??, ??, 1–16 (??)

Automatic detection of human nudes

D.A. FORSYTH
daf@cs.berkeley.edu
Computer Science Division, University of California at Berkeley,Berkeley, CA 94720

M.M. FLECK
fleck@cs.hmc.edu
Computer Science Department, Harvey Mudd College, Claremont, CA 91711

Received ??; Revised ??

Editors: ??

Abstract.
This paper demonstrates an automatic system for telling whether there are human nudes present in an image. The system marks skin-like pixels using combined color and texture properties. These skin regions are then fed to a specialized grouper, which attempts to group a human figure using geometric constraints on human structure. If the grouper finds a sufficiently complex structure, the system decides a human is present. The approach is shown to be effective for a wide range of shades and colors of skin and human configurations. This approach offers an alternate 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 human nudes, mostly obtained from the internet, and 4289 assorted control images, drawn from a wide variety of sources.

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

Several typical collections containing over ten million images are listed in [16]. In the most comprehensive field study of usage practices (a paper by [16] surveying the use of the Hulton Deutsch collection), there is a clear user preference for searching these collections on image semantics; typical queries observed are overwhelmingly oriented toward object classes ("dinosaurs," p. 40, "chimpanzee tea party, early," p. 41) or instances ("Harry Secombe," p. 44, "Edward Heath gesticulating," p. 45). 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. The primary recognition problem in this application is *finding*, where the image components that result from a single object are collected together (rather than *naming*, where the particular name of a single isolated object is determined).

The technology does not exist to build programs that can find images based on complex semantic

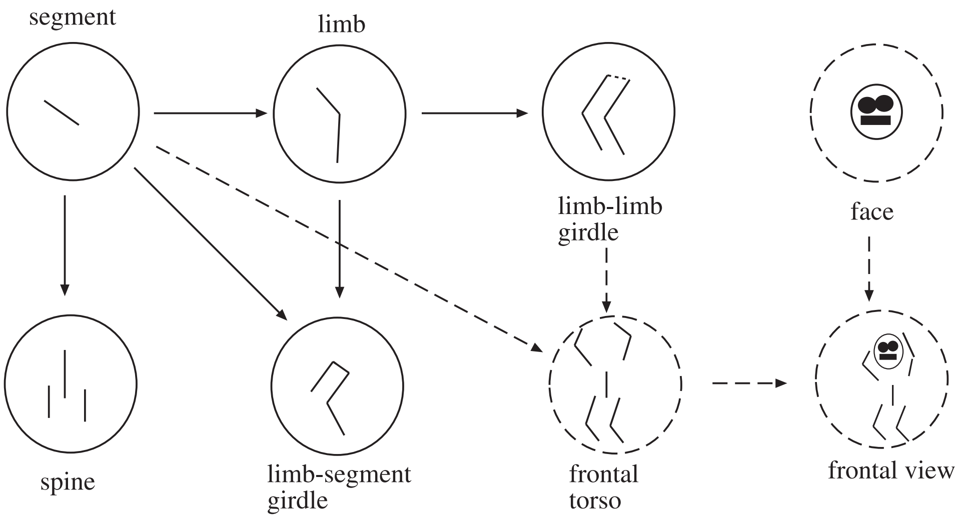

Fig. 1. The grouping rules (arrows) specify how to assemble simple groups (e.g. body segments) into complex groups (e.g. limb-segment girdles). These rules incorporate constraints on the relative positions of 2D features, induced by constraints on 3D body parts. Dashed lines indicate grouping rules that 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.

notions of content ("the pope, kissing a baby") with high recall. However, there are many applications where low recall is not a problem. In the Enser study of a stock photo agency, for example, requesters are seldom burdened with more than 10 pictures, whatever the subject matter. As another example, consider filtering internet connections for offensive images; as long as a manager can be reasonably certain that any protracted flow of such images will result in an alarm — say 50% recall — the tool is usable. Usually, low precision is a greater problem because it will confuse and annoy a user.

Determining whether an image contains a human nude is a natural problem: this is a form of semantic content which produces strong cultural responses; marking such images based on textual or contextual cues can be embarassingly unreliable[1]; it is a difficult recognition problem, which stresses abstraction over geometric matching; and there is some prospect of producing a usable solution.

1. Background

There is an extensive literature on finding images using features such as colour histograms, texture measures and shape measures (e.g. [1, 19, 21, 25, 27, 32, 33, 38, 39, 42, 48, 54, 65, 66, 67, 68, 72]). Typically, work in this area considers whole image matches rather than semantic ("a fish is present") matches (with the exception of the face matching module in Photobook [53]). This is largely because current object recognition methods cannot handle the demands presented by semantic queries.

Current approaches to recognition rely on detailed geometric models (e.g. [17, 22, 26, 31, 35, 40, 43, 44, 59, 60, 69, 70, 71]) or on parametric families of templates (e.g. [34, 45, 46, 55, 62, 63, 56]). For more complex objects, models consist of composites of primitives, which are themselves either parametric families of templates (e.g. [13, 30, 37, 49, 57]) or surfaces chosen from "nice" families (e.g. [8, 9, 12, 41, 47]).

Modelling humans and animals as assemblies of cylinders has an established history (for example, [15, 23, 29, 50, 58, 60, 11]). Early work by Marr and Nishihara [41] viewed recognition of humans as a process of obtaining cylinders at a variety of scales; at the coarsest scale, the whole form would be a cylinder; at finer scales, sub-assemblies would emerge. Their approach assumes that *only the components of the image corresponding to the person sought actually contribute to this process* – i.e. that finding has already occurred.

