Journal of Web Engineering, Vol. 3, No.2 (2004) 153-168

ON THE IMAGE CONTENT OF A WEB SEGMENT: CHILE AS A CASE STUDY

A. JAIMES[1], J. RUIZ-DEL-SOLAR[2], R. VERSCHAE[1,2],
R. BAEZA-YATES[1], C. CASTILLO[1], D. YAKSIC[1,2], E. DAVIS[1]

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

Received May 18, 2004
Revised October 31, 2004

We propose a methodology to characterize the image contents of a web segment, and we present an analysis of the contents of a segment of the Chilean web (.CL domain). Our framework uses an efficient web-crawling architecture, standard content-based analysis tools (to extract low-level features such as color, shape and texture), and novel skin and face detection algorithms. In an automated process we start by examining all websites within a domain (e.g., .cl websites), obtaining links to images, and downloading a large number of the images (in all of our experiments approx. 383,000 images that correspond to about 35 billion pixels). Once the images are downloaded to a local server, our process automatically extracts several low-level visual features (color, texture, shape, etc.). Using novel algorithms we perform skin and face detection. The results of visual feature extraction, skin, and face detection are then used to characterize the contents of a web segment. We tested our methodology on a segment of the Chilean web (.cl), by automatically downloading and processing 183,000 images in 2003 and 200,000 images in 2004. We present some statistics derived from both sets of images, which should be of use to anyone concerned with the image content of the web in Chile. Our study is the first one to use content-based tools to determine the image contents of a given web segment.

Key words: Web characterization, Web image analysis.
Communicated by: Y Deshpande & S Murugesan

1 Introduction

Web The web is growing at an increasingly rapid pace. More importantly, faster computers and network connections are allowing creators of web content more freedom to add, with fewer constraints, larger quantities of images, graphics, and video. At the same time, people's interest in using images from the web has also increased. The web in Latin America is no exception, and many Internet websites in countries like Chile are full of multimedia content. In a recent study [3], for instance, it was found that in 2001 the keyword "fotos" (photos) was the second most searched keyword in the Chilean search engine TodoCL [3, 15]

Given the trend to enrich websites with multimedia, it becomes increasingly important to be able to characterize a given segment of the web according to the multimedia elements that it contains. This type of information is of great importance for Internet service providers (who can determine required levels of regional service), for content producers, for researchers in content-based retrieval, and for web search application developers.

Characterizing the multimedia contents of the web, however, is a challenging technical problem. First, one must deal with huge amounts of distributed data. Second, it is necessary to use media-specific content-based analysis tools to be able to determine the content of multimedia data (i.e., not just using metadata). With images and video, this means developing tools to automatically determine visual characteristics using features that represent color, texture, shape, etc. More interestingly, it implies using algorithms to automatically detect objects of interest (e.g. persons). Obviously, given the large amounts of data, manual characterization of web content is not an option.

In recent years, advances in content-based image and video retrieval research [12] have produced a large number of tools for automatic content-based analysis. In spite of such advances, there has been little progress in automatic detection of objects and extraction of features at the semantic level in general multimedia collections. Although many low-level feature extraction algorithms have been proposed and implemented (i.e., those that characterize *syntactic* elements such as color and texture), their use in real application scenarios such as the web, has been very limited. One of the reasons for this is that in terms of search, people are most interested in semantic content (searching for images *of* someone or *about* something [7]). The semantic gap between what low level features can represent in general collections and what people are looking for is often quite large, so most search engines only offer the retrieval of images and video on the web using text-only queries.

One way to alleviate the semantic gap problem is to construct algorithms for specific applications and for specific types of content. This is a strategy often adopted in the Computer Vision community. In the case of the web, this is extremely difficult in part because there have been no efforts to characterize what types of content are contained in specific web collections. This is of course a very challenging task given the wide range and very large quantities of multimedia content available.

In this context, we propose a starting point for characterizing the multimedia content of specific segments of the web (or any very large heterogeneous general collection) for several purposes including the following: (1) aid researchers working in content-based retrieval by proposing a methodology to create a snapshot of specific segments of the web; (2) aid developers of search engines gain insight into how to characterize the image contents of a web collection using standard tools; (3) characterize the image contents of the web in Chile for future research purposes.

We propose a methodology and a framework that uses an efficient web-crawling architecture, standard content-based analysis tools (to extract low-level features such as color, shape and texture), and novel skin and face detection algorithms.

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) [2]. Low-level visual features are extracted using standard techniques from content-based retrieval—several feature extraction algorithms for color, shape and texture are used to compute content-dependent statistics across a segment of the web. Finally, automatic detection of faces is carried out by sequentially combining a skin detection algorithm and a state of the art face detector. The skin detection algorithm uses color analysis for determining if a given image pixel corresponds to skin or not. As many similar algorithms it is based on the fact that skin colors form a cluster in color space, but it also employs neighborhood information for better determining the skin image areas [10]. The employed face detection algorithm corresponds to a cascade asymmetrical Adaboost detector [17]. The algorithm uses simple, rectangular feature face detectors (a kind of Haar wavelets), the integral image [17] for fast computation of these feature detectors, asymmetrical Adaboost as a boosting strategy for the training of the classifiers, and a cascade structure for combining successively more complex classifiers.

We present a general methodology for characterizing the contents of a segment of the web, and test the tools we have developed. We obtained more than 3.5 million web page links, from which we downloaded a sample of 479,455 pages. From the 479.455 pages we obtained a random sample of 200.000 images. We present statistics obtained from the web pages processed, as well as from the images downloaded in this study and our previous study (183.000 images in 2003).

1.1 Related Work

The authors of [14] presented a system for automatically indexing images collected from the web. Over 500,000 images and videos were catalogued, but general statistics on the visual content of the images in the entire collection (or a subset of the collection using a pre-defined criterion such as our .cl domain) were not presented. In that work images are collected automatically and assigned to categories based on surrounding text. Visual features are extracted from the images to construct a

search engine that allows search by keyword and visual content. The authors of [5] implemented a similar system, in which face detection is used in addition to low-level visual features.

Our work differs from [5] and [14] in several ways. First, we have developed novel skin and face detection algorithms. Second, our specific algorithms for web crawling, and feature extraction are also different from those in [14, 5]. Finally, we do not construct a search engine, but rather analyze the contents of a web segment by computing content-based statistics. To our knowledge, this is the first study analyzing statistics of content-based features in a large segment of the web.

