# Characterizing Objectionable Image Content (Pornography and Nude Images) of specific Web Segments: Chile as a case study

*J. Ruiz-del-Solar*[1,2], *V.Castañeda*[1,2], *R. Verschae*[1,2], *R. Baeza-Yates*[2], *and F. Ortiz*[2]

[1]Department of Electrical Engineering, Universidad de Chile
[2]Center for Web Research, Department of Computer Science, Universidad de Chile

## Abstract

In this paper we perform a characterization of the objectionable image content (pornography and nude images) of the Chilean web (.cl domain). This characterization is based on a novel objectionable image filtering system that detects porno and nude images. In an automated process we examine all .cl websites, and download a large number of the available images. Then, we analyze the images' contents and we detect human-skin blobs. Afterwards, special features are derived from these skin blobs (size, orientation, position, solidity, eccentricity, etc.). Finally, a statistical classifier is employed for classifying the images in three classes: normal, porno and nude. In addition to these, we characterize the three classes by their faces content. We present statistics of use to anyone concerned with the objectionable image content of the web in Chile. We believe that the described methodology can be employed for carrying out similar studies in other segments of the web, and also for the automatic discovering of new pornographic sites.

## 1. Introduction

One of the main problems of the Internet is the presence of harmful (e.g. pornographic) or even illegal (e.g. pedophilic) contents. The amount of this non-desired material is growing at an increasing rate. According to a recent estimation [15], 28% of the web servers contain pornographic contents, mostly images. (This study was performed by analyzing 9,800 randomly chosen websites, and using 94 pornographic stop words). On the other hand, the amount of objectionable images received by email, mainly by spam email, is increasing. In 2003 spam made up 55% of all email traveling over the Internet, 19% of this spam corresponds to adult messages [15]. For these reasons, the effective filtering of images in Internet (email and websites) is of paramount importance. This is especially relevant for protecting children from objectionable contents. Therefore, pornography filters for children are becoming popular in schools and homes. Moreover, these filters can also be used for controlling adults' access to Internet (for example, employees' control during working hours). Pornographic filters work under one of the following three strategies: (i) keywords for filtering the text contents associated with pornographic contents, (ii) lists of adult web site addresses for filtering the access to these prohibited sites, and (iii) content-based analysis of images for detecting offensive material. This last alternative has the advantage of allowing a flexible, text-independent, and dynamic filtering of pornographic contents. A good content-based filter should also distinguish between pornographic and nude images. In a given context it can be chosen to filter just one type of images.

However, filtering is not the only possible action to be taken in order to deal with the increasing amount of pornographic material in the Web. On the one hand, the automatic pornographic sites' detection using content-based analysis can allow the discovering of new pornographic sites that can be added to lists of prohibited sites, which can be then filtered using traditional methods or even detailed analyzed for law enforcement purposes. On the other hand, a characterization of the pornographic contents in a given segment of the web can be important for determining users' behaviors and user bandwidth requirements in the case that only nude images are analyzed.

In this article we perform a characterization of the objectionable image content (pornography and nude images) of the Chilean web (.cl domain). In an automated process we examine all .cl websites and download a large number of the images available. These images are then automatically analyzed and classified into three classes, porno, nude and normal using a novel, content-based objectionable image filtering system that detects porno and nude images. Then, statistics of the porno and nude images are obtained. Using these statistics the Chilean web is characterized in terms of objectionable image content. It should be emphasized that, using the described strategy new pornographic sites can also be discovered. However, at the moment this is not the purpose of our work.

0-7695-2471-0/05 $20.00 © 2005 IEEE

One of the main contributions of this work is the proposed objectionable image filtering system. Its main characteristics are: (i) classification of images into three classes: normal, nude and porno, (ii) operational tuning to Internet images (training and testing was performed using 11,795 Internet images, 6,799 for training and 5,996 for testing), and (iii) 76% detection rate of objectionable images (porno + nude) with an 11% of false positive rate. (The detection rate quantifies the number of correctly detected objects, in this case objectionable objects. The false positive rate quantifies the number of wrongly detected objects. Both rates are used for characterizing object recognition or detection systems.)

The here-presented characterization study can be seen as an extension of our former studies about the image content of the Chilean web [1][2]. In that studies we characterize the images' content in terms of: (i) image size and format, (ii) color, shape and textures image features, (iii) image category (graphics or pictures), and (iv) skin and face image information, which allows to estimate the presence of human beings in the images.

The article is structured as follows. In section 2 some related works are outlined. In section 3 the proposed methodology and tools for analyzing the objectionable image content of a Web collection are described. In section 4 is presented the nude and pornographic image detection system. In section 5 we presented statistics of the pornographic image content of the Chilean web. Finally, in section 6 some conclusions of this work are given.

## 2. Related Work

Successful content-based pornography filtering approaches are based on the use of statistical classifiers that solve the pornography detection problem as a two-class classification problem, pornography vs. non-pornography. Most of the them use skin-derived color features for performing the classification [4][5][6][7][8][9][10][11][14]. Skin detection algorithms achieve high processing speed (due to its low-level nature), and invariance against rotations, partial occlusions and pose changes. State of the art skin detection techniques are also robust enough for dealing with complex backgrounds [18]. However, it is clear that not all images containing human skin correspond to pornographic images. Therefore, computed features should account not only for the presence of skin (amount of skin pixels) in a given image, but also for the characteristics of the skin patches/blobs (size, number, shape, etc.). Most of the already proposed image content-based pornography detection systems do not successfully distinguish between pornographic images and nude images. This is not a minor problem because the filtering of non-porno nude images blocks the access to some art material. Therefore a good pornography filtering system should allow the user to enable/disable the filtering of nude images.

