

On the detection of pornographic digital images

R. Schettini^a, C. Brambilla^b, C. Cusano^{ac}, G. Ciocca^{ac}

^aDISCO, Università degli Studi di Milano Bicocca, Via Bicocca degli Arcimboldi 8, 20126
Milano Italy

^bIMATI, Consiglio Nazionale delle Ricerche, Via Bassini 15, 20133 Milano, Italy

^cITC, Consiglio Nazionale delle Ricerche, Via Bassini 15, 20133 Milano, Italy

ABSTRACT

The paper addresses the problem of distinguishing between pornographic and non-pornographic photographs, for the design of semantic filters for the web. Both, decision forests of trees built according to CART (Classification And Regression Trees) methodology and Support Vectors Machines (SVM), have been used to perform the classification. The photographs are described by a set of low-level features, features that can be automatically computed simply on gray-level and color representation of the image. The database used in our experiments contained 1500 photographs, 750 of which labeled as pornographic on the basis of the independent judgement of several viewers.

Keywords: adult image detection, CART, decision forests, image classification, low-level features, Support Vectors Machines.

1. INTRODUCTION

Most of the web filters currently in use are based on text or contextual cues, and can be easily deceived, for very often neither text nor context are reliable indicators of pornographic sites. We have, instead, addressed the problem of distinguishing between pornographic and non-pornographic images on image content alone. Several similar attempts have been reported in recent years. Forsyth and Fleck have developed a system for detecting the presence of nudes in an image by analyzing the structure of the skin regions¹; Wang et al. have described a system capable of classifying an image as objectionable or benign using a combination of color histograms, texture filters and a wavelet-based shape matching algorithm²; Johns and Rehg have combined a photo detector with standard text analysis to identify adult web sites³.

We have experimented two different strategies for classifying images as pornographic or non-pornographic: decision forests of trees built according to the CART (Classification And Regression Trees) methodology and Support Vectors Machines (SVM). These strategies are briefly described in Section 2. Our experimental results, on a database of 1500 photographs, are presented in Section 3. The photographs are described by a set of low-level features referring to color and edge distribution, texture and composition of the image, and the distribution of skin regions. We show that the results can be further improved by introducing a rejection option in the classification process.

2. CLASSIFICATION STRATEGIES

We have chosen to experiment with decision forests consisting of trees built according to CART (Classification And Regression Trees) methodology^{4,5}, and Support Vectors Machines (SVM)^{5,6} since both are non-parametric, i.e. they do not require any distributional assumption about the features, and have provided good generalization accuracy in other imaging applications, even when the dimension of the feature space is high^{7,8}.

In both approaches we have also experimented the application of an ambiguity rejection option⁹. This has the advantage that ambiguous images, that is images that may be labeled differently by different viewers, are likely to be rejected, while the classification accuracy of non-ambiguous images is improved.

2.1 Decision forests

Generally speaking, CART trees are trees constructed by recursively partitioning the input space, the splits being determined by conditions related to the values of the input variables. Each subset corresponds to a node of the tree: the whole input space corresponds to the root node, the subsets of the final partition correspond to the terminal nodes. Once a tree has been constructed, a class is assigned to each the terminal node. When a new case is processed by the tree, the class associated with the terminal node in which the case ends up is its predicted class. In problems where it is feasible to assume that the cost of misclassifying a class j case as a class i case is the same for all $i \neq j$, the class assigned to each terminal node t is the class i for which $p(i|t) = \max p(j|t)$, where $p(j|t)$ is the resubstitution estimate of the conditional probability of class j in node t (the probability that a case found in node t is a class j case). This rule maximizes the resubstitution estimate of the accuracy inside the node, given by $p(i|t)$. When it is not realistic to assume equal misclassification costs, the class assigned to each terminal node of the tree is the class for which the estimated misclassification cost inside the node is minimized. In our studies to date we have assumed equal misclassification costs.

The critical problems of the splitting process are essentially two: how to identify candidate splits, and how to define the goodness of the splits. Candidate splits are generated by a set of admissible questions regarding the values of the input variables, which differ according to the nature of the variables themselves. The central idea of the goodness of the splits, is that the splits be selected so that the data in the descendant nodes are purer than the data in the original ones. To do so, a function of impurity of the nodes is introduced, and the decrease in its value produced by a split is taken as a measure of the goodness of the split itself.

