LARGE SCALE IMAGE-BASED ADULT-CONTENT FILTERING

Henry A. Rowley
Google, Inc., Mountain View, CA, USA

Yushi Jing
College of Computing, Georgia Institute of Technology, Atlanta, GA, USA

Shumeet Baluja
Google, Inc., Mountain View, CA, USA
Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, USA

Keywords: Image processing and applications, computer vision, adult-content detection, skin color detection.

Abstract: As more people start using the Internet and more content is placed online, the chances that individuals will encounter inappropriate or unwanted adult-oriented content increases. This paper presents a practical and scalable method to efficiently detect many adult-content images, specifically pornographic images. We currently use this system in a search engine that covers a large fraction of the images on the WWW. For each image, face detection is applied and a number of summary features are computed; the results are then fed to a support vector machine for classification. The results show that a significant fraction of adult-content images can be detected.

1 INTRODUCTION

As more people start using the Internet and more content is placed online, the chances that individuals will encounter inappropriate or adult-oriented content increases. Search engines can exacerbate this problem by aggregating content from many sites and summarizing it into a single result page. Many existing methods for detecting adult-content currently attempt to classify web pages based on their text content. If the text content of a page is classified as adult-content, this information can be propagated to linked images and pages. However, keyword and other text-based approaches have significant limitations. First, they are language specific and require a tremendous amount of manual work to construct (either directly, or by labeling training data for all languages). Second, many adult-content pages do not contain enough text for reliable classification. Third, the text on the page may be intentionally obfuscated (i.e. encoded in an image).

This paper looks at practical ways to detect adult-content in the images themselves, on a scale which can be applied to a search engine covering a large fraction of the images on the WWW. The focus is on efficient and robust techniques, such as color classification and face detection, which together can detect many pornographic images with little computational cost. The remainder of this paper is organized as follows. Section 2 describes some previous work in the field of image-based adult-content detection. Section 3 describes our system in more detail, followed by a description of our training and test data in Section 4 and test results in Section 5. A summary is given in Section 6.

2 BACKGROUND

Previous work on adult-content image detection can be divided into two broad categories: those that are based mainly on skin color or texture, and those which analyze the shapes of skin colored regions to determine their similarity to human figures. Representative work is described below.

Early work on detecting skin color in images made the observation that once brightness is normalized out, skin colors across different races form a fairly tight cluster, which can be modeled with a Gaussian distribution (Hunke, 1994). Later work (Jones and Rehg, 1998a) used a much larger set of data to build a three-dimensional histogram of skin color using hand labeled pixels in a large number of images collected from the WWW. One deficiency of this approach is

A. Rowley H., Jing Y. and Baluja S. (2006).
LARGE SCALE IMAGE-BASED ADULT-CONTENT FILTERING.
In *Proceedings of the First International Conference on Computer Vision Theory and Applications*, pages 290-296
DOI: 10.5220/0001377002900296

that both very bright and very dark pixels must be included in the skin color model. Since it is unlikely that skin color will have both extremes in the same image, one can build separate color models for different average image brightness levels (Zeng et al., 2004) or illumination conditions, as a weak form of color constancy. In addition to looking at color, the size and rough shape of skin colored blobs and their textures have been used to refine the decision of whether a region is actually skin (Zeng et al., 2004; Kruppa et al., 2002; Arentz and Olstad, 2004; Wang et al., 1998; Zheng et al., 2004).

Other than texture, another way to verify whether a patch of skin colored pixels corresponds to a person is to analyze the shape of the region. One of the earliest papers taking this approach was (Fleck et al., 1996). A number of authors have followed that work by developing methods to match the variety of shapes a human figure can take, typically with the assumption that all the parts have been detected by a skin color classifier, as in (Ioffe and Forsyth, 1999). One recent piece of work has attempted to segment the image using color and texture rather than skin color alone, in order to detect clothed people (Sprague and Luo, 2002).

3 PROPOSED SYSTEM

Because of the vast number of images that we will need to process (on the order of 10^9), a primary constraint is the processing speed; the goal in our work is to develop an efficient image classification system. Because many shape classification techniques can take a few seconds or even minutes per image, these methods are not currently applicable to such a large number of images. Below we describe the set of features we use, along with their motivations. The contribution of each set of features will be explored in Section 5.