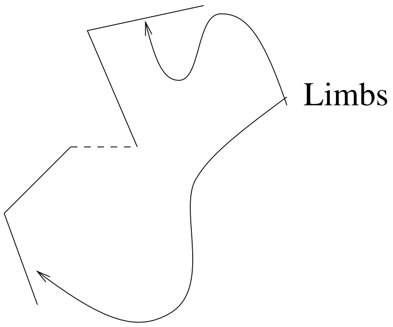

Fig. 2. A configuration that is prohibited by geometric constraints on assembling groups. The dashed line represent the girdle. Neither arms nor legs can be configured to look like this. There are configurations that are possible for the one girdle, but not the other.

Other systems group regions into stick figures using reasoning about gravity [36] or knowledge about the background [7, 73]. Gesture recognition is now widely studied; constraints on the colour of skin are commonly exploited [3, 4, 6, 64], though we are not aware of any use of texture constraints. Advanced template matching strategies work extremely well at finding faces [52, 62, 63, 56] and standing pedestrians (who have a characteristic "lollipop-like" appearance [51]).

Our 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 nudes 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. We therefore detect nudes 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 nudes.

2. Finding Skin

The color of human skin results from a combination of blood (red) and melanin (yellow, brown) [61]. Human skin has a restricted range of hues and is 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 offwhite. 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.

2.1. 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. [24]) as follows:

$$\begin{aligned} L(x) &= 105\log_{10}(x+1+n) \\ I &= L(G) \\ R_g &= L(R)-L(G) \\ B_y &= L(B)-\frac{L(G)+L(R)}{2} \end{aligned}$$

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. The resulting values are a type of median absolute deviation estimate of variation (see [28]). These operations use a fast multi-ring approximation to the median filter [18].

The texture amplitude and the smoothed R_g and B_y values are then passed to a tightly-tuned skin filter. It marks as probable 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 desaturated 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.

3. 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.

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 [14] 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 [10] are found by a straightforward algorithm. Sets of points forming regions with roughly straight axes ("ribbons" [12]) are found using an algorithm based on the Hough transform. 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^2 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 8 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

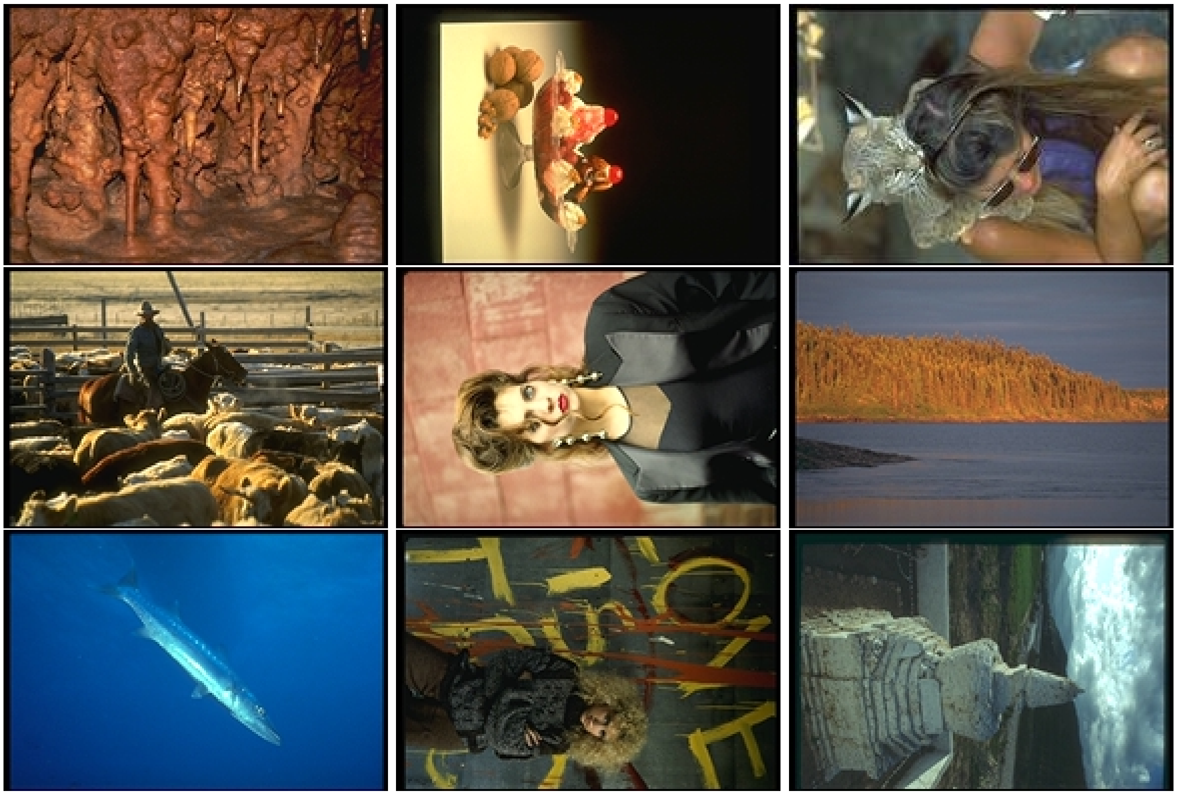

Fig. 3. Typical control images. The images in the first row are incorrectly classified as containing nudes; those in the second row pass the skin test, but are rejected by the geometric grouping process; and those in the third row are rejected by the skin test.

