Body Plans

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 representation for people and animals, called a body plan, which is adapted to segmentation and to recognition in complex environments. The representation is 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 can be learned from image data, using established statistical learning techniques. The approach is illustrated with two examples of programs that successfully use body plans for recognition: one example involves determining whether a picture contains a scantily clad human, using a body plan built by hand; the other involves determining whether a picture contains a horse, using a body plan learned from image data. In both cases, the system demonstrates excellent performance on large, uncontrolled test sets and very large and diverse control sets.

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.* The vast majority of user queries involve determining which images in a large collection depict some particular type of object. Typical current systems, which are reviewed briefly along with user requirements in [10], do not support this type of query. Current object recognition algorithms cannot handle queries as abstract as "find people." In turn, existing content based retrieval systems perform poorly at finding objects. Building satisfactory systems requires automatic segmentation of significant objects from complex backgrounds. Typical recent systems for finding people or animals typically simplify segmentation using either motion cues or a known or simplified background (e.g. [15]). 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; 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 the use of a representation that emphasizes assemblies of a constrained class of primitive; typical versions of this idea appear in [3, 4, 2, 16]. Another version appears in [11], which represents people and animals by cylinders at a variety of scales; they suggest finding a person by finding a large extended cylinder, which is then resolved into smaller cylinders forming limbs and torso, and so on to fingers and toes. The approach is impractical, not least because the models contain little information to support segmentation and little actual constraint.

In an image, segments must be coherent, extended and have near parallel sides with an interior that appears to be hide or skin; because the 3D relationships between segments are constrained, there are relatively few acceptable assemblies of image segments. Thus, for a person or animal to be present there must be 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 (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.

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 statistical learning technique to infer an approximate representation of possible configurations from a variety of example views, producing a classifier that could, given an assembly, tell whether it represented a possible view. The advantage of this approach is that techniques for building effective classifiers quite efficiently are well established (e.g. [14]), and that variations from individual to individual could be captured with a sufficiently large data set.

Statistical learning theory is notoriously unconcerned with the computational efficiency of the classifiers constructed (the introduction in [8] is fairly typical). This is a serious problem: telling whether an image contains a horse, for example, appears to require groups of at least four straight ribbons, and searching over all groups of four straight ribbons is impractical for typical images. However, a body plan can be viewed as a sequence of classifiers, where each predicate is a classifier for some sub-assembly. Building classifiers for various sub-assemblies ensures that only very few groups are tested at the final stage. The hierarchical structure has the advantage that, if is not possible to add segments to an assembly, there is still a working hypothesis about the identity of the assembly. As figure 1 shows, body plans are efficient.

3 Learning a body plan

Rigorous statistical principles for learning have been established over the last 20 years; good introductions appear in [14, 8]. There are two principles: that samples of a distribution provide a representation that converges quite quickly in probability to that distribution, and that formalising the effects of changes in parameter on a decision boundary using the Vapnik-Chervonenkis dimension results in a prediction of the

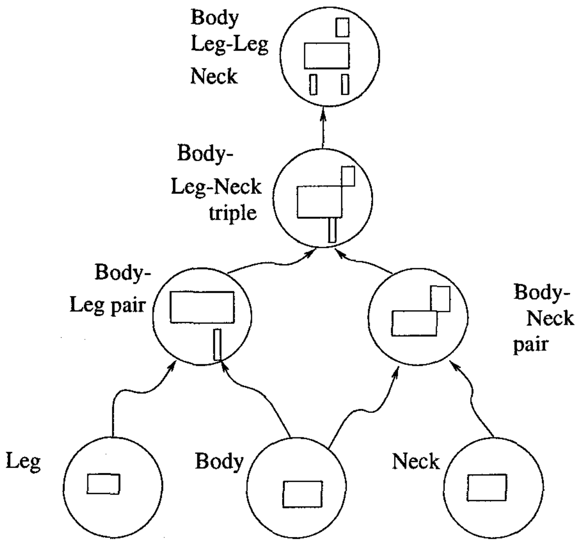

Figure 1: *The body plan used for horses. Each circle represents a classifier, with an icon indicating the appearance of the assembly. An arrow indicates that the classifier at the arrowhead uses segments passed by the classifier at the tail. Constraints exist between groups, too; for example, a body-leg-neck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

future risk of using the classifier that also converges in probability.

As a result, it is in priniciple possible to produce a classifier that results in low risk both on the training set (known as *empirical risk*) and predicted for future use; typically such classifiers are trained using a large number of samples compared to the V-C dimension of the class of decision boundaries used. An important principle is to keep the V-C dimension of the class of decision boundaries used as small as possible, involving both thoughtful selection of features, and the incorporation of as much *a priori* knowledge as possible; this point alone justifies representations in terms of primitives.

