© International Journal of Combinatorial Optimization Problems and Informatics, Vol. 7, No. 2, May-Aug 2016, pp. 13-21. ISSN: 2007-1558.

Filter for Web Pornographic Contents Based on Digital Image Processing

Luis Enrique Colmenares-Guillén[1, 2], Francisco Javier Albores Velasco[2]
[1]*Benemérita Universidad Autónoma de Puebla. Facultad de Ciencias de la Computación. Apartado postal J-32 Ciudad Universitaria, Puebla, México.*
lecolme@gmail.com
[2]*Universidad Autónoma de Tlaxcala, Facultad de Ingeniería y Tecnología Calzada Apizaquito s/n. C.P. 90300, Tlaxcala, México.*
Javier.albores@gmail.com

Abstract. In this paper, the problem to be solved is the visualization of undesirable contents on the Internet, both by children and young people. The average age at which a child first watches pornography on the Internet, is 11 years. Part of a solution, is the proposal to filter pornographic content on the Internet based on digital image processing, which is a software tool, designed for Internet users looking for safe navigation. For detecting nudity in digital imaging the RSOR algorithm (Recognition, Selection and Operations Regions) was implemented. When digital images recognized as nudity represent less than 40%; the filter classifies the URL website as secure, and allows the visualization of the site. However, if the percentage is greater than 40%, then, such a website is considered unsafe, and it is redirected to a secure one, which has been predefined by the administrator of the filter.

Keywords: Digital Images, Segmented image, Face Detection.

1 Introduction

Automated processes to detect nudity on digital images are applied in the designing of tools to help preventing access to pornographic material on the Internet. The development of tools for filtering pornographic content on the Internet, allows us to detect producers and distributors of illegal contents which have operated for decades. The Internet pornography industry in the USA generates about 2.3 billion Euros per year worldwide, and about 4 billion Euros. Every day, 2.5 billion pornographic emails —equivalent to 8% of the total emails— circulate on the network. 25% of all searches carried out in browsers are related to pornography. That is 68 million daily searches; 35% out of all Internet downloads is pornographic [1]. Today, nevertheless, this pornographic industry income has been reduced, allowing the generation of free pornographic content that circulates on the network browsers without any security control and it is available to all the Internet users. But most worrisome of all it is the lack of restrictions for children and young people to access these contents through their mobile devices, without any limitation or verification to check whether they are old enough to access them.

In 2012; Tru Research carried out, 2,017 online interviews with young people aged 13-17 and their parents, which revealed the following: 32% of the youths admitted they intentionally accessed nudity or pornographic content online; 43% did so weekly, and only 12% of their parents was aware of these pornographic access of their children [2].

In 2008, YouGov conducted a survey among 1,424 British youth of 14-17 years old, which reflected the following information; 58% of them said they have watched pornography, 71% of those sexually active young people have also watched pornography, and more than a quarter of young men displayed pornography at least once a week, although 5% of them do it every day [3].

According to a Symantec study, after analyzing 3.5 million online searches performed from February 2008 to July 2009, the word "sex" was the fourth most commonly used; and the term "porn" was the sixth. This result was obtained by children who use Norton Family from their homes [4].

Some data published as infographic by Online MBA, shows the following information: 34% of Internet users have been unwillingly exposed to pornography either through pop-ups, misleading links or emails. 116,000 daily searches related to child pornography are performed. The average age at which a child first watches pornography on the Internet is 11.

Internet's concern regarding unlimited access to unwanted content produced the opportunity to propose, as the main objective of this research, to design of a software tool for filtering pornographic content on the Internet; so that children and young people

Received Nov 29, 2015 / Accepted May 21, 2016
Editorial Académica Dragón Azteca (EDITADA.ORG)

cannot access these types of content. This filter is based on digital image processing that implements the RSOR algorithm [5], which scans images and finds a degree of nudity considered unsuitable for displaying.

2 RSOR algorithm development

In recognition of nudity in digital images, the pixels of the images are used as a source of information to determine whether the images represent a potential nudity [6], [7], [8], [9] and [10]. The most relevant pixels are those corresponding to the color of human skin within a color model, which, in this case, it is the HSV model. To narrow the range found in the skin pixels within the model HSV (Hue, Saturation, Value) it was necessary to rely on existing HSV segmentation techniques, and to process the image, segmenting those pixels found within the range corresponding to human skin. In addition; results obtained determined that this filter is useful for detecting pornographic sites, obtaining a 4.7% false positive [5].