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: search intervals extending from their ends 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 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 are forbidden configurations 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.

4. Experimental protocol

The performance of the system was tested using 565 target images of nudes and 4302 assorted control images, containing some images of people but none of nudes. 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[3] by collecting about 100-150 images per sample on several occasions. The origin of the test images was not recorded[4]. 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 widely 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 test 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):

- 1241 images sampled[5] 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 [2];
- 11 assorted personal photos, re-photographed with our CCD camera;
- 47 pictures of objects and textures taken in our laboratory for other purposes;
- 1200 pictures, consisting of the complete contents of a series of CD-ROM's in the Corel stock photo library (titles in the appendix);
- 41 images from CD-ROM 135000 (bald eagles) in the Corel stock photo library;
- 1660 pictures, consisting of the every fifth image from a series of CD-ROM's in the second edition of the Corel stock photo library (see the appendix for titles).

The DWR images and Corel images were available at a resolution of 128 by 192 pixels. The images from other sources were automatically resized (and, if necessary, rotated) to obtain the minimum reduction that fit the image into a block this 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. 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 nude is 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 con-*

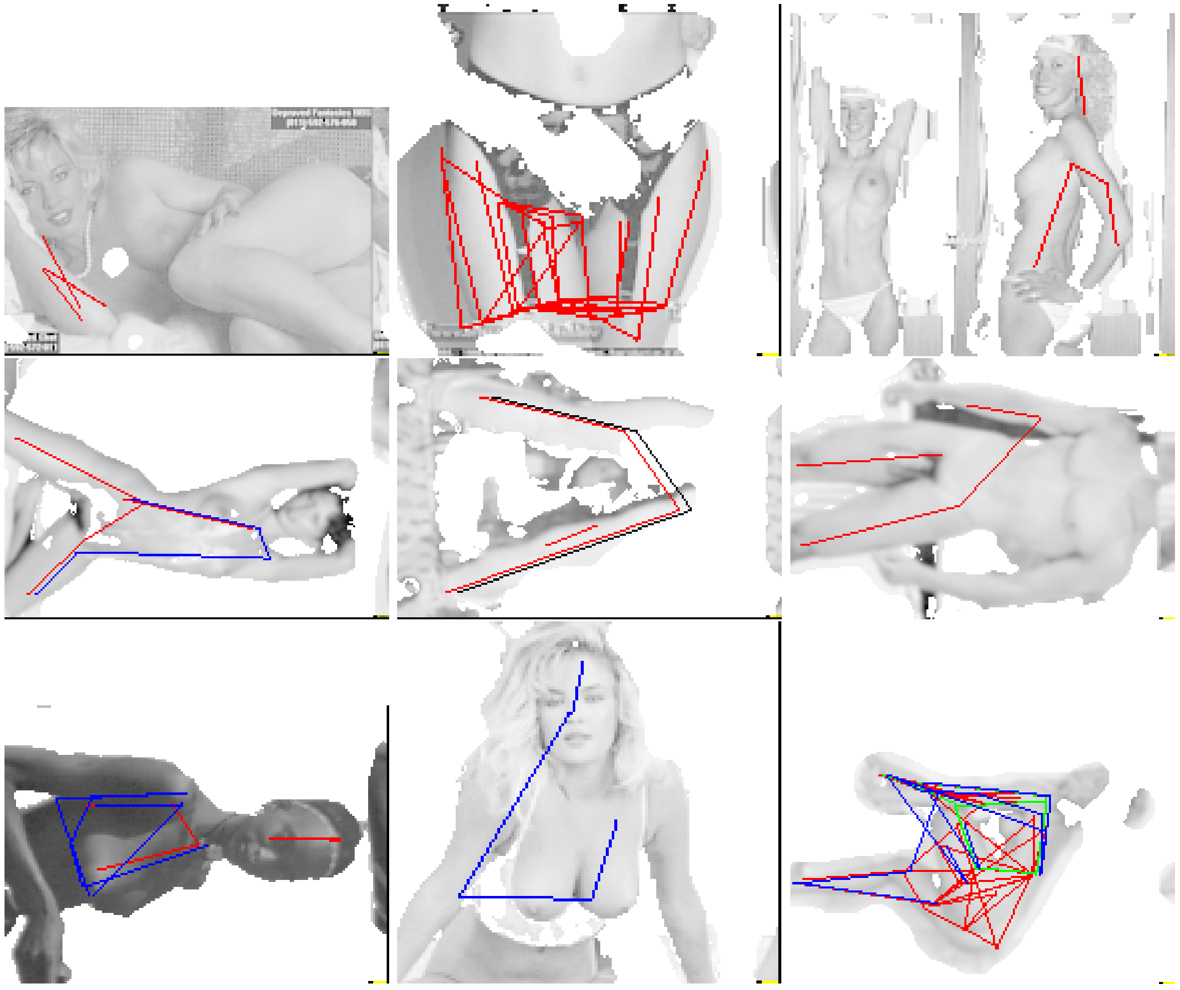

Fig. 4. Typical images correctly classified as containing nudes. The output of the skin filter is shown, with spines overlaid in red, limb-limb girdles overlaid in green, and limb-segment girdles overlaid in blue. Notice that there are cases in which groups form quite good stick figures; in which the groups are wholly unrelated to the limbs; in which accidental alignment between figures and background cause many highly inaccurate groups; and in which 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.

figuration, a version of the grouper which requires that a girdle or spine group be present for a nude 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