The size of a tree is treated as a tuning parameter, and the optimal size is adaptively chosen from the data. A very large tree is grown and then pruned, using a cost-complexity criterion which governs the tradeoff between size and accuracy, or cost. This eliminates both the risk of large trees which overfit the training data, and that of small trees that do not capture important information. The pruning process generates a sequence of subtrees decreasing in size; these are evaluated in terms of their accuracy, or misclassification cost, and the best subtree is then selected. When, as is the case here, large sets of data are available the accuracy, or misclassification cost, of the subtrees are usually estimated on the basis of a test set.

Although the pruning process prevents the danger of trees too tailored to the training data, there is still overfitting due to instability, a phenomenon inherent in the hierarchical nature of the construction process of trees.

To overcome this problem instead of using single trees as classifiers in our experimentation we have used decision forests. A forest consists of trees obtained by bootstrapping the training set a number of times and using the bootstrap replicates as new training sets.

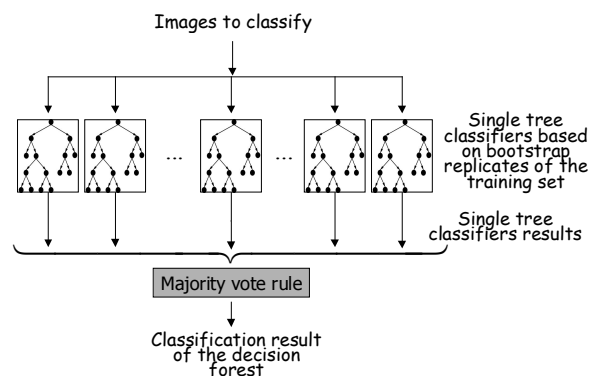


Figure 1. A decision forest.

The classification results obtained with the trees of the forest are combined by means of a majority vote rule. Figure 1 shows the scheme of a decision forest.

To provide a measure of confidence in the classification results and, at the same time, achieve greater accuracy, we applied an ambiguity rejection rule to the decision forest: the classification obtained by means of the majority vote is rejected if the percentage of trees that contribute to it is lower than a given threshold. In this way only those results to which the classifier assigns a given confidence, are accepted.

2.2 Support Vector Machines

The central idea of SVM is the adjustment of a discriminating function so that it optimally uses the separability information of the boundary cases. Given a set of cases which belong to one of two classes, training a linear SVM consists in searching for the hyperplane that leaves the largest possible number of cases of the same class on the same side, while maximizing the distance of either class from the hyperplane. If the training set is linearly separable, then a discriminant hyperplane will satisfy the inequalities:

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1, \quad (1)$$

where $\mathbf{x}_i \in \mathcal{R}^d$ is a vector of the training set, d being the dimension of the input space, and $y_i \in \{-1, +1\}$ is the corresponding class. Among the separating hyperplanes, the SVM approach selects the one for which the distance to the closest point is maximal. Since such a distance is $1/\|\mathbf{w}\|$, finding the hyperplane is equivalent to minimizing $\|\mathbf{w}\|^2$ under constraints (1). The points closest to the hyperplane are called *Support Vectors*, and the quantity $2/\|\mathbf{w}\|$ is called the *margin* (see Figure 2); it can be considered a measure of the generalization ability of the SVM: the larger the margin, the better the generalization is expected to be.

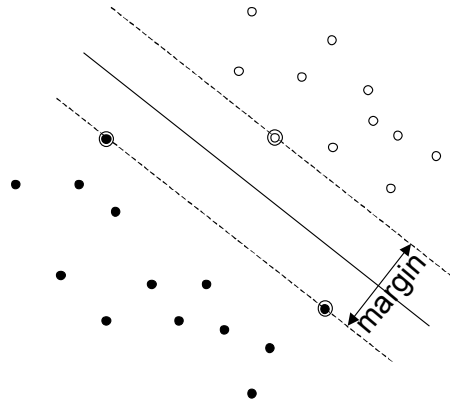


Figure 2. Separating hyperplane. The Support Vectors are circled.

