Finding People and Animals by Guided Assembly

D.A. Forsyth

Computer Science Division
U.C. Berkeley
Berkeley, CA 94720
daf@cs.berkeley.edu

M.M. Fleck

Department of Computer Science
University of Iowa
Iowa City, IA 52240
mfleck@cs.uiowa.edu

Abstract

This paper describes a new representation for people and animals, called a body plan. The representation is an organized collection of grouping hints obtained from constraints on color, texture, shape, and geometrical relations. Body plans can be learned from image data, using established statistical learning techniques.

Body plans are well adapted to segmentation and recognition in complex environments, such as the huge libraries of digitized images now becoming widely available. Two specific applications of body plans are presented: an algorithm that determines whether an image depicts a scantily clad human and an algorithm that learns and uses a body plan to find pictures of horses. Both algorithms demonstrate excellent performance on large, poorly controlled input data.

Keywords: *Object Recognition, Computer Vision, Content based retrieval, Image databases, Learning in vision*

1 Introduction

The recent explosion in internet usage and multimedia computing has created a substantial demand for algorithms that perform *content-based retrieval*. 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. In other words, they require algorithms which can perform object recognition, using large, general model-bases, to which new classes of object or action can easily be added.

Typical content-based retrieval systems, reviewed briefly along with user requirements in [6], model images as collections of two dimensional coloured and textured shapes. This paradigm has motivated extensive work on user interfaces that support image recovery. However, these systems perform poorly at finding objects, because they do not represent object shape in a way that compensates 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. Nor can their impoverished shape representations be used to learn combinations of shape features diagnostic for particular objects.

Current object recognition algorithms also cannot handle abstract queries such as as "find people." They are based on a search for correspondences between geometric details. But these details are not preserved between different views of an object, or between different objects of the same class. Robust performance on large image collections requires a more flexible theory of shape.

Building satisfactory systems also requires automatic segmentation of significant objects. Typical recent systems for finding people or animals typically simplify segmentation using either motion cues or a known or simplified background (e.g. [8], which segments by subtracting a known background). The automatic segmentation literature has traditionally concentrated on describing images as regions of coherent colour or texture, whereas the notion of segmentation appropriate to our present application is: "find the image regions that come from a single object of the required class," a process that is impossible without model information. Content-based retrieval requires segmentation of very general images. Therefore, our approach attempts to marshal as much model information as possible at each segmentation stage.

2 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 representing objects as assemblies of a constrained

class of primitive: cylinders or generalized cylinders. Typical versions of this idea appear in [1, 2, 3, 7, 9]. Previous algorithms, however, have displayed little practical success, because their models contain insufficient information to support robust segmentation.

The key to success is for segmentation and recognition to use all the information and constraints available from the object model. People, and many animals, are made up of segments which have a coherent material properties (color, texture, and shading), which have extended and near parallel sides, and whose interior appears to be hide or skin. Furthermore, constraints on the relationships between segments in the 3D object model imply that relatively few assemblies of 2D segments are consistent with a particular type of object. 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 rules, 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 (using the plan given in figure 1) 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.

When applying each rule, 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 kinematic constraints on mammalian joints imply that the predicate will fail on many false groups. We use a simple learning strategy for learning these predicates. A more detailed description of the implementation appears in in [5].

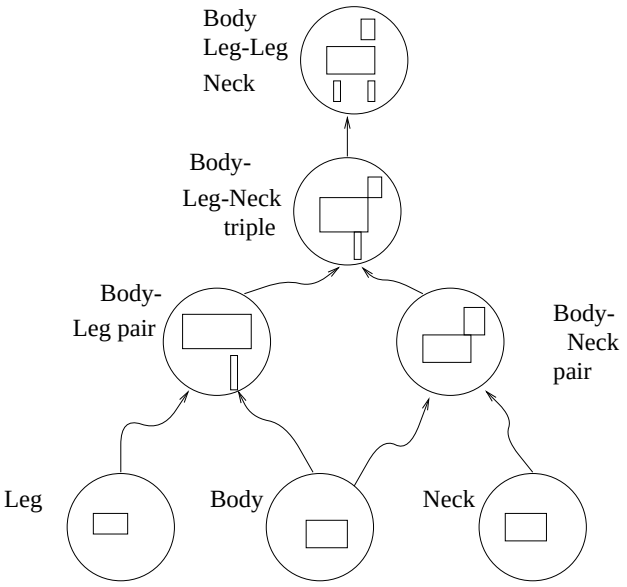

Figure 1: *The body plan used for horses. Each circle represents a classifier, with an icon indicating the appearance of the assembly. An arrow indicates that the classifier at the arrowhead uses segments passed by the classifier at the tail. Note that constraints exist between groups, too; for example, a body-leg-neck classifier will attempt to form triples only out of pairs that share the same body.*

3 Experimental results

We have built two systems to demonstrate the new 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.

