Finding objects by grouping primitives

D.A. Forsyth S.Ioffe J.Haddon

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

Abstract

We describe the use of a representation, called a body plan, to segment and to recognize people and animals 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.

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. The mechanism of recognition by assembly is very general; we describe recent work on finding clothing by marking folds and then assembling groups of folds.

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

Several typical collections containing over ten million images are listed in [3]. There is an extensive literature on obtaining images from large collections using features computed from the whole image, including colour histograms, texture measures and shape measures. However, in the most comprehensive field study of usage practices (a paper by Enser [3] surveying the use of the Hulton Deutsch collection), there is a clear user preference for searching these collections on image semantics. An ideal search tool would be a quite general recognition system that could be adapted quickly and easily to the types of objects sought by a user. Building such a tool requires a much more sophisticated understanding of the process of recognition than currently exists.

Object recognition will not be comprehensively solved in the foreseeable future. Solutions that are good enough to be useful for some cases in applications are likely, however. Querying image collections is a particularly good application, because in many cases no other query mechanism is available — there is no prospect of searching all the photographs by hand. Furthermore, users are typically happy with low recall queries - in fact, the output of a high-recall search for "The President" of a large news collection would be unusable for most application purposes. Our research focuses on areas that form a significant subset of these queries where useful tools can reasonably be expected.

1 Body plans

A natural implicit representation to use for people and many animals is a *body plan* — a sequence of grouping stages, constructed to mirror the layout of body segments. These grouping stages assemble image components that could correspond to appropriate body segments or other components (as in figure 1, which shows the plan used as an implicit representation of a horse). Having a sequence of stages means the process is efficient: the process can start with checking individual segments and move to checking multi-segment groups, so that not all groups of four (or however many for the relevant body plan) segments are presented to the final classifier. We have done extensive experiments with two separate systems that use the same structure:

- Images are masked for regions of appropriate colour and texture.
- Roughly cylindrical regions of appropriate colour and texture are identified.

0-7803-5148-7/98/$10.00©1998 IEEE

- Assemblies of regions are formed and tested against a sequence of predicates.

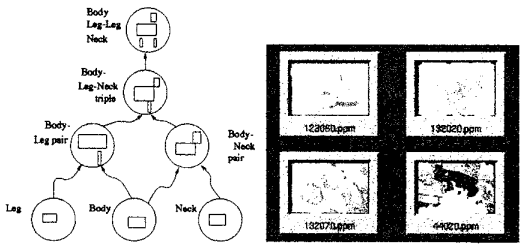

Figure 1: *On the left, the body plan used for horses. Each circle represents a classifier, with an icon indicating the appearance of the assembly. An arrow indicates that the classifier at the arrowhead uses segments passed by the classifier at the tail. The topology was given in advance. The classifiers were then trained using image data from a total of 38 images of horses. On the right, typical images with large quantities of hide-like pixels (white pixels are not hide-like; others are hide-like) that are classified as* not *containing horses, because there is no geometric configuration present. While the test of colour and texture is helpful, the geometric test is important, too, as the results in figure 2 suggest.*

The first example identifies pictures containing people wearing little or no clothing, to finesse the issue of variations of appearance of clothing. This program has been tested on an usually large and unusually diverse set of images; on a test collection of 565 images known to contain lightly clad people and 4289 control images with widely varying content, one tuning of the program marked 241 test images and 182 control images (the performance of various different tunings is indicated in figure 2; more detailed information appears in [7, 5]). The recall is comparable with full-text document recall [1, 2, 12] (which is surprisingly good for so abstract an object recognition query) and the rate of false positives is satisfactorily low. In this case, the representation was entirely built by hand.

