Finding Naked People

David A. Forsyth and Margaret Fleck

Abstract:

This paper demonstrates an automatic system for telling whether there are naked people present in an image. The approach combines color and texture properties to obtain a mask for skin regions, which is shown to be effective for a wide range of shades and colors of skin. These skin regions are then fed to a specialized grouper, which attempts to group a human figure using geometric constraints on human structure. This approach introduces a new view of object recognition, where an object model 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 system demonstrates excellent performance on a test set of 565 uncontrolled images of naked people, mostly obtained from the internet, and 4289 assorted control images, drawn from a wide collection of sources.

Keywords: Object Recognition, Computer Vision, Erotica/Pornography, Internet, Color, Content Based Retrieval.

1 Introduction

The recent explosion in internet usage and multi-media computing has created a substantial demand for algorithms that perform *content-based retrieval*—determining which images in a large collection depict some particular type of object. Such collections might be as structured as library databases, where images are stored with substantial amounts of secondary information including the results of earlier queries, or as chaotic as the entire internet. In either case, there is a need for systems that can estimate the content of an image, automatically and based purely on image information; in what follows, we use the term *image content assessment* for this process. This must be done quickly and automatically, because there is so much irrelevant material, but it need not be done with precise accuracy, because later processing (by the computer or by a user) will re-examine the selected data.

Identifying images depicting naked or scantily-dressed people is a natural problem for image content assessment. Typically, there are no textual or contextual cues to the content of these images. The existence of internet sites that sell images of naked people suggests that some people seek them, and would welcome a program that could find them automatically, perhaps by attaching an image content assessment program to a web robot (along the lines suggested by [51]). Equally, sales of programs such as "NetNanny" suggest that another significant group of people would prefer to prevent images of this form arriving on their computers. At present, each class seeks or avoids images based purely on the origin of the image, rather than on its content. As a result, seekers miss pictures, and avoiders miss

sites containing information they would like to have. Such incongruities occasionally receive media attention; in a recent incident, a commercial package for avoiders refused to allow access to the White House childrens' page[20].

2 Background

Determining which of a large set of pictures contain naked people can be seen as a problem in object recognition or in content based retrieval. Neither categorisation is wholly satisfactory: current object recognition systems do not deal with scenes as complex, categorisations as abstract, or datasets as large as those required by this problem; and current content based retrieval systems must exploit internal database structure or textual information to make abstract queries.

2.1 Objects and materials

Many notions of image content have been used to organize collections of images [25]. Relevant here are notions centered on objects; the distinction between materials and objects is particularly important. A material (e.g. skin) is defined by a homogeneous or repetitive pattern of fine-scale properties, but has no specific spatial extent or shape. An object (e.g. a ring) has a specific size and shape. This distinction and a similar distinction for actions, are well-known in linguistics and philosophy (dating back at least to [61]) where they are used to predict differences in the behavior of nouns and verbs (e.g. [13, 55, 56]).

To a first approximation, 3D materials appear as distinctive colors and textures in 2D images, whereas objects appear as regions with distinctive shapes. Therefore, one might attempt (following, for example, Adelson) to identify materials using low-level image properties, and identify objects by analyzing the shape of 2D regions. Indeed, materials with particularly distinctive color or texture (e.g. sky) can be successfully recognized with little or no shape analysis, and objects with particularly distinctive shapes (e.g. telephones) can be recognized using only shape information.

In general, however, too much information is lost in the projection onto the 2D image for such a strategy to be successful. The typical material, and thus the typical color and texture, of an object is often helpful in recognizing it. The shapes into which it is typically formed can be useful cues in recognizing a material. For example, a number of other materials have the same color and texture as human skin, at typical image resolution. Distinguishing these materials from skin requires using the fact that human skin typically occurs in human form.

2.2 Object Recognition

Current object recognition systems represent models either as a collection of geometric measurements—typically a CAD or CAD-like model—or as a collection of images of an object. This information is then compared with image information to obtain a match. Comparisons can be scored by using a feature correspondence either to backproject object features into an image or to determine a new view of the object and overlay that on the

image. Appropriate feature correspondences can be obtained by various forms of search (for example, [19, 17]). Alternatively, one can define equivalence classes of features, each large enough to have distinctive properties (invariants) preserved under the imaging transformation. These invariants can then be used as an index for a model library (examples of various combinations of geometry, imaging transformations, and indexing strategies include [14, 26, 48, 52, 54, 60, 30, 24]).

Each case described so far models object geometry exactly. Systems that recognize an object by matching a view to a collection of images of an object proceed in one of two ways. In the first approach, correspondence between image features and features on the model object is either given a priori or established by search. An estimate of the appearance in the image of that object is then constructed from the correspondences. The hypothesis that the object is present is then verified using the estimate of appearance [58]. An alternative approach computes a feature vector from a compressed version of the image and uses a minimum distance classifier to match this feature vector to feature vectors computed from images of objects in a range of positions under various lighting conditions [34].

All of the approaches described rely heavily on specific, detailed geometry, known (or easily determined) correspondences, and either the existence of a single object on a uniform, known background (in the case of [34]) or the prospect of relatively clear segmentation. None is competent to perform abstract classification; this emphasis appears to be related to the underlying notion of model, rather than to the relative difficulty of the classification vs. identification. Notable exceptions appear in [7, 5, 35, 64], which attempt to code relationships between various forms of volumetric primitive, where the description is in terms of the nature of the primitives involved and of their geometric relationship. None of these systems is well-suited for determining whether a naked person is present in typical test images, which depict a wide range of body part, body and multiple-body configurations.

2.3 Content based retrieval from image databases

Algorithms for retrieving information from image databases have concentrated on material-oriented queries, and have implemented these queries primarily using low-level image properties such as color and texture. Object-oriented queries search for images that contain particular objects; such queries can be seen either as constructs on material queries [42], as essentially textual matters[44], or as the proper domain of object recognition. A third query mode looks for images that are near iconic matches of a given image (for example, [21]); this is clearly not relevant to the task in hand, and such systems will not be reviewed here.

This application highlights the problems created by attempting to identify objects primarily on the basis of material properties. Although color and texture are useful aids in identifying an object, its shape must also be correct. As the results below show, it is insufficient to search for naked people by looking for skin alone; the skin needs to be in pieces of the right shape, which are attached to one another in the right ways.

The best-known image database system is QBIC [36], which allows an operator to specify various properties of a desired image. The system then displays a selection of potential matches to those criteria, sorted by a score of the appropriateness of the match. The operator

can adjust the scoring function. Region segmentation is largely manual, but the most recent versions of QBIC [3] contain simple automated segmentation facilities. The representations constructed are a hierarchy of oriented rectangles of fixed internal color and a set of tiles on a fixed grid, which are described by internal color and texture properties. However, neither representation allows reasoning about the shape of individual regions, about the relative positioning of regions of given colors or about the cogency of geometric coocurrence information, and so there is little reason to believe that either representation can support object queries.

Photobook [39] largely shares QBIC's model of an image as a collage of flat, homogenous frontally presented regions, but incorporates more sophisticated representations of texture and a degree of automatic segmentation. A version of Photobook ([39], p. 10) incorporates a simple notion of object queries, using plane object matching by an energy minimisation strategy. However, the approach does not adequately address the range of variation in object shape and appears to require images that depict single objects on a uniform background. Further examples of systems that identify materials using low-level image properties include Virage [59], Candid [10, 23] and Chabot [37]. None of these systems code spatial organisation in a way that supports object queries.

Variations on photobook [42, 33] use a form of supervised learning known in the information retrieval community as "relevance feedback" to adjust segmentation and classification parameters for various forms of textured region. When a user is available to tune queries, supervised learning algorithms can clearly improve performance given appropriate object and image representations. In the most useful applications of our algorithms, however, users are unlikely to want to tune queries. Those who object to pictures of naked people are unlikely to want to spend time looking at such pictures to help tune a learning algorithm, though one might speculate that seekers could sell tuning services to avoiders.