2.1 RSOR Algorithm

The RSOR (Recognition, Selection and Operations Regions) algorithm performs the recognition and selection of the largest region within the segmented image [5] besides; it evaluates the information obtained using percentage calculation operations, to decide whether the content of the original image is nudity. The RSOR algorithm consists of three main parts: Recognition of regions, Selecting regions and Operations in the regions.

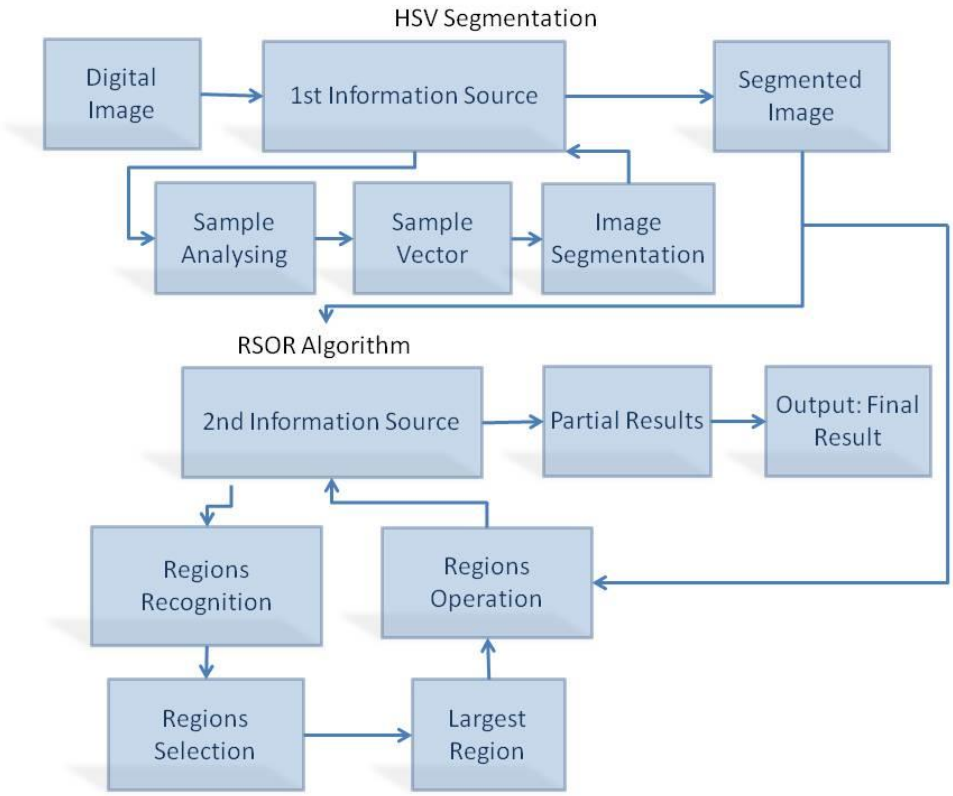

Fig. 1. Block diagram of the digital image processing using the RSOR algorithm.

The diagram on Figure1 shows the process that the digital image will be given to determine whether or not there is nudeness. It begins with HSV segmentation, which identifies the skin pixels contained in a section of various segments of human skin; Caucasian, Asian or African, in order to obtain a range where the largest number of pixels recognized as skin may be. Once the range has been defined, those pixels contained in the original image, which correspond to human skin, are separated. The segmented image is used by the RSOR algorithm wherein the skin regions found are recognized, carrying out a labeling; and subsequently, the selection of the region with the highest content of pixels of skin, defined as Largest Region is carried out, and it will become the second source of information.

2.2 Recognition and selection using regions

In the field of artificial vision, recognition using region refers to techniques aimed at detecting points or regions, light or dark, in digital images. Therefore, the possibility of selecting objects in a binary image it is especially useful. For the latter, two defined functions are used in MatLab to expedite these processes, *bwlabel* = to recognize and label the regions and *bwselect* = to select the regions.

The *bwlabel* function performs labeling of existing components in the segmented image, and can be used to determine how many regions are present in digital images. The function has the following format: *ImageR = bwlabel (ImageS, connectivity)*. Where *ImageR*, is the resulting image, containing regions labeled with the corresponding number to each region; *ImageS* is the segmented image where regions are obtained, and the connectivity is a numeric value which can be either 4 or 8.

