# USING COLOUR FEATURES TO BLOCK DUBIOUS IMAGES

*Yi Chan, Richard Harvey* and *J. Andrew Bangham*
School of Information Systems, University of East Anglia,
Norwich, NR4 7TJ, UK.
Tel: +44 1603 593257; Fax: +44 1603 593345
{yc,rwh,ab}@sys.uea.ac.uk

**ABSTRACT**

This paper describes a vision system that classifies web-images containing people. It works by identifying skin-coloured regions, extracting very simple features from these regions and making a classification decision. A two-stage skin filtering algorithm using likelihood matrices in HSV space followed by some local clustering works well. Our conclusion is that a simple approach using low-level features can work as well as much more complicated methods.

## 1 CONTEXT

Many enterprises are concerned to detect and restrict the circulation of material considered pornographic. The concerns are often practical and centre on the loss of communications bandwidth, disk storage and staff time as a result of unwanted material circulating in the organisation but also legal liability may arise because, if the content is pornographic or indecent then, in some countries, there is a possibility for prosecution. An exposition of these issues can be found elsewhere [1].

## 2 SKIN FILTERING

Algorithms to identify human skin form a common module in many computer vision systems ( [2–4] for example) and are usually based on colour. The objective is to choose a colour space in which the skin pixel (pel) cluster is as compact as possible and Table 1 lists several colour spaces have been proposed for this task.

| *Colour space* | *Components* |
|---|---|
| RGB | $r, g, b$ |
| HSV [5] | $h, s, v$ |
| Log opponent [6] | $I, R_g, B_y$ |
| Normalised RGB [7] | Two of $\bar{r}, \bar{g}, \bar{b}$ |
| Comprehensive [8] | Two of $\tilde{r}, \tilde{g}, \tilde{b}$ |

Table 1: Colour-space conventions. For the normalised RGB and the comprehensive normalisation intensity variation is removed so one colour component is a linear combination of the other two.

The HSV colour space [5] may be derived from the RGB space as

$$v = \max(r, g, b), \tag{1}$$

$$s = d/v, \tag{2}$$

$$h = \begin{cases} \frac{g-b}{6d} & r = v \\ \frac{2-r+b}{6d} & g = v \\ \frac{4-g+r}{6d} & b = v \end{cases} \tag{3}$$

where $d = \max(r, g, b) - \min(r, g, b)$. The log opponent space [4]

$$I = \log g, \tag{4}$$

$$R_g = \log r - \log g, \tag{5}$$

$$B_y = \log b - \frac{\log g + \log r}{2} \tag{6}$$

is an attempt to attempt to model the human vision system's opponent colour representation [9]. The contention is that at least one of the log-opponent channels is insensitive to melanin content [4].

Alternatives to three-channel spaces derive from colour constancy algorithms in which the aim is to remove variations in colour due to either illuminant angle or colour. We examine two: a simple normalised RGB space known as chromaticity space that is popular in skin filtering [7] which removes the effect of lighting geometry

$$\bar{r} = \frac{r}{r+g+b}, \bar{g} = \frac{g}{r+g+b}, \bar{b} = \frac{b}{r+g+b} \tag{7}$$

and also an iterative comprehensive scheme [8] that removes the effects of lighting geometry and illuminant colour. In the first stage

$$r' = \frac{r}{r+g+b}, g' = \frac{g}{r+g+b}, b' = \frac{b}{r+g+b} \tag{8}$$

and in the second stage

$$\tilde{r} = \frac{2r'}{\sum_{\text{all pels}} r'}, \tilde{g} = \frac{2g'}{\sum_{\text{all pels}} g'}, \tilde{b} = \frac{2b'}{\sum_{\text{all pels}} b'} \tag{9}$$

The algorithm iterates (8) and (9) until the maximum variation in $\tilde{r}$, $\tilde{g}$ or $\tilde{b}$ from one stage to the next is less

Figure 1: Left: original; Right: corresponding likelihood image normalised in the range 0 (black) to 1 (white).

than 1%. In practice this usually takes only a couple of iterations provided pixels that have near zero intensity are masked-out of the normalisation [1].

The pixels that are labelled as skin in the training set may be projected into each colour space to form a skin cluster which may itself be normalised via a conventional Mahalonobis projection (principal component analysis). If the column vector $\boldsymbol{e}_i$ is the $i$th eigenvector of the correlation matrix of the colour vector, $\boldsymbol{c}$. Then $i$th component of the normalised colour is

$$c_i = \frac{1}{\sqrt{\lambda_i}} \left(\boldsymbol{c} - \mathrm{E}\left\{\boldsymbol{c}\right\}\right)^T \boldsymbol{e}_i \tag{10}$$

where $\boldsymbol{c} = [r, g, b], [h, s, v]$, two of $[\bar{r}, \bar{g}, \bar{b}]$ or two of $[\tilde{r}, \tilde{g}, \tilde{b}]$. Thus the skin cluster is transformed to one centred on $\mathbf{0}$. Choosing all pixels that have a projection in this new space of length less than some value is a method of identifying skin pixels and it has been shown that once such a threshold is selected the choice of colour space is not critical [1]. This is a somewhat surprising conclusion given previous reports but images acquired from the web have often been heavily processed "by eye" to, for example, reduce the number of colours and to correct for skin tone and this might account for the difference. A further observation is that intereflections within images can often be significant enough to mask the advantages of colour constancy algorithms [10].

