# Finding Naked People

Margaret M. Fleck[1], David A. Forsyth[2], and Chris Bregler[2]

[1] Department of Computer Science, University of Iowa, Iowa City, IA 52242
[2] Computer Science Division, U.C. Berkeley, Berkeley, CA 94720

**Abstract.** This paper demonstrates a content-based retrieval strategy that can tell whether there are naked people present in an image. No manual intervention is required. The approach combines color and texture properties to obtain an effective mask for skin regions. The skin mask 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 geometric properties such as the structure of individual parts, and the relationships between parts, and constraints on color and texture. The system is demonstrated to have 60% precision and 52% recall on a test set of 138 uncontrolled images of naked people, mostly obtained from the internet, and 1401 assorted control images, drawn from a wide collection of sources. **Keywords:** Content-based Retrieval, Object Recognition, Computer Vision, Erotica/Pornography, Internet, Color

## 1 Introduction

The recent explosion in internet usage and multi-media computing has created a substantial demand for algorithms that perform *content-based retrieval*—selecting images from a large database based on what they depict. Identifying images depicting naked or scantily-dressed people is a natural content-based retrieval problem. These images frequently lack textual labels adequate to identify their content but can be effectively detected using simple visual cues (color, texture, simple shape features), of the type that the human visual system is known to use for fast (preattentive) triage [19]. There is little previous work on finding people in static images, though [9] shows that a stick-figure group can yield pose in 3D up to limited ambiguities; the work on motion sequences is well summarised in [4].

Several systems have recently been developed for retrieving images from large databases. The best-known such system is QBIC [15], 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. Searches employ an underlying abstraction of an image as a collection of colored, textured regions, which were manually segmented in advance, a significant disadvantage. Photobook [17] 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 ([17], p. 10) incorporates a simple notion of objects, using plane matching by an energy minimisation strategy. However, the approach does not adequately address the range of variation in object shape and appears to require manually segmented images for all but trivial cases. Appearance based matching is also used in [8], which describes a system that forms a wavelet based decomposition of an image and matches based on the coarse-scale appearance. Similarly, Chabot [16] uses a combination of visual appearance and text-based cues to retrieve images, but depends strongly on text cues to identify objects. However, appearance is not a satisfactory notion of content, as it is only loosely correlated with object identity.

Current object recognition systems represent models either as a collection of geometric measurements or as a collection of images of an object. This information is then compared with image information to obtain a match. Most current systems that use geometric models use invariants of an imaging transformation to index models in a model library, thereby producing a selection of recognition hypotheses. These hypotheses are combined as appropriate, and the result is back-projected into the image, and verified by inspecting relationships between the back-projected outline and image edges. An extensive bibliography of this approach appears in [12].

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 points and points on the model of some object is assumed known and an estimate of the appearance in the image of that object is constructed from correspondences. The hypothesis that the object is present is then verified using the estimate of appearance [20]. 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 [13]. Neither class of system copes well with models that have large numbers of internal degrees of freedom, nor do they incorporate appropriate theories of parts. Current part-based recognition systems are strongly oriented to recovering cross-sectional information, and do not treat the case where there are many parts with few or no individual distinguishing features[22].

Typical images of naked people found on the internet: have uncontrolled backgrounds; may depict multiple figures; often contain partial figures; are static images; and have been taken from a wide variety of camera angles, e.g. the figure may be oriented sideways or may viewed from above.

## 2 A new approach

Our system for detecting naked people illustrates a general approach to object recognition. The algorithm:

- first locates images containing large areas of skin-colored region;

– then, within these areas, finds elongated regions and groups them into possible human limbs and connected groups of limbs, using specialised groupers which incorporate substantial amounts of information about object structure.

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

### 2.1 Finding Skin

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) [18]. Therefore, human skin has a restricted range of hues. Skin is somewhat saturated, but not deeply saturated. Because more deeply colored skin is created by adding melanin, the range of possible hues shifts toward yellow as saturation increases. 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 colour planes omitting locations within 10 pixels of the image edges, to avoid potentially significant desaturation. The input $R$, $G$, and $B$ values are transformed into a log-opponent representation (cf e.g. [6]). If we let $L$ represent the log transformation, the three log-opponent values are $I = L(G)$, $R_g = L(R) - L(G)$, and $B_y = L(B) - \frac{L(G)+L(R)}{2}$. The green channel is used to represent intensity because the red and blue channels from some cameras have poor spatial resolution.

