Content-based Image Retrieval and Characterization on Specific Web Collections*

R. Baeza-Yates[1], J. Ruiz-del-Solar[1,2], R. Verschae[1], C. Castillo[1], C. Hurtado[1]
[1]Center for Web Research, Department of Computer Science, Universidad de Chile, CHILE
[2]Department of Electrical Engineering, Universidad de Chile, CHILE

Abstract. One of the challenges in image and video retrieval is the content-based retrieval of images and videos in the web. Less work has been done in this area, mainly due to scalability issues. For this reason, in this paper we investigate this problem by presenting tools for the characterization of the visual contents on specific web collections and a strategy for the search of faces in the web using visual and text information. A case study is also presented in a specific web domain.

1. Introduction

Content-based image and video retrieval is a fast growing and increasing relevant research area. The research community recognizes the following main challenges in this field [8]: the bridging of the semantic gap (understanding the meaning behind the query), the content-based retrieval of videos (finding a video similar to another one), and the increasing huge amount of digital data, produced by digital consumer devices (e.g. digital cameras) and computational devices (hard disks, CD-ROMs, etc.), which needs a semantic understanding and also produces a scale problem. In addition to that, we believe that the content-based retrieval of images and videos in the web is an important and challenging area where less research has been done, probably because of technical and practical reasons. As all of us know, popular search engines allow the retrieval of images on the web using only text queries. This situation should be improved and we think we should start to develop methods and strategies for the content-based retrieval of information on the web, the largest and most used multimedia database in the world.

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 words *pictures* and *pics* are among the most queried terms). Furthermore, given the trend to enrich websites with multimedia, it becomes increasingly important to be able to characterize a given collection 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, 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

* This research was funded by Millenium Nucleus Center for Web Research, Grant P01-029-F, Chile.

is necessary to use media-specific content-based analysis tools to be able to determine the content of the multimedia elements. With images and video, this means developing tools to automatically determine their visual characteristics: 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 classification is not an option.

In this context, this paper studies the content-based retrieval of images on specific web collections (for practical reasons the whole web can not be studied at the moment), and also the characterization of the visual contents on these collections. For doing that we have developed tools for: efficient web-crawling, content-based image analysis (low-level features such as color, shape and texture), skin segmentation, face detection and web pages' clustering using text information. For developing and testing these tools we have analyzed more than 4 millions web pages, processed more than 383 thousand images (about 35 billion pixels!) and clustered the text of more than two thousand web pages.

This article is structured as follows. Related work is presented in section 2. In section 3 a strategy for the content-based retrieval of faces using visual and text information is proposed, motivating our image characterization results. Tools for processing and analyzing the images of a web collection are described in section 4. In section 5 we present a characterization of the image contents of the .CL domain as a case study. Finally, we conclude in section 6.

2. Related Work

The content-based retrieval of images and videos in the web is an underdeveloped area. However, some preliminary work has been done. Two of the most important early works are here outlined. In [9] is presented a system for automatically indexing images collected from the web. Images are automatically collected and assigned to categories based on text surrounding the images. In addition, visual features are extracted from the images to construct a search engine that allows search by visual content. However, the content-based analysis performed in this work is restricted to color histograms. In [4] is implemented a similar system, which in addition uses automatic face detection to index images on the web. This work differs from ours in the specific processing tools being used (our skin and face detection algorithms are much faster), and also in the fact that for solving actual technical problems (bandwidth, response time, etc.), we split the retrieval process in two: the off-line creation of the image database for a specific web collection and the on-line retrieval of the images. Concerning web characterization, to our knowledge, there have not been any studies of web content that use content-based features to characterize the images on the web. In the work of [9] for example, 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 criteria such as our .CL domain) were not presented. Finally, the first version of our characterization study was presented in a regional conference [5]. In that study only 83,000 images were employed, text information was not analyzed and no distinction was done between home page

images and inner page images. All this processing is performed in the current version of this study, citing the mentioned work for the low level feature image analysis.