In information retrieval, it is traditional to describe the performance of algorithms in terms of *recall* and *precision.* The algorithm's recall is the percentage of test items marked by the algorithm. Its precision is the percentage of test items in its output. Unfortunately, the precision of an algorithm depends on the percentage of test images used in the experiment: for a fixed algorithm, increasing the density of test images increases the precision. In our application, the density of test images is likely to vary and cannot be accurately predicted in advance.

To assess the quality of our algorithm, without dependence on the relative numbers of control and test images, we use a combination of the algorithm's recall and its *response ratio.* The response ratio is defined to be the percentage of test images marked by the algorithm, divided by the percentage of control images marked. This measures how well the algorithm, acting as a filter, is increasing the density of test images in its output set, relative to its input set.

3.1 Sparsely clad humans

The basic structure of our system is described in [4], which describes the body plan used; the experimental results given here are more recent. 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 collected from the internet and by scanning or re-photographing images from books and magazines. Images were not pre-screened for particular poses or photographic conditions. However, only images encoded using JPEG compression were selected because GIF compression, the other common encoding method, reproduces colors poorly. Test images were automatically reduced to fit into a 128 by 192 window, and rotated as necessary to achieve the minimum reduction. 4302 assorted images were used as controls.

If images are selected only on the basis of the number of skin pixels only (see [4]), 448 test images are marked (a recall of 79 %) but 485 control images are marked. Thus, the response ratio is 7. When geometrical information is also included, 241 test images are marked (a recall of 42 %) but only 182 control images, producing a response ratio of 10. The selectivity of the system increases as the geometric complexity of the groups required to identify a person increases, suggesting that the presence of a sufficiently complex geometric group is an excellent guide to the presence of a person. The limited recall probably reflects the fact that the current implementation does not include all important geometrical structures (e.g. those found in side views).

3.2 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, which is shown schematically in figure 1 was learned using a bounding box classifier; the topology of the body plan was given in advance. The classifier was learned using 102 acceptable groups, drawn from 38 images; the risk associated with a false negative was assumed to be zero, so that the classifier is simply the bounding box of this set.

Performance was tested using 100 target images selected from CD 113000 ("Arabian horses") in the Corel stock photo library, and 1086 unrelated control images from the Corel stock photo library. All test and control images fit into a 128 by 192 window. A hide filter, similar to the skin filter but using different parameter settings, marks pixels that are likely to be hide. Images which contain insufficient hide pixels are excluded from consideration.

Ribbon finding for horse images is complicated by the need to find legs, which are relatively narrow. Looking for narrow ribbons can generate very large numbers of local symmetries, to the point where our current ribbon grouping algorithm is overwhelmed. This occurred on 13 test images and 116 control images that had already passed the hide filter, an unusually large number. Performance of the hide filter is estimated including these images. Performance of the grouper is estimated excluding these images, as well as images used to learn the geometrical models (34 test images). Overall performance is estimated by multiplying the two separate recall and precision figures.

For the case of people, the classifier asserts that a person is present if a sufficiently complex geometric group is present. In the case of horses, a considerable improvement in performance can be obtained by noting that, if a sufficiently large set of segments is passed to the final classifier (for example, for an image of a horse in front of a fence, where many ribbons must be found), it is likely to mark a horse erroneously. Thus, for a picture to be marked as containing a horse, we require that (a) at least one body-leg-leg-neck group be present and (b) that the ratio of the number of such groups to the number of groups presented to the final stage, be larger than a parameter, which for convenience we call the robustness parameter.

If images are recovered purely on the basis of number of hide pixels, 85 test images and 260 control images are marked, for a recall of 85 % and a response ratio of 3.5. When geometrical information is also used, the system displays a recall of 15% and a response ratio of 23. The high response ration means that the system effectively extracts image semantics. Figure 2 shows the images marked as containing horses. Note the horses returned are in a variety of aspects. Nevertheless, only very few control images are returned and one of them contains an animal that looks a lot like a horse. Table 1 shows the body plan is efficient.