Content-based pornography detection is not a new idea. First systems were proposed already in 1996-1997 [14][12]. However, we believe that still there are some important improvements to be done in this area. Usually, the obtained detection results are not well characterized; in many works is presented just a single operational point (detection rate vs. false positive rate) of the whole ROC curve (Receiver Operating Characteristic), which makes very difficult to characterize and compare the different systems and to tune the system to a specific application. Furthermore, each research group uses its own image database (different complexity and size) for testing its systems (obviously there is no public pornography/nude databases), which makes very difficult the comparison between the different approaches. Additionally, the different groups have different definitions and names from what they consider non-desirable material: nude, naked, porno, objectionable, offensive, non-acceptable, adult, heavy porno, etc. Nevertheless, the most important works in this area are described in the following paragraphs.

The first proposed content-based pornography detection system was based on the use of skin analysis and the detection of elongated regions (possible limbs) using edge detection, symmetry measures and the Hough transform for detecting naked people [14]. Main drawbacks of this approach are the low detection rate (60% precision and 52% recall on a test set of 138 images with naked people and 1401 assorted benign images), and the low processing speed (in 1996 about 6 minutes on a workstation). An improved version of that work achieved 79.3% detection rate and 11.3% false positive rates using a dataset of 4,289 normal images and 565 nude images [10].

The WIPE system [12][13] uses a manually specified color histogram model for pre-filtering (yellow is used as shortcut). Images that contain considerable skin (yellow) pass on to a final stage of analysis where they are classified using a combination of Daubechies wavelets, normalized central moments, and color histograms to provide semantically meaningful feature vector matching. Using WIPE high detection rates are obtained (96% in a test set of 1,096 objectionable images), but also high false positive rates (9% in a test set of 10,809 non-adult images) [13]. The WIPE system does not use a proper skin detector, and is based mainly in the power of wavelet analysis. Its main drawback is that it cannot distinguish between porno and nude images.

In [11] is proposed a nude detection system, which is based on the use of skin derived simple features and a K-NNclassifier. This system was tested only on 140 images and the obtained results are very poor (55% detection rate; 65% when using hand-labeled skin pixels). Bosson et al. [7] proposed a skin-based pornography detection system that uses a commercial face finder to detect and localize faces, which are also used for pornography detection. Using mutual information of a class given a single feature, they select the following features to be employed in the classification process: the fractional area of the largest skin blob, the number of skin segments, the fractional area of the largest skin segment, the number of colors in the image, and the fractional area of skin that is accounted for a face. For performing the detection experiments two meta-classes are defined non-acceptable images (porno and nude) and acceptable images. Best results are obtained using a Multilayer Perceptron (87.2% accuracy) and a K-NN classifier (87% accuracy). Some experiments are also performed using porno, nude and acceptable images, however no definitive results (ROC curve) are given for those experiments.

Jones and Rehg system [4] used skin color detection based on skin and non-skin pixel models. To detect adult images, some simple skin-blob derived features are extracted. The discrimination performance based solely on skin is rather good for such simple features (85.8% detection rate and 7.5% false positive rate using a test set consisting on 5,241 adult images and 13,970 non-adult images).

Other recently proposed systems achieve high detection rates. However, its main drawback is the no differentiation between porno and nude images. In [9] is proposed a detection system based on skin as well as color, texture and shape analysis. Two classes are defined, offensive and non-offensive. An 89% correct classification is obtained on a dataset of 800 offensive images and 500 non-offensive images. When classifying portraits the percentage of wrongly classified images is 26.5%. In [6] is proposed another detection system based on skin, color, texture and shape analysis. For this system the performance employing a SVM and a C4.5 classifier is compared. Best classification results are obtained using the SVM; on a test set of 110,657 natural images and 11,351 adult images is obtained a 76.5% detection rate in adult images and 95% detection rate in natural images. In [8] is proposed a detection system based on skin analysis, where most of the obtained features are computed over the largest blob, and shape analysis (Hu moments) is also employed. Obtained results are characterized in terms of FAR (False Acceptance Rate) and FRR (False Rejection rate). Using a test set of 800 porno images and 800 non-porno images, the obtained results are FAR=15% and FRR=19.5%. In [5] is described another detection system based on color analysis (HSV moments and skin statistics), texture analysis (wavelets and neighbor gray tone difference matrix), edge analysis (Canny edge detector and edge directional histograms) and composition analysis (fragmentation and symmetry). On a test set of 500 porno and non-porno images, best results were obtained using SVM (90.4%/88.4% true positive rate for the porno/non-porno class).

# 3. Analysis Tools

## 3.1 Proposed Methodology

Our methodology consists of five basic steps: (1) selection of a web segment (what part of the web to study); (2) definition of scope criteria (for each page, how many levels of links to examine and what types of content to process); (3) image downloading and filtering; (4) content-based pornography detection; and (5) analysis.