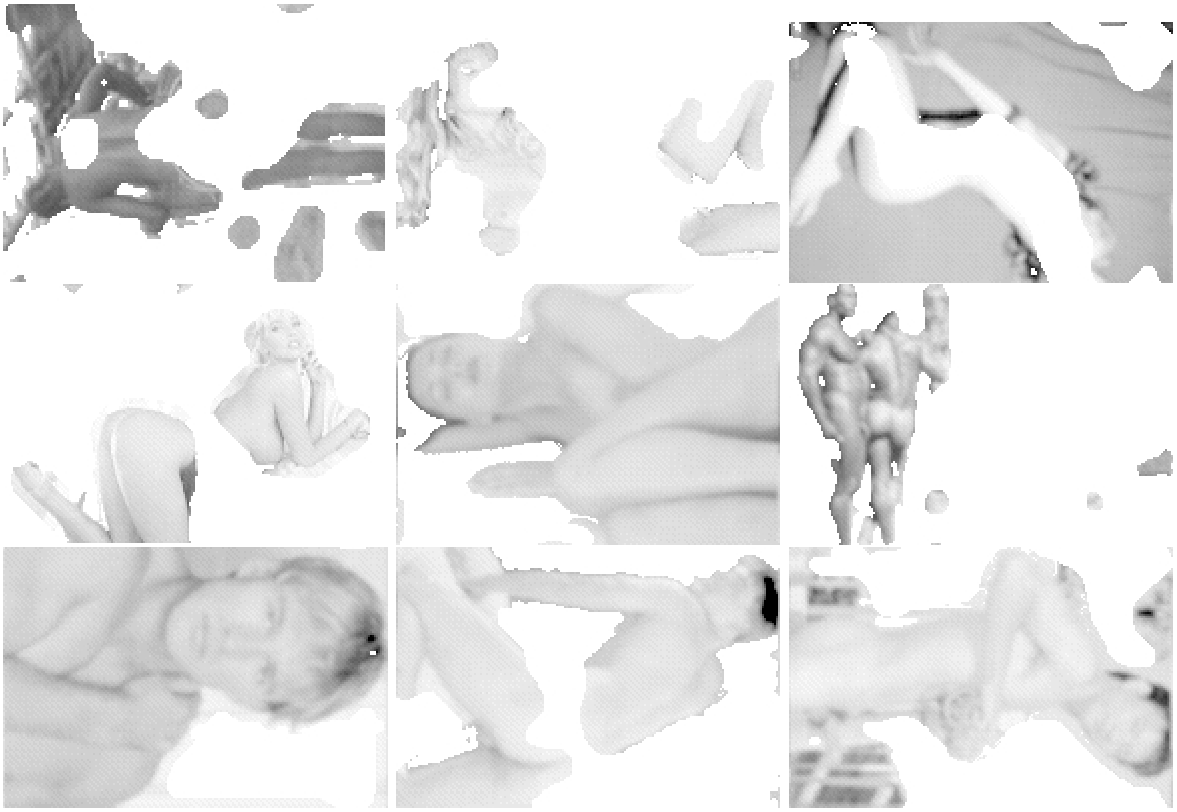

Fig. 5. Typical false negatives: the skin filter marked significant areas of skin, but the geometrical analysis could not find a girdle or a spine. Failure is often caused by absence of limbs, low contrast, or configurations not included in the geometrical model (notably side views, head and shoulders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

well the system, 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.

5.1. 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 and texture 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. This approach cannot yield a useful tool on its own, because of the high rate of false positives.

Mistakes by the skin filter occur for several reasons. In some test images, the nudes 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

