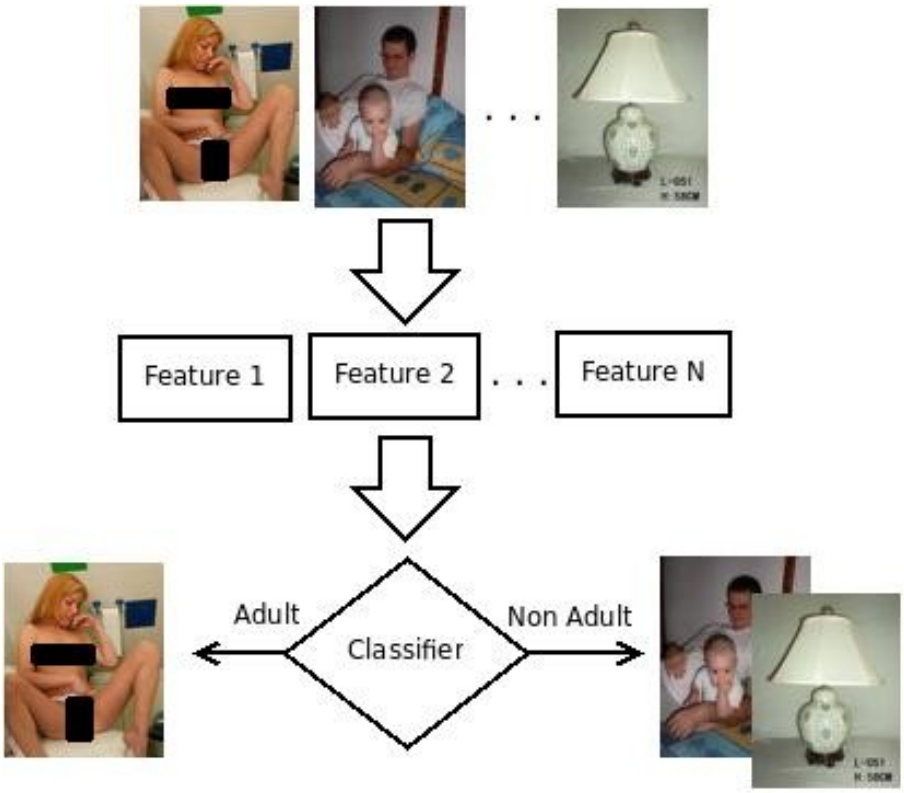
CHALMERS



Large Scale Adult Image Filtering

Master of Science Thesis

AXEL LYCKBERG

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
Göteborg, Sweden, April 2010

The report examines an automated process of filtering out unwanted photographic adult material from the index of a modern image web search engine, using image analysis algorithms.

AXEL LYCKBERG

Examiner: PETER DAMASCHKE

Chalmers University of Technology
University of Gothenburg
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Cover:
Abstract overview of the automated image filtering process. Further detailed in chapter 2.

Department of Computer Science and Engineering
Göteborg, Sweden April 2010

Abstract

In the environment of a modern web picture search engine, enormous amounts of data are being accumulated and indexed. Any manipulation on such amounts of data are bound to be highly efficient to be scalable. Especially when it comes to the rather computational field of image processing. The problem this thesis work have addressed is more specifically; to with a high degree of confidence filter out and separate explicit adult picture material from such a search engine's index.

The work described by this thesis report is two sided. Firstly previous research in the topic have been reproduced in a base line system. By feeding a support vector machine a number statistical measures recorded from each image, a practical level of classification confidence is reached. Secondly a set of new image analysis functionality is evaluated and argued aid the classification rates even further. The proposed system is therefor claimed to show results on par with the state of the art adult image classifiers.

Sammanfattning

I en modern webbildsökmotor insamlas och indexeras enorma mängder data. All slags manipulation av sådana mängder data är bunden att vara högeffektiv för att kunna skala upp. Än mer då det tas i beaktning att bildmanipulation är ett relativt beräkningskrävande område. Det problem som denna rapport adresserat är mer specifikt; att med hög träffsäkerhet filtrera bort stötande pornografiskt material från en sådan sökmotors index.

Arbetet presenterat i denna rapport är tvåsidigt. För det första har tidigare forskning återskapats i ett bassystem. Genom att mata en stödvektormaskin med statistiska mått dragna ur enskilda bilder har en praktiskt användbar nivå av korrekt klassifikation nåtts. För det andra har ny bildanalysfunktionalitet evaluerats och framledes förordats ytterligare förbättra klassifikationsnivån. Det föreslagna systemet gör således anspråk på att visa resultat i linje med de bästa klassificerarna av pornografiska bilder.

Preface

The thesis work this report covers was done at a full time pace from October 2009 through March 2010 at Picsearch AB's office in Stockholm, Sweden. I would like to thank Picsearch for granting me work space and resources needed for this thesis. A special thanks to my supervisor Rickard Cöster who even though he constantly had his hands full with other tasks made time to discuss, encourage and help keep my work on the right track during the course of the work.

I would like to thank Chalmers University of Technology and my examiner there, Peter Damaschke for his time and efforts.

Last but not least I would like to thank the many researchers and developers on who's work the by this thesis proposed system rests upon. Without your contributions the accomplishments of this thesis work would have been but a fraction of what it became. Especially I would like to thank the developers behind the excellent OpenCV and LIBSVM software libraries and Dr. Michael Jones of Massachusetts Institute of Technology, for aiding this work with a copy of his substantial set of labeled skin pixel data.

Contents

Abstract **iii**

Preface **v**

Contents **vii**

1 Introduction **1**
1.1 Motivation 2
1.2 Thesis Outline 2

2 Method **4**
2.1 Goals 4
2.2 Demarcations 5
2.3 Means of Evaluation 6

3 Background **8**
3.1 Image Analysis 8
3.1.1 Universal Feature Detection 9
3.1.1.1 Bag-of-Features 9
3.1.1.2 Discussion 10
3.1.2 Specific Feature Detection 13
3.1.2.1 Skin Color Detection 13
3.1.2.2 Skin Region Shape Anomalies 14
3.1.2.3 Skin Texture Detection 16
3.1.2.4 Identifying Artificial Content 17
3.1.2.5 Face Detection 17
3.1.2.6 Image Size and Dimensionality 18
3.2 Classification 18

4 Image Features **19**
4.1 Base Features 19
4.1.1 Skin Detection 20
4.1.2 Skin Shape 20
4.1.3 Skin Texture 21

4.1.3.1 Hough Lines . . . 21
4.1.4 Image Clutter . . . 22
4.1.5 Size & Dimensionality . . . 22
4.1.6 Entropy . . . 22
4.1.7 Face Detection . . . 22
4.2 Base System Results & Discussion . . . 23
4.2.1 System Disparities . . . 23
4.2.2 Evaluation . . . 24
4.2.3 Concluding Remarks . . . 26
4.2.3.1 No Recalculation of Skin Dependent Measures 26
4.2.3.2 Pre-Filtering Small Images . . . 27
4.3 Further Improvement . . . 28
4.3.1 Further Artificial Discrimination . . . 28
4.3.1.1 Number of Colors Used . . . 28
4.3.1.2 Saturation . . . 30
4.3.2 Illumination Adaptive Skin Detection . . . 32
4.3.2.1 Histogram Equalization . . . 32
4.3.2.2 Clustering . . . 34
4.3.3 Enriching Skin Shape Analysis . . . 35
4.3.3.1 Hu Moments . . . 35
4.3.4 Additional Skin Component Features . . . 36
4.3.5 Hough Line Length . . . 37

5 Design & Implementation 38
5.1 Base System . . . 38
5.1.1 Skin Color Detection . . . 39
5.1.1.1 Skin Histograms . . . 40
5.1.1.2 Skin Probability Histogram . . . 41
5.1.1.3 Testing . . . 41
5.1.2 Skin Component Labeling . . . 41
5.1.3 Texture Analysis . . . 42
5.1.4 Face Detection . . . 42
5.2 Final System . . . 43
5.2.1 Counting Independent Colors . . . 44
5.2.2 Illumination Adaptive Skin Detection . . . 44
5.2.3 Moments . . . 45

6 Experiments & Results 46
6.1 Training & Testing . . . 46
6.2 Results . . . 46
6.3 Large Scale Test . . . 48
6.4 Comparison with Related Research . . . 48

7 Conclusions **50**
7.1 Full Size Images 50
7.2 Further Work 51
7.2.1 Adaptive Skin Color Detection 51
7.2.2 Addressing Graphical & Grayscale Adult Content . . 51
7.2.3 Explicit Area Detection 52

Bibliography **53**

Chapter 1

Introduction

The content of the world wide web has grown out of imaginable proportions. In the case of images the openly reachable or crawlable, currently amount to several billions. A significant portion of this content is arguably considered to be adult[1] photo material. It goes without a saying that it can be of value to be able to separate these images from the other non adult images. Examples of reasons can be obeying applicable laws or gaining the goodwill of being able to market a search engine as porn free. As the content of the image is not explicitly labeled within the image file itself the task must be considered to be non-trivial.

A classic strategy in the field of classifying adult material has been to consider the text material of a given site. Quite effective techniques have been made to filter out adult sites using keywords and other text-based approaches. However this technique only works so far, as sites can be completely made out of images. Or if an otherwise non-adult site contains a single or a few adult images without further stating it. Even a category of sites made to fool these systems is imaginable, as these could gain the privilege of being wrongfully indexed in a non-adult search engine.

Given the increasing speed of modern computers the possibility of using techniques for looking at the content of an image has become more and more interesting even in these large scale settings. In later years a rather substantial amount of research has been made into specifying such systems. Pioneering work was made by (Fleck, Forsyth, & Bergler, 1996) and following that a number of different systems more or less adressing this very problem has been proposed including (Jones & Rehg, 1999; Zheng, Zeng, & Wang, 2004; Arentz & Olstad, 2004; Liang, Scott, & M.Waqas, 2004; Lee, Kuo, & Chung, 2006; Rowley, Jing, & Baluja, 2006; Deselaers, Pimenidis, & Ney, 2008; Lienhart & Hauke, 2009).

[1]For clarification of what is considered to be an adult image in this report. An image is considered adult if it reveals naked explicit areas of the human body. Or if willing, the general western/American definition of adult, without further regard to the context.

1.1 Motivation

In reviewing the current state of research in this topic it can be noted that some of the more recent system are reaching quite confident rates of correct classification of adult images. However when working with filtering the web from this content, every small progress reflect large absolute numbers. This thesis work argues that such progress is still possible. Both in regard to the level of refinement of the existing image analysis algorithms, as well as accounting for the possibilities of improved computational speed. Below a short introductory outline of some topics considered for improvements in regard to prior research is given.

A ground reliance in most previous classifiers is skin detection. Typically a series of different measurements including, skin coverage, shape analysis, and texture analysis are made on the identified skin. This has proven a reliable indicator for adult content. However, a heavy skin dependence is known to, and must be considered a source of some common misconceptions. For one photographs of scantily dressed people, such as bikini or lingerie models are likely to be misclassified as adult. On the other hand adult pictures of significantly clothed people will likely be hard to identify as well. A not previously examined approach to help lift this dependancy, could be to try to identify the explicit areas themselves. Genitals and or the naked female chest could proposedly be identified using techniques similar to that of face detection.

A field which consideration has grown in more recent systems is the identification of images that surely aren't adult. One such category that has been introduced to some extent is artifical image content. One could also imagine other categories worth examining such as scenery photography.

As the topic is and has been quite well researched for some 15 years to date, it has enjoyed a good flora of innovative image analysis methods. Some of the systems have made use of combining a significant portion of these innovations. The possibilities of further synthesization should however not be left unexamined. For one incorporation of one of the several possible automatic image feature detection methods, such as SIFT[2] or SURF[3], could be evaluated. But as well a further combination of different texture and shape description techniques, might turn out to aid inference rates.

1.2 Thesis Outline

The report is split up into six chapters, the first one being this introductory overview. Chapter 2 covers the goals and demarcations of the project. Chapter 3 takes a deeper look into what has been done in previous research

[2]Scale-Invariant Feature Transform
[3]Speeded Up Robust Features

and aims to draw conclusions from that work. Chapter 4 presents the functionality and a reasoning of the system proposed in this work. Chapters 5 and 6 covers the implementation and evaluation of the system respectively. Finally chapter 7 concludes the contributions made from the work as well as suggests possible future enhancements.

Chapter 2

Method

The overall purpose of this project has by a short explanation been, to build a highly efficient classification system for identification of adult photo material. The proposed method of doing this is firstly to extract fast enough image analysis algorithms to find characteristics of such an adult photo. Secondly the characteristics are translated into a vector of features, that can be fed into a machine learning classifier. Finally such a classifier is trained with features extracted from a set of hand labeled images to arrive at as confident rate of binary inference as possible, dividing the adult from the non adult images. An overview of such a system is shown in figure 2.

It is important to consider that the topic is not an exact science, but very much a real world problem and a non-trivial such. In order to make this quite large field of research graspable by a half year full time project, demarcation and prioritization was of course important.

2.1 Goals

The utter most goal of the project has been to with as high degree of success as possible, correctly separate adult from non-adult images. This with regard to a good enough efficiency to be able to work against the image index of a web search engine.

