Building systems to block pornography

Yi Chan, Richard Harvey, and Dan Smith
School of Information Systems, University of East Anglia,
Norwich, NR4 7TJ, UK.
Email: {yc,rwh,djs}@uea.ac.uk

Abstract

Experience and recent lawsuits have led large internet users to search for ways to filter email and web traffic by content. This paper reviews the prospects for this research for specific domain: pornography. We present results on a particularly challenging image-only database. We use an approach that relies on training – we hand-segment skin regions in the training set and use these to compute the likelihood that a particular colour is associated with skin. Pixels that are identified as skin are grouped together to form blobs and simple features extracted from these blobs are used to train a nearest neighbour classifier. Comparisons with hand-labelled data show that the skin detection algorithms are almost as good as a human operator but the overall performance on the pornography detection problem falls far short of that of a human. We therefore describe how these image-only methods might be fused with text-based analysis to produce a composite system with superior performance to any single media approach.

1 Introduction

This paper is about a problem that is perceived to be of increasing importance: the automatic detection of multimedia documents, in particular web pages, that contain pornography. Recent court rulings (see [1] for example) combined with surveys of internet traffic have led to internet users, particularly large corporations, to search for solutions to the automatic detection of pornography. Two serious concerns are expressed: denial of service and *cyber-liability*.

Denial of service arises when legitimate traffic is blocked or lost due to large quantities of illegitimate traffic. Liability may arise because email is treated as "tantamount to sending written correspondence" [2]. Furthermore if the content is pornographic or indecent then, in the UK, there is a possibility for prosecution under at least seven acts of parliament (see [2] for a review.). In the United States these concerns have coalesced into new legislation (the Communications Decency Act (CDA) 1996). Whether these concerns are justified and whether the CDA is defensible has been the subject of speculation and controversy (see [3, 4, 5] for some examples). These moral and political questions are outside the scope of this paper. However we note that pornography is a concept that is not susceptible to a precise definition that might be operationalised to distinguish between images which are acceptable and those which are not. In this paper we use the term "pornography" loosely, to refer to unwanted material with a sexual content.

Our approach to identifying this material is to:

1. identify images that contain large areas of skin and have the features discussed in Section 2;
2. analyse the associated text;
3. apply a weighting scheme to exclude unwanted material.

We hope that this approach will enable us to identify sexually explicit documents (e.g. those depicting human genitalia and designed to be erotic) with a high degree of precision, and to distinguish them from documents that, for example, promote the sale of underwear or swimwear or that are used for medical education. However, there remains a "grey area" of images and web sites of uncertain purpose. Typically these sites concentrate on subjects such as images of women in bikinis, but which do not sell swimwear. We believe that some organisations may tolerate the free circulation of such images, but that others may wish to block them. Here we outline an approach designed to support a wide range of organisational policies on the circulation of images containing large areas of skin and sexually related vocabulary.

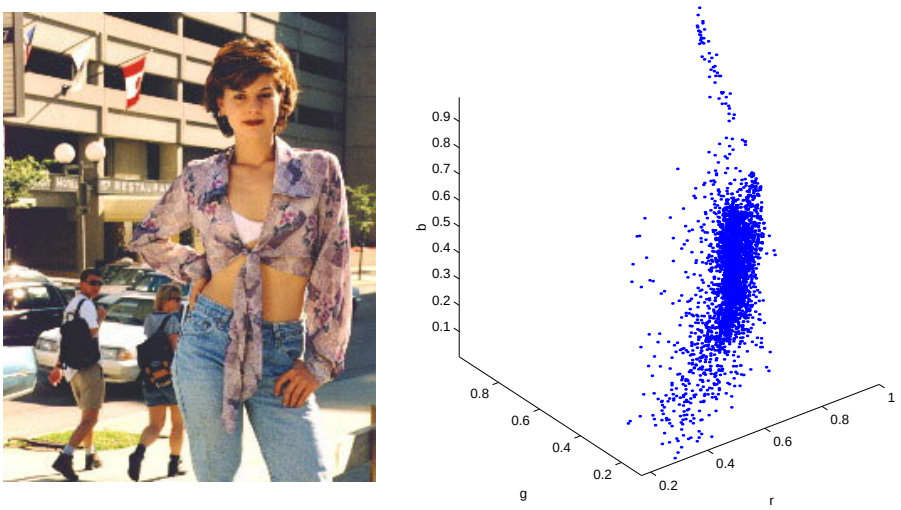

Figure 1: Left: original skin with skin region highlighted; Right: RGB plot of skin pixels

2 Image-based analysis

Algorithms to identify skin form a common module in many computer vision systems ([6, 7, 8] for example). This Section compares these algorithms and illustrates some possible high-level features that might be useful for classifying images containing people.

