Combining Text and Image Analysis in The Web Filtering System "WEBGUARD"

Mohamed Hammami, Youssef Chahir, Liming Chen

COMBINING TEXT AND IMAGE ANALYSIS IN THE WEB FILTERING SYSTEM "WEBGUARD"

Mohamed Hammami
LIRIS, Ecole Centrale de Lyon
36, Av Guy de Collongue, 69131 Ecully-France

Youssef Chahir
GREYC - URA CNRS 6072
Campus II - BP 5186 Université de Caen
14032 Caen Cedex

Liming Chen
LIRIS, Ecole Centrale de Lyon
36, Av Guy de Collongue, 69131 Ecully-France

ABSTRACT

Web applications increasingly utilize search techniques that heavily rely on content-based text and image analyses. For example, for parental site filtering, it is necessary to identify adult sites. These applications must rely on a semantic analysis of images in the process of identification where text analysis alone is insufficient. In this article, we describe our site filtering system "WebGuard" and show the importance of image analysis in such system. Our results show that it can detect and filter adult content effectively.

KEYWORDS

Web filtering, Data-mining, Image analysis, Skin color model, Text mining, Semantic web

1. INTRODUCTION

Nowadays, Internet takes a place growingly pivotal in everyday life. The Internet community has been not only in an ever increasing number, but it is also getting increasingly younger. In fact, children find each day an easier access to Internet, which may cause socio-cultural problems. According to a study carried out in May 2000, 60% of the interviewed parents are anxious when their children navigate on internet, in particular because of the presence of adult material. In addition, according to the lookup of Forrester, a company which examines contained of internet, the sum of the sales of pornography on line corresponds to 10% of the total amount of the sales on line. This problem concerns parents as much as companies. For example, the company *Rank Xerox* laid off in October 1999 forty employees who navigate on pornographic sites during their working hours. To avoid this kind of abuse, the company installed program packages to supervise what its employees visit on the Net.

Some companies have proposed solutions to Web site filtering. Their products concentrated on IP-based filtering, and their classification of Web sites is mostly manual. But, as we know, the Web is a highly dynamic information source. Not only do many Web sites appear everyday while others disappear, but site content (including links) is updated frequently. Thus, manual classification and filtering systems are largely impractical. The highly dynamic character of the Web calls for new techniques designed to classify and filter Web sites and URLs automatically.

In this paper, we propose an adult content detection and filtering system (called WebGuard) that extends adult content detection accuracy through both image signature and textual clues of adult material. Compared to other system, WebGuard has the advantage of combining image analyses and text analyses. Image analyses complement text analyses by detecting adult content incorporated inside images.

The remainder of this paper is organized as follows. The WebGuard architecture is presented in Section 2. The extraction of feature vectors from Web pages is reviewed in Section 3. The classification of URLs through Data Mining techniques is discussed in Section 4. Fuzzy clustering and Skin-color image segmentation is presented in Section 5. An experimental evaluation and comparison results are presented in Section 6. Finally Section 7 summarize the WebGuard approach.

2. WEBGUARD ARCHITECTURE

The web filter system (WebGuard) aims to block those sites with pornographic or other nudity, and sexually explicit language. It provides Internet content filtering solutions and Internet blocking of pornography, adult material, and many more categories. The Internet will thus become more controllable and therefore safer for both adults and children.

