Adult Content Filtering
Using Text and Image Analysis

Halidu Sule

Submitted to the
Institute of Graduate Studies and Research
in partial fulfillment of the requirements for the Degree of

Master of Science
in
Electrical and Electronics Engineering

Eastern Mediterranean University
September 2012
Gazimağusa, North Cyprus

Approval of the Institute of Graduate Studies and Research

Prof. Dr. Elvan Yılmaz
Director

I certify that this thesis satisfies the requirements as a thesis for the degree of Master of Science in Electrical and Electronic Engineering.

Assoc. Prof. Dr. Aykut Hocanın
Chair, Department of Electrical and Electronic Engineering

We certify that we have read this thesis and that in our opinion it is fully adequate in scope and quality as a thesis for the degree of Master of Science in Electrical and Electronic Engineering.

Assoc. Prof. Dr. Erhan A. İnce
Supervisor

Examining Committee

1. Assoc. Prof. Dr. Cem Ergün
2. Assoc. Prof. Dr. Erhan A. İnce
3. Assoc. Prof. Dr. Hasan Demirel

ABSTRACT

The working principle of the Internet is such that anyone who sets up a server computer and connects it to the local area network in their neighborhood becomes equipped to shear with the world any type of information they deem appropriate. Generally, some of this information dispatched is not appropriate for viewing of our children and some steps should be taken to help the society so that classification and controlled access become possible.

Throughout this thesis, we designed and implemented a text and image based web-page filtering system that makes use of web page parsing, HTML tags removal and string in string search procedures along with various other criteria for processing images downloaded from a web site using a custom written JAVA program.

For the text, there are some words and phrases that are common to pornographic sites and are rarely seen in regular sites. To find out such words and phrases, a survey was done on a number of sites. With the words and phrases determined, our expectation is that any site which may contain pornographic oriented text will have in it some of these words and phrases. Hence, once the tested web page was parsed and the pure text string was obtained from the downloaded HTML code the string would be searched for the type of words and phrases previously determined and final decision would be made based on the frequency of words detected.

From literature survey, everyone seems to agree that pornographic images have too much skin exposure which is why detecting skin is generally the starting point. To find out the amount of skin in an image, improved YC_bC_r color segmentation was implemented. The improved YC_bC_r segmentation would satisfactorily segment out the skin from the other regions but some skin like objects would still be falsely detected. Therefore, texture property was used to differentiate bearing in mind that skin is generally smooth and most others textures aren't (many are more coarse). In order to classify a web site from which images have been extracted through the help of a JAVA program, criteria such as face detection, lacunarity, edge sum, uniformity, entropy and percentage of skin region have been employed and when three or more of the criteria were met this was taken as an indication for containing adult nature material. Final decision was made by computing percentages for the results obtained for both the text and image analysis and comparing the average of the two to some previously selected threshold ranges.

For the five randomly selected adult content containing web sites that were used for test purposes the text analysis always gave 95-100% accuracy and the image analysis resulted in 56.83, 54.83, 52.63, 57.14, 66.67 % accuracy respectively for sites 1-5 as detailed in chapter 5. In chapter five it was also shown how the two results (text and image analysis) can be combined to get an average percentage. For the five different web sites considered the lowest average percentage obtained was 73.82%.

Keywords: HTML parsing, skin color segmentation, texture analysis, lacunarity.

ÖZ

İnternetin çalışma prensipleri, bir bilgisayarı server olarak kullanıp komşuluğundaki yerel ağ bağlantısına bağlayan herkesin uygun gördüğü her türlü bilgiyi dünya ile paylaşmak için gerekli donanıma sahip olacağı bir ortam oluşturmaktadır. Genel olarak çocuklarımızın paylaşılan bu bilgilerin bir kısmına erişimleri uygun olmayıp sınıflandırma ve kontrollü erişimin sağlanması amacıyla topluma yardımcı olmak adına bazı çalışmaların yapılması gerekmektedir.

Bu tez çalışmasında, geliştirilen bir JAVA programı sayesinde bir web sitesinden indirilen görüntülerin işlenmesi için çeşitli diğer kriterlerin yanında web sitesi ayrıştırılması, HTML etiketlerinin kaldırılması ve digi içinde dizgi araştırma prosedürlerini uygulayan metin ve görüntü bazlı bir web site filtreleme sistemi geliştirilmiştir.

Metin ile ilgili olarak genellikle normal sitelerde nadiren görülen ve pornografik siteler arasında ortak olan bazı kelime ve terimler bulunmaktadır. Bu kelime ve terimlerin belirlenmesi ve saptanması amacıyla birkaç site üzerinde bir anket çalışması yapılmıştır. Belirlenen kelime ve terimlerden yola çıkılarak, beklentimiz pornografik odaklı metinleri içeren sitelerde bu kelime veya terimlerin bazılarının bulunacağı yönündedir. Dolaysıyla test edilen web sitesinin ayrıştırılıp indirilen HTML kodlarından saf metin dizelerinin elde edilmesinden sonra bu dizeler daha önceden belirlenen kelime ve terimler açısından

araştırılacak olup nihai kararlar belirlenen kelimelerin kullanım sıklıkları dikkate alınarak verilecektir.

Literatür çalışmasından, herkesin pornografik içerikli sitelerde yüksek deri gösterim oranlarının bulunduğu yönünde hem fikir olduğu belirlenmiş olup bu gerçek ise cilt belirlemesinin bir başlangıç noktası olarak kabul edilmesinin nedenini oluşturmaktadır. Bir görüntüdeki cilt oranının belirlenmesi için geliştirilmiş YC_bC_r renk ayrıştırmama algoritması uygulanmıştır. Bu yöntem cildin diğer kısımlardan ayırt edilmesinde iyi sonuçlar doğurmuş olup ancak cilt ile benzer özelliklere sahip olan bazı diğer kısımlar da yanlıkşlıkla ayırt edilmiştir. Dolaysıyla cildin genellikle diğer dokuların bir çoğu ile kıyas ile daha yumuşak olduğu (birçoğu daha kabadır) izleniminin göz önünde bulundurulması amacıyla doku özelliklerinden yararlanılmıştır. Bir JAVA programından yararlanılarak görüntülerin çıkarıldığı bir web sitesinin sınıflandırılması amacıyla yüz tanıma, lakunarite, kenar toplamları, tekdüzelik, cilt alanı entropi ve yüzdesi gibi bazı kriterler dikkate alınmış olup bu kriterlerin en az üçünün sağlandığı durumlarda yetişkenlere özel içeriklerin bulunduğu yönünde bir işaret olarak kabul edilmiştir. Nihai kararlar hem metin hem de görüntü analizlerinden elde edilen sonuçların yüzdelerinin hesaplanması ve bu iki faktörün ortalamasının daha önceden belirlenen bir eşik değer ile karşılaştırılması sonucunda verilmiştir.

Yetişkinlere özel içeriklere sahip olup test amacıyla kullanılan ve gelişigüzel bir şekilde seçilen beş web sitesi için metin analizleri her zaman 95-100% oranında doğruluk göstermiş olup görüntü analizleri ise 1-5 olarak adlandırılan ve 5.bölümde detaylı bir şekilde açıklanan web siteleri için sırasıyla 56.83, 54.83, 52.63, 57.14 ve 66.67% olarak sonuçlanmıştır. Beşinci bölümde ayrıca ortalama

bir yüzde oranının elde edilmesi için zikredilen iki sonucun (metin ve görüntü analizleri) nasıl birleştirilebileceği belirtilmiştir. Dikkate alınan beş farklı web sitesi için elde edilen en düşük ortalama yüzde oranı 73.82% olarak bulunmuştur.

Anahtar Kelimeler: HTML ayrıştırması, deri rengi ayrıştırması, doku analizi, lakunarite

To My Family & everyone who have aided me in one way or another to get to where I am now

ACKNOWLEDGMENT

I would like to thank Assoc. Prof. Dr. Erhan A. İnce for his continuous support and guidance in the preparation of this study. Without him, all my efforts would have been miss guided.

Assoc. Prof. Dr. Hasan Demirel is the vice chairman at the Department of Electrical and Electronics Engineering, helped me with various issues during the thesis and I am grateful to him. Also worthy of mention is the inadvertent support I got from a number of friends which I am very appreciative of.

I owe quit a lot to my family who allowed me to travel all the way from Nigeria to Cyprus and supported me all throughout my studies. I would like to dedicate this study to them as an indication of their significance in this study as well as in my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

Θ	Threshold of skin probability
|	Such that
| x |	Absolute x

ST	Skin Texture
T	Text analysis
T_n	Total none-skin pixels
T_s	Total skin pixels
TPAS	Total Pixels Attributed to Skin
TSL	Tint Saturation Luminance
URL	Uniform Resource Locator
WWW	World Wide Web
YIQ	Luma In-phase Quadrature
YC_bC_r	Intensity, Chrominance-blue, Chrominance-red

Chapter 1

INTRODUCTION

1.1 Aim

Lots of innovations have been made throughout time. In recent times, the bulk of the advancements were felt in technological aspect of human existence. To a large extent, it is right to say that all such innovations have their demerits even though they were intended to better the lives of people. The “watch word” in this thesis is controlled application since we know that if such technologies are not used properly, the aim will be defeated.

There is one such innovation that the aforementioned is directly applicable to and that is the “Internet”. The Internet has been in existence for quite a while now and its importance can’t be over emphasized. Since the Internet is built up of interconnected servers, a client computer would request the content of the site accessed from an unknown and/or known server and it will then acquire and display the information if the hand shake is successful. The downloaded information at times may contain some material that may make it undesirable for some audience. By common knowledge, content of web-sites can be accepted as desirable if more people are in favor of it and not desirable if the reverse is the case.

In the case of children viewing the Internet, there are some contents that could be considered undesirable. These include the following: brutal graphic images, games that can manipulate them to act in an un-desired manner, sites that can feed them with terrible information (i.e. terrorism), pornographic sites etc. Apparently, there are levels of tolerance and right now web servers consider the age of the child involved when trying to decide.

Classification of web sites is essential to many tasks in web information retrieval as mentioned by Xiaoguang Qi and Brian D. Davison [1]. They went ahead to state that maintaing web directories and focused crawling as perfect examples where web site clasification is applicable. Going even futher, they also mentioned that the uncontrolled nature of web content poses additional challenges to web page classification therefore, we can not understand what kind of content a website contains just by looking at its URL adress. Like in the URL of a pornograthic site, most have “xxx” as part of their domain names and usually domain name for such web sites end in “.com”. If this by itself serves as our bases for site filtering, it will allow some porn sites that should not pass due to a small variation in the sub-domain name (i.e. xxnx). Therefore, traditional text classification is required and here it is combined with image analysis to checkmate websites that have very little text content but lots of images.

This thesis used both MATLAB and JAVA programs to evaluate and classify any chosen website by carrying out text and image based content analysis. This is very relevant because many studies have estimated that of the world's 42 million websites, 12% contain adult content. Another statistical result estimated 70% of teens, for instance, have inadvertently come across pornography on the web. This

is alarming and the numbers are expected to go up since the age at which children are exposed to computers is reducing. In some cases, children can't read but they already know how to navigate to websites at a very tender age. Clearly, this necessitates the urgency of such a research.

1.2 Literature Survey

There are many web sites on the World Wide Web that are not suitable for children's viewing. But some of this "adult only sites", ask only for a confirmation that the user is over 18 without any more proof than having the visitor select "Enter" or "Exit". This means that a child can get to any site that is not black listed by their parents. There are several ways to block websites. One such method is by setting browser to abort any request to open a website by an administrator. Generally the browser history is checked since it keeps track of sites visited and restrictions can be put in for future viewings. But this means that the child must have already visited the site. Also the problem with this method is that the child might be knowledgeable enough to delete the browser's history. This is not so efficient so a method where information from the site will not be seen but processed internally and a decision on whether or not the site is okay can be made is required and this is exactly what researchers in this area have been trying to do for some time now.

In order to classify a web site as fit for all to view or fit for just adults, a couple of techniques can be adopted. These methods have been implemented on text, images and videos obtained from the web. Lots of research has been carried out where researchers have applied their strategies to one, two or all of these three recourses (if we consider the extract from the supposed website to be a resource).

In this work, two of the recourses were used “text” and “images”. Having downloaded the content of a home page, text and image analysis were carried out on them borrowing ideas from Xaiming, Xiaodong and Lihua [2]. In [2] the authors had tried to detect adult images by considering color, texture and geometrical features of an image. The idea that was implement in the research paper which is adopted here was, color filtering the image to determine candidate skin regions, then the coarse degree of pixels of candidate skin regions was calculated for each pixel and lastly, fractal dimension of all the rest big enough skin regions was calculated and after a couple of iterations, a threshold was picked to use in decision making. Having implemented this method with a few inclusions, the result were (by my observation) not good enough since the amount in error was a little too much to be ignored. This is not to say that the method is not good since ideas mentioned in it were very helpful in understanding texture properties of skin and skin like objects in an image.

It appears that more researchers in this topic emphasized on making decision based on pictures and video analysis and evidence to this is the availability of lots of research papers that have discussed these topics in those regard. Forsyth and Fleck [3] proposed an automatic system which helps decide whether or not human nude is present in an image. Their system marks skin-like pixels by the use of color and texture properties (which are similar to the method implemented here) after which the skin regions are then fed to a specialized grouper. The grouper then attempts to group a human figure using geometric constraints on human structure and then, if the grouper is able to match it to a predefined structure, the system decides a human nude is present. Still on detection of nude images; in 1999 came a research conducted by Jones & Regh [4]. This research

has been considered mind blowing by many because of the thing they were able to achieve. In their work, two three dimensional histograms were produced. One for skin and another for non-skin pixels, representing one color channel in each dimension. Each labeled pixel was put in the correct bin in the correct histogram. Upon dividing each bin with the total number of pixels in each histogram correspondingly, the conditional probability histograms equation (1.1) and (1.2) were presented. Where (rgb) denotes pixel value, s and n shortens skin and non skin, and T_s and T_n is the total amount of pixels in all histogram bins respectively.

1.1

1.2

From these histograms a Bayes classifier (which is not intelligent enough) was produced as could be seen in equation (1.3). With

considers the naked body which is composed by trunk, limb and face as the object to be recognized. Body is taken to be a combination of predefined key rectangles. If this is present and lots of skin detection is made by forward propagation neural network. Yue Wang [6] proposed the combination of Ada-boost algorithm which means rapid speed in object detection and the robustness of nipple features for adaptive nipple detection. It was able to achieve this by locating nipple-like region followed by detection. There are numerous adult images with no nipple exposure suggesting weakness in the result that will was obtained. Wonil Kim [7] proposed a neural network based adult image classification where HSV color model is used for the input images for the purpose of discriminating elements that are not human skin, then the image is filtered using by checking how much large the exposed skin is. Ours improved skin segmentation method not only did this but included some other criteria for obtaining better result.