The bwselect function allows selecting the region to be segmented just by pinpointing its location within the tagged image. The format of the function is: *ImageC = bwselect (ImageR, r, c).* Where *ImageR*, is the image containing the region to be selected; *ImageR*, is the labeled image containing *r* and *c* regions, representing the coordinates of the region to select [11] and [12].

2.3 Operations in regions

The next step is to conduct operations in the regions detected as skin. First of all, the total number of pixels within the segmented image, which is a binary image, is counted, and the total result of pixels corresponding to skin is:

$$Skin = \sum_{x=0}^{n} \sum_{y=0}^{m} ImageS(x, y) \qquad (1)$$

In equation (1), ImageS represents the *n x n* matrix, which contains the segmented image. Once the number of pixels recognized as skin within the segmented image is obtained, the RSOR algorithm calculates the percentage of the number of pixels in regard to the size of the original image.

The image has the values of the following variables; Size = number of pixels in the original image, Skin leather = number of pixels in segmented image and Percentage = (Skin * 100) / Size. Where Percentage is the value corresponding to the percentage of pixels on skin with respect to the original image, and it proceeds to decide whether there is a nudity under the following criteria: If the percentage> 25%, it represents a nudeness, if the percentage <25, it does not represent a nudeness.

To carry out the selection of the Largest Region, an algorithm, which can be obtained from [7] is implemented. From this algorithm, the value of MaxArea, representing the largest Region area is obtained; as well as the value corresponding to the Region variable, which is a pointer to the largest Region that is used to find its location, using the following MatLab function; [*r, c*] = *find (bwlabel(ImageR) == Region*). Where [r, c] are elements of the matrix containing the coordinates of the region to be selected and *find* is the MatLab function, which returns the location of the Largest Region within ImageR. Once [r, c] values are obtained, the Largest Region is selected and the bwselect function is used with each *r y c* values, so, a new binary image is obtained.

The value of MaxArea is used to obtain the percentage corresponding to the largest region, with respect to the segmented image and the following formula is used*; P_LargestRegion = (Max_Area*100)/Piel.* Where P_LargestRegion, is the percentage corresponding to the Largest Region with respect to the segmented image and it is evaluated on the following criterion; if P_LargestRegion> 35%, it is a nudity, if P_LargestRegion <35%, it is not a nudity.

From the previous process two indicators are obtained; the percentage of skin within the segmented region, and the percentage of skin within the largest region that will be used to decide whether or not the original image depicts a nude, the second criterion, which will be more important to generate the final result; the proportion represents 40% of the two methods.

3 Filter for web pornographic contents on the Internet based on digital image processing

The architecture of the filter for pornographic content on the Internet based on digital image processing (known as Digital ImageFilter or *DigImFilt*) consists of two applications based on client-server model; Client (*DigImFilt) and* Server

(*DigImFiltServices)*. Both applications make the system respond in a relatively quick time interval. The client provides information about each *URL* visited and in turn the server processes each image detected inside the *URL* accessed.

3.1 Add-on

A Complement is understood as an extension or addition of software that runs independently and can be referred to as a complement to browsers, which represents an installable upgrading for the Internet browsers. They are also known as extensions. Complement or add-on, and, they are programs that work only when attached to another and contribute to increase or complement their functionality.

3.2 DigImFilt and DigImFiltServices

DigImFilt, is the browser complement that has references to the services provided by the server in real-time. *DigImFilt* tasks are described by Figure 2.