We train body plans to achieve a minimum of risk on the training set (the criterion is usually known as *empirical risk*). In general, the individual classifiers in a body plan cannot be trained separately using this criterion, because determining the effect of a change in a given classifier's parameters on overall risk requires knowing (a) what later classifiers will do with the assembly the given classifier accepts and (b) what the distribution of assemblies leaving earlier classifiers looks like.

Since the individual classifiers in a body plan are defined by sub-assemblies of the main group, one can train them all simultaneously to get a minimum of empirical risk by constructing an augmented feature vector, containing all data that all classifiers will see. A single classifier is then trained on this augmented feature vector; once it has been trained, by projecting its decision boundary onto the features associated with each separate assembly, we obtain the component sub-classifiers.

The process can be described formally using the following notation: the final assembly is a group of k elements; there is a function f_i which computes the feature vector associated with a group of i elements (as section 4 indicates, this function will change with the number of elements but is independent of the elements themselves); the j'th example is g_j^{k1}; and the l'th subgroup of i elements drawn from g_j^{k1} is g_j^{il}.

Now consider the augmented feature vector for example j given by:

$$\mathbf{v}_j = (f_1(g_j^{11}), f_1(g_j^{12}), \ldots f_2(g_j^{21}), \ldots f_i(g_j^{il}), \ldots f_k(g_j^{k1}))$$

and write the projection of this vector onto the space spanned by the terms corresponding to $f_i(g_j^{il})$ as $\pi_{il}(\mathbf{v}_j)$. The elements of this vector are the feature vectors for all i-fold combinations taken from the group. Assume that a classifier is trained to obtain a minimum of empirical risk on a set of such vectors, yielding a decision surface $S = 0$.

Now consider a point in the space $\pi_{il}(\mathbf{v}_j)$; a classifier for sub-assemblies should accept this point if, by attaching any other assemblies, it was possible to obtain a group for which $S >= 0$, and should reject it otherwise. But such a classifier can be obtained by projecting S into this space; the singular set under this projection forms a set of possible components for the decision boundary, which must be sorted to ensure that the criterion described holds.

For S algebraic, this is cylindrical algebraic decomposition [1]. By constructing this decomposition of S, we obtain a set of sub-assembly classifiers that achieves the same empirical risk as S does, but is potentially computationally more efficient. Using classifiers that project well is wise. Classifiers that have decision boundaries that consist of unions of axis-aligned boxes are known to have low V-C dimension, perform well ([8], chap. 20), and project particularly easily.

4 Describing shape

We use colour and texture properties, documented in greater detail in [9] to identify image regions which could be skin or hide. A version of Canny's [7] edge detector, with relatively high smoothing and contrast thresholds, is applied to the resulting areas to obtain a set of connected edge curves. Pairs of edge points with a near-parallel local symmetry [5] are found by a straightforward algorithm, and sets of points forming regions with roughly straight axes ("straight ribbons," after [6]) are found using an algorithm based on the Hough transformation.

The ribbons are abstracted as oriented rectangles, whose width is given by the average width along the ribbon, and whose length and axis come from the ribbon spine, hiding individual variations in segment cross-section. In this case, scaled orthography is an acceptable camera model for all practical views. Shape measurements for segment groups are then obtained using a canonical frame (as in, say, [17]). A distinguished segment - usually a body segment - is chosen to have its center of gravity at the origin, and is rotated and flipped so that (a) it is axis aligned and (b) other segments lie in particular quadrants. Any measurement in such a frame is invariant; in these frames, the variations between shapes is surprisingly small.

Variations in appearance with changes in viewpoint are a primary difficulty in object recognition. Body plans are intrinsically relatively robust to these effects, as our experimental results show (see figure 4). This robustness comes from two main sources: firstly, the underlying primitives have no significant view-variation in appearance; secondly, the kinematics of the assemblies are such that complex inter-primitive occlusions are not possible, suppressing a rich source of difficulties

For example, foreshortening between a lateral and a three-quarter view of a horse is of the same order of magnitude as the noise in obtaining the length of segments, while the layout of a frontal view and a lateral view is basically the same. However, in the horse example, views that are notably absent are head-and-shoulders views and overhead views; the approach would fail to isolate horses in such images. The deficit can be dealt with by adding classifiers to the body plan. We have no results on how many such classifiers are required but believe that relatively few are required.

5 Experimental results

We have built two systems to demonstrate the approach. The first can very accurately tell whether an image contains a naked person; 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.