More significantly, the representations used in these supervised learning algorithms do not code spatial relationships. Thus, these algorithms are is unlikely to be able to construct a broad range of effective object queries. While relevance feedback can be effective at adjusting a metric by which image relevance is scored, it is hard to see that supervised learning would be the technique of choice for establishing such intricate constructs as the variations in appearance associated with different views of a body plan. It is extremely hard to see what appropriate representations might be and how the learning process might be controlled.

2.4 Finding people

A variety of systems have been developed specifically for recognizing people or human faces. There are several domain specific constraints in recognizing human bodies: humans are made out of parts whose shape is relatively simple; there are few ways to assemble these parts; and, when one can measure motion, the dynamics of these parts are limited, too.

Most previous work emphasizes motion, though [27] shows that structural constraints on humans yield pose, if a stick-figure group is available. The constraints on human dynamics can be exploited to locate moving people in images [43] or to track them [46, 18] The resulting figures can then be labelled to various degrees of granularity, leading to inferences

about what the person being tracked is doing. Typically, such systems are engineered to simplify segmentation, by, for example, constraining the contrast or the appearance of the background. Additional robustness is achieved by maintaining a pixel-based statistical color model of the person being tracked and of the background. Finally, an assumption that the figure is viewed in a frontal pose makes it possible to label head and hands, and thereby interpret simple gestures like pointing or waving using space-time templates matched with a warping technique. It is usual to deal with images where segmentation is essentially trivial, or where the background is known in advance (as in PFind [62]). Typical systems group regions into stick figures by combining reasoning about gravity[28] or knowledge about background material [2] with motion cues, to form and fuse image ribbons. A good survey of related approaches that deal primarily with motion features and recognition of gestures, gaits, and other high level categories appears in [11].

Face and hand finding/tracking is a related problem where textural, color, and structural constraints are exploited. The main features on a human face appear in much the same form in most images, enabling techniques based on principal component analysis or neural networks proposed by, for example, [41, 53, 49, 8]. Face finding based on affine covariant geometric constraints is presented by [29]. Hands have been segmented using color [1] and tracked using on a kinematic model [45]. Color information is used in [50] to segment skin regions for face identification.

2.5 Summary

Coding the appearance of individual regions is not a satisfactory notion of content. To identify 3D objects, or even materials, requires representing shape properties of regions, and the relative spatial disposition of regions. None of the systems described contains a coding of object shape, able to compensate for variation between different objects of the same type (e.g. several dogs), changes in posture (how any flexible parts or joints are arranged), and variation in camera viewpoint, and so none can perform object queries of the type described. Automatic segmentation at a satisfactory level remains an extremely difficult problem for object recognition or image database systems. Finally, work on finding people typically concentrates either on motion cues or on specific body parts like faces and hands; there is little work on segmentation.

None of these systems is suitable for analyzing typical images of naked people found on the internet, which

- have uncontrolled backgrounds, from which the figure elements must be extracted,
- may depict multiple figures,
- often contain partial figures,
- are still images (no motion information), and
- have been taken from a wide variety of camera angles.

Clearly, solving this problem will be require effective segmentation in quite complex images. For a variety of reasons, segmentation has to date been regarded as independent of recognition. Because the present application requires segmentation in very general images, our approach attempts to marshal as much model information as possible at each segmentation stage, to control segmentation problems. We detect naked people by:

1. determining which images contain large areas of skin-colored pixels;
2. within skin colored regions, finding regions that are similar to the projection of cylinders;
3. grouping skin coloured cylinders into possible human limbs and connected groups of limbs.

Images containing sufficiently large skin-colored groups of possible limbs are then reported as containing naked people.

3 Finding Skin

Images of naked people may vary in content but, by definition, they contain significant regions of skin, typically a large fraction of the image area. Skin can be identified as regions that have no texture and satisfy a collection of color constraints, as the appearance of skin is tightly constrained.

The color of a human's skin is created by a combination of blood (red) and melanin (yellow, brown) [47]. Therefore, human skin has a restricted range of hues and is somewhat saturated, but not deeply saturated. Because more deeply colored skin is created by adding melanin, one would expect the saturation to increase as the skin becomes more yellow, and this is reflected in our data set. Finally, skin has little texture; extremely hairy subjects are rare. Ignoring regions with high-amplitude variation in intensity values allows the skin filter to eliminate more control images.

Detection of skin is complicated by the fact that skin's reflectance has a substantial non-Lambertian component. It often (perhaps typically) has bright areas or highlights which are desaturated. Furthermore, the illumination color varies slightly from image to image, so that some skin regions appear as blueish or greenish off-white. We have not encountered internet images which show skin with strong skews in hue derived from illumination. We believe that information providers manually enhance their images to avoid these effects, which are notably unaesthetic.

3.1 Color and texture processing

The skin filter starts by subtracting the zero-response of the camera system, estimated as the smallest value in any of the three color planes omitting locations within 10 pixels of the image edges, to avoid potentially significant desaturation. The input R, G, and B values are then transformed into log-opponent values I, R_g, and B_y (cf. e.g. [15]) as follows:

$$L(x) = 105 \log_{10}(x + 1 + n)$$

$$I = L(G)$$

$$R_g = L(R) - L(G)$$

$$B_y = L(B) - \frac{L(G) + L(R)}{2}$$

The green channel is used to represent intensity because the red and blue channels from some cameras have poor spatial resolution. In the log transformation, 105 is a convenient scaling constant and n is a random noise value, generated from a distribution uniform over the range $[0, 1)$. The random noise is added to prevent banding artifacts in dark areas of the image. The log transformation makes the R_g and B_y values intensity independent.

Next, smoothed texture and color planes are extracted. The R_g and B_y arrays are smoothed with a median filter. To compute texture amplitude, the intensity image is smoothed with a median filter, and the result subtracted from the original image. The absolute values of these differences are run through a second median filter. These operations use a fast multi-ring approximation to the median filter [12].

The texture amplitude and the smoothed R_g and B_y values are then passed to a tightly-tuned skin filter. It marks as probably skin all pixels whose texture amplitude is small, and whose hue and saturation values are appropriate. (Hue and saturation are simply the direction and magnitude of the vector (R_g, B_y).) The range of hues considered to be appropriate changes with the saturation, as described above. This is very important for good performance. When the same range of hues is used for all saturations, significantly more non-skin regions are accepted.

Because skin reflectance has a substantial specular component, some skin areas are de-saturated or even white. Under some illuminants, these areas appear as blueish or greenish off-white. These areas will not pass the tightly-tuned skin filter, creating holes (sometimes large) in skin regions, which may confuse geometrical analysis. Therefore, the output of the initial skin filter is expanded to include adjacent regions with nearly appropriate properties.

Specifically, the region marked as skin is enlarged to include pixels many of whose neighbors passed the initial filter (by adapting the multi-ring median filter). If the resulting marked regions cover at least 30% of the image area, the image will be referred for geometric processing. Finally, the algorithm unmarks any pixels which do not satisfy a less tightly tuned version of the hue and saturation constraints.

4 Grouping People

The human figure 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 constraints on the 3D parts induce grouping constraints on the corresponding 2D image regions. These induced constraints provide an appropriate and effective model for recognizing human figures.

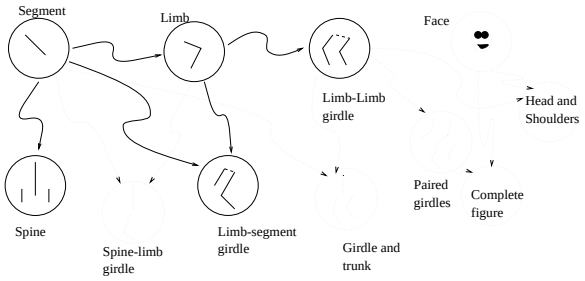

Figure 1: The grouping rules (arrows) specify how to assemble simple groups (e.g. body segments) into complex groups (e.g. limb-segment girdles). These rules incorporate constraints on the relative positions of 2D features, induced by constraints on 3D body parts. Dashed lines indicate grouping rules that are not yet implemented, but suggest the overall structure of the kind of model we are advocating. A complete model would contain information about a variety of body parts; occlusion and aspect information is implicit in the structure of the paths through the grouping process.