The second example used a representation whose combinatorial structure — the order in which tests were applied — was built by hand, but where the tests were learned from data. This program identified pictures containing horses, and is described in greater detail in [6]. Tests used 100 images containing horses, and 1086 control images with widely varying content. The performance of various different configurations is shown in figure 2. For version "F", if one estimates

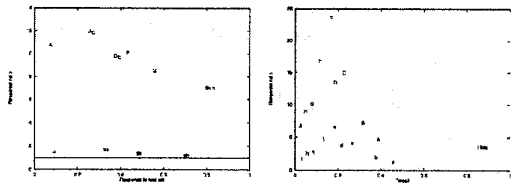

Figure 2: *The response ratio, (percent incoming test images marked/percent incoming control images marked), plotted against the percentage of test images marked, for various configurations of the two finding programs. Data for the nude human finder appears on the left, for the horse finder on the right. Capital letters indicate the performance of the complete system of skin/hide filter and geometrical grouper, and lower case letters indicate the performance of the geometrical grouper alone. The label "skin" (resp "hide") indicates the selectivity of using skin (resp hide) alone as a criterion. For the human finder, the parameter varied is the type of group required to declare a human is present — the trend is that more complex groups display higher selectivity and lower recall. For the horse finder, the parameter being varied is the maximum number of that will be considered.*

performance omitting images used in training and images for which the segment finding process fails, the recall is 15% — i.e. about 15% of the images containing horses are marked — and control images are marked at the rate of approximately 0.65%. In our test collection, this translates to 11 images of horses marked and 4 control images marked[1].

Finding using body plans has been shown to be quite effective for special cases in quite general scenes. It is relatively insensitive to changes in aspect [6]. It is quite robust to the relatively poor segmentations that our criteria offer, because it is quite effective in dealing with nuisance segments — in the horse tests, the average number of four segment groups was $2,500,000$, which is an average of forty segments per image. Nonetheless, the process described above is crude: it is too dependent on colour and texture criteria for early segmentation; the learning process is absent (humans) or extremely simple (horses); and there is one recogniser per class.

[1]These figures are *not* 15 and 7, because of the omission of training images and images where the segment finder failed in estimating performance.

2 Learning assembly processes

We have been studying processes for learning to assemble primitives. The recognition processes described above have a strong component of correspondence: in particular, we are pruning a set of correspondences between image segments and body segment labels by testing for kinematic plausibility. The search for acceptable correspondences can be made efficient by using *projected classifiers*, which prune labelings using the properties of smaller sub-labelings (as in [9], who use manually determined bounds and do not learn the tests). Given a classifier C which is a function of a set of features whose values depend on segments with labels in the set $L = \{l_1 \ldots l_m\}$, the projected classifier $C_{l_1 \ldots l_k}$ is a function of of all those features that depend only on the segments with labels $L' = \{l_1 \ldots l_k\}$. In particular, $C_{l_1 \ldots l_k}(L') > 0$ if there is some extension L of L' such that $C(L) > 0$.The converse need not be true: the feature values required to bring a projected point inside the positive volume of C may not be realized with any labeling of the current set of segments $1, \ldots, N$. For a projected classifier to be useful, it must be easy to compute the projection, and it must be effective in rejecting labelings at an early stage. These are strong requirements which are not satisfied by most good classifiers; for example, in our experience a support vector machine with a positive definite quadratic kernel projects easily but typically yields unrestrictive projected classifiers.

We have been using an axis-aligned bounding box, with bounds learned from a collection of positive labellings, for a good first separation, and then using a boosted version of a weak classifier that splits the feature space on a single feature value (as in [8]). This yields a classifier that projects particularly well, and allows clean and efficient algorithms for computing projected classifiers and expanding sets of labels (see [11]).

The segment finder may find either 1 or 2 segments for each limb, depending on whether it is bent or straight; because the pruning is so effective, we can allow segments to be broken into two equal halves lengthwise, both of which are tested.

2.1 Results

The training set included 79 images without people, selected randomly from the COREL database, and 274 images each with a single person on uniform background. The images with people have been scanned from books of human models [13]. All segments in the test images were reported; in the control images, only segments whose interior corresponded to human skin in colour and texture were reported. Control images, both for the training and for the test set, were chosen so that all had at least 30% of their pixels similar to human skin in colour and texture. This gives a more realistic test of the system performance by excluding regions that are obviously not human, and reduces the number of segments in the control images to the same order of magnitude as those in the test images.