Over the last few years many approaches have been developed to index images by content (what appears in the images) [6, 13, 9]. A large number of such approaches compare images using similarity measures between low-level visual features such as color, shape, and texture. Many approaches also seek to classify images at the scene (e.g., indoor, outdoor, etc.) and object levels (e.g., face, sky, etc.).

Finally, two previous studies have been performed on the contents of the web in Chile [3, 4]. None of them focused on the image content of the web, and only statistics on the image types (GIF vs. JPG, etc.) were given.

In our previous characterization of the image content of the web in Chile [6], only 183,000 images were employed, and no distinction was done between home page images and inner page images (we make this distinction in the current analysis). The second data set used in this paper was used in [1], but a detailed analysis was not presented. For details on previous work on web crawlers please see [2].

The rest of this article is structured as follows. In section 2 we describe tools for analyzing the images of a web collection. In section 3 we present a characterization of the image contents of the .CL domain as a case study. We discuss our results in section 4 and we conclude in section 5.

2 Tools for Analyzing the Images of a Web Collection

2.1 Proposed Methodology

Our methodology consists of four 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) selection of content-based features (*which visual features to extract*); and (4) analysis.

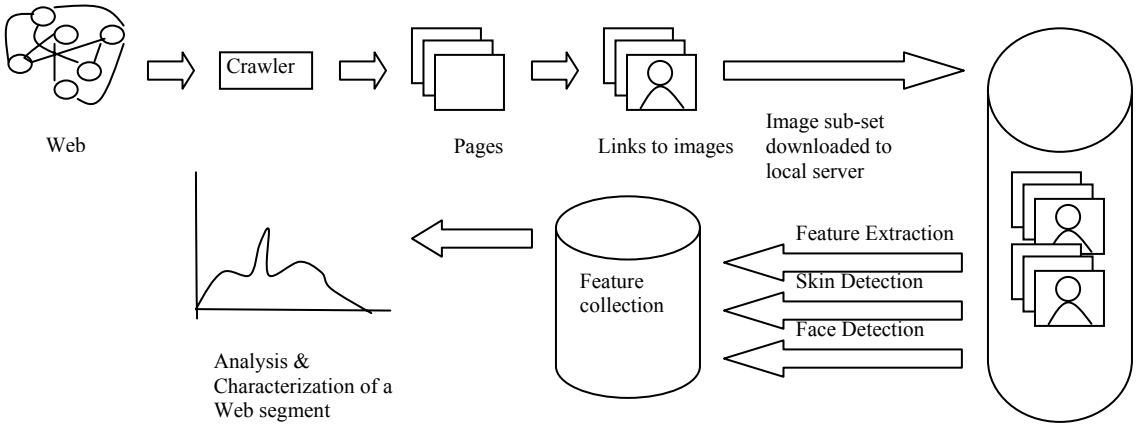


Figure 1. Overall architecture of our system.

Once the web segment has been selected and the scope criteria are defined, we proceed by automatically collecting web pages using a web crawler (section 2.2), and extracting the links to images in each of these pages. A random subset of these images is then downloaded to a local server and features are extracted for analysis (Figure 1). In the extraction of features we distinguish between low-level visual features (section 2.3.1), mid-level features (skin in section 2.3.2), and high-level features (face in section 2.3.3). This distinction is important in understanding the type of information we can obtain from the analysis at each level. In the following sections we describe our web crawling architecture and content-based analysis methodology.

2.2 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) [2]. 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.

2.3 Content-based Analysis

After downloading each of the images we extract visual features (section 2.3.1), detect skin areas (section 2.3.2) and apply the face detection algorithm we have developed (section 2.3.3). We use these results to characterize the overall visual content of the segment of the web of interest.

Table 1. Visual Features (72). Dimension of each feature space and type are given in brackets (c: color; s: shape; t: texture).

Feature	**Description**
Area (1)	Number of pixels in the image.
Aspect ratio (1, s)	Image aspect ratio.
Most dominant and second most dominant colors in HSV color space (6, c)	The HSV color space is quantized to 166 levels [14]. The most frequent and second most frequent colors in the color-quantized image are used here.
Average SV components (2, c)	Average S (Saturation) and V (Value) in HSV space.
Number of most frequent color intensities (H component) (1, c)	Color intensities (H) with number of pixels larger than 2% of image area [6].
Average and standard deviation in the R, G and B histograms (6, c)	Average and standard deviation in R, G and B histograms.
Percentiles (2%, 5%, 10%, 50%, 90%) in the R, G and B accumulated histograms (15, c)	Percentiles in the R, G and B accumulated histograms [8].
Edge Histogram in 0°, 45°, 90° and 135° (4, s, t)	Number of edge pixels in the horizontal, vertical, and diagonal (2) directions.
Extension of shape primitives (5, s)	Shape primitives (short, long, etc.) [11].
Texture features from co-occurrence matrices (20, t)	Energy, entropy, contrast, homogeneity and first moment calculated in 0°, 45°, 90° and 135° co-occurrence matrices.
Percentiles (2%, 5%, 10%, 50%, 90%) in the LBP accumulated histogram (5, t)	Percentiles in the LBP (Local Binary Pattern) accumulated histogram [8].
Texture feature MMD (Mean Maximum Difference) (1, t)	Mean value of the maximum difference between each pixel and its neighbors [11].
Texture feature MTV (Mean Total Variation) (1,t)	Mean value of the minimal difference between groups of directional neighbor pixels.
Texture feature LD (Local Deviation) (4, t)	Mean value of standard deviation of 5 direct neighbor pixels, in four different directions (0°, 45°, 90° and 135°).

2.3.1 Content-based Analysis

We extract 72 visual features that represent color, shape and texture. Although some of these features are fairly simple, they are useful in giving a snapshot of the visual content of images in the web. We give a brief overview of the features in Table 1 (in depth descriptions and implementation details can be found in the corresponding references).

We extract color features in the RGB and in the HSV color space. While RGB is often used, it is not a perceptually uniform color space, thus statistics information from the two color spaces is complementary. Mean values, standard deviations and histogram percentiles are computed in each color band. Both color spaces were quantized to 166 colors [14] (a quantization deemed sufficient to represent the majority of important colors). Edge direction histograms are computed by extracting edges from the images in different directions, keeping only the strongest edges, and counting the number of resulting edge pixels. These serve as a coarse measure of shape (and texture) and have been used effectively by other researchers in the construction of several classifiers (such as indoor and outdoor) [13]. Texture features are derived from co-occurrence matrices and LBP (Local Binary Pattern) [8].