The current system models a human as a set of rules describing how to assemble possible girdles and spine-thigh groups (Figure 1). The input to the geometric grouping algorithm is a set of images, in which the skin filter has marked areas identified as human skin. Sheffield's version of Canny's [9] edge detector, with relatively high smoothing and contrast thresholds, is applied to these skin areas to obtain a set of connected edge curves. Pairs of edge points with a near-parallel local symmetry [6] are found by a straightforward algorithm. Sets of points forming regions with roughly straight axes ("ribbons" [7]) are found using an algorithm based on the Hough transformation. The number of irrelevant symmetries recorded is notably reduced by an assumption that humans in test images will appear at a relatively small range of scales; this assumption works fairly well in practice[1] but appears to limit performance.

Grouping proceeds by first identifying potential segment outlines, where a segment outline is a ribbon with a straight axis and relatively small variation in average width. Ribbons are checked to ensure that (a) their interior contains mostly skin-coloured pixels, and (b) that intensity cross-sections taken perpendicular to the axis of the ribbon are similar from step to step along the axis. While this approach is successful at supressing many false ribbons, the local support of the intensity test means that ribbons that contain texture at a fairly coarse scale (with respect to the size of the ribbon) are not rejected; as figure 7 indicates, this is a significant source of false positives. Ribbons that may form parts of the same segment are merged, and suitable pairs of segments are joined to form limbs. An affine imaging model is satisfactory here, so the upper bound on the aspect ratio of 3D limb segments induces an upper bound on the aspect ratio of 2D image segments corresponding to limbs. Similarly, we can derive constraints on the relative widths of the 2D segments.

Specifically, two ribbons can only form part of the same segment if they have similar widths and axes. Two segments may form a limb if their search intervals intersect; there is skin in the interior of both ribbons; their average widths are similar; and in joining their axes,

[1] The iconography of pornography is such that subjects typically occupy most of the image.

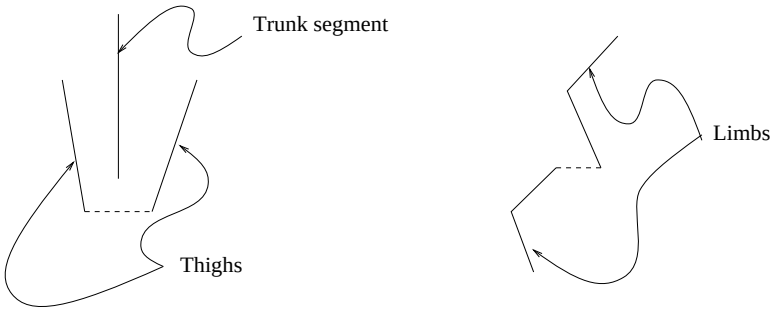

Figure 2: A configuration that is prohibited by geometric constraints on assembling groups. The dashed line represent the girdle. Neither arms nor legs can be configured to look like this. There are configurations that are possible for the one girdle, but not the other.

not too many edges must be crossed. There is no angular constraint on axes in grouping limbs. The output of this stage contains many groups that do not form parts of human-like shapes: they are unlikely to survive as grouping proceeds to higher levels.

The limbs and segments are then assembled into putative girdles. There are grouping procedures for two classes of girdle, one formed by two limbs, and one formed by one limb and a segment. The latter case is important when one limb segment is hidden by occlusion or by cropping. The constraints associated with these girdles are derived from the case of the hip girdle, and use the same form of interval-based reasoning as used for assembling limbs.

Limb-limb girdles must pass three tests. The two limbs must have similar widths. It must be possible to join two of their ends with a line segment (the pelvis) whose position is bounded at one end by the upper bound on aspect ratio, and at the other by the symmetries forming the limb and whose length is similar to twice the average width of the limbs. Finally, occlusion constraints rule out certain types of configurations: limbs in a girdle may not cross each other, they may not cross other segments or limbs, and there is a forbidden configuration of limbs (see figure 2). A limb-segment girdle is formed using similar constraints, but using a limb and a segment.

Spine-thigh groups are formed from two segments serving as upper thighs, and a third, which serves as a trunk. The thigh segments must have similar average widths, and it must be possible to construct a line segment between their ends to represent a pelvis in the manner described above. The trunk segment must have an average width similar to twice the average widths of the thigh segments. The grouper asserts that human figures are present if it can assemble either a spine-thigh group or a girdle group.

The ratio of average widths of two cylinders, measured in an image, is a meaningful measurement under an affine imaging model unless the two cylinders do not occur at vastly different depths in space. We regard cases of the latter type as accidental alignments: they are increasingly unlikely to survive as grouping proceeds to higher levels. As a result, the relative widths of two ribbons can be compared meaningfully, to allow joining pairs of ribbons together into limbs. Two ribbons may form a limb if:

- their search intervals intersect;

- there is skin in the interior of both ribbons;
- their average widths are similar; and
- in joining their axes, not too many edges must be crossed.

There is no angular constraint on axes in grouping limbs.

The present grouper then proceeds to assemble limbs and segments into putative girdles. It has grouping procedures for two classes of girdle; one formed by two limbs, and one formed by one limb, and a segment. The latter case is important when one limb segment is hidden by occlusion or by cropping. The constraints associated with these girdles are derived from the case of the hip girdle, and use the same form of interval-based reasoning as used for assembling limbs. It is not possible to reliably determine which of two segments forming a limb is the thigh: the only cue is a small difference in average width and this is unreliable when either segment may be cropped or foreshortened. Therefore, both orientations of each limb are considered.

5 Experimental protocol

The performance of the system was tested using 565 target images of naked people and 4302 assorted control images, containing some images of people but none of naked people. Most images encode a (nominal) 8 bits/pixel in each color channel. The target images were collected from the internet and by scanning or re-photographing images from books and magazines. They show a very wide range of postures and activities. Some depict only small parts of the bodies of one or more people. Most of the people in the images are Caucasians; a small number are Blacks or Asians. Images were sampled from internet newsgroups[2] by collecting about 100-150 images per sample on several occasions. The origin of the test images was not recorded[3]. There was no pre-sorting for content; however, only images encoded using the JPEG compression system were sampled as the GIF system, which is also often used for such images, has poor color reproduction qualities. Test images were automatically reduced to fit into a 128 by 192 window, and rotated as necessary to achieve the minimum reduction.

It is hard to assess the performance of a system for which the control group is properly all possible images. In particular, obvious strategies to demonstrate weaknesses in performance may fail. For example, images of clothed people, which would confuse the grouper, fail to pass the skin filter and so would form a poor control set. Furthermore, a choice of controls that deliberately improves or reduces performance complicates assessing performance. The only appropriate strategy to reduce internal correlations in the control set appears to be to use large numbers of control images, drawn from a wide variety of sources. To improve the assessment, we used seven types of control images (figure 3):

[2]Specifically, alt.binaries.pictures.erotica, alt.binaries.pictures.erotica.male, alt.binaries.pictures.erotica.redheads and alt.binaries.pictures.erotica.female.

[3]In retrospect, this is an error in experimental design. It appears to be the case that the material posted by each individual typically has significant correlations as to content; a record of who posted which image would have improved our understanding of the statistics of the test set.