The features we use can be broadly grouped into those which make use of color information, and those which do not. We first describe those which make use of a skin color model, as these features are most commonly used in this domain. We then describe some additional features which are not dependent on color. As will be seen, one of these features, the face detector, can be used to customize the skin color model to each image.

Many of the features values below are computed for a region of interest (ROI) in the image, which is a rectangle centered in the image, inset by $1/6$ of the original image dimensions on all sides. The size of this ROI was determined empirically.

Figure 1: Example of skin probability map. (Left) Input image. (Middle) Initial skin map. (Right) After erosion and dilation (note that thin lines of skin color are removed).

3.1 Skin-Dependent Features

The first set of 9 features are those dependent on color.

Skin Color Detection (2 features) This work used the skin color model from (Jones and Rehg, 1998a). The model is built from many images in which pixels have been hand labeled as skin or not. We have not focused on gray-scale images in this work; handling such images would require other segmentation and image analysis approaches.

To evaluate whether a pixel in the image is skin color, its skin and non-skin probabilities are computed from the model of (Jones and Rehg, 1998a), then combined with Bayes' rule using a prior for skin color of 20% to generate a skin map containing $P(\text{skin}|\text{color})$, as shown in Figure 1.

Noise pixels and gaps are reduced by performing gray-scale versions of erosion followed by dilation, resulting in the skin probability map that will be used for the remaining features, as shown at the right side of Figure 1.

Because not all the images we apply the filter to are color balanced, in some cases the skin pixels will not be detected. Although not explored in this work, one possibility would be to use a color constancy algorithm such as (Rosenberg et al., 2003) to correct for lighting. We will discuss later how a face detector can be used to improve the skin color model in this situation.

The first two features computed from the skin map are the mean and standard deviation of the skin map in the ROI.

Connected Component Analysis (4 features) To further clarify which pixels may belong to skin regions, connected components of pixels where the skin map has a probability over 50% are located. The number of resulting connected components is recorded as a feature. In addition, the mean and standard deviation of the skin map within the skin components (and

Figure 2: Example of using edge pixels as a texture measure. (Left) Example input image containing a textured skin-toned door, which is detected as skin color. (Right) Edge pixels in this image.

also in the ROI) are recorded as features.

As a simple measure of the compactness of the connected components, the log of the weighted average perimeter squared over the area of each component is computed. The hope is that skin colored regions which may occur in the non-person background of an image will be less compact, in general, than those that are actually skin on humans.

In subsequent sections, the skin connected components will be referred to as *skin blobs* for brevity.

Skin Texture Features (2 features) Another useful way to distinguish skin from other similarly colored regions is with texture. In this work, texture is approximated by the OpenCV implementation (Bradski, 2000) of the Canny edge detector (Canny, 1986) applied to the gray-scale image, as illustrated in Figure 2. We record the following ratio:

$$\frac{\text{number of edge pixels in skin blobs in ROI}}{\text{number of pixels in skin blobs in ROI}} \quad (1)$$

as a measure of skin texture, and the following ratio:

$$\frac{\text{number of edge pixels in skin blobs in ROI}}{\text{number of edge pixels}} \quad (2)$$

as a measure of how much of the image texture can be attributed to skin-colored pixels. The specific ratios used above were arrived at experimentally.

Lines (1 feature) Often, parts of buildings (like wooden doors or bricks) appear to have skin color. One way to distinguish these from skin is that they often have long straight edges which are not typical of skin.

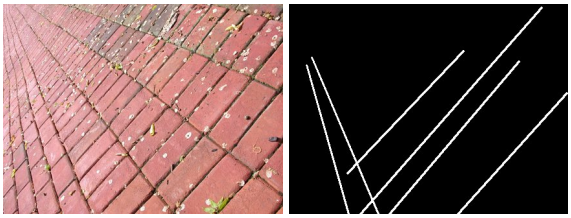


Figure 3: Example of line extraction. (Left) Example input image with straight edges on the skin color regions. (Right) Detected lines.

We use the OpenCV implementation (Bradski, 2000) of the probabilistic Hough transform (Kiryati et al., 1991) computed on the edges of the skin color connected components. The number of distinct lines detected that are not within 5 pixels of the image edge is recorded. An example of the detected lines is shown in Figure 3.