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. [3]

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. 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, output of the initial skin filter is refined to include adjacent regions with almost appropriate properties.

Specifically, the region marked as skin is enlarged to include pixels many of whose neighbors passed the initial filter. If the marked regions cover at least 30%

[3] All operations use a fast multi-ring approximation to the median filter [5].

of the image area, the image will be referred for geometric processing. Finally, to trim extraneous pixels, the algorithm unmarks any pixels which do not pass a more lenient version of the skin filter, which imposes no constraints on texture amplitude and uses less exacting constraints on hue and saturation.

## 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. 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.

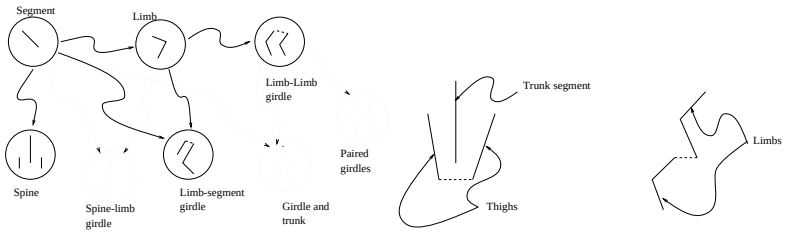


**Fig. 1.** Left: grouping rules (arrows) specify how to assemble simple groups (e.g. body segments) into complex groups (e.g. limb-segment girdles). These rules incorporate constraints on the relative positions of 2D features, induced by constraints on 3D body parts. Dashed lines indicate grouping rules that are not yet implemented. Middle: the grouper rejects this assembly of thighs and a spine (the dashed line represents the pelvis) because the thighs would occlude the trunk if a human were in this posture, making the trunk's symmetry impossible to detect. Right: this hip girdle will also be rejected. Limitations on hip joints prevent human legs from assuming positions which could project to such a configuration.

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 [3] 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 [1] are found by a straightforward algorithm. Sets of points forming regions with roughly straight axes ("ribbons" [2]) 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.

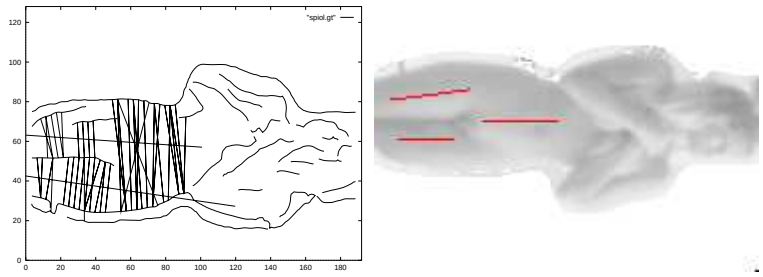


**Fig. 2.** Grouping a spine and two thighs: *Top left* the segment axes that will be grouped into a spine-thigh group, overlaid on the edges, showing the upper bounds on segment length and the their associated symmetries; *Top right* the spine and thigh group assembled from these segments, overlaid on the image.

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 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. There must be a line segment (the pelvis) between their ends, 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 kneecaps (see figure 1). 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 seg-

ment must have an average width similar to twice the average widths of the thigh segments. Finally, the whole configuration of trunk and thighs must satisfy geometric constraints depicted in figure 1. The grouper asserts that human figures are present if it can assemble either a spine-thigh group or a girdle group. Figure 2 illustrates the process of assembling a spine-thigh group.

## 4 Experimental Results

The performance of the system was tested using 138 target images of naked people and 1401 assorted control images, containing some images of people but none of naked people. Most images were taken with (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. Some depict several people, sometimes intertwined. 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.

Five types of control image were used

- 1241 images sampled[4] from an image database produced by California Department of Water Resources (DWR), 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 CD included with an issue of MacFormat [11],
- 11 assorted personal photos, re-photographed with our CCD camera, and
- 47 pictures of objects and textures taken in our laboratory for other purposes.