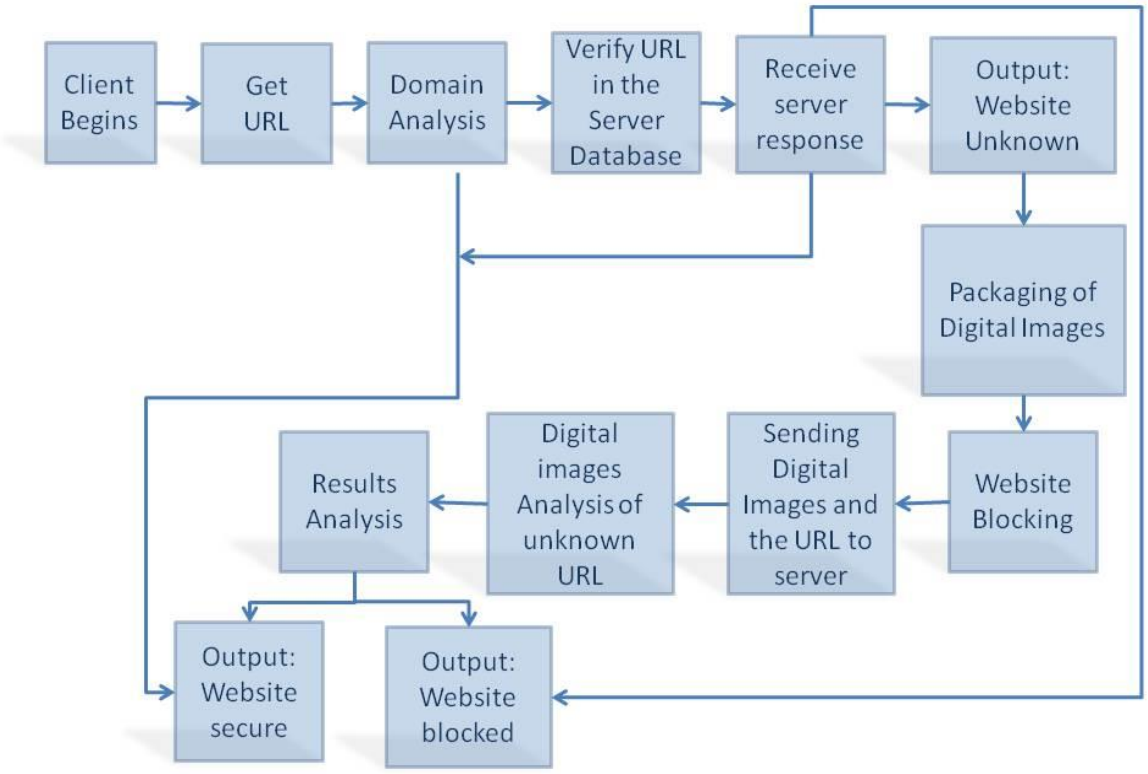

Fig. 2. Block diagram of the Client (DigImFilt).

In the client, block diagram, the steps of the behaviour of the filter by the client are shown. The client starts by running the web browser, in which the URL of a site is entered by the user; the URL is captured by the add-on, in order to be analysed. In this process, the URL is decomposed to analyse the domain extension, which may be insecure sites such as: organizations (.org); government (. Gov); education (.edu); in these cases the pages are displayed directly on the system.

If the domain is not among the safe sites, the URL is sought within the cache (URLs database server). The response can be any of the following states: 0, 1, or -5; in the case of state 0, the website will be displayed directly. If the status is 1, it represents a blocked site and so, it is redirected to a safe one, which has been predefined by the system. Finally, if the state is -5, it indicates this is an unknown URL, this being the case, all the images on the website are packed, and the site remains temporarily blocked to prevent visualizing inadequate information; immediately, the images are sent to the server and they are stored in a folder representing the name of the URL for its analysis within the server. In the server, the results for each digital image are generated

allowing determining whether or not the URL accessed is a safe web site, if it is, it will be assigned the state 0, and its viewing will be permitted without restriction. If it were not, it is considered an insecure website and will be given the state 1; therefore, the website is blocked and then it is redirected to a secure website, predefined by the system administrator.

In Figure 3, the block diagram of the server is displayed. The diagram begins with the running of the system on the server, then, the connection is opened waiting for a client; when a client makes a connection, the URL of the website visited is received within the browser; the URL is verified in the cache of those blocked URLs (URLs database). If there is a URL, the state of such URL is sent to the client; where 1, signifies a blocked website. In case the URL is not registered, the state -5 is sent to the client; the packet of digital images is received from the client immediately in the cache or database of URLs, creating a temporary folder with the ID of the URL, then, the package of images is decompressed to run the digital image processing algorithm RSOR; once the processing of digital images is completed the result is analyzed. If each of the images analyzed contains less than 40% nudity, then the URL of the website is recorded in the URLs cache with the status 0, indicating that the site is secure. If the overall percentage is greater than 40% of the website is registered in the database of URLs with the state 1, indicating that the website visited is unsafe. Finally, the state of the URL is sent to the client to determine whether or not to visualize the website visited.