2.1 Skin filtering

A number of competing approaches have been proposed for the identification of pixels that are skin coloured. The problem is illustrated in Figure 1 which shows, on the left, an image that contains some skin regions and on the right a random sample of pixels taken from these regions and plotted in R, G, B space. From Figure 1 it is evident that, for this image. the skin pixels fall into a reasonably well defined banana-shaped region running from black to white through pink. The curved banana shape is characteristic of images found on the web and is due to, often poor, attempts at gamma correction.

Two questions present themselves. Firstly, is it possible to learn the distribution of skin and non-skin pixels in colour space? Secondly, which colour space is most appropriate for this problem? We provide preliminary answers to these questions by constructing sets of training and test images and measuring skin detection performance. The test and training data consisted of 140 images acquired from the web. Half the images contained clothed people and half contained naked people. The smallest image measured 80 by 35 pixels and the largest 810 by 542 pixels and the images were in a variety of compressed and uncompressed formats. Images were randomly selected from the two sub-sets to form two training sets each containing 35 images. For each training image the skin regions were segmented manually. Figure 2 shows some example clothed training images.

Our first objective was to choose a colour space in which the skin region was as compact as possible. Each pixel, r, g, b in the training set is transformed to one of the colour spaces shown in Table 1.

The HSV colour space [10] may be derived from the RGB space as

$$v = \max(r, g, b), \qquad s = d/v, \qquad h = \begin{cases} \frac{g-b}{6d} & r = v \\ \frac{2-r+b}{6d} & g = v \\ \frac{4-g+r}{6d} & b = v \end{cases} \tag{1}$$

Figure 2: Example clothed images from the training set

Colour space	*Components*
RGB	r, g, b
HSV	h, s, v
Normalised RGB	Two of $\bar{r}, \bar{g}, \bar{b}$
Log opponent [8]	I, R_g, B_y
Comprehensive [9]	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.

where $d = \max(r, g, b) - \min(r, g, b)$. The log opponent space [8]

$$I = \log(g), \qquad R_g = \log(r) - \log(g), \qquad B_y = \log(b) - \frac{\log(g) + \log(r)}{2} \tag{2}$$

is an attempt to attempt to model the human vision system's opponent colour representation [11] – the contention is that at least one of the log-opponent channels is insensitive to melanin content.

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. Here we examine only two algorithms: first, a simple normalised RGB space, popular in skin filtering [12], which removes the effect of lighting geometry

$$r_n = \frac{r}{r+g+b}, \qquad g_n = \frac{g}{r+g+b}, \qquad b_n = \frac{b}{r+g+b} \tag{3}$$

and, second, a new iterative comprehensive scheme [9] that removes the effects of lighting geometry and illuminant colour. In the first stage

$$r' = \frac{r}{r+g+b}, \qquad g' = \frac{g}{r+g+b}, \qquad b' = \frac{b}{r+g+b} \tag{4}$$

and in the second stage

$$\tilde{r} = \frac{2r'}{\sum_{\text{all pixels}} r'}, \qquad \tilde{g} = \frac{2g'}{\sum_{\text{all pixels}} g'}, \qquad \tilde{b} = \frac{2b'}{\sum_{\text{all pixels}} b'} \tag{5}$$

Here the algorithm iterates 4 and 5 until the maximum variation in $\tilde{r}$, $\tilde{g}$ or $\tilde{b}$ from one stage to the next is less than 1% (usually a couple of iterations).