Text analyses have not been much of interest to researcher. Evidence to this observation is the fact that all the research papers mentioned so far which happen to have done a good job with this topic, did not mention it. The reason is not so far from the obvious fact that such sites do not contain much text. This is not to say it is not relevant. As a matter of fact, it is since such sites contain more of pictures and videos where they are always with their titles, the developers or managers of the site always want the title to say a lot about any picture or video that is to be viewed. Such titles need to be straight to the point because in most cases, people are required to pay to view or copy it. Hence, this thesis took advantage of the available resource to make better judgment. The text analysis method implemented here is new to this field of research.

1.3 Description of Work Carried Out

Many responsible parents are using expensive Internet filtering software's to protect their children from accidentally accessing sites with adult content. This thesis entails a long and careful research to develop a system that will classify web sites. Once the user specifies a web address, the process begins. First, the image and text contents are downloaded from the address provided. As the images are downloaded and saved to a folder for future analysis parallel to this text based content analysis will be carried out from the parsed HTML code and the text wise classification will be finalized. Following the download of the set of images, a MATLAB program is invoked which handles the images analysis.

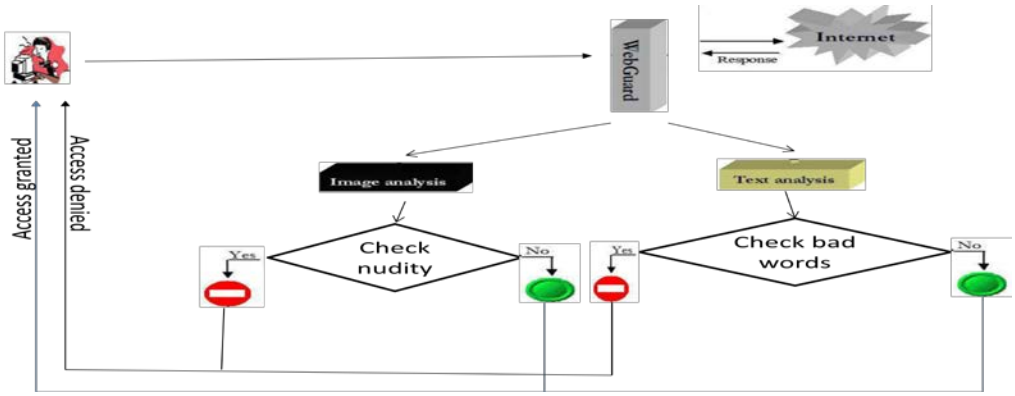


Figure 1.1: System Implementation.

Figure 1 shows the processing steps for the web site that the user is trying to access.

Before the analysis step a web crawler is employed to extract sample images from various links of an http address. The image analysis that follows makes use of both skin color filtering and texture filtering concepts. In addition, lacunarity

(ratio of the second and first moments of pixel intensity distribution) and face detection in the candidate skin regions will be carried out to distinguish some scenery pictures which are difficult to distinguish using normal features (i.e. deserts or beaches which all have similar color and texture with real human skin).

This thesis will provide a comparison between what was obtained from the proposed approach and results that were obtained by other researchers.

1.4 Thesis Organization

In Chapter 1 provides a general overview for web page filtering. This is then followed by a summary of the literature survey and a paragraph about how the thesis is organized. Chapter 2 highlights segmentation for both skin and non-skin regions. It introduces different color models and gives details about how each can be made use of. Comparison among the different color space based results is also provided. In chapter 3, the criteria which were used to classify the images downloaded from a site were explained. Chapter 4 carries on by presenting the JAVA program developed for web page parsing, text search, and automatic downloading of images. The results obtained from simulation were presented in Chapter 5 and lastly Chapter 6 provides conclusions and gives some directions for future work.

Chapter 2

SKIN COLOR SEGMENTATION

The need for better image interpretation gave rise to image processing. So far, a lot has been achieved in this field which is credited to research motivated by huge market demand for products such as computers, mobile phones and IP cameras that incorporates image processing ideas. Since these products are constantly improved and made cheaper more and more pictures are available for sharing on the WWW. Also web developers can get pictures easily and hence web sites without pictures are very rare.

The concern in this thesis is about digital images and how it can be manipulated to obtain information about its content. The information obtained will serve as the basis for decision making to whether or not the image contains lots of exposed skin. How much can be considered as “excess skin exposure” is subject to debate. Segmenting out the skin color like regions and making a decision as to when the pictures contain nudity or deciding if these skin parts are not sufficient to say that the image contains nudity was a vital part of this research. Every procedure that was involved was aimed at harnessing the level of correctness of this decision.

While trying to decide on the content of an image, color property was the most important among the different information pieces considered. But this alone was

not strong enough for making a trustworthy verdict since it is possible for some other images to appear to contain skin areas when in reality they do not. In the field of image processing, such images are called false positives.

To avoid having many of these false positives our approach was to group the images as follows;

- Nude images
- Standard everyday type images
- Beach and desert images
- Lion images

These grouping are not just a coincident but a logical and intelligent selection as could be seen in Xaiming, Xiaodong and Lihua [2]. The rationale behind this is to come up with different criteria that would be satisfied more often by images in the nude images group than the others.

Since color property implementation alone is not good enough, texture analysis will also be exploited later in chapter 3. For this reason, the portions of the image that are detected as skin after color segmentation are called candidate skin regions as illustrated in figure 2.1.

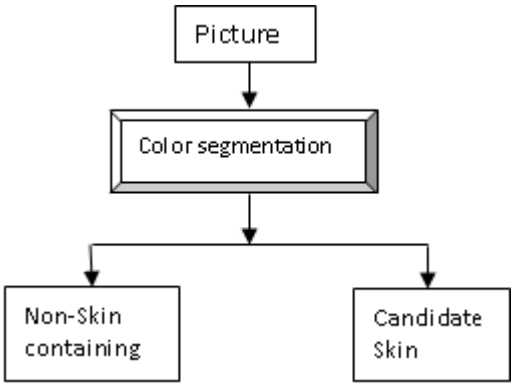


Figure 2.1: Color Segmentation Procedure.

2.1 Image Manipulation

Image processing is pretty vital to this thesis. Image or picture processing is basically matrix manipulation in 2 or 3 dimensions. When the image is in a computer, adjustments could be done by transposing a grid on the image. This grid forms boundaries for small or tiny squares called pixels. The value of each pixel is averaged so that each will represent one digital value. Each value used in a digital image can also be assigned a number therefore making it an array of integer values. The horizontal rows of pixels are called lines and the vertical columns are called samples. For example, pixel in the top left corner in the array is line 1, sample 1. The image could be viewed as a whole, within a neighborhood or pixel wise. The last two are more important to this work since the color and texture property (which will be seen later) of skin are better exploited in these forms.

2.2 Color Property

The values of the intersection of a lines and samples contain the color information in colored images. In the different color models, it usually contains 3 layers of equal line and sample size. Corresponding lines and samples intercept for each layer is varied to give the needed color. Therefore operation could be

performed pixel wise. Color is a very important feature of an image thus it was used in determining skin regions. There are lots of color models, some of which are derived from others. The following are some of the widely known color models:

- Normalized RGB
- HSI, HSV, HSL (Fleck HSV)
- TSL
- YC_bC_r
- Perceptually uniform colors (CIELAB, CIELUV)
- Others (YES, YUV, YIQ, CIE-xyz)

2.2.1 RGB Color Model

The RGB color model is an additive color model in which the primary colors red, green, and blue light are added together in various ways to reproduce a broad array of colors. The name comes from the initials of the three colors Red, Green, and Blue and black is simply the absence of light. The RGB color model is shown in figure 2.2.

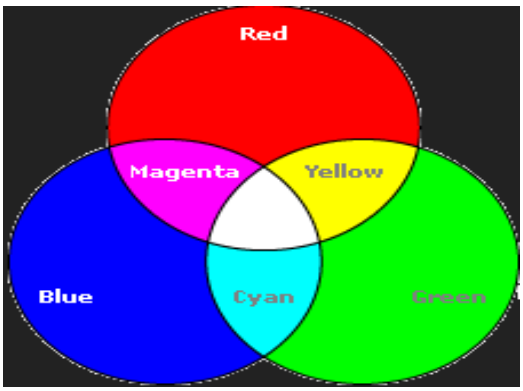


Figure 2.2: RGB Color Model [8]

The easiest model to work with is the RGB model and this is because the related operations are linear. Each layer that is part of the combination can be treated separately. An example of what it looks like is seen in figure 2.3. In the image, the composite slice shows what the image looks like to the eyes. The other three layers are primary color images which show the concentration of each primary color. For these, the deeper the color (red, green or blue) the more of that color the pixel equivalent has in the composite image and the other colors are close to zero combination. Precisely speaking, say a pixel appears to look really red in the composite image, when it is separated to its constituent color we will see something similar to R: 255, G: 10 and B: 5.

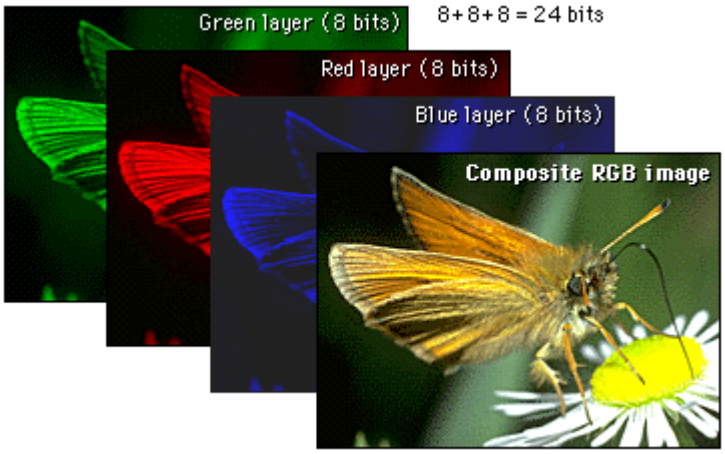


Figure 2.3: RGB layers of an insect's image and its composite RGB image [9]

In this model, we have seen that the primary colors are red, green, and blue and that it is an additive model, in which colors are produced by adding components, with white having all colors present and black being the absence of any color. In as much as working in RGB is rewarding due to speed as regards performance and easy programming requirement. It is actually not so good compared to other color models since it doesn't consider intensity separately. This is very important and not to be ignored because pictures on web sites that will be worked with are

taken under varied light intensities. For this reason, it is not advisable to work in RGB also because computation time is not necessarily better.

2.2.2 HSI Color Model

Another color model which is at our disposal is the HSI model. The letters are abbreviations for Hue, Saturation and Intensity or sometime referred to as HSV this time the V stands for value which is same as I, while H and S maintain their original meanings. HSI is related to the RGB model via the set of equations 2.1, 2.2 and 2.3.

2.1

2.2

2.3

The HSI model addresses the laps that is present in the RGB model but the result is not still as good enough as YC_bC_r model when we compare result that are obtained using the two color spaces with the following intervals adopted from a research paper.

2.4

2.5

2.6

In one of the research papers that used this range of values captured in equation 2.4, 2.5 and 2.6, they had obtained the values by carefully observing figure 2.4.

Here, H, S and V are in the range 0 to 1 as specified in the same paper by Jorge, Gualberto, Gabriel, Linda, Héctor and Enrique [10].

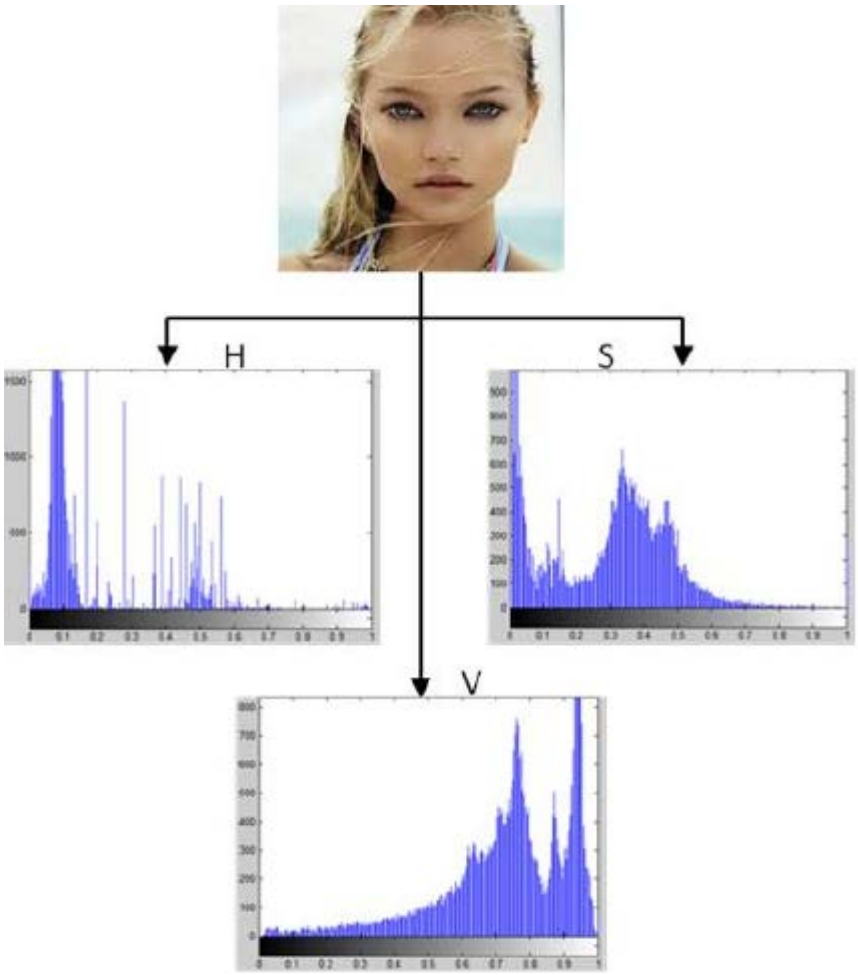


Figure 2.4: Histogram plot of HSV model [10]

Figure 2.4 shows a properly illuminated image of a human face and histogram plot of its corresponding values in HSV color space. The plots in the image gave clues for deciding on the range of values where skin pixels fall (specified in the equations 2.4 through to 2.6). This was done by picking numbers within the boundary that contains values that display almost all of the skin pixels when specified. Very importantly, [8] did not use only the values from this image but performed the same experiment for different other skin tones before arriving at the specified ranges.