It has been shown that the needed speed of classification in order to address billions of images in a rather sizable CPU-cluster, is a throughput in the order of several images per second and CPU-core (Rowley et al., 2006, Sec. 5.2). Hence the throughput aimed for in this report is at a similar level.

The confidence rate of the classification is regarded as an unbound factor to optimize as far as possible. Notable in regard to this is that the project takes a quite pragmatic approach. Reuse of previously developed algorithm implementations and ideas are considered to large extent. The main focus instead goes into understanding, reusing, mixing and tuning well known libraries and techniques to conclude greater rates of correct classification.

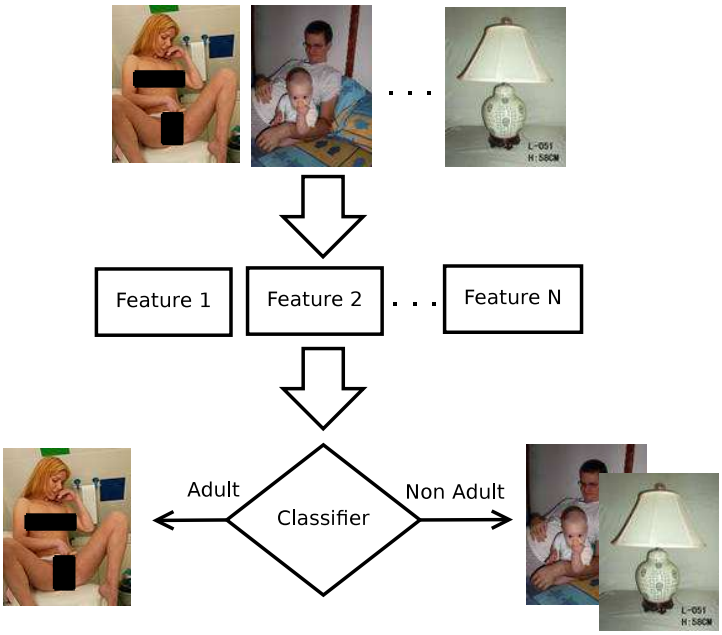


Figure 2.1: Overview of work flow in proposed adult image classification system.

In order to reach good classification rates a general prioritization was made to focus on the image analysis during the course of this work. This was decided due to that results of different machine learning classifiers tended to be rather close, when compared in the reviewed work.

2.2 Demarcations

The aim of the implemented system this report covers was to become a fully working adult image classifier, working in the area of an existing picture search engine[1]. However the only real demarcation made to be able to comply with this environment was that it had to be developed for a *nix[2] environment. Which is probably a reasonable choice anyhow.

Another more relevant demarcation made was to only consider colored photographic images. This makes a separation where two major categories of adult material falls outside the grasp of the project.

The first category covers grayscale or black and white images. The reason these are not considered is that the backbone of the image analysis uses techniques to identify pixels by skin color, see section 3.1.2.1. On the plus side these images can easily be separated from the other as they only contain

[1] `http://www.picsearch.com/`
[2] A Unix or Linux like operating system.

one color channel, the intensity. This oppose to the three or four which colored images use.

The second category, not classified by the system, is the space of artificial or synthetic images. The charcteristics needed to classify these images are too different to the ones proposed in this system. There are however fairly succesfull techniques for separating these images as well, a topic evaluated in the concluding remarks, section 7.2.2.

By separating the artificial and the gray scale images respectively, stand alone classification systems addressing the unique nature of their problems could be developed. However, even if such a separation is possible it is not made in this work. Instead all such images are simply regarded as being non adult.

2.3 Means of Evaluation

The dataset used for test and training purposes was retrieved out of a subset of a crawl[3] of the world wide web. The content was presented in two roughly pre-labeled groups of adult and non adult images, made by text and site classification. Each group consisting of 20000 images. A number of 4450 images was identified as adult from the adult group, and respectively 19984 was found to be non adult from the non adult group.

As stated in the introduction, the prerequisite for an image being adult is the inclusion of a visible explicit region of the human body. The large number of excluded images from the adult group was primarily made up of banners, logotypes and other irrelevant images presumably caught by site classification. However to some extent also consistent of images found not to match the prerequisites of being adult. That is images not showing any explicit areas of the body. Some further characteristics of interest in regard to image set should be noted.

- All images are uniformly converted to JPEG-format[4] and scaled to at most $128x128$ pixels. Both the compression and down scaled representation reduces the image quality. For example the palette used is altered and new colors introduced. Another defect is that sharp color transitions, such as edges, are typically blurred out.

- Although a rough separation was made during labeling, near duplicate images are identified to still exist to some extent. The most noted occurrence of a single such image is five times.

- As a result of the proposed demarcations, graphical and grayscale adult image content has not been included. Both types of content is

[3] An algorithm allowing automatic download of web content by traversing hyperlinks.

[4] JPEG or Joint Photographic Experts Group is a group of standards of lossy image compression, named after the group who created it.

however included in the non adult category. Also worth clarifying is that all various semi-graphical adult images, such as seen in figure 2.2, are included.

All in all the image set is believed to a good degree, reproduce the vast difference in both diversity of motives and photographic conditions of the web.

(a) (b) (c)

Figure 2.2: Examples of semi-graphical images.

Chapter 3

Background

As stated in the introduction, this is a field of research that have had a not unsubstantial amount of input over the last couple of decades. Regarding this matter it would of course be unwise not to make a deep and thorough literature study prior to the implementation of this kind of system. Even more so considering that the area of research is quite large and given the projects pragmatic approach, explained in section 2.1.

This chapter presents a summary of prior work made in the field of adult image filtering. It aims to make a good presentation of what adult image filtering methods and classification techniques that have been used and how they work. It also tries to make relevant argumentation as to what their strengths and weaknesses are. It is not supposed to be a complete in depth guide as to all possible and currently known techniques as that would be a task of vast proportions. Instead it should be considered as an overview of the more influential work in the area.

3.1 Image Analysis

The image analysis is the sole core of any image classification system. It is of utter importance of the problem at hand to get a good representation of image features. The features has to be discriminative to separate an adult from a non-adult image but yet correctly match images inside the two classes as such. Each stand alone feature and ultimately the complete set of features from an image has to be extracted in a computationally quick enough manner. The amount of features must also be kept at a minimum such that the classification step can be made in a computationally tractable way when considering the order of billions of images.

Two different, more or less exclusive approaches to the problem of finding characteristics have been proposed by prior research. A shallow presentation is here given about how they work respectively.

The classic approach have been to by hand made algorithms extract a

set of independent features. The different features are chosen by a human reasoning of what can discriminate an adult picture from another or vice versa. This is typically made combining and tuning known image analysis techniques. A lot of research has been put into system developed in this fashion, working under the same or similar conditions as this project. It has been shown to be able to address the problem to a varying, but arguably high degree of confidence. Due to the combination of single observations of the image content, a modular approach of implementation and measurement can be enjoyed. Ultimately this was the approach chosen for this thesis work that is covered in later chapters. Section 3.1.2 covers this methodology.

In more recent research an alternate approach have been proposed. As its core it considers automatic approaches of finding general image characteristics. These features are then further generalized into bins or bags considering several very similar but not equivalent characteristics. Histograms are created counting the amount of these features and finally compressed by one of a number of techniques to reduce the amount of dimensions that has to be considered by the classifier. These techniques argues for a great increase in confidentiality in the classification, (Lienhart & Hauke, 2009, Sec. 6) and (Deselaers et al., 2008, Sec. 4). However they were not realized or considered outside this background chapter because of a number of reasons, discussed in 3.1.1.2. An overview of the technical bits of this family of systems is presented below.

3.1.1 Universal Feature Detection

A number of different sets of techniques used for automatic extraction of local features in images have been presented in research. They tend to be built up in a similar kind of fashion. First of they locate possible points of interest using one of several techniques. Secondly they manipulate a small area around these points in order to yield a representation of the local interest point that is as characteristic as possible but still invariant to scale, rotation and/or illumination. Looking at these patches independently invariance towards absolute position in the original image is also retrievable. Thirdly the calculated information is stored in a high-dimensional feature vector which can be used for later classification. An early and still highly considered such feature detector is the popular SIFT algorithm, (Lowe, 1999; Lowe, 2004).

3.1.1.1 Bag-of-Features

In order to efficiently classify using the large amount of features extracted from these universal techniques, a model commonly considered in text classification has been borrowed. Introductory work in image processing with the model, called bag-of-words, was made in the field of texture recognition

(Cula & Dana, 2001). It has later been popularized by its success in the field of scene classification, automatically categorizing scenery photographs after topics (Fei-Fei & Perona, 2005; Sivic, Russell, Efros, Zisserman, & Freeman, 2005).

The analogy of the model is to consider the unsorted set of words (here features), without regard to location inside the document (image in this context). In order to retrieve a feasible set of visual word equivalents some sort of discretization or clustering has to be made to the feature description vectors. The set is also referred to as the vocabulary.

The purpose of considering the bag-of-features model is that a number of well known techniques can be utilized in classifying the topic of the document. An example of an early successful application of such techniques would be spam filtering. Where typically a naive Bayes classifier is used to calculate the probability of spam, conditional to the word frequencies. Other more complex classification methods have since been developed for the model. Two of these, pLSA[1] and LDA[2], have been used in the reviewed work. Both of them automatically generates hidden topics to which sets of specific words are reduced. The set of words extracted from a document will hence be reduced by a vector of probabilities that the image falls into the different categories.

The resultant vector has historically been used directly to infer multiple categories. In the reviewed work of adult image inference however, it is used as input for a final binary classifier. The purpose of the first classifier is instead to reduce dimensionality for and argued to improve the results of the final classifier (Lienhart & Slaney, 2007).

3.1.1.2 Discussion

As stated initially in the current chapter, this family of classifiers of universal features was chosen not to be considered in the implementation of the system. The reason for this was due to a number of uncertain factors that would have to be tackled. Weighed against the limited time span of this project. Instead the more well documented and modular approach presented in section 3.1.2 was chosen for the implementation. Below a shorter explanation of considered uncertainties are presented.

Scalability Problems

Generally considering this family of universal feature descriptors it should be argued that they are likely to be more CPU-demanding than their alternative. Which would be to extract features in regard to the specific purpose of the classification.

[1] Probabilistic Latent Semantic Analysis
[2] Latent Dirichlet Allocation

First of, the extraction of universal descriptors would likely have to be of high dimensionality to describe any of a large number of possible local characteristics. Considering the same generic nature, a larger amount of samples would likely have to be chosen to confidently discriminate classes of images apart.

Secondly the larger amount of features would propagate to the stage of classification. As noted earlier in 3.1.1.1, this has been argued to boost the confidentiality of the results even more. But nevertheless it would still make for increased computation time and maybe even a considerable memory footprint.

In (Lienhart & Slaney, 2007) a larger scale system utilizing the presented techniques is considered. However no large scale runs or thorough discussion on the actual scalability of the system is proposed. In the final chapter a statement about wanting to verify the result on a dataset in the order of millions of images was presented as a conclusion. Nevertheless classifying a few millions of images is not the same task as classifying tens of billions.

To summarize. Given the systems presented in (Arentz & Olstad, 2004; Rowley et al., 2006), both presenting absolute figures of image classifications per second and CPU-core. Comparing their respective methods of feature extraction and means of classification. It is reasonable to believe that the universal approaches could have problems to scale up to the ad hoc systems.

Possible Thumbnail Issues

In non of the research papers found on the topic is it explicitly stated what kind of dimensions the images of the training and test sets are having. Judging implicitly from their content it is however likely to conclude that they are using more or less high resolute picture material. (Lienhart & Slaney, 2007) considers pictures from a web based picture gallery and the others hand labeled sets of pictures foretelling the scene of the picture. Irrespectively it is unlikely that they regard thumbnail images.

In the context of this report the image sets are at a maximum of 128 pixels in any dimension. This must as well be considered the general context when regarding image filtering in the indexes of an all purpose picture search engine. Likewise (Rowley et al., 2006) considers thumbnails of 150 pixels when considering their sets in Google's image index. The universal feature representations presented is stated to be scale invariant. However it must be a relevant question to ask to what extent.

As an example regarding (Lienhart & Hauke, 2009). The work claiming the yet highest degree of correct classification. In their successful blend of techniques they extract about 500 points of interest. At each point of interest a feature is created by testing how similar a $5x5$ square of the pixel is regarding a neighborhood of $42x42$ pixels. Considering the largest possible square image dimension of $128x128$ pixels approximately:

- 3% of the image pixels would be points of interest.
- An amount equivalent of 76% of the image pixels would be considered as interest regions.
- The equivalent surface that the neighborhood pixels could produce, would be able to contain the thumbnail almost 54 times around.

To conclude; its probable that the extracted feature data would be quite redundant.

Uncertain Factors of Implementation

The system based on the principals above carries a greater number of uncertain factors of implementation time. To get a first up and running system, a substantial chain of tools would have to be linked together.

- A proper feature detector, including point of interest localization.
- Some kind of discretization or clustering of the features to visual words during learning.
- Distance algorithms for deducting what visual word an extracted feature represents.
- A middle classifier, reducing the visual words to a smaller set of topics. Including algorithms for both training and testing.
- A final classifier.

