs and generalizing to many other types of dogs. A computer system might return other four legged creatures, but it has no hope of rejecting cats or horses. There is simply not enough information in these images. The missing component is visual experience. People encounter many hundreds of dogs, having seen them from every conceivable angle. Given two examples, the human observer can immediately retrieve much of this visual experience, using it to correctly determine detailed information about the class of dogs. Our visual memory is likely to include useful negative examples such as cats and horses, which can be used to further refine our notion. All of this information is used when a human searches an image database. On the contrary the computer must work from only two example images.

A visual encyclopedia can supply some of this missing information. The encyclopedia will likely have an entry for dogs containing hundreds of examples, pointers to similar negative examples like cats, and pointers to distant negative examples like whales. Clearly such an encyclopedia can have significant impact on image database retrieval. Before retrieving from a database, the encyclopedia can be used to both expand and refine a query.

In addition to developing the technology for constructing and using a visual encyclopedia we intend to address several generic questions in image and video database retrieval. How can the system be made more reliable and easier to use? How can the system be made more efficient? How can the class of capabilities be extended to new domains? At the same time, by working with colleagues from NTT, we will leverage our ongoing work into practical demonstrations and applications.

Technology

In past work we have explored two complementary approaches: i) A flexible template approach that recognizes images based on the qualitative spatial relationships between regions; and ii) A Feature Based Approach that decomposes images into a very broad set of multi-scale texture features.
The first is good for image classes that are best described by the configuration of components. For example a ``snowy mountain scene'' contains a sky region which is above a snow region which is above the mountain region (see Figure 1 for an example).  The second is best for complex and unpredictable patterns in images.  For example an image of a ``French gothic cathedral'' can be described by the textural properties of the ornate columns and flying buttresses (see Figure 2 for an example). We believe that these techniques are synergistic and will lead to more powerful tools for classifying images and video.