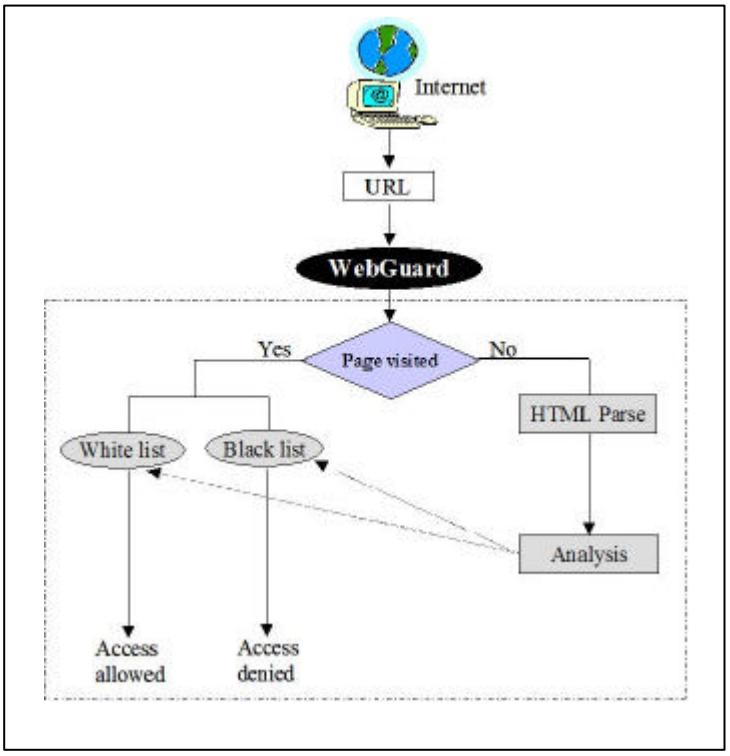


Figure1. WebGuard architecture

The formulation of the « Web Guard » is as follows:
-Fully automated adult content detection and filtering
-Categorization into “black list” (access denied) and “white list” (access allowed) to speed up navigation
-If the site is not recorded on the "black list" or "white list" the engine will then analyses both the visual and textual information and makes a further decision on the sites access allowed/denied status. The black list/white list file is then updated.

In order to rapidly detect and filter the Web pages with adult sexual content in real-time, we must first have some knowledge about adult sexual content, such as suspected URLs, stored in the knowledge base. Hence, our Web Based Audit Content Detection and Filtering System is comprised of two parts. The first

part is designed to create and accurately Update the Knowledge Base (CUKB), the second part is designed to Detect and Filter (D&F) the Web pages with adult sexual content dynamically when younger browsers view them. Figure 2 is the overview of the system architecture.

In CUKB, as show in Figure 3, we have four facilities: the Web Crawler, the Temporary Database, the Data Mining Tools, and the Updating Trigger, used to create and update the Knowledge Base. The Web Crawler is used to periodically search adult sexual images and web pages on the Internet, download suspect images or web pages, put them in the temporary database, and then trigger the Data Mining Tool. The Data Mining Tool uses a data mining method to extract the features of adult sexual images or web pages stored in the temporary database, to discover the suspect URLs, to classify the features, and to trigger Updating Trigger. The Updating Trigger uses predefined strategies to add newly discovered adult sexual content and suspect URLs to the Knowledge Base. To date, we have created the Knowledge Base - and can periodically update it - and have established the fundamentals of our Web based adult content detection and filtering system.

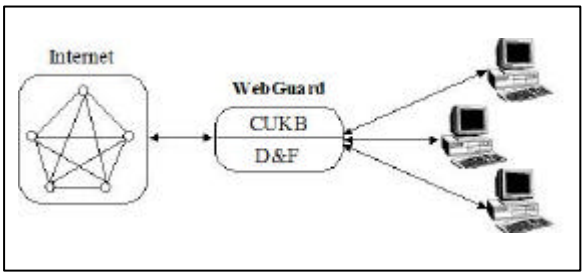


Figure 2. The overview of the system architecture

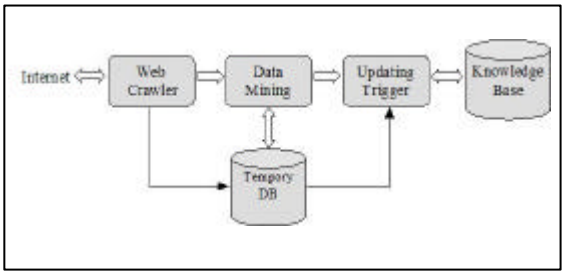


Figure 3. The components of CUKB

In D&F, as shown in Figure 4, we have three facilities to detect and filter browsing activity: the Activity Monitor, the Decision Engine and the Knowledge Base. The Activity Monitor captures active users' URLs in real-time, and compares these URLs with the suspected URLs stored in the Knowledge Base. If such URLs are in the Knowledge Base the Decision Engine is informed. According to the strategies stored in the Knowledge Base, the Decision Engine filters the adult content or disconnects the connection. Apart from classified features and suspected URLs, any anti-browse measures or management information which have been defined by ISPs or generated by the system are also stored in the Knowledge Base.

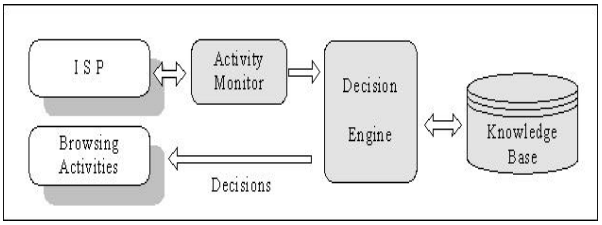


Figure 4. The components of D&F

3. WEB PAGE FEATURE VECTOR EXTRACTION

Before detecting and filtering the URLs with adult content, we need to know which URLs are sex oriented and which are not. This is quintessentially a problem of URL classification.