3. Towards a Face Search Engine

As established in the introduction, the content-based retrieval of images and videos in the web is an important and challenging task that should be addressed. However, due to technical limitations (bandwidth, storage capacity, processing time, etc.) a general retrieval system of images for the web cannot be build at this time. This doesn't means that nothing can be done for the moment. On the contrary, this task can be addressed in an incremental way. To start we propose: (i) to restrict the domain of operation of the retrieval system to a certain web collection to build a vertical image search engine. We have chosen to work on the .CL domain, whose characteristics and dimensions allow the implementation of a prototype; (ii) to create an image database where the search process will be carried out. This database (a cache of images) is created off-line, using the crawling tool described in section 4.2, for solving the problems related with the required time for the gathering of the images. Thus, the on-line process of image retrieval is performed on this database; (iii) to filter the images to be stored in the database according with the functionality of the retrieval system to be built (for dealing with the storage capacity limitations). In our case we want to build a person search engine; this means that graphics, images non-containing skin and images non-containing faces should be filtered, in addition with repeated images, all these filters are described in section 4; (iv) to process and label the images to be stored in the database. In the case of the person search engine that means to store the position of the faces detected in each image and the web page class of the text associated with this image (the page clustering algorithm employed is described in 4.6).

Webfaces, the person search engine under construction, is based on the use of face and text information. For searching a given person, the user should provide a picture of the person and optionally a related text (a group of keywords). The search system will provide a set of database images where the person can be present, and a confidence value for each image. The text information will be used to determine the associated web page class and therefore to restrict the search process to a given portion of the database (the set of images with this associated class). The face contained in the provided picture is used to do a similarity search with all the faces containing in the selected subset of the database using a face recognition algorithm. Using this information, the text clustering information will be used to recommend clusters related to the similar images as well as keywords that can help to improve the query. All relevant subsystems of Webfaces are already built. We are working on the implementation of fast similarity algorithms based in metric spaces and on solving scale and orientation issues of the face recognition process, to finish the integration of all the mentioned components.

4. Tools for Analyzing the Images of a Web Collection

4.1 Proposed Methodology

For developing and testing the image retrieval and analysis tools, which are the same tools employed for characterizing the contents of certain web collections, we employed real web data (images and text) sampled from the .CL top level domain (4 millions web pages, 383,000 images and the text of 200,000 web pages containing the selected images.). The processes employed for obtaining and processing this data are: (i) web-crawling for sampling a given web collection and obtaining image links (I-URLs) and web pages associated with these images, i.e. the web pages where the images are found (W-URLs); (ii) color, edge and texture low-level visual analysis for characterizing different kinds of images and for constructing image filters, such as photographs vs. graphics and indoor vs. outdoor; (iii) skin segmentation algorithms for detecting image areas where humans and human-body parts are present; (iv) face detection algorithms for detecting humans; and (v) tools for clustering web pages using the text information associated with the processed and selected images.

4.2 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) [1]. First we obtain a list of the domains of interest (all the domains registered under .CL) 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 is employed for the developing and testing of the tools. The crawling of the .CL domain was performed in May 2003, August 2003 and January 2004. Each time about 1.3 million web pages were analyzed and the downloaded images were 100,000 in May 2003, 83,000 in August 2003 and 200,000 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.

4.3 Low-Level Visual Analysis

A set of 72 visual features that represent color, shape and texture was extracted (see feature description in [5]). Although some of these features are fairly simple, they are useful in giving a snapshot of the visual content of images in the web and in the construction of image filters. Using these basic features we build a photograph vs. graphics filter. This filter was implemented using a support vector machine classifier and 5 from the 72 features (aspect ratio, standard deviation in the R histogram, average of the S component, percentile 90% in the R histogram and the texture feature LD in 0°), which

were automatically determined using forward selection [13]. The performance of the obtained classifier is 94.5%.

4.4 Skin Segmentation

This functionality was implemented using *SkinDiff* [7], a robust skin segmentation algorithm that uses neighborhood information. 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 needs 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.