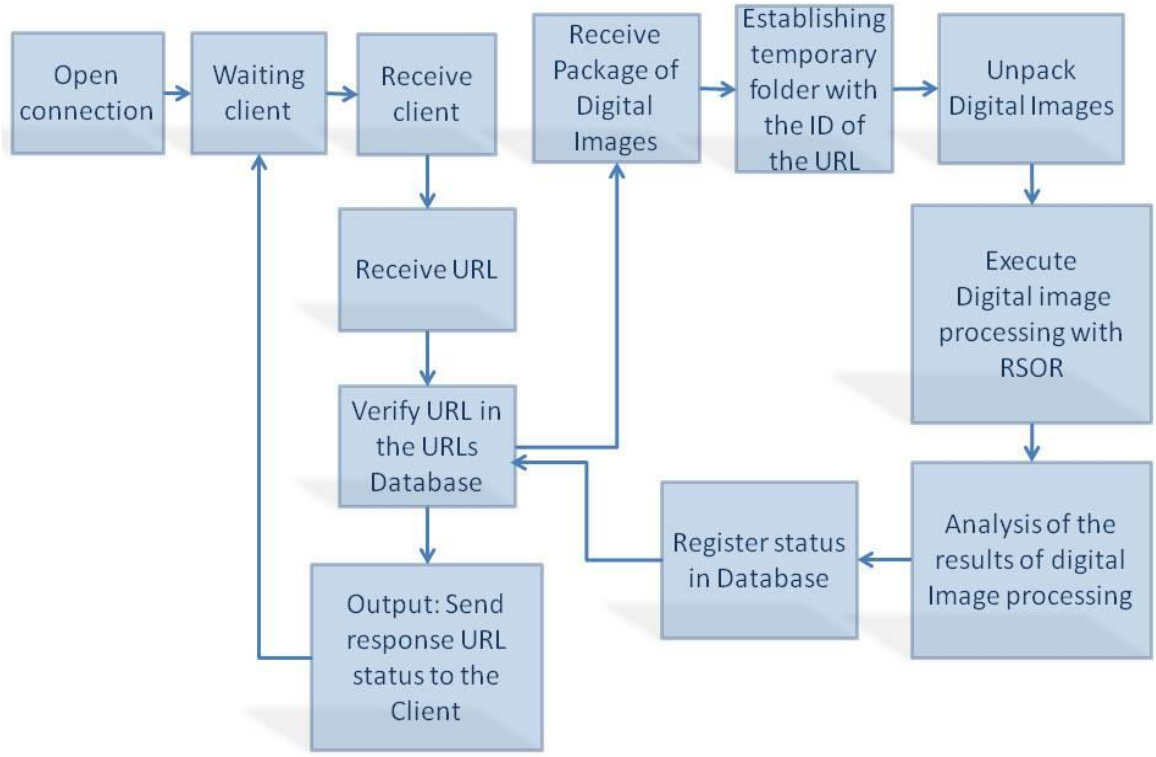

Fig. 3. Block diagram of the Server (DigImFiltServices).

The Client (DigImFilt) and the Server (DigImFiltServices) are implemented under C #, programming language, which is an object-oriented language that is part of the .NET platform, developed by Microsoft and based on the C and C ++ languages; in addition to the fact that the Add-in Express libraries were integrated into the Visual Studio. C # is a language that incorporates features of other programming languages that makes easy the implementation of the filter for pornographic content on the Internet based on digital image processing.

In addition, the framework Windows Communication Foundation (WCF), which is an environment developed for building service-oriented applications, and is designed to facilitate the creation of web services, and web service clients with interoperability in their functions, it is used by the Client.

In the future, it is foreseen using free software, Qt as a development environment, and using the C language with computer vision libraries, which at the present time have OpenCV, aggregating MatLab mathematical functions on a library defined for OpenCV. It will also integrate with a new technology using GPU-CUDA (Graphics Processing Unit- Compute Unified Device Architecture) for faster processing of some filters. Besides, others such as the bilateral Filter, Viola-Jones, can be implemented in the programming language C ++ [14].

The implementation of the RSOR algorithm is carried out in MatLab, using the digital image processing toolbox, consequently minimizing the time of deployment by using scientifically proven, robust algorithms.

The image processing toolbox contains a set of functions of the best known algorithms for binary image processing, geometric transformations, morphology and colour manipulation, which, along with the functions already integrated in MatLab, allow performing image analysis and transformations in the domain of the frequency using the Fourier and Wavelet Transform [13].

4 Experimental results

System testing was performed, where out of 68 websites visited (51.47% were safe sites while 48.53% were unsafe) —all of them unknown to the system— to carry out the statistics shown below.