Once the web segment has been selected and the scope criteria are defined, we proceed by automatically collecting web pages using a web crawler, and extracting the links to images in each of these pages. A random, large subset of these images is then downloaded to a local server. After application of a size filter, these images are carefully analyzed and objectionable content is detected.

## 3.2 Web Crawling

Our web-crawling architecture is based on a long-term schedule for collecting sites and a short-term schedule that worries about network politeness and use of resources (CPU, bandwidth) [3]. First we obtain a list of the domains of interest (all the domains registered under .CL in this case), and then we use our crawler to obtain the web pages in each of the selected domains. The next step consists of automatically extracting the links to the images (I-URLs) and the links to the associated web pages (W-URLs). For practical purposes (processing time and storage capacity) the total amount of links is sampled and a statistical representative subset of them (e.g. uniformly distributed) is employed for the developing and testing of the tools.

## 3.3 Image Downloading and Filtering

Images are downloaded using the sampled subset of I-URLs. In our former studies we realize that small images (smaller than 50x50 pixels) mostly corresponds to graphics

[1]. Therefore, after downloading, images with size under 50x50 pixels are discarded.

### 3.4 Content-based Pornography Detection

Downloaded images (after size filtering) are analyzed using our content-based pornography detection system (objectionable image filtering system). The system detects porno and nude images by using a 3-class image statistical classifier (porno, nude and normal). The system includes three main processing steps: (i) skin detection using a state of the art skin detection algorithm, (ii) feature extraction using heuristics applied over the skin blobs, and (iii) feature classification using SVM (Support Vector Machine). In this system high classification accuracy is obtained thanks to the selection of very discriminative skin-derived features and the use of a robust classifier (SVM).

### 3.5 Face detection

The here employed face detection system is based on the boosted cascade system proposed by Viola and Jones [21], with the later improvements proposed by Wu et al. [22]. These two cascade face detectors work on grey scale images and outperform previous systems in terms of processing speed, by keeping a high detection rate. Our face detection system employs simple rectangular features (a kind of Haar wavelets) [21] together with LBP-based features [23], a nested-cascade of filters [22] that discard non-face images, the integral image for a fast computation of the rectangular features [21], weak partitioning real Adaboost as a boosting strategy for the training of the detectors [22], and LUTs (Look-up Tables) for a fast evaluation of the weak classifiers. For more details refer to [24].

### 3.6 Analysis

The porno and nude images are analyzed, and some statistics are computed.

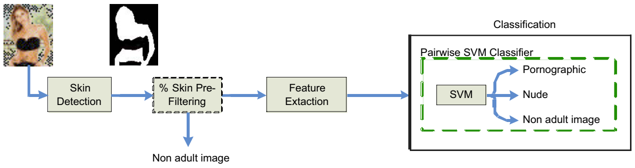


**Fig. 1.** Block diagram of the objectionable image filtering system.

## 4. Nude and Pornographic Image Detection System

### 4.1 System Overview

The block diagram of the employed content-based pornography detection system is presented in figure 1. As it can be observed the system is composed by the following modules: *skin detection*, *percentage of skin pre-filtering*, *feature extraction* and *classification*. The objective of the *skin detection* module is the localization of connected pixels (blobs) that correspond to areas where human-skin occur in the image. This module is implemented using the *skindiff* algorithm, which uses a MoG (Mixture of Gaussians) skin model and local spatial context. The *percentage of skin pre-filtering* module discards some non-adult images using information about the amount of skin contained in the image under analysis. In the *feature extraction* module 28 features are computed. These features correspond to local measures in the three largest skin blobs (size, orientation, position, solidity, etc.). Finally, in the *classification* module the images are classified in three classes, porno, nude and non-adult. In order to classify the feature vectors we used a statistical classifier: a two-class cascade SVM.

The employed class definitions are: (i) porno images: exposed genitals as center of interest in the image (male or female) or explicit sex acts with two or more participants, (ii) nude images: nude bodies with explicit female breast or pubic zones, and (iii) non-adult images: non-porno and non-nude images.

### 4.2 Skin Detection

Skin detection is a very popular and useful technique for detecting and tracking human-body parts, specially faces and hands. This kind of technique is based on the analysis of color information in images, and in the assumption that the perceived human-skin color does not change very much among different individual pixel's instances. Its most attractive properties are: (i) high processing speed due to its low-level nature (color processing), and (ii) invariance against rotations, partial occlusions and pose changes. However, standard pixel-wise skin detection algorithms are not robust enough for dealing with some real-world conditions, such as changing lighting conditions and complex backgrounds containing surfaces and objects with skin-like colors. This situation can be improved by incorporating context information in the skin detection process, as humans do. Based on this idea, in [18] is proposed a skin detection approach that uses neighborhood information (local spatial context). In this approach two main processing stages are defined, pixel-wise classification and spatial diffusion. The pixel-wise classification stage determines the pixel's *skinness*, i.e. the belonging of a pixel to the skin class (a real number between 0 and 1). The spatial diffusion stage refines the pixel-wise skin detection results, by incorporating neighborhood information for determining smooth skin segments using a region growing algorithm (see details in [18]). This approach is efficiently implemented by the so-called *skindiff* algorithm, which achieves a robust segmentation of skin regions at a high processing speed.