To assess the quality of our algorithm, without de-

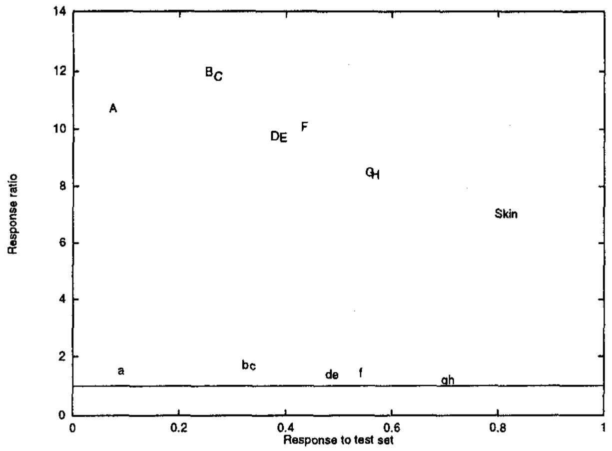

Figure 2: *The response ratio, (percent incoming test images marked/percent incoming control images marked), plotted against the percentage of test images marked, for various configurations of the naked people finder. Labels "A" through "H" indicate the performance of the entire system of skin filter and geometrical grouper together. The label "skin" shows the performance of the skin filter alone. The labels "a" through "h" indicate the response ratio for the corresponding configurations of the grouper; because this number is always greater than one, the grouper always increases the selectivity of the overall system. The cases differ by the type of group required to assert that a naked person is present. The horizontal line shows response ratio one, which would be achieved by chance. The response ratio increases, and the recall decreases, as the geometric complexity of the groups required to identify a person increases, suggesting: (1) finding a sufficiently complex geometric group yields the object (2) that the body plan used omits important geometric structures.*

pendence 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.

5.1 Naked humans

The basic structure of our system is described in [9], which describes the body plan used; the experimental results given here are new and much more comprehensive. 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 naked people collected from the internet and by scanning or re-photographing images from books and magazines. There was no pre-sorting for content. Test images were automatically reduced to fit into a 128 by 192 window, and rotated as necessary to achieve the minimum reduction. The system was controlled against a total of 4302 assorted control images, containing some images of people but none of naked people.

Figure 2 graphs response ratio against response for a variety of configurations of the grouper. The recall of a skin-filter only configuration is high, at the cost of poor response ratio. Configurations G and H require a relatively simple configuration to declare a person present (a limb group, consisting of two segments), decreasing the recall somewhat but increasing the response ratio. Configurations A-F require groups of at least three segments. They have better response ratio, because such groups are unlikely to occur accidentally, but the recall has been reduced. The selectivity of the system increases, and the recall decreases, as the geometric complexity of the groups required to identify a person increases, suggesting that our representation used in the present implementation omits a number of important geometric structures and that the presence of a sufficiently complex geometric group is an excellent guide to the presence of an object.

5.2 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, that was projected as described above to yield appropriate subclassifiers; the topology of the body plan was given in advance. The body plan uses geometric measurements in a canonical frame to describe segment groups. Each classifier is a bounding box classifier - segment groups are accepted if they lie in an axis aligned bounding box, and are rejected otherwise.

The body plan is trained using augmented feature vectors, and *assuming that false positives carry no risk* (an assumption that simplifies training the classifier, and appears to be justified by the tight constraint placed on the groups). This box is then projected onto the feature spaces defined by the subgroups, and the resulting boxes define the individual assemblies in the body plan. This approach makes training extremely simple, and yields an effective representation. The classifier was learned using a total of 102 acceptable groups, drawn from 38 images.

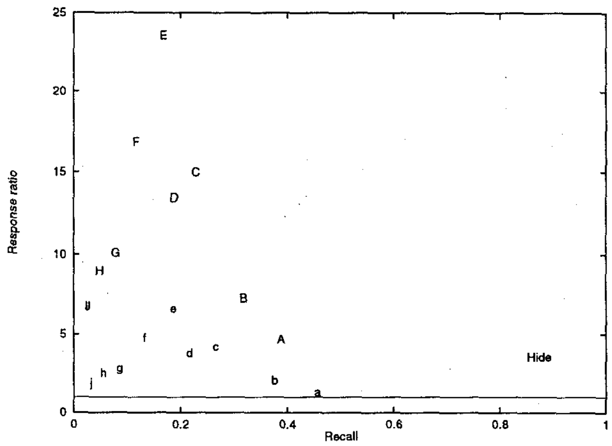

Figure 3: *The response ratio, (percent incoming test images marked/percent incoming control images marked), plotted against the percentage of test images marked, for various configurations of the horse finder. Labels "A" through "J" indicate the performance of the entire system of hide filter and geometrical grouper together; each label corresponds to a different value of the robustness parameter described in the text, where the parameter value increases from "A" to "J" in even steps. The label "hide" shows the performance of the hide filter alone. The labels "a" through "j" indicate the response ratio for the corresponding configurations of the grouper alone; because this number is always greater than one, the grouper always increases the selectivity of the overall system. The horizontal line shows response ratio one, which would be achieved by chance. The grouper displays relatively low recall, but the groups are clearly extremely distinctive.*

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 that library. All test and control images fit into a 128 by 192 window. Our system first excludes images which contain insufficient hide pixels, leaving 85 test images and 260 control images.