In order to sort the URLs into two classifications, sex-oriented and non sex-oriented, we first decide which features of a URL can be used as its defining features. Considering many sex-oriented Web pages have picture galleries with little or no text at all, we use both image signature and textual clues as the features of a URL. At the same time, many sex-oriented URLs have some pop-up windows and if a Web page links to another Web page it is possible this Web page also has sexual content. Consequently, the number of pop-up windows on a Web page and the nature of a Web page's links (sex relevant or not) are also important features of an URL. The URLs of many sex-oriented Web sites contain sexually explicit words which is another a

clear indication that the site contains sexual content. To summarize the above, we give the feature vector of a Web site as following:

$$\overrightarrow{VoW} = [bSEW, nWD, nWDwS, nLNK, nLNKwS, nIMG, nIMGwS, nPW]^T$$

Where, bSEW is the flag of whether or not the current URL contains sexually explicit words, nWD is the number of words on the current Web page, nWDwS is the number of sexually explicit words on the current Web page, nLNK is the number of links on the current Web page, nLNKwS is the number of the current Web page's links with adult sexual content, nIMG is the number of images on the current Web page, nIMGwS is the number of the current Web page's images with adult sexual content, nPW is the number of pop-up windows on the current Web page.

Using the Web crawler we create the feature vector $\overrightarrow{VoW}$ of a URL. From the definition of the $\overrightarrow{VoW}$ we can know that in order to set up the feature vector $\overrightarrow{VoW}$ of a URL, we should first decide whether or not the Web pages that this Web page linked to are sex relevant. So, we must traverse the Web site corresponding to this URL and get the leaf-URL of this Web site, then construct the feature vectors of all leaf-URLs, then construct the feature vectors of their parent URLs, after which we construct the feature vectors of their grandparent-URLs. Finally, we set up the feature vector of the given URL. Obviously, it is a process of computing from bottom to top. And, in this process, we used stack as the data structure.

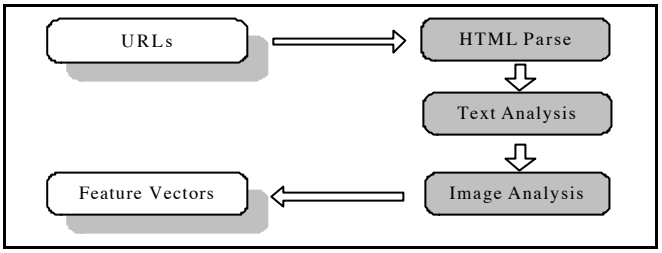


Figure 5. The preparation of feature

As shown in Figure 5, at each step in the computing process, we first parse the HTML, deleting the HTML tags; after that, we analyse the textual content of the HTML, gathering the textual information; and then we analyze the images appearing in the HTML, deciding whether or not they are sex relevant. Finally, based on the obtained information, we create the feature vector of the given URL.

4. USING DATA MINING TECHNIQUES TO CLASSIFY URLS

Once the feature vectors of all the URLs have been constructed, the task is to construct a classifier to classify these URLs into two classes: adult sexual URLs and other URLs.

A number of classification techniques from the statistics and machine learning communities have been proposed [6, 7, 8, 10]. A well-accepted method of classification is the induction of decision trees [2, 6, 10]. A decision tree is a flow-chart-like structure consisting of internal nodes, leaf nodes, and branches. Each internal node represents a decision, or test, on a data attribute, and each outgoing branch corresponds to a possible outcome of the test. Each leaf node represents a class. In order to classify an unlabeled data sample, the classifier tests the attribute values of the sample against the decision tree. A path is traced from the root to a leaf node which holds the class predication for that sample.