2.3.2 Skin Detection

Skin detection or segmentation is a very popular and useful technique for detecting and tracking human-body parts, specially faces and hands. It's most attractive properties are: (i) high processing speed due to its low-level processing, and (ii) invariance against rotations, partial occlusions and changes in pose. However, skin detection is not robust enough for dealing with complex environments. Changing lighting conditions, and complex background containing surfaces and objects with skin-like colors are major problems, limiting its use in practical real-world applications. For solving the mentioned drawbacks we use context information in the skin detection process. This idea was incorporated into *SkinDiff* [10], a robust skin segmentation algorithm. The decision about the pixel's class is taken using a spatial diffusion process that employs context information. In this process a given pixel will belong to the skin class if and only if its Euclidean distance, calculated in a given color space, with a direct diffusion-neighbor that already belongs to the skin class, is smaller than a certain threshold (T_{diff}). The seeds of the diffusion process are pixels with a high probability of being skin, i.e. the skin probability is larger than a certain threshold (T_{seed}). The extension of the diffusion process is controlled using a third threshold (T_{min}), which defines the minimal probability allowed for a skin pixel. *SkinDiff* uses the RGB color space (normally images in the web use this color space) and a *Mixture of Gaussians* (MoG) model for determining the skin probabilities. For a fast computation, the MoG is implemented using look up tables (LUTs). It is not necessary to store the skin probabilities in the LUT, 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 need to be stored. For an adequate implementation of the LUTs, the colors in each channel are quantized to 64. Using SkinDiff a 320x280 image is processed in about 0.2 seconds.

2.3.3 Face Detection

Our algorithm detects frontal faces with small in-plane rotations. The detector corresponds to a cascade of filters, where each filter discards non-faces and let's face candidates pass to the next stage of the cascade. With this architecture we seek a high detection rate, considering the fact that only a few faces are to be found in an image, while almost all of the image area corresponds to non-faces. This fast detection is achieved in two ways: (i) having low complexity in the first stages of the cascade, and (ii) using simple rectangular features (the filters), which are quickly evaluated using a representation of

the image called the integral image [17]. Each of the filters of the cascade is trained using the Adaboost training algorithm [17]. The images are analyzed using 24x24 pixel windows. Each window corresponding to a color image is pre-processed (filtered) using the skin segmentation algorithm described in 2.3.2. The number of skin pixels in each window is counted, and if this number is smaller than 50% of the pixels of the window, then this window is discarded, otherwise, it is further processed. With this procedure, face detection time was reduced by a factor of 2 and the number of false detections was reduced considerably with an increase in the face detection rate. The increase in the detection rate was achieved by reducing the number of stages in the cascade when the detector was applied to color images (in grey scale images 49 stages were used, while in color images only 42). Additionally, the cascade processing was complemented using a statistical classifier added in parallel at the end of the cascade. The idea behind this procedure is the following: when fewer stages in the cascade are implemented, the detection rate increases but the false detection rate also rises (remember that each cascade stage filters non-face windows). On the other hand, a statistical classifier of face and non-face windows, implemented using low-level color and texture features, decreases the detection rate of the cascade, but also the false detection rate. Thereafter, a best compromise can be found between the obtained detection rate and false detection rate, by placing the statistical classifier at the end. After many trials it was found that the best place to put the classifier was after stage number 35. The selected classifier was the SVM and the low-level features determined using forward selection [18]. The selected features are the average of the B channel, standard deviation of the G channel, average V component, number of colors greater than 2% of the image area, percentile 50 of the G channel, percentile 10 of the B channel, and the number of edge pixels in 45° greater than the average edge of the window. Finally, the obtained detections (detected face regions) are fused for determining the size and position of the final detected faces. Overlapping detections are processed for filtering false detections and for merging correct ones. All detections are separated in disjoint sets using the heuristic described in [15]

3. Case Study: Characterization of the Images of .cl Domain

We employed real web data (images and text) sampled from the .CL top-level domain (479.455 pages obtained by randomly sampling 3.5 million web pages links and 183,000 images from 2003 and 2004).

The following process was employed for obtaining and processing this data: (i) web-crawling to sample the Chilean web collection and obtain image links (I-URLs); (ii) extraction of low-level visual features (color, edge and texture); (iii) skin detection to find image areas that contain humans and human-body parts; and (iv) face detection.

3.1 Crawling

The crawling of the .CL domain was performed in May 2003 and August 2003 for a first version of this study [6], and in January 2004 for this final version. We downloaded 100,000 images in May 2003, 83,000 images in August 2003, and finally 200,000 images in January 2004. Text information was processed only in January 2004, and the total amount of web pages downloaded for this processing was 200,000. The results of the text processing analysis are reported elsewhere [1].

We use the same terminology of [2]. A *page* is a document indexed by the crawler. A *site* is a logical web server identified by a sub-domain (e.g., *dcc.uchile.cl* which belongs to the *uchile.cl* domain). A *domain* is any name of the form *x.y* where, in this paper *y=cl*.

3.1.1 Domains and Pages

In the most recent official study of the .CL domain [3] almost 2 million pages were found in 38,307 sites in 34,867 domains. Current estimations of TodoCL [15] point out that the Chilean Web has 5 million pages ± 10% and that the number of sites and domains is 80,000 ± 10%.

From the 3.5 million page links used for this final version of our study (we considered up to 5 levels of links), we obtained two samples: one of home pages and one of inner pages. We distinguish between home pages and inner home pages to analyze the differences in content between them. A home page is the starting page of a site. An inner home page is a page of the same site, but that can be acceded from the home page (i.e., it is linked from the home page or one of its siblings). Home pages are the pages at the level 0 of a site, while inner pages can be placed at level 2 (2 clicks), level 3 (3 clicks) and so on.

3.1.2 Home Page Images

This collection was obtained from 36,455 home pages (a sub-sample from the set of 3.5 million links); from those home pages, 23,523 (64%) had objects or links to non-textual URLs. In total 338,963 links were found, 208,066 (62%) of them unique. From the unique links, 60.0% were to GIF images, 26.8% to JPG images, 7.7% to Flash animations, 2.6% to style-sheets, and 0.7% to PNG images; the rest of the links were mostly to PDF or Word documents (see figure 2). The total number of GIF, JPG and PNG images was 183,669, from which, 100,000 were randomly selected and downloaded for analysis.