An alternative approach to modelling the data with an elliptically shaped Gaussian distribution ((10) amounts to this) is to compute the likelihood

$$L(\boldsymbol{c}|\mathrm{skin}) = \frac{\Pr\{\boldsymbol{c}|\mathrm{skin}\}}{\Pr\{\boldsymbol{c}|\mathrm{not\ skin}\}} \tag{11}$$

for a quantized colour space. Figure 1 shows the likelihood of pixel colours for an example image using a likelihood histogram with $25^3$ bins. Likelihood images such as the one shown on the right of Figure 1 may be used to produce segments that represent regions of skin by thresholding the likelihood image at the odds set by the ratio of the priors. However care is needed to avoid two common problems. The first is that an image may contain many isolated pixels that have the same colour as skin but are associated with the background (examples of such pixels can be seen to the left of the woman's head in the image on the right of Figure 1). The second problem is that for any particular image the likelihood distribution is not guaranteed to contain the mode of the training set likelihood distribution which can cause likelihood values to be unfeasibly low. In the image on the right of Figure 1 for example, the skin segments associated with the woman's torso do not appear to have high likelihood values and skin pixels can be missed. However a legitimate assumption is that skin regions are of reasonable area compared to the total image area and contain a locally maximum likelihood value. We therefore use a region-growing algorithm that uses as its seed points likelihood local maxima above a certain threshold. The regions are grown out to a lower likelihood threshold.

The region growing algorithm used here is based on a morphological scale-space process referred to as a *sieve* [11, 12]. The algorithm operates by identifying extremal regions in an image and "slicing-off" these extremal regions to the next most extreme value. The differences between successive stages are called *granules* and correspond closely to the region of support for sharp-edged objects. Since small granules are contained within large ones they form a tree [13]. The putative skin regions correspond to the largest area granules with an underlying likelihood above a lower likelihood threshold.

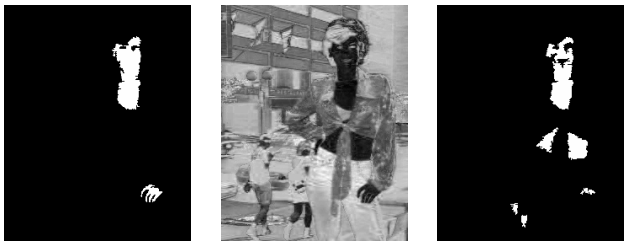

Figure 2: The left image shows the skin mask produced by thresholding the sieved likelihood image (shown on the right of Figure 1). This mask is then used to define a colour cluster which is then projected using (10). The Mahalanobis distance is shown in the centre image (black is zero indicating the pixel is similar to the mean skin colour). This distance is then thresholded to give (right) the final mask

These regions are then used to build a new local definition of skin and non-skin regions and hence a new segmentation may be computed. This second step improves the results for images containing skin types that are under-represented in the training set. In the final

[1] The masking level is usually set by eye so here we present the best results obtained over varying the masking level from 0 to 0.3 in steps of 0.05.

operation each skin segment is forced to have zero Euler number by flood filling any interior regions. Figure2 illustrates this sequence of operations.

This likelihood segmentation approach has been tested using a second database consisting of 950 training images and 950 test images manually segmented to provide the ground truth. The performance may be summarised through two-class confusion matrices:

$$\boldsymbol{P} = \begin{bmatrix} p(\bar{s}|\bar{s}) & p(s|\bar{s}) \\ p(\bar{s}|s) & p(s|s) \end{bmatrix} = \begin{bmatrix} 0.82 & 0.17 \\ 0.18 & 0.83 \end{bmatrix} \quad (12)$$

where, for example, $p(\bar{s}|s)$ denotes the probability that a pixel from a skin region is classified as one from a non-skin region. Equation (12) gives a typical result for a $25^3$ bin system operating in HSV space which is our current preferred compromise between performance and storage. We use this system for the remainder of the paper.