- 1241 images sampled[4] from an image database originating with the California Department of Water Resources (DWR), showing environmental material around California, including landscapes, pictures of animals, and pictures of industrial sites,
- 58 images of clothed people, a mixture of Caucasians, Blacks, Asians, and Indians, largely showing their faces, 3 re-photographed from a book and the rest photographed from live models at the University of Iowa,
- 44 assorted images from a photo CD that came with a copy of a magazine [32],
- 11 assorted personal photos, re-photographed with our CCD camera, and
- 47 pictures of objects and textures taken in our laboratory for other purposes.
- 1241 pictures, consisting of the complete contents of CD-ROM's 10000 (Air shows), 113000 (Arabian horses), 123000 (Backyard wildlife), 130000 (African speciality animals), 132000 (Annuals for American gardens), 172000 (Action sailing), 173000 (Alaskan wildlife), 34000 (Aviation photography), 38000 (American national parks), 44000 (Alaska), 49000 (Apes) and 77000 (African antelope), and 41 images from CD-ROM 135000 (Bald eagles) in the Corel stock photo library.
- 1660 pictures, consisting of the every fifth image in CD-ROM's 186000 (Creative crystals), 188000 (Classic Antarctica), 190000 (Interior desighn), 191000 (Clouds), 195000 (Hunting), 198000 (Beautiful women), 202000 (Beautiful Bali), 207000 (Alps in spring), 208000 (Fungi), 209000 (Fish), 212000 (Chicago), 214000 (Gardens of Europe), 218000 (Caverns), 219000 (Coast of Norway), 220000 (Cowboys), 221000 (Flowers close up), 225000 (Freestyle skiing), 226000 (Amateur sports), 227000 (Greek scenery), 228000 (Authumn in Maine), 230000 (Canada), 234000 (Decorated pumpkins), 237000 (Construction), 238000 (Canoeing adventure), 240000 (Arthropods), 243000 (Acadian Nova Scotia), 246000 (Bhutan), 250000 (Industry and transportation), 251000 (Canadian farming), 255000 (Colorado plateau), 261000 (Historic Virginia), 263000 (Antique postcards), 267000 (Hiking), 268000 (African birds), 275000 (Beverages), 276000 (Canadian rockies), 279000 (Exotic Hong Kong), 281000 (Exploring France), 282000 (Fitness), 285000 (Fire fighting), 291000 (Devon, England), 292000 (Berlin), 294000 (Barbecue and salads), 297000 (Desserts), 298000 (English countryside), 299000 (Images of Egypt), 302000 (Fashion), 304000 (Asian wildlife), 308000 (Holiday sheet music), 310000 (Dog sledding), 311000 (Everglades), 314000 (Dolphins and whales), 318000 (Foliage backgrounds), 322000 (Fruits and nuts), 325000 (Car racing), 327000 (Artist textures), 329000 (Hot air balloons), 332000 (Fabulous fruit), 333000 (Cuisine), 336000 (Cats and kittens), 340000 (English pub signs), 341000 (Colors of autumn), 344000 (Canadian national parks), 346000 (Garden ornaments and architecture), 350000 (Frost textures), 353000 (Bonsai and Penjing), 354000 (British motor collection), 359000 (Aviation photography II), 360000 (Classic aviation), 363000 (Highway and street signs),

[4]The sample consists of every tenth image; in the full database, images with similar numbers tend to have similar content.

Figure 3: Typical control images. The images in the first row are incorrectly classified as containing naked people; those in the second row pass the skin filter, but are rejected by the geometric grouping process; and those in the third row are rejected by the skin filter.

367000 (Creative textures), 369000 (Belgium and Luxembourg), 371000 (Canada, an aerial view), 372000 (Copenhagen, Denmark), 373000 (Everyday objects), 378000 (Horses in action), 382000 (Castles), 384000 (Beaches), 394000 (Botanical prints), 396000 (Air force), 399000 (Bark textures) and 412000 (Bobsledding) in the second edition of the Corel stock photo library[5]

The DWR images and Corel images were available at a resolution of 128 by 192 pixels. The images from other sources were automatically reduced to approximately the same size.

On thirteen of these images, our code failed due to implementation bugs. Because these images represent only a tiny percentage of the total test set, we have simply excluded them from the following analysis. This reduced the size of the final control set to 4289 images.

[5]Both libraries are available from the Corel Corporation, whose head office is at 1600 Carling Ave, Ottawa, Ontario, K1Z 8R7, Canada.

6 Experimental results

Our algorithm can be configured in a variety of ways, depending on the complexity of the assemblies constructed by the grouper. For example, the process could report a naked person present if a skin-colored segment was obtained, or if a skin-colored limb was obtained, or if a skin-colored spine or girdle was assembled. Each of these alternatives will produce different performance results. Before running our tests, we chose as our *primary configuration*, a version of the grouper which requires that a girdle or spine group be present for a naked person to be reported. All example images shown in figures were chosen using this criterion. For comparison, we have also included summary statistics for several other configurations of the grouper.

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.

As the configuration of the algorithm is changed, the recall and response ratio both change. It is not possible to select one configuration as optimal, because different users may require different trade-offs between false positives and false negatives. Therefore, we will simply graph recall against response ratio for the different configurations.

6.1 The skin filter

Of the 565 test and 4289 control images processed, the skin filter marked 448 test images and 485 control images as containing people. As table 1 shows, this yields a response ratio of 7.0 and a test response of 79%. This is surprisingly strong performance for a process that, in effect, reports the number of pixels satisfying a selection of absolute color constraints. It implies that in most test images, there are a large number of skin pixels; however, it also shows that *simply marking skin-colored regions is not particularly selective*.

Mistakes by the skin filter occur for several reasons. In some test images, the naked people are very small. In others, most or all of the skin area is desaturated, so that it fails the first-stage skin filter. It is not possible to decrease the minimum saturation for the first-stage filter, because this causes many more responses on the control images. Some control images pass the skin filter because they contain (clothed) people, particularly several close-up portrait shots. Other control images contain material whose color closely resembles that

of human skin. Typical examples include wood, desert sand, certain types of rock, certain foods, and the skin or fur of certain animals.

All but 8 of our 58 control images of faces and clothed people failed the skin filter primarily because many of the faces occupy only a small percentage of the image area. In 18 of these images, the face was accurately marked as skin. In 12 more, a recognizable portion of the face was marked. Failure on the remaining images is largely due to the small size of the faces, desaturation of skin color, and fragmentation of the face when eye and mouth areas are rejected by the skin filter. A combination of the skin filter with filters for eye-like and mouth-like features might be able to detect faces reliably. These face images contain a wider range of skin tones than our images of naked people: the skin filter appears to perform equally well on all races.

6.2 The geometric filter

The geometrical filter ran on the output of the skin filter: 448 test images and 485 control images. The primary grouper marks 241 test images and 182 control images, meaning that the entire system composed of primary grouper operating on skin filter output displays a response ratio of 10.0 and a test response of 43%. Considered on its own, the grouper's response ratio is 1.4, and the selectivity of the system is clearly increased by the grouper. Table 1 shows the different response ratios displayed by various configurations of the grouper. Both girdle groupers and the spine grouper often mark structures which are parts of the human body, but not hip or shoulder girdles. This presents no major problem, as the program is trying to detect the presence of humans, rather than analyze their pose in detail.

False negatives occur for several reasons. Some close-up or poorly cropped images do not contain arms and legs, vital to the current geometrical analysis algorithm. Regions may have been poorly extracted by the skin filter, due to desaturation. The edge finder may fail due to poor contrast between limbs and their surroundings. Structural complexity in the image, often caused by strongly colored items of clothing, confuses the grouper. Finally, since the grouper uses only segments that come from bottom up mechanisms and does not predict the presence of segments which might have been missed by occlusion, performance is notably poor for side views of figures with arms hanging down.

Some of the control images were typically classified by the skin filter as containing significant regions of possible skin, actually contain people; others contain materials of similar color, such as animal skin, wood, or off-white painted surfaces. The geometric grouper wrongly marks spines or girdles in some control images, because it has only a very loose model of the shape of these body parts. The current implementation is frequently confused by groups of parallel edges, as in industrial scenes, and sometimes accepts ribbons lying largely outside the skin regions. We believe the latter problem can easily be corrected.

In the Corel CD-ROM database, images are grouped into sets of images with similar content. False positives tend to be clustered in particular sets. Table 2 lists the sets on which the system showed the strongest response. These images depict objects with skin-like colors and elongated (limb-like) structures. We believe that these examples could be eliminated by a more sophisticated grouper.