Let the set of Web Sites be O

$$C : \Omega \rightarrow \phi = \{\text{ suspect URLs , normal URLs}\}$$
$$W \rightarrow C(w)$$

The observation of C(w) is not easy; therefore we are looking for mean value f to describe class C. The process of graph construction is as follows: We begin with a sample of sites, both suspect URLs and normal URLs and look for the particular attribute which will produce the best partition. We repeat the process for each node of the new partitions. The best partitioning is obtained by maximizing the variation of uncertainty

$\Im_\lambda$ between the current partition and previous partition. As $I_\lambda(S_i)$ is a measure of entropy for partition S_i and $I_\lambda(S_{i+1})$ is the measure of entropy of the following partition S_{i+1}.
The variation of uncertainty is:

$$\Im_\lambda(S_i) = I_\lambda(S_i) - I_\lambda(S_{i-1})$$

For $I_\lambda(S_i)$ we use the quadratic entropy(a) or Shannon entropy(b):

$$I\lambda(Si) = \sum_{j=1}^{K} \frac{n_j}{n}\left(- \sum_{t=1}^{m} \frac{nij+\lambda}{n_j+m\lambda}\left(1 - \frac{nij+\lambda}{ni+m\lambda}\right)\right) \quad (a) \qquad I_\lambda(S_i) = \sum_{j=1}^{K} \frac{n_j}{n}\left(- \sum_{i=1}^{m} \frac{n_{ij}+\lambda}{n_j+m\lambda} \log_2 \frac{n_{ij}+\lambda}{n_j+m\lambda}\right) \quad (b)$$

Where n_{ij} is the number of elements of class I at the node S_j with I ∈{Suspect URLs, Normal URLs}; n_i is the total number of elements of the class i, $n_i = \sum_{j=1}^{k} n_{ij}$; n_j the number of elements of the node S_j $n_j = \sum_{i=1}^{2} n_{ij}$; n is the total number of elements, $n = \sum_{i=1}^{2} n_i$; $m = 2$ is the number of classes {suspect URLs, normal URLs).
As λ is a variable controlling effectiveness of graph construction. The algorithm stops if no changes in uncertainty occur.

In our system "WebGuard" , we use several classification methods (ID3, C4.5, SIPINA) that can be combined in order to ensure a high degree of accuracy. In addition, the user can configure the blocking degree to a level that suites his/her cultural background. Furthermore, the user can protect his/her configuration through a password. Figure 7 shows the configuration interface.

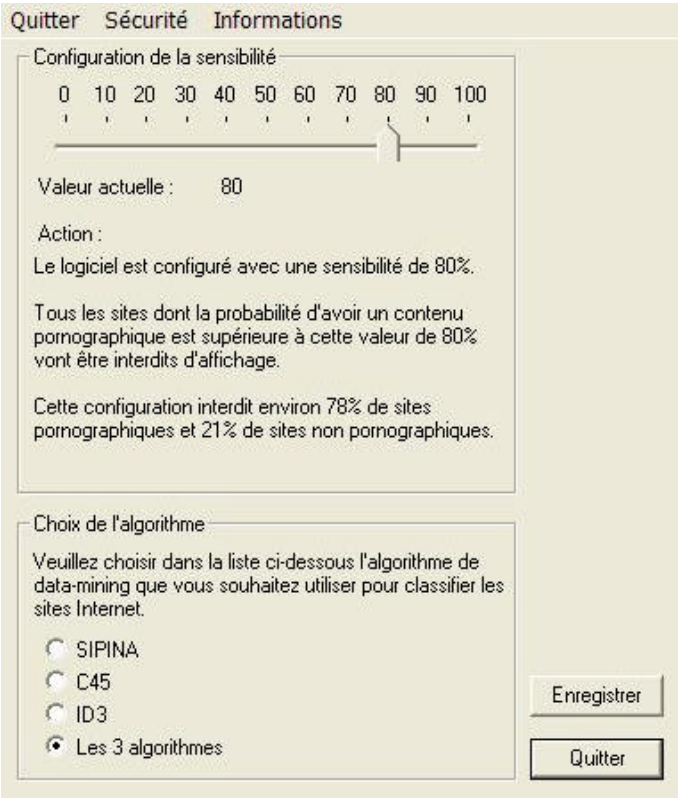


Figure 7. Configuration interface

5. FUZZY CLUSTERING AND SKIN-COLOR IMAGE SEGMENTATION

One of the driving applications for skin-color model construction of large datasets is data-mining. The main goal is to find an original structure of the data, select the most significant part of this structure and describe it with a set of compact decision rules [4,5].

An other important step in the image classification process is color segmentation of the image into skin-color regions and non-skin color regions. Each skin-color image in the database was proceed in the following manner: first similar regions were automatically labeled using a fuzzy clustering, then the skin color regions are extracted and identified by the dedied process.