Generally the design of such systems would be more complex and less modular than the systems described in section 3.1.2. For example, in the later system it would be easier to both extend and substitute a single feature extraction method. In the former system the glue code such as the distance algorithms would probably have to be significantly altered upon a substitution of feature detector. Furthermore, software libraries of the later system are sparse and to a greater extent undocumented. For example, no suitable implementation of the pLSA-technique was found.

To make a conclusion, there are higher degrees of uncertainty in how many hours of implementation this system would require. Even a produced system would likely be quite unmodular, which could be devastating if it's shown that the used techniques are too slow or does not produce the expected results.

3.1.2 Specific Feature Detection

In difference to the previously examined methodology of universal features, most of the research in the topic have leaned towards extracting specific discriminating features from images. The features are hand made for the purpose, with a reasoning of why they would help separating the adult from the non adult content. This will typically result in a small number of such characteristic features, which can be fed into one of a number of possible machine learning classifiers.

The identification of naked human skin has proven invaluable in this methodology. In fact all the reviewed research from the pioneering work of (Fleck et al., 1996) through (Rowley et al., 2006) uses skin detection at a foundational level of their system. Further some analysis of the areas identified as skin is usually present.

The big benefits of this methodology include a provenly scalable image classification (Rowley et al., 2006, Sec. 5.2). As well as the benefits of modularity in terms of extending and interchanging parts of the system, allowing varying implementation time, and being able to run different tests of individual feature contributions.

3.1.2.1 Skin Color Detection

It has been shown that the color of skin resides in a small and compact cluster of the total color space (Jones & Rehg, 1999, p. 5-7). This pleasant fact has made it possible to with a rather high degree of confidence being able to classify single pixels, as skin or non-skin at a very low, constant time cost. Larger research fields such as face and people detection and recognition have made use for it. Even more so, it has been shown to alone be a very good indicator in adult image classification. This has made it a core element in many such classifiers and in all provenly web scalable such classifiers reviewed in this work.

Because the skin pixel detection has proven itself useful for addressing a number of problems, substantial research has been made in this topic alone. One of the more popular and fairly successful methods, have been to divide color channels and accumulate by value into single color histograms. From this try to produce a set of rules to foretell in what ranges skin lies. Different color spaces for this have been proposed in order to get more accurate rules. For example the YCrCb[3] color space have been proposed as a good color space, due to the fact that some or all of the lighting conditions can be disregarded (Arentz & Olstad, 2004; Mahmoud, 2008). However rules of matching confidence rate in RGB has been proposed as well (Kovač, Peer, & Solina, 2003).

[3] A linear transformation from the usual RGB color space. Represents brightness (luma) in one channel and uses two channels for color.

The work of (Jones & Rehg, 1999) has by many been considered ground breaking. In this work, two three dimensional histograms were produced. One for skin and one 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 histograms correspondingly, the conditional probability histograms (3.1) and (3.2) 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.

$$P(\text{rgb}|\text{skin}) = \frac{s[\text{rgb}]}{T_s} \tag{3.1}$$

$$P(\text{rgb}|\neg\text{skin}) = \frac{n[\text{rgb}]}{T_n} \tag{3.2}$$

From these histograms a naive Bayes classifier could be produced as in (3.3). With Θ denoting a chosen threshold of acceptance. The results from this classifier was argued to outperform prior research and rules sets. One important argument being to include non-skin probabilities. Finally the authors presented an approximative model consisting of sixteen Gaussian curves. This model was shown to almost but not quite produce a result as good as the straight ahead histogram classifier (Jones & Rehg, 1999, p. 11).

$$\frac{P(\text{rgb}|\text{skin})}{P(\text{rgb}|\neg\text{skin})} \geq \Theta \tag{3.3}$$

Noticeable about these methods is that the classification gives a real scalar as classification medium. This opposed to the binary yes or no of the priorly discussed rule approaches. The later would have to manipulate their rules to obtain varying rates of precision and recall and thereby lose their simplicity. Possibilities of a rate against a binary classification could be for example; parameterized change of acceptance and computation of more complex mathematical features for the final classifier.

In taking the research field further a popular approach have been to include illumination conditions in the inference. In the work of (Zeng, Gao, Zhang, & Liu, 2004) the images are divided in three groups, given their average brightness, prior to accumulation in histograms. In another approach proposed by (Zheng et al., 2004), the images divided into groups based on both average brightness and chromaticity[4] using the k-means algorithm.

3.1.2.2 Skin Region Shape Anomalies

A rational question to raise, after retrieving a skin mask of an adult picture, would probably be if it is possible to draw any discriminating conclusions

[4]In that work defined as the two dimensional value $T = (R/B, G/B)$

out of the shape of the skin components in the mask. Not surprisingly a number of different techniques for analyzing the geometric shape of pixel components exist. The theory behind component shape analysis spans from simple ratios between perimeter and area, to fitting such components into geometric figures and to scale, rotation and translation invariant moment calculations[5]. It is included in some way or another in all the reviewed work to varying degrees of extent. In the early work of (Fleck et al., 1996) a quite complex method for detecting human limbs was developed, however producing rather fair results. In the more recent work of (Zheng et al., 2004) a review of different shape descriptors suitable to extract as discriminating features was presented.

Shape Description Algorithms

As the only shape descriptor used in (Rowley et al., 2006) a measure of the compactness of the skin components is included. The compactness measure is calculated by recording a weighted average of the ratios between the length of the components perimeters and their areas. The thought behind the feature is that pixels identified as skin of unclothed human bodies would probably tend to be more compact then falsely identified pixel areas, such as for example brick buildings.

Except a compactness measure (Zheng et al., 2004) proposes two other simple shape descriptors. The first one is eccentricity, here calculated as the length ratio between the major and minor axes of a component. In this context eccentricity can be seen as a measure of how much the component, seen as an elliptic shape, deviates from being perfectly circular. Not stated in the paper is from what point the diameter lengths to the axes are recorded. Both the figures centroid (or center of mass) and the center of the best fitting ellipse would be reasonable points to choose. As the second simple shape descriptor a ratio between the areas of the figure and its bounding box, in example its best fitting rectangle, is recorded.

A different approach for extracting shape features is to make various moment calculations of the skin mask in the components. The moment of an image is a mathematical moment of second dimension. For a binary image the moment generating function is defined as in (3.4), where x and y are pixel coordinates, $I(x, y)$ is the intensity of a given pixel (here either 0 or 1) and M_{ij} is the order of the moment. Noticeable is that M_{00} equals the area and $\{\bar{x}, \bar{y}\} = \{M_{10}/M_{00}, M_{01}/M_{00}\}$ denotes the centroid of the component.

[5]Such transformations are typically used in automatic image similarity methods such as SIFT and SURF for representing the content found in the points of interest. See prior section 3.1.1

$$M_{ij} = \sum_x \sum_y x^i y^j I(x,y), \quad I(x,y) \in [0,1] \tag{3.4}$$

Moments of a component can instead of being calculated absolutely, be calculated relative to their centroid. This gives the property of being translation invariant, that is not regarding absolute position of component, stated by (Hu, 1962, Sec. II-C). Such moments are known as central moments, generated by the function (3.5). Where μ_{pq} is the order of the moment, $\bar{x}$ and $\bar{y}$ are the centroid coordinates.

$$\mu_{pq} = \sum_x \sum_y (x-\bar{x})^p (y-\bar{y})^q I(x,y), \quad I(x,y) \in [0,1] \tag{3.5}$$

From the central moments, (Hu, 1962) showed that it was possible to derive moments that had the properties of being invariant to rotation, position and scale. He proposed a set of seven such moments for discriminating different shapes. The set has been named after him, simply called the Hu-moments. Up to this set, a second one known as Zernike-moments was proposed by (Zheng et al., 2004). They are built using another rationale but are as the Hu-moments translation and scale invariant.

The results suggest that the two moments do not produce completely redundant information, as they are shown to aid each other in the classification. The addition of simple geometric shape descriptors boosts the confidence even a bit more. Nevertheless its shown that the Hu-moments makes the single most important contribution to the classification rates and outperform the Zernike-moments where comparable.

3.1.2.3 Skin Texture Detection

Another approach of discriminating true skin components from other components with skin color classified by the skin color detection, is to look at their texture. The hypothesis is that areas of skin will turn out to be rather smooth, whilst for example rocks, bricks or wood have a more rough texture. The suggested ways for discriminating these surfaces are quite numerous.

The work of (Zeng et al., 2004) used a map of skin probabilities produced by a Bayesian histogram classifier. By computing the local variance of a small kernel around every pixel probability, the map was thresholded at a statistically arrived level of accepted variance. The produced binary skin mask was then suggestively more confident to contain human skin. This is utilized for computing better results in a later component shape analysis step.

(Fleck et al., 1996) proposed reducing misinterpreted skin pixels by running the intensity image through a smoothing median filter. By subtracting the filtered image from the original intensity image, the recorded difference

act as a measure of texture amplitudes. An area with a lot of high texture amplitudes would suggest a rough texture.

Some of the reviewed research including (Rowley et al., 2006; Liang et al., 2004) makes use of the popular Canny edge detection algorithm (Canny, 1986). In the former paper, the amount of edge pixels found in the skin areas is weighed against the total amount of pixels in the same areas. This is argued to give a measure of how rough a surface the identified skin area would have.

An approach analyzing the co-occurrence matrix of the intensity image is suggested by (Liang et al., 2004). Even yet another strategy is used in (Arentz & Olstad, 2004).

Locating Straight Lines

To further aid the texture analysis, (Rowley et al., 2006) propose identifying long straight line segments in the Canny edge map inside the skin components. The thought is that human skin will typically contain round-shaped edges, as opposed to for example smooth wooden panels or brick patterns. The location of such lines is proposedly found using the probabilistic Hough transform algorithm of (Kiryati, Eldar, & Bruckstein, 1991).

3.1.2.4 Identifying Artificial Content

Historically several measurements have been proposed to infer whether a given image contains photographic material or if its made artificially. In fact this is considered an interesting image retrieval topic alone. Regarding the arbitrary nature of images sampled from the web, it must also be considered relevant for an adult image classifier.

One strategy is to analyze the entropy of intensities present in an image. This can be done by accumulating the intensities in a histogram and record the spread of their distribution as a feature (Rowley et al., 2006). The reasoning being that graphical content have a tendency to favor a fewer amount of intensity levels.

3.1.2.5 Face Detection

Detecting if a person is present in an image can obviously be aided by trying to identify if faces are present. (Rowley et al., 2006) propose using face detection for two purposes. First of they count the amount of identified faces in the picture to record as a feature. Secondly they remove the area recorded as a face. Using this they recalculate their skin dependent features, aiming to reduce the false classifications of portrait pictures and other images with a lot of face skin.

Used to attain this task is a highly efficient face detection algorithm developed by (Lienhart & Maydt, 2002) which is merely an extension on the

work of (Viola & Jones, 2001). The algorithm records Haar-like features[6] by transforming the image into what they call an integral image. The Haar-like features are used to build a cascade of weak classifiers, to which the AdaBoost algorithm is applied to learn if a face is present. Extracting the features of the integral image is fast enough to do at different scales and positions. Which allows for an algorithm to search for a face region locally in the image. To conclude; a computationally efficient algorithm that identifies face regions locally within an image, while working completely invariant to skin color detection.

3.1.2.6 Image Size and Dimensionality

Typically dimensionality information of the image is recorded somehow into the adult image classification system. First of all its argued reasonable to conclude that a certain image size is needed to be able to identify nudity and also that such small images are typically icons anyway. Different image sizes have been proposed, spanning from $10x10$ to $32x32$ pixels. Secondly certain dimensions are probably more countable towards artificial images, such as those of bars and banners. A proposed feature to include has therefor been the aspect ratio.

3.2 Classification

In the reviewed research a broad variety of machine learning algorithms have been proposed to address the task of classification. The more frequent classifiers include artificial neural networks (ANN), C4.5 decision trees with boosting, and support vector machines (SVM). The few works that include comparisons show that the performance rates on equal data are rather close, (Zeng et al., 2004; Zheng et al., 2004). Based on this observation and belief a general decision to focus primarily on feature extraction was made early on.

A decision was made to use SVM-classifier with a radial basis function kernel. As this setup both had proven successful in (Rowley et al., 2006) as well as being available in a well suited software package.

[6] The features bear resemblance to the Haar Basis function.

Chapter 4

Image Features

As a first step of development a decision was made to conclude the work of (Rowley et al., 2006), in a baseline system. The reasons of this choice of research are several. Firstly the system is quite recent, which has allowed the authors to sum up and consider several of the older systems for their research. Secondly they have incorporated a broad variety of approaches for feature extraction. Thirdly they have a well specified training and test set, containing thumbnail images downloaded from the web in similarity to this thesis works setting. Fourthly they have a results discussion, including comparison of their and previous systems results. Fifthly they have reasonable enough level of explanation of implementation details, to be able to reproduce most of their system.

This chapter covers the logic and individual results of the image features. Section 4.1 covers the details of the base system features. The following section 4.2 concludes the performance of the base system and presents a discussion about possible shortcomings. The final section 4.3, is introduced by an empiric study of tendencies in the misclassified images. In regard to this, a number of extending contributions are proposed, individually evaluated and discussed. This section is to be considered to contain the primary source of contribution of this research.