When the training set is not linearly separable, the problem is relaxed by introducing a set of slack variables e_i and a penalization for the points that are misclassified. This leads to the optimization problem

$$\min \left(\|\mathbf{w}\|^2 + C \sum_{i=1}^N e_i \right), \quad (2)$$

under the constraints

$$y_i (\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1 - e_i, \quad e_i \geq 0. \quad (3)$$

The parameter C determines a trade-off between the error on the training set and the separation of the two classes.

The SVM approach can be extended to non-linear decision surfaces through a non-linear function Φ which maps the original feature space \mathfrak{R}^d into a higher dimensional space H . Since the only operation needed on H is the inner product, if we have a kernel function

$$k : \mathfrak{R}^d \times \mathfrak{R}^d \rightarrow \mathfrak{R}, k(\mathbf{x}', \mathbf{x}'') = \Phi(\mathbf{x}') \cdot \Phi(\mathbf{x}''), \quad (4)$$

Φ mapping is never explicitly used. The kernel will exist, provided some non stringent conditions are fulfilled. Examples of widely used kernel functions are the polynomial kernel:

$$k(\mathbf{x}', \mathbf{x}'') = (\mathbf{x}' \cdot \mathbf{x}'' + 1)^p, \quad (5)$$

and the gaussian kernel:

$$k(\mathbf{x}', \mathbf{x}'') = \exp\left(-\frac{\|\mathbf{x}' - \mathbf{x}''\|^2}{\sigma}\right). \quad (6)$$

We also applied a reject option for the SVM classifier: a case is rejected when its distance from the decision surface is lower than a given threshold.

3. EXPERIMENTAL RESULTS

A database of 1500 photographs, 750 of which labeled as pornographic on the basis of the independent judgement of several viewers was used. The non-pornographic class included indoor and outdoor images of different subjects. Out of the 1500 photographs 1000 were randomly selected to form the training set in such a way that the two classes were equally represented.

3.1 Image description

The photographs are described by a set of low-level features, that is features that can be automatically computed simply on the basis of the gray-level and color representation of the image. The features refer to color and edge distribution, texture and composition of the image, and the distribution of skin regions.

To describe the color distribution we computed the first three moments, mean, standard deviation, and skewness, of each color channel of the HSV color space, derived by transforming the R, G, and B color coordinates. The color distribution of an image can in fact be considered a probability distribution and can therefore be characterized uniquely by its central moments alone, as can any probability distribution¹⁰.

To compute the features related to image composition, the HSV color space was partitioned into eleven color zones corresponding to basic color names (red, orange, yellow, green, blue, purple, pink, brown, black, gray and white). This division was defined and validated empirically by three different groups of examiners. Since after quantization the image presented noisy points due to regions of non-uniform color, a max filter was applied to remove these points. Based on the uniform color regions formed by the filtering process, four spatial composition features were computed: fragmentation (the number of color regions), distribution of the color regions with respect to the center of the image, and distribution of the color regions with respect to the x axis, and with respect to the y axis¹¹.

To detect the presence of skin in the image, we used a statistical skin color detector based on a probability model of the r , g chromaticities of the pixels. The model was trained with a set of 30000 color skin data of three different human races: African, Caucasian, and Indian¹². The percentage of skin pixels has already been used in a previous study as a feature for discriminating between indoor, outdoor and close-up photographs¹³. To identify pornographic images, we now refined the skin detector by labeling as skin only those pixels with a small texture amplitude. This was computed as:

$$A = med_2(|I - med_1(I)|), \quad (7)$$

where I is the luminance, while med_1 and med_2 are two median filters of different size (5x5, and 7x7 respectively)¹⁴. After labeling the regions of connected skin pixels, the following descriptors were computed:

- the number of skin regions and the total area;
- the percentage of skin pixels belonging to small, medium size, and large regions (skin regions covering less than 1% of the image area were considered small; those covering more than 5% of the image were considered large);
- the average distance between the baricenter of the regions and the center of the image itself;
- the horizontal and vertical dispersion of skin pixels.

Figure 3 shows an example of the skin detection and labeling process.