The fuzzy algorithm detects automatically the number of classes. The local minima of the contrast in the image give a set of seeds regions which are exploited subsequently by an discrete segmentation method such the continuous watershed method. The region growing process of the segment is simulated by linear diffusion, with a diffusion coefficient that depends on local image properties similar to the boundary indicator criteria.

We apply a modified fuzzy c-mean algorithm [1] to classify all pixels in a given image into C classes by minimizing the following objective function:

$$J = \sum_{i=1}^{C} \sum_{j=1}^{N} (u_{ij}) md2(xj, vi) - \alpha \sum_{i=1}^{C} pi \log(pi)$$

u_{ij} is the is the membership value at pixel j in the class i such that $\sum_{i=1}^{C} uij = 1 \forall j \in [0, N]$.

$p_i = \frac{1}{N} \sum_{j=1}^{N} u_{ij}$ is interpreted as the probability of object (pixel) j to belong to the class i.

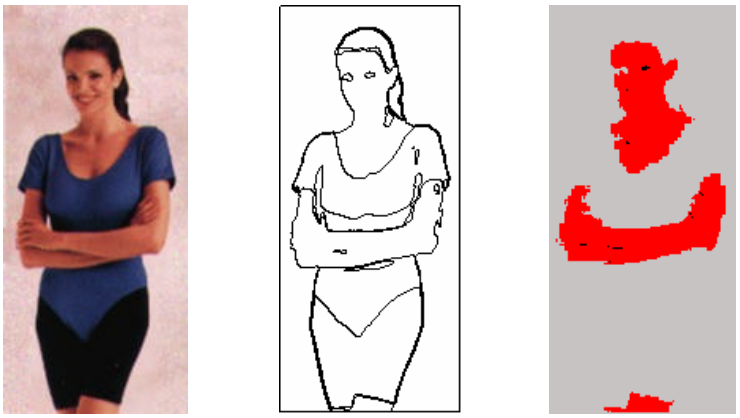
v_i is the centroid of class i, N is the total number of pixels in image, $d^2(x_j, c_i)$ is the standard Euclidian distance and m is a weighting exponent on each fuzzy membership, we take m=2.

In the algorithm the number of classes, C can be known, or determined automatically by choosing a high value of C and eliminating the class i with the smallest probability p_i.

At this step, we have a fuzzy membership function f, which can be considered as multi-valued image $f_i(x_1, x_2): R2 \rightarrow RC$, where $f_i(x_1, x_2)$ is the membership of class i at pixel location (x_1, x_2). This multi-valued image contains contour and regions information. The gradient amplitude of obtains the contour information f defined as follows:

$$\left| \nabla f \right| = \sqrt{\lambda + - \lambda -}$$

$\lambda+, \lambda-$ are the largest, respectively. smallest eigenvalues of the quadratic form associated to f. The local minima of this contrast image give a set of seeds regions Gi, placed nearly symmetrically with respect to the object boundaries. These seeds are classified according to their intensity or color value, and characterized by region information, which is given by mean and variance of each class i : μ_i. Figure 8 shows an example of the automatic segmentation of the image into skin regions.



a. Original image b. Image contour c. Segmented image

Figure 8. Automatic segmentation of a color image into skin-color regions

After using the fuzzy classification and image segmentation, we expurgate areas that contains false skin regions by morphological operations. We aim to filter out as few skin region candidates as possible while losing none of the real skin regions. The output is skin areas candidates for testing and classification .

6. EVALUATION AND COMPARISON RESULTS

Our technique is based on analysing both the textual and visual information to efficiently detect and filter adult content on the WWW. However, there are still several Web sites which could escape our vigilance. These use textual information which does not have a direct relationship to the contents.

In such sites several textual indicators do exist but often these are included within images as in the case of web site "www.france2.com". This makes accurate automatic adult content detection and filtering difficult. In such situations, the analysis requires another phase of text detection and text recognition in an image [9,3].

Most of the existing systems that rely only on URL and textual information can be readily outsmarted by adult content providers. Still, textual information remains the most significant index for fast filtering and will be used as the first stage in the detection and the filtering of contents of adult sites.

When our method of text based analysis is used we are able to filter more than 90% of web sites. To improve the performance we use image analysis which is based on our skin color pixel model. According to the percentage of skin color pixels in the image we can decide if the image is suspect. This increases accuracy to 95%.

We evaluated our technique using textual analysis only (A) and then textual + visual analysis using data-mining based skin-color model (B)[4]. For the purposes of our experiment we used 1000 web sites which were manually classified into adult and non-adult sets. There were 500 non-adult web sites and 500 adult web sites. Table 1 shows the improvement of WebGuard system by the use of image analysis.