The DWR images are 128 by 192 pixels. The images from other sources were reduced to approximately the same size. Table 1 summarizes the performance of each stage of the system.

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. Some control images pass the skin filter because they contain people, particularly several close-up portrait shots. Other control images contain material whose color closely resembles that of human skin: particularly wood 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 substantial portion of the face was marked, suggesting that the approach provides a useful pre-filter for programs that mark faces. 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.

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

Figures 3-4 illustrate its performance on the test images. Configurations marked by the spine-thigh detector are typically spines. The girdle detector often marks 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. Figures 5-6 show typical performance on control images. 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.

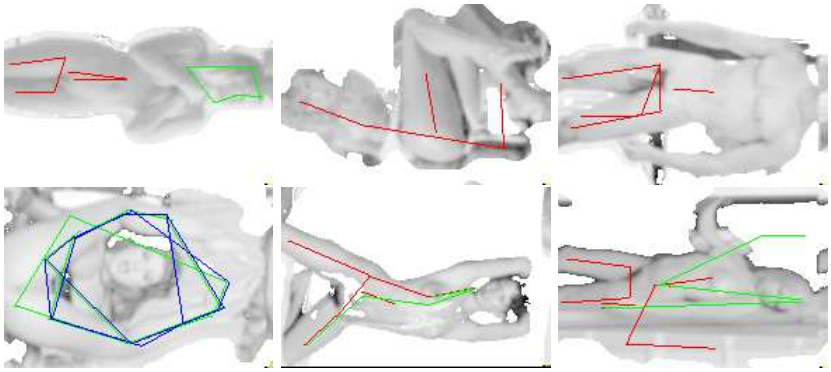

**Fig. 3.** Typical images correctly classified as containing naked people. The output of the skin filter is shown, with spines overlaid in red, limb-limb girdles overlaid in blue, and limb-segment girdles overlaid in blue.

| | eliminated by skin filter | eliminated by geometrical analysis | marked as containing naked people |
|---|---|---|---|
| test images | 13.8% (19) | 34.1% (47) | 52.2% (72) |
| control images | 92.6% (1297) | 4.0% (56) | 3.4% (48) |

**Table 1.** Overall classification performance of the system.

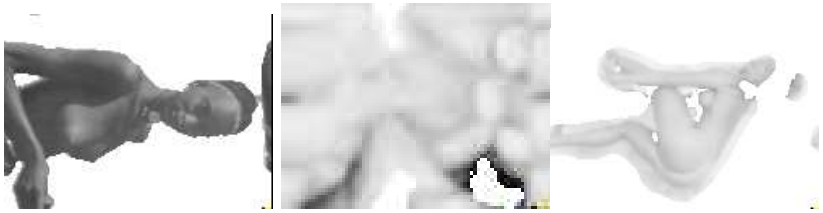

**Fig. 4.** Typical false negatives: the skin filter marked significant areas of skin, but the geometrical analysis could not find a girdle or a spine. Failure is often caused by absence of limbs, low contrast, or configurations not included in the geometrical model (notably side views).

**Fig. 5.** A collection of typical control images which were correctly classified as control images by our system. All contain at least 30% skin pixels, and so would be classified as containing naked people if the skin filter were used alone.

## 5 Discussion and Conclusions

From an extremely diverse set of test images, this system correctly identifies 52.2% as containing naked people. On an equally diverse and quite large set of control images, it returns only 3.4% of the images. In the terminology of content-based retrieval, the system is displaying 52% recall and 60% precision against a large control set[5]. Both skin filtering and geometric processing are required for

[5] *Recall* is the percentage of test images actually recovered;*precision* is the percentage of recovered material that is desired.

**Fig. 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 or girdles. The grouper is frequently confused by groups of parallel edges, as in the industrial image.

this level of performance: the skin filter by itself has better recall (86.2%), but returns twice as many false positives. This is an extremely impressive result for a high-level query ("find naked people") on a very large (1539 image) database with no manual intervention and almost no control over the content of the test and control images.