3.2 Skin-Independent Features

The remaining 9 features do not use color information.

Image Features (3 features) The first set of non-color-based features describe the image shape and size.

Images larger than 10×10 pixels have a flag feature set to one. If the image is larger than 10×10, we also compute the remainder of the features described here. Images that are smaller than 10×10 pixels have the flag and all other features set to zero; most of the features cannot be computed for such small images, which are unlikely to contain recognizable adult-content in any case.

Two additional features contain the log of the number of pixels in the ROI defined earlier, and the aspect ratio of the image.

Entropy Features (2 features) One feature that may be valuable in distinguishing adult-content images from icons or other artificial images is the entropy of the intensity histogram of the image, as follows:

$$-\sum_{i=0}^{255}(P(i) \cdot \log_2(P(i))) \quad (3)$$

where $P(i)$ is the fraction of pixels in the image with intensity i. Artificial images tend to have fewer distinct intensities present, so their entropies will be lower. We compute the entropy for both the full image and a 10 pixel wide border around the image, to aid in the detection of framed images.

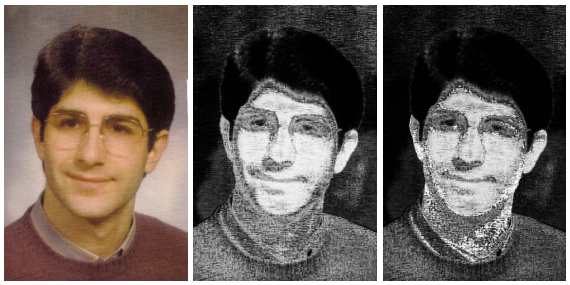


Figure 4: Example of using face detection to improve the skin color model. (Left) Sample input image. (Middle) Skin map without customization using the face detector. (Right) More accurate skin map produced by using face detection (note that the neck is brighter and sweater is darker).

Clutter Features (2 features) Just as an edge detector was used above to measure the amount of texture in the skin colored pixels, it can also be used as an overall measure of the amount of clutter in an image. We record the fraction of pixels in the ROI which are edges (a measure of central clutter in the image), and the fraction of edge pixels in the whole image which are in the ROI (a measure of the peripheral clutter relative to the central clutter).

Face Detection (2 features) One of the most reliable signals that an image contains a person is whether or not a face is present. For this work, we used the OpenCV face detector (Bradski, 2000), an implementation of the detector described in (Lienhart and Maydt, 2003), which is in turn based on the work in (Viola and Jones, 2001). From the face detection results, two features are used: the number of faces in the image, and the fraction of the image pixels which are in the largest detected face. For speed, the face detector is run on a half-resolution version of the input image.

As mentioned earlier, the face detection results can be used to customize the skin color model to a particular image, as illustrated in Figure 4. Once a face is detected, pixels in the face region are used to build a 3-channel Gaussian skin color model. Then, when detecting skin, each pixel's skin probability is estimated both from the model of (Jones and Rehg, 1998a) and the face skin model, and these estimates are averaged.

In the complete system, all of the skin color dependent features are computed twice, once with the face present in the skin map, and once with the face region set to zero probability in the skin map, as shown in Figure 5. The idea here is that large portrait photos of people's faces will contain a large amount of skin color on the face, but this skin should not count against the image when trying to decide whether it contains adult-content.

Figure 5: Example of blanking out the face region of the skin map. (Left) Example input image. (Middle) Skin map showing the face region. (Right) Skin map after face region is set to zero.

3.3 Classification

In total, we employ 27 features; 9 which do not use the skin map, and 9 which do (and are computed for two versions of the skin map, with and without faces present). Once the features have been computed for an image, they are fed into a support vector machine (SVM) for classification. In this work, the LIBSVM library was used to train and run the SVM (Chang and Lin, 2001).

4 TRAINING AND TEST DATA

The training and test data consisted of images found on the WWW. A set of 17,300 examples was hand labeled for training as either adult-content or not. The images in this set were selected from websites known to be family-safe and those known to contain adult content; images from the latter set were then manually labeled the authors as to whether each was adult-content or not. Of these 17,300 samples, 812 were adult-content and 16,488 were not.