Skindiff uses the RGB color space, MoG models implemented using LUTs (Look-Up Table) as pixel-wise classification algorithm, and the mentioned spatial diffusion. Skindiff works directly in the RGB color space, avoiding the format conversion of images to other color spaces (most of the web images are originally in RGB). In the diffusion process a given pixel will belong to the skin class if and only if its similarity (calculated in the RGB color space) with a direct diffusion-neighbor that already belongs to the skin class, is larger than a certain threshold ($T_{diff}$). The seeds of the diffusion process are pixels with a large skinness value, i.e. their probability of being skin or their membership degree to the skin class is larger than a given threshold ($T_{seed}$). The extension of the diffusion process is controlled using a third threshold ($T_{min}$), which defines the minimal skinness allowed for a skin pixel. By analyzing the pixel-wise classification and diffusion algorithms it can be noticed that both can be joined in a single processing step. If this is done, for each RGB possible color it is also not necessary to store the skin probabilities in the LUTs, but only the information concerning the following three situations: skin probability larger than $T_{seed}$, smaller than $T_{min}$ or in [$T_{seed}$, $T_{min}$]. Therefore for each possible RGB combination, only 2 bits needs to be stored: LUT[].seed ("1" means >$T_{seed}$) and LUT[].min ("1" means >$T_{min}$). The LUTs are filled off-line using the following procedure. First, the MoG parameters are obtained. Second, for each RGB combination $rgb_i$, the mixture density function is evaluated, and LUT[$rgb_i$].seed and LUT[$rgb_i$].min are set. For keeping the size of the LUTs under control and at the same time obtaining high detection results, LUTs are quantized to 64 bins (6 bits) per color channel. This results in two LUTs of 32 Kbyte size. Skindiff processes a typical image of 400x250 pixels in about 100 ms (Pentium 4 1.8Ghz, 512 RAM, Linux Red Hat 3.2.3.).

After all image pixels are classified as skin or non-skin pixels, skin blobs, i.e. connected skin pixels, are determined.

### 4.3 Feature Extraction

The adequate selection of the features to be employed in the classification process is the key point for obtaining a good detection system. Therefore we performed a detailed analysis of the features already employed by similar systems in the literature, and we also carried out several experiments with real images. By looking at many skin-segmented porno, nude and normal images, we realized that a global measure of the amount of skin in an image, even though useful for a rough image classification, is not enough for a fine discrimination among them. Local measures of the skin blobs (size, orientation, etc.) needs to be employed. We also realized that for characterizing human nude images, either porno or pure nude, it is necessary to characterize at least the three largest skin blobs. Therefore, we decided to extract features from the three largest blobs and also to employ a global skin feature. The selected features are the following:

- **Relative Blob Area (RBA)**: Porno and nude images contain relative large uncovered (exposed) human parts. Therefore, skin blobs in these images should have relative large sizes. The RBA feature measures the relative area of the blob (number of pixels) with respect to the image area (size).

- **Relative Blob Position (RBP):** Porno and nude images contain exposed human parts in central positions (human parts correspond to the most important objects in that images). Therefore, skin blobs in these images should be placed at central positions. We define the position of a blob as the centroid of the best ellipse than contains the blob. The relative blob position (RBP) is the position of the centroid, normalized by the image width and height.

- **Relative Blob Orientation (RBO):** Porno and nude images contain uncovered human parts whose orientation matches the main orientation of the image (standard or landscape). The reason seems to be that in most of these images the image orientation is selected to fit the body orientation of the nude humans they contain. Therefore, the skin blobs' orientation in porno and nude images should match the image orientation. We define the blob orientation as the orientation of the best ellipse that contains the blob. This relative blob orientation (RBO) is calculated by subtracting the blob orientation and the image orientation.

- **Blob Eccentricity (BE):** Porno and nude images contain exposed human parts that tend to be elongated.

Therefore, skin blobs in these images also tend to be elongated objects. By considering the best ellipse than contains the blob, we define the blob eccentricity (BE) as the ratio of the minor and major ellipse axis.

- **Blob Solidity (BS):** Porno and nude images contain naked human parts that tend to be solid. That means that skin in these human parts is not covered by other objects such as cloth. This seems to be an effective way for distinguishing a naked human from a human with underwear or swimsuit. Therefore skin blobs in porno and nude images should be solid or dense objects. We define the blob solidity (BS) as the ratio between the effective blob area, that is the number of blob pixels that correspond to skin, and the total blob area, defined by the number of image pixels delimited by the external blob perimeter. In a solid object the effective blob area has the same value as the total blob area.

- **Blob Average Color (BAC)**: Exposed human parts can also be characterized by their perceived color. Due to the skin detection algorithms are not perfect, some pixels that do not correspond to skin are miss classified as skin (e.g. pixels corresponding to hair near face areas or to skin-like cloth). For this reason it is important to analyze the pixel color information in image areas that correspond to the detected skin blobs. This analysis allows also discriminating when certain skin blobs are composed by pixels corresponding to the skin of more than one human being (in porno images this situation occurs very oft). For characterizing the blob color information we define the blob average color (BAC) feature, as the average color of a blob. The BAC feature is calculated in each RGB channel (BAC_R, BAC_G, BAG_B).