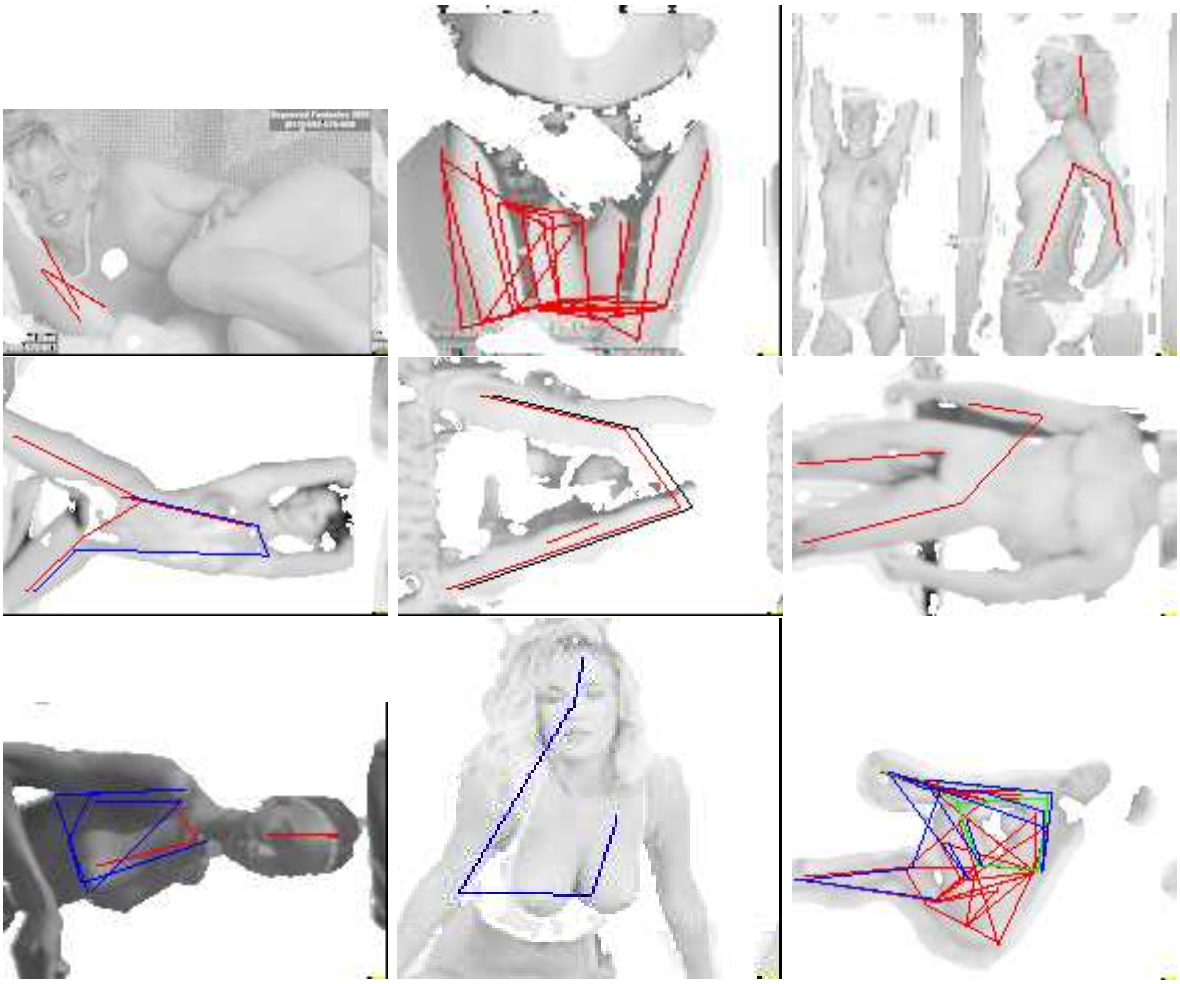

Figure 4: Typical images correctly classified as containing naked people. The output of the skin filter is shown, with spines overlaid in red, limb-limb girdles overlaid in blue, and limb-segment girdles overlaid in blue. Notice that there are cases in which groups form quite good stick figures; where the groups are wholly unrelated to the limbs; where accidental alignment between figures and background cause many highly inaccurate groups; and where other body parts substitute for limbs. Assessed as a producer of stick figures, the grouper is relatively poor, but as the results below show, it makes a real contribution to determining whether people are present.

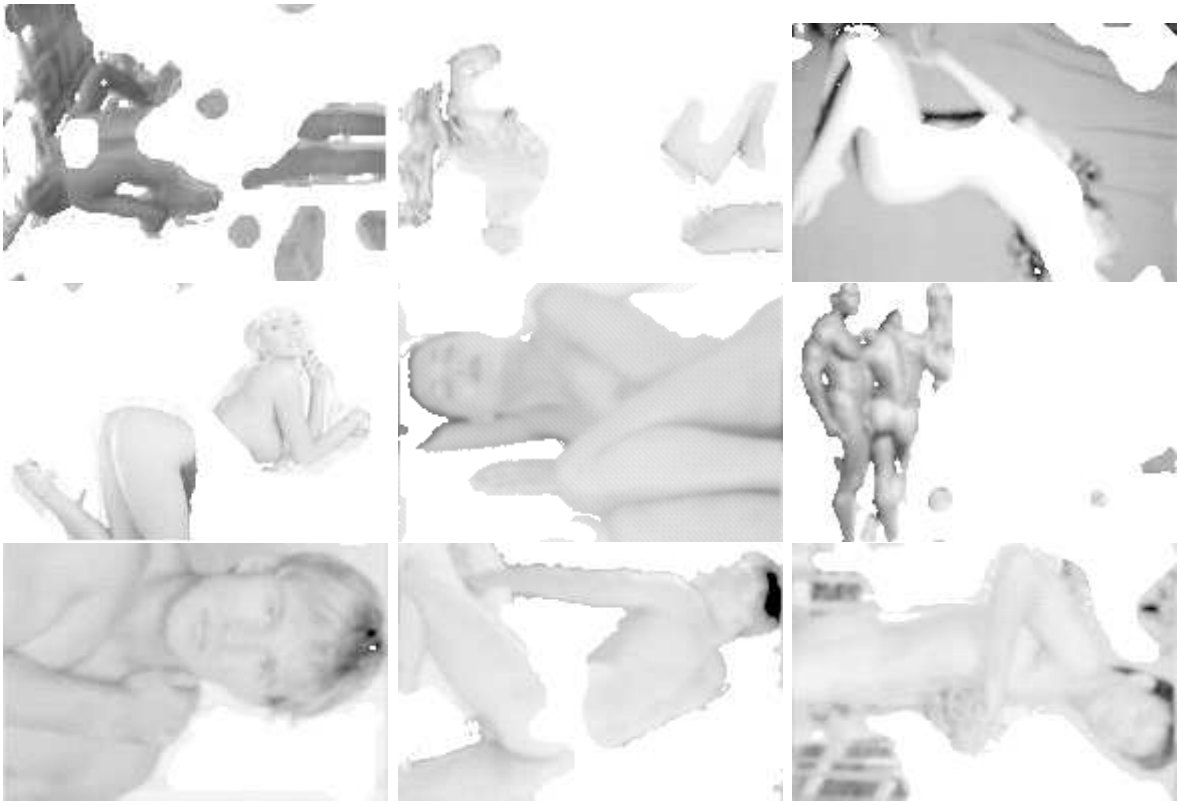

Figure 5: Typical false negatives: the skin filter marked significant areas of skin, but the geometrical analysis could not find a girdle or a spine. Failure is often caused by absence of limbs, low contrast, or configurations not included in the geometrical model (notably side views, head and shoulders views, and closeups).