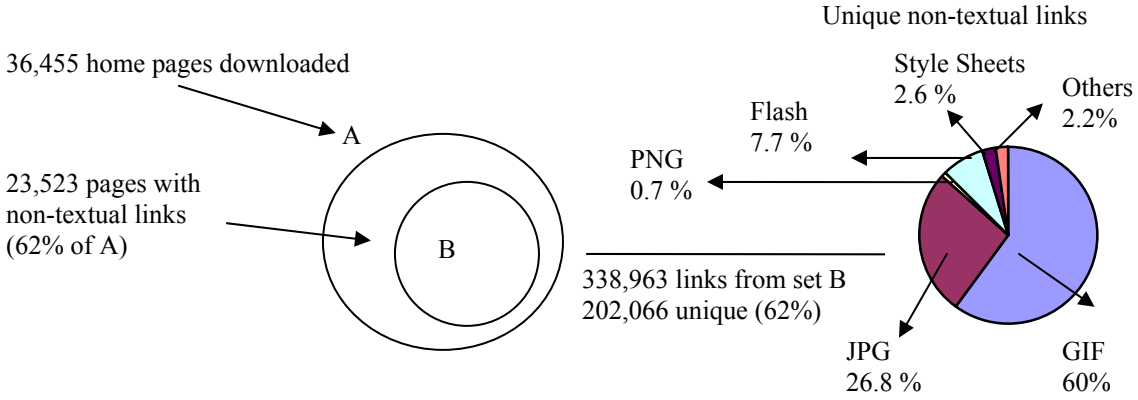


Figure 2. Distribution of pages and non-textual links for home pages.

3.1.3 Inner page Images

The sample of inner pages was obtained in 8 hours of crawling, with 443,000 pages downloaded (also a sub-sample of the 3.5 million links). We discarded all the pages that were at depth greater or equal to 5 in the websites, and all the pages without non-textual links, obtaining a sample of 311,589 pages. We believe that this sample is representative of what a user sees while browsing the web; and using pages at deeper levels would bias the sample towards large, dynamic websites. These pages contained 9,148,115 links to images, and only 926,781 (10%) were unique, relatively fewer unique links than in the home page collection (in the homepage collection 62% of the links were unique). Our interpretation is that web site owners usually have a small set of images, which are repeated across their entire websites. From the unique links, 53.9% were to GIF images, 35.4% to JPG images, 2.8% to Flash animations, 2.2% to style-sheets, and 0.8% to PNG images (see figure 3). There is a significant reduction of animations in the inner pages. The total number of images was 842.902, where 823.357

were GIF, JPG, and PNG images. From those, 100,000 were randomly selected and downloaded for analysis.

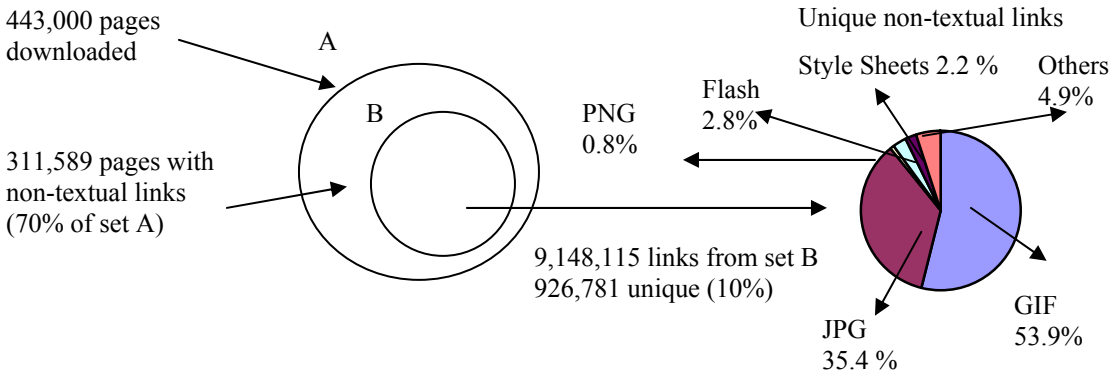


Figure 3. Distribution of pages and non-textual links for inner pages.

3.1.4 Home Page vs. Inner Page Images

There are several interesting differences between the two collections. As mentioned above, inner web pages have significantly fewer unique image links (62% vs. 10%). In addition, the percentage of JPG images is greater in inner pages (35.4% vs. 26.8%, see Table 2) and the percentage of GIF images is lower in inner pages (53.9% vs. 60%). One possible explanation for this is that JPG images tend to be photographs and are larger. In many websites the home page is basically an entry point to many other pages, so it seems natural to place more, smaller images that link to other pages, than to have fewer larger images (e.g., photographs). The number of links to Flash animations is significantly higher in home pages (7.7% vs. 2.8%). This seems natural, as it is common to for designers to use Flash animations as introductions to the page. Finally, the percentage of links to other types of content (mostly PDF and word documents) is much higher for Inner pages than for Home pages (4.9% vs. 2.2%). This makes sense if we assume most of the content is linked from inner pages.

Although these results make sense intuitively, caution must be taken and further analysis is needed. We are comparing the number of images in the home pages vs. the number of images of inner pages, where inner pages are those reached through a given home page. Thus while two home pages might be completely independent, two inner pages may not.

Table 2. Comparison of image types in home pages vs. inner pages.

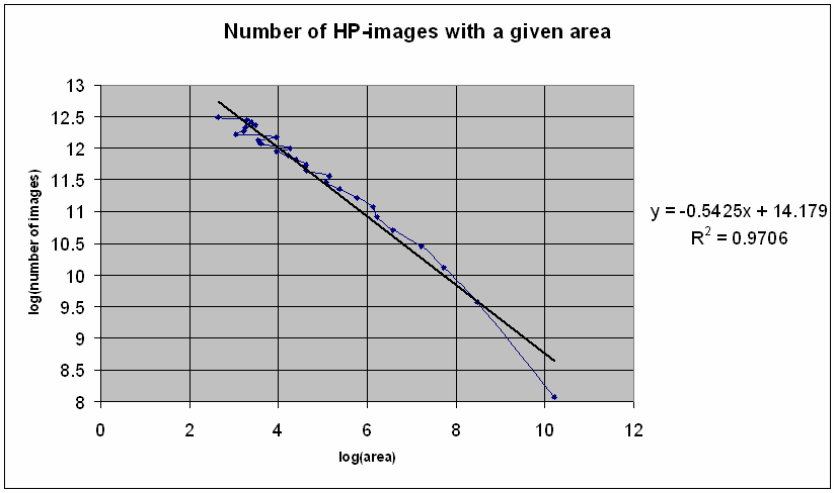
	GIF	JPG	PNG	Flash	Style Sheets	Others
Home pages	60.0%	26.8%	0.7%	7.7%	2.6%	2.2%
Inner pages	53.9%	35.4%	0.8%	2.8%	2.2%	4.9%