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 clustering (principal component analysis). If the column vector $\boldsymbol{e}_i$ is the ith eigenvector of the correlation matrix of the colour vector, $\boldsymbol{c}$. Then ith component of the normalised colour is

$$c_i = \frac{1}{\sqrt{\lambda_i}} \left(\boldsymbol{c} - \mathrm{E}\left\{\mathrm{c}\right\}\right)^T \boldsymbol{e}_i \tag{6}$$

Figure 3 shows an example of this transformation applied to the image and RGB representation shown in Figure 1. The skin cluster on the right of Figure 1 is, in Figure 3, transformed to one centred on **0**. Choosing all pixels that have

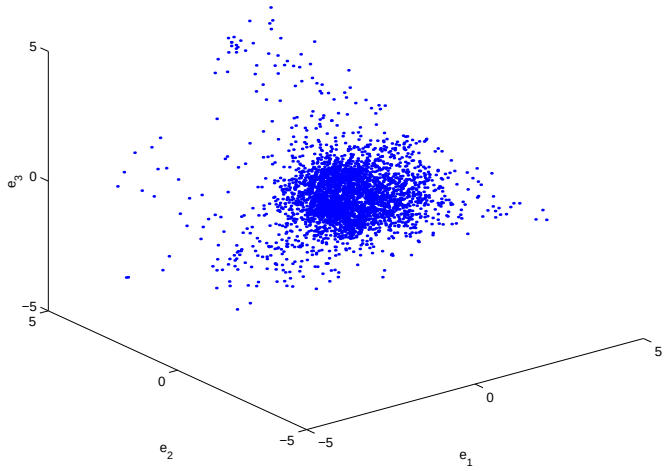

Figure 3: Skin pixels in Figure 1 after re-projection along the principal axes

a projection in this new space of length less than some value is a reasonably principled method of identifying skin pixels. Table 2 summarises the two-class (skin/not skin) recognition results on the previously unseen test data for a variety of colour spaces with $R_0 = 0.1$ and $R_0 = 0.9$ which corresponds to the radii necessary to select 10% and 90% of the pixels labelled as skin in the training set.

		$p(s\|s)$	$p(\bar{s}\|\bar{s})$	$p(s\|\bar{s})$	$p(\bar{s}\|s)$
$R_0 = 0.1$	RGB	0.12	0.99	0.01	0.88
	Log opponent	0.13	0.99	0.01	0.87
	HSV	0.14	0.99	0.01	0.86
	Normalised RGB	0.13	0.98	0.02	0.87
	Comprehensive	0.12	0.97	0.03	0.88
$R_0 = 0.9$	RGB	0.89	0.56	0.44	0.11
	Log opponent	0.91	0.60	0.40	0.09
	HSV	0.92	0.69	0.31	0.08
	Normalised RGB	0.90	0.36	0.64	0.10
	Comprehensive	0.92	0.23	0.78	0.08

Table 2: Elements of two-class confusion matrices for a variety of colour spaces and threshold radii of 0.1 and 0.9 (also compared, but not shown here were $R_0 = 0.3, 0.5, 0.7$).

Table 2 illustrates that once a threshold is selected the choice of colour space is not critical. We note however that other authors recommend the use of colour constancy or log-opponent spaces, so the issue requires further study.

A potential disadvantage of the re-projection approach is that it encourages implementations that require three real multiplies and two real additions per pixel. We adopt an alternative approach that encourages the use of table lookup which is computationally attractive. Using the training data it is simple to construct the likelihood score $L(\boldsymbol{c}|\text{skin}) = \Pr\{\boldsymbol{c}|\text{skin}\}/\Pr\{\boldsymbol{c}|\text{not skin}\}$ for a binned colour space. Figure 4 shows the colours associated with these likelihood histograms. Figure 5 shows the result of computing the likelihood in an RGB colour space quantized into 26 bins.

The likelihood image may be used to produce segments that represent regions of skin but care is needed to avoid