2.2.3 YC_bC_r Color Model

A third and final color model that was experimented in this thesis was the YC_bC_r. It was chosen and used in the color segmentation solely because it gave better results when compared with other color spaces. The RGB color space is the default color space for most available image formats and as was the case for HIS, YC_bC_r can also be obtained from it via a transformation. The color space transformation is assumed to decrease the overlap between skin and non-skin pixels which in turn, makes the process robust thereby aiding skin-pixel classification under a wide range of illumination conditions.

YC_bC_r is an encoded nonlinear RGB signal, commonly used by European television studios and for image compression works. Here, the color is represented by luma (which is luminance or brightness), computed from nonlinear RGB constructed, as a weighted sum of the RGB values and two color difference values C_b and C_r that are formed by subtracting the luma value from red and blue components of RGB model.

2.7

2.8

2.9

This model is intended for use under strictly defined conditions within closed systems. The Y component describes brightness and the other two values describe a color difference rather than a color, making the color space unintuitive. The transformation simplicity and explicit separation of luminance and chrominance components makes this color space attractive for skin color modeling. The YC_bC_r color model was developed as part of the ITU-R

Recommendation B.T.601 for digital video standards and television transmissions. It is a scaled and offset version of the YUV. In YC_bC_r the RGB components are separated into luminance (Y), chrominance blue (C_b) and chrominance red (C_r). The transformation used to convert from RGB to YC_bC_r color space is shown in the equation as thus;

2.10

Since this color model is luma independent, it is a better choice when trying to compare with the RGB model. The cluster region is after constructing a histogram (very similar to Figure 2.4) of the different component is given as thus:

2.10

2.11

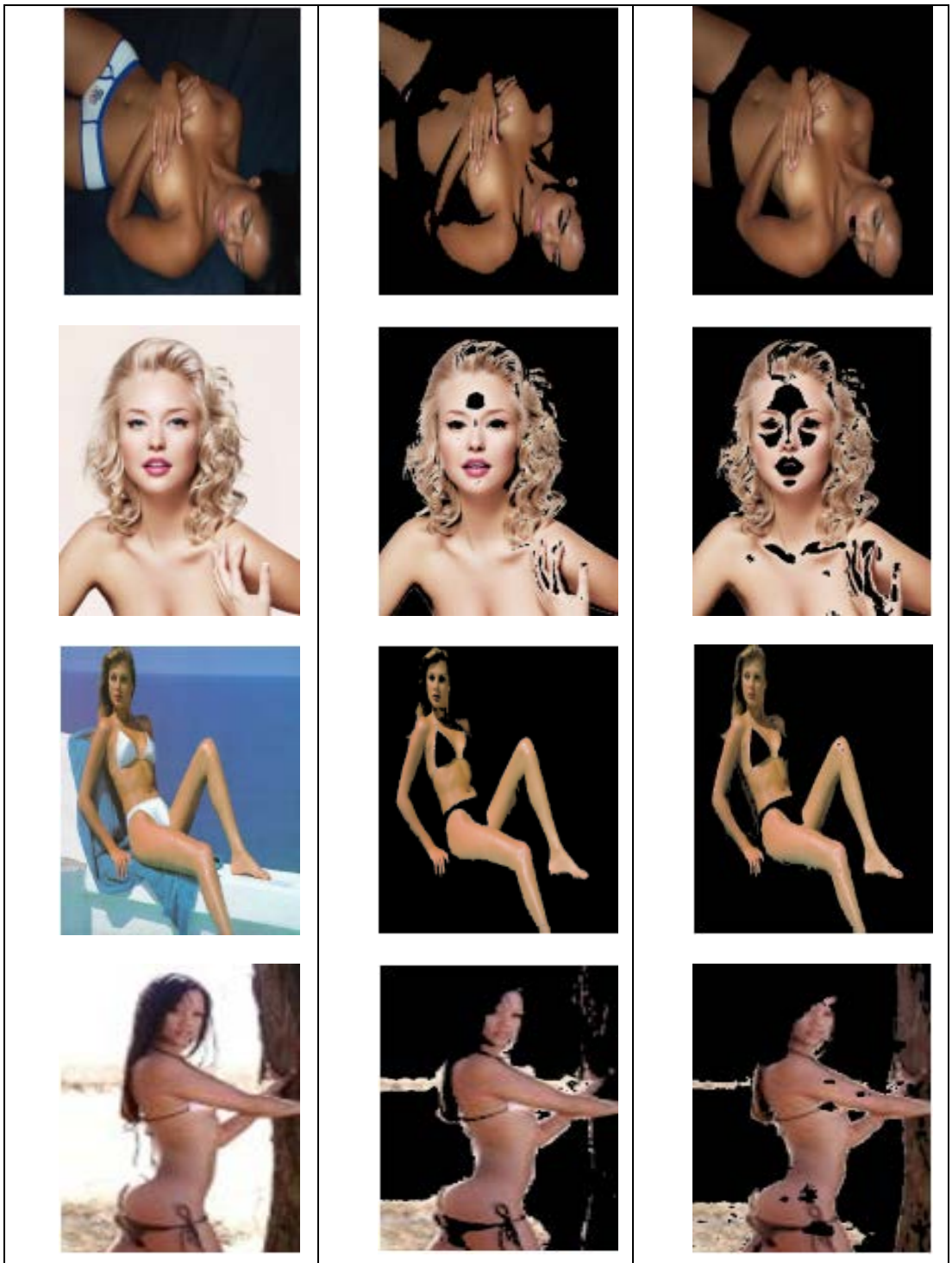
2.12

Y, C_b and C_r values for each pixel in image is in the range 0 to 255. These values are similar to those presented by Tarek in [11]. After conducting couple of experiments we increased the range of C_b from 55 to 85. This made the images of beaches a lot bigger in general and aiding us when implementing lacunarity in chapter 3.

2.2.4 HSI versus YC_bC_r

As mentioned earlier while discussing the HSI model, both the HSI and the YC_bC_r models are by intuition supposed to have a better performance when compared to the RBG model. This is because of the luminance condition which is different since pictures are taken under different lighting conditions and varied camera setting. To pick the method that gives the better performance the two

techniques have been compared using MATLAB and the results of skin region segmentation is shown in Figure 2.5.



(a) Original Image

(b) YC_bC_r segmented skin region

(c) HSI segmented skin region

Figure 2. 5: Skin region Segmentation using YC_bC_r and HSI color models.

Figure 2.5 show the segmented skin regions for four different images where both YC_bC_r and HSI color segmentation have been employed. For the first image it

can be stated without doubt that HSI domain based segmentation provides the better output. The YC_bC_r segmented skin region is fairly close to that of HIS's. When the subject is a colored person the YC_bC_r segmentation appears to be inferior to that of HSI. For the remaining three images, clearly YC_bC_r gives better results. This could be observed by comparing the face, hair, and collarbone and hand areas depicted in the two segmented skin regions. For the fourth subject, also note that no portion of the hair was considered as skin and only little patches of the tree was incorrectly considered as skin. After the application of both methods on a lot images obtained from the Internet and also judging by the results presented here, it appears that YC_bC_r model is a better model for skin segmentation.

2.2.5 Improved YC_bC_r Color Segmentation

Still not satisfied with the segmented out regions via the use of the YC_bC_r model, we looked into ways of improving this. A major problem that had to be solved was related to the intensity. Generally, the intensity of light falling on object that are not smooth, changes more compared to those of smooth objects. In general the background of all scenes is either smooth or rough. When it is a smooth surface, we expect almost no noticeable change in light intensities. When these backgrounds have colors similar to skin or grayish purple (in YC_bC_r color space), the candidate regions appear the same way that skin color appears. This is why the YC_bC_r segmentation miss-judges them and picks them out as skin regions.

By trial and error, we found out that such errors could be minimized by combining the National Television Standards Committee (NTSC) color space and the YC_bC_r and forming a new segmentation mask. NTSC is the mode of television broadcast in the United States which operates in YIQ color space. The

discriminating power of NTSC among color and intensity has been noted by Blinn in [12]. The components of the NTSC color space are Y (the luminance component), I (the cyan-orange component), and Q (the green-purple component).

The new segmentation method we developed here is quite simple and involved only the second layer which is cyan-orange from YIQ. It is achieved by first finding average of the matrix for the second layer and counting the number of pixels that are greater than this average. We check if this count is greater or equal to half the size of the matrix then find out those pixels that are greater than 0.3 times the average and make them white. If the count is neither grater nor equal to half the size of the matrix, we pick those pixels that are greater than 0.9 times the average and make those pixels white. All other pixels are made black. The binary mask from the procedure described above is then combined with that of the YC_bC_r mask with an AND operation and the final improved binary mask is obtained.

2.2.6 Segmentation using Standard and Improved YC_bC_r Schemes

From the results depicted in figure 2.6 we see that when the background color has some similar tones to human skin, the YC_bC_r segmentation would classify a big part of the background as skin regions. This would at times, double up the exposed skin detected from the image and perhaps lead to wrong classification. It is clear from figure 2.6(c) that the improved YC_bC_r method will enhance the segmentation greatly and big portions of the background will no more be misclassified as skin regions.

(a) (b) (c)

Figure 2. 6: Skin regions using standard and improved YC_bC_r schemes

(a) Original images, (b) Standard YC_bC_r segmented skin regions, (c) improved YC_bC_r segmented skin regions

Chapter 3

NUDE PICTURE CLASSIFICATION

In previous chapter, we used the color property to detect skin regions after stating that image processing is vital to this work. Also stated was the fact that using color property of skin alone can't give a reliable result because we will have problems when a reasonable amount of the images (collected from a web site to be evaluated) are lions, beach scene and/or objects with skin like colors. Making use of the texture property of skin and non-skin images, we can manipulate pixels either individually or in groups to decide whether or not an image contains exposed skin.

Since the texture of skin is generally smooth and that of skin like images are not smooth (most of the time), the texture property was used in combination with color information and area of candidate skin blob to determine the content of a picture. The following mainly used texture property to further determine actual skin regions from candidate skin regions:

- Fractal dimension
- Lacunarity
- Edge Analysis
- Co-Occurrence Matrix Based Entropy and Uniformity
- Face detection in excess skin exposed images
- Categorizing Based on Analysis of Candidate Skin Regions

All six concepts require finding a threshold to make a decision on whether or not the image is adult content containing. Next, we will be discussing each one of these six different criteria in more detail. It is important to bare in mind that threshold were empirically chosen.

3.1 Fractal Dimensionality (FD)

Fractal geometry was presented by Mandelbrot [13] as studying irregular and disordered figures which cannot be described by Euclidean geometry. A fractal dimension shows the ratio of a statistical index of complexity comparing the detail in a pattern. In exact terms, a fractal pattern changes with the scale at which it is measured. In accordance to human perception, it could be seen as the measure of an objects contour. Geometries with similar irregularities have similar fractals dimension and the fractal dimension between different regularities are usually long distance apart.

3.1.1 The Mandelbrot Set

The history of fractals dates far back but was really brought to light by Benoît Mandelbrot. He coined the name “fractal dimension” so for this reason and things he was able to achieve in his research on the topic, he is referred to as the father of fractal geometry. Very importantly, he described the Mandelbrot set which is as follows:

3.1

For starters, we put the value of “*C*” (which could be different values when iterating) in the equation. Each complex number is actually a point in a 2-dimensional plane. The equation gives an answer ’Z_{new}’. we repeat the process but this time, inserting ’Z_{new}’ as ’Z_{old}’ and calculate ’Z_{new}’ again. The rational here is to see what happens for different starting values of ‘C’.

When a number is squared, it gets bigger (except in a few cases i.e. numbers between (−1, 1)) and then if you square the answer, it gets even bigger. Eventually, the answer goes to infinity. This is the fate of most starting values of ‘C’. For those values of ‘C’ that do not get bigger, they actually do the opposite (gets smaller), or alternate between a set of fixed values. These are the points inside the Mandelbrot Set, which correspond to the black colors in figure 3.1. Also visible from the image is that; outside the set all the values of ‘C’ cause the equation to go to infinity and the colors are proportional to the speed at which they expand.

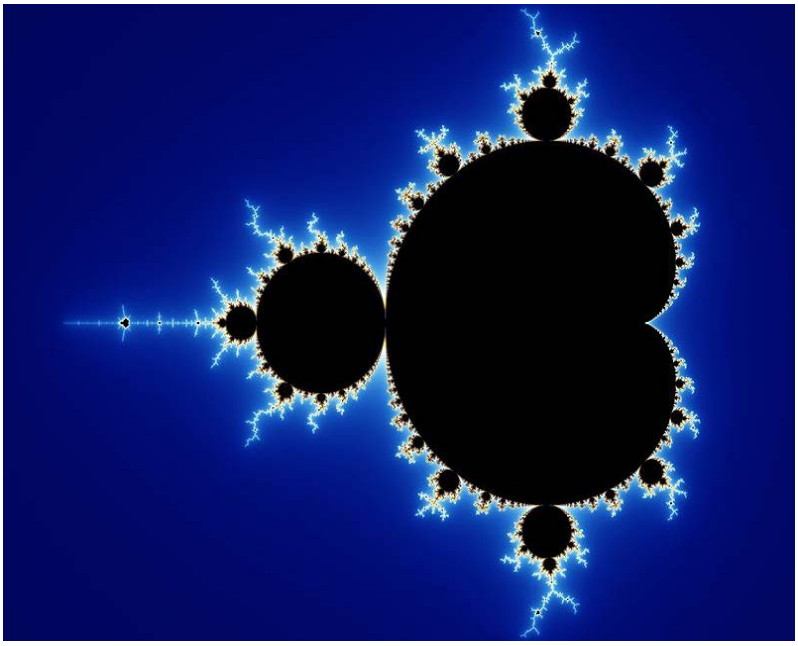


Figure 3.1: A Typical Mandelbrot set image [14]

The border of the shape is very important because if we zoom in much closer, we see that the same shape is reproduced over and over. Also, the computation becomes more cumbersome.

3.1.2 Geometric Fractals

Fractals can be found all over nature in an enormous range of scales. We find the same patterns repeating themselves again and again, from the tiny branching of

our blood vessels and neurons to the branching of trees, lightning bolts, and river networks. Regardless of scale, these patterns are all formed by repeating a simple branching process and the derivation process of a particular set was presented earlier in the Mandelbrot set. That branch of the mathematics is referred to as fractal algebra. Our interested is to find a way to measure lengths of fractal patterns. To do so, Richardson-Mandelbrot plot was utilized.