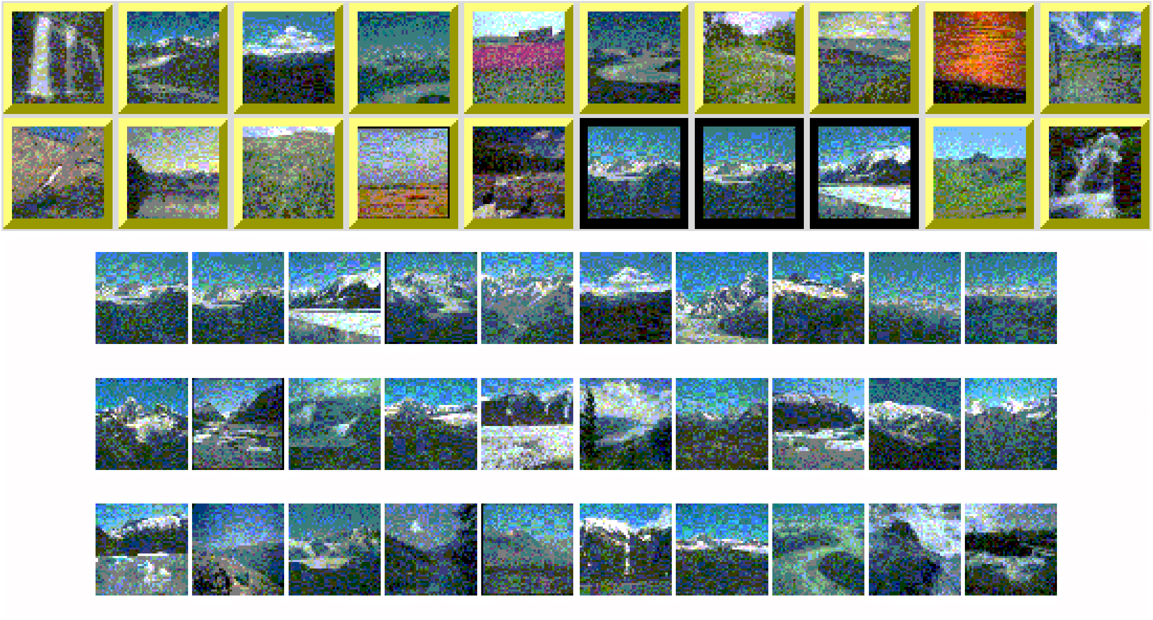

Figure 1. This figure illustrates the flexible template method retrieving images of ``snowy mountains''. The top two rows show a set of examples from which the users has selected three (highlighted in black). The bottom three rows show a larger collection of images which the system has selected as being similar to the two query images.

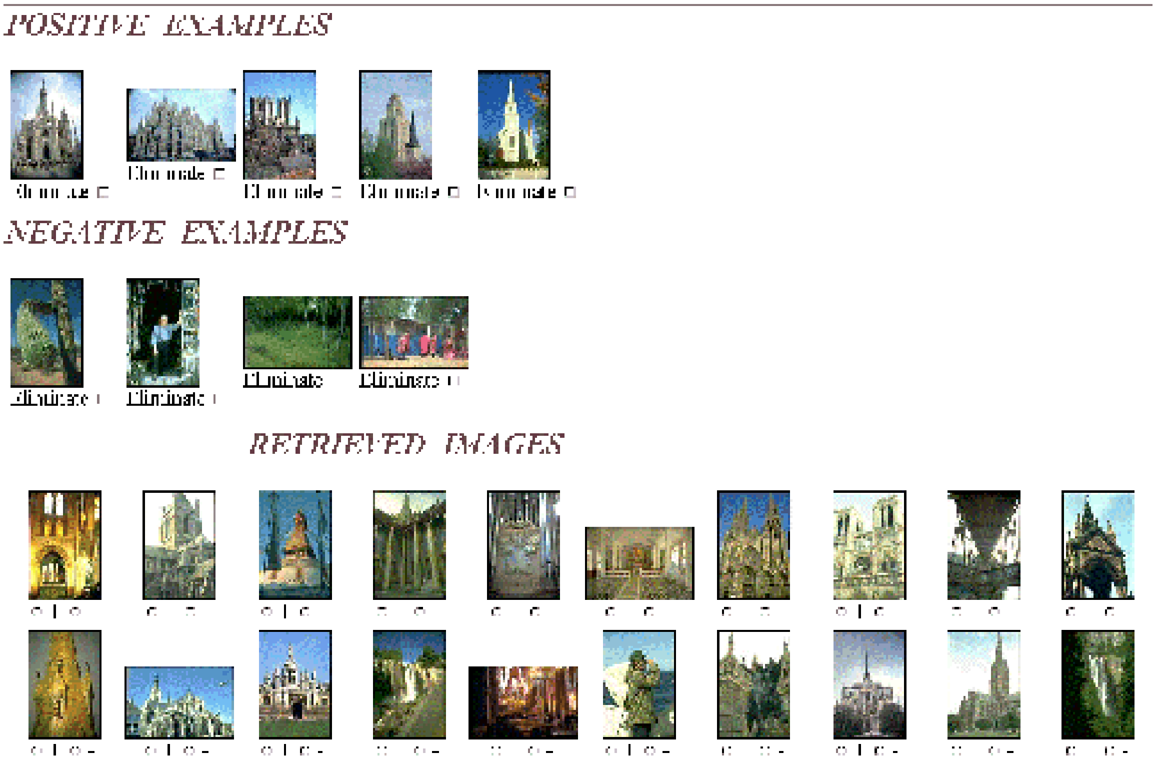

Figure2. This figure illustrates the complex feature method retrieving images of ``French gothic cathedrals''. At the top are the set of positive examples selected by the user. Below that are a set of negative examples selected by the user. At the bottom is a larger collection of images which have been returned by the system as matching the query.