As already mentioned, all features are computed for the three largest skin blobs. Therefore, the feature vector is composed by the following 27 components: {RBA_1, RBP_X_1, RBP_Y_1, RBO_1, BE_1, BS_1, BAC_R_1, BAC_G_1, BAC_B_1, RBA_2, RBP_X_2, RBP_Y_2, RBO_2, BE_2, BS_2, BAC_R_2, BAC_G_2, BAC_B_2, RBA_3, RBP_X_3, RBP_Y_3, RBO_3, BE_3, BS_3, BAC_R_3, BAC_G_3, BAC_B_3}. Where "1" denotes the largest blob, and "2"/"3" denotes the second/third largest blob.

### 4.4. Classification

For implementing the feature classification module we use a SVM (Support Vector Machine), due to its inherent classification robustness. A SVM in its simplest form (linear and separable) is defined as the hyperplane that separates the vector sets belonging to different classes with the maximum distance (margin) to its closest samples, called support vectors. The problem is solved using Lagrange multipliers. A SVM in its general form (non-linear and non-separable) is very similar to its simplest form. Non-separable cases are solved by adding an upper bound to the Lagrange multipliers [19], while non-linear cases are solved by increasing the dimensionality of the feature space, in such a way that in the high-dimensional space the problem is linearly separable. The kernel trick is employed for avoiding the computations in the high dimensional space [19].

In its original form a SVM is a two-class classifier. However, in our case we need to solve a 3-class classification problem (porno vs. nude vs. normal). There are two basic approaches to solve $q$-class classification problems using a two-class classifier. In the first approach, the so-called *one-vs-all classes* approach (or *cascade* approach), $q$ two-class classifiers are trained. Each of the classifiers separates a single class from all remaining classes. In the second approach, the so-called *pairwise* approach, $q(q\text{-}1)/2$ two-class classifiers are trained, and then, the estimated class probabilities are combined to get a joint probability estimation for the $q$-class problem [16].

In this work we implemented the pairwise approach, because when processing Internet images empirically we obtained better classification results using this approach. We employed standard SVMs trained using the Weka package [17]. The training of the SVM was made using the SMO (Sequential Minimal Optimization) algorithm.

### 4.5. Percentage of Skin Filtering

In images that include dressed humans or no humans at all, the amount of skin pixels is small or even cero. Therefore, many non-adult images can be detected using a global measure of the amount of skin in the image. We implemented a percentage of skin pre-filter for making this detection. The pre-filter is placed before the features extraction module, which speeds up the whole classification process because the non-adult images discarded by the pre-filter (usually a large amount images) are not further processed by the other system modules. The drawback of using a pre-filter is that due to its lower discriminability respect to the SVM classifier, some adult images can be detected as non-adult images in the pre-filter and discarded. Therefore, when using the pre-filter there is a tradeoff between accuracy and processing speed.

### 4.6. Implemented Nude and Pornography Detection System

The implemented detection system was tuned to operate in the Web. Therefore, it was trained using 6,799 normal, nude and porno images obtained from the Internet. The training was performed under the following conditions: (i) skin detector with operational point DR=62% (detection rate) and FPR=4.34% (false positive rate), and (ii) pre-filtering threshold of 1%.

We tested the pairwise and the cascade approaches in the classification module, using polynomial and RBF SVMs implemented with different kernel and capacity parameters. After many trials, best results were obtained using a pairwise SVM classifier with RBF kernel of parameter 0.1 and C=10.

This system was tested using a set of 4,996 Internet images. We obtained a 76% DR) of objectionable images (porno + nude) with an 11% of FPR, i.e. the number of normal images classified as objectionable. The True Positive Rate (TPR) of normal, nude and porno images is, 89%, 57% and 64%, respectively.

## 5. Statistics of the Pornographic Contents

### 5.1. Chilean Web Images

Current estimations of TodoCL [20] point out that the Chilean Web has 5 million pages ± 10% and that the number of sites and domains is 80,000 ± 10%. For performing this study we employed a random selected subset of these images. The crawling of the .CL domain was performed in March 2005. From more than 5 million images' links we download and processed 1,649,900 images.

Many of these images correspond to animations (ej. animated gif images) or graphics (graphics can be detected because they usually correspond to very small images, smaller than 50x50 pixels). Moreover, several file images present corruption or format errors. There are also many repeated files (same image or animation is used in several web pages and therefore they are linked from different places). We filtered all these images (see table 1), and we obtained 766,107 non-corrupted pictures to be analyzed by our pornography detection system.

### 5.3. Pornography Detection

We applied our nude and pornography detection system to the 766,107 images under analysis. The obtained statistics are presented in table 1. As it can be observed, from the analyzed images: 81.72% correspond to normal images (non-adult), 17.98% to nude images, and 0.3% to pornographic images. It is remarkable that adult images (nude + porno) account for about 18.3% of the analyzed images. If we take into account all processed images (1,649,900), 8.48% correspond to adult images. Given that we processed a very large number of images, we can conclude that about 8.5% of the images in the Chilean web correspond to adult images!

It is also interesting to note the adult images' distribution between different file formats. In the JPG images, which correspond to the largest subset of analyzed images (75%), 23.29% of the images correspond to nude images and 0.38% to porno images. That means that almost 24% of the JPG images correspond to adult images! When analyzing the GIF images (24% of the analyzed images), we observe a very different situation: a 99.92% of the images correspond to normal images. The reason seems to be that GIF images correspond mostly to graphics and not to pictures. Finally, it can be seen that in the PNG and BMP image groups, the adult contents is 4.31% and 9.39%, respectively.