3.1.3 Box Counting Method

This method gives a reasonably good estimate of the fractal dimension for a binary image. The procedure is as follows: the image is first covered with a grid of squared cells of size '*r*', for binary images it is much easier because the cell size is expressed as number of pixels and this is the reason it was implemented in this work. Sierpinsky's gasket which is stored in a 688 × 612 matrix and gridded is shown in figure 3.2 for illustration.

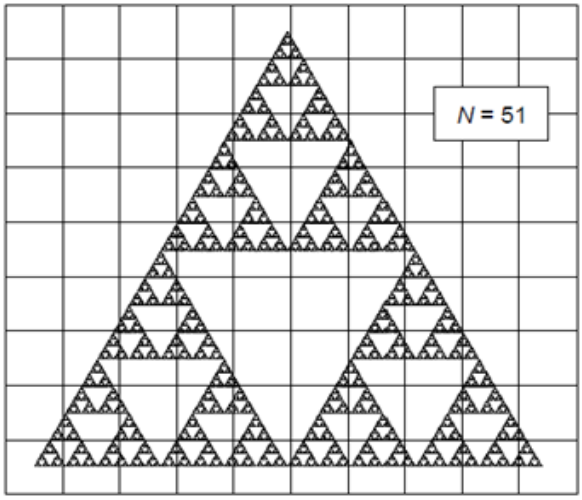


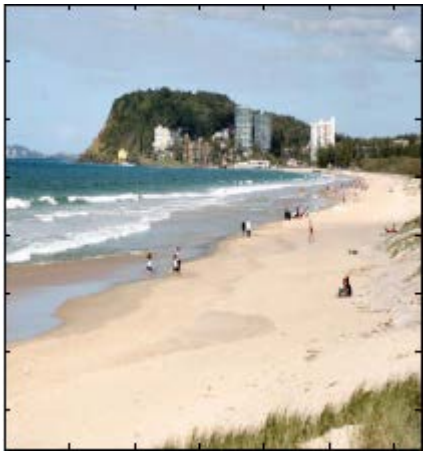
Figure 3.2: Sierpinsky's gasket [15]

The number of squares $N(r)$ needed to cover the structure is giving by a power which is as follows:

3.2

For which

From equation 3.2, the total area



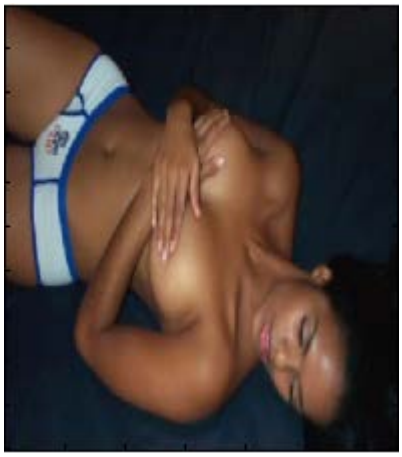


Figure 3.4: Candidate skin region.

This is an indication that even our improved skin segmentation would not always give us good results and we need some other criteria for correctly categorizing the test images. This other criteria could perhaps be the fractal dimension.

A first round computation of the fractal dimension for nude images and images which has tones that resemble skin color (beaches, deserts, lion furs etc.) showed that the distribution of the fractal dimension values for real skin and distribution of fractal dimension values for the remaining images would tend to overlap when one uses the largest connected component among the candidate skin regions as is (overlap reduces the ability to separate). To avoid erroneous decisions, further investigations were carried out and it was noted that due to tone changes in the real nude images, binarization of the candidate skin regions with a selected threshold value would make the skin regions segmented out from nude images loose some pixels but the loss from beach and dessert images would be minimal. This change of size in the largest components for skin regions segmented out from nude images would reduce the overlap between the two distributions and would help achieve better seperability.

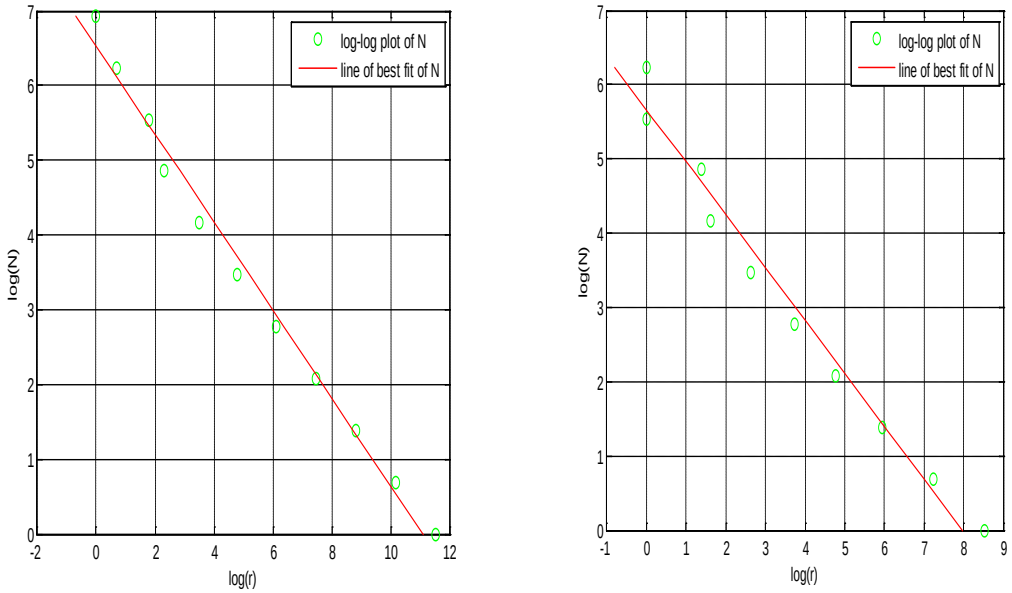
Figures 3.5 and 3.6 depict the binarization with a threshold value of 0.4 for the candidate skin regions in Figure 3.4 and the selection of the largest component among the binarized regions.

Figure 3.5: Binarization of candidate skin regions using threshold value of 0.4.

Figure 3.6: Binary image of the largest candidate skin blob.

The binarization is also advantageous since for the box counting algorithm, the image must be binary type. While trying to compute the fractal dimention we keep splitting the candidate skin blob in equivalent squares with normalized sizes r = 1/2,1/4,1/8,…, and for each scale compute the number of squares covered by the object. The FD is then obtained by plotting log(N) versus log(r) and finding

the slope of the line that best fits the pair of points as depicted in figure 3.7. Fractal dimension which equals unity minus the slope of this line is known to give a measure of the roughness of a surface. Intuitively, the larger the fractal dimension, the rougher the texture would be.



(a) (b)

Figure 3.7: log (N) versus log (r) for largest candidate skin blobs

(a) for largest connected component from beach image (b) for largest connected component from nude image

The FD values for the two test images are 1.7107 for nude image and 1.5877 for the beach image. We repeated this process for a large number of nude images and found out that most of the images would have a FD value in the range 1.6 - 1.82.

3.2 Lacunarity

Mandelbrot has previously stated that lacunarity which is an important spatial characteristic of data sets would be a competitive alternative to fractal dimension usage. Lacunarity in fact complements fractal dimension by taking it a step further since it better describes the appearence of selected objects. Many

researchers have chosen to work with fractal dimension, but fractal dimension alone does not completely describe the appearance of an object it only considers the space filling characteristics of the data. This was pointed out in [16] by Charles and MaJunkin.

Since fractal dimension does not fully quantify texture, for different images it is possible to get FD values that fall in distribution ranges of images from other categories. This makes proper categorization of images more difficult. Empirical work carried out shows that lacunarity measure would tend to give better margins between the various distributions for images from different categories when compared to FD.

In the literature the most commonly used method for estimating lacunarity is the gliding box algorithm which takes into account the localized mass and is very similar to the box counting algorithm which the FD uses. It is implemented by picking a box of size *r* and counting the number of skin pixels that fall into it everytime it is moved to a new location within the largest candidate skin component obtained via the use of our improved YC_bC_r skin segmentation and binarization. The distribution of box masses, *B(p, r)*, where *B* is the number of boxes with *p* skin pixels. This distribution is then converted into a probability distribution, *Q(p, r)*, by dividing it by the total number of boxes of size *r*.

Then the value of p and the probability distribution are used to calculate the first and second moments of the box mass using equations 3.4 and 3.5:

3.4

3.5

Finally one can determine the gliding box lacunarity as in equation 3.6,

3.6

3.3 Edge Analysis

Another useful criteria for distinguishing skin regions from other similar colored region is by making use of skin texture features pointed out in [17] by Henry, Jin and Balujah.

The two ratios defined in [17] and shown below in equations 3.7 represent the measure of skin texture and measure of how much of the image texture can be attributed to skin-colored pixels.

3.7

In our filtering algorithm only the TPAS ratio defined in 3.7 was used. Computing this ratio for various images in each category, we found that the ratio for skin images had values less than 6.5. This value was later selected as the threshold to be used in out MATLAB implementation of the algorithm.

3.4 Co-Occurrence Matrix Based Entropy and Uniformity

In [18], Tuceryan and Jain pointed out that one way to analyze texture was via the use of spatial distribution of gray values. A large number of texture features have been proposed and the prominent ones include energy, uniformity,

homogeneity, contrast, entropy and correlation. Our experimentation with the above said features have pointed out that entropy and uniformity features used together would provide another distinguishing criterion for our web content filtering system.

Given any image for which the gray level co-occurrence matrix (GLCM) $P_d(i,j)$ is determined, the entropy and uniformity values can be calculated using equation 3.8 and 3.9.

$$\text{Entropy} = \qquad 3.8$$

$$\text{Uniformity} = \qquad 3.9$$

Gray level co-occurrence matrix is useful for estimating image properties related to second order statistics as stated in [18] and in Woods & Gonzalez [19]. The gray level co-occurrence matrix for a certain displacement vector $d=(d_x,d_y)$ can be written as

3.10

In this work if the absolute value of the difference between uniformity and average entropy was less than 0.5 and at the same time if uniformity was greater than 0.2 the candidate skin region tested would be considered to belong to the nude category.

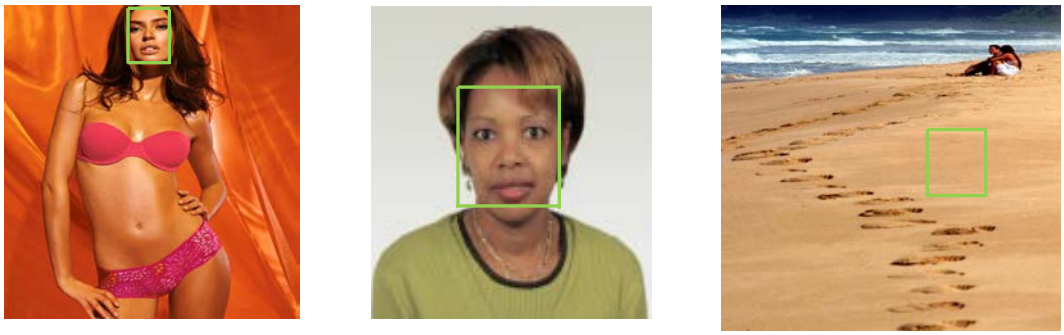
3.5 Face Detection in Excess Skin Exposed Images

Another criterion we used while trying to improve the level of correctness of the decision made on test images was face detection. Reason for involving face

detection is solely hinged on the fact that we expect real skin images to sometimes, contain face and non-skin images not to contain face all of the time. From literature, we were able to lay hands on face detection algorithm and code as developed by Viola and Jones [20]. Incorporating this code as part of our filtering criterion, the outcome was in three forms;

- Actual nude image with detected face
- Face detected for close up image
- Wrongly detected face regions

Figure 3.8 depicts all three scenarios.



(a) (b) (c)

Figure 3.8: Face detection in different images.

Usually, we are faced with a problem when the image is a close up face image or in situations where face is detected when it actually is not. The idea upon which this face detection works is such; if face is detected, consider it a nude, if not, then is not nude. Therefore, we do not want figure 3.8 (b) to be connived by our program to be face nor should figure 3.8 (c) be.

We implemented two procedures to make our program consider only figure 3.8 (a) as nude.

First procedure was based on the face that face is oval therefore should have very little pixel at the 4 corners of the rectangle that bounds the supposed face region. Figure 3.9 depicts our expectation for faced image.

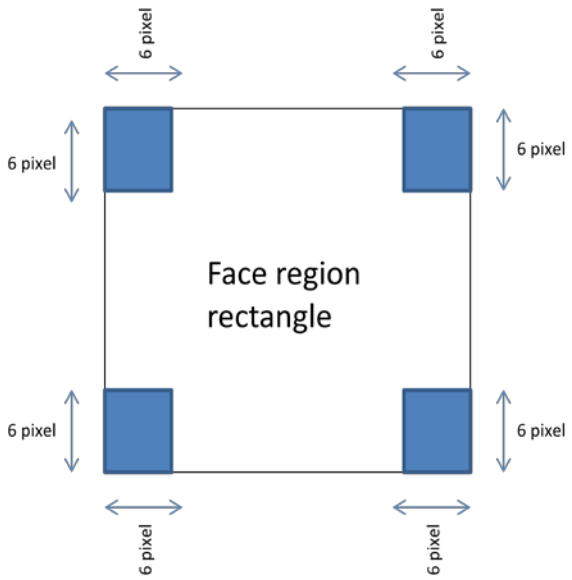


Figure 3.9: Extracted face region.

Picking 6×6 pixels on the 4 corners, we were able to deduce that, of the 144 pixels (6×6 + 6×6 + 6×6 + 6×6) only 52 or less pixels are expected to be white bearing in mind that the image is in binary form. The threshold value 52 was arrived at empirically. This procedure was geared towards eliminating situations like those in figure 3.8 (c).

The second procedure considered percentage of face pixel count and captured in equation 3.11.

3.11

This is intended to eliminate the chances of selecting images like in figure 3.8 (b).

For those images that are nude but not detected, we expect that they satisfy other criteria.

3.6 Categorizing Based on Analysis of Candidate Skin Regions

In our quest to get even better results, another method was adopted. It mainly considered area occupied by the candidate skin's mask. Therefore, an understanding of area as regards grouping of pictures needs to be understood. Grouping of pictures here refers to the major categories that have been considered so far. They are; nude images, lion images, beach sceneries and normal images. The reason for this is same as earlier. We want to decide whether or not an image contains too much exposed skin and in this case, how much of it forms a connected region (skin blob). This section combines general ideas which are hinged on pixel count and appearance (taking skin and non-skin into account) by using different constraints where test images are passed and each goes through all constraints one after the other. When a constraint is satisfied, it is not nude otherwise, the next constrain is implemented. In the end, if no constraint is satisfied, the image is nude.