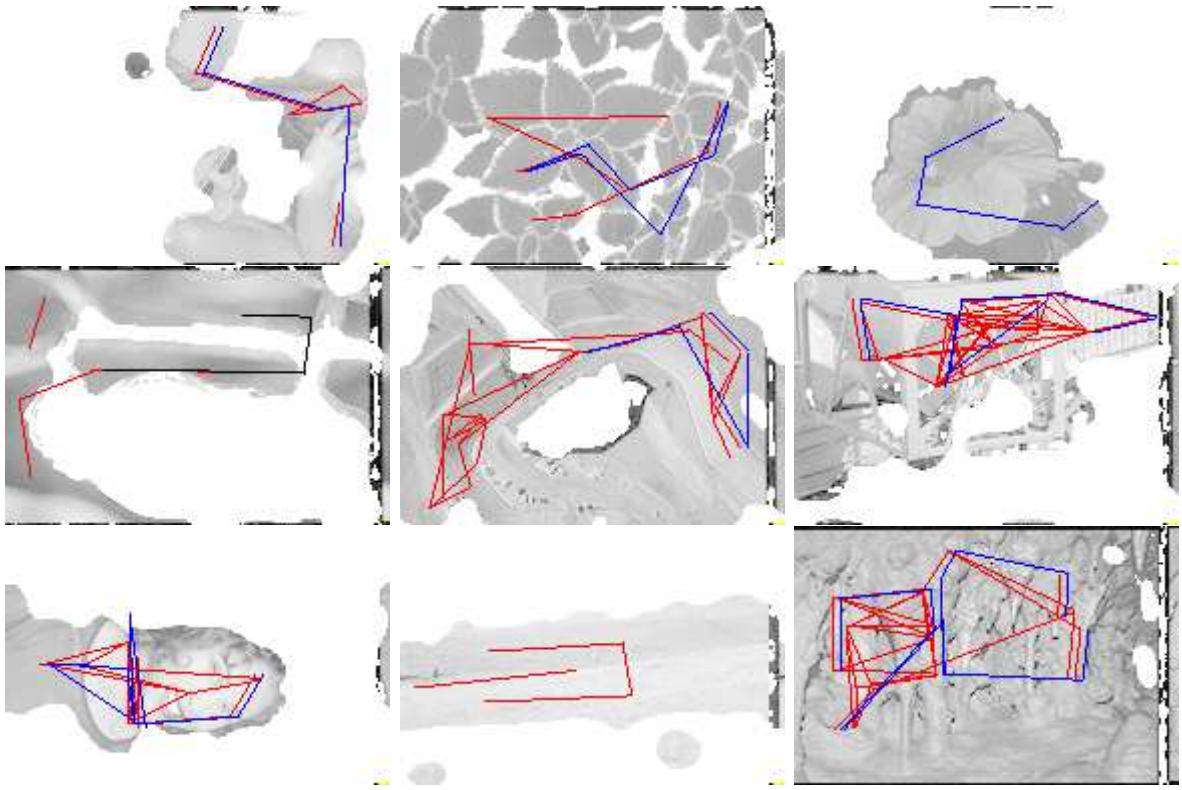

Figure 6: Typical control images wrongly classified as containing naked people. These images contain people or skin-colored material (animal skin, wood, bread, off-white walls) and structures which the geometric grouper mistakes for spines (red) or girdles. Limb-limb girdles are shown in blue, limb-segment girdles in blue. The grouper is frequently confused by groups of parallel edges, as in the industrial images. Note that regions marked as skin can contain texture at a larger scale than that measured by the texture filter. An ideal system would require that limbs not have texture at the scale of the limb, and would be able to automatically determine an appropriate scale at which to search for limbs.

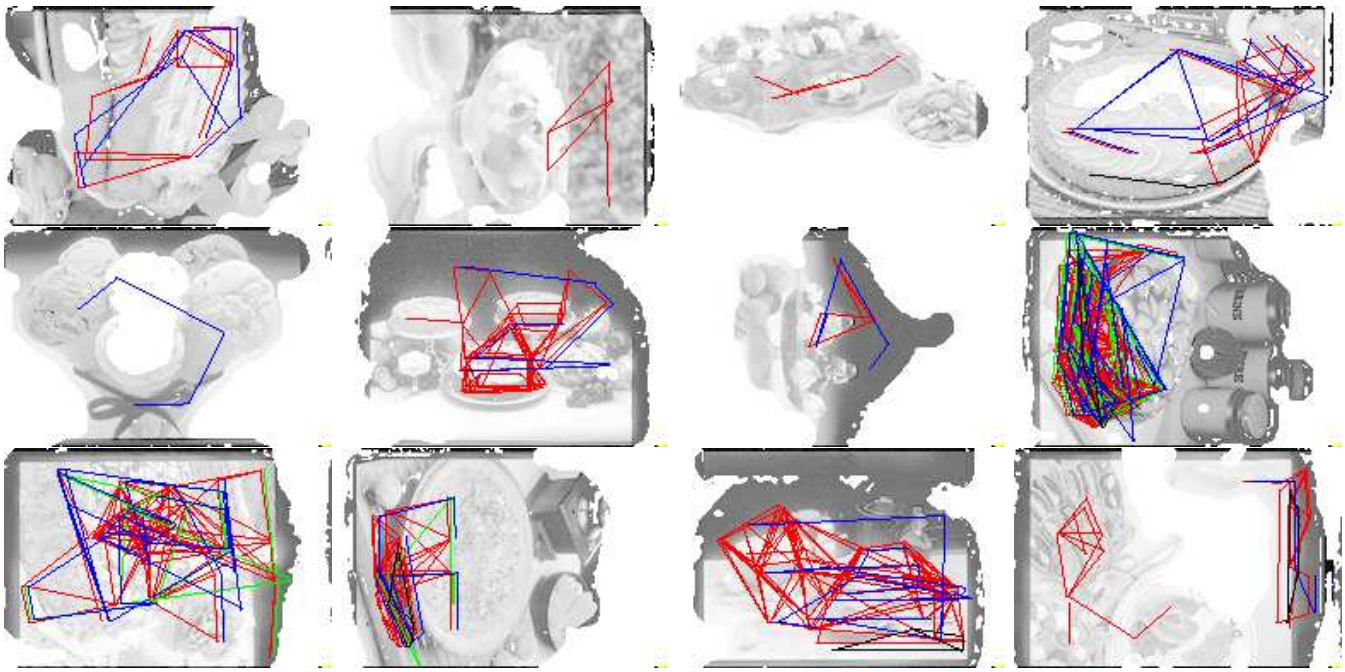

Figure 7: All the control images in the dessert sequence that were marked by the system, with groups overlaid. These images contain large regions of skin-coloured material, with texture at a scale invisible to the skin filter. Since there are many edges in the skin filter output, a large collection of symmetries appears and limb or girdle groups are virtually guaranteed. As many ribbons contain coarse scale texture features, these false positives suggest that a local verification mechanism that looked more carefully at the intensities in a ribbon at an appropriate scale, would improve the performance of the system.

system configuration	response ratio	test response	control response	test images marked	control images marked	recall	precision
skin filter	7.0	79.3%	11.3%	448	485	79%	48%
A	10.7	6.7%	0.6%	38	27	7%	58%
B	12.0	26.2%	2.2%	148	94	26%	61%
C	11.8	26.4%	2.2%	149	96	26%	61%
D	9.7	38.6%	4.0%	218	170	39%	56%
E	9.7	38.6%	4.0%	218	171	39%	56%
F (primary)	10.1	42.7%	4.2%	241	182	43%	57%
G	8.5	54.9%	6.5%	310	278	55%	53%
H	8.4	55.9%	6.7%	316	286	56%	52%

Table 1: Overall classification performance of the system, in various configurations, to 4289 control images and 565 test images. Configuration F is the primary configuration of the grouper, fixed before the experiment was run, which reports a naked person present if either a girdle, a limb-segment girdle or a spine group is present, but not if a limb group is present. Other configurations represent various permutations of these reporting conditions; for example, configuration A reports a person present only if girdles are present. There are fewer than 15 cases, because some cases give exactly the same response.

Figure 8 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.