The flexible templates approach constructs high-level global representations of image concepts -- such as a waterfall, cityscape or grassy field. The basic idea of the template is to break the image up into major regions (e.g., sky, sun, trees, fields, mountains). To represent the associated class of images, each region in the template can deform spatially to account for changes in position of the major components of an image. At the same time, regions of the template include relative photometric relationships which account for the spectral layout of an image while allowing for illumination variations. The amount of deformation, both spatial and photometric, needed to match a template to an image is an indication of the similarity of the images.

The complex feature approach attempts to simulate the processing that goes on in the visual cortex of human beings and primates. In the visual cortex there are many thousands of cells which each compute a different complex feature of the image. For example, cells have been found that respond to green texture on the right side of an image. This type of feature might be useful for finding forested scenes. Other cells detect faces in the center of the image. These are useful for finding a crowd of people. Since each of these complex features is selective for a small set of scenes, there must be many thousands to classify every scene. Our Complex Feature approach computes forty thousand features for every image. A particular query, for example ``find all images of jet aircraft'', is specified visually by selecting a small set of images containing jet aircraft. These images are analyzed to determine which of the 40,000 features consistently respond to jet aircraft (typically there are about 1000). These 1000 features are then used to screen the entire database.

Proposed Work

Acquiring a Visual Encyclopedia

The research challenges for constructing a visual encyclopedia are many. What is to be included? Where does the data come from? What are the natural entries for the encyclopedia? The early stages of encyclopedia construction will rely on two key insights: continuity and similarity. We believe that much can be learned by simply observing the world (or perhaps watching movies and TV). Continuity tells us that if we are looking at an object in one frame of a movie it is likely to be visible in the next frame, very likely. We can use this fact to observe many different angles of the same object. Similarity tells us that many classes of objects have instances that are similar in appearance (Mercedes Benz's and BMW's look reasonably similar). We can use this fact to cluster objects together into classes. Of course this is not invariably true: tables come in many different types. But it provides a great deal of helpful constraint. How then does anyone ever discover the that cats and dogs are different animals—since they look reasonably similar?

At some point the system must be told that cats are not the same as dogs, even though they have a similar appearance. This sort of information is available from user interaction during query refinement. Imagine that a user selects 2 or 3 dog images, and is returned images of dogs and cats. During refinement of this query the user selects some of the cats as negative examples. We can use this information to further refine the encyclopedia entry for dogs, an entry cannot be correct when it contains both negative and positive examples.

We propose to explore algorithms that will automatically construct an encyclopedia of visual concepts. These will work both in an unsupervised fashion by watching movies and television and in a semi-supervised fashion by observing user interaction during query refinement.

Synergy

Interestingly the flexible template approach and the complex feature approach are quite complementary. We propose to combine these technologies into a single unified system — *Flexible Templates of Complex Features*. The current Flexible Template technology uses relatively simple descriptions of image properties in its representation of image class. We propose to instead use the Complex Feature representation, which is capable of representing and discriminating a wider set of images properties. Such a unified system should provide a considerable improvement in the accuracy and sensitivity of our indexing methods.

Segmentation

For many types of queries we are not interested in retrieval using the entire image. Instead we want to retrieve images which contain particular objects, for example automobiles, lion cubs or suspension bridges. In these cases the confounding influence of the background must be removed before retrieval can succeed.

We propose to investigate a number of techniques for segmenting images into multiple regions. Recently Malik and Shi have shown that images can be decomposed into a dominant set of regions even before recognition has been performed (there has been similar work by Weiss and Adelson). One of the largest missing components from segmentation approaches has been an adequate theory of texture. Most segmentation algorithms assume that regions contain homogeneous or smoothly varying image properties. Our recent work (DeBonet and Viola) provides a new methodology for representing texture. We believe that these new insights into texture will significantly boost the performance of segmentation algorithms.