3.2 Content-based Processing

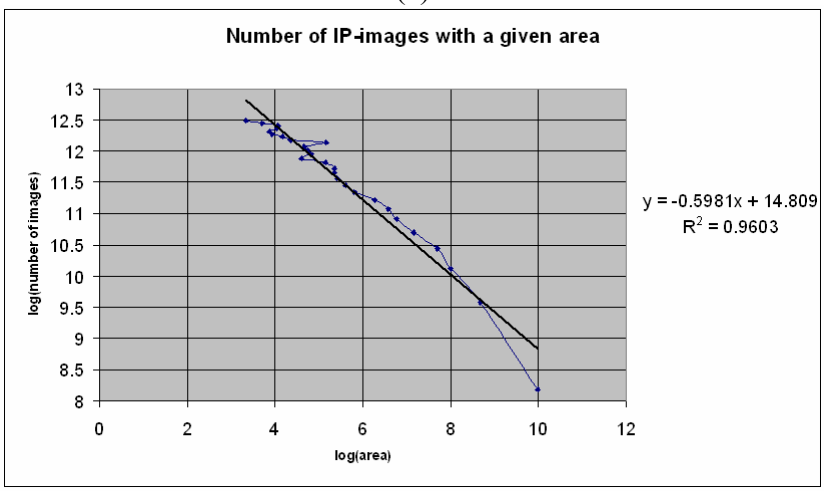
3.2.1 Visual Features

The analysis was split between home pages images (HP-images) and inner pages images (IP-images). We extracted all 72 visual features presented in Table 1. Figure 4 shows statistics about the area of the analyzed images (log-log graph) for both images sets (100,000 HP- and 100,000 IP-Images). The

graphs show that the distribution of the number of images according to area is close to a Power law in both cases.



(a)



(b)

Figure 4. Area statistics. A log-log graph of the number of images vs. area. (a) HP-Images and (b) IP-Images.

In our previous study (183,000 images in [6]) we divided the images collected into two sets: (1) set of images smaller than 50x50 pixels; (2) images larger than 50x50 pixels. We found that the percentage of large/small images is 21.10% large and 78.89% small. This means that for each large image there are approximately four small images.

Figures 5 to 7 display statistics of some selected general, color, shape, and texture features for the set of 183.000 images used in [6], where we found that most of the small images do not contain faces. Furthermore, the features of large images are very different from the ones of small images—the content of small images, mainly graphics, differs from the content of large images, mainly pictures. This characteristic has been used effectively to construct algorithms that automatic distinguish between photographs and graphics images [14]. In [14] was found that 19.2% of the images correspond to photographs and the rest to graphics.

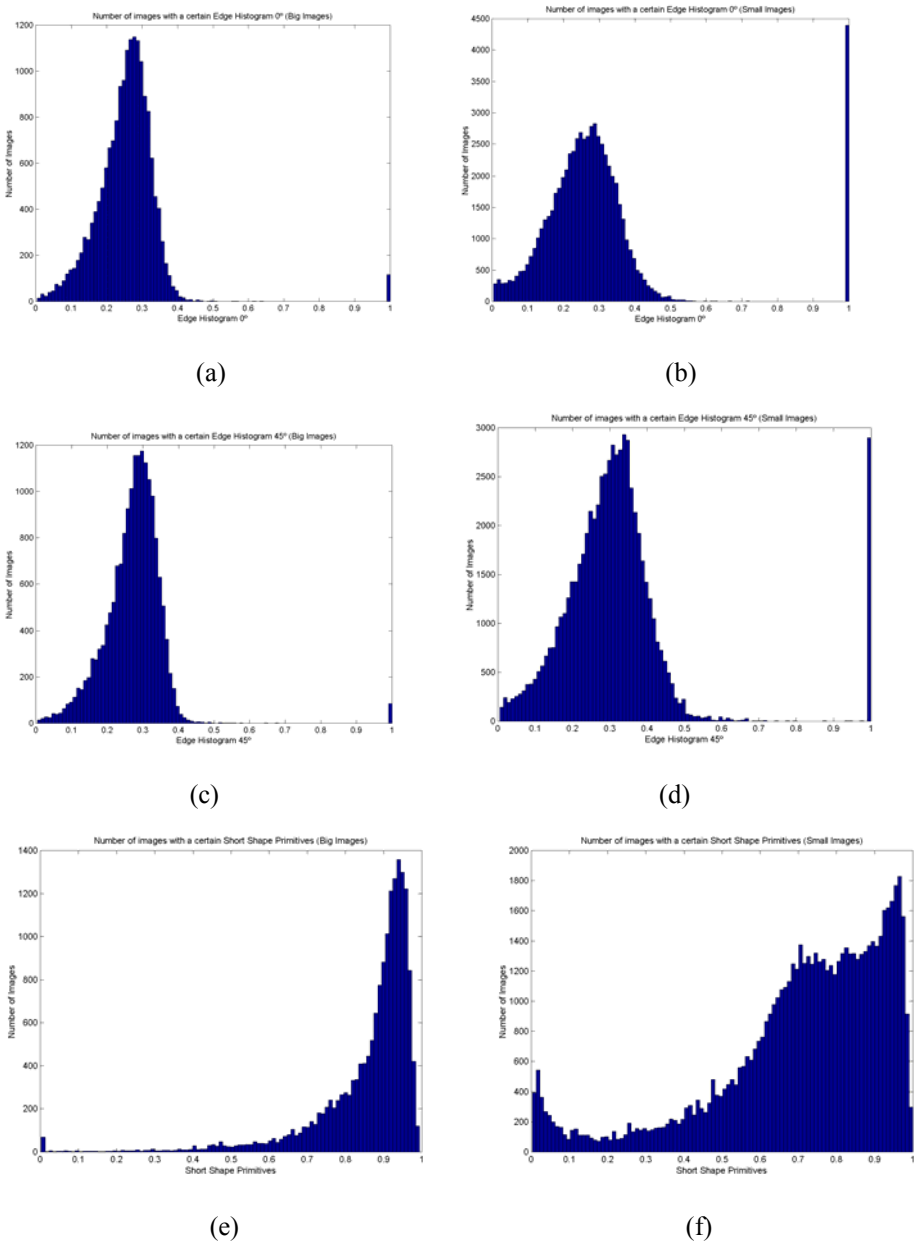


Figure 6. Selected shape features (see Table 1): Edge histogram 0° (a-b), Edge histogram 45° (c-d), and shape primitives (short) (e-f). Graphs on the left are for images that are larger than 50x50 pixels. Graphs on the right are for larger images.