The system determines the status of the website (1 = insecure, 0 = safe) based on the following criterion:

$$state = 0; \quad if\left(\sum imgDes > \left(\frac{(40 * (\sum imgsS - \sum imgsD))}{100}\right)\right) then\ state = 1; \tag{2}$$

In equation (2), *imgDes*; represents the images detected as nudities by the RSOR algorithm, and *imgsS*, are images enclosed on the *imgsD* website, and they represent the number of images that were discarded by the RSOR algorithm.

The filter for pornographic content on the Internet based on digital image processing, obtained an 8.78% error rate —getting about four unsafe sites as safe ones, while a safe site was considered unsafe— against a 91.22% success rate. Sites where no images to be analysed were obtained, were discarded; in these cases it cannot be determined whether a site is safe or not. In the future, a word analysis is intended for helping to distinguish if a certain site can be related to pornography.

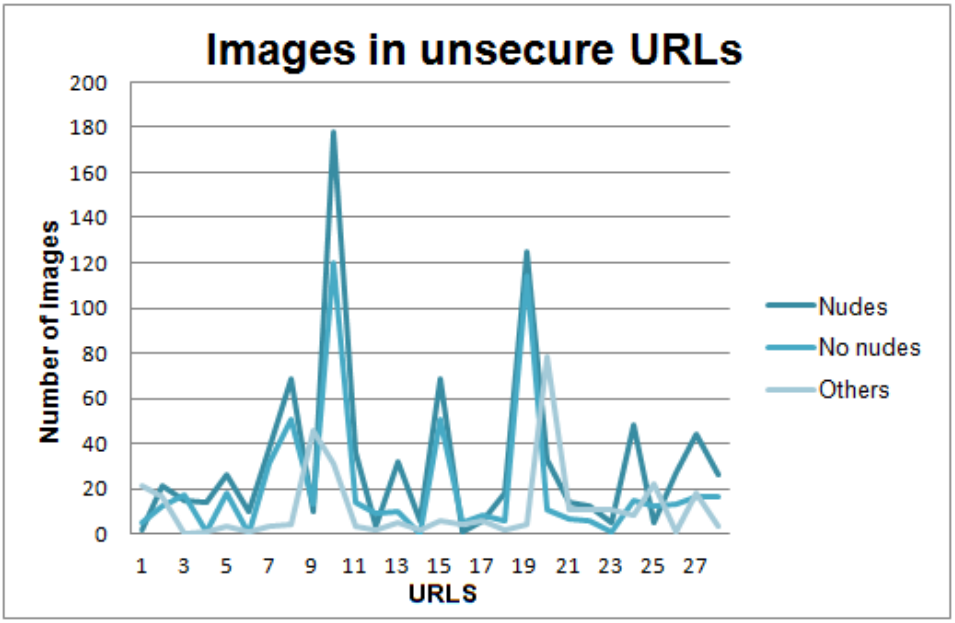

Fig. 4. Three types of images for each insecure site analyzed.

Another point to be considered by the filter for pornographic content on the Internet is that it takes into consideration three types of images from web sites: images with nudes, images without nudes, and other images —of less than 48x48 pixels in size— that are not considered in the analysis. In Figure 4, the number of each of these three types of images, on each unsafe site analysed is represented. It should be mentioned that images of "nudities" on unsafe places, correspond to the highest peaks on the graph. In this case, the numbers of images with greater nudity are considered. Figure 5, shows the corresponding graph of safe sites, where images of "Others" and “No nudes" represent the highest peaks in the graph.