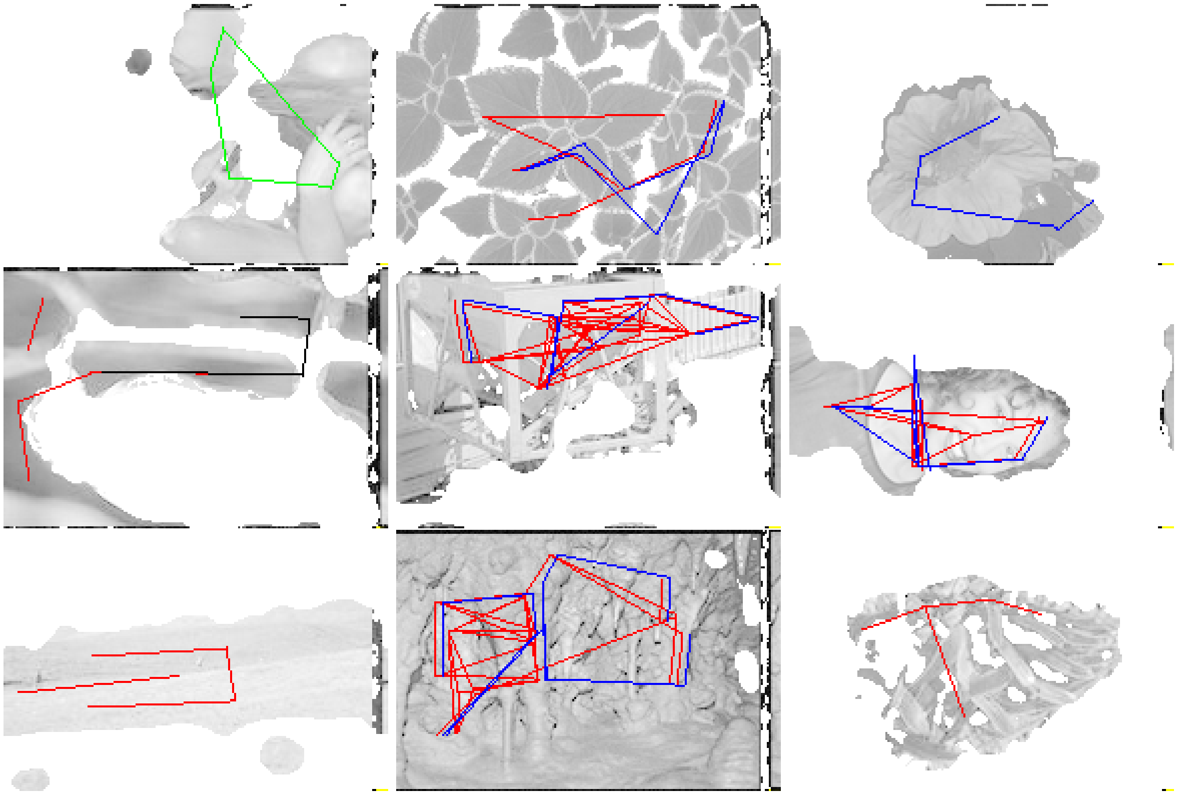

Fig. 6. Typical control images wrongly classified as containing nudes. These images contain people or skin-colored material (animal skin, wood, bread, off-white walls) and structures which the geometric grouper mistakes for spines (red) or girdles. Limb-limb girdles are shown in blue, limb-segment girdles 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.

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 nudes: the skin filter appears to perform equally well on all races.

5.2. The grouping process

The grouper ran on images with sufficient skin pixels, a total of 448 test images and 485 control images. The primary grouper marked 241 test images and 182 control images, meaning that the entire system composed of primary grouper operating on skin filter output displayed a response ratio of 10.0 and a test response of 43%. Considered on its own, the grouper's response ratio was 1.4, and the selectivity of the system was 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 marked structures which are parts of the human body, but not hip or shoulder girdles. This presents no major problem, as the program was trying to detect the presence of humans, rather than analyze their pose in detail.

False negatives occured for several reasons. Some close-up or poorly cropped images did 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 can fail due to poor contrast between limbs and their surroundings. Structural complexity in the image, often caused by strongly colored items of clothing, can confuse 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 was notably poor for side views of figures with arms hanging down.

Some of the control images which were 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 marked 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 7 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.

6. Discussion and Conclusions

This paper has shown that images of nudes 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, but it is performing a much more abstract task (“find a nude” 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; 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 performance would be improved significantly by 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.

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 nude 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.

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%

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 present an effective answer to the usual critique of bottom up vision, that segmentation is too hard in that framework. 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 used primarily for grouping is more attractive than the view in which the detailed geometric structure of the volumetric primitive identifies an object. Recent work has shown that representations of this form can be learned from image data [20]. Important cues that are currently absent from our representation, such as the characteristic variation in width along a body segment due to musculature, might be incorporated with some sort of probabilistic model once the grouping process is over.

This purely bottom up approach has some weaknesses:

- **Resurrecting dead hypotheses** is impossible. Once one test has rejected an assembly, no hypothesis containing that assembly can succeed. Even a relatively small error rate in the early tests can lead to substantial problems. This is a standard problem in bottom-up recognition, which is usually dealt with by setting a high standard for hypothesis rejection and then doing a lot of work on verification.
- **Handling ambiguity** is difficult. We expect some types of image group to be more easy to confuse than others. Typically image evidence varies in different ways depending on the hypothesis. For example, a skin coloured region could be a (closed) eye, but never a nostril.