7 Discussion and Conclusions

This paper has shown that images of naked people can be detected using a combination of simple visual cues—color, texture, and elongated shapes—and class-specific grouping rules. The algorithm successfully extracts 43% of the test images, but only 4% of the control images. This system is not as accurate as some recent object recognition algorithms. However, this system is performing a much more abstract task ("find naked people" rather than "find an object matching this CAD model"). It is detecting jointed objects of highly variable shape, in a diverse range of poses, seen from many different camera positions. Both lighting and background are uncontrolled, making segmentation very difficult. Furthermore, the test database is substantially larger and more diverse than those used in previous object recognition experiments. Finally, the system is relatively fast for a query of this complexity;

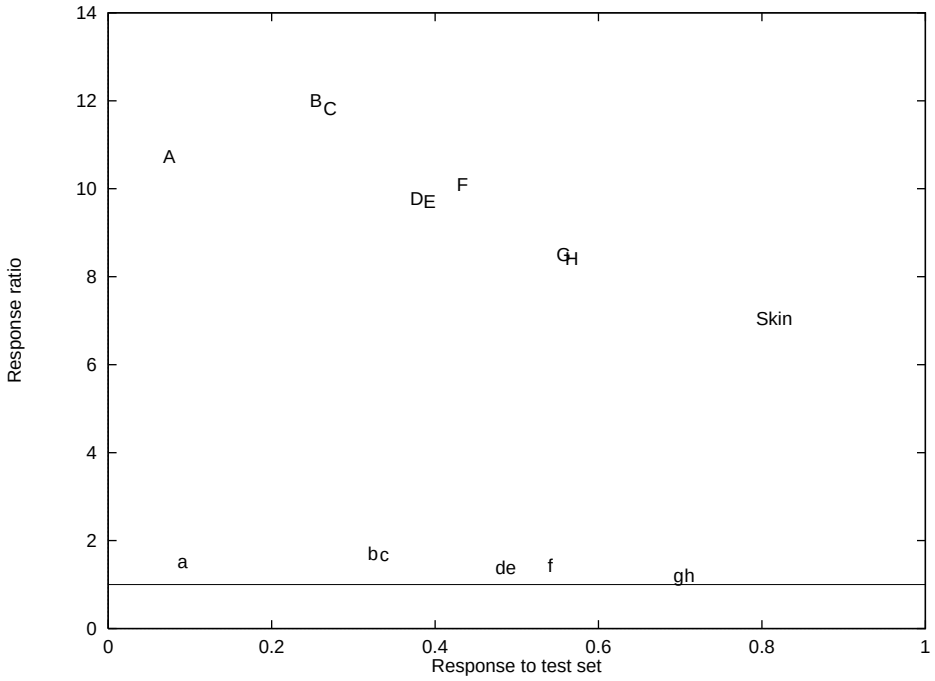

Figure 8: The response ratio, (percent incoming test images marked/percent incoming control images marked), plotted against the percentage of test images marked, for various configurations of the naked people finder. Labels "A" through "H" indicate the performance of the entire system of skin filter and geometrical grouper together, where "F" is the primary configuration of the grouper. 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, where "f" is again the primary configuration 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. While the grouper's selectivity is less than that of the skin filter, it improves the selectivity of the system considerably. **Key:** A: limb-limb girdles; B: limb-segment girdles; C: limb-limb girdles or limb-segment girdles; D: spines; E: limb-limb girdles or spines; F: (two cases) limb-segment girdles or spines and limb-limb girdles, limb-segment girdles or spines; G, H each represent four cases, where a human is declared present if a limb group or some other group is found.

System response	Title (s)
60%	Desserts
40%	Caverns, Colorado plateau Cuisine
35%	Barbecue and salads
25%	Fabulous fruit, Colors of autumn
20%	Decorated pumpkins, Fashion Copenhagen–Denmark
15%	Beautiful women, Fungi, Cowboys, Flowers close up Acadian Nova Scotia, Antique postcards Fire fighting, Images of Egypt Fruits and nuts

Table 2: The titles of CD-ROM's in the Corel library to which the system responded strongly, tabulated against the response. In each case, the sample consisted of 20 images out of the 100 on the CD-ROM. Figure 7 shows the skin filter output for the marked images from the dessert series, with groups overlaid.

skin filtering an image takes trivial amounts of time, and the grouper - which is not efficiently written - processes pictures at the rate of about 10 per hour.

The current implementation uses only a small set of grouping rules. We believe its performance could be improved substantially by techniques such as

- adding a face detector as an alternative to the skin filter, for initial triage,
- making the ribbon detector more robust,
- adding grouping rules for the structures seen in a typical side view of a human,
- adding grouping rules for close-up views of the human body, and/or
- extending the grouper to use the presence of other structures (e.g. heads) to verify the groups it produces.
- improving the notion of scale; at present, the system benefits by knowing that people in the pictures it will encounter occupy a fairly limited range of scales, but it is unable to narrow that range based on internal evidence. Inspecting the result images suggests that a process that allowed the system to (i) reason about the range of scales over which texture should be rejected and (ii) narrow the range of scales over which symmetries are accepted, would improve performance significantly.

Finally, once a tentative human has been identified, specific areas of the body might also be examined to determine whether the human is naked or merely scantily clad.

This system is an example constructed to illustrate a modified concept of an object model, which is a hybrid between appearance modelling and true 3D modelling. Such a

model consists of a series of predicates on 2D shapes, their spatial arrangements, and their color and texture. Each predicate can be tuned losely enough to accomodate variation in pose and imaging conditions, because selection combines information from all predicates. For efficiency, the simplest and most effective predicates (in our case, the skin filter) are applied first.

In this view of an object model, and of the recognition process, model information is available to aid segmentation at about the right stages in the segmentation process in about the right form. As a result, these models are present an effective answer to the usual critique of bottom up vision, that segmentation is too hard in that framework. In this view of the recognition process, the emphasis is on proceeding from general statements ("skin color") to particular statements ("a girdle"). As each decision is made, more specialised (and thereby more effective) grouping activities are enabled. Such a model is likely to be ineffective at particular distinctions ("John" vs "Fred"), but effective at the kind of broad classification required by this application—an activity that has been, to date, very largely ignored by the object recognition community. In our system, volumetric primitives enable a grouping strategy for segments, and object identity comes from segment relations. As a result, the recognition process is quite robust to individual variations, and the volumetric constraints simplify and strengthen grouping. In our opinion, this view of volumetric primitives as abstractions primarily for grouping is more attractive than the view in which the detailed geometric structure of the volumetric primitive identifies an object.

A great deal of work is required to fully elaborate and test this model of modelling and recognition, but there is reason to believe that it will extend to cover at least animals assembled according to the same basic body plan as humans. Our view of models gracefully handles objects whose precise geometry is extremely variable, where the identification of the object depends heavily on non-geometrical cues (e.g. color) and on the interrelationships between parts. While our present model is hand-crafted and is by no means complete, there is good reason to believe that an algorithm could construct a model of this form, automatically or semi-automatically, from a 3D object model or from a range of example images.

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. Anonymous referees drew our attention to relevance feedback, suggested checking intensity variation along a ribbon, and provided a number of references.

References

[1] Ahmad S., (1995) "A Usable Real-Time 3D Hand Tracker," *28th Asilomar Conference on Signals, Systems and Computers*, IEEE Computer Society Press.

[2] Akita, K., "Image sequence analysis of real world human motion," *Pattern Recognition*, **17**, 1, 73-83, 1984.

[3] Ashley, J., Barber, R., Flickner, M.D., Hafner, J.L., Lee, D., Niblack, W. and Petkovich, D. "Automatic and semiautomatic methods for image annotation and retrieval in QBIC," *SPIE Proc. Storage and Retrieval for Image and Video Databases III*, 24-35, 1995.

[4] Bajcsy, Ruzena (1973) "Computer Identification of Visual Surfaces," *Comp. Vis. Im. Proc.* 2/2, pp. 118–130

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

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

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

[8] Burel G., Carel D., (1994) "Detecting and localization of face on digital images" *Pattern Recognition Letters 15* pp 963-967.

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

[10] Candid home page at `http://www.c3.lanl.gov/ kelly/CANDID/main.shtml`

[11] Cedras C., Shah M., (1994) "A Survey of Motion Analysis from Moving Light Displays" *Computer Vision and Pattern Recognition* pp 214-221.

[12] Fleck,Margaret M. (1994) "Practical edge finding with a robust estimator," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 649–653.

[13] Fleck, Margaret M. (1996) "The Topology of Boundaries," in press, *Artificial Intelligence*.

[14] Forsyth, D.A., J.L. Mundy, A.P. Zisserman, A. Heller, C. Coehlo and C.A. Rothwell, "Invariant Descriptors for 3D Recognition and Pose," *IEEE Trans. Patt. Anal. and Mach. Intelligence*, **13**, 10, 1991.

