Identifying nude pictures

D.A. Forsyth

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

M.M. Fleck

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

Abstract

This paper 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, Content Based Retrieval.*

1 Introduction

The recent explosion in internet usage and multimedia 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. 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. Seeking or avoiding such images based on origins alone (as commercial software does) leads to incongruities; in a recent incident, a commercial package for avoiders refused to allow access to the White House childrens' page[?].

Similarly, coding the appearance of individual regions by colour or texture properties 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. Existing content based retrieval systems (e.g. [?, ?, ?, ?]) do not contain codings of object shape that are 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; furthermore, because these properties are not coded, combinations diagnostic for particular objects cannot be learned. Equally, current object recognition algorithms (e.g. [?, ?, ?]) perform poorly for queries as abstract as "Find naked people." Automatic segmentation at a satisfactory level remains an extremely difficult problem for object recognition or image database systems. Work on finding people typically concentrates either on motion cues or on specific body parts like faces and hands against a known or simplified background (e.g. [?]); there is little work on segmentation. None of these systems is suitable for analyzing typical images of naked people found on the internet.

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. Images of naked people may vary in content but must contain significant regions of skin pixels, which must be organised into limb segments. Finally, the relative configurations of segments must satisfy geometric constraints. These observations suggest the use of a representation that emphasizes assemblies of a constrained class of primitive; typical versions of this idea appear in [?, ?, ?].

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.

2 Finding Skin

The color of a human's skin is created by a combination of blood (red) and melanin (yellow, brown) [?]. 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.

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. 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 [?].

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 fall within an appropriate range. 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, which 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. 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 [?] 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 [?] are found by a straightforward algorithm. Sets of points forming regions with roughly straight axes ("ribbons" [?]) are found using an algorithm based on the Hough transformation.

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 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, not too many edges must be crossed. There is no angular constraint on axes in grouping limbs. 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 a forbidden configurations of limbs. 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 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 by collecting about 100-150 images per sample on several occasions. The origin of the test images was not recorded. 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. 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:

- 1241 images sampled 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,
- 44 assorted images from a photo CD that came with a copy of a magazine [?],
- 11 assorted personal photos, re-photographed with our CCD camera, and
- 47 pictures of objects and textures taken in our laboratory for other purposes.
- a total of 2901 pictures sampled from the Corel stock photo libraries I and II.

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 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.

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 figure **??** 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.

5.2 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. Figure **??** 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.

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

6 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 by detecting jointed objects of highly variable shape, in a diverse range of poses, seen from many different camera positions. 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; 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; extending the grouper to use the presence of other structures (e.g. heads) to verify the groups it produces; and improving the notion of scale. 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.

Finally, this work demonstrates that object models quite different from those commonly used in computer vision offer the prospect of effective recognition systems that can work in quite general environments. In this approach, an object is modelled as a loosely coordinated collection of detection and grouping rules. The object is recognized if a suitable group can be built. Grouping rules incorporate both surface properties (color and texture) and shape information. This type of model 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.

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. IRI-9209728, IRI-9420716, IRI-9501493,

References

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

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

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

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

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

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

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

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

[9] 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.

[10] 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.

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

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

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

[14] 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.

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

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

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

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

[19] 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.

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

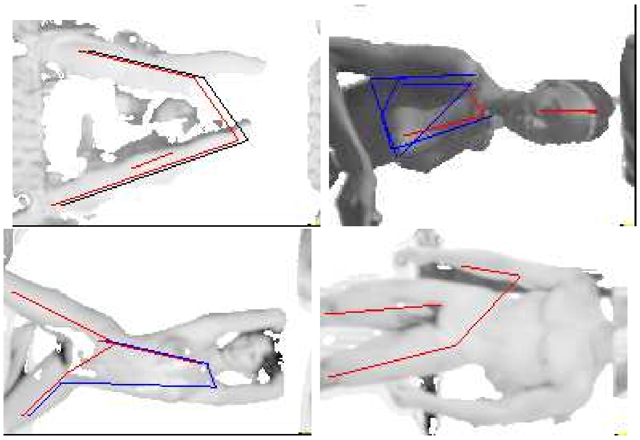

Figure 1: *Typical images correctly classified as containing naked people. The output of the skin filter is shown, with spines, limb-limb girdles, and limb-segment girdles overlaid. Notice that there are cases in which groups form quite good stick figures; where the groups are wholly unrelated to the 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.*

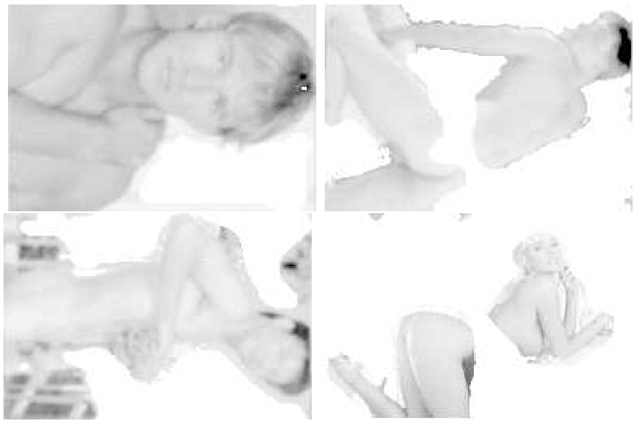

Figure 2: *Typical false negatives: the skin filter marked significant areas of skin, but the geometrical analysis could not find a girdle or a spine. Failure is often caused by absence of limbs, low contrast, or configurations not included in the geometrical model (notably side views, head and shoulders views, and close-ups).*

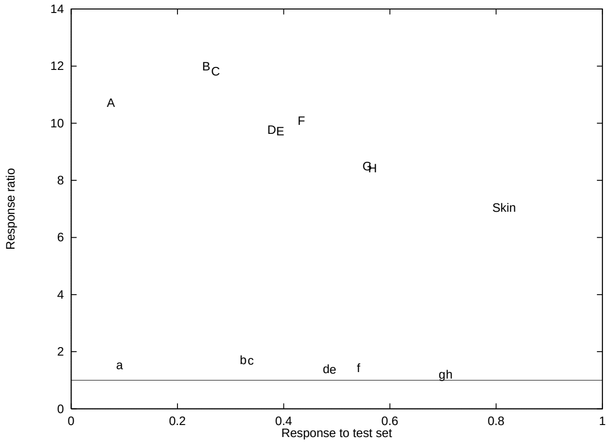

Figure 3: *The response ratio, (percent incoming test images marked/percent incoming control images marked), plotted against the percentage of test images marked, for various configurations of the naked people finder. Labels "A" through "H" indicate the performance of the entire system of skin filter 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. There is an important trend here; the response ratio increases, and the recall decreases, as the geometric complexity of the groups required to identify a person increases. This suggests (1) that the presence of a sufficiently complex geometric group is an excellent guided to the presence of an object (2) that our representation used in the present implementation omits a number of important geometric structures.* **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.*