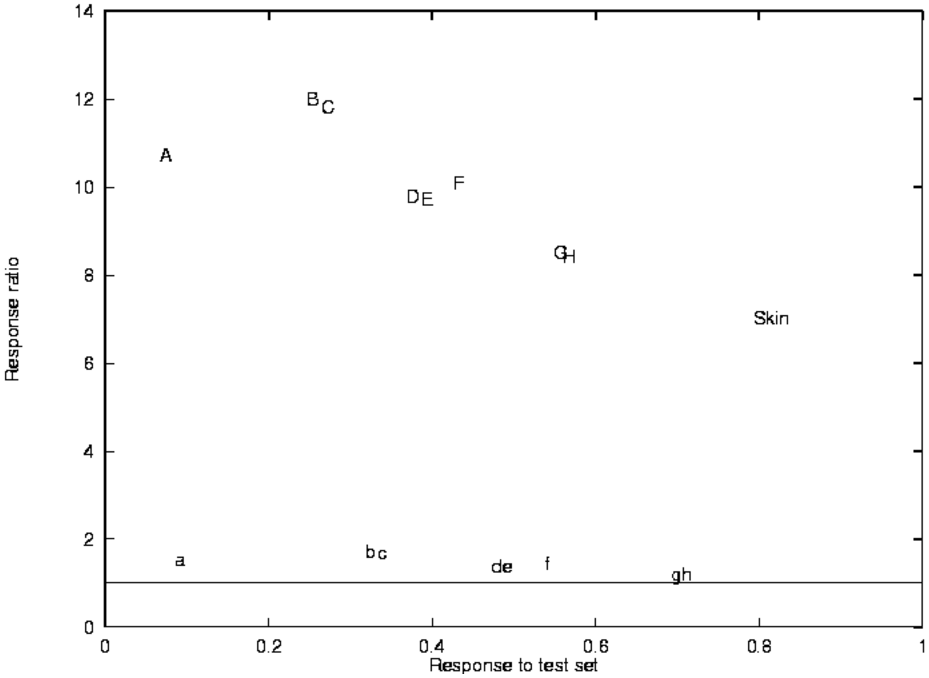

Fig. 7. The response ratio, (percent incoming test images marked/percent incoming control images marked), plotted against the percentage of test images marked, for various configurations of the nude finder. Labels "A" through "H" indicate the performance of the entire system which looks for configurations of skin-like ribbons. Here "F" is the primary configuration of the grouper, and other labels correspond to asserting a nude is present based on different groups. The label "skin" shows the performance obtained by checking the number of skin-like pixels 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 nude person is present. The horizontal line shows response ratio one, which would be achieved by chance; the grouper beats chance significantly. While the grouper's selectivity is less than that of the skin test, 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.

- It does not **fuse or split groups**. One instance may lead to many image groups. For example, lighting may cause one body segment to appear as two image segments; similarly, a straight arm may appear as only one image segment representing two body segments. These groups should be fused or split, respectively.
- **Prioritizing hypotheses** is important; not all acceptable groups are equally likely. The current mechanism generates all acceptable groups, but cannot determine whether any are superior to others. A natural enhancement is to determine the value of the posterior for each acceptable group; but this does not deal with the difficulties above.
- **Uniqueness** is a problem; there is no mechanism that prevents the simultaneous use of two groups that contain almost the same segments. The solution to this problem is to see the groups as *evidence* rather than *instances* — thus, several similar groups may be evidence for one, rather than many, instances. Uniqueness reasoning at the evidence level is notoriously unreliable — one image segment may genuinely represent evidence of two body segments. However, uniqueness reasoning at the instance level is much more plausible — two people cannot occupy the same volume of space, for example, and visibility reasoning is possible, too.

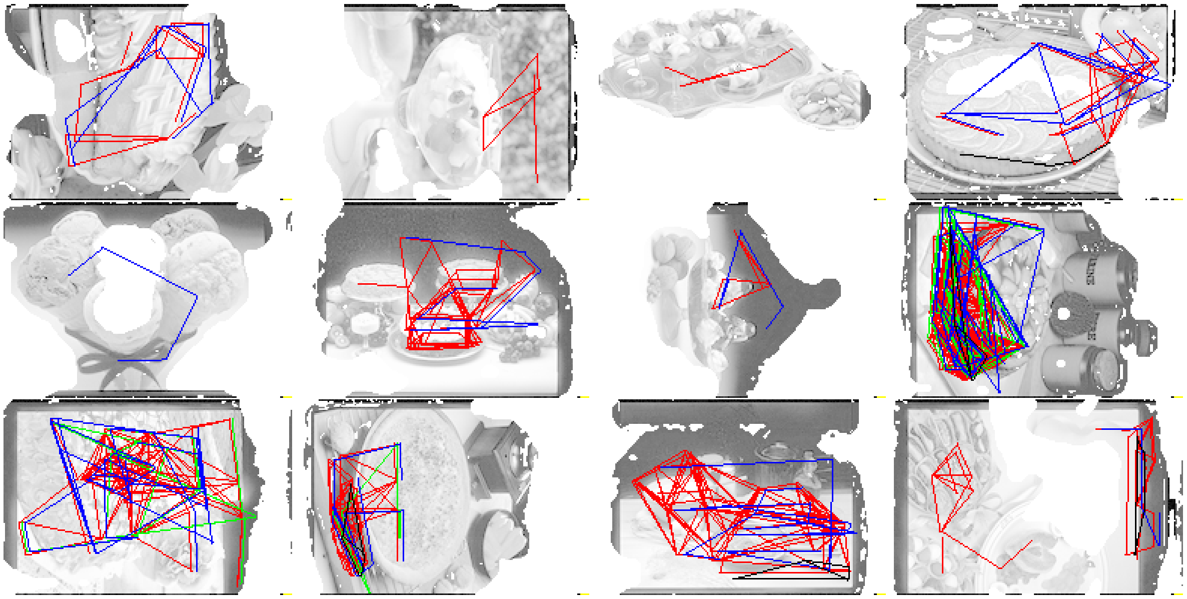

Fig. 8. All the control images in the dessert sequence that were marked by the system, with groups overlaid. These images contain large regions of skin-coloured material, with texture at a scale invisible to the skin filter. Since there are many edges in the skin filter output, a large collection of 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.

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 8 shows the skin filter output for the marked images from the dessert series, with groups overlaid.

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

All these difficulties point to the need for a top-down revision mechanism that can advance groups past tests that they have failed, fuse and split groups, and collect many groups as evidence for one instance. We are currently experimenting with the use of Markov chain Monte Carlo methods for this purpose.

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. This paper would probably not have appeared without Robert Wilensky's patience and diplomacy.

Appendix Control CD titles

We used each image from: 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), in the **first series** of CD's.

We used every tenth image from: 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 (Autumn 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), 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 series** of the Corel stock photo library[6]

Notes

[1] Incongruities occasionally receive media attention; in a recent incident, a commercial package for avoiders refused to allow access to the White House childrens page[5].

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

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

[4] 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. The "clumpy" nature of this sort of content could be used as a cue to improve recognition.