4.1 Base Features

The here proposed base system has a total of twenty-seven features recorded in the image analysis in one case. Of these, eighteen are uniquely implemented leaving the first nine to be recorded twice, the second time after face removal. An overview of differences and possible misconceptions in regard to the original work is presented in section 4.2. The reader interested in technical or implementary details of the system is referred to chapter 5.

4.1.1 Skin Detection

As stated in the background chapter, the skin color detection can be viewed as the backbone of most adult classification systems. This one is no exception. In this proposed base system not less than eighteen of the total twenty-seven features depend on the classification of skin pixels.

To detect skin the base system utilizes a straight forward implementation of the Bayesian skin color histogram approach of (Jones & Rehg, 1999). As explained in the background chapter each pixel is being considered independently. A probability of given pixel being skin is inferred by looking at statistically accumulated conditional probability of its color combination being skin, against it not being skin. To build the histograms of conditional probabilities needed by this model a training set of images with labeled skin areas and a set of non skin images was needed. Luckily the set used in the original model was kindly shared by its authors, for the benefit of this thesis work.

A map of per pixel skin probabilities is induced from the given input picture. The map is thresholded at a rather accepting level, only to reduce the most obvious false positives. A rate of 90% true positives was chosen, at which 16% false positives was being let through. Two features are recorded from this map. The mean skin color probability and the standard deviation. Apart from indicating high skin probability its reasoned to help characterize the supposedly small probability deviations of human skin.

4.1.2 Skin Shape

A second skin probability map is considered for the purpose of skin shape analysis. This map is thresholded more tightly, as it aims for a higher level of confidence of human skin. The rate of 9% false positives yielded 80% true positives in a test run, a setting chosen for the task. To further reduce noise and small gaps in the skin areas, the map is run through a step of grayscale erosion followed dilation.[1] The resultant skin map is used to identify and label all connected components. Four features are recorded from the skin components.

- The total amount of connected components.
- The mean skin probability of the skin pixels inside the components.
- The standard deviation of the skin probabilities inside the components.

[1] Erosion and dilation are two opposite morphological operations commonly used in image processing. Erosion computes, for each pixel, the local minimum of its neighboring pixels. Dilation, instead computes the maximum. In this work the neighborhood was considered a square of $3x3$ pixels.

- A weighted average of compactness for skin blobs larger then 100 pixels. The compactness is measured as the ratio between a component's area and its perimeter. To promote larger components the average is defined as in equation (4.1). Where C denotes the set of skin components with an area greater than 100 pixels. The reasoning is that large areas of naked human skin is likely to be compact and hence weigh in for a large number.

$$F_{\text{compactness}} = \sum_{c \in C} \frac{\text{area}_c}{|C| \cdot \text{perimeter}_c} \tag{4.1}$$

4.1.3 Skin Texture

As a measure to discriminate the smooth texture of human skin from other more rough skin colored surfaces, the system relies on the Canny edge detection algorithm (Canny, 1986). The idea of a pixel being an edge is basically a considerable change of intensity[2] in regard to neighboring pixels. The Canny algorithm involves multiple steps in order to confidently identify as many edges as possible.

Without further going into technical details it uses three parameters. Out of these two represent a high and low threshold of finding and linking edges. As the image qualities of the web differs greatly, a quite accepting setting of 10 and 100 were chosen for these parameters. The third and more relevant parameter regulates the size of the Gaussian filter. A relatively small filter of size $3x3$ was chosen as smaller filters aids location of sharper edges.

Two features are recorded from within the skin components described in the previous section. The first one measures the percentage of pixels in the skin components being edge pixels. The second one records the percentage of edge pixels that resides within the skin components.

4.1.3.1 Hough Lines

One additional feature is recorded from the edge map inside the skin components. The amount of long straight line segments located by the probabilistic Hough transform (Kiryati et al., 1991). The idea is to discriminate sharp edged objects from the supposedly round shapes of the human body. A parameter choice of counting lines longer than 30 pixels found in a 10 pixel gap was found as a reasonable compromise by an empiric overview of sample images.

[2]Intensity is the single value of a grayscale pixel. An intensity image is an equivalent term to a grayscale image.

4.1.4 Image Clutter

To measure the general clutter in the image, two additional calculations on the Canny edge map are made. Firstly, to record the central clutter, the image is cropped by 1/6 of its size at each side. That is, the image is cropped by 1/6 of the width from left and right and 1/6 of the height from the top and bottom. The fraction of pixels being edge pixels in the central region is recorded as a feature. Secondly the fraction of edge pixels of the whole image that's residing in the central region is recorded.

4.1.5 Size & Dimensionality

Images of sizes smaller than 10 pixels in any dimension are considered too low resolute to be able to depict any content that could be argued adult. A binary feature is set to take note of any such tiny pictures. All hereafter following features are only calculated for the larger images.

Two more features are recorded at this stage. Firstly the image size is recorded as the logarithm of the number of pixels in the image. Secondly the aspect ratio is measured as the image width divided by its height.

4.1.6 Entropy

To aid separation of graphical or artificial images such as cartoons and logotypes from natural, a measure of intensity entropy is made. This is done by summing up the products of the occurrence frequency of each intensity level and its logarithm, as noted in equation (4.2). Where H denotes the grayscale histogram and H_i a given histogram bin or intensity frequency.

The measure is recorded twice, the second time inside the region produced of cropping each side by 1/4 of the image size. This is done to help discriminate any artificial background images with centered photographic scenery.

$$F_{\text{entropy}} = \sum_{i=0}^{255} H_i \cdot log_2(H_i) \tag{4.2}$$

4.1.7 Face Detection

The finding of a face in an image must inarguably be considered a good indicator human presence. The highly efficient and skin color invariant face detection algorithm presented in the works of (Lienhart & Maydt, 2002; Viola & Jones, 2001), is used for two purposes. Firstly by recording the total amount of detected faces for a given image. Secondly the algorithm allows identification of a rectangular region around the faces. This functionality is used to record the fractions of image pixels residing in the largest face

rectangle. It is also used to recalculate all skin dependent features for images with one or more faces present, this time with face regions blackened out.

4.2 Base System Results & Discussion

This section aims to give a full analysis of the base system made in this work. Below a discussion is provided of the known differences between the systems. This is followed by a replication of the experiments made in the original work. Finally in section 4.2.3, two conclusive suggestions are made to yield the system used as base for the extensions proposed by this thesis work.

4.2.1 System Disparities

As stated earlier, the aim of the base system was to conclude the research of (Rowley et al., 2006). In general the above presented system is believed to amount to this task. Most of their features or elsewhere general concepts of their work have been reproduced. However a full replication of their exact system can not be claimed. Such a task would be impossible as not all parameters are given, nor is the test data equal.

The cropping of the image into a central region of interest have not been made in the shape, texture, and image size measurements. For one because of the smaller thumbnail size than their setting. Secondly because an empiric overview of the adult content argues that the full area of the picture is quite frequently used. Hence its believed that a cropping could well remove seemingly relevant data.

The weighed average compactness measure presented in this model (4.1), differs from the vaguely stated original. It's however argued to give a corresponding measure. The exclusion of small components was reasoned to remove some noise.

A more important difference to the original system however, may be the exclusion of adaptive skin color detection. As a part of face detection they proposed the creation of a 3-channel Gaussian mixture model built on the pixels found in the face region. The reasons to be skeptic are several to the aiding degree of such an inclusion. Firstly the square region of the detected face is often too large, including areas of non-skin pixels. Examples of this are shown in figure 4.1. Secondly the face region on a thumbnail picture must be argued to contain a troubling small amount of pixels. Already in (Jones & Rehg, 1999, Sec. 3.3) it's shown that small amounts of training data significantly lowers the rates of correct classification when considering the full set. Varying lighting conditions could arguably turn out to be a problem, even when considering different regions of the same person in the same picture. Thirdly a small statistical survey shows that only about

10.1% of the adult pictures are found to have one or more faces. To conclude; it was believed that this skin detection model generally would have a hard time identifying skin properly. Specifically it was reasoned that the space of face detected images with previously undetected skin that would now be corrected, would turn out to be near negligible. Weighing these reasons against what was believed to be a rather tedious implementation, the functionality was chosen not to be considered.

(a) (b) (c)

Figure 4.1: Examples of successfully detected face regions that include non skin pixels.

4.2.2 Evaluation

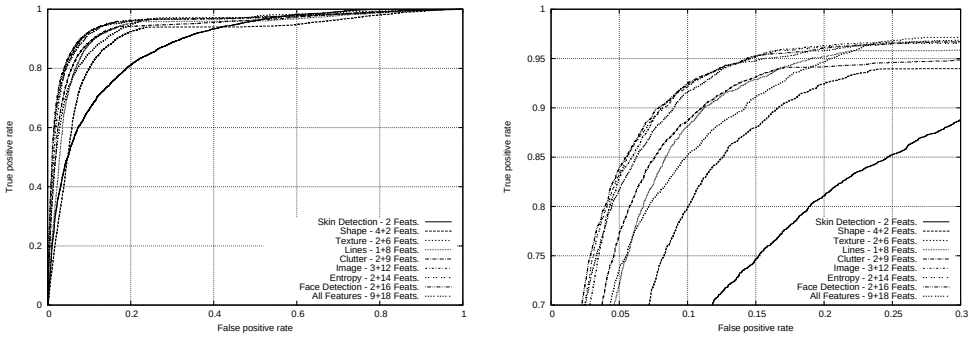
To conclude the results presented in (Rowley et al., 2006, Sec. 5) a test setup was built up in a corresponding manner. All accuracy measurements presented below have been recorded using five-fold cross validation[3] on the full dataset. Using a weight parameter of 4, in favor of skin to approximately even out the proportion of negative to positive examples. Note that absolute rates can not be compared in a fair manner due to the differences in datasets. As opposed to this works dataset, explained in section 2.3, the authors of the original work supposedly labeled all their sample images as either adult or non adult and as well included grayscale and graphical adult content. Although the difference certainly contributes to greater absolute inference rates here, the tendencies of the systems should still be comparable.

[3]K-fold cross validation is a strategy used to validate the performance of a statistical model, without separating an independent set used only for testing. Firstly the full set of samples are randomly split up into K subsets. Secondly K different models are trained, each model considering 1 subset for testing and the other K-1 for training. Finally the results of the K models can be combined to produce a single estimation.

Features	Amount	Extrac. (s/kimg) Adult	Extrac. (s/kimg) ¬Adult	Class. (s/kimg) Adult	Class. (s/kimg) ¬Adult	Thr.put (img/s)	Support Vectors
Skin	2	1.54	1.11	0.84	0.86	490	10852
Shape	4+2	2.60	1.88	0.73	0.71	370	6640
Texture	2+6	3.48	2.62	0.75	0.71	290	5930
Lines	1+8	7.17	3.12	0.78	0.77	220	6201
Clutter	2+9	7.33	3.19	0.85	0.79	210	5270
Image	3+11	7.40	3.27	1.18	1.15	190	6577
Entropy	2+14	7.41	3.30	1.49	1.49	180	7468
Faces	2+16	19.7	12.4	1.63	1.64	65	7733
All	9+18	20.3	12.7	2.31	2.26	61	8290

Table 4.1: Comparison of the impact of adding features to the SVM-model. Features are added per row basis. *Note that timing measures were made with images in cache, not considering cold disk read. In order to make a fair comparison, the median out five runs was used in all time measures.*

Figure 4.2 presents a ROC plot[4] of classification rates by the addition of features. A corresponding plot can be found in the original work. In addition, some other statistics of interest regarding the models have been recorded in table 4.1.



(a) (b)

Figure 4.2: ROC curves showing the contributions of features by addition in the baseline system. Figure 4.2b depicts a zoomed in version of same plot.

Hardly visible in the plot, the "All Features" model, which re-records all skin dependent features is not the best performing model. Instead the "Face Detection" model is shown to give the overall best classification accuracy. As seen in the table it is also a computationally lighter model, in terms of the number of feature vectors used as support vectors. Two primary possibilities

[4]ROC or reciever operating characteristic is a plot of the rate of true positives vs. false positives, as its discriminatory threshold is varied.

of this contradictory behavior has been reasoned. First and most important must be that the face adaptive skin color detection not implemented would indeed aid the results. This in contrary to the argumentation previously presented in section 4.2. Secondly the additional types of content included in the original test set should be accounted.

4.2.3 Concluding Remarks

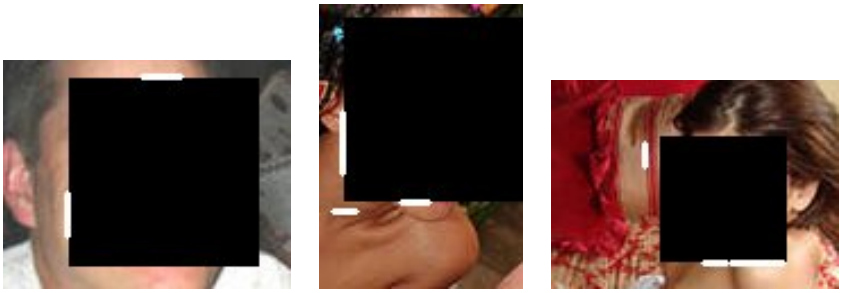
As stated by the original authors the base system shows to produce good classification rates at a speed feasible for large scale purposes. Nevertheless a few of the used conceptions and features can be argued to be of less importance for the good performance. In order to test new concepts and techniques it is also reasoned as good practice to try to eliminate any under performing features. Primarily to reduce the complexity of the final classifier, but as well to some extent reduce computation time during feature extraction.