4 Discussion and Conclusions

We have demonstrated a representation for people and animals in terms of primitives and their geometric relations. The representation provides grouping information at the image level; we have demonstrated that this representation can be learned from examples, is robust to variations in aspect, and is effective at quite abstract recognition queries because it emphasizes within-class similarities of structure over geometric detail. These 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.

Much remains to be done. The present system involves one classifier for horses, and another for people.

Figure 2: *All images returned from a control set of 1086 and a test set of 100 images. The first line of horse images comes from the training set; the rest from the test set. A further four control passed the hide filter but overwhelmed the ribbon finding algorithm.*

While the structure of the classifiers contains many teasing analogies, it is not yet obvious how one uses these similarities to build a single process that, as ribbons are accreted into an assembly, can tell a horse from a person, while using the same underlying set of activities.

Finding a group does not extract all information. In future versions of this program we expect to be able to tell not only that a person or animal is present, but, by looking in greater detail at the segment relationships, what they are doing.

Acknowledgements

We thank Joe Mundy for suggesting that the response of a grouper may indicate the presence of an object and Jitendra Malik for many helpful suggestions. Portions of this research were supported by the National Science Foundation under grants IRI-9209728, IRI-9420716, IRI-9501493, under a National Science Foundation Young Investigator award, an NSF Digital Library award IRI-9411334, and under an instrumentation award CDA-9121985.

References

[1] Connell, Jonathan H. and J. Michael Brady "Generating and Generalizing Models of Visual Objects," *Artificial Intelligence* 31/2, pp. 159–183, 1987

[2] Binford, T.O., " Visual perception by computer," *Proc IEEE Conf. Systems Control*, 1971.

$\overline{n_4}$	$\overline{n_c}$	$\overline{n_c}/\overline{n_4}$	$\overline{(n_c/n_4)}$
2,500,000	511	0.0002	0.006

Table 1: *Body plans are efficient; the number of segment groups handled by the final classifier is very much less than the total number of four segment groups. Efficiency of body plans can be measured in two ways; n_4 is the number of four segment groups in an image, n_c is the number of calls to the final classifier of the body plan, and an overbar denotes the mean over all test and control images that could be presented to the grouper. $\overline{(n_c/n_4)}$ tends to underestimate the efficiency, because it penalises images where there are very few groups. Clearly, by either statistic, body plans are a significant improvement over simple classifiers, at no cost in empirical risk.*

[3] Binford, T.O., "Body-centered representation and perception," *ProceedingsObject Representation in Computer Vision.*, Hebert, M. *et al.* (eds), Springer Verlag, 1995.

[4] M.M. Fleck, D.A. Forsyth and C. Bregler, "Finding naked people," *Proc. European Conf. on Computer Vision*, Edited by: Buxton, B.; Cipolla, R. Berlin, Germany: Springer-Verlag, 1996. p. 593-602

[5] Forsyth, D.A. and Fleck, M.M., "Body Plans," *Proc. CVPR-97*, 1997.

[6] Forsyth, D.A., Malik, J., Fleck, M.M., Greenspan, H., Leung, T., Belongie, S., Carson, C. and Bregler, C., "Finding pictures of objects in large collections of images," *Proc. 2'nd International Workshop on Object Representation in Computer Vision*, April, 1996.

[7] Marr, D., and Nishihara, H.K., "Representation and Recognition of the Spatial Organization of Three-Dimensional Shapes", *Proc. Roy. Soc. B*, **B-200**, 269-294, 1977.

[8] Wren, C., Azabayejani, A., Darrell, T. and Pentland, A., "Pfinder: real-time tracking of the human body," MIT Media Lab Perceptual Computing Section TR 353, 1995.

[9] Zerroug, M. and Nevatia, R., "Three-dimensional part-based descriptions from a real intensity image," *Proceedings of 23rd Image Understanding Workshop*, 1994.

[10] Zisserman, A., Forsyth, D.A., Mundy, J.L., Rothwell, C.A., and Liu, J.S., "3D Object Recognition using Invariance," *Artificial Intelligence*, **78**, 239-288, 1995.