Generally, it is expected that the binary mask corresponding to a skin region should not be composed of scattered patches of small connected components. But for some animals like lions whose skin color resembles human skin color under certain illumination such behavior is observed (a false positive situation). As expected, there are lots of shadowy areas due to roughness of the fur which are perceived to be none skin within the area. This makes such candidate skin images to appear to have lots of scattered patches of connected components. Note that lion's fur is not skin but is detected as skin most of the time because its color falls

in the range that our skin segmentation model considers as skin. To eliminate such images from being classified as nude, we count the number of separate patches detected and if this count is more than 300, the image is taken to be non-nude (first constrain). Else, we continue by checking other conditions.

The second constraint checks the differences between the pixel values of the neighboring pixels (excluding zero pixels) in each layer of the RGB image. It mainly tries to determine if an image contains lots of varying neighboring pixel values.

0	0	0	0	0
0	0	a	b	0
0	0	c	c	0
0	0	e	0	0
0	d	0	f	g

Figure 3.10: A 5×5 matrix representing a segmented image.

Let us assume that the 5×5 area in figure 3.10 has been segmented from a color image and pixel positions a-g are actually skin pixel values. First, each pixel is checked to see if it has values other than 0. If this condition is satisfied, then the pixel is compared with neighboring non-zero pixels by taking the absolute difference between itself and each non-zero pixel in the 3×3 square surrounding it. The maximum and minimum among these difference values are recorded and then we start to accumulate the difference between the max and min values for each layer and each position of the kernel in each layer. Once the kernel is moved over the entire image the accumulated difference for each layer would be obtained by dividing the accumulated differences by the number of candidate

skin pixels. Equation 3.12 and 3.13 can be used to classify an image as none adult image if the average of the accumulated max and min differences is as indicated.

3.12

3.13

Our third constraint considered the largest object. This constraint is based on the fact that the said blob's pixel count must not be 1/50 times less than the size of the image.

The final constraint used here was obtained experimentally by Rigan in [21]. [21] States that if the percentage of skin pixels relative to the tested image size is less than 15 percent, the test image is none nude. If none of the mentioned constraint is satisfied, this criterion considers the test image to contain too much exposed skin and hence would classify it as nude.

Figure 3.11 depicts a flowchart showing how the downloaded images are processed in a sequential manner using the 5 different criteria previously discussed in chapter 3. Note that with the exception of the face detector all the other criteria are based on the candidate skin parts obtained from the improved YC_bC_r segmentation algorithm or a binarized version of it. For each criterion a count is held and whenever one of the five criteria is met the corresponding count for that run (image processed) is incremented. Once all checks are done if the count is greater or equal to three the image is classified as adult content containing. This process is repeated till all images in the download folder are processed. Finally a percentage value saying how confident we are that the page has nudity is obtained.

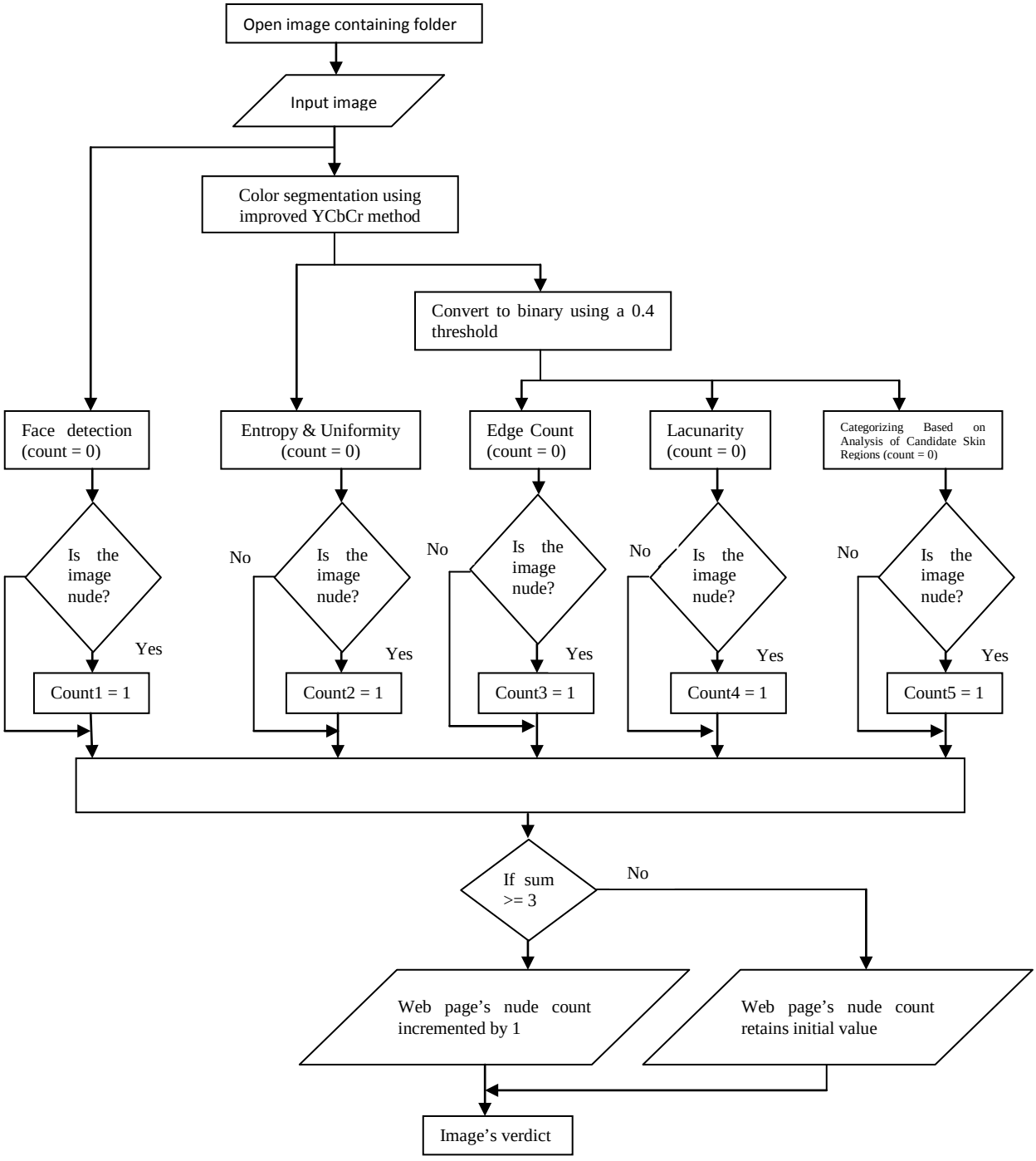


Figure 3.11: Image processing Flowchart.

Chapter 4

TEXT ANALYSIS AND JAVA CODE

"*Content is king*" is a popular quote in the web development world and there is a logical reasoning behind it. Without content (text, image, video, audio, etc.), visitors to a web site will have nothing to read, look at or listen to, that will help them learn about the message a site is trying to convey. This is very important for developers because, if web users cannot find what they are looking for, chances are that they will not visit the site again.

Because of this many commercial sites will use text, supporting images and even sound to have their message(s) passed across. We can exploit this fact when considering adult sites. In adult site, usually they want people to pay via credit card to watch videos, meet new people, look through picture galleries, read articles etc. We are interested in the 'read article' (text part) here and there are words that are peculiar to such sites. Generally speaking, adult sites do not have much text but rather, they contain more of videos and pictures. The videos are usually "title explanatory" meaning a visitor or registered member can know what to expect when they load a video they want to view. This suggests that in our bead to detect such site, text wise analysis is necessary.

The motive in this work still remains trying to better decide on whether or not a website is adult content containing. Since we are going to base our final decision

using both text and image based analysis and image analysis is already presented in the preceding chapter, this chapter will detail processes involved in our text analysis and highlight other related topics.

4.1 Words and Phrase Selection

There are words that are pertaining to adult sites and we are faced with the challenge of finding out such words. These words will be included in the JAVA code which was used to download images so that while the JAVA code is downloading the images it will also simultaneously downloads the HTML content and after removing the tags will carry out string in string search to determine words and their frequencies for the tested site. If the frequency of the words individually sum up to more than a selected threshold, the site is classified to be adult content containing otherwise, the site is considered safe.

In this thesis we considered couple of ways to find out such words. One approach was to randomly copy the text parts of a number of adults site and check the frequency of occurrence of individual words. After which a database of words pertaining to such sites will be created. But, this method was not adopted because it was computationally intense since one has to repeat the process for a large number of sites.

The second approach involved a general survey of lots of adult sites. But to do so, we did not look for all kind of word & phrases exactly but instead found clues as to “what to look for” by considering blacklisted word like those in Google [22]. Once such words and phrases were determined, the next thing the program did, was to search for it in a string which had the text content of a web site under

evaluation. There was a problem which came up as a result of the fact if a word which is part of our collection of word is found in a longer word and this longer word has a different meaning, the program detected it and would increment the count. To solve this, most of the words given to the program to search had once space inserted after it. The words and phrases that were used in the program will not be mentioned here due to their explicit nature.

4.2 Java Program and Algorithm to Parse Text from URL Address

The JAVA code that was written for this work was to achieve two things. The first was to collect the text content of the web page and search through as depicted in figure 4.1.

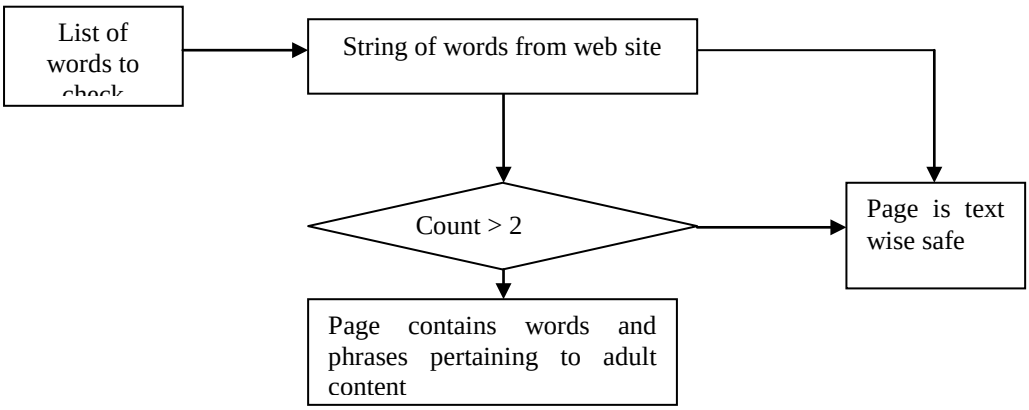


Figure 4.1: Text analysis mechanism.

It gives an insight as to how the text analysis was done here. Firstly, the list of words was generated and we had picked them based on individual observation of web pages. When the web site is typed in, the crawler saves all the text content of the home page as a long string of words. The program then picks a word from the list (1 after the other) and check if it exists in the string. If it does, a counter is incremented. This process is repeated for each word on the list after which a final

count is obtained. This count is a total of counts each word in the list and we have set a threshold of 2. The number of words must not be more than 2. This number is so low because we want the program to indicate that a site is of adult content at the slightest detection of such word. Shortening of the execution time can possibly be achieved by skipping frequently occurring English words. Such words can be seen in Richard [23] and we have a few highlighted here in table 4.1.

Table 4.1: Frequently appearing words

The	Have	On	But	As	If	Her	Make
Be	It	With	From	We	Their	Find	Who
Of	For	Do	They	An	Go	Come	Such
And	I	At	His	Say	What	Me	Out
A	That	By	She	Will	All	My	Up
In	You	Not	Or	Get	Would	People	See
To	He	This	Her	Can	Which	Your	Know
Year	Than	No	Also	May	About	These	Think
Into	More	Other	Well	Way	Because	Very	New
Last	How	Give	Any	Look	When	Use	
Time	Take	Them	Some	So	Could	Him	
Then	Now	Just	Only	Like	Should	Good	

4.3 Algorithm to Download Images from given URL

The second thing the java code aimed to achieve was to download images from the index page of the web site we are interested in analyzing. These images will in turn serve as input images for the image analysis that was highlighted in chapter 3.

Figure 4.2 shows the algorithmic steps for the JAVA code developed for downloading image on a specified web site.

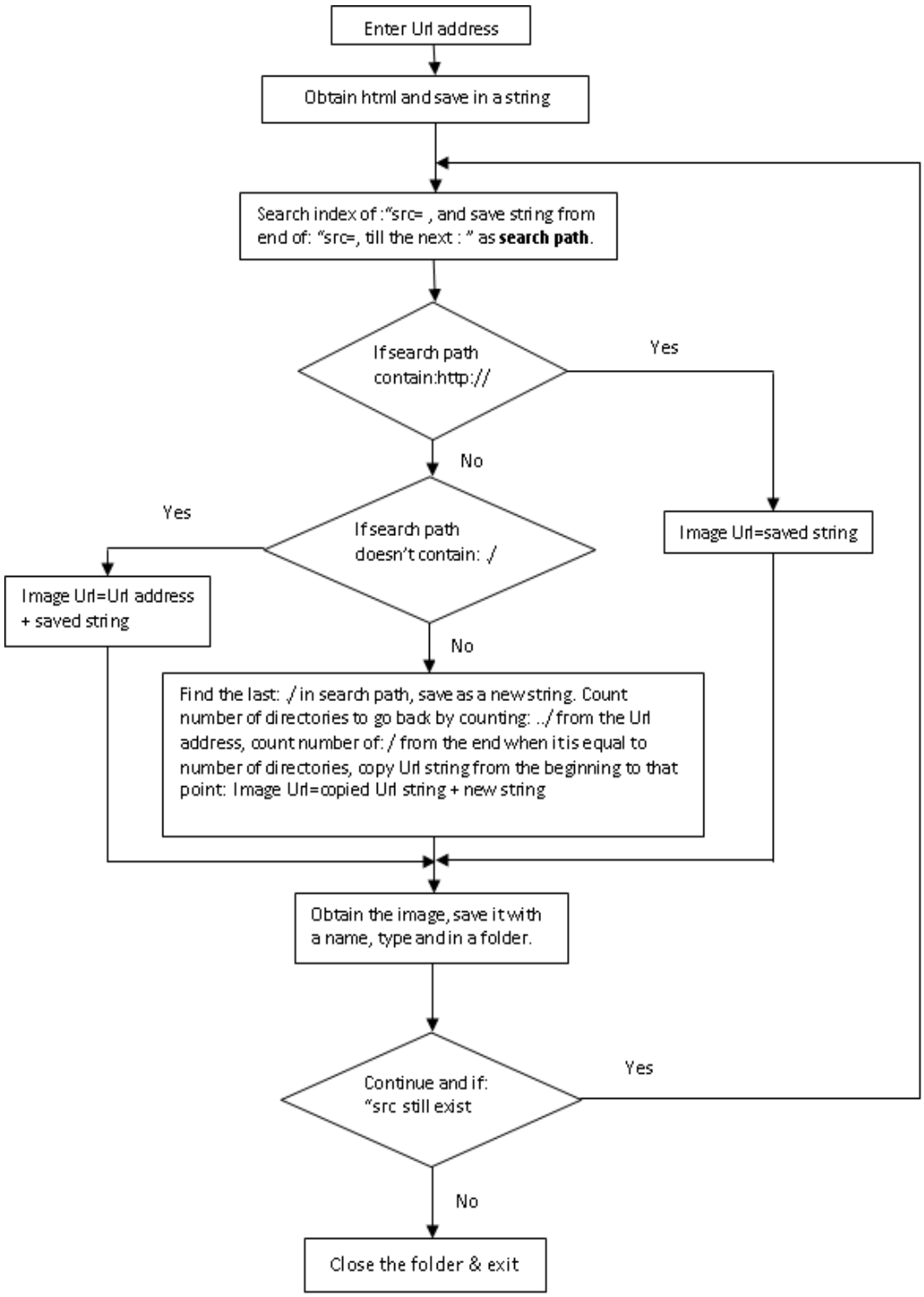


Figure 4.2 Flowchart for automated image downloading.