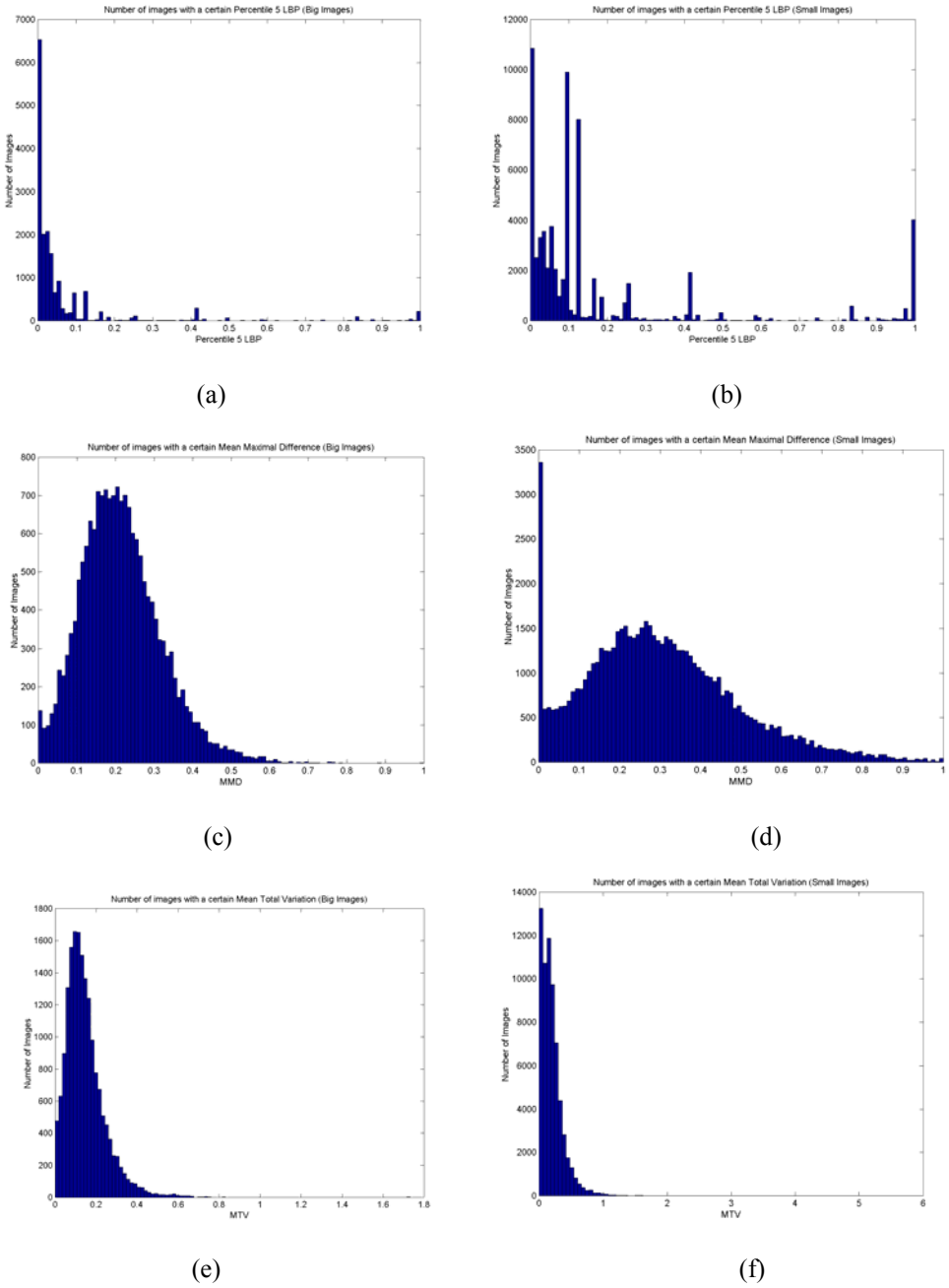


Figure 7. Selected texture features (see Table 1): Percentile 5% LBP (a-b), MMD feature (c-d), and MTV feature (e-f). Graphs on the left are for images that are larger than 50x50 pixels. Graphs on the right are for larger images.

3.2.2 Skin Detection

Skin was detected in 6.5% of the HP-images and in 7.9% of the IP-images. The average number of skin clusters is 3.73 in the HP-Images and 4.14 in the IP-Images. These differences seem to occur due to the larger size of IP-images and the larger proportion of photographs in this set. Table 3 summarizes some statistics of skin detection in both HP, and IP-images. It can also be noticed that larger skin clusters (as a percentage of the image area) and also larger skin cluster area (as a percentage of the image area) are found in HP-images.

Table 3. Summary of statistics from skin detection obtained for the HP-Images and the IP-Images.

	HP-Images		**IP-Images**	
	Average	Maximum	Average	Maximum
Number of skin clusters per image *	3.73	46	4.14	57
Average size of the skin clusters per image (in pixels) *	3167.22	262235	3121.48	220743
Average size of the skin clusters per image (in pixels) *, normalized by the area of the image	0.0745	1	0.0683	1
Size of the largest skin cluster *	5907.7	407125	6558.96	366253
Size of the largest skin cluster *, normalized by the area of the image	0.1145	1	0.1135	1
Percentage of the image that contain skin (%)	14.03	100	14.33	100

* For images where faces/skin was detected.

3.2.3 Face Detection

We found that 2.07% of the HP-images and 2.12% of the IP-images contained faces. The average number of faces per image (from those images containing faces) is 2.1167 and 2.1162 for the HP- and IP-images, respectively. The maximum number of faces found in a single image was 89 for the HP-images and 39 for the IP-images[a]. We also found that the distribution of the number of faces in both image sets (considering only the images that contained faces) is close to a Power law. For example, for the 2003 set and HP-images, considering cases from 2 to 10 faces, the parameter of the distribution is -2.13 (figure 8 shows this).