The models are all wearing either swim suits or no clothes, otherwise segment finding fails; it is an open problem to segment people wearing loose clothing. There is a wide variation in the poses of the training examples, although all body segments are visible. The sets of segments corresponding to people were then hand-labeled. Of the 274 images with people, segments for each body part were found in 193 images. The remaining 81 resulted in incomplete configurations, which could still be used for computing the bounding box used to obtain a first separation. Since we assume that if a configuration looks like a person then its mirror image would too, we double the number of body configurations by flipping each one about a vertical axis. The bounding box is then computed from the resulting 548 points in the feature space, without looking at the images without people.

The boosted classifier was trained to separate two classes: the $193 \times 2 = 386$ points corresponding to body configurations, and 60727 points that did not correspond to people but lay in the bounding box, obtained by using the bounding box classifier to incrementally build labelings for the images with no people. We added 1178 synthetic positive configurations obtained by randomly selecting each limb and the torso from one of the 386 real images of body configurations (which were rotated and scaled so the torso positions were the same in all of them) to give an effect of joining limbs and torsos from different images rather like childrens' flip-books. Remarkably, the boosted classifier classified each of the real data points correctly but misclassified 976 out of the 1178 synthetic configurations as negative; the synthetic examples were unexpectedly more similar to the negative examples than the real examples were.

The test dataset was separate from the training set and included 120 images with a person on a uniform background, and varying numbers of control images, reported in table 1. We report results for two classifiers, one using 567 features and the other using a subset of 367 of those features. Table 1b shows the false positive and false negative rates achieved for each of the two classifiers. By marking 51% of test images

Features	# test images	# control images
367	120	28
567	120	86

a

Features	False Negatives	False Positives
367	37 %	4 %
567	49 %	10 %

b

Table 1: *(a) Number of images of people and without people processed by the classifiers with 367 and 567 features. (b) False negative (images with a person where no body configuration was found) and false positive (images with no people where a person was detected) rates.*

and only 10% of control images, the classifier using 567 features compares extremely favourably with that of [4], which marked 54% of test images and 38% of control images using hand-tuned tests to form groups of four segments. In 55 of the 59 images where there was a false negative, a segment corresponding to a body part was missed by the segment finder, meaning that the overall system performance significantly understates the classifier performance.

3 Finding clothing

Finding clothed people is a far more subtle problem than finding naked people, because the variation in colour, texture and pattern of clothing defeats a colour segmentation strategy. Clothing does have distinctive properties: the patterns formed by folds on clothing appear to offer cues to the configuration of the person underneath (as any textbook on figure drawing will illustrate). These folds have quite distinctive shading patterns [10], which are a dominant feature of the shading field of a person clad in a loose garment, because, although they are geometrically small, the surface normal changes significantly at a fold. Folds are best analysed using the theory of buckling, and arise from a variety of causes including excess material, as in the case of a full skirt, and stresses on a garment caused by body configurations. Folds appear to be the single most distinctive, reliable and general visual cue to the configuration of a person dressed in a cotton garment.

3.1 Grouping folds

We apply the simple fold finder described in [10] to the image at twelve different orientations. Using these twelve response maps, we use non-maximum supression to find the centre of the fold, and follow this maximum along the direction of maximum response to link all points corresponding to a single fold. The linking process breaks sharp corners, by considering the primary direction of the preceding points along the fold.

After finding all of the folds in the image, the next step is to find pairs which are approximately parallel, and in the same part of the image. If the projections of the two folds onto their average direction are disjoint, they are considered to belong to different parts of the image.

From the theory, we expect that multiple folds will be at regularly spaced intervals. Thus, we look for pairs which have one common fold, and consistent separations. (The separations should either be the same, or one should be double the other—if a single fold gets dropped, we do not want to ignore the entire pattern.) The separation between folds is required to be less than the maximum length of the folds. Finally, some of these groups can be further combined, if the groups have almost the same set of folds.

3.2 Results

The program typically extracts 10–25 groups of folds from an image. Figure 3 shows typical behaviour; groups clearly corresponds to the major folds across the torso in the image. This is in fact a segmentation of the image into coherent regions consisting of possible pieces of cloth. Other groups can appear, too — for example, venetian blinds fit our criteria well. We expect these extra segments can be easily dealt with by higher level processes, because they are not laid out in the same way as human body segments.

