Finding images in large collections

David Forsyth, Jitendra Malik, Margaret Fleck, Serge Belongie and Chad Carson

U.C. Berkeley,
Berkeley,
CA 94720
USA

Abstract

Digital libraries can contain hundreds of thousands of pictures and video sequences. Typically, users of digital libraries wish to recover pictures and videos from collections based on the objects and actions depicted: this is object recognition, in a form that emphasizes large, general modelbases, where new classes of object or action can be added easily.

We first describe a representation - the ``blobworld'' representation - that uses an image segmentation in terms of novel colour and texture features to represent an image in terms of a small number of coherent regions of colour and texture. The blobworld representation allows a powerful image retrieval paradigm at the composition level in which the user is allowed to view the internal representation of the submitted image and the query results.

We then show how one can use coherent regions to recover people and animals, using a representation called a body plan. This representation is adapted to segmentation and to recognition in complex environments, and consists of an organized collection of grouping hints obtained from a combination of constraints on color and texture and constraints on geometric properties such as the structure of individual parts and the relationships between parts. Body plans are part of a more general scheme of representation for object recognition, where images are segmented into regions that have a stylised structure in shape, shading, texture or motion; objects and actions are recognised by reasoning about the spatio-temporal layout of these primitives.

We will illustrate these ideas with examples of systems running on real collections of images.

Introduction

The recent explosion in internet usage and multi-media computing has created a substantial demand for algorithms that perform content-based retrieval. The vast majority of user queries involve determining which images in a large collection depict some particular type of object. Typical current systems abstract images as collections of simple statistics on colour properties; there is much work on user interfaces that support image recovery in this abstraction. Instead, we see the problem as focussing interest on poorly understood aspects of object recognition, particularly classification and top-down flow of information to guide segmentation.

Current object recognition algorithms cannot handle queries as abstract as ``find people,'' because all are based around a search over correspondence of geometric detail, whereas typical content-based-retrieval queries require abstract classification, independent of individual variations. Existing content based retrieval systems perform poorly at finding objects, because they do not contain codings of object shape that are able to compensate for variation between different objects of the same type (e.g. a dachshund and a dalmatian), changes in posture (e.g. sitting or standing), and changes in viewpoint. Furthermore, because of the poor or absent shape representation, combinations diagnostic for particular objects cannot be learned.

Blobworld

Building satisfactory systems requires automatic segmentation of significant objects. Natural segmentations should produce regions that have coherent colour and texture. We use the Expectation-Maximization (EM) algorithm to perform automatic segmentation based on image features. EM iteratively models the joint distribution of color and texture with a mixture of Gaussians; the resulting pixel-cluster memberships provide a segmentation of the image into regions where colour and texture are coherent.

After the image is segmented into regions, a description of each region's color, texture, and spatial characteristics is produced. Regions are represented as blobs of colour and texture; an image is a composite of blobs. In a querying task, the user can access the regions directly, in order to see the segmentation of the query image and specify which aspects of

promising because blobworld captures the important elements of an image---the objects it contains---rather than simply encoding overall stuff properties.

In our system, the user composes a query by submitting an image to the segmentation/feature extraction algorithm in order to see its blobworld representation, selecting the blobs to match, and finally specifying the relative importance of the blob features. The user may also submit blobs from several different images. (For example, a query might be the disjunction of the blobs corresponding to airplanes in several images, in order to provide a query that looks for airplanes of several shades.)

We define an ``atomic query'' as one which specifies a particular blob to match (e.g., ``like-blob-1''). A ``compound query'' is defined as either an atomic query or a conjunction or disjunction of compound queries (``like-blob-1 and like-blob-2''). We might expand this definition to include negation (``not-like-blob-1'') and to allow the user to specify two blobs with a particular spatial relationship as an atomic query (``like-blob-1-left-of-blob-2''). Once a compound query is specified, we score each database image based on how closely it satisfies the compound query.

We then rank the images according to overall score and return the best matches, indicating for each image which set of blobs provided the highest score; this information will help the user refine the query. After reviewing the query results, the user may change the weighting of the blob features or may specify new blobs to match.

Blobworld is a significant improvement on colour histogram based methods, both because it allows a more detailed representation of image properties and layout, and because there is a clear path for building more complex queries.

Body plans

People and many animals can be viewed as an assembly of nearly cylindrical parts, where both the individual geometry of the parts and the relationships between parts are constrained by the geometry of the skeleton and ligaments. These observations suggest the use of a representation that emphasizes assemblies of a constrained class of primitive.

Much information is available to support segmentation and recognition: firstly, segments must be coherent, extended and have near parallel sides with an interior that appears to be hide or skin; secondly, because the 3D relationships between segments are constrained, there are relatively few assemblies of 2D segments. As a result, it is possible to tell whether a person or animal is present by determining whether there is an assembly of image segments that (a) have the right colour and texture properties and (b) form an assembly that could be a view of an acceptable configuration.

A body plan is a sequence of grouping stages, constructed to mirror the layout of body segments in people and animals. To tell whether a picture contains a person or an animal, our program attempts to construct a sequence of groups according to the body plan. For example, in the case of horses the program first collects body, neck and leg segments; it then constructs pairs that could be views of a body-neck pair, or a body-leg pair; from these pairs, it attempts to construct triples and then quadruples.

At each stage of the plan, a predicate is available which tells whether a group could correspond to some view of the segments described. For a sufficiently large collection of segments, the fact that such predicates are non-trivial follows from the existence of kinematic constraints on mammalian joints. We use a simple learning strategy for learning these predicates.

We have built two systems to demonstrate the approach. The first can very accurately tell whether an image contains a person wearing little or no clothing; the second can tell whether an image contains a horse. In each case, the approach involves pure object recognition; there is no attempt to exploit textual cues or user interaction.

Lightly clad people

The system segments human skin using colour and texture criteria, assembles extended segments, and uses a simple, hand built body plan to support geometric reasoning. A prefilter excludes from consideration images which contain insufficient skin pixels. Performance was tested using 565 target images of sparsely clad people. The system was controlled against a total of 4302 assorted control images. If images are selected on the basis of the number of skin pixels only, 448 test images are marked, but 485 control images are marked. The most selective choice of geometric test marks 241 test images and only 182 control images - almost twice as selective.

Horses

The horse system segments hide using colour and texture criteria and then assembles extended segments using a body plan to support the geometric reasoning. This body plan was learned using a bounding box classifier; the topology of the body plan was given in advance. If images are recovered on the number of hide pixels alone, 85 test images and 260

recovers 11 test images and only 4 control images. While the recall is relatively low, the selectivity is very high, meaning that the system effectively extracts image semantics.

The results are good, taking into account the abstraction of the query and the generality of the control images. The program is a practical, but not perfect, tool for extracting semantics.