Figures 9 to 11 show some other face statistics for the 2003 image set. Figure 9 presents a histogram of the number of images containing faces with a certain size (face area). As the figure shows, most of the images contain middle-size faces and no large faces. Figure 10 shows the average of the number of faces contained in an image vs. image size. The analysis is performed for both HP- and IP-Images. It can be seen that IP-images contain more faces, specially the middle size images. Figure 11 presents the relation between face area, as a percentage of the image area, and position, as a percentage of the image size (x or y). As figure 11b shows, faces are usually centered in the images.

[a] A group photo in www.bradford.cl had 89 faces in 2003.

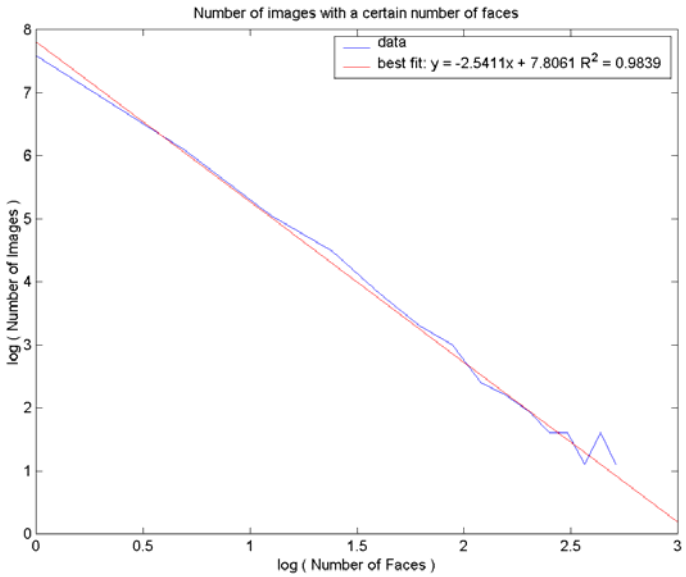


Figure 8. A log–log graph of the number of images containing a certain number of faces.

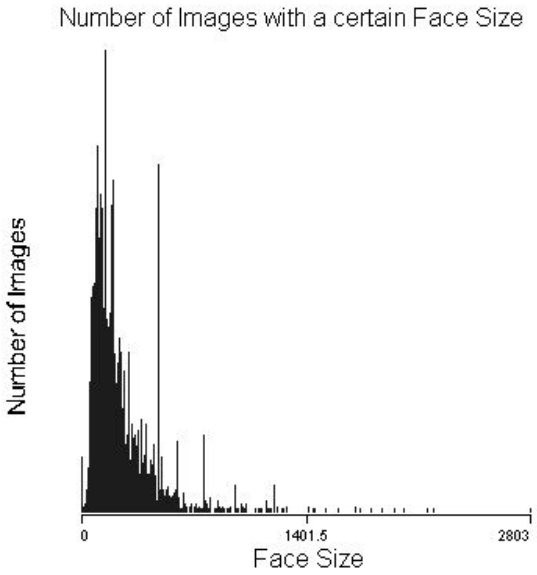
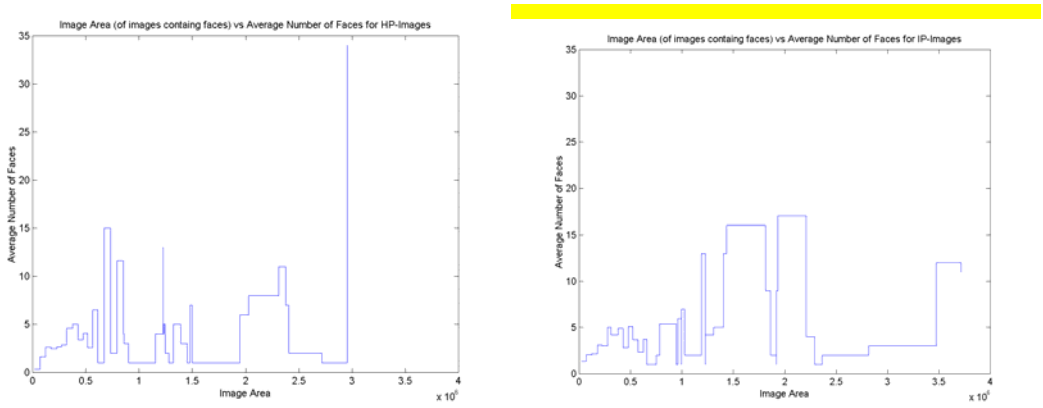
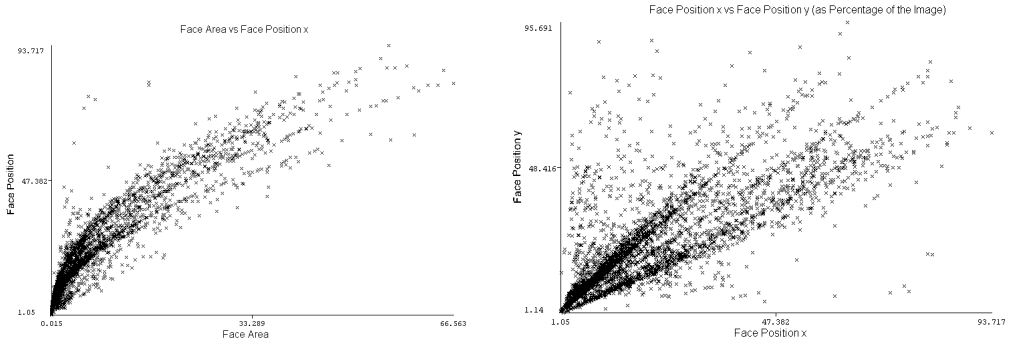


Figure 9. Histogram of the number of images containing faces with a certain size (face area).



(a) (b)

Figure 10. Image area against average number of contained faces. In (a) only HP-images considered, while in (b) only IP-Images are considered.



(a) (b)

Figure 11. Relations between face area, as a percentage of the image area, and position, as a percentage of the image size (x or y). Face area against face position in the image (a). Face position x against face position y (b).