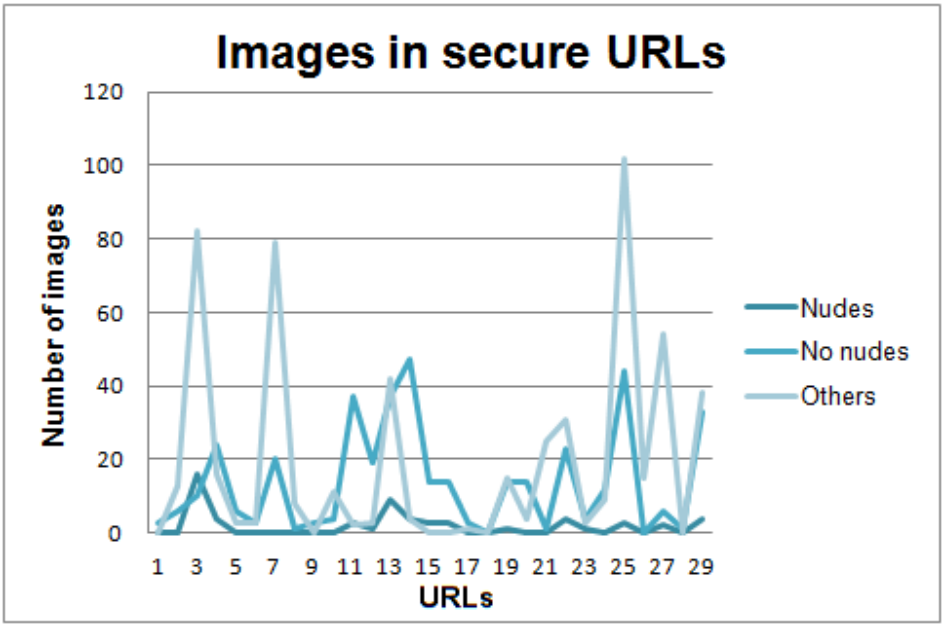

Fig. 5. Three types of images for each secure site analyzed.

On Table 1, the algorithm RSOR is visualized showing that it obtains a low rate of false-positives, compared to other existing systems; therefore, it becomes the best choice for detecting nudity in digital images, thanks to a more efficient detection of those images representing nudity, more details in [5].

Table 1. Comparison of results obtained with existing algorithms.

Algorithm/System	Correct Detection (%)	False-Positives (%)
SVM	86	35
WIPE	97.5	18.4
Jones-Rehg	86.7	9
Fleck-Forsyth	43	4.2
TheNudityDetection Algorithm	94.32	5.98
RSOR Algorithm	92.12	4.7

5 Conclusion and future work

The main objective of the present research is developing a filter for pornographic contents on the Internet, which will become an efficient software tool for filtering pornography. The RSOR algorithm that is implemented for detecting nudity in digital

imaging uses a combination of techniques for digital image processing to determine whether the images on a web site contain any nudity, then, such information is used to determine pornography on a website. The filter is contributing to the following:

1. Innovation in developing a tool, other than the existing ones, based on a filter for digital image processing.
2. A low level of false-positive on the analysis of URLs images in real-time, obtaining a 7.88% error rate, proving to be an efficient tool in filtering Internet pornography.
3. Development and implementation of the RSOR algorithm carried out in a (add-on) for a browser, and it offers support to analyze nudity detection in digital images.

In future work, a new face analyzer of digital images will be integrated in order to identify whether a face is that of an adult or an infant. This new research involves various techniques for analyzing faces, on five areas: face recognition [16], [18], facial expressions [19], gender classification [20], face detection in real time [21] and facial features [17], [22].

The functions that will help us process in GPU-CUDA (Compute Unified Device Architecture) for object detection; Haar on adaptive cascade, for example, face, nose, eyes, mouth and LBP (Local Binary Patterns), utilization of the Histogram of Oriented Gradients (HOG) detector. Some of the advantages of using GPU are that it is highly parallel to hundreds of simple cores [14] and [15].

Finally, the objective of this research is to put up a system to prevent the visualization of pornography on the Internet, an everyday growing problem, for which the Internet is the main source of distribution. The proposed RSOR algorithm and the creation of the filter for pornographic content on the Internet, founded on digital image processing, are the first steps towards the implementation of this system.

Acknowledgments

The support provided by the Dirección de Innovación y Transferencia del Conocimiento de la Benemérita Universidad Autónoma de Puebla, for the development of this project, along with the collaboration of Prof. José Luis Luna Govea for the reviewing of English of this document is appreciated.

Legal disclaimer

This work is a first approach to a patent obtained his title this year 2015, with the name: "Proceso para Detectar Desnudez en Imágenes Digitales" with the request number in the Instituto Mexicano de la Propiedad Intelectual MX /a/2012/003672.