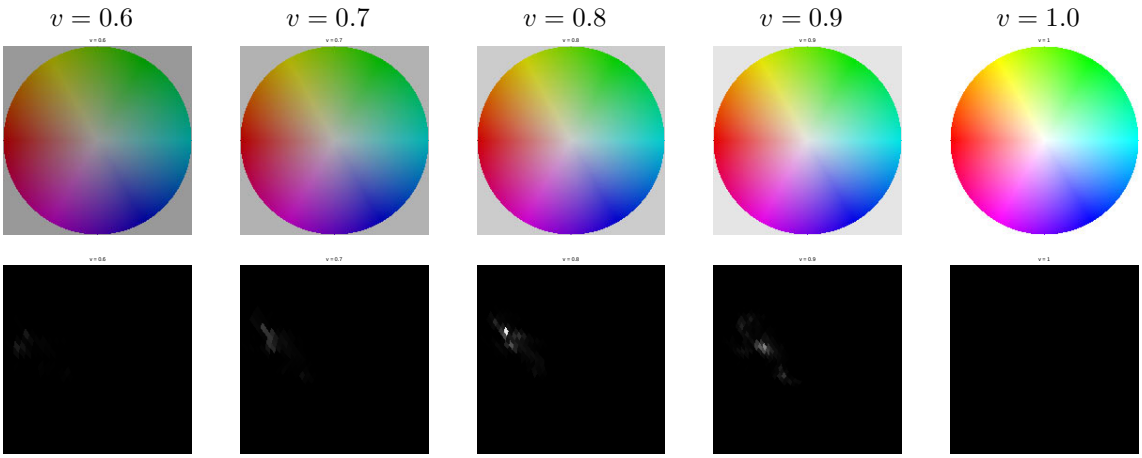

Figure 4: Colour wheels showing hue (angularly) and saturation (radially) for varying value. Underneath are shown the corresponding likelihood values normalised so that maximum likelihood over all colours is 1 (white) and the minimum 0 (black).

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 on the right of Figure 5). The second problem is that the per image likelihood distribution is not guaranteed to contain the mode of the training set likelihood distribution. This means that the pixels with the highest likelihood in a test image may not identify all of the likely skin segments. A solution to these problems is to use a region-growing algorithm that uses as its seed points local likelihood maxima above a certain threshold

Here we use a segmentation algorithm based on a new morphological processor called the *sieve* [13, 14]. 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 [15]. Here the likelihood images are thresholded at some lower fraction of the peak likelihood and connected sets above some fractional area are then inspected to see if they contain a high likelihood value – if so, they are retained. These large regions are then used to build a new local definition of skin and non-skin regions and hence a new localised likelihood is computed. This is then used to re-segment the image. Each segment in the final image is forced to have 0 Euler number by flood filling any interior regions.

Figure 6 shows some example segmentations for clothed images in the test set. Qualitatively the results are acceptable and, for skin colours in the training set, the algorithms produces segmentation that are close to manual ones.

2.2 Towards high-level features

We have tested three simple features: (i) the ratio of skin area to image area; (ii) the ratio of the area of the largest skin segment to the image area; (iii) the number of segments in the image. The classifier chosen was the well known k-nearest neighbour classifier with $k = 1, 3, 5, 7,$ or 9 [16]. We estimate the effectiveness of each system by computing the fraction of images identified correctly (a meaningful figure if pornographic and non-pornographic images are equiprobable). Using the automatic classifier and the test and training data described in Section 2.1 the best result is around 55% correct. This is worryingly close to chance and surprising when one considers that all but one of the automatic segmentations on the training and test data are acceptable. When the same features are generated from hand-segmented data the best result is around 65%. Several observations follow:

1. For this database where every image contains a person, many in swimsuits, is is necessary to not only to have a reliable skin segmentation algorithm but also to extract useful features from these regions. Furthermore if good