Table 1. Performance of adult web site detection with textual and visual information

Method	*Site identified*	*Site no identified*
A	*792 (0.792)*	*208 (0.208)*
B	*988 (0.988)*	*12 (0.012)*

We have also compared the WebGuard with other Web based adult content detection and filtering systems. The comparison chart is shown in Figure 9. The selected systems are Cyber Patrol, Norton Internet Security, Pure Sight, Cyber sitter, Net Nanny, IE (Internet Explorer).

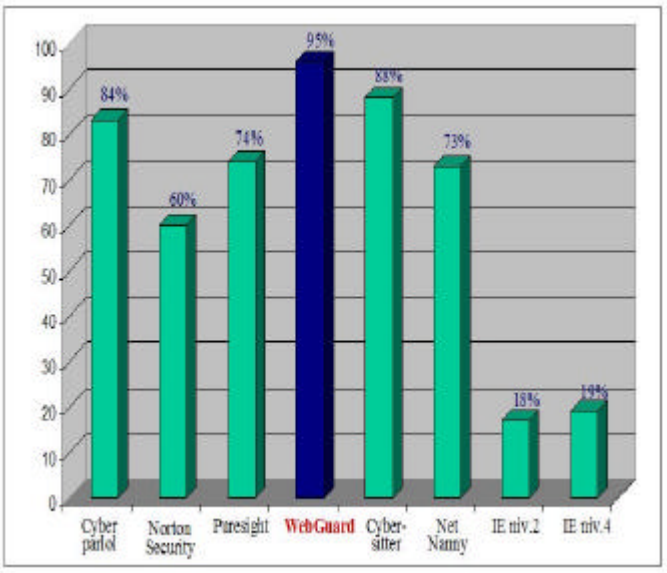


Figure 9. Comparison chart

The comparison has been conducted on 2000 Web sites, including 1000 adult Web sites and 1000 non-adult Web sites. The least effective results come from IE with 18% and 19% success rates while our system is the best with a 95% success rate. Other systems gives success rates between 60% (Norton Security) and 88% (Cyber Nanny).

7. CONCLUSION

In this paper, we have presented the new system WebGuard for detecting and filtering Web pages with adult (image and textual) content in real time. WebGuard combines image and textual analysis with an adjustable sensibility degree. The textual analysis uses several classification approaches that can be combined to give higher accuracy rates. The image analysis combines the fuzzy clustering approach and skin-color segmentation to detect efficiently nude areas. Our experimental evaluation shows the importance of image analyses(i.e., nude area detection) in such systems.

REFERENCES

[1] C. Bezdek, L.O. Hall, L. P. Clarke, “Review of MR image segmentation techniques using pattern recognition”.

[2] L. Breiman, J. Friedman, R. Olshen, and C. Stone, 1984. *Classification of Regression Trees*. Wadsworth.

[3] Y. Chahir et al. , "Détection et extraction automatique de texte dans une vidéo: une approche par morphologie mathématique", MediaNet2002, Ed Hermès, pp :73 - 82 , 2002.

[4] M. Hammami, Y. Chahir, L. Chen, D. Zighed, Janvier 2003. "Détection des régions de couleur de peau dans l’image" revue RIA-ECA vol 17, Ed.Hermès, ISBN 2-7462-0631-5, , pp.219-231.

[5] M. Hammami, L. Chen, D. Zighed, Q. SONG, "Définition d’un modèle de peau et son utilisation pour la classification des images", Ed. Hermès, ISBN 2-7462-0500-9, Juin 2002, pp.186-197.

[6] J. R. Quinlan, 1986. Induction of decision trees. *Machine Learning*, 1:81–106.

[7] J. R. Quinlan, 1993. C4.5: Programs for Machine Learning. Morgan Kaufmann.

[8] S. M. Weiss and C. A. Kulikowski, 1991. Computer Systems that Learn: Classification and Prediction Methods from Statistics,Neural Nets, Machine Learning, and Expert Systems. Morgan Kaufman.

[9] S. Schüpp, Y. Chahir and A. Elmoataz, Extraction d’informations textuelles dans une vidéo par une approche morphologique robuste , International Conference on Vision Interface VI 2003 , Halifax, Canada Canada

[10] D.A.Zighed et R.Rakotomala, 1996. A method for non arborescent induction graphs. Technical report, Laboratory ERIC, University of Lyon2.