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

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

References

1. Virage home page at `http://www.virage.com/`.
2. Macformat, issue no. 28 with cd-rom, 1995.
3. *Proceedings of the First Int. Workshop on Automatic Face- and Gesture-Recognition*, 1995.
4. *Proceedings of the Second Int. Workshop on Automatic Face- and Gesture-Recognition*, 1996.
5. White house 'couples' set off indecency program. *Iowa City Press Citizen*, 24 Feb., 1996.
6. S. Ahmad. A usable real-time 3d hand tracker. In *28th Asilomar Conference on Signals, Systems and Computers*, 1995.
7. K. Akita. Image sequence analysis of real world human motion. *Pattern Recognition*, 17(1):73–83, 1984.
8. I. Biederman. Recognition-by-components: A theory of human image understanding. *Psych. Review*, 94(2):115–147, 1987.
9. T.O. Binford. Visual perception by computer. In *Proc. IEEE Conference on Systems and Control*, 1971.

10. J.M. Brady and H. Asada. Smoothed local symmetries and their implementation. *International Journal of Robotics Research*, 3(3), 1984.
11. C. Bregler and J. Malik. Tracking people with twists and exponential maps. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 8–15, 1998.
12. R.A. Brooks. Symbolic reasoning among 3-D models and 2-D images. *Artificial Intelligence*, 17(1-3):285–348, 1981.
13. M.C. Burl, T.K. Leung, and P. Perona. Face localisation via shape statistics. In *Int. Workshop on Automatic Face and Gesture Recognition*, 1995.
14. J.F. Canny. A computational approach to edge detection. *IEEE Trans. Patt. Anal. Mach. Intell.*, 8(6):679–698, November 1986.
15. Z. Chen and H.-J. Lee. Knowledge-guided visual perception of 3d human gait from a single image sequence. *IEEE T. Pattern Analysis and Machine Intelligence*, 22(2):336–342, 1992.
16. P.G.B. Enser. Query analysis in a visual information retrieval context. *J. Document and Text Management*, 1(1):25–52, 1993.
17. O.D. Faugeras and M. Hebert. The representation, recognition, and locating of 3-D objects. *International Journal of Robotics Research*, 5(3):27–52, Fall 1986.
18. M.M. Fleck. Practical edge finding with a robust estimator. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 649–653, 1994.
19. M. Flickner, H. Sawhney, W. Niblack, and J. Ashley. Query by image and video content: the qbic system. *Computer*, 28(9):23–32, 1995.
20. D.A. Forsyth and M.M. Fleck. Body plans. In *IEEE Conf. on Computer Vision and Pattern Recognition*, 1997.
21. D.A. Forsyth, J. Malik, M.M. Fleck, H. Greenspan, T. Leung, S. Belongie, C. Carson, and C. Bregler. Finding pictures of objects in large collections of images. In *Proc. 2'nd International Workshop on Object Representation in Computer Vision*, 1996.
22. D.A. Forsyth, J.L. Mundy, A.P. Zisserman, C. Coelho, A. Heller, and C.A. Rothwell. Invariant descriptors for 3d object recognition and pose. *PAMI*, 13(10):971–991, 1991.
23. D.M. Gavrila and L.S. Davis. 3d model-based tracking of humans in action: a multi-view approach. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 73–80, 1996.
24. R. Gershon, A.D. Jepson, and J.K. Tsotsos. Ambient illumination and the determination of material changes. *J. Opt. Soc. America*, A-3(10):1700–1707, 1986.
25. M.M. Gorkani and R.W. Picard. Texture orientation for sorting photos "at a glance". In *Proceedings IAPR International Conference on Pattern Recognition*, pages 459–64, 1994.
26. 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.
27. A. Hampapur, A. Gupta, B. Horowitz, and Chiao-Fe Shu. Virage video engine. In *Storage and Retrieval for Image and Video Databases V – Proceedings of the SPIE*, volume 3022, pages 188–98, 1997.
28. D. C. Hoaglin, F. Mosteller, and J. W. Tukey, editors. *Understanding Robust and Exploratory Data Analysis.* John Wiley, 1983.
29. D. Hogg. Model based vision: a program to see a walking person. *Image and Vision Computing*, 1(1):5–20, 1983.
30. C-Y. Huang, O.T. Camps, and T. Kanungo. Object recognition using appearance-based parts and relations. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 877–83, 1997.
31. D.P. Huttenlocher and S. Ullman. Object recognition using alignment. In *Proc. Int. Conf. Comp. Vision*, pages 102–111, London, U.K., June 1987.
32. C.E. Jacobs, A. Finkelstein, and D.H. Salesin. Fast multiresolution image querying. In *Proc SIGGRAPH-95*, pages 277–285, 1995.
33. P.M. Kelly, M. Cannon, and D.R. Hush. Query by image example: the comparison algorithm for navigating digital image databases (candid) approach. In *SPIE Proc. Storage and Retrieval for Image and Video Databases III*, pages 238–249, 1995.
34. J.J. Koenderink and A.J. Van Doorn. The internal representation of solid shape with respect to vision. *Biological Cybernetics*, 32:211–216, 1979.
35. D.J. Kriegman and J. Ponce. On recognizing and positioning curved 3D objects from image contours. *IEEE Trans. Patt. Anal. Mach. Intell.*, 12(12):1127–1137, December 1990.
36. M.K. Leung and Y-H. Yang. First sight: a human body labelling system. *IEEE T. Pattern Analysis and Machine Intelligence*, 17(4):359–377, 1995.
37. T.K. Leung, M.C. Burl, and P. Perona. Finding faces in cluttered scenes using random labelled graph matching. In *Int. Conf. on Computer Vision*, 1995.
38. P. Lipson, W.E. L. Grimson, and P. Sinha. Configuration based scene classification and image indexing. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 1007–13, 1997.
39. F. Liu and R.W. Picard. Periodicity, directionality, and randomness: Wold features for image modeling and retrieval. *IEEE T. Pattern Analysis and Machine Intelligence*, 18:722–33, 1996.
40. D. Lowe. Three-dimensional object recognition from single two-dimensional images. *Artificial Intelligence*, 31(3):355–395, 1987.
41. D. Marr and K. Nishihara. Representation and recognition of the spatial organization of three-dimensional shapes. *Proc. Royal Society, London*, B-200:269–294, 1978.
42. T.P. Minka and R.W. Picard. Interactive learning with a "society of models". *Pattern Recognition*, 30:465–481, 1997.
43. J.L. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision.* MIT Press, Cambridge, Mass., 1992.
44. J.L. Mundy, A. Zisserman, and D. Forsyth. *Applications of Invariance in Computer Vision*, volume 825 of *Lecture Notes in Computer Science*. Springer-Verlag, 1994.
45. H. Murase and S. Nayar. Visual learning and recognition of 3D objects from appearance. *Int. J. of Comp. Vision*, 14(1):5–24, 1995.
46. S.K. Nayar, S.A. Nene, and H. Murase. Real time 100 object recognition system. In *Int. Conf. on Robotics and Automation*, pages 2321–5, 1996.