4.5 Face Detection

This algorithm detects frontal faces with small in-plane rotations. The detector corresponds to a cascade of filters, where each filter discard non-faces and let face candidates pass to the next stage of the cascade. This architecture seeks to have a fast detection, 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 a small 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 [12]. Each of the filters of the cascade is trained using the Adaboost classifier [12]. The images are analyzed using 24x24 pixel windows. Each window corresponding to a color image is preprocessed (filtered) using the skin segmentation algorithm described in 4.4. The number of skin pixels is each window is counted, and if this number is smaller than 50% of the pixels of the window, then this window 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 gray 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 color and texture low-level 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 its end. After many trials it was found that the best place to put the classifier was after the stage number 35. The selected classifier was the SVM and the low-level features determined using forward selection [13] were average of the B channel, standard deviation of the G channel, average V component, number of colors greater than 2% of 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 [11].

4.6 Text Clustering

Images in the web are inserted into web pages using the IMG html tag. The attribute ALT of the IMG tag allows us to specify a text alternative to the image, which is automatically displayed when the browser cannot display the image. Some images are included within a hypertext anchor: in this case an image may behave as a button linked to other documents or resources. The text in the ALT attribute, along with the text inside the hypertext hidden meanings. This motivated us to use the whole text anchors are candidate descriptors for the image. However, only a small fraction of the images in our collection have such descriptions. Furthermore, the quality of these descriptions is low; many of them have few words which sometime refer to file names with in the web pages as the accompanying text for the images. The text in web pages gives us some approximated context for each image. We left as future work the discovering of better descriptors for images. Such task may consider heuristics for extracting data from anchor text, ALT tags, or other parts of the html page that includes the image. We ran a clustering process over text in web pages of the images. Our goal is to discover clusters that define textual contexts for the images. Such clusters are the basis in our approach for integrating textual contexts in our image retrieval tool. Clusters centroids can be used to model textual contents, and user queries, specified as list of terms, can be compared against the centroids to determine the relevant contexts users are searching for. When the cluster associated to a query is found, the search for images can be focused on the images of the cluster. The clustering process is achieved by an implementation of a k-means algorithm provided by the clustering toolkit CLUTO [3]. We used a k-means algorithm for its simplicity and low computational cost. In addition, it has proved to be very effective for clustering collection of documents [14].

5. Case Study: Characterization of the Images of .CL Domain

Due to the limited extension of this article in this section we present a very condensed part of our characterization results. The complete results, including histograms, graphics and a complete statistical analysis can be found in our website [15].

5.1 Crawling

Domains and pages. In the most recent official study of the .CL domain [2] almost 2 million pages were found in 38,307 sites in 34,867 domains. Current estimations of TodoCL [10] point out that the Chilean Web has 5 millions of pages ± 10% and that the number of sites and domains is 80,000 ± 10%. From the 3.5 million pages used for this final version of our study (we considered up to 4 levels of links), we obtained two samples: one of home pages and one of inner pages.

Home Page Images. This collection was obtained from 36,455 home pages; from those home pages, 23,523 had objects or links to non-textual URLs. In total 338,963 links were found, 208,066 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 was mostly to PDF or Word documents. The total number of GIF, JPG and PNG images was 183,669, from those, 100,000 were randomly selected.

Inner Page Images. The sample of inner pages was obtained in 8 hours of crawling, with 443,000 pages downloaded. We discarded all the pages that were at depth greater or equal to 5 in the websites, and all the pages without links to images, 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 were unique, relatively much less unique links than in the home page collection. 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. There is a significant diminution of animations in the inner pages. The total number of GIF, JPG, and PNG images was 842,902. From those, 100,000 were randomly selected.

5.2 Image Processing and Analysis

Visual Features. We extracted all 72 visual features mentioned in section 4.3 from the 200,000 images that were processed. It was found that 19.2% of the images correspond to photographs and the rest to graphics. It is interesting to mention that the number of images with a certain area follows a Power law distribution. The analysis was split between home pages images (HP-images) and inner pages images (IP-images).