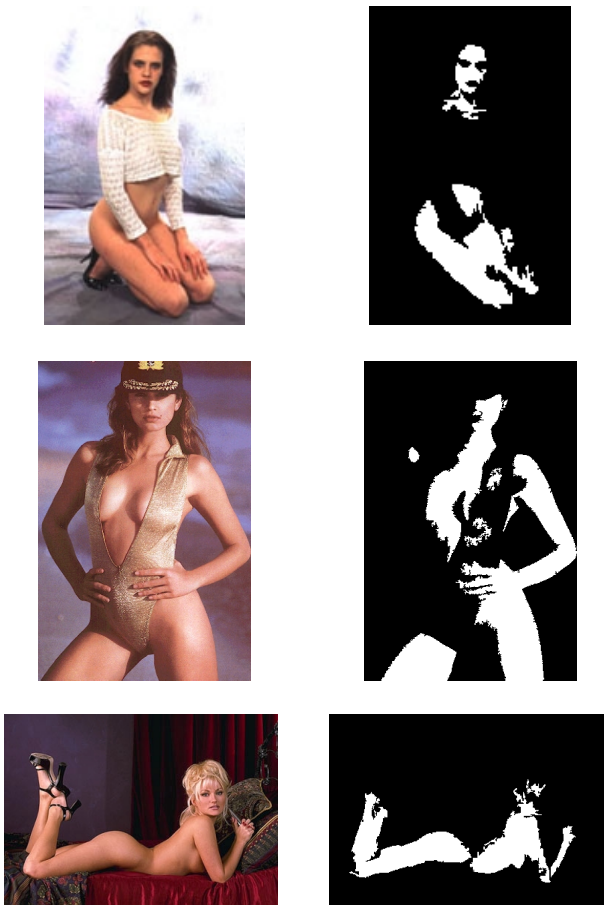

Figure 3: Example segmentations

Figure 3 shows some example segmentations for images in the test set. As expected the high resolution images (second from left on top and bottom of Figure 3) give qualitatively better results than the low resolution images but, provided the test images contain skin colours that were in the training set the automatic segmentations are close to those obtained manually. Having identified areas of skin it is necessary to extract higher level features on which to distinguish the classes of image.

## 3 CLASSIFICATION

The data have been hand-classified into five categories: pornography (nude pictures that show genitalia or sexual acts); nude; people (showing people in all poses not covered in other categories showing people); portraiture (which is restricted to head and shoulders portraits of a type prevalent on the web); miscellaneous and graphics (containing computer generated web graphics, buttons and so on). There are suggestions for high-level features based on grouping of skin segments [4] that might distinguish these classes but here we have a requirement to process the images speedily so are interested to try simpler features. For each blob in the image we have computed: area; $x$-centroid; $y$-centroid; the length of the major axis of an ellipse with the same second-order moments as the blob; the length of the minor axis of the same ellipse; the eccentricity of the ellipse; the orientation of the ellipse; the area of a convex hull fitted to the blob; the diameter of a circle with the same area is the blob; the Solidity (the proportion of the convex hull area accounted for by the blob) and the Extent (the proportion of the area of a rectangular bounding box accounted for by the blob). These features are ranked using the mutual information of the class given the single feature. Doing this gives the subset of features that we use: the area of the largest blob; its centroid co-ordinates and the major and minor length of the fitted ellipse and its orientation. The performance of these features is evaluated using a conventional $k$-nearest neighbour classifier with $k = 1, 3$ or 5 implemented, for speed, via the Multiedit and Condense algorithm [14]. A typical confusion matrix ($k = 1$) is

$$\boldsymbol{P} = \begin{bmatrix} 0.73 & 0.12 & 0.01 & 0.02 & 0.11 & 0.01 \\ 0.50 & 0.12 & 0.02 & 0.08 & 0.26 & 0.02 \\ 0.03 & 0.08 & 0.15 & 0.05 & 0.51 & 0.18 \\ 0.28 & 0.11 & 0.08 & 0.13 & 0.34 & 0.06 \\ 0.20 & 0.07 & 0.05 & 0.13 & 0.46 & 0.09 \\ 0.01 & 0.01 & 0.06 & 0.03 & 0.09 & 0.80 \end{bmatrix} \quad (13)$$

where $\boldsymbol{P}$ has elements $[\boldsymbol{P}]_{ij}$ which is the probability of classifying as class $j$ given that the image is in class $i$. The results for $k > 1$ are slightly worse but do not differ much from those above. There are many non-serious confusions (nude pictures (class 2) are often classified as pornography (class 1)) but also a few serious ones. Portraits (class 4) are often classified as pornography. Although this is worrying, a solution may exist by using a conventional face finder.

Defining two meta-classes as "unwanted" (porn (1) and nude (2)) which we denote $u$ and safe (all other

classes), $s$, gives the two-class confusion matrix

$$\boldsymbol{P} = \begin{bmatrix} p(u|u) & p(s|u) \\ p(u|s) & p(s|s) \end{bmatrix} = \begin{bmatrix} 0.81 & 0.19 \\ 0.23 & 0.77 \end{bmatrix} \quad (14)$$

## 4 CONCLUSIONS AND FURTHER WORK

This paper has provided evidence of a successful skin segmentation algorithm and suggested how this might form part of an automated pornography detector with a performance that compares favourably to more complex alternatives [4].