47. R. Nevatia and T.O. Binford. Description and recognition of complex curved objects. *Artificial Intelligence*, 8:77–98, 1977.

48. V.E. Ogle and M. Stonebraker. Chabot: retrieval from a relational database of images. *Computer*, 28:40–8, 1995.

49. M. Oren, C. Papageorgiou, P. Sinha, and E. Osuna. Pedestrian detection using wavelet templates. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 193–9, 1997.

50. J. O'Rourke and N. Badler. Model-based image analysis of human motion using constraint propagation. *IEEE T. Pattern Analysis and Machine Intelligence*, 2:522–546, 1980.

51. E. Osuna, R. Freund, and F. Girosi. Training support vector machines: an application to face detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 130–6, 1997.

52. A. Pentland, B. Moghaddam, and T. Starner. View-based and modular eigenspaces for face recognition. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 84–91, 1994.

53. A. Pentland, R. Picard, and S. Sclaroff. Photobook: content-based manipulation of image databases. *Int. J. Computer Vision*, 18(3):233–54, 1996.

54. R.W. Picard, T. Kabir, and F. Liu. Real-time recognition with the entire brodatz texture database. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 638–9, 1993.

55. H. Plantinga and C. Dyer. Visibility, occlusion, and the aspect graph. *Int. J. of Comp. Vision*, 5(2):137–160, 1990.

56. T. Poggio and Kah-Kay Sung. Finding human faces with a gaussian mixture distribution-based face model. In *Asian Conf. on Computer Vision*, pages 435–440, 1995.

57. A.R. Pope and D.G. Lowe. Learning object recognition models from images. In *Int. Conf. on Computer Vision*, pages 296–301, 1993.

58. J. Rehg and T. Kanade. Model-based tracking of self-occluding articulated objects. In *ICCV-95*, pages 612–617, 1995.

59. L.G. Roberts. Machine perception of three-dimensional solids. In J.T. Tippett et al., editor, *Optical and Electro-Optical Information Processing*, pages 159–197. MIT Press, Cambridge, 1965.

60. K. Rohr. Incremental recognition of pedestrians from image sequences. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 9–13, 1993.

61. H. Rossotti. *Colour: Why the World isn't Grey.* Princeton University Press, Princeton, NJ, 1983.

62. H.A. Rowley, S. Baluja, and T. Kanade. Human face detection in visual scenes. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing 8*, pages 875–881, 1996.

63. H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 203–8, 1996.

64. D. Sanger, H. Haneishi, and Y. Miyake. Method for light source discrimination and facial pattern detection from negative colour film. *J. Imaging Science and Technology*, 2:166–175, 1995.

65. S. Santini and R. Jain. Similarity queries in image databases. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 646–651, 1996.

66. M. Stricker and M.J. Swain. The capacity of color histogram indexing. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 704–8, 1994.

67. M.J. Swain. Interactive indexing into image databases. In *Storage and Retrieval for Image and Video Databases – Proceedings of the SPIE*, volume 1908, pages 95–103, 1993.

68. M.J. Swain and D.H. Ballard. Color indexing. *Int. J. Computer Vision*, 7(1):11–32, 1991.

69. D.W. Thompson and J.L. Mundy. Three-dimensional model matching from an unconstrained viewpoint. In *IEEE Int. Conf. on Robotics and Automation*, pages 208–220, Raleigh, NC, April 1987.

70. S. Ullman. *High-level Vision: Object Recognition and Visual Cognition.* MIT Press, 1996.

71. S. Ullman and R. Basri. Recognition by linear combination of models. *IEEE Trans. Patt. Anal. Mach. Intell.*, 13(10):992–1006, 1991.

72. D.A. White and R. Jain. Imagegrep: fast visual pattern matching in image databases. In *Storage and Retrieval for Image and Video Databases V – Proceedings of the SPIE*, volume 3022, pages 96–107, 1997.

73. C. Wren, A. Azabayejani, T. Darrell, and A. Pentland. Pfinder: real-time tracking of the human body. Mit media lab perceptual computing section tr 353, MIT, 1995.