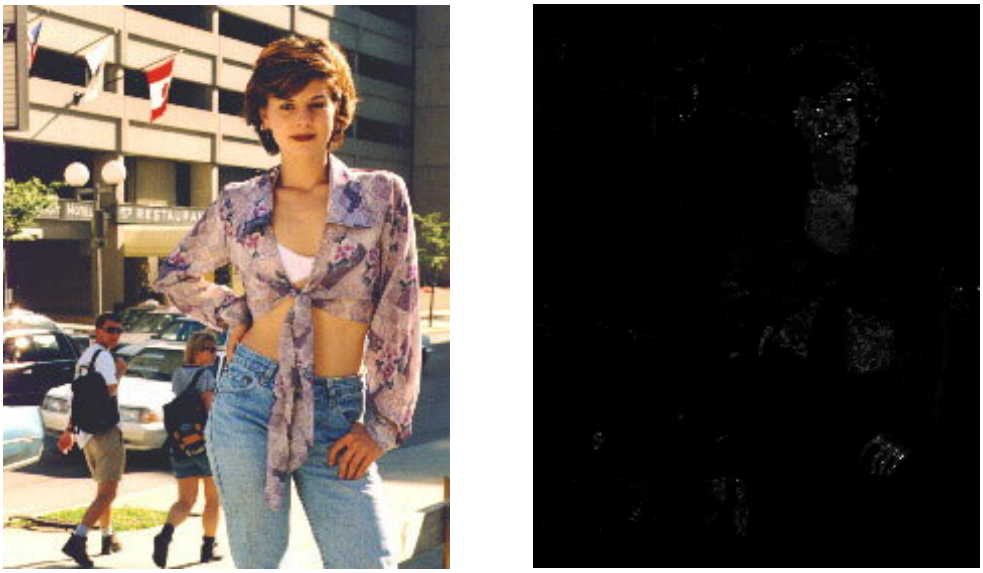

Figure 5: Likelihood image corresponding to an image not in the training set

Figure 6: Example segmentations for the images in Figure 2

features do not emerge it may become essential to incorporate additional modalities, such as text, to classify pornographic documents.

2. The automatic skin region extractor is nearly as good as a manual segmentation.

3. Since the skin detector is reliable it would be easy to manipulate the performance figures by adding images that contain no people to the non-pornographic test set.

This last point is important, not only for the unscrupulous, since it suggests that the construction of standardized test sets with classes that are meaningful to potential users will be essential. Unfortunately, as will be seen later, our initial discussions with users have implied user-dependent class priors and costs. Our solution to this problem has been to collect a larger database consisting of 1400 images with the images classified into classes: fashion; pornography; nudity; logos; portraits and miscellaneous. Each image has been annotated by hand to identify skin regions so we hope to be able to collect meaningful statistics for a variety of user policies. There is of course an interest in the development of standardised task for visual information retrieval but whether these are relevant to this problem is an open question.

3 Text-based approaches

The evidence from the image analysis implies that to substantially improve image-only performance we must develop new features and take into account other sources of information. A number of approaches to identifying pornographic or obscene web pages using text alone have also been attempted, with generally poor results. Examples include:

- a search for strings such as "sex" – which fails to distinguish sex education or zoology from pornography;
- a search for obscenities, as whole or part words – fails by treating "Scunthorpe then added a fourth ..." as an obscene, rather than a soccer reference.

Some current commercial pornography filters have incorporated simple text searching to augment image-based filtering strategies.

We have started by considering text strings associated with images that may be pornographic. There are three principal classes of relevant text: information supplied in <META> tags and titles, descriptive commentary, and disclaimers.

Many pornographic web sites use words with strong sexual connotations in <META> tags to attract visitors; substantial numbers of other sites do the same, for the same reasons, although their motive is to boost their visitor count and therefore appear more important than they would otherwise be. Any filtering based on the contents of <META> tags or other meta-data supplied voluntarily by the site owners is liable to spoofing and other misrepresentations of a page's content – if it is supplied at all. However, at present the contents of <META> tags are a useful aid to identifying pornographic pages through single whole or hyphenated words.

Pornographic web pages generally have very little associated text on the page, but certain words and phrases are indicative of pornographic content. An initial search of a number of sites suggests that terms such as "Miss <month>", "image of <name>", "interracial pix" and similar phrases are generally indicative of pornographic content.

Disclaimers of the general form "You must be at least 18 years old or the legal age in your area to view this adult material" are common on pornographic web sites, but rare elsewhere, so form another text indicator of likely content.

In order to achieve a good precision in the classification we must distinguish between pornography and sex education, medical or fashion pages. It is difficult to distinguish images advertising swimwear or underwear from pin-up images using image processing techniques alone. Here, the associated text is different in the terminology and phrasing used to describe the images. It is an open question whether the presence of terms, either single words or short phrases, is sufficient to distinguish these types of site, or whether some stylistic analysis is necessary.

Sex education and medical sites have a much higher ratio of text to images than either fashion or pornography. The language used is also substantially different, although many of the terms used in popular sex education pages, such as Dr Ruth or Cosmopolitan, may not be.