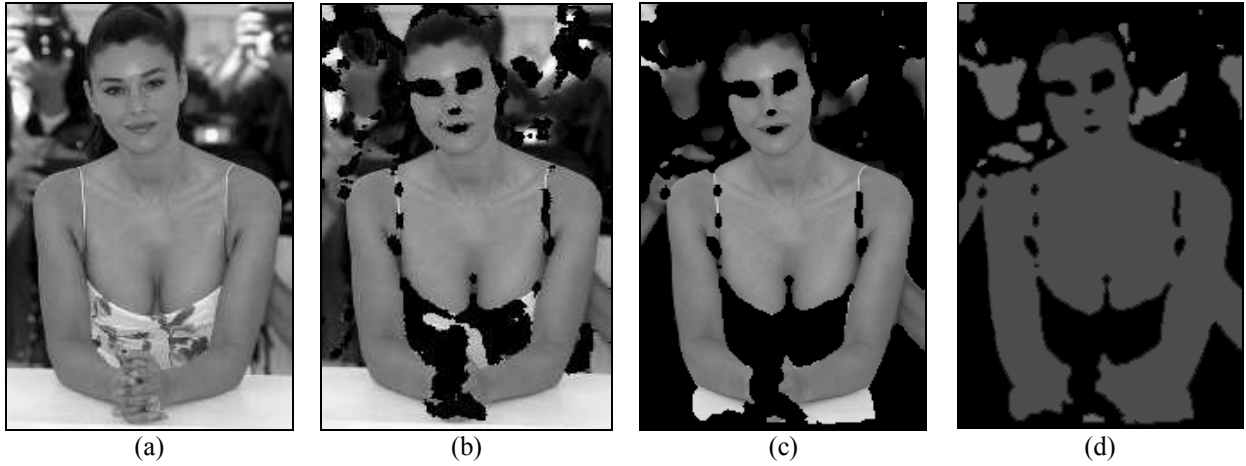


Fig. 3. Skin detection and labeling: input image (a), after texture filter (b), after color filter (c), labeled regions (d).

Multiresolution wavelet analysis was used to provide representations of the image data in which both spatial and frequency information are present. In multiresolution wavelet analysis we have four bands for each level of resolution resulting from the application of a low-pass filter (L) and an high-pass filter (H). The filters are applied in pairs in the four combinations LL, LH, HL and HH, followed by a decimation step that halves the resulting image size. Each band (the output of the filtering and decimation processes) corresponds to a coefficient matrix one-fourth the size of the processed image. The final image, of the same size as the original, contains a smoothed version of the original (LL band) and three bands of details (LH, HL and HH). In our procedure the features were extracted from the luminance image using three-step Daubechies multiresolution wavelet expansion to produce ten sub-bands¹⁵. Two energy features, the mean and variance, were computed for each sub-band.

The statistical information on image edges was extracted by Canny's algorithm¹⁶. The Canny operator works in a multi-stage process. It begins by computing first the luminance image. Then, after smoothing the image by a Gaussian convolution, a simple 2D first derivative operator is applied to the smoothed image in both the x and y directions to reveal the edges with high first derivative values. With these values the algorithm computes the gradient magnitude for each pixel. Edge pixels have higher gradient values than non-edge ones. The algorithm tracks the pixel with higher gradient value, setting at zero all pixels that are not on the edge line, and producing in output an image with lines only 1-pixel wide. This process is called *non-maximal suppression*. After which all the non-zero pixels are tagged as being possible edge pixels. To find the real edge, an *hysteresis* phase is applied: starting from the possible edge pixels, the algorithm searches for the next edge pixels. This ensures that the edges have a smooth behavior. All together, eleven features ere derived by the use of the Canny algorithm.

Texture features are related to local spatial changes in the intensity of the image pixels. We based their computation on the Neighborhood Gray Tone Difference Matrix (NGTDM) proposed by Amadasun and King¹⁷ and Tamura et al.¹⁸. In this matrix each element is the difference between the intensity of a pixel and the average intensity of the neighboring pixels. The features computed were related to five texture properties: coarseness, contrast, busyness, complexity, and strength.

Table 1 summarizes the features used.