A very important aspect is to analyze how accurate are these results. It should be kept in mind that our pornography detection system has a 76% DR. That means that the detected objectionable or adult images correspond to the 76% of adult images present in the 766,107 images under analysis. On the other hand, given that the system has an 11% FPR, about 11% of the total amount of analyzed normal images are wrongly classifies as adult.

We also studied the face contents on the analyzed images using the (frontal) face detector described in section 3.5. In figure 2 is presented a graph showing the percentage of images containing a certain number of faces. As it can be noticed, almost 80% of the on-adult images, 60% of porno images and 54% of nude images do not contain any (frontal) face. That means that, as expected, about 40% of porno and 36% of nude images contains at least one frontal face. Moreover, an important number of adult images contain just one or two faces (31% nude and 33% porno).

Finally and just for demonstrative purposes, in figure 3 we show histograms of some selected features employed for performing the pornography detection. As it can be observed, the three image groups under analysis, non-adult, nude and porno, have very different histograms contents, and therefore can be used for performing the content-based analysis here-presented.

**Table 1.** Statistics of the analyzed images. After filtering repeated file images, files with format errors or corrupted, as well as images containing graphics, the pornographic content of 766,107 images was analyzed.

| | | | | | |
|---|---|---|---|---|---|
| **Processed File Images** | 1,649,900 | | | | |
| **Repeated Files + File Format Errors + Animations + Graphics** (images smaller than 50x50 pixels) | 881,913 | | | | |
| **Pictures** (images larger than 50x50 pixels) | 767,987 | | | | |
| **Corrupted Pictures** | 1,880 | **JPG** | **GIF** | **PNG** | **BMP** |
| **Analyzed Pictures** | 766,107<br>(100%) | 574,622<br>(100%) | 183,934<br>(100%) | 6,742<br>(100%) | 809<br>(100%) |
| **- Normal Images** | 626,073<br>(81.72%) | 435,105<br>(75.72%) | 183,784<br>(99.92%) | 6,451<br>(95.68%) | 733<br>(90.61%) |
| **- Nude Images** | 137,770<br>(17.98%) | 137,257<br>(23.29%) | 148<br>(0.08%) | 290<br>(4.30%) | 75<br>(9.27%) |
| **- Porno Images** | 2,264<br>(0,30%) | 2,260<br>(0.39%) | 2<br>(0.00%) | 1<br>(0.01%) | 1<br>(0.12%) |

### 5.3. Processing Time

The time required for processing the 766,107 analyzed images is: (i) pornography detection: 17 hours (14 hours for feature extraction and 3 hours for image classification); face detection: 48 hours. As it can be seen, the whole processing can be done is less than 3 days.

## 6. Conclusions

We carried out a characterization of the objectionable image content (pornography and nude images) of the Chilean web (.cl domain). This characterization was based on a novel objectionable image filtering system that detects porno and nude images. In addition to these, we characterize the three classes by their faces content. We presented statistics of use to anyone concerned with the objectionable image content of the web in Chile. We believe that the described methodology can be employed for carrying out similar studies in other segments of the web, and also for the automatic discovering of new pornographic sites.

## Acknowledgements

This research was funded by Millenium Nucleus Center for Web Research, Grant P04-067-F, Mideplan, Chile.

## References

[1] A. Jaimes, J. Ruiz-del-Solar, R. Verschae, D. Yaksic, R. Baeza-Yates, E. Davis, C. Castillo, “On the Image Content of the Web in Chile”, *Proc. of the First Latin American Web Congress*, 72 – 83, Nov. 10 – 12, 2003, Santiago, Chile.

[2] A. Jaimes, J. Ruiz-del-Solar, R. Verschae, R. Baeza-Yates, C. Castillo, D. Yaksic, E. Davis, “On the image content of a Web Segment: Chile as a case study”, *Journal of Web Engineering*, Vol. 3, No.2 (2004) 153-168, Rinton Press.

[3] R. Baeza-Yates, and C. Castillo, “Balancing collection volume, quality and freshness in a web crawler”, in A. Abraham. J. Ruiz-del-Solar, M. Köppen (Eds.), *Soft-Computing Systems: Design, Management and Applications*, *Frontiers in Artificial Intelligence and Applications* 87, IOS Press, pp. 565 – 572, 2002.

[4] M.J. Jones, and J.M. Rehg, “Statistical color models with application to skin detection”, *Int. Journal of Computer Vision* 46(1): 81-96, 2002.

[5] R. Schettini, C. Brambilla, C. Cusanoa, G. Ciocca, “On the detection of pornographic digital images”, Visual Communications and Image Processing 2003. Edited by Ebrahimi, Touradj; Sikora, Thomas. *Proceedings of the SPIE*, Volume 5150, pp. 2105-2113 (2003).

[6] W. Zeng, W. Gao, T. Zhang and Y. Liu, “Image Guarder: An Intelligent Detector For Adult Images”, *Asian Conference on Computer Vision – ACCV 2004*, Jeju Island, Korea, Jan.27-30, 2004, pp. 198-20.