The test set consists of 51,960 sample images uniformly collected from the Internet and manually labeled by the authors. Of these images, 1,331 were labeled as adult-content, and the remaining 50,629 as not adult-content. The test data was also manually labeled by authors. The next section will show the classification results for this data.

5 EXPERIMENTAL RESULTS

In this section, we describe the training of the SVM, experiments to evaluate the contributions of the features, and present comparisons with related work.

5.1 Training

Because the number of positive adult-content training examples greatly outweighs the negative samples, LIBSVM's weight setting for the positive examples was set to 18 during training. While this does not exactly match the proportion of ratio of positive to negative examples for historical reasons, the results of the training do not seem very sensitive to this parameter, as long as the positive examples are given a fairly large weight. RBF kernels with a gamma parameter of 0.05 were used, and all other parameters to LIBSVM were left at their defaults. LIBSVM used a total of 3,102 support vectors for the final classifier.

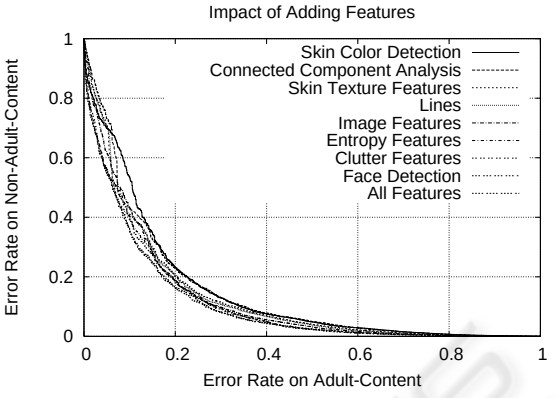
5.2 Testing

To see the impact of each set of features on the performance of the system, a number of tests were conducted. We began by training an SVM with only the basic skin color features (the mean and standard deviation of the skin map in the ROI), and then used the SVM to produce scores for each of the test images. From this data, and ROC curve plotting the number of errors on adult-content vs. the number of errors on family safe images could be produced. This curve is the upper curve in the graph of Figure 6a. For this type of graph, a better result means the curve comes closer to the origin.

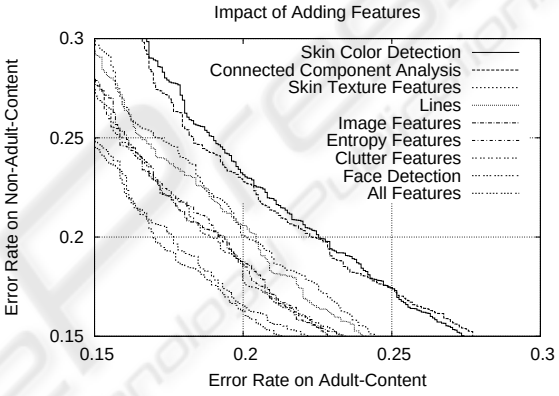
In steps, we then added batches of features, retrained the SVM and computed the resulting ROC curve. The results are shown together in Figure 6a (full range) and 6b (zoomed). Each curve represents adding one batch of features (described in a paragraph in Section 3.1) to the set of features used by the previous curve, with the exception of the last two curves. The "Face Detection" curve adds features for the number of faces and largest face, modifies the skin color model using the face pixel samples, and also modifies the skin map by removing the face rectangles. The "All Features" curve uses both versions of the features, with and without the faces removed from the skin map, which improves accuracy slightly.

As can be seen from Figure 6, as more features are added, the accuracy improves. However, there appears to be diminishing returns from the types of straightforward features we are using.

We evaluated the speed of the algorithm on a corpus of around 1.5 billion ($1.5 \cdot 10^9$) thumbnail images of less than 150×150 pixels. Processing the entire corpus took less than 8 hours using 2,500 computers, for an overall throughput of around 20 images per second per computer.



(a)



(b)

Figure 6: This figure shows the ROC curves as more features are added to the overall feature set. (a) shows the curves at their original scale, while (b) is zoomed in to show the differences between the feature sets more clearly.