3.1 Experimental work

To test our hypothesis that a weighted combination of text terms and skin detection is substantially more effective than either approach alone, we performed three small experiments in October 1998. In the first, we analysed the first 50 URLs returned by AltaVista [17] from the simple search "Miss April",which had 707 hits; the first 50 URLs from a similar search on "Miss September", which had 419 hits were also analysed and the results are shown in Table 3

porn content	10
unavailable	9
titillation	5
unrelated	76

Table 3: Content of first fifty hits from searches on "Miss April" and "Miss September"

The unrelated sites in these searches were dominated by American military history (42 references), as AltaVista strips out punctuation in its indexing, giving many hits relating to events in Mississippi. Sites devoted to college football (4) pets (3), fashion (2), marriage bureaux (2) and a variety of other topics also appeared. The category

	Much skin	Some skin	No skin	Totals
Many explicit terms	37.50	6.25	21.88	65.63
Few explicit terms	3.13	0	0	3.13
No text	18.75	0	12.5	31.25
Totals	59.38	6.25	34.28	100

Table 4: Percentages of pornographic references with skin and explicit language.

"titillation" poses problems and covers a wide spectrum of material. These pages are mostly of the "model in bikini on a beach" type of image, with no explicit sexual references, in poses that are not generally regarded as provocative in Western Europe or North America. As such they illustrate the imprecision of definitions of pornography. Nine of the pornographic references had images with large amounts of skin and five also had numerous sex words; the other four had few sex words. Thus, if we wished to implement a policy that rejected pages in our pornographic class but accepted those in our titillation class (which shares many characteristics with fashion web pages) we would reject over half by filtering on skin alone; conversely filtering on text alone would let several pornographic pages through.

Our third experiment was an AltaVista search on the term "nude", which registered 5.7 million hits. We examined the first 50 references, of which 70% were pornographic. A breakdown of the content of these pages, in terms of images with large amounts of pink skin, and in the number of sexually related words is given in Table 4.

Factors that complicate the analysis of this table are that several of the references have text in languages other than English, the four that have no text or skin are all gateway pages to pornographic sites, but are not themselves offensive. If these are removed from the analysis, the combination of recognising sexual terms and skin detection would filter at least 78% of these pages. The proportion would be higher if we included hidden text. Of the sites analysed, there was only one false positive (a calendar page).

From these experiments it is clear that this multimedia approach is capable of substantially higher precision than current image or text based approaches.

3.2 Technical approach

The fundamental problem is a classification of web pages into pornographic or non-pornographic. Our technical approach is to search for all the elements we have identified as possible indicators of pornographic content and, using suitable weights, to reject material which is probably pornographic.

Document classification has been extensively researched [18]. The problem considered here is the inverse of the normal one, in that we are trying to classify and deselect, or block, the target document set.

The limitations of document classification based on single words have been well documented [19]. Information extraction [20] relies on context to restrict the range of meanings a text fragment may have and to restrict the number of fragments of interest. Since we already have a restricted domain – text associated with pictures containing large areas of skin – the approach is well suited to this problem. The small text volumes and restricted domain also facilitate a whole-text approach, which facilitates high precision classification [21]. An informal analysis of a sample of pornographic, fashion and sex education sites suggests that word clustering techniques [22] perform well.

The normal approach to document classification problems is to create an annotated corpus and then to devise a weighting scheme that correctly classifies the set, then to use this scheme to classify the unseen examples. The approach we are adopting is to derive a general weighting scheme from a large training set and then to systematically modify the weights to include or exclude certain classes of images at several levels of precision, using modified versions of the InfoExtractor and InfoDistributor prototypes described in [23, 24]. Automatic classification techniques have been extensively researched (e.g. [25, 26, 18, 26]). We believe that an automatic approach provides the best method for devising a general weighting scheme, but that the requirements of different organisations vary sufficiently that we must be able to customize the weighting scheme to reflect organisational policies and requirements – in short different organisations have different priors so it is not possible to produce a minimum risk classifier without knowledge of the organisation's policy. Additionally this allows us to produce customised weighting schemes without the need to collect large volumes of data to train the software for each new set of detailed requirements. The rationale

for this is analogous to that for libraries of reusable components in other middleware projects [27].

4 Conclusions

This paper has provided evidence of a successful skin segmentation algorithm and suggested how this might form part of an automated pornography detector. We are currently addressing the following areas:

- The extension to different skin types. The current system has some robustness to melanin content but not enough – we are extending our training set to incorporate more skin types.
- The refinement and extension of the training and test databases. We have collected a much larger database of over one thousand images. The images are classified into some additional categories such as "fashion" and "logos" which, as indicate previously, may have special significance for some users.
- We are developing new features including a face finder – we believe this we be useful in avoiding some of the false positives we see with, for example, full-face portraits.
- Integrating our existing methods to allow multimedia recognition.