Presented below is a group of changes and feature exclusions found reasonable to consider for varying reasons. In making these changes a reduction from at most 27 down to 17 features can be reached. In order to conclude that an acceptable loss was made in terms of classification rates, new models were trained in the same manner as with the base system evaluation.

4.2.3.1 No Recalculation of Skin Dependent Measures

In figure 4.2 it was shown that the inclusion of the nine recalculated skin dependent features did not prove to contribute to better accuracy in the base model. Two reasons were presented as to why this result could not be concluded in the followed discussion, the most important being the lack of adaptive skin color detection, based on the face rectangle. Nevertheless the contribution in the original system must be argued to be quite small in proportion to the increased dimensionality of the feature vector. With regard to these facts the second recording of skin dependent features were chosen to be excluded from the final model.

The inclusion of texture measures and line segments should as well be seen as of an unobvious or far-fetched nature after face removal. Given that their stated purpose is to discriminate areas falsely detected as skin, such as bricks or wooden panels. As seen in the example images in 4.3, the zeroed face rectangles proved to produce false Hough line segments. The effect is of course reasonable as a black rectangle in a typical case can induce a significant intensity drop, identified as edges by the Canny edge detector.



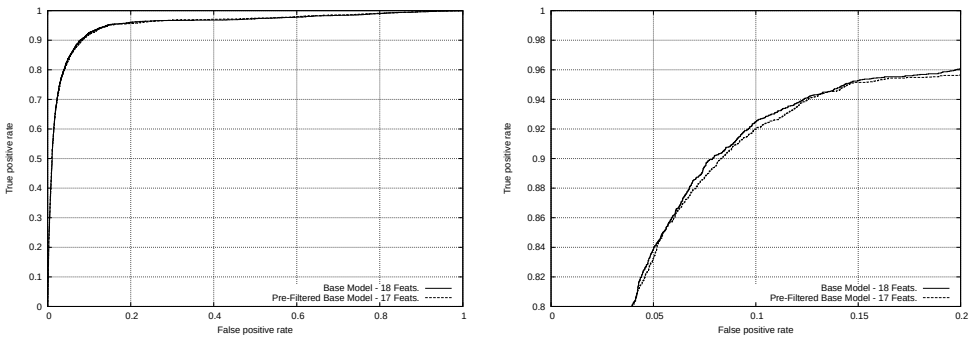
(a) (b) (c)

Figure 4.3: Examples of detected line segments, falsly induced by removed face rectangle. Detected lines are shown in white, pictures 4.3b and 4.3c have been cropped to conceal explicit areas.

4.2.3.2 Pre-Filtering Small Images

The feature recording a binary flag to be unset for images smaller than 10 pixels in any dimension, is only considered to increase the complexity of the model. As it has been argued that no images of such small dimensions can contain visible adult material, a proposition was made to filter them out all in all prior to feature extraction. Thus reducing the complexity of the feature vector by one more feature.

In filtering out all the small images, the non adult category proved to account for 755, while the adult category luckily contained none. To make fair future comparisons, a ROC curve of the now smaller set of extracted feature vectors was conducted. A comparison to the previous model is shown in 4.4. The amount of support vectors increased from 7733 to 8307.



(a) (b)

Figure 4.4: ROC-plot concluding changes in classification rates, due to exclusion of small sized images.

4.3 Further Improvement

In making a study of misclassified images, some of the priorly argued conceptions of possible weaknesses could be identified. For example both problems with clothed adult material and scantily dressed non adult material can be found 4.5a and 4.5b. Also some graphical content carrying areas with skin like colors have managed to fool the classifier 4.5c. The probably most common source of erroneously inferred adult pictures seems to involve poor lighting and or camera conditions 4.5d. These yield non or very low skin pixel probability. Moreover, pictures of smooth textured skin color are still a problem, such as animals, buildings, wood, rope and sand 4.5e. Finally the category of pictures with mixed content, depicting only small subareas of adult content proves very hard to correctly classify 4.5f.

Below a set of extending features are evaluated. Most have been chosen to address the tendencies found above and in regard to the reviewed research of the background chapter 3. Individual tests are made to conclude their discriminatory abilities. As concluded by the tests, a subset of features showed to perform well enough to be included in a final classifier. The performance of the final system is evaluated in chapter 6.

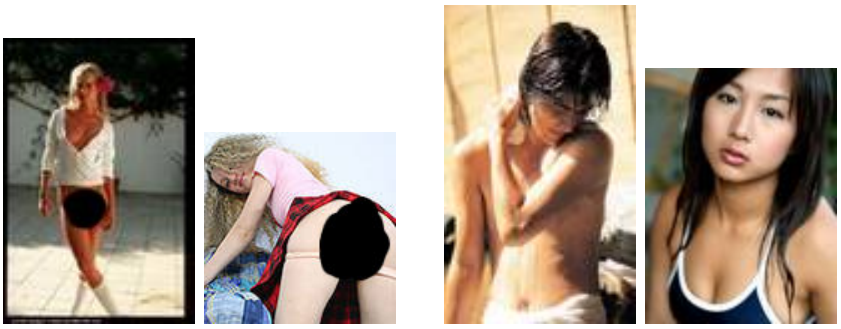
4.3.1 Further Artificial Discrimination

As earlier stated the topic of artificial discrimination is alone an interesting image retrieval research topic. In the work of (Garbarino, 2008) a total of eight different techniques of artificial discrimination are examined. Out of these the intensity entropy measure is proven to be the single best performing. Nevertheless it's also shown to be rather erroneous if used as a stand alone such discriminator. To try to further reduce the group of misclassified artificial images found, three additional features from (Garbarino, 2008) was evaluated. Two of these were found good enough to be included in the final classifier of this work. The fraction of individual pixel color values and the average pixel saturation value in a given image.

4.3.1.1 Number of Colors Used

One idea is to record the fraction of individual colors used. The reasoning is that the amount of colors used in photographic or natural images tend to fall into a fixed span. Making a separation of images using only a few colors, such as cartoons that typically have large one colored areas. Or in another extreme case unnaturally many colors could be used in some image effect.

It is obvious that this kind of measure is preferable on full scale uncompressed images. As stated in section 2.3; the JPEG compression algorithm typically introduce a significant amount of previously non existent color values. This fact is further stressed by the often heavily down scaled size, where previously several pixels have to be combined to one.



(a) Clothed Adult Content

(b) Scantily Dressed Content

(c) Graphical Content

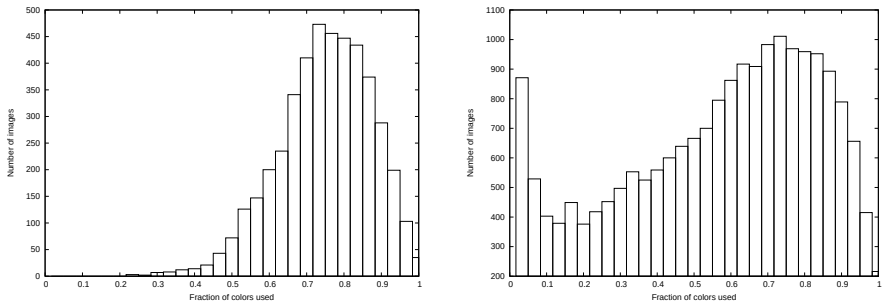
(d) Poor Photographic Conditions

(e) Skin Colored Surfaces

(f) Semi Graphical Content

Figure 4.5: Examples of different types of misclassified images. Adult images have been blackened out to conceal explicit areas.

A test run was made considering the full dataset of adult and non adult images. Individual separation of artificial content was not made as it was noted as common enough to show in results anyhow. As seen in figure 4.6 the non adult images prove to contain a significantly larger portion of images using a small fraction of individual colors. Especially a category of images using very few colors shows present. This category suggestively contains content such as banners and logotypes.



(a) Adult (b) Non Adult

Figure 4.6: Fraction of different color pixel values used. Recorded individually for each image.

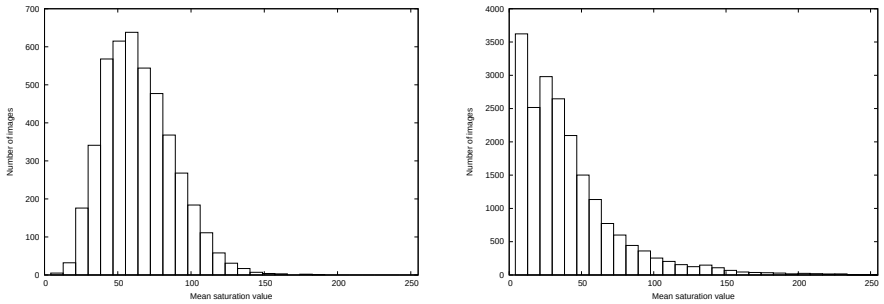
In conclusion. By measuring the fraction of individual pixel color values compared to the total amount of pixels. A contribution to the separation of artificial content can be made, even considering compressed images scaled down to typical thumbnail size.

4.3.1.2 Saturation

A highly saturated color can be thought of as pure. Saturation is the difference of a color and its own brightness. The proposed definition used in this work is $S = \max(R, G, B) - \min(R, G, B)$. where R, G and B is the color values of the red, green and blue channels respectively. The colors in photographic images are most commonly at quite low saturation levels. To put it in other words one could say that pure colors is quite sparse in nature. On the other hand the use of sharp colors must be regarded common for a variety of synthetic image categories such as cartoons, maps, charts, logotypes and other commercial graphics. Sharp colors will typically insist on attention of the human eye.

Two different saturation measures for artificial discrimination from the earlier stated work are here considered. Again tests were run using the full adult and non adult image data. In figure 4.7 the mean saturation of individual images is accumulated. Noted is a tendency of non adult images more frequently having lower mean saturation levels. This appearance was

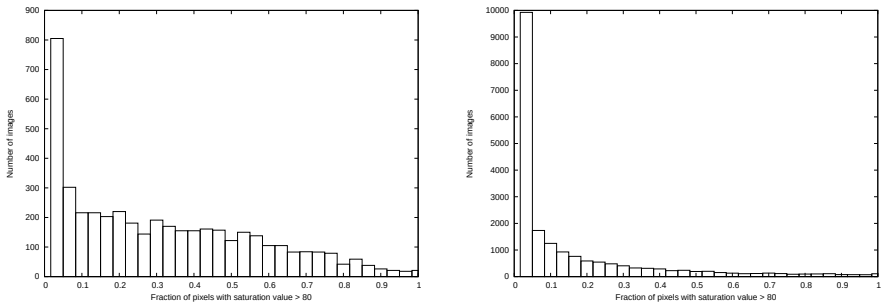
previously unexpected. In reasoning it's believed to be a combination of relatively high natural saturation level of skin areas and due to frequent use of large very low saturated near black and near white areas in the non adult image set. However to some extent the original theory of non adult images showing unnaturally high saturation levels can also be seen in the figure.



(a) Adult (b) Non Adult

Figure 4.7: Mean image saturation value.

In figure 4.8 the fraction of highly saturated pixels, with a saturation value greater than 80, is demonstrated in a manner corresponding to the previous. As hinted by the average saturation histograms, adult images prove to contain quite large areas of such highly saturated pixels.



(a) Adult (b) Non Adult

Figure 4.8: Fraction of image pixels with saturation value > 80.

The measure of highly saturated pixels was disregarded as a feature in the final classifier. For once because the results proposed in the original work could not be reproduced. Secondly because it's believed to be somewhat redundant to the measure of average saturation at such low threshold levels. Thirdly and in combination with the previous argument, no time could be prioritized to separate an independent test set of artificial images in order to search for a better threshold level.

4.3.2 Illumination Adaptive Skin Detection

As earlier noted pictures taken under poor photographic conditions have a tendency to be misclassified. Two methods were evaluated to help aid the classification of such images. Histogram equalization and clustering, prior to the training of skin color histograms. As presented below, none of the models proved to outperform the original Bayesian skin color histograms.

In conclusion it was however reasoned that the histogram equalization could come to terms with some evenly dark and light images. Due to this fact it was included separately. Thresholded at a rate of 15% false positives a rate of 90% true positives proved to be retrieved in a small test run, using the test set described in section 5.1.1.3. Using this threshold three features are recorded; the mean and standard deviation of the skin color probability as made during the original skin color detection and as well the fractional skin cover, measured as the amount of detected skin pixels through the total amount of pixels.

4.3.2.1 Histogram Equalization

Histogram equalization is a simple image processing method used to attain brightness normalization. For a given grayscale image a series of four steps are worked through. To consider colored images an extension can easily be made by transforming the image to a color space which keeps the brightness in one of its channels. For example in this work, the image was transformed into a YCrCb representation, the intensity channel Y equalized, and then retransformed into RGB.