5.3 Comparison with Related Work

Although we have not been able to locate any two papers which use the same test sets, it may be useful to examine the reported error rates which have been published. A numerical summary of related results is shown in Table 1. Each of these numbers is drawn from the paper, or estimated from an ROC curve presented in the paper. The same results are plotted on a graph in Figure 7, along with the ROC curve from our system. In addition, the ROC curve for "Zheng (our data)" evaluates the work of (Zheng et al., 2004) using its open source implementation in (di Linguistica Computazionale, 2004) applied to our test set.

The results presented by (Arentz and Olstad, 2004) are interesting because in addition to showing the accuracy for a general set of images, results are also reported for a test set containing portrait pictures of people. As might be expected, the false alarm rate on these pictures is higher, as they contain a significant

Table 1: A summary of accuracy results (on different test sets) from related work.

Citation	Positive error rate	Negative error rate
(Ioffe and Forsyth, 1998)	49%	10%
(Fleck et al., 1996)	56%	2%
(Wang et al., 1998)	4%	9%
(Arentz and Olstad, 2004) (all images)	4.6%	12%
(Arentz and Olstad, 2004) (portraits)	4.6%	26.5%
(Duan et al., 2002)	19.3%	10%
(Zeng et al., 2004)	23.5%	5%
(Jones and Rehg, 1998b)	14.2%	7.5%
(Zheng et al., 2004)	9%	20%

amount of skin color while not being adult-content. It is expected that our system's use of a face detector may eliminate some of these false alarms.

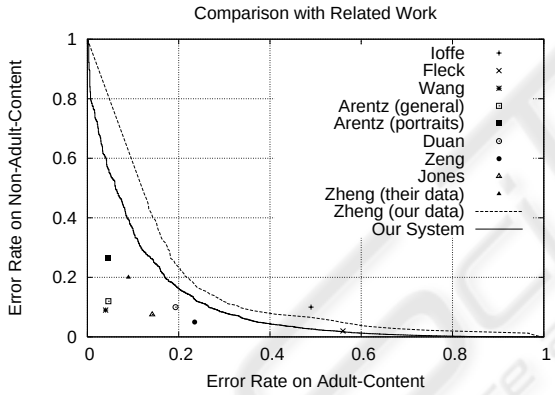


Figure 7: Graphical summary of accuracy results (on different test sets) from related work, as well as the ROC curve from our system for comparison purposes. The only direct comparison which can be made is between Our System and Zheng (our data), and it can be seen that our system has significantly better accuracy.

As can be seen, there is a wide variety in the reported accuracies of the systems that have been published. Given that many of these systems are heavily dependent on skin color models and features that are very similar across systems, it is quite likely that the wide range of differences in accuracy can be attributed to differences in the training and test set distributions. This underscores the importance of creating a consistent test set with which to measure progress in this field. Using the WWW as the source of test images as we did creates a very difficult, real-world test case. Not only is the content extremely diverse, but the quality, resolution, color balance, and brightness vary greatly per image—increasing the difficulty of this problem dramatically. It is worth noting that many of the false negatives in our results are due to grayscale or cartoon adult-content images, which cannot be detected by the algorithms described here.

6 SUMMARY AND FUTURE WORK

The results above indicate that the system is able to detect roughly 50% of the adult-content images in a small test set, with roughly 10% of the safe images being incorrectly marked as adult-content; or at a different threshold detecting 90% of adult-content images with a false alarm rate of 35%. However, since safe images significantly outnumber those with adult-content, this leads to a large number of false alarms, indicating there is much work remaining to be done.

The system described above has been incorporated into Google's adult-content filtering infrastructure, and is now in active use for image safe-search.

There are a number of potential ways to improve this work. Face image processing can determine the age (Lanitis et al., 2004) and gender (Moghaddam and Yang, 2000; Baluja and Rowley, 2005) of people in the image, which may turn out to be useful features, perhaps simply as priors. More elaborate representations of the shape of the skin color blobs, such as those used in (Wang et al., 1998) may be helpful without significantly increasing the computational cost. Better measures of texture beyond simple edge detection may improve the recognition of skin regions. Finally, image similarity matching against a database of adult-content image signatures (using ideas similar to (Tieu and Viola, 2000; Lowe, 2004)) may be useful. Any image-based system can be used to complement the results of a keyword based filter, since each approach has its own strengths and weaknesses.