4.4 Java Code for Parsing a Web Site

```
import java.awt.Graphics2D;
import java.awt.Image;
import java.awt.image.BufferedImage;
import java.net.*;
import java.util.*;
import java.io.*;
import java.net.HttpURLConnection;
import javax.imageio.ImageIO;
 class Etracturl_stringsearch {
      public static void main(String[] args) {
        System.out.println("enter the URL address: ");
        Scanner input = new Scanner(System.in);
        String url = input.nextLine();
        String html = getText(url);
        String ary = html;
        String searchKey = "img";
        String invComma = "\"";
        String imgName, imgUrl, srcPath;
        int i=1, imgIndex, srcIndex, index1=0, index2, firstImg;
        String[]  collection  =  new  String[]{"##","#####   ","###","#####","###
","#### ","######## ","#### ######"};
        int count = 0;
        int index, imgNo = 1;
        int tot_cnt = 0;

        for(String word: collection)
        {
            count = 0;
            index = 0;

            while (index < ary.length() && (index = ary.indexOf(word, index)) >=
0)
            {
                    count++;
                    index = index + word.length();
            }
            System.out.println(count);
                tot_cnt = tot_cnt + count;
        }
                        System.out.println(tot_cnt);
        if(tot_cnt > 1){
                System.out.println("it is an adult site");
                }
        else {
                System.out.println("it is okay for children");
            }
        imgIndex = firstImg = html.indexOf(searchKey,index1);
        while(firstImg <= imgIndex)
        {
            srcIndex = html.indexOf("src", imgIndex);
            index1 = html.indexOf(invComma, srcIndex);
            index2 = html.indexOf(invComma, index1+1);
            srcPath = html.substring(index1+1, index2);
            System.out.println(srcPath);
            int dirCount = 0;
            if(srcPath.lastIndexOf("../") != -1)
            {
                imgName =
srcPath.substring(srcPath.lastIndexOf("../")+2,srcPath.length()).trim();
                for(i=0; i<srcPath.length(); i++ ) {
                    if( srcPath.charAt(i) == '/' ) {
                        dirCount++;
                    }
                }
            }
            else if(srcPath.startsWith("/"))
                   imgName = srcPath.substring(1,srcPath.length()).trim();
            else
                   imgName = srcPath;
            String domain = url.substring(0,(url.indexOf("/", 7)));
            if(srcPath.startsWith("http://"))
                imgUrl = srcPath;
            else
```

```
            {      if(dirCount == 0)
                          imgUrl = url.substring(0, url.lastIndexOf("/")) + "/" +
imgName;
                   else
                   {      int numOfDir = 0;
                   for(i = domain.length() + 1; i<url.length(); i++ ) {
                        if( url.charAt(i) == '/' ) {
                          numOfDir++;
                        }
                    }
                   int parentUrlIndex = (count == 0) ? 0:1;

                   for(i=1; i <= numOfDir - count; i++ ) {
                    parentUrlIndex               =               url.indexOf("/",
domain.length()+parentUrlIndex);
                  }
                  if (parentUrlIndex == 0)
                     imgUrl = domain + "/" + imgName;
                  else
                     imgUrl = url.substring(0, parentUrlIndex) + "/" + imgName;
                     }
            }
            System.out.println(imgUrl);
            try{
                URL urlnew = new URL(imgUrl);
                Image image = ImageIO.read(urlnew);
                BufferedImage cpimg=bufferImage(image);
                File f1 = new File("../../../Desktop/WebImages/image_" + imgNo +
".png");
                ImageIO.write(cpimg, "png", f1);
                imgNo++;
            }
             catch(Exception e){
               System.out.println(e);
          }
           imgIndex = html.indexOf(searchKey,index2);
        }
    }
    public static String getText(String fn) {
        StringBuilder text = new StringBuilder();
        try {
        URL page = new URL(fn);
        HttpURLConnection conn =
        (HttpURLConnection) page.openConnection();
        conn.connect();
        InputStreamReader in = new InputStreamReader(
        (InputStream) conn.getContent());
        BufferedReader read = new BufferedReader(in);
        String line;
        do {
        line = read.readLine();
        if (line != null)
        text.append(line +"\n");
        } while (line != null);
        } catch (IOException e) {
        return "Error - :" + e.getMessage();
        }
        return text.toString();
      }
   public static BufferedImage bufferImage(Image image) {
        return bufferImage(image,BufferedImage.TYPE_INT_RGB);
     }
   public static BufferedImage bufferImage(Image image, int type) {
        BufferedImage  bufferedImage  =  new  BufferedImage(image.getWidth(null),
image.getHeight(null), type);
        Graphics2D g = bufferedImage.createGraphics();
        g.drawImage(image, null, null);
        return bufferedImage;
        }
}
```

Note that the words and phrases that were used to determine the nature of a web page have been replaced in this code with “####” of various length so as not to be vulgar and allow this work to be readable by people of all ages.

Chapter 5

SIMULATION AND RESULTS

In order to show the accuracy of classification through the use of the proposed text and image analysis techniques testing with some pre-selected web sites were carried out. Figure 5.1 below shows the two parallel processes that a computer equipped with a JAVA compiler and the MATLAB programming environment would carry out.

The URL address of the website which needs to be evaluated is entered when the java code is executed. The text from the web site is saved as a string variable in the java environment. The text is then checked for existence of certain words that are pertaining to adult web sites. Simultaneously the JAVA code download and saves copies of images from the site to a folder. Once this is achieved the MATLAB code then takes the content of this folder and evaluates the images one after the other for the existence of nudity.

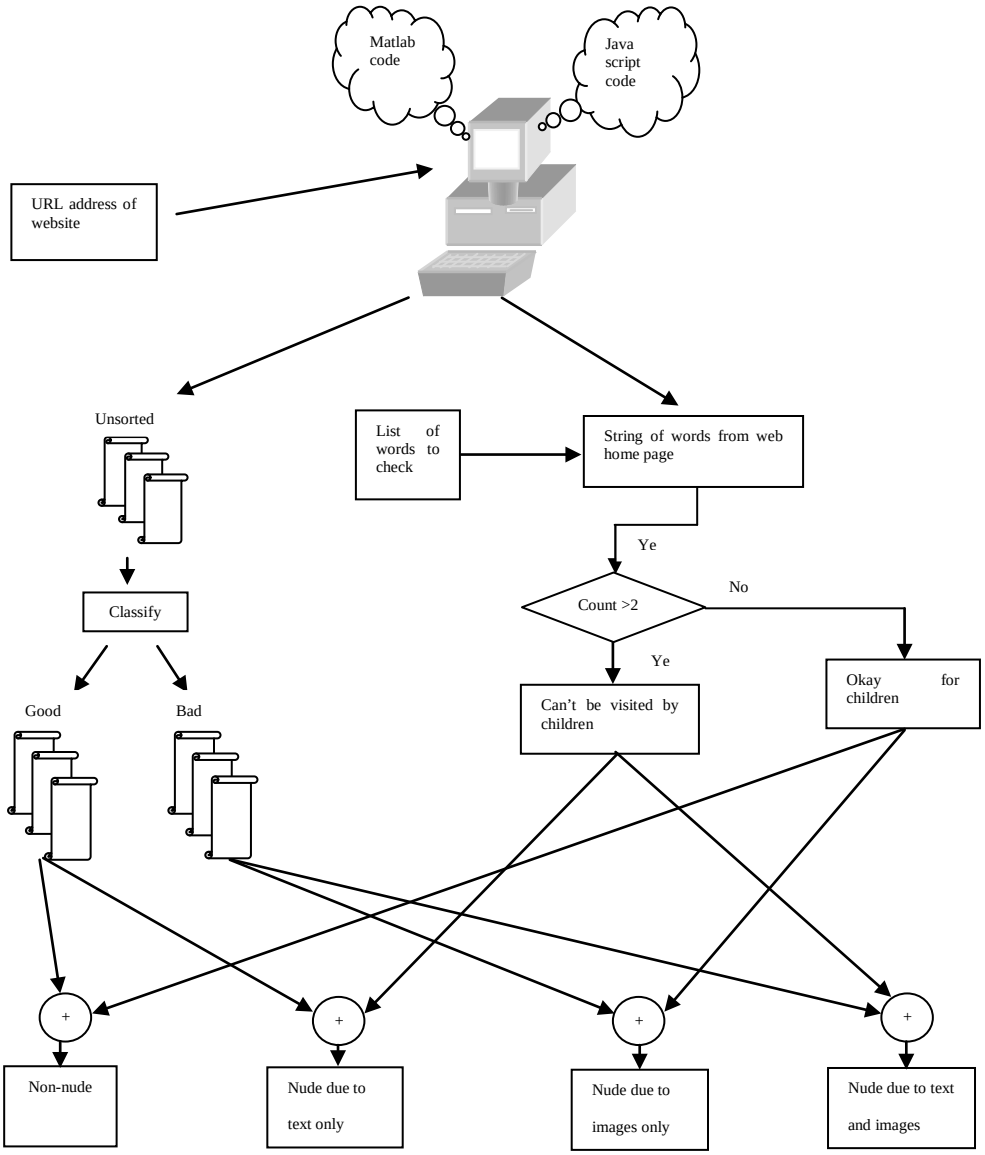


Figure 5.1: System Operation

The threshold values adopted for the different criteria discussed in chapter 3 were obtained by training with a mixed image set composed of four image categories; namely nude pictures, lion photos, beach scenes and regular everyday pictures. For each category the number of images was as follows: 47 nude, 24 lion, 38 scenery and 48 regular everyday pictures. Once the testing was finalized five different porn sites were picked for the ultimate test. The URL address for each

site and results showing how each criterion has been triggered by each image downloaded from the particular site are provided in tables 5.1 to 5.5.

5.1 First Web Site

Table 5.1: Image analysis for www.bondagester.com.

Image #	Lacunarity	Face detected	Entropy and Uniformity	Analysis based on size of candidate	Edge sum	Total score
1	1	1	1	1	0	4
2	1	0	0	1	0	2
3	1	0	1	1	1	4
4	1	1	1	1	1	5
5	1	0	1	0	0	2
6	0	0	0	0	0	0
7	0	0	1	1	0	2
8	1	0	0	1	0	2
9	1	0	1	1	0	3
10	1	0	1	1	0	3
11	1	0	1	1	0	3
12	1	0	1	1	1	4
13	1	0	1	1	0	3
14	1	0	1	0	0	2
15	1	0	0	0	0	1
16	1	0	1	1	0	3
17	1	0	1	1	0	3
18	1	0	1	1	0	3
19	1	0	0	0	0	1
20	1	0	0	0	0	1
21	1	0	1	1	0	3
22	1	1	1	1	0	4
23	1	0	1	1	1	4
24	0	0	1	1	0	2
25	1	0	0	1	0	2
26	1	1	1	1	0	4
27	1	0	1	1	1	4
28	1	1	1	0	0	3
29	1	0	0	1	0	2
30	1	0	1	1	1	4
31	1	0	1	1	0	3
32	1	0	1	1	0	3
33	1	0	1	1	0	3
34	0	0	1	1	0	2
35	0	0	1	0	0	1
36	1	0	1	0	0	2
37	0	1	1	1	0	3

38	1	0	1	0	0	2
39	1	0	1	0	0	2
40	0	0	0	1	0	1
41	1	0	1	0	0	2
42	1	0	1	1	0	3
43	1	0	0	1	0	2
44	1	1	0	1	0	3
45	1	0	0	1	0	2
46	1	0	1	1	1	4
47	0	0	1	1	0	2
48	1	1	1	1	0	4
49	0	0	1	1	1	3
50	1	0	1	1	0	3
51	1	0	1	1	1	4
52	0	0	1	1	1	3
53	1	0	1	0	1	3
54	0	0	1	1	1	3
55	0	0	1	1	1	3
56	1	0	1	1	0	3
57	1	0	1	1	1	4
58	1	0	0	0	0	1
59	1	0	1	1	0	3
60	1	0	1	1	1	4
61	1	0	0	1	0	2
62	1	0	1	1	0	3
63	1	0	1	1	0	3
64	1	0	0	1	0	2
65	0	0	0	0	0	0
66	0	0	0	0	0	0
67	1	0	1	1	0	3
68	1	0	0	0	0	1
69	0	0	0	1	1	2
70	1	0	1	1	1	4
71	0	0	0	1	1	2
72	1	0	0	0	0	1
73	1	0	1	1	0	3
74	0	0	0	1	1	2
75	1	0	1	1	1	4
76	1	0	1	1	0	3
77	1	0	1	0	0	2
78	1	0	1	1	0	3
79	1	0	1	1	1	4
80	0	0	1	1	1	3
81	1	0	0	0	0	1
82	1	0	1	1	0	3
83	1	0	1	1	1	4
84	0	0	0	1	1	2
85	1	0	0	1	0	2