3.3 Processing Time

Page Gathering: It took about 8 hours for the 400,000 pages using a single, standard PC running Linux. With this setting, the page recollection of the whole Chilean web takes about five days. *Feature Extraction*: The process of automatic extraction of the 72 visual features on the 200,000 images under analysis takes about 47 hours on a single, standard PC running Linux. *Skin Segmentation* and *Face Detection*: The process of skin segmentation and face detection on the 200,000 images took about 10 and 40 hours, respectively, using a single, standard PC running Linux. Obviously, any of these processes can be sped up using more than 1 PC, and can be done in a reasonable time, as it is an off-line task.

4. Discussion

We processed a total of 383,000 images, 200,000 for this last version of our study and 183,000 for a first version, and obtained some statistics to characterize their content. We found that filtering

decisions (to reduce the amount of data to be processed) are crucial and must be made carefully. Obtaining the basic statistics is trivial once the images have been downloaded, but analysis of the results requires visual inspection of the content (or further automatic analysis). It is interesting to determine the types of images that contribute to certain statistics (e.g., determine the percentage of "small" images that are photographs and not graphics, or that contain faces).

Another interesting issue related to the analysis is the accuracy of automatic results. In most approaches to automatic classification of scenes or detection of objects (e.g., faces), a small database is used as a test base. The accuracy of the results can be easily measured once the test set is manually labeled. When automatic detection algorithms are applied to large sets of data, it becomes more difficult to determine their accuracy. One way to approximate the accuracy measurement is to observe the types of errors that the algorithm makes in a large collection, in comparison to the types of errors that it makes in a smaller collection in which it is possible to take a full quantitative measurement of the algorithm's performance. In [15] our performance analysis yielded an 80% detection rate. Since the data and the types of errors are similar in [16] and in this collection, detection accuracy on the web segment we have evaluated is likely to be around 80%. Using this measurement, the rate of false-positives (wrongly detected faces) is about 1/1.000.000 for each processed window (for analyzing an image a high amount of windows is processed), for black and white images (the skin filter is not used). For color images (a skin filter is used), we obtain about one false positive every 5-7 images, for typical 320x240 images. Altogether (color and non-color images), we obtain about one false positive every 20 images, for typical 320x240 images.

Our results also show that distributions related to images are also power laws as many other measures in the web.

5. Conclusions

We presented a general framework for characterizing the image content of a segment of the web. The proposed process consists of automatically collecting web pages and downloading images from the links contained in the pages. For content-based analysis we use standard content-based retrieval tools, and novel skin and face detection algorithms we have developed. In a previous study we analyzed 183,000 images. In this study we used 3.5 million pages to obtain and analyze a set of 200,000 images A first application of these tools is the characterization of the image contents of the .CL domain. For carrying out this study a statistical representative subset of the total number of images of the .CL domain was employed. We presented statistics of the web in Chile (.cl domain) using the results of automatic low-level visual feature extraction, skin detection, and face detection.

Although we computed interesting statistics further analysis is necessary. Future work includes the implementation of more complex feature extraction algorithms and the use of text to characterize the contents of a segment of the web (i.e., the text surrounding an image).

Acknowledgements

This research was funded by Millennium Nucleus Center for Web Research, Grant P01-029-F, Mideplan, Chile. We would like to thank Mor Naaman for his comments on the first version of this study [6]. This work was done at the University of Chile, but now the first author is at FXPAL Japan, Fuji Xerox Co. Ltd., Japan and we would like to thank Jun Miyazaki and Toru Tanaka for supporting his activity at Fuji Xerox while making the revisions to this article.

References

[1] R. Baeza-Yates, J. Ruiz-del-Solar, R. Verschae, C. Castillo, and C. Hurtado, "Content-based Image Retrieval and Characterization on Specific Web Collections," *Lecture Notes in Computer Science* 3115, Springer, 189 – 198, 2004.

[2] 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.

[3] R. Baeza-Yates, B.J. Poblete, and F. Saint-Jean, *Evolución de la Web Chilena 2001-2002* (*Evolution of the Chilean Web 2001 - 2002*), Center for Web Research, Department of Computer Science, Universidad de Chile, January 2003 (in Spanish).

[4] R. Baeza-Yates and C. Castillo, "Relating Web Characteristics with Link Based Web Page Raking," *Proc. of SPIRE 2001*, IEEE CS Press, Laguna San Rafael, Chile, pp. 21-32, Nov. 2001.

[5] C. Frankel, M.J. Swain and V. Athitsos, *WebSeer: An Image Search Engine for the World Wide Web*, University of Chicago Technical Report TR-96-14, July 31, 1996.

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

[7] A. Jaimes, "Conceptual Structures and Computational Methods for Indexing and Organization of Visual Information," Ph.D. Thesis, Department of Electrical Engineering, Columbia University, February 2003.

[8] M. Niskanen, O. Silven, and H. Kauppinen, "Color an Texture based Wood Inspection with non-supervised Clustering," *Proc. of the 12th Scandinavian Conf. on Image Analysis* - SCIA 2001, 336 - 342, Bergen, Norway, June 11-14, 2001.

[9] Y. Rui, T.S. Huang, and S.-F. Chang, Image Retrieval: Current Directions, Promising Techniques, and Open Issues, *Journal of Visual Communication and Image Representation*, No. 10:1-23, 1999.

[10] 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.

[11] J. Russ. The Image Processing Handbook, 3rd Edition. CRC Press, Boca Raton, Florida, 1999.

[12] N. Sebe, M. Lew, X. Zhou, T. Huang and E. Bakker, The State of the Art in Image and Video Retrieval, *Lecture Notes in Computer Science 2728* (*Image and Video Retrieval 2003*) 1 – 8, 2003.

[13] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-Based Image Retrieval at the End of the Early Years", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, pp. 1349-1380, Dec. 2000.

[14] J.R. Smith and S.-F. Chang, An Image and Video Search Engine for the World-Wide Web, *Proc. of SPIE Storage & Retrieval for Image and Video Databases V*, Vol. 3022, pp. 84-95, San Jose, CA, Feb. 1997.

[15] TodoCL Search Engine (http://www.todocl.cl/), 2000-2004.

[16] R. Verschae and J. Ruiz-del-Solar, A Hybrid Face Detector based on an Asymmetrical Adaboost Cascade Detector and a Wavelet-Bayesian-Detector, *Lecture Notes in Computer Science* 2686, Springer, 742-749, 2003.

[17] 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, Cambridge, MA, 2002.

[18] I. Witten and E. Frank, Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations, Morgan Kaufmann, 1999. Weka homepage: http://www.cs.waikato.ac.nz/~ml/weka/