Figure 7: Left: original image from an "innocent" source. Right: associated skin mask.

However even with these additions it is certain that there will be images containing a large amount of skin, such as that shown in Figure 7, that appear in a texturally ambiguous context, and yet are actually from an well-known young women's magazine.

References

[1] Ian Traynor. Child porn verdict stuns net lawyers. *The Guardian*, May 1998. 29th May.

[2] Graham J H Smith, editor. *Internet Law and Regulation*. FT Tax & Law, 21–27 Lamb's Conduit St. , London, UK., 2 edition, 1997.

[3] Marty Rimm. Marketing pornography on the information superhighway. *Georgetown Law Journal*, 83(5):1849–1934, 1985.

[4] Phillip Elmer-DeWitt. Cyberporn. *Time magazine*, 146(1):1–10, June 1995.

[5] Wendy M. Grossman. *net.wars*. NYU Press, 1997.

[6] Alex P.Pentland. Smart rooms: machine understanding of human behavior. In Roberto Cipolla and Alex Pentlan, editors, *Computer vision for human-machine interaction*, pages 3–21. Cambridge University Press, 1998.

[7] K.C.Yow and R.Cipolla. Feature-based human face detection. *Image and Vision Computing*, 15(9):713–735, 1997.

[8] Margaret M. Fleck, David A. Forsyth, and Chris Bregler. Finding naked people. In *European Conference on Computer Vision*, volume II, pages 593–602. Springer-Verlag, 1996.

[9] Graham D. Finlayson, Bernt Schiele, and James L. Crowley. Comprehensize colour normalisation algorithm. In *European Conference on Computer Vision*, pages pp 475–490, 1998.

[10] James D. Foley, Andries van Dam, Steven K. Feiner, and John F. Hughes. *Fundamentals of interactive computer graphics*. Addison-Wesley, 2 edition, 1994.

[11] Andrew B. Watson. *Computer vision for human systems*. Addison-Wesley, 1994.

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

[13] J.A. Bangham, P.W. Ling, and R. Harvey. Nonlinear scale-space causality preserving filters. *IEEE Trans. Patt. Anal. Mach. Intelli*, 18:520–528, 1996.

[14] J.A.Bangham, R.Harvey, and P.D.Ling. Morphological scale-space preserving transforms in many dimensions. *J. Electronic Imaging*, 5(3):283–299, July 1996.

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

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

[17] L. Perrochon. A quick tutorial on searching and evaluating internet resources. *IEEE communications magazine*, 35(6):142–145, 1997.

[18] J.P.Callan D.D.Lewis, R.E.Shapire and R.Papka. Training algorithms for linear text classifiers. In *Proc. ACM SIGIR Conf.*, pages 298–315, 1996.

[19] E.Riloff and W.Lehnert. Information extraction as a basis for high-precision text classification. *ACM TOIS*, 12(3):296–333, 1994.

[20] J.Cowie and W.Lehnert. Information extraction. *CACM*, 39(1):80–91, 1996.

[21] E.Riloff. Little words can make a big difference for text classification. In *Proc. 18th Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 130–136, 1995.

[22] L.D.Baker and A.K.McCullum. Distributional clustering of words for text classification. In *Proc. 21st Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, pages 96–103, 1998.

[23] D. Smith and M. Lopez. Information extraction for semi-structured documents. In *Proc. Workshop on Mgt. of Semi-structured Data, Tucson*, 1997. http://www. research.att.com/ suciu/workshop-papers.html.

[24] D. Smith and M. Lopez. Information finding and filtering for collections of semi-structured documents. In *Proc. INFORSID XV, Toulouse*, pages 353–367, 1997.

[25] G. Salton and A. Singhal. Automatic text theme generation and the analysis of text structure. Technical Report TR 94-1438, Cornell U., 1994.

[26] E.Riloff and J.Shoen. Automatically acquiring conceptual patterns without an annotated corpus. In *Proc. 3rd Workshop on Very Large Corpora*, pages 148–161, 1995.

[27] L. Raschid A. Tomasic and P. Valduriez. Scaling heterogeneous database and the design of disco. In *Proc. Int. Conf. on Distributed Computing Systems, Hong Kong*, 1996.