Group	Features	Components
Color	HSV Moments	9
	Skin Statistics	8
Texture	NGTDM	5
	Wavelets	20
Edge Distribution	Canny Edge Statistics	11
	Edge Direction Histogram	18
Composition	Fragmentation	1
	Symmetry	3

Table 1. Summary of the features used to describe the images.

3.2 Classification results

Table 2 shows the results obtained on the test set by using a single tree classifier.

		Predicted class	
		Not Pornographic	Pornographic
True class	Not Pornographic	0.772	0.228
	Pornographic	0.208	0.792

Table 2. Confusion matrix obtained on the test set by using a single tree.

Table 3 shows the results obtained on the test set by using a decision forest of 25 CART trees. As expected, the use of the decision forest produced a marked improvement in generalization accuracy, of 3% and 7% for non-pornographic and pornographic classes respectively. No improvement was obtained by increasing the size of the forest.

		Predicted class	
		Not Pornographic	Pornographic
True class	Not Pornographic	0.801	0.199
	Pornographic	0.138	0.862

Table 3. Confusion matrix obtained on the test set by using a decision forest with 25 trees.

Tables 4 shows the results achieved on the test set by using a SVM with a gaussian kernel ($\sigma=5$, $C=250$). The generalization accuracy achieved by the SVM is better for both classes: the improvement over the results shown in Table 2 is about 8% for non-pornographic images and about 4% for pornographic ones.

		Predicted class	
		Not Pornographic	Pornographic
True class	Not Pornographic	0.884	0.116
	Pornographic	0.096	0.904

Table 4. Confusion matrix obtained on the test set by using the SVM.

Figures 4 and 5 show how the generalization accuracy increases as the rejection rate increases when the rejection option is applied, in both the CART and the SVM approach. In the CART approach the rejection rule experimented states that a classification result obtained by means of the majority vote inside a decision forest is rejected if the percentage of trees that contribute to it is less than a given threshold. In the SVM approach the rule states that a classification result is rejected if the distance of a feature vector from the decision surface is less than a given threshold. Both rules are global, in the sense that they do not include local information regarding the feature space.

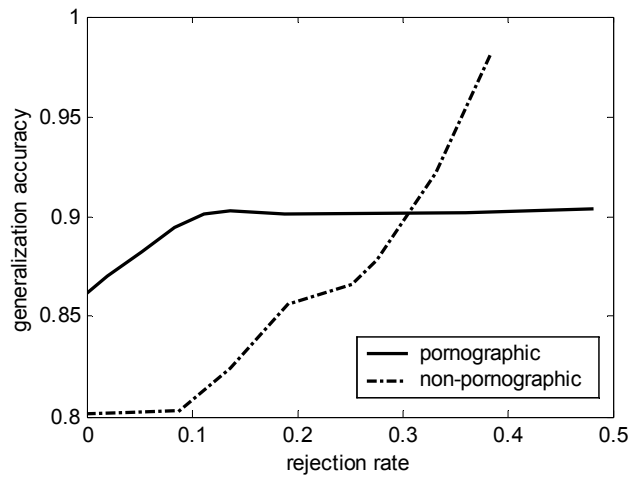


Figure 4. Plot of the generalization accuracy vs. the rejection rate, when the decision forest is applied.

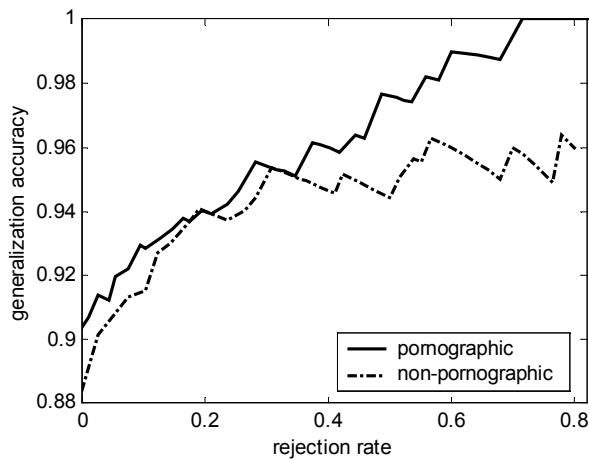


Figure 5. Plot of the generalization accuracy vs. the rejection rate, when SVM is applied.