This system demonstrates detection of jointed objects of highly variable shape, in a diverse range of poses, seen from many different camera positions. It also demonstrates that color cues can be very effective in recognizing objects that whose color is not heavily saturated and whose surfaces display significant specular effects, under diverse lighting conditions, without relying on preprocessing to remove specularities. While the current implementation uses only very simple geometrical grouping rules, covering only poses with visible limbs, the performance of this stage could easily be improved. In particular: the ribbon detector should be made more robust; detectors should be added for non-ribbon features (for example, faces); grouping rules for structures other than spines and girdles should be added; grouping rules should be added for close-up views of the human body.

The reason we have achieved such good performance, and expect even better performance in the future, is that we use object models quite different from those commonly used in computer vision (though similar to proposals in [2, 14]). In the new system, 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. These grouping rules incorporate both surface properties (color and texture) and simple shape information. In the present system, the integration of different cues is simple (though effective), but a more sophisticated recognizer would integrate them more closely. 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. 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.

Finally, as this paper goes to press, a second experimental run using a substantially improved version of the grouper has displayed 44 % recall and an extraordinary 74 % precision on a set of 355 test images and 2782 control images from extremely diverse sources.

**Acknowledgements:** We thank Joe Mundy for suggesting that the response of a grouper may indicate the presence of an object and Jitendra Malik for helpful suggestions. This research was supported by the National Science Foundation under grants IRI-9209728, IRI-9420716, IRI-9501493, under a National Science Foundation Young Investigator award, an NSF Digital Library award IRI-9411334, and under CDA-9121985, an instrumentation award.

## References

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

2. Brooks, Rodney A. (1981) "Symbolic Reasoning among 3-D Models and 2-D Images," *Artificial Intelligence* 17, pp. 285–348.
3. Canny, John F. (1986) "A Computational Approach to Edge Detection," *IEEE Patt. Anal. Mach. Int.* 8/6, pp. 679–698.
4. Cedras C., Shah M., (1994) "A Survey of Motion Analysis from Moving Light Displays" *Computer Vision and Pattern Recognition* pp 214-221.
5. 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.
6. 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.
7. Gong, Yihong and Masao Sakauchi (1995) "Detection of Regions Matching Specified Chromatic Features," *Comp. Vis. Im. Underst.* 61/2, pp. 263–269.
8. Jacobs, C.E., Finkelstein, A., and Salesin, D.H., "Fast Multiresolution Image Querying," *Proc SIGGRAPH-95*, 277-285, 1995.
9. Lee, H.-J. and Chen, Z. "Determination of 3D human body postures from a single view," CVGIP, **30**, 148-168, 1985
10. Lowe, David G. (1987) "The Viewpoint Consistency Constraint," *Intern. J. of Comp. Vis,* 1/1, pp. 57–72.
11. MacFormat, issue no. 28 with CD-Rom, September, 1995.
12. Mundy, J.L. and Zisserman, A. *Geometric Invariance in Computer Vision*, MIT press, 1992
13. Murase, H. and Nayar, S.K., "Visual learning and recognition of 3D objects from appearance," to appear, *Int. J. Computer Vision*, 1995.
14. Nevatia, R. and Binford, T.O., "Description and recognition of curved objects," *Artificial Intelligence*, **8**, 77-98, 1977
15. 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.*
16. Ogle, Virginia E. and Michael Stonebraker (1995) "Chabot: Retrieval from a Relational Database of Images," *Computer* 28/9, pp. 40–48.
17. 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.
18. Rossotti, Hazel (1983) *Colour: Why the World isn't Grey*, Princeton University Press, Princeton, NJ.
19. Treisman, Anne (1985) "Preattentive Processing in Vision," *Com. Vis. Grap. Im. Proc.* 31/2, pp. 156–177.
20. Ullman, S. and Basri, R. (1991). Recognition by linear combination of models, *IEEE PAMI*, **13**, 10, 992-1007.
21. 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.*
22. Zerroug, M. and Nevatia, R. "From an intensity image to 3D segmented descriptions," ICPR, 1994.

This article was processed using the LaTeX macro package with ECCV'96 style