1. The given image I is accumulated into histogram H, such that H records the amount of occurrences of all intensities in image I.
2. Histogram H is normalized into H_N, so that the sum of all bin values equal 255.
3. A new histogram H_{cdf} is created to record the cumulative distributive function of the normalized histogram. Explained in equation (4.3), where i and j denotes respective histogram bins.
4. Image I is finally transformed into its equalized version I_e, by using H_{cdf} as a look-up table for new intensity values. For every pixel; $I_e(x, y) = H_{cdf}(I(x, y))$, where (x, y) denotes pixel location.

$$H_{cdf}(i) = \sum_{i=0}^{j} H_N(j) \tag{4.3}$$

By histogram equalizing all images prior to training and inference, a new model of inferring skin pixel probabilities could be created. As shown

in figure 4.9 the overall classification rates of the pre-equalized skin detection was slightly outperformed by the original in terms of total area under curve. Although better performing at some levels of acceptance.

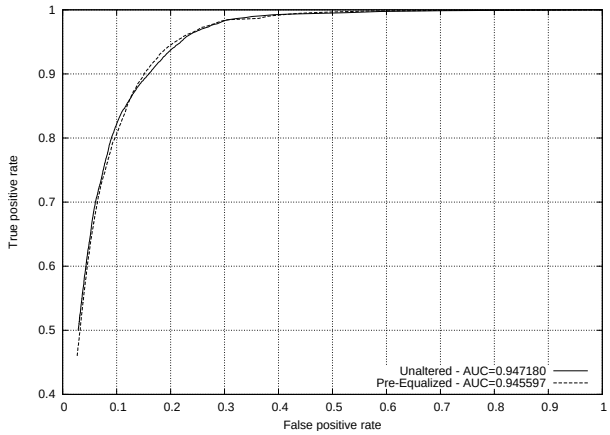
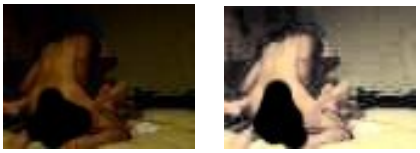


Figure 4.9: ROC plot comparing pre-equalized to ordinary, unaltered skin color detection.

The performance increase left unproven is most likely dependent to the fact that images with very dark or light backgrounds are heavily overcompensated in foreground areas. As the algorithm regards normalizing globally. For example portrait and model photography is commonly made against white or black backgrounds, such as example 4.10. The enhancement made due to generally poor lighting conditions is however believed significant, such as example 4.11.

(a) Before (b) After

Figure 4.10: Example of an image with bright colored background, before and after histogram equalization.



(a) Before (b) After

Figure 4.11: Example of a generally dark image, before and after histogram equalization.

4.3.2.2 Clustering

As stated in the background section 3.1.2.1, two of the previous systems examined the possibilities of including brightness data in the skin pixel classification. In (Zeng et al., 2004) images were divided (in an unspecified manner) into three groups based on their average brightness. Three different Bayesian skin color classifiers were then trained on respective dataset. Naturally introducing the average brightness as a factor in querying images for skin. In their results they showed and argued for a slight increase of accuracy. (Zheng et al., 2004) developed the concept further by including a measure of average chromaticity, to further separate different photographic conditions. They proposed using K-means clustering (with an unspecified amount of clusters) for pre-grouping skin training data. Their results as well argued to outperform the original model.

In order to enhance skin classification and try some of the concepts above, two pre-grouping models were tested. First a straight forward geometric model, dividing all the training examples in equally large sets based on their average brightness. The second model instead use K-means to cluster the sets. Furthermore a size of 32 bins per dimension were used for the Bayesian histogram models. Both because the smaller memory footprint it produces and as well accounting for less amount of training data for each model.

Test runs of the models were made on the skin dataset described in section 5.1.1.3. It was also reasoned due to early indicators, that the image background often had a too big impact on the brightness average. Therefor the best sized image center region for brightness average was searched as an additional parameter. The results of this first set of test runs are shown in table 4.2. As noted in the table, one setting produced a better AUC-value than the original model.

However as the test results was varying in a somewhat fluctuating manner it was believed that the test set could be too small. More specifically that the test results could vary by hitting qualitative thresholds of a single or a few images. To address this matter a new test set was produced by separating a set of 395 skin and 364 non skin full scale images from the

AUC	k-Means			Equal Split		
	3	4	6	3	4	6
0%	0.946255	0.942945	0.944704	0.944728	0.944836	0.944294
30%	0.946859	0.945817	0.947171	0.946528	0.947604	0.945385
40%	0.946636	0.945067	0.944507	0.945520	0.946968	0.942043
Original 0.947180						

Table 4.2: Area under ROC-curve comparison, clustered by average illumination. Left column shows percentage of image cropped from each side. Top row shows amount of clusters.

% False Positives	0.05	0.1
Original	0.742	0.893
4-Equal Split Crop 30%	0.741	0.881
6-Means Split Crop 30%	0.738	0.882

Table 4.3: Skin pixel classification recall, compared at two false positive acceptance rates. Using a large test set.

training set. The best performing models were again tested and compared against the original non clustered Bayesian skin histogram model. Due to the huge amount of pixel data a smaller comparison was made, computed at two relevant error rates. The results are shown in 4.3. As figured, the original model showed slightly more accurate using this set.

4.3.3 Enriching Skin Shape Analysis

As noted in the background chapter, the shape analysis of the proposed base system is rather thin. In the summary of (Rowley et al., 2006) "More elaborate shape representation of the shape of the skin color blobs" is proposed as a field of possible improvements.

Two strategies have been identified from other research to attain this task. Firstly and prior to the component labeling, a more thorough step of skin map enhancing algorithms may be run. Removing not only noise such as the current erosion and dilation step, but also filtering out rough textures, see section 3.1.2.3. Secondly a much more extensive flora of shape descriptors have been presented, including moment and other geometric shape measures, see 3.1.2.2.

4.3.3.1 Hu Moments

As stated earlier in the background chapter, the seven Hu-moments have been claimed to give a decent classification contribution. Inclusion have been proposed in two of the reviewed adult classification systems. It is also

known to be rather quick regarding computational efforts.

However two concerns have to be addressed regarding inclusion of the moments. Firstly a decision have to be made of what to measure. In the work of (Liang et al., 2004) the moments are calculated for the single largest blob. Regrettably the same details have been omitted from the work of (Zheng et al., 2004). However a weighted average approach could also be suggested.

Secondly and more importantly, possible issues with such a large dimensional increase of the feature vector have to be reasoned. Including the Hu moments in a straight forward sense into the base system would increase its complexity by 7 more features. One could also suggest including only a chosen subset of the moments. For example the seventh moment is known to disambiguate mirrored shapes (Hu, 1962, Sec. IV-C) while the first and the second has been shown to measure "the spread of the shape relative to its area square" and "the degree of elongation of a best fit ellipse on the shape" respectively (Zheng et al., 2004, Sec. 4.1).

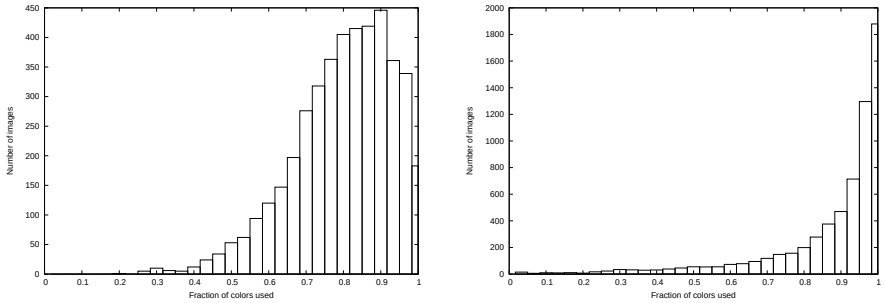
A decision was made to consider the Hu moments due to their proven results. However in order not to contribute too much to the complexity of the classifier only the first two moments are recorded. The argument being that only these measurements functionality is known, which is also known not to already be measured in the model. The two features are included in the shape analysis, being recorded for the single largest component.

4.3.4 Additional Skin Component Features

Two more simple features regarding the skin components were further proposed for inclusion.

Firstly; a simple method to further aid the identification of rough texture could be to count the amount of internal components. That is islands found not to be skin within the skin components. This measure is reasoned to help discriminate surfaces of unpure or noisy skin.

Secondly the previously presented feature from section 4.3.1.1 is proposed to be recorded inside the skin components. That is, recording the fractional amount of pixels found to have unique colors inside the skin components. It is specifically reasoned to further aid the discrimination of artificially made skin colored areas. In general and in similar with the skin color probability deviation measure, it is thought that a combination of lighting conditions and round shapes of the human body could to some extent produce a distinguished span of quite high fractions. Making separation of naturally flat skin colored surfaces possible as well. Figure 4.12



(a) Adult (b) Non Adult

Figure 4.12: Fraction of different color pixel values used inside skin components. Images with a detected skin cover less than 2% of total image size are disregarded.

4.3.5 Hough Line Length

Picking parameters for the Hough line transform isn't necessarily an easy choice. In practice the vast difference of adult image scenery and quality can sometimes yield identification of as well short as long straight line segments. To further aid separation of true skin areas, an additional measure of the total length of all found line segments is proposed to be recorded as a feature. The idea is for one that the total length measure is by itself an indicator of false skin areas. But also that the average line length can be retrieved through the geometric combination of the total length and the already recorded amount of line segments.

Chapter 5

Design & Implementation

During the implementation of the adult image classification system a struggle have been made to keep the amount of third party software dependencies low. As well as considering only reliable and well maintained tools. At an early stage in development planning OpenCV[1] (Bradski, 2000) and LIBSVM[2] (Chang & Lin, 2001) were considered to be good such software to build towards. It proved that these were the only needed dependencies upon completion. Both these libraries deliver mappings to the most commonly used programming languages. As both efficiency and modularity was seen as important for the system, C++[3] was chosen as the main programming language of implementation.

A goal during the implementation of the system has been to make it as modular as possible. Even so when in cases it would produce small extra computational costs. Due to the form of the task at hand of finding often quite independent discriminating image features, modularization has proven quite successful. Of course the benefits of modularity and separation of concerns are countable. But the easy inclusion and exclusion of features was aimed for especially, as it opens up better possibilities of making varying runs to measure individual feature contributions and computational time cost.

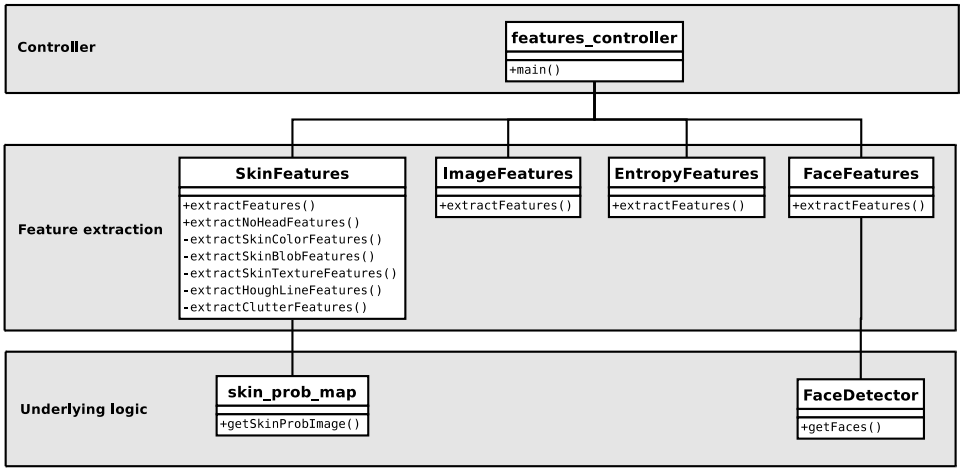
5.1 Base System

As depicted in figure 5.1 below, a quite natural hierarchy of the system logic could be achieved. At the top tier the features classifier serves as an interface for retrieving the resulting feature vectors for the SVM classifier. The middle tier presents logic for extraction of the different features. Here the different feature's extraction methods have been spread out to as many classes as

[1] An extensive library for computationally fast image analysis and processing.
[2] A highly considered support vector machine classifier library.
[3] An imperative, object oriented programming language based on the C-language.

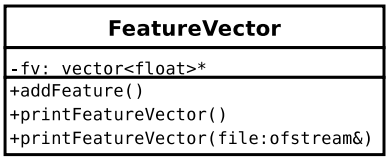
possible. However the bundle of skin dependent functionality proved hard to separate in a natural manner due to shared dependencies. As noted in the shown class operators in the figure they were instead abstracted into functions sharing the same instance data. The tier at the very bottom was used to separate logic needed for extracting more advanced features. In general, OpenCV proved to contain logic that solved a lot of the feature extractions in only a few lines. Why only a couple of features needed classes in this back-end tier.

Figure 5.1: Hierarchical overview of class structure in base system.



Not shown in the hierarchical figure for clarity reasons is the container class of the feature vector, see figure 5.2. It was implemented to produce an easy interface for adding new features from the feature extraction tier, as well as for the controller by presenting printing and file writing functionality.

Figure 5.2: Feature vector container class.



5.1.1 Skin Color Detection

The used Bayesian skin color histograms was to most extent hand implemented, as no reliable software package was found that delivered this functionality out of the box. OpenCV however, proved to contain a proficient histogram data structure that shipped with a rather useful set operations for the task. The implementation was split up into two major units of logic.

Bins	AUC
32^3	0.946090
64^3	0.947180
128^3	0.943340

Table 5.1: A comparison of classification rate at different histogram dimension sizes. The area under the produced ROC-curve, noted AUC, is used as the single measure.