[7] A. Bosson, G. C. Cawley, Y. Chan, R. Harvey, “Non-retrieval: Blocking Pornographic Images”, Int. Conf. on Image and Video Retrieval, *Lecture Notes in Computer Science*, vol. 2383, 50-60, Springer (2002).

[8] K.M. Liang, S.D. Scott, M. Waqas, “Detecting pornographic images”, *Asian Conference on Computer Vision – ACCV 2004*, Jeju Island, Korea, 27-30 January 2004, pp. 497-502.

[9] W. A. Arentz and B. Olstad, "Classifying offensive sites based on image content", *Computer Vision and Image Understanding* 94, 2004, 295-310.

[10] D. Forsyth, and M. Fleck, `”Automatic Detection of Human Nudes," *International Journal of Computer Vision* 32 , 1, 63-77, August,1999.

[11] Y. Chan, R. Harvey and D. Smith, “Building Systems to Block Pornography", Challenge of Image Retrieval, *BCS Electronic Workshops in Computing series* (1999) 34-40.

[12] J. Wang, J. Li, G. Wiederhold, O. Firschein, “System for screening objectionable images using daubechies’ wavelets and color histograms”, *4th Int. Workshop on Interactive Distributed Multimedia Systems and Telecommunication Services*, 20 – 30, 1997.

[13] J. Wang, J. Li, G. Wiederhold, O. Firschein, “System for screening objectionable images”, *Computer Communications*, Vol.21, No. 15, pp. 1355-1360, Elsevier, 1998.

[14] M. Fleck, D. Forsyth, C. Bregler, “Finding Naked People”, *4th European Conference on Computer Vision - ECCV 96*, *Lecture Notes in Computer Science*, vol. 1065, 593 – 602, 1996.

[15] How Much Informarion? 2003 Project Web site. Berkely Univserity. Available on 26 Nov. 2004 http://www.sims.berkeley.edu/research/projects/how-much-info-2003/internet.htm

[16] T. Hastie, R. Tibshirani, “Classification by pairwise coupling”, *The Annals of Statistics 1998*, Vol. 26, No. 2, 451–471

[17] I.H. Witten and E. Frank, "Data Mining: Practical machine learning tools with Java implementations," Morgan Kaufmann, San Francisco, 2000.

[18] J. Ruiz-del-Solar and R. Verschae, “Skin Detection using Neighborhood Information”, *6th Int. Conf. on Face and Gesture Recognition – FG 2004*, 463 – 468, Seoul, Korea, May 2004.

[19] C. Cortes and V. Vapnik, “Support Vector Networks”, *Machine Learning*, 20, pp. 273-297, 1995.

[20] TodoCL Search Engine (http://www.todocl.cl/), 2000-2005.

[21] P. Viola and M. Jones, “Fast and robust classification using asymmetric adaboost and a detector cascade”, *Advances in Neural Information Processing System 14*. MIT Press, 2002.

[22] B. Wu, H. AI, C. Huang, and S. Lao, “Fast rotation invariant multi-view face detection based on real Adaboost”, *6th Int. Conf. on Face and Gesture Recognition - FG 2004*, 79–84, Seoul, Korea, May 2004.

[23] B. Fröba and A. Ernst, “Face detection with the modified census transform”, *6th Int. Conf. on Face and Gesture Recognition - FG 2004*, 91–96, Seoul, Korea, May 2004.

[24] R. Verschae, J. Ruiz-del-Solar, M. Köppen, R.V. Gracia, “Improvement of a Face Detection System by Evolutionary Multi-Objective Optimization”, *5th Int. Conf. on Hybrid Intelligent Systems - HIS 2005*, Nov. 6-9, 2005, Rio de Janeiro, Brazil (in press).