4 Summary

Representations based on spatial association between primitives can be used to perform recognition at acceptable levels for difficult tasks, and are widely applicable. The assembly process can be learned from data. Important problems remain: firstly, we need a theory that allows clean integration of cues from different sources; and secondly, we need a theory that predicts which primitives will be useful. We are currently

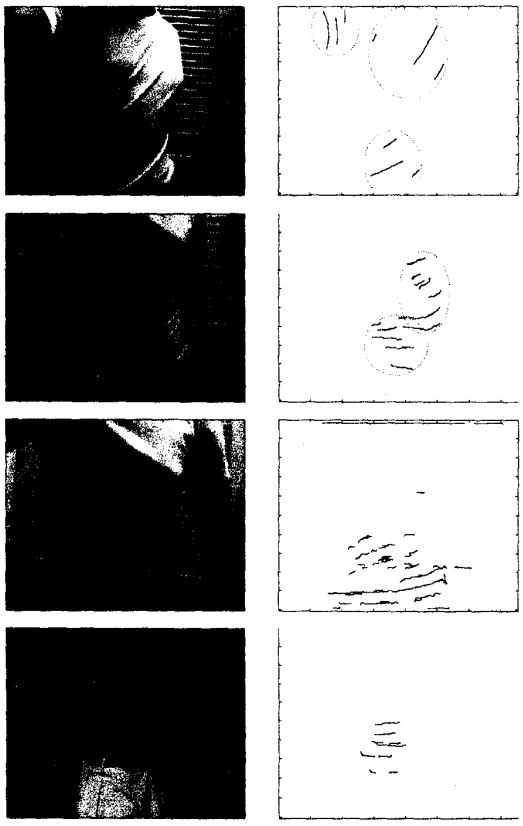

Figure 3: *Examples of segmentations produced by our grouping process. The figures show groups of fold responses, for the torsional (b,f) and axial (d,h) cases. In some cases, more than one group should be fused to get the final extent of the torso — these groups are separated by circles in the image. In each case, there are a series of between 10 and 25 other groups, representing either aliasing effects, the venetian blinds, or other accidental events. Each group could be a region of clothing; more high-level information is required to tell which is and which is not.*

investigating the use of large-scale Bayesian models to address these problems.

References

[1] D.C. Blair. Stairs redux: thoughts on the stairs evaluation, ten years after. *J. American Soc. for Information Science*, 47(1):4–22, 1996.

[2] D.C. Blair and M.E. Maron. An evaluation of retrieval effectiveness for a full text document retrieval system. *Comm. ACM*, 28(3):289–299, 1985.

[3] P.G.B. Enser. Query analysis in a visual information retrieval context. *J. Document and Text Management*, 1(1):25–52, 1993.

[4] M. M. Fleck, D. A. Forsyth, and C. Bregler. Finding naked people. In *European Conference on Computer Vision 1996, Vol. II*, pages 592–602, 1996.

[5] D. A. Forsyth and M. M. Fleck. Identifying nude pictures. In *IEEE Workshop on Applications of Computer Vision 1996*, pages 103–108, 1996.

[6] D.A. Forsyth and M.M. Fleck. Body plans. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.

[7] D.A. Forsyth, M.M. Fleck, and C. Bregler. Finding naked people. In *European Conference on Computer Vision*, 1996.

[8] Y. Freund and R.E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning - 13*, 1996.

[9] W.E.L. Grimson and T. Lozano-Pérez. Localizing overlapping parts by searching the interpretation tree. *IEEE Trans. Patt. Anal. Mach. Intell.*, 9(4):469–482, 1987.

[10] J. Haddon and D.A. Forsyth. Shading primitives. In *Int. Conf. on Computer Vision*, 1997.

[11] S. Ioffe and D.A. Forsyth. Learning to find pictures of people. In *In review — NIPS*, 1998.

[12] G. Salton. Another look at automatic text retrieval systems. *Comm. ACM*, 29(7):n649–657, 1986.

[13] unknown. *Pose file*, volume 1-7. Books Nippan, 1993-1996. A collection of photographs of human models, annotated in Japanese.