Looking for narrow ribbons (legs) can generate very large numbers of local symmetries, to the point where ribbon grouping is overwhelmed. This occurred for a total of 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 and excluding images used for learning (a total of 34 test images); and overall performance is estimated by multiplying the two separate recall and precision figures. This is

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

Table 1: *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 images 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 more efficient than handling all four segment groups, at no cost in empirical risk.*

the fairest approach to estimating performance in this case, where the difficulty is clearly an implementation error, and the training set is usually better excluded in estimating performance.

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.

For a good choice of the robustness parameter, the system displays a recall of 15% and a response ratio of 23; while the recall is relatively low, the response ratio is very high, meaning that the system effectively extracts image semantics. The effects of ribbon finding difficulties make it hard to represent the result set exactly, but figure 4 shows the images recovered for this case. Note the horses returned are in a variety of aspects, for a large control set there are very few control images returned (the high response ratio ensures this) and that one of the control images returned contains an animal that looks a lot like a horse.

As figure 3 shows, increasing the robustness parameter leads to decreased recall, but better response ratio; we envisage a user setting a value according to whether they require many test images, but are tolerant of false positives, or would prefer a more focussed set of responses.

6 Discussion and Conclusions

The results are good, taking into account the abstraction of the query and the generality of the control images; for example, a group of 1000 control and 100 test images (a realistic test) presented to the horse program would result in about 15 test images and 7

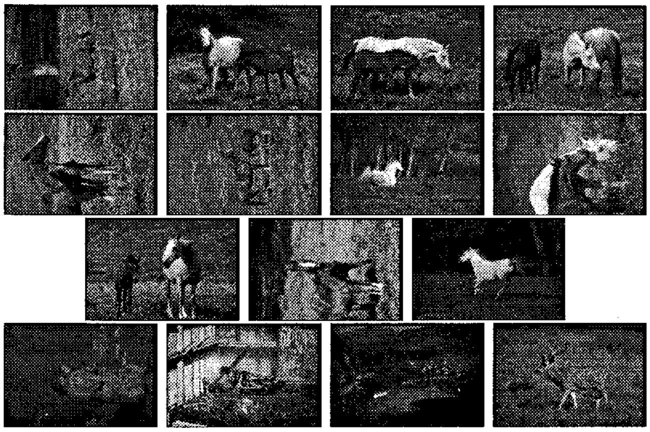

Figure 4: *All images returned from a control set of 1086 and a test set of 100 images, for the horse query with robustness parameter set to the most selective value. The first line of horse images comes from the training set; the rest from the test set. A further four control images could be expected to come from the images that passed the hide filter, but overwhelmed the ribbon finding algorithm. The test images recovered contain horses in a wide range of aspects; one control image contains an animal that might reasonably pass for a horse.*

control images returned, meaning the program is a practical, but not perfect, tool for extracting semantics. Much remains to be done. The description of primitives is impoverished, and incorporates no shading information. In particular, the main differences between, say, leopards and horses is in the appearance of their pelts; clearly, segmenting leopards from general backgrounds requires further work. Some promising lines of attack on this problem are sketched in [10].

The present system involves one classifier for horses, and another for people. 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. As our results show, even in the present quite primitive form, body plans enhance model information by organising it into a form that aids segmentation and grouping, and simplifies learning; the result is a representation that is clearly capable of extracting semantic information from an image for two difficult and abstract cases.

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] Arnon, D.S.; Collins, G.E.; McCallum, S. "Cylindrical algebraic decomposition. I. The basic algorithm." *SIAM Journal on Computing*, **13**, 4:865-77, 1984.

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

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

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

[5] Brady, J. Michael and Haruo Asada (1984) "Smoothed Local Symmetries and Their Implementation," *Int. J. Robotics Res.* 3/3, 36–61.

[6] Brooks, Rodney A. (1981) "Symbolic Reasoning among 3-D Models and 2-D Images," *Artificial Intelligence* 17, pp. 285–348.

[7] Canny, John F. (1986) "A Computational Approach to Edge Detection," *IEEE Patt. Anal. Mach. Int.* 8/6, pp. 679–698.

[8] Devroye, L., Gyorfi, L., and Lugosi, G., *A probabilisitic theory of pattern recognition*, New York : Springer, 1996.

[9] 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

[10] 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.

[11] 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.

[12] Minka, T.P.; Picard, R.W., "Interactive learning with a "Society of Models"." *CVPR-96*, 447-452, 1996

[13] Picard, R.W. and Minka, T. "Vision texture for annotation," *J. Multimedia systems*, **3**, 3-14, 1995.

[14] Vapnik, V., *The nature of statistical learning theory*, Springer-Verlag, 1996.

[15] 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.

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

[17] 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.