We are currently addressing the extension to different skin types – the current system has some robustness to melanin content but not enough – and the investigation of special detectors for known failure modes. Face detectors are an obvious example and we have implemented a number of systems using commercial face finders.

A further refinement is to consider the prior class probabilities. Currently the test and training sets do not have equal numbers of data in each class. This is the correct training strategy [15] provided that the training set contains data that occur at rates that are representative of real situations. These data were collected during real browsing sessions but it is probable that there may be significant variations in priors between users. We are currently investigating this.

A final observation is that even quite simple use of image side-data can further improve the performance. An example is the coding method: web designers tend to use jpg and gif codings for rather different types of image. This is illustrated in equations 15 and 16 which give the confusion matrices when the features are augmented with a feature that is the length of the image colourmap.

$$C = \begin{bmatrix} 0.84 & 0.04 & 0.00 & 0.01 & 0.11 & 0.00 \\ 0.65 & 0.04 & 0.00 & 0.09 & 0.22 & 0.00 \\ 0.41 & 0.03 & 0.00 & 0.08 & 0.49 & 0.00 \\ 0.31 & 0.07 & 0.01 & 0.18 & 0.41 & 0.01 \\ 0.36 & 0.03 & 0.01 & 0.10 & 0.50 & 0.00 \\ 0.00 & 0.00 & 0.00 & 0.00 & 0.00 & 1.00 \end{bmatrix} \quad (15)$$

and

$$M = \begin{bmatrix} 0.85 & 0.15 \\ 0.31 & 0.69 \end{bmatrix} \quad (16)$$

The number-of-colours feature reduces the false alarm rate on the graphics class to zero which gives error reductions for the other classes.

## References


[1] Y. Chan, R. Harvey, and D. Smith, "Building systems to block pornography," in *2nd UK Conference on Image Retrieval: The Challenge of Image Retrieval (CIR'99)* (J. Eakins and D. Harper, eds.), pp. 34–40, BCS Electronic Workshops in Computing series, Feb 1999.

[2] A. P. Pentland, "Smart rooms: machine understanding of human behavior," in *Computer vision for human-machine interaction* (R. Cipolla and A. Pentland, eds.), pp. 3–21, Cambridge University Press, 1998.

[3] K.C.Yow and R.Cipolla, "Feature-based human face detection," *Image and Vision Computing*, vol. 15, no. 9, pp. 713–735, 1997.

[4] M. M. Fleck, D. A. Forsyth, and C. Bregler, "Finding naked people," in *European Conference on Computer Vision*, vol. II, pp. 593–602, Springer-Verlag, 1996.

[5] J. D. Foley, A. van Dam, S. K. Feiner, and J. F. Hughes, *Fundamentals of interactive computer graphics*. Addison-Wesley, 2 ed., 1994.

[6] R. Gershon, A. D. Jepson, and J. K. Tsotos, "Ambient illumination and the determination of material changes," *Journal of the optical society of America A – optics image science and vision*, vol. 3, no. 10, pp. 1700–1707, 1986.

[7] B. Schiele and A. Waibel, "Gaze based tracking based on face-color," in *International workshop on automatic face- and gesture-recognition*, June 1995.

[8] G. D. Finlayson, B. Schiele, and J. L. Crowley, "Comprehensize colour normalisation algorithm," in *European Conference on Computer Vision*, pp. pp 475–490, 1998.

[9] A. B. Watson, ed., *Digital images and human vision*. Bradford, MIT Press, 1993.

[10] G. Finlayson and G. Tian, "Color normalization for color object recognition," *International Journal on Pattern Recognition and Artificial Intelligence*, pp. 1271–1285, 1999.

[11] J. Bangham, P. Ling, and R. Harvey, "Nonlinear scale-space causality preserving filters," *IEEE Trans. Patt. Anal. Mach. Intelli*, vol. 18, pp. 520–528, 1996.

[12] J.A.Bangham, R.Harvey, and P.D.Ling, "Morphological scale-space preserving transforms in many dimensions," *J. Electronic Imaging*, vol. 5, pp. 283–299, July 1996.

[13] J.A.Bangham, J.R.Hidalgo, G.C.Cawley, and R.W.Harvey, "Analysing images via scale-trees," in *British Machine Vision Conference*, 1998.

[14] P.A.Devijver and J.Kittler, *Pattern recognition: a statistical approach*. Prentice Hall, 1982.

[15] E. Davies, "Training sets and a priori probabilities with the nearest neighbour method of pattern recognition," *Pattern Recognition Letters*, vol. 8, pp. 11–13, July 1998.