One class for training new skin color histograms which purpose is to produce an instance of the other one. The second class contain a histogram of pre-calculated probabilities of color values being skin. Naturally, only the second one is used for classifying adult images.

5.1.1.1 Skin Histograms

The skin histograms internals are made up of two histogram data structures, one for recording the color values of skin pixels and one for the non-skin pixels. Except for training functionality that accumulate data into these histograms, some basic statistics and querying is provided. The one parameter needed to create a new instance of the class is the size in amount of bins of each dimension in the histograms. That is how many bins each color should be accumulated into. The parameter can be set to any evenly divisible number of the maximum size of 256 bins. Using the same test set as described in section 5.1.1.3, an optimum parameter value of 64 was experimentally arrived at, as shown in table 5.1. A result found interesting as it contradicts the dimension size of 32, proposed as optimum in the original work.

The class provides the possibility to create a skin probability histogram. This is made by firstly normalizing the skin histograms to even out differences in the amount of accumulated training data. Secondly the two histograms are merged into a single one of prior probabilities of pixels being skin, as shown in (5.1). Here σ_s and σ_n denote parameters to tune the span of possible inference rates. In practice a σ_n 10-20 times larger than σ_s tends to produce a proper span. This is likely due to the large concentrations of near black and white pixel color values that is contained within the non-adult training set (Jones & Rehg, 1999, p. 6).

$$H_{\text{prob}}(i) = \begin{cases} 0 & H_{\neg\text{skin}}(i) = 0 \\ 1 & H_{\neg\text{skin}}(i) \neq 0 \text{ and } H_{\text{skin}}(i) > H_{\neg\text{skin}}(i) \\ \frac{H_{\text{skin}}(i) \cdot \sigma_s}{H_{\neg\text{skin}}(i) \cdot \sigma_n} & H_{\neg\text{skin}}(i) \neq 0 \text{ and } H_{\text{skin}}(i) \leq H_{\neg\text{skin}}(i) \end{cases} \quad (5.1)$$

5.1.1.2 Skin Probability Histogram

The produced skin probability histogram is basically a lookup table for probabilities of pixel values being skin. Its main purpose of operation is to query pixel values as shown in (5.2). Where q_{RGB} is the value found in the histogram and Θ is a chosen threshold of acceptance. For convenience it provides possibilities to query full images and return either a binary pixel mask or a map of the queried skin probabilities.

$$r = \begin{cases} q_{\text{rgb}} & q_{\text{rgb}} \geq \Theta \\ 0 & q_{\text{rgb}} < \Theta \end{cases} \tag{5.2}$$

As the querying only reads from an indexed histogram structure it operates in constant time. Correspondingly for an image it operates in $O(n)$, where n denotes the number of pixels in the image. Also worth noting is that since the data of the histogram is never manipulated. Only one instance of the class is ever needed during the execution of the system.

5.1.1.3 Testing

A small test set containing 42 images of humans with labeled skin masks and 100 images not containing any human skin, was separated. All the images were taken from the web and scaled to at most $128x128$ pixel thumbnails. The skin set was biased to contain a reasonable spread of different human ethnicities and to some extent photographic conditions. The images in the non skin set was chosen randomly from the larger dataset of non adult images, introduced in section 2.3.

The results presented in the original work stated a true positive rate of 80% at a false positive rate of 8.5%. Using the set described above this result could almost be reproduced. As can be read from the curve in figure 5.3; the same true positive rate gives approximately 9% false positives. The difference is most likely due to the diversity biased skin set. However the down scaled image size could arguably also contribute to the fact.

For comparing reasons, three of the rule based skin detection techniques presented in the reviewed work are included in the figure. As seen, they are greatly outperformed by the proposed statistical approach.

5.1.2 Skin Component Labeling

To segment skin areas into connected components one of OpenCV's contour finding algorithms, based on the work of (Suzuki & Abe, 1985) was used. Instead of labeling full components, it only records their contours as chains of codes. The codes are single digits assigned between 0-7, stating in what direction the previous and next neighbor are located. The algorithm considers 8-connectivity, meaning that it accepts both straight and diagonally

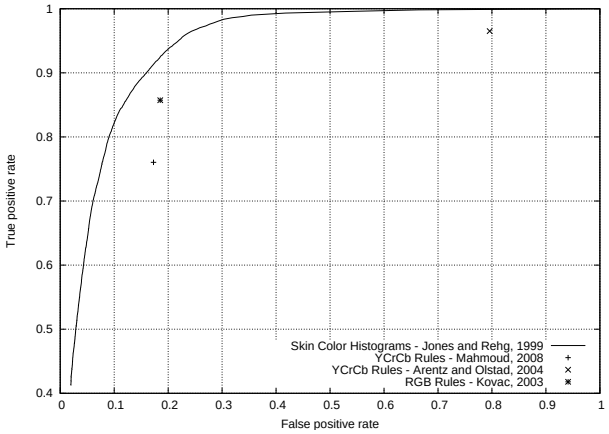


Figure 5.3: ROC plot comparing performance of different skin color detecting techniques.

attached pixels as neighboring pixels. The algorithm also claims functionality to record and divide inner contours, that is holes inside the components. Furthermore data structures and functionality for convenient manipulation of the contours is included in the library.

Regarding computational speed, the following is stated in OpenCV's reference manual (Intel, 2000). "All algorithms make a single pass through the image; there are, however, rare instances when some contours need to be scanned more than once. The algorithms do line-by-line scanning."

5.1.3 Texture Analysis

As noted earlier the texture analysis features were implemented using John Canny's edge detection algorithm. Using the produced binary edge map, straight line segments could be identified using a probabilistic Hough transform. Implementations of both these algorithms was found as a part of the OpenCV library.

5.1.4 Face Detection

An overview of the concepts behind the used face detection algorithm was given in the background chapter, section 3.1.2.5. The face detection algorithm is arguably considered one of the most sophisticated algorithms in the OpenCV library.

Four different cascades of weak classifiers, trained on the Haar-like features of frontal faces of humans was included in the distribution. Due to this pleasant fact a small test had to be conducted, to evaluate the best suited classifier. A test set of 100 images depicting one or more faces and 100 images not depicting any human face was separated from the final dataset in

an arbitrary manner.

In a first interactive test run, it was noted how many of the face images had their face regions correctly detected. Correspondingly a count of the amount of false positives on the non face images were recorded. As shown in figure 5.4, two of the cascades proved to outperform the other in terms of recall. Also noted in the figure is that the cascades were run on full scale images. The possibility to run the algorithm on half scale images was used for speed concerns in (Rowley et al., 2006). Making runs at both sizes on a patch of 1000 images proved that the full scale version was approximately 3 times slower. This extra computation time was however seen acceptable as the half scale versions greatly reduced the recall rate.

A second test made was conducted to state that the recorded amount of faces of a given image was correct. Using the face images from the dataset it could be concluded that the "default" cascade outperformed the "alt". With the first scoring 64% correct face amounts to 53% of the second, which tended to multiply locate the same faces. In conclusion; the "default" cascade was chosen for face detection, considering full scale images.

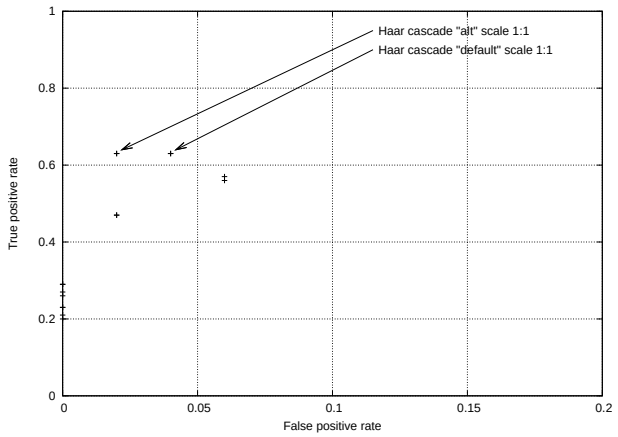


Figure 5.4: ROC plot of face region detection. Comparing performance of different cascades and parameters.

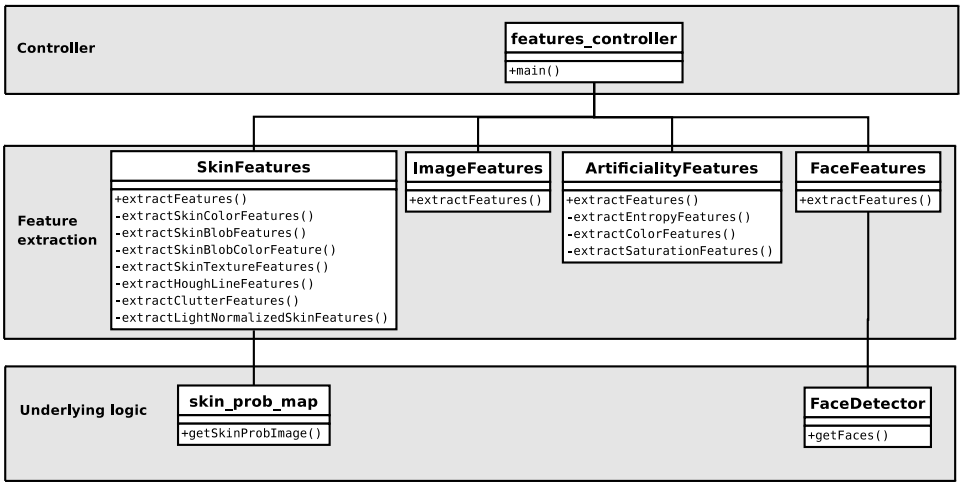
5.2 Final System

The final system contains a total of 27 features. Contributing by an additional 10 features to the reduced base system, as described in section 4.2.3. A class overview of the system can be seen in figure 5.5. A smaller number of differences are notable when compared to the class hierarchy of the base system. In the skin dependent logic the head removed functionality is no longer included. Furthermore two new function constructors are introduced for new measurements. Another significant change is also the extended

artificial discriminating logic in addition to the entropy features. Other additional functionality introduced in previous chapter, such as additional skin shape measurements, have been placed in already existent class methods.

The sections below presents any interesting implementary details of the extending functionality. Left out features are either thought explained in enough detail in previous chapter or not introducing any out of the ordinary in terms of technology.

Figure 5.5: Hierarchical overview of class structure in final system.



5.2.1 Counting Independent Colors

The fractional amount of individual color values are recorded as two of the additional features. Once globally for discrimination of artificial images and again inside the skin components.

In order to count the total amount of unique color values in an image an efficient data structure have to be used. For example, in considering an index table or histogram of all possible value combinations this would have to hold 256^3 storages or bins. However, a feasible representation is most likely too cumbersome in memory allocation and deallocation time and probably in terms of memory footprint as well. Instead an unordered hash set was reasoned a rational choice for the task and hence used. Although insertion can in the extreme case turn out to be linear in time complexity for this data structure, it's considered constant for the typical case.

5.2.2 Illumination Adaptive Skin Detection

Both the histogram equalization algorithm and the k-means clustering algorithm used was found as part of the OpenCV library. While the functionality of the former class proved small enough to be included in the existing skin

histogram classes, described in section 5.1.1. A new class had to be implemented to give an interface for the latter approach of clustering, as this classifier will typically train, use and hold several skin histogram instances at once.

5.2.3 Moments

Again, logic to make moment and specific Hu moment calculations was found in the OpenCV distribution.

Chapter 6

Experiments & Results

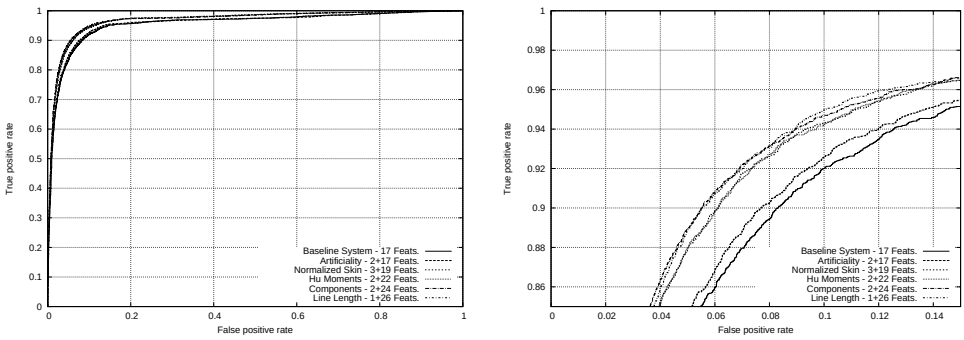
6.1 Training & Testing

As noted earlier the machine learning algorithm used by this work for final binary inference of adult picture content has been support vector machines (SVM). Of the few varieties of the SVM algorithm this work has considered using a radial basis function (RBF) kernel, as this has been proven successful in prior research. Without further stating technical details, an SVM-model with a RBF-kernel presents two parameters. Firstly the constant penalty parameter C and secondly the kernel specific parameter γ. In order to find reasonable parameter values, a grid search were conducted by training individual models at 11 values of C and 10 values of γ. The best classification accuracy of the final model was found to be at $C = 2.0$ and $\gamma = 2.0$.

Training and testing was conducted in the same manner as made in the evaluation of the base system, section 4.2.2. Using five fold cross validation on the full 24434 thumbnail image dataset, detailed in section 2.3. Again LIBSVM's weight parameter was set to four in favor of the adult images, to approximately even out the impact of the larger amount of non adult images.