[15] Gershon, Ron, Allan D. Jepson, and John K. Tsotsos (1986) "Ambient Illumination and the Determination of Material Changes," *J. Opt. Soc. America* A 3/10, pp. 1700–1707.

[16] Gong, Yihong and Masao Sakauchi (1995) "Detection of Regions Matching Specified Chromatic Features," *Comp. Vis. Im. Underst.* 61/2, pp. 263–269.

[17] Grimson, W.E.L. and Lozano-Pérez, T., "Localising overlapping parts by searching the interpretation tree", *PAMI*, **9**, 469-482, 1987.

[18] Hogg, D., "Model-based vision: a program to see a walking person," *Image and Vision Computing*, **1**, 1, 5-19, 1983.

[19] Huttenlocher, D.P. and Ullman, S., "Object recognition using alignment," *Proc. ICCV-1*, 102-111, 1986.

[20] Iowa City Press Citizen, "White House 'couples' set off indecency program," 24 Feb. 1996.

[21] Jacobs, C.E., Finkelstein, A., and Salesin, D.H., "Fast Multiresolution Image Querying," *Proc SIGGRAPH-95*, 277-285, 1995.

[22] Jacobs, Gerald H. (1981) *Comparative Color Vision,* Academic Press, New York.

[23] Kelly, P.M., Cannon, M., Hush, D.R., "Query by image example: the comparison algorithm for navigating digital image databases (CANDID) approach," *SPIE Proc. Storage and Retrieval for Image and Video Databases III*, 238-249, 1995.

[24] Kriegman, D. and Ponce, J., "Representations for recognising complex curved 3D objects," *Proc. International NSF-ARPA workshop on object representation in computer vision*, LNCS-994, 89-100, 1994.

[25] Layne, S.S., "Some issues in the indexing of images," *J. Am. Soc. Information Science*, **45**, 8, 583-588, 1994.

[26] Lamdan, Y., Schwartz, J.T. and Wolfson, H.J. "Object Recognition by Affine Invariant Matching," Proceedings CVPR, p.335-344, 1988.

[27] Lee, H.-J. and Chen, Z. "Determination of 3D human body postures from a single view," CVGIP, **30**, 148-168, 1985

[28] Leung, M.K., and Yang, Y.-H., "First sight: a human body labelling system," *PAMI*, **17**, 4, 359-377, 1995.

[29] Leung, T.K., Burl M.C., Perona P. (1995) "Finding faces in cluttered scenes using random labelled graph matching, " *International Conference on Computer Vision* pp 637-644.

[30] J. Liu, J.L. Mundy, D.A. Forsyth, A.P. Zisserman and C.A. Rothwell, "Efficient Recognition of rotationally symmetric surfaces and straight homogenous generalized cylinders," *IEEE conference on Computer Vision and Pattern Recognition '93*, 1993.

[31] Lowe, David G. (1987) "The Viewpoint Consistency Constraint," *Intern. J. of Comp. Vis,* 1/1, pp. 57–72.

[32] MacFormat, issue no. 28 with CD-Rom, September, 1995.

[33] Minka, T., "An image database browser that learns from user interaction," MIT media lab TR 365, 1995.

[34] Murase, H. and Nayar, S.K., "Visual learning and recognition of 3D objects from appearance," to appear, *Int. J. Computer Vision*, 1995.

[35] Nevatia, R. and Binford, T.O., "Description and recognition of curved objects," *Artificial Intelligence*, **8**, 77-98, 1977

[36] Niblack, W., Barber, R, Equitz, W., Flickner, M., Glasman, E., Petkovic, D., and Yanker, P. (1993) "The QBIC project: querying images by content using colour, texture and shape," *IS and T/SPIE 1993 Intern. Symp. Electr. Imaging: Science and Technology, Conference 1908, Storage and Retrieval for Image and Video Databases.*

[37] Ogle, Virginia E. and Michael Stonebraker (1995) "Chabot: Retrieval from a Relational Database of Images," *Computer* 28/9, pp. 40–48.

[38] O'Rourke, J. and Badler, N., "Model-based image analysis of human motion using constraint propagation," PAMI, **2**, 6, 522-536, 1980.

[39] Pentland, A., Picard, R.W., and Sclaroff, S. "Photobook: content-based manipulation of image databases," MIT Media Lab Perceptual Computing TR No. 255, Nov. 1993.

[40] Perez, Frank and Christof Koch (1994) "Towards Color Image Segmentation in Analog VLSI: Algorithm and Hardware," *Intern. J. of Comp. Vis.* 12/1, pp. 17–42.

[41] Pentland A., Moghaddam, B., Starner T., (1994) "View-based and modular eigenspaces for face recognition," in *Computer Vision and Pattern Recognition*, pp 84-91.

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

[43] "Detecting Activities" (1993) Polana R., Nelon R., in *Computer Vision and Pattern Recognition* pp 2-13.

[44] Price, R., Chua, T.-S., Al-Hawamdeh, S., "Applying relevance feedback to a photo-archival system," *J. Information Sci.*, **18**, 203-215, 1992.

[45] Rehg, J.M, Kanade, T., (1995) "Model-based tracking of self-occluding articulated objects," *International Conference on Computer Vision* pp 612-617.

[46] Rohr, K., "Towards model-based recognition of human movements in image sequences," CVGIP-IU, **59**, 1, 94-115, 1994

[47] Rossotti, Hazel (1983) *Colour: Why the World isn't Grey*, Princeton University Press, Princeton, NJ.

[48] Rothwell, C.A., A. Zisserman, J.L. Mundy and D.A. Forsyth, "Efficient Model Library Access by Projectively Invariant Indexing Functions," *Computer Vision and Pattern Recognition 92*, 109-114, 1992.

[49] Rowley, H., Baluja, S., Kanade, T. (1996) "Human Face Detection in Visual Scenes" To Appear in: *Neural Information Processing Systems 8.*

[50] Sanger, D., Haneishi, H. and Miyake, Y., "Method for light source discrimination and facial pattern detection from negative colour film," *J. Imaging Science and Technology*, **39**, 2, 166-175, 1995.

[51] Sclaroff, S. "World wide web image search engines," Boston University Computer Science Dept TR95-016, 1995.

[52] Stein, F. and Medioni, G., "Structural indexing: efficient 3D object recognition," *PAMI-14*, 125-145, 1992.

[53] Sung, K.K, Poggio, T., (1994) "Example-based Learning from View-based Human Face Detection" MIT A.I. Lab Memo No. 1521.

[54] Taubin, G. and Cooper, D.B., "Object recognition based on moment (or algebraic) invariants," in J.L. Mundy and A.P. Zisserman (ed.s) *Geometric Invariance in Computer Vision*, MIT Press, 1992.

[55] B. Taylor, Tense and Continuity, Linguistics and Philosophy 1 (1977) 199–220.

[56] C. L. Tenny, Grammaticalizing Aspect and Affectedness, Ph.D. thesis, Linguistics and Philosophy, Massachusetts Inst. of Techn. (1987).

[57] Treisman, Anne (1985) "Preattentive Processing in Vision," *Com. Vis. Grap. Im. Proc.* 31/2, pp. 156–177.

[58] Ullman, S. and Basri, R. (1991). Recognition by linear combination of models, *IEEE PAMI*, **13**, 10, 992-1007.

[59] Virage home page at `http://www.virage.com/`

[60] Weiss, I. "Projective Invariants of Shapes," Proceeding DARPA Image Understanding Workshop, p.1125-1134, April 1988.

[61] Whorf, Benjamin Lee (1941) "The Relation of Habitual Thought and Behavior to Language," in Leslie Spier, ed., *Language, culture, and personality, essays in memory of Edward Sapir*, Sapir Memorial Publication Fund, Menasha, WI.

[62] 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.

[63] Zisserman, A., Mundy, J.L., Forsyth, D.A., Liu, J.S., Pillow, N., Rothwell, C.A. and Utcke, S. (1995) "Class-based grouping in perspective images",*Intern. Conf. on Comp. Vis.*

[64] Zerroug, M. and Nevatia, R. "From an intensity image to 3D segmented descriptions," ICPR, 1994.