To end on a practical note, because of the ubiquity of the Internet, search engines, and the widespread proliferation of electronic images, adult-content detection is an extremely important problem to address. To improve the rate of progress in this field, it would be useful to establish a large fixed test set which can be used by both researchers and commercial ventures. Another avenue to pursue is the creation of an annual competition for filtering systems. Unfortunately, the images used in this paper were obtained from the WWW, and likely contain copyrighted content, so cannot be redistributed. Nonetheless, the authors look forward to collaborating with other researchers, both in terms of the algorithms and approaches taken, as well as ideas to promote more interest and faster progress in this field.

REFERENCES

Arentz, W. A. and Olstad, B. (2004). Classifying offensive sites based on image content. In *Computer Vision and Image Understanding*.

Baluja, S. and Rowley, H. A. (2005). Boosting sex identification performance. In *Innovative Applications of Artificial Intelligence*, Pittsburgh, PA, USA.

Bradski, G. (2000). *Programmer's Toolchest: The OpenCV Library*. Software available at `http://www.intel.com/research/mrl/research/opencv/index.htm`.

Canny, J. (1986). A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6).

Chang, C.-C. and Lin, C.-J. (2001). *LIBSVM: a library for support vector machines*. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

di Linguistica Computazionale, I. (2004). *POESIA: Public Open-source Environment for a Safer Internet Access*. Software available at `http://www.poesia-filter.org`, alpha 2 release, `poesieasoft_snapshot_25july2003.tgz`.

Duan, L., Cui, G., Gao, W., and Zhang, H. (2002). Adult image detection method base-on skin color model and support vector machine. In *Asian Conference on Computer Vision*, pages 797–800, Melbourne, Australia.

Fleck, M. M., Forsyth, D. A., and Bregler, C. (1996). Finding naked people. In *European Conference on Computer Vision*.

Hunke, H. M. (1994). Locating and tracking of human faces with neural networks. Master's thesis, University of Karlsruhe.

Ioffe, S. and Forsyth, D. (1998). Learning to find pictures of people. In *Neural Information Processing Systems*.

Ioffe, S. and Forsyth, D. (1999). Finding people by sampling. In *International Conference on Computer Vision*.

Jones, M. J. and Rehg, J. M. (1998a). Stastical color models with applications to skin detection. *International Journal of Computer Vision*, 46(1):81–96.

Jones, M. J. and Rehg, J. M. (1998b). Stastical color models with applications to skin detection. Technical report, Compaq Cambridge Research Laboratory.

Kiryati, N., Eldar, Y., and Bruckstein, A. M. (1991). A probabilistic Hough transform. *Pattern Recognition*, 2(4):303–316.

Kruppa, H., Bauer, M. A., and Schiele, B. (2002). Skin patch detection in real-world images. In *Annual Pattern Recognition Symposium DAGM*.

Lanitis, A., Draganova, C., and Chirstodoulou, C. (2004). Compariong difference classifiers for automatic age estimation. *IEEE Transactions on Systems, Man, and Cybernetics*, 34(1).

Lienhart, R. and Maydt, J. (2003). Empirical analysis of detection cascades of boosted classifiers for rapid object detection. In *Annual Pattern Recognition Symposium DAGM*.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

Moghaddam, B. and Yang, M.-H. (2000). Sex with support vector machines. In *Neural Information Processing Systems*.

Rosenberg, C., Minka, T., and Ladsariya, A. (2003). Bayesian color constancy with non-gaussian models. In *Neural Information Processing Systems*.

Sprague, N. and Luo, J. (2002). Clothed people detection in still images. In *International Conference on Pattern Recognition*.

Tieu, K. and Viola, P. (2000). Boosting image retrieval. In *Computer Vision and Pattern Recognition*.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition*.

Wang, J. Z., Li, J., Wiederhold, G., and Firschein, O. (1998). System for screening objectionable images. In *Computer Communications Journal*.

Zeng, W., Gao, W., Zhang, T., and Liu, Y. (2004). Image guarder: An intelligent detector for adult images. In *Asian Conference on Computer Vision*, pages 1080–1084, Jeju Island, Korea.

Zheng, H., Daoudi, M., and Jedynak, B. (2004). Blocking adult images based on statistical skin detection. *Electronic Letters on Computer Vision and Image Analysis*, 4(2):1–14.