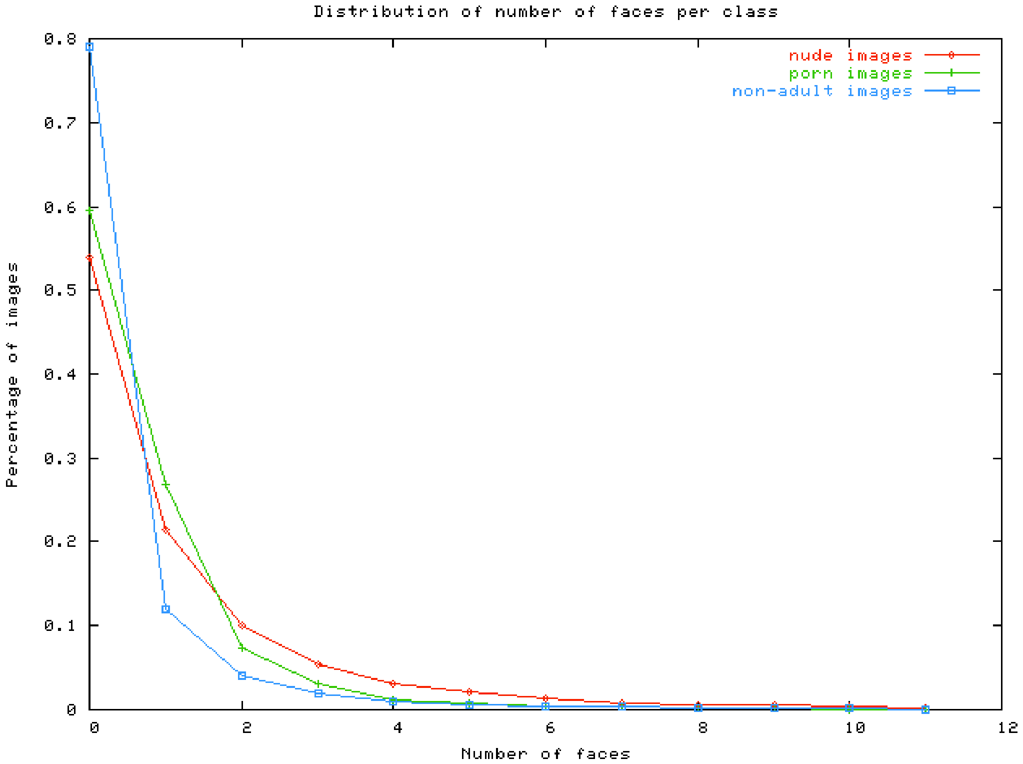
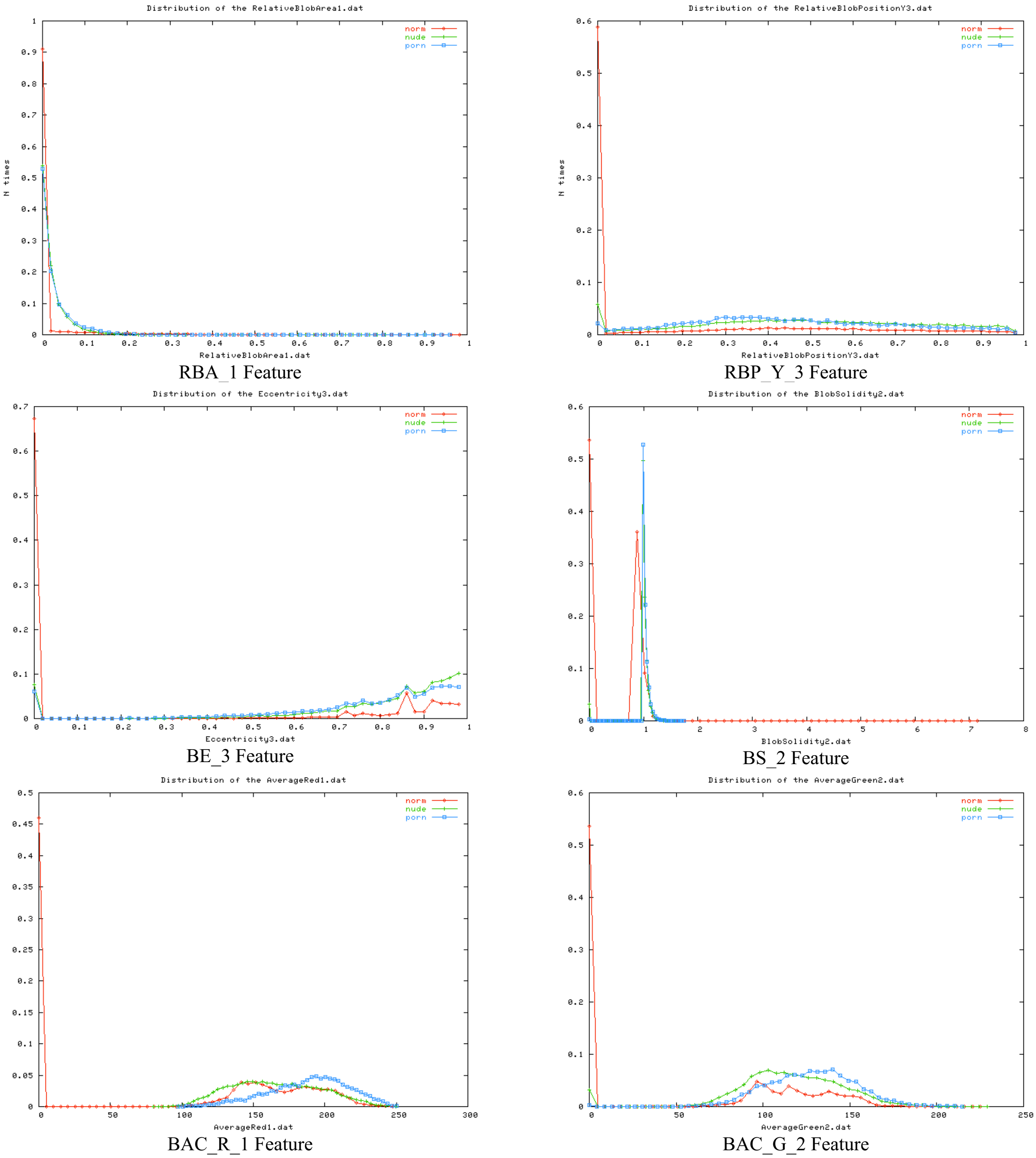


**Fig. 2.** Percentage of images containing a certain number of faces. These percentages are calculated over the analyzed images, 626,073 normal, 137,770 nude and 2,264 porno (see table 1).



RBA_1 Feature

RBP_Y_3 Feature

BE_3 Feature

BS_2 Feature

BAC_R_1 Feature

BAC_G_2 Feature

**Fig. 3.** Selected features' histograms calculated in the JPG images set. Features definition can be found in section 4.3. Different histograms are displayed for normal, nude and porno images.