References

1. Pornography Statistics, 250+ facts, quotes, and statistics about pornography use, http: //www.covenanteyes.com/pornography-facts-and-statistics/, 2015 Edition. Accessed November 20, 2015.
2. Jamie Le, The Digital Divide: How the Online Behavior of Teens is Getting Past Parents, McAfee.com. June 2012. http://www.mcafee.com/us/resources/misc/digital-divide-study.pdf. 2012. Accessed October 10, 2015.
3. SexperienceUK, Highlights from YouGov's Sex Education survey, http://sexperienceuk.channel4.com/teen-sex-survey. 2015. Accessed October 9, 2015.
4. [20] BBC News, Kids top searches include Porn, Aug. 12, 2009. http://news.bbc.co.uk/2/hi/technology/8197143.stm. 2009. Accessed October 10, 2015.
5. Tello, Colmenares and Niño: Approach of RSOR algorithm using HSV color model for nude detection in digital images (2011) http://dx.doi.org/10.5539/cis.v4n4p29.
6. Johnson I. A., Lok B., Wong Y. and Da Silva S., Automatic Online Porn Detection and Tracking, in Proceedings of the 12th International Conference on Telecommunications, Publisher with The Institute of Electrical and Electronic Engineers Inc (IEEE), Cape Town, South Africa, pp. 1-7, (ICT 2005).
7. Ap-apid R., An Algorithm for Nudity Detection, in Proceedings of the 5th Philippine Computing Science Congress, Rafael Saldana and Caslon Chua, Editors. Published by the Computing Society of the Philippines (CSP), ISSN 1908-1146 March 4-5, 2005 University of Cebu (Banilad Campus), Cebu City, Philippines, pp. 201-205. (2005).
8. Fleck M., Forsyth D., and Bregler C., Finding Naked People. European Conference on Computer Vision, Cambridge, UK, Vol. 2, pp. 592-602, (1996).
9. Mahjoub M. A., Improved FCM Algorithm applied to Color Image Segmentation. In Canadian Journal on Image Processing & Computer Vision February, ISSN 1923-1717**,** vol. 2, No. 2, pp 16-19, (2011)**.**

10. Forsyth,D.A. & Fleck, M.M Automatic Detection of Human Nudes. International Journal of Computer Vision, Kluwer Academic Publishers Hingham, MA, USA ISSN 0920-5691, Volume 32 Issue 1, Pages 63 – 77, (Aug. 1 1999).
11. García Santillán Iván Danilo. Visión Artificial y Procesamiento Digital de Imágenes usando Matlab, ISBN: 978-9942-01-790-1.Editorial Ibarra. Ecuador, (2008).
12. Computer Vision System Toolbox, http: //www.mathworks.com/products/image/, 2015. Accessed November 20, 2015
13. Cuevas Jimenez Erik Valdemar, Zaldivar Navarro Daniel. Visión por Computador utilizando MATLAB Y el Toolbox de Procesamiento Digital de Imágenes.*2007*, http://es.scribd.com/doc/23371/Procesamiento-de-imagenes-con-Matlab. Accessed november 25, 2015.
14. Shalini Gupta, Shervin Emami, Frank Brill. Digital image processing with GPU, http://on-demand.gputechconf.com/gtc/2013/webinar/opencv-gtc-express-shalini-gupta.pdf, 2013. Accessed November 27, 2015.
15. Manuel Ujaldón Martínez, Procesamiento de imágenes en GPUs mediante CUDA, http://www.fing.edu.uy/inco/cursos/gpgpu/clases2012/Ujaldon1.pdf, 2012. Accessed November 25, 2015.
16. Smith Kelly, Face Recognition, http://www.biometrics.gov/Documents/FaceRec.pdf, 2006. Accessed 28 November 2015.
17. Yaniv Taigman, Ming Yang, Marc' Aurelio Ranzato, Lior Wolf, DeepFace: Closing the Gap to Human-Level Performance in Face Verification. (2015).
18. W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, ACM Computer Survey 35 (4) 399–458, (2003).
19. T. Kanade, J. Cohn, Y. Tian., Comprehensive database for facial expression analysis, in: Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition, pp. 46–53, (2000).
20. J. Hayashi, M. Yasumoto, H. Ito, Y. Niwa, H. Koshimizu, Age and gender estimation from facial image processing, in: Proceedings of the 41st SICE Annual Conference, vol. 1, pp. 13–18, (2002).
21. P. Viola, M.J. Jones, Robust real-time face detection, International Journal of Computer Vision 57 (2) 137–154, (2004).
22. A. Bastanfard, M. Abbasian Nik, M.M. Dehshibi, Iranian face database with age, pose and expression, in: Proceedings of IEEE International Conference on Machine Vision, pp. 50–55, (2007).