Once again, SVM gives a better performance. For example, when the rejection rate is 20%, the generalization accuracy is approximately 94% for both classes; that of the decision forest is less than 90%.

4. CONCLUSIONS

The experimental results presented here show that the use of only low-level features to describe digital images can provide for satisfactory detection of pornographic images. Our investigation is still at an early stages and a number of issues deserve future study. First of all we intend to go on working with both CART trees and SVM classifiers. CART trees have properties which may be very useful for the design of image filters for the WEB, and the difference in performance is not such as to exclude the feasibility of using them. Qualitative features could be dealt with as easily as quantitative features, and any different misclassification costs could be readily handled. Moreover, since CART methodology allows for the presence of missing values both in the training set and in new cases to be classified, we could exploit ancillary information which may not always be available.

Taking advantage of the fact that trees provide a clear understanding of the conditions driving the classification process, we may gain insight into the role that the different features play in the classification process. In particular, the effects of the interaction among the skin features and the other features are of great interest. The analysis of the role of the features in the classification process will allow a careful re-examination of the features set and the detection of any possible redundancy: while noisy features do not have a detrimental effect on trees performance, support vector machines are adversely affected by them.

We also plan to design a more powerful skin detector that will be insensitive to changes in lighting.

REFERENCES

1. D. A. Forsyth and M. M. Fleck. "Identifying nude pictures.", *Proc. IEEE Workshop on Application of Computer Vision*, 103-108, 1996.
2. J. Z. Wang, J. Li, G. Wiederhold, O. Firschein, "System for Screening Objectionable Images", *Computer Communications Journal*, **21**(15), 1355-1360, 1998.
3. M. J. Jones and J. M. Rehg, "Statistical color models with application to skin detection", *Proc. IEEE Computer Vision and Pattern Recognition*, **1**, 274-280, 1999.
4. L. Breiman, J.H. Friedman, R.A. Olshen, C.J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks/Cole, 1984.
5. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2001.
6. V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, 1995.
7. R. Schettini, C. Brambilla, G. Ciocca, A. Valsasna, M. De Ponti, "A Hierarchical Classification Strategy for Digital Documents", *Pattern Recognition*, **35**, 1759-1769, 2002.
8. K. Goh, E. Chang, K. Cheng, "Support Vector Machine Pairwise Classifiers with Error Reduction for Image Classification", <http://www1.acm.org/sigs/sigmm/MM2001>, 2001.
9. Vailaya and A. Jain, "Reject option for VQ-based bayesian classification", *15th International Conference on Pattern Recognition*, Barcelona, Spain, September, 2000.
10. M.A Stricker, M. Orenge, "Similarity of color images", *SPIE Storage and Retrieval for Image and Video Databases III Conference*, 1995.
11. P. Ciocca, R. Schettini, "A relevance feedback mechanism for content-based retrieval", *Information Processing and Management*, **35**, 605-632, 1999.
12. R. Schettini, A. Valsasna, C. Brambilla and M. De Ponti, "A new classification strategy for color documents", *Internet Imaging II, Proc of SPIE*, **4311**, 70-78, 2001.
13. R. Schettini, C. Brambilla, and C. Cusano, "Content-Based Classification of Digital Photos", *Proc. MCS 2002*, 272-281, 2002.
14. M. Fleck, D. Forsyth, and C. Bregler, "Finding Naked People", *Proc. European Conf. on Computer Vision*, **2**, 592-602, 1996.
15. P. Scheunders, S. Livens, G. Van de Wouwer, P. Vautrot, D. Van Dyck, "Wavelet-based texture analysis", *International Journal Computer Science and Information management*. wcc.ruca.ua.ac.be/~livens/WTA/, 1997.

16. J. Canny, "A computational approach to edge detection", *IEEE Trans. On Pattern Analysis and Machine Intelligence*, **IEEE-8**, 679-698, 1986.
17. M. Amadasun, R. King, "Textural features corresponding to textural properties", *IEEE Transaction on System, Man and Cybernetics*, **19**(5), 1264-1274, 1989.
18. H. Tamura, S. Mori, T. Yamawaki, "Textural features corresponding to visual perception", *IEEE Transaction on System, Man and Cybernetics*, **8**, 460-473, 1978.