86	1	0	1	0	1	3
87	1	0	1	0	0	2
88	1	0	0	0	0	1
89	1	0	1	1	0	3
90	0	0	0	0	0	0
91	1	0	1	1	0	3
92	1	0	1	1	0	3
93	1	1	1	1	0	4
94	1	0	1	1	1	4
95	1	0	1	1	0	3
96	0	0	0	1	1	2
97	0	0	0	1	1	2
98	1	0	1	1	0	3
99	1	0	1	1	0	3
100	1	0	1	1	0	3
101	1	0	1	1	0	3
102	1	1	1	1	0	4
103	1	0	1	1	0	3
104	1	1	1	1	0	4
105	1	0	1	0	0	2
106	1	0	1	0	0	2
107	0	0	1	1	0	2
108	1	0	1	0	0	2
109	1	0	0	1	0	2
110	0	0	0	1	1	2
111	1	0	1	1	0	3
112	0	1	1	1	1	4
113	1	0	1	1	1	4
114	1	0	1	1	1	4
115	1	0	0	0	0	1
116	1	0	1	1	1	4
117	0	1	1	1	0	3
118	1	0	1	1	0	3
119	0	0	1	1	1	3
110	1	0	1	1	1	4
121	0	0	0	1	1	2
122	1	0	1	1	0	3
123	1	0	1	1	1	4
124	1	0	1	1	0	3
125	1	0	1	0	0	2
126	1	0	1	1	0	3
127	1	0	0	1	0	2
128	1	1	0	0	0	2
129	1	0	1	1	0	3
130	1	0	1	1	0	3
131	0	0	1	1	1	3
132	1	0	1	0	0	2
133	1	0	1	1	0	3

134	1	0	1	0	0	2
135	1	0	1	0	0	2
136	0	0	0	1	1	2
137	1	1	1	0	0	3
138	1	0	1	1	1	4
139	1	0	1	1	0	3
140	1	0	0	1	0	2
141	1	0	0	1	0	2
142	1	1	1	1	0	4
143	1	0	1	0	0	2
144	1	1	1	1	0	4
145	0	0	0	0	0	0
146	0	0	0	0	0	0
147	0	1	0	1	1	3
148	0	0	0	1	1	2
149	1	0	1	0	0	2
150	1	0	1	0	0	2
151	1	0	0	0	0	1
152	1	0	1	1	0	3
153	1	0	1	1	0	3
154	0	0	1	1	1	3
155	1	0	1	1	1	4
156	0	0	0	1	1	2
157	0	0	1	1	1	3
158	1	1	1	1	0	4
159	1	1	1	1	0	4
160	1	0	1	0	0	2
161	1	0	1	1	0	3
162	1	0	1	0	1	3
163	1	0	0	1	0	2
164	1	0	1	0	0	2
165	0	0	0	1	1	2
166	1	0	0	0	0	1
167	1	0	1	1	0	3
168	1	0	1	1	0	3
169	1	0	1	1	1	4
170	1	0	1	1	0	3
171	1	1	1	1	1	5
172	1	0	0	0	0	1
173	1	0	1	1	0	3
174	1	0	0	1	0	2
175	1	0	1	0	0	2
176	0	0	1	1	1	3
177	1	0	1	1	0	3
178	0	0	0	1	1	2
179	1	0	1	1	1	4
180	1	0	0	1	0	2
181	0	0	0	1	1	2

182	0	1	1	0	0	2
183	0	0	0	1	1	2

183 images were obtained from the home page of "www.bondagester.com" and classified as seen in table 5.1. When three or more of the five criteria were satisfied the classification percentages and the percentage of false negatives and positives were as depicted in Table 5.2.

Table 5.2: Detection percentages for www.bondagester.com

Nude percentage	56.8
Non-nude percentage	43.2
False negative	23.5
False positive	14.8

When each criteria was individually evaluated for its contribution to classifying an image as nude, the lacunarity gave 76.5%, face detection 12.0%, entropy and uniformity 73.2%, analysis based on size of candidate skin regions 64.5% and edge sum 30.1%.

5.2 Second Web Site

Table 5.3: Image analysis for www.spankwire.com.

Image #	Lacunarity	Face detected	Entropy and Uniformity	Analysis based on size of candidate skin regions	Edge sum	Total score
1	0	0	1	1	1	3
2	0	0	1	0	0	1
3	1	0	1	1	0	3
4	1	0	1	1	1	4
5	0	0	1	1	1	3
6	1	0	1	1	0	3
7	0	0	1	1	1	3
8	0	0	1	1	0	2
9	0	0	0	1	1	2
10	1	0	1	1	0	3
11	1	0	1	1	0	3
12	1	0	1	1	0	3

13	1	0	0	1	0	2
14	1	0	1	1	0	3
15	0	0	1	1	1	3
16	0	0	0	1	1	2
17	1	0	1	1	0	3
18	1	0	1	0	0	2
19	0	0	1	1	1	3
20	1	0	0	1	0	2
21	0	0	1	1	0	2
22	1	0	1	1	0	3
23	0	0	0	1	1	2
24	0	0	0	1	1	2
25	1	0	1	1	0	3
26	0	0	0	1	1	2
27	1	0	1	1	0	3
28	1	0	0	1	0	2
29	0	0	0	0	1	1
30	0	0	0	1	1	2
31	1	0	1	1	0	3

Table 5.3 shows results for 31 images obtained from "www.spankwire.com". When three or more of the five criteria were satisfied the classification percentages and the percentage of false negatives and positives were as depicted in Table 5.4.

Table 5.4: Detection percentages for www.spankwire.com

Nude percentage	54.8
Non-nude percentage	44.2
False negative	25.8
False positive	25.8

When each criteria was individually evaluated for its contribution to classifying an image as nude, the lacunarity gave 51.6%, face detection 0.0%, entropy and uniformity 67.7%, analysis based on size of candidate skin regions 90.3% and edge sum 41.9%.

5.3 Third Web Site

Table 5.5: Image analysis for www.tubegalore.com.

Image #	Lacunarity	Face detected	Entropy and Uniformity	Analysis based on size of candidate skin regions	Edge sum	Total score
1	0	0	1	1	1	3
2	1	0	1	1	0	3
3	1	0	1	1	1	4
4	1	0	1	1	0	3
5	1	0	1	1	0	3
6	1	0	1	1	1	4
7	0	0	1	1	0	2
8	0	0	0	1	1	2
9	1	0	1	1	0	3
10	1	0	1	1	0	3
11	1	0	1	1	0	3
12	1	0	1	1	0	3
13	1	0	1	1	0	3
14	1	0	1	1	1	4
15	1	0	1	1	0	3
16	0	0	0	1	1	2
17	1	0	1	1	0	3
18	0	0	0	1	1	2
19	0	0	0	1	1	2
20	0	0	0	1	1	2
21	0	0	1	1	0	2
22	1	0	0	0	0	1
23	1	0	1	1	0	3
24	1	0	1	1	0	3
25	1	0	1	1	0	3
26	0	0	0	0	0	0
27	1	0	1	1	0	3
28	0	0	0	1	1	2
29	0	0	0	0	0	0
30	1	0	1	0	0	2
31	0	0	1	1	1	3
32	1	0	1	0	0	2
33	1	0	1	1	1	4
34	0	0	1	1	0	2
35	1	0	1	0	0	2
36	1	0	1	1	0	3
37	0	0	0	0	0	0
38	1	0	1	1	1	4
39	1	0	1	1	0	3
40	1	0	0	1	0	2
41	1	1	1	1	0	4
42	1	0	0	0	0	1

43	1	0	1	1	0	3
44	1	0	1	1	0	3
45	1	0	1	1	0	3
46	1	1	0	1	0	3
47	1	0	1	1	0	3
48	1	0	0	0	0	1
49	0	0	0	0	0	0
50	0	0	1	1	1	3
51	0	0	1	1	1	3
52	0	0	1	1	1	3
53	1	0	1	1	1	4
54	1	1	1	1	0	4
55	0	0	1	1	0	2
56	1	0	1	1	0	3
57	1	1	1	1	0	4
58	1	0	1	1	0	3
59	1	0	1	0	0	2
60	1	0	0	0	0	1
61	0	0	0	0	0	0
62	1	0	1	1	0	3
63	1	0	1	1	1	4
64	1	0	1	1	0	3
65	0	0	0	1	1	2
66	1	0	1	1	0	3
67	1	0	1	1	0	3
68	0	0	0	1	1	2
69	0	0	0	0	0	0
70	1	0	1	1	0	3
71	0	0	0	1	0	1
72	0	0	0	1	1	2
73	1	0	0	0	0	1
74	0	0	0	1	1	2
75	1	0	1	0	0	2
76	0	0	0	1	1	2
77	1	0	1	1	0	3
78	1	0	1	1	0	3
79	0	0	0	1	1	2
80	0	0	0	0	1	1
81	1	0	1	1	0	3
82	0	0	0	1	1	2
83	1	0	1	1	1	4
84	0	0	0	1	1	2
85	0	0	1	1	0	2
86	1	0	1	1	1	4
87	1	1	1	1	0	4
88	1	1	1	0	0	3
89	0	0	1	1	1	3
90	0	0	0	1	1	2

91	1	0	0	1	0	2
92	0	0	1	1	1	3
93	0	0	1	0	0	1
94	1	0	0	1	0	2
95	0	0	0	1	1	2
96	0	0	1	1	1	3
97	1	0	1	1	0	3
98	1	0	0	0	0	1
99	0	0	0	0	0	0
10	0	0	0	1	1	2
10	1	0	1	0	0	2
10	0	0	1	1	0	2
10	0	0	0	1	0	1
10	1	0	1	1	1	4
10	0	0	0	1	1	2
10	1	0	1	1	0	3
10	0	0	0	1	1	2
10	0	0	1	1	1	3
10	1	0	1	1	0	3
11	0	0	0	1	1	2
11	1	0	1	0	0	2
11	0	0	0	1	1	2
11	1	0	1	1	1	4
11	0	0	1	1	1	3

Table 5.5 contains values obtained on analyzing 114 images from "www.tubegalore.com". When three or more of the five criteria were satisfied the classification percentages and the percentage of false negatives and positives were as depicted in Table 5.6.

Table 5.6: Detection percentages for www.tubegalore.com

Nude percentage	52.6
Non-nude percentage	47.4
False negative	16.7
False positive	21.9

When each criteria was individually evaluated for its contribution to classifying an image as nude, the lacunarity gave 57.9%, face detection 5.3%, entropy and

uniformity 64.0%, analysis based on size of candidate skin regions 79.8% and edge sum 37.7%.

5.4 Fourth Web Site

Table 5.7: Image analysis for www.xnxx.com.

Image #	Lacunarity	Face detected	Entropy and Uniformity	Analysis based on size of candidate skin	Edge sum	Total score
1	0	0	1	1	1	3
2	0	0	0	1	0	1
3	1	0	1	1	1	4
4	0	0	0	0	1	1
5	1	0	0	1	0	2
6	1	0	1	1	0	3
7	1	0	0	0	0	1
8	1	0	1	1	0	3
9	1	0	1	1	1	4
10	0	0	0	1	1	2
11	1	0	1	1	0	3
12	1	0	1	1	1	4
13	0	0	0	1	1	2
14	0	0	0	0	0	0
15	1	0	1	1	1	4
16	1	0	0	1	0	2
17	1	0	1	1	0	3
18	1	0	1	1	0	3
19	1	0	1	1	0	3
20	1	0	0	1	0	2
21	1	0	1	0	0	2
22	1	0	1	1	0	3
23	0	0	0	1	1	2
24	1	0	1	1	1	4
25	1	0	1	1	1	4
26	1	0	1	1	1	4
27	1	0	0	1	0	2
28	0	0	0	1	1	2
29	1	0	1	1	1	4
30	0	0	0	1	1	2
31	1	0	1	1	0	3
32	1	0	0	1	0	2
33	1	0	1	1	0	3
34	1	0	1	1	1	4
35	1	0	1	1	0	3
36	1	0	1	1	1	4
37	1	0	1	1	0	3
38	1	0	1	1	0	3

39	0	0	0	1	0	1
40	1	0	1	1	1	4
41	1	0	1	1	1	4
42	1	0	1	1	0	3
43	1	0	1	1	0	3
44	1	0	0	1	0	2
45	0	0	0	1	1	2
46	1	0	1	1	0	3
47	0	0	0	1	1	2
48	1	0	0	1	0	2
49	0	0	1	1	1	3
50	0	0	0	0	1	1
51	1	0	1	1	0	3
52	1	0	0	0	0	1
53	1	0	1	0	0	2
54	0	0	1	1	1	3
55	1	1	1	1	0	4
56	0	0	0	1	1	2
57	0	0	0	1	0	1
58	1	0	0	1	0	2
59	1	0	1	1	1	4
60	1	0	1	1	0	3
61	1	1	1	1	0	4
62	1	0	0	0	0	1
63	0	0	1	1	1	3
64	0	1	0	1	1	3
65	0	0	0	1	1	2
66	1	0	1	1	0	3
67	1	0	1	0	0	2
68	0	0	0	1	1	2
69	1	0	1	1	0	3
70	1	0	1	1	1	4

Table 5.7 contains values obtained on analyzing 70 images from "www.xnxx.com". When three or more of the five criteria were satisfied the classification percentages and the percentage of false negatives and positives were as depicted in Table 5.8.

Table 5.8: Detection percentages for www.xnxx.com

Nude percentage	57.1
Non-nude percentage	42.9
False negative	21.4
False positive	17.1

When each criteria was individually evaluated for its contribution to classifying an image as nude, the lacunarity gave 70.0%, face detection 4.3%, entropy and uniformity 60.0%, analysis based on size of candidate skin regions 87.1% and edge sum 44.3%.

5.5 Fifth web site

Table 5.9: Image analysis for www.stileproject.com

Image #	Lacunarity	Face detected	Entropy and Uniformity	Analysis based on size of candidate skin regions	Edge sum	Total score
1	0	0	0	1	1	2
2	0	0	1	1	1	3
3	1	0	1	1	1	4
4	0	0	0	0	1	1
5	0	0	1	1	1	3
6	0	0	1	1	1	3
7	1	0	1	1	0	3
8	1	0	1	1	0	3
9	1	0	1	1	0	3
10	1	0	1	1	0	3
11	0	0	1	1	1	3
12	1	0	1	1	1	4
13	0	0	0	1	1	2
14	0	0	0	1	1	2
15	0	0	0	1	1	2
16	1	0	1	1	0	3
17	0	0	0	1	1	2
18	1	0	1	1	0	3
19	0	0	0	0	1	1
20	1	0	1	1	0	3
21	0	0	1	1	1	3
22	1	0	1	1	1	4
23	0	0	0	0	0	0

24	0	0	0	1	1	2
25	0	0	1	1	1	3
26	1	0	1	1	1	4
27	0	0	1	1	1	3
28	0	0	0	1	1	2
29	1	0	1	0	1	3
30	1	0	1	1	0	3
31	1	0	1	1	0	3
32	1	0	1	1	1	4
33	0	0	1	1	0	2

Table 5.9 contains values obtained on analyzing 33 images from "www.stileproject.com". When three or more of the five criteria were satisfied the classification percentages and the percentage of false negatives and positives were as depicted in Table 5.10.