Skin Detection. It was detected skin in the 6.5% of the HP-images and in the 7.9% of the IP-images. The reason of these different percentages seems to be the larger size of IP-images and the larger proportion of photographs in this set. The average size of the skin clusters is 3167/3121 pixels, and the mean number of skin cluster in each image is 3.73/4.14, for the HP- and IP-images, respectively.
Faces Detection. We found that 2.07% of the HP-images, while 2.12% of the IP-images contained faces. The average number of faces per image (from those images containing faces) is 2.1167/2.1162 for the HP- and IP-images, respectively. The maximum number of faces found in a single image was 89/39 for the HP-/IP-images[2]. It was also found that the distribution of the number of faces in both image sets (considering only the images that contain faces) is close to a Power law. For example, for HP-images considering cases from 2 to 10 faces, the parameter of the distribution is -2.13.

5.3 Text Analysis

We consider two sets of 1,965 images each, corresponding to images that arise in web pages. The first set has images with high probability of having portions of human skin, and the second set contains images with human faces. The *face* images belong to 1,480 web pages, while the *skin* images belong only to 748 different web pages. These images were obtained using the algorithms already described. We model the associated web pages as term-weight vectors using a vocabulary of 20,600 words. *Stopwords* were eliminated from this vocabulary. The cosine similarity function between vectors was employed.

The clustering process is guided by a score function that measures the overall clusters quality. The score function used is the total sum of the average similarities between the vectors and the centroids of the clusters that are assigned to. Each run of the algorithm computes k clusters. Thus, in order to study adequate values of k we run the algorithm several times. We reach a quality score of 0.80, which reflects high similarities between objects in each clusters, at k=250 and k=300 for the home-skin and home-face dataset, respectively. In [15], we show some figures that depict the quality of the clusters found for different values of k, for the *face* and *skin* web page sets. These figures also show curves with the incremental gain of the overall quality of the clusters, and a histogram for the number of clusters per average intra-cluster similarity for the two datasets at the aforementioned values of k. Many clusters that represent clearly defined contexts for images were found. In [15], we present also tables with some of the found clusters. These clusters allowed us to discover semantic connections between web pages having faces. An example of that is a cluster containing web pages related to movies, musicals, DVDs, etc. Also, good search keywords can be detected using word frequency for each cluster.

[2] A group photo in www.bradford.cl has 89 faces!

5.4. 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 hours and 40, respectively, using a single, standard PC running Linux. *Webpage Clustering*: The text of 2228 web pages was clustered. It took 5 minutes to compute 300 clusters using a standard PC, running Windows XP. Obviously, any of these processes can be speeded up using more than 1 PC, and can be done in a reasonable time, as is an off-line task.

6. Conclusions

We investigated the content-based retrieval of images on specific web collections and also the characterization of the visual contents on these collections. For doing that we presented tools for: efficient web-crawling, content-based image analysis (low-level features such as color, shape and texture), skin segmentation, face detection and web pages' clustering using text information. For developing and testing these tools we analyzed more than 4 millions web pages, processed more than 383,000 images and clustered the text of more than 2,200 web pages. 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. In the final version of this article we plan to also include results on clustering a larger sample of only the text segments surrounding the selected images.

In this article we also presented a strategy for the content-based search of persons using visual and text information is proposed. All relevant components of this system, including a face recognition subsystem, are already built. We are working in the system final integration, which will be reported in a near future.

References

[1] 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.

[2] 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).

[3] CLUTO Home page: http://www-users.cs.umn.edu/~karypis/cluto/

[4] 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.

[5] 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.

[6] 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.

[7] J. Ruiz-del-Solar and R. Verschae, Robust Skin Segmentation using Neighborhood Information, ICIP 2004, submitted.

[8] 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.

[9] 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.

[10] TodoCL Search Engine (http://www.todocl.cl/)

[11] 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.

[12] 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.

[13] 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/

[14] Y. Zhao and G. Karypis, Comparison of Agglomerative and partitional document clustering algorithms, *SIAM Workshop on Clustering High-dimensional Data and its Applications*, 2002.

[15] http://www.cwr.cl/chile-images/