6.2 Results

Figure 6.1 presents ROC-curves of the incremental contributions made by the addition of new features. The final base system as concluded in section 4.2.3 stands as the foundation of comparison. To recapture this is argued a rough replication of the system of (Rowley et al., 2006). However not considering a second recording of skin dependent features after face removal and as well pre-filtering tiny images instead of flagging them as a feature. Other statistics of possible interest regarding the models have been collected into table 6.1.



(a) (b)

Figure 6.1: ROC curves showing the contributions of extending features by addition for the final system. Figure 6.1b shows zoomed in version of same plot.

Features	Amount	Thr.put (img/s)	Support Vectors	Area Under ROC-Curve
Base System	17	64	8307	0.9553
Artificiality	2+17	49	10622	0.9573
Normalized Skin	3+19	48	5006	0.9679
Hu Moments	2+22	47	5007	0.9681
Components	2+24	47	5104	0.9687
Line Length	1+26	47	5122	0.9689

Table 6.1: Comparison of the impact of extending functionality to the final system. Sets of features are added on a per row basis. *Note that timing measures were made with images in cache, not considering cold disk read. In order to make a fair comparison, the median out of five runs was used in all time measures.*

As can be seen in the figure the illumination adaptive skin color detection proved to by far yield the most important contribution. An argument backed by the fact that it more than halves the amount of support vectors used in the trained model, indicating that it turns out to be a remarkably good descriptor for separation of adult content. An at least partial explanation of the success is reasonably ascribed to the measure of total skin cover. As this measure did not have any direct equivalence in the base system.

The two artificiality measures and the component features also has to be argued to make a considerable accuracy rate boost. As seen in the table, the former proves to add to some measurable CPU-time consumption during feature extraction. It is however believed that this can be reduced to some extent with better initialization of the used hash set data structure. Regarding the two extending skin component features and their very different rationales, it should be stated of interest to make individual tests of actual contribution to the shown accuracy boost. This is however left as an uncertainty by this work.

In accounting for the features shown to significantly improve the classifier, it can by exclusion be concluded that the the two Hu moments and the recording of total line segment length did not prove to be as important. Both are in review reasoned to address quite niched problem areas. The total line length could probably be interleaved, if a proper parameter search on an independent test set was to be conducted for the first Hough transform feature. The one counting the amount of line segments. The Hu moments on the other hand could be evaluated using a weighted average over the skin components, instead of the current single measure of the largest component. Irrespective of the small significance of the contributions, the features were decided to be kept in the final model. Both because they prove to give some small accuracy boost and in regard to their small computational impact.

6.3 Large Scale Test

In order to test the stated speed and scalability restrictions, a test run of the final system was made on a corpus of $4.7 \cdot 10^7$ images. The classification was run on a machine with an Intel Core 2 Quad processor. About 31 hours was needed to finish the run, equivalently each CPU-core proved a throughput of more than 100 images per second. A rate that confidently reaches the goal limit of several images per second.

6.4 Comparison with Related Research

A comparison of accuracy rates could be presented as such figures have been stated in all the reviewed systems. However it must be argued that not much is to gain of such a comparison of absolute rates, as the test

and training images shows significant variations in composition, source of retrieval, compression, dimensionality and different regards of the definition of adult. As argued in (Rowley et al., 2006, Sec. 5.3), using the web as the source of test content must be regarded as a very hard, real world case. The same work scores significantly lower absolute rates than the by this thesis very similarly composed system. This is most likely due to inclusion of cartooned and grayscale images in their tests. But it can as well be worth speculating in if they have been able to show for even smaller dependencies between training and test data.

A relative comparison of classification rates in regard to (Rowley et al., 2006) was nevertheless seen as the best mean of comparison possible. By making the rather bold supposition that the presented base system of this work would be equivalent to the "Face Detection" system outlined in the original work, such relative measures could be computed. An average accuracy increase of 1 percentage point for the full featured system, noted as "All features", must be considered a generous estimation in reading their ROC-plot. Which would equivalently account for a relative reduction of error rates of about 6%. The corresponding absolute accuracy increase by the extending features in this work accounts for 2.2%, which corresponds to an error rate reduction of about 32%. However important to note is that neither the comparison of absolute rates, nor the relative reduction of error rates should be considered preferable measures due to the varying natures of the test data. The graphical and grayscale content is for example considered to be unclassifiable in the original system, even though it is included.

Chapter 7

Conclusions

This thesis work has for one concluded the major parts of the large scale adult image classification system by (Rowley et al., 2006). From the results of the implemented system a series of extensions have been proposed and evaluated. It is presented and argued that at least three of these extensions do make measurable incrementation of classification accuracy. These featuers are considered new contributions to the research field and the system at large is reasoned to perform in line with the best adult classification systems presented in research. In a test run approximately 94.2% of the adult content could successfully be identified at a rate of 10% false positives. Also it is shown that the proposed system stays well below the specified image throughput limit and must hence be regarded purposeful for use at web scale.

7.1 Full Size Images

An interesting question to consider is the possible benefits that could be reached in considering classification of full scaled instead of thumbnail size images. A real world application for such a setting could for example be if the classification was to be made live during the downloading of new image content. In reviewing the functionality and discussions made in this and the previous work, some identified space for improvements is worth noting. Firstly the down scaling of images have previously been stated to reduce the image quality. In the work of (Garbarino, 2008), it is argued that this quality drop can yield less accurate measures when recording the fraction of unique colors in the image. Secondly the face detection is reasoned to perform more accurately in larger images, as the technique is more or less limited to detect faces larger than about $20x20$ pixels. Thirdly the thumbnail size was argued as one of the reasons not to investigate the universal feature detection methodology. A decision that should be rethought given the full images.

A final note should be stated about the loss of computational speed in considering full scale images. A review of downloaded web image content in (Garbarino, 2008) stated that the typical image size was found to be $(516 \pm 139)x(593 \pm 165)$ pixels. This would equally correspond to around an eighteen-fold size increase in amount of pixels. A fact that can't be seen as insignificant considering that most of the extracted features in the presented system works in linear time to the amount of image pixels.

7.2 Further Work

7.2.1 Adaptive Skin Color Detection

Two quite novel methods for adapting to illuminatory conditions of photographs have been evaluated, both shown to yield rather fair individual results. Nevertheless detection of skin areas continue to be the by far single most important source of greater classification rates. With the possibility of more accurate skin pixel detection not only could the sizable group of adult images misclassified due to unfound skin be corrected, but the accuracy of the shape measures would also increase.

In the survey of (Kakumanu, Makrogiannis, & Bourbakis, 2007) a series of 11 different approaches used to adapt to varying photographic conditions during skin detection is reviewed. Noted are for example two other simple color constancy algorithms, namely Gray World and White Patch. These could be evaluated in the same manner as the histogram equalization. With implementary time and complexity being less of a factor, the algorithm presented in (Rosenberg, Minka, & Ladsariya, 2003) could also be considered.

7.2.2 Addressing Graphical & Grayscale Adult Content

As decided at a starting stage of this project only colored photographic content was to be considered. Although a substantial amount of both graphical and black and white image content was identified in the dataset. As noted in the method chapter, section 2.2, both the separation of grayscale and to some extent graphical content can be made, the latter shown in (Garbarino, 2008). In making such separations prior to feature extraction, the system can be extended to include a single classifier for each of these adult categories. In addressing the individual tasks the adult classifier could be relieved by some of its complexity for separation of synthetic content.

However the identification of adult images grayscale or graphical images is very much a non-trivial topic and a less researched such. As discussed in the background chapter, the bag-of-words methodology uses universal feature detection methods. Some of these methods do only consider grayscale settings. For example the work of (Lienhart & Hauke, 2009) state a true positive classification rate of 90.8% achieved at an impressive rate of 2% false

positives considering grayscale images only. Even considering the discussion presented in section 3.1.1.2 of this universal feature detection methodology, it could show purposeful to examine for at least this type of images.

Given a thorough inspection of the graphical or synthetic content, some visible and measurable tendencies would reasonably be revealed for coarser filters. Nevertheless given the vastly greater possibilities of image color, form, and content of this material it's most probably very hard to filter out at higher rates. Again, the possibility of applying a universal feature detection method could show to aid the problem.

7.2.3 Explicit Area Detection

Another possibility explored in the introductory motivation section 1.1, would be to build detectors for explicit areas of the human body. This could typically be done using OpenCV's framework for rapid object by a cascade of boosted classifiers, as used for face detection in this system. In detecting these areas, improved classification rates of the adult system was believed to be reachable and under conditions invariant to skin color.

However the creation of an acceptable such classifier would at least turn out to be a time consuming task. An overview of the complexity of training Haar-cascades can be given by reviewing (Seo, 2006). Here a step by step replication of the frontal face detector is presented. Noticeable is that the samples typically need to be the same size, have close cropping and illumination conditions, and furthermore amount to a couple of thousand samples. Even overcoming the labeling of such a set, the training include some parameter tuning and a training run takes a not inconsiderable amount of time. Finally the consideration of explicit areas would to the knowledge of this thesis be breaking new ice for the object detection technique. Possibilities of tackling uncertainties such as not enough shared patterns between samples should not be left unaccounted for.

To conclude; it is argued for the wiser to examine other methods for adult image discrimination such as universal feature detection, before trying to train Haar-cascade classifiers.

Bibliography

Arentz, W. A., Olstad, B. (2004). Classifying offensive sites based on image content. *Computer Vision and Image Understanding*.

Bradski, G. (2000). Programmer's toolchest: The opencv library. *Doctor Dobb's Journal*. (Software available at `http://opencv.willowgarage.com/`)

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

Chang, C.-C., Lin, C.-J. (2001). *LIBSVM: a library for support vector machines.* (Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`)

Cula, O. G., Dana, K. J. (2001). Compact representation of bidirectional texture functions. In *Ieee computer society conference on computer vision and pattern recognition.*

Deselaers, T., Pimenidis, L., Ney, H. (2008). Bag-of-visual-words models for adult image classification and filtering. In *18th international conference on pattern recognition.*

Fei-Fei, L., Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Ieee computer society conference on computer vision and pattern recognition.*

Fleck, M., Forsyth, D., Bergler, C. (1996). Finding naked people. In *European conference on computer vision.*

Garbarino, M. (2008). *Automatic classification of natural and synthetic images*. Unpublished master's thesis, Royal Institute of Technology, School of Computer Science and Communication.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *IRE transactions on Information Theory.*

Intel. (2000). *Open source computer vision library: Reference manual.*

Jones, M. J., Rehg, J. M. (1999). *Statistical color models with application to skin detection* (Tech. Rep.). Cambridge Research Laboratory.

Kakumanu, P., Makrogiannis, S., Bourbakis, N. (2007). A survey of skin-color modeling and detection methods. *Pattern Recognition.*

Kiryati, N., Eldar, Y., Bruckstein, A. M. (1991). A probabilistic hough transform. *Pattern Recognition.*

Kovač, J., Peer, P., Solina, F. (2003). Human skin colour clustering for face

detection. In *Eurocon - the international conference on "computer as a tool".*

Lee, J.-S., Kuo, Y.-M., Chung, P.-C. (2006). The adult image identification based on online sampling. In *International joint conference on neural networks.*

Liang, K. M., Scott, S. D., M.Waqas. (2004). Detecting pornographic images. In *Asian conference on computer vision.*

Lienhart, R., Hauke, R. (2009). Filtering adult image content with topic models. In *Ieee international conference on multimedia and expo.*

Lienhart, R., Maydt, J. (2002). An extended set of haar-like features for rapid object detection. In *Ieee international conference on image processing.*

Lienhart, R., Slaney, M. (2007). plsa on large scale image databases. In *Ieee international conference on acoustics, speech and signal processing.*

Lowe, D. G. (1999). Object recognition from local scale-invariant features. In *International conference on computer vision.*

Lowe, D. G. (2004). Distinctive image features from scale-invariant key points. *International Journal of Computer Vision.*

Mahmoud, T. M. (2008). A new fast skin color detection technique. *World Academy of Science, Engineering and Technology.*

Rosenberg, C., Minka, T., Ladsariya, A. (2003). Bayesian color constancy with non-gaussian models. In *Neural information processing systems.*

Rowley, H. A., Jing, Y., Baluja, S. (2006). Large scale image-based adult-content filtering. In *International conference on computer vision theory and applications.*

Seo, N. (2006). *Tutorial: Opencv haartraining (rapid object detection with a cascade of boosted classifiers based on haar-like features).* (Article available at `http://note.sonots.com/SciSoftware/haartraining.html`)

Sivic, J., Russell, B. C., Efros, A. A., Zisserman, A., Freeman, W. T. (2005). Discovering objects and their location in images. In *Ieee international conference on computer vision.*

Suzuki, S., Abe, K. (1985). Topological structural analysis of digitized binary images by border following. *Computer Vision, Graphics, and Image Processing.*

Viola, P., Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Ieee computer society conference on computer vision and pattern recognition.*

Zeng, W., Gao, W., Zhang, T., Liu, Y. (2004). Image guarder: An intelligent detector for adult images. In *Asian conference on computer vision.*

Zheng, Q.-F., Zeng, W., Wang, W.-Q. (2004). Shape-based adult images detection. In *International conference on image and graphics.*