Table 5.10: Detection percentages for www.stileproject.com

Nude percentage	66.7
Non-nude percentage	33.3
False negative	24.2
False positive	15.2

When each criteria was individually evaluated for its contribution to classifying an image as nude, the lacunarity gave 45.5%, face detection 0.0%, entropy and uniformity 69.7%, analysis based on size of candidate skin regions 87.9% and edge sum 66.7%.

5.6 Combined Decision

Since we have so far separately obtained percentages for text and image analysis this section will show how we can combine the two results and do filtering of a web site based on both percentages. Table 5.11 shows a set of results that could

potentially be obtained from text and image analysis when considering best and worst scenario.

Table 5. 11: Combining Text and Image Analysis

Text (T%)	**Image (I%)**	**Verdict**
0	I% < 45	Non-nude
95	I% <45	Nude due to text only
0	45< I% <100	Nude due to images only
95	45< I% <100	Nude due to text and images

As could be seen from the first row of the table, when text analysis fails (0% detection) and image analysis is less than 45% this indicates that the site is non-adult content containing. From the second row, we see the second possibility, where text is detected (is considered to be 95% because the accuracy level for the text analysis was deduced to be so after implementing it on about 25 web sites) and image analysis is below 45%. For such a scenario, the nudity is said to be due to text only. The third row captures the third possibility. Here, nudity is said to be image wise detected since we have 0% text detection and above 45% nudity for pictures obtained from the test web site. The final scenario occurs when the text classification is at 95% and the images detected as nude are more than 45% this would mean the tested site contains both text and images that are adult content containing.

5.7 Comparison with Results in Literature

In order to get a full understanding of how successful the efforts in this thesis has been, it will be best to compare the results here with those obtained from previous works but first, a little insight on the outcome of this work. The result

obtained from this work as presented for (approximately) best case scenario (all the images obtained from the web site contain too much skin exposure) is 80.9% nude and false positive of 22.9%. The number of images in the nude folder and non-nude folder were 47 and 48 respectively. Images were randomly picked from the Internet and the reason for having these amounts is because the number of images downloaded from a web home page is generally in those ranges. Also important to note is the difference between the results presented in this section compared to those in previous sections which showed result for 5 different web sites. The reason for the difference is because a good amount of images from earlier are not nude. Such images can be people kissing with cloth on, people with transparent cloth or worst case sex scene with participants still in their cloth.

From the Duan [24], titled adult image detection based-on skin color model and support vector machine, a detection rate of 80.7% was observed for adult images and a 10% false positive. In another research by Zheng [25], titled shape-based adult image detection, 89.2% adult images were detected and a 15.3% false positive. Lastly, Henry, Jin and Balujah's [17] paper titled large scale image-based adult content filtering, had a 50% adult image detection with a 10% false positive but when the threshold was changed, 90% adult imaged were detected and a false positive of 35%.

By careful observation, it appears the method here is only preferable to the result in the last research paper (by Henry, Jin and Balujah's [17]). But, this is the case before combining the text analysis. For text, we have a 95% detected correctly and 5% wrong. By averaging both text and image, an 88.0% nudity detection is

obtainable with 14.0% for false positives. With this detection rate, the proposed method is worth considering.

Chapter 6

CONCLUSION AND FUTURE WORKS

6.1 Conclusion

This thesis made use of text and image content of a web site to decide whether or not the "said web site" is for adults only. In the text part, a word search was implemented to see if the searched words existed and with what frequency. The image analysis part was much more involving. For starters, skin pixels were detected by color segmentation where two segmentation methods were combined to produce a method called "*improved* YC_bC_r". Since the results were not good enough due to the existence of high amount of skin-like colored objects in pictures on web sites, lacunarity, face detection, entropy, uniformity, edge count and percentages for candidate skin region were also used as new criteria. These classifying criteria had YC_bC_r segmented image as their input except for face detection which made use of the original image. There were five classifiers and each had 20% contribution to decisions made.

As explained at the beginning of chapter 5, the four image categories; namely nude pictures, lion photos, beach scenes and regular everyday pictures were used to optimize the threshold values for the different criteria used in this thesis. After the optimization the classifier would classify nude images with 80.85% accuracy for 47 such images whereas the lion, the scenery and normal every day like pictures would respectively be classified with 83.33%, 60.53% and 79.17%

accuracy haven considered 24, 38 and 48 images respectively. This implies that the miss classification for skin like parts was minimized after optimization of the thresholds.

For the five different adult content containing web sites that was tested the text analysis always gave 95-100% accuracy and the image analysis results for sites 1-5 respectively were 56.83, 54.83, 52.63, 57.14, 66.67 percent accurate.

6.2 Future Works

In a bead to improve this work, there are a few things that could be done. In the text analysis part, the program will not detect anything (even if they exist) if the text to be considered is not written in English. Therefore, translator could be incorporated for future works.

Also important but not considered here are gray images. Clearly, we will run into trouble when classifying a gray image since the program was written to handle colored images. It is not much of a big deal if the site has just a few of the images in gray format because such images are considered to not contain skin. But when we have many (which is rare), then the classification result will be bad. Therefore some future work may be carried out to deal with gray level images that may be of adult nature.

We have noted that a new nude image classification algorithm based on the navel and other body features has been proposed in [5] by Xiaoyin Wang. Since most nude images would have some features like the navel exposed it would be wise to include a new criteria based on body features and further improve the results provided by this thesis.

REFERENCES

[1] X. Qi, B. D. Davison, “Web page classification: features and algorithm,” Lehigh University, June 2007.

[2] H. Yin, X. Xu, L. Ye, “Big skin detection for adult image identification,” *Workshop on digital media and digital content management*, Jiaxing University, China, 2011.

[3] Seminar on skin detection. Retrieved February 16, 2012, from the World Wide Web http://www.slideshare.net/alimadooei/seminar-on-skin-detection

[4] M. J. Jones and J. M. Regh, “Statistical color models with application to skin detection”, In Proc. Of the CVPR ’99, Vol. 1, pp. 274-280, 1999.

[5] X. Wang, C. Hu, S. Yao, “An adult image recognizing algorithm based on naked body detection,” *International Colloquim* on *Computing, Communication, Control, and Management*, Beijing, China, pp. 197-200, 2009.

[6] Y. Wang, J. Li, H. L. Wang, Z. J. Hou, “Automatic Nipple Detection Using Shape and Statistical Skin Color Information”, Advances In Multimedia Modeling, LNCS, Vol. 5916/2010, pp. 644 – 649, 2010.

[7] W. Kim, “A neural network based adult image classification,” *Design Signal & Image Processing : An International Journal(SIPIJ)*, Vol.1, No.2, December 2010.

[8] Mike's Sketchpad, Color Model. Retrieved January 8, 2012, from the World Wide Web http://www.sketchpad.net/basics4.htm.

[9] Color display primer, Retrieved February 17, 2012, from the World Wide Web http://vesta.astro.amu.edu.pl/Library/WWW/Tutorial1/graphics/display_primer.html.

[10] J. A. M. Basilio, G. A. Torres, G. S. Pérez, L. K. T. Medina, H. M. P. Meana, E. E. Hernadez, "Explicit content image detection," in *Signal & Image Processing: International Journal(SIPIJ)*, Vol. 1, No. 2, pp. 47-58, Dec 2010.

[11] T. M. Mahmoud, "A new fast skin color detection technique," in *World academy of science, engineering and technology*, Vol. 43, pp. 501-505, 2008.

[12] J. F. Blinn, "Ntsc: Nice technology, super color: IEEE Computer Graphics and Applications", Vol. 13, No. 2, pp.17-23, 1993.

[13] B. B. Mandelbrot, *The fractal geometry of nature*, New York: W. H. Freeman, pp. 188-189, 1983.

[14] Mandelbrot set, Retrieved February 17, 2012, from the World Wide Web http://en.wikipedia.org/wiki/Mandelbrot_set.

[15] R. Kraft, J. Kaner, "Estimating the fractal dimension from digital images," *Statistics and data processing institute*, D-85850 Freising, Germany, February 1995.

[16] C. R. Tolle, T. R. McJunkin, "Lacunarity definition for ramified data sets based on optimal cover," in *Physica D179 , Idaho Falls, ID 83415-2210, USA*, pp. 129-152, Jan 2003.

[17] H. A. Rowley, Y. Jing, S. Baluja, "Large scale image-based adult-content filtering," USA.

[18] M. Tuceryan and A. K. Jain, "Handbook of pattern recognition and computer vision," pp. 235–276, World Scientific, 1993.

[19] R. Woods, R. C. Gonzalez, *Digital Image Processing Using MATLAB*, 3rd ed.: Prentice Hall, 2007.

[20] P. Viola, M. Jones, "Robust Real-time Object Detection," 2nd international workshop on statistical and computational theories of vision –modeling, learning, computing, and sampling, 2001.

[21] R. Ap-apid, "An algorithm for nudity detection," De la Salle University, Manila, Philippines.

[22] 2600: The Hacker Quarterly, "Google blacklist: Words that Google instant doesn't like," Retrieved February 28, 2012, from the World Wide Web http://www.2600.com/googleblacklist/

[23] R. Nordguist, “The 100 most commonly used words in English,” *Grammar and Compilation*, Retrieved March 14, 2012, from the World Wide Web http://grammar.about.com/od/words/a/100freqused07.htm

[24] L. Duan, G. Cui, W. Gea, H. Zhang, “Adult image detection method based on skin color model and support vector machines,” *Asian conference on computer vision*, pp. 23-25, Australia, Jan 2002.

[25] Q. Zheng, W. Zheng, G. Wen, W. Wang, “Shape-based adult images detection,” *Proc. of the 3rd Inter. Conf. on Image and Graphics*, pp. 150-153, Dec 2004.

[26] B. D. Zarit, B. J. Super, and F .K. H. Quek, "Comparison of five color models in skin pixel classification". In Int. Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems, pp. 58-63, Corfu, Greece, Sep 1999.

[27] B. Randell, “The originals of digital computers,” 3rd Ed, *Selected papers*, Springer-Verlag Berlin Heidelberg Germany.

[28] D. A. Forsyth and M. M. Fleck, “Identifying nude pictures,” *IEEE Workshop on the Applications of Computer Vision*, pp. 103-108, 1996.

[29] D. Chai, and A., Bouzerdoum, “A Bayesian Approach to Skin Color Classification in YC_bC_r Color Space,” *In Proc. of IEEE Region Ten Conference*, vol. 2, pp. 421- 4124, 1999.

[30] D. Chai, and K. N. Ngan, "Face segmentation using skin-color map in videophone applications," *IEEE Trans. on Circuits and Systems for Video Technology*, 9(4): pp.551-564, Jun 1999.

[31] F. A. Merchant, K. R. Castleman, Q. Wu, *Microscope Image Processing*, in *Academic Press*, Apr 2008.

[32] H. Dong, S. C. Hui, Y. He, "Structural analysis of chat messages for topic detection," School of Computer Engineering, Nanyang Technological University, Nanyang Ave, Singapore, May 2006.

[33] H. Zheng, M. Daoudi, B. Jedynak, "Blocking adult images based on statistical skin detection," *Electrical Letters on computer vision and image analysis*, France, Nov 2004.

[34] I. Aldasouqi, and M. Hassan, "Human Face Detection System Using HSV", *In Proc. Of 9th WSEAS Int. Conf. on Circuits, Systems*. Electronics, Control & Signal Processing (CSECS'10), pp. 13-16, 2010.

[35] J. A. Marcial-Basilio, G. Aguilar-Torres, G. Sánchez-Pérez, L. K. Toscano-Medina, and H. M. Pérez-Meana, "Detection of pornographic digital images," *International journal of computers* , issue 2, Vol. 5, pp. 298-305, 2001.

[36] J. C. Russ, *The image processing handbook*, 5th Ed, Taylor & Francis Group, 2007.

[37] J. Canny, *A computational approach to edge detection. Readings in Computer Vision: Issues, Problems, Principles and Paradigms*, 1986.

[38] J. G. Proakis, *Digital Communications*, Fourth Edition, McGraw Hill, 2000.

[39] L. Fei-Fer, P. Perona, "A Bayesian hierarchical model for learning natural scene categories," *IEEE computer society conference on computer vision and pattern recognition*, Vol.2, pp. 524-531, Jun 2005.

[40] M. Fleck, D. Forsyth, C. Bergler, "Finding naked people," *European conference on computer vision*," 1996.

[41] M. Garbarino, "Automatic classification of natural and synthetic images," *Master's thesis*, Royal Institute of Technology, School of Computer Science and Communication, 2008.

[42] M. H. Yang and N. Ahuja, "Detecting human faces in color images ", In International Conference on Image Processing (ICIP), Vol. 1, pp. 127-130, Oct 1998.

[43] M. Lamar, M. Bhuiyant, "Hand Alphabet Recognition Using Morphological PCA and Neural Networks", *Proceedings of International Joint Conference on Neural Networks*, Washington, USA, pp. 2839-2844, 1999

[44] N. K. Ngan, D. Chai, "Locating Facial Region of a Head-and-Shoulders Color Image," in *Third IEEE International Conference on Automatic Face and Gesture Recognition* , Nara, Japan, pp. 124-129, Apr 1998.

[45] O. G. Cula, K. J. Dana, "Compact representation of bidirectional texture functions," *In IEEE computer society conference on computer vision and pattern recognition*, 2001.

[46] Pornography Statistics. Retrieved February 16, 2012, from the World Wide Web http://www.familysafemedia.com/pornography_statistics.html

[47] V. De Witte, S. Schulte, E. E, Kerre, A. Ledda, W. Philips, "Morphlogical image interpolation to magnify image with sharp edges," pp.381-393, 2006

[48] V. Vezhnevets, V. Sazonov, A. Andreeva, "A Survey on Pixel-based Skin Color Detection Techniques," *In Proceedings of the GrapiCon*, pp. 85-92, 2003.

[49] Z. Hussain, "Digital image processing," *Practical application of parallel processing techniques*, Redeood press, Melkshan, Wiltshire, 1991.