Learning Structure in Feature Space

The complex feature approach is unique in part because of the very large number of features, which are computed from an image. Much of its effectiveness is due to the very rich representation for images. Nevertheless because of its size, feature space can be very difficult to understand, modify, or control. For example, a user may wish to modify a query so that it focus on a few critical image properties, like low frequency texture, spatial organization, or color. In a framework where there are 40,000 features, it can be difficult to express this sort of simple constraint.

We propose to explore low dimensional representations for feature space. One possibility is principal components analysis (PCA), which works by projecting the data onto a few large variance eigenvectors. PCA has proven effective both in face recognition and text retrieval (e.g. Turk). Unfortunately, even the best algorithms for computing this representation would take weeks to compute on a single processor workstation. Our collaborators have recently constructed a scientific computing package that parallelizes Matlab operations across a large cluster of workstations (Husbands et al.). Using this code on a cluster of workstations we have computed eigenvector representations for text retrieval -- where the original vector space contain 100,000 dimensions (Isbell and Viola). More interestingly in the same publication we have proposed a number of other low dimensional representations: one based on the notion of independent components analysis (ICA) and the other based on clustering (e.g., Bell and Sejnowski). These representations yield superior retrieval performance in a text retrieval task and we intend to explore their utility image retrieval.

Potential Applications

Travel Agent / Tour Guide

Virtual reality content is rapidly expanding to include tours of museums, cities and perhaps countries. One advantage of virtual reality is that the user can view imagery in a natural way, by walking around and through it. One potential disadvantage is that it can be difficult to find exactly what you want. For example it will be possible to view the cathedral called Notre Dame in Paris using virtual reality. It might be more difficult to find other cathedrals which appear similar. Image database technology can be used to search virtual reality content much in the same way that it can be used to search for images.

Virtual Catalogs

There are many types of products which are distinguished principally based on appearance. In the fashion world there are many hundreds of manufacturers which produce shirts. Each of these is very similar in function, but potentially different in appearance. In the future it is likely that sales companies will put together virtual catalogs that contain the products of every known producer. The difficulty arises when the users tries to find the perfect shirt pattern among the many thousands which are available. Image database technology can be used both to organize shirts based on appearance, and retrieve the similar shirts if the selected shirt is no longer available. This technology should be similarly useful for pants, shoes, furniture, draperies, carpets, dishes and other housewares.

It should also be possible to search databases of real estate based on appearance, using images both of the building exterior/landscape as well as building interior. Given a very large number of images it may also be possible to construct a virtual reality tour of the house.

Trademarks and Copyrights

Companies often place great value in their trademarks as a way of differentiating their products. It is important that a new trademark be significantly different from previous trademarks. We believe image database technology can be used to automatically search for trademarks which are too similar to previous trademarks, where the comparison is based on visual content and visual similarity.

Similarly it is difficult to search out people who are infringing copyrighted material (e.g. a work of art or photograph). We believe image database technology can be used to detect copyright infringement. This is potentially of considerable value to museum curators, holders of photograph collections, and other owners of visual material.

Collaboration with NTT

The research will be conducted by researchers and students from the Learning and Vision Group, under the direction of Prof. Paul Viola. Within MIT there will be very close collaboration with Prof. Grimson and his research group. In addition we expect that we will collaborate closely with researchers at the Human Interface Laboratories of NTT. Dr. Shoji Kurakake of HIL has been proposed by Dr. Tohkura. His interests are quite similar to our own. We have attempted to contact him, but in the very short time given for proposal preparation, he has not been able to respond. Dr. Kurakake has proposed a similar project wherein text from video segments is to be automatically detected and read. This technology is very complementary to our own and we expect a fruitful collaboration.

The Director of the NTT Information Science Research Laboratory, Dr. Kenichiro Ishii, has also published a number of papers, which are directly related to this research. Dr. Hiroshi Murase, also at NTT-BRL, is also working on related problems in image database retrieval and computer vision. We believe that there are collaboration possibilities with researchers both in the Human Interface Laboratory and NTT-BRL.