Multiway Approach to Content Recognition on Internet

Ing. Miroslav Vacura
vacura@vse.cz
Department of Information and Knowledge Engineering
Faculty of Computer Science and Statistics, VSE Praha
nm. Winstona Churchilla 4, 130 67 Praha 3, Czech Republic

December 30, 2001

Abstract

Every application that works with multimedia data on Internet - web pages with text, images, sound etc. must deal with problem of large amount of unusable information. One of most unuseful information types that is wide spread on Internet in pornography. This paper presents effective method for filtering out pornographic web pages. Presented method uses multiple ways of analyzing web pages: analyzes URL address of web page, analyzes images included on the page, analyzes HTML code of page to find out type of structure of site that is given web page part of and few other characteristics. These methods can be easily used also for recognition of another content types of multimedia documents on Internet.

1 Introduction

Every application that works with multimedia data on Internet - web pages with text, images, sound etc. must deal with problem of large amount of unusable information. One of most unseful (for most purposes) information types that is wide spread on Internet in pornography. This paper presents effective method for filtering out pornographic web pages on Internet using multiple ways of analysis.

Presented approach uses number of methods for analyzing web pages: analyzes URL address of web page, analyzes images included on the page, analyzes HTML code of page to find out type of structure of site that is given web page part of and few other characteristics. These methods can be easily used also for recognition of another content types of multimedia documents on Internet.

2 Image analysis

Purpose of this work was find image analysis method that would be not only very effective in recognition images with pornographic content but also very fast, due to need to process large amounts of Internet data. Fulfilling both these necessary attributes - effectiveness and speed, color based analysis seems most appropriate.

2.1 Choosing color space

First step is choosing right color space that will be used to analyze images. There is lot of possible color spaces, such as RGB, YIQ, YUV, OPP and others. I have chosen HSV color space because it has some very useful properties:

- Uniformity - metrical distance of codes of two different colors is in relation to their similarity as they are perceived by human.
- Completeness - color space contains all perceivable colors.
- Compactness - all colors in color space are different.
- Naturality - color space colors are naturally represented by 3 perceivable properties hue, saturation and lightness.

Relation between best known color space RGB (representing every color by amounts of red, green and blue) and color space HSV (representing every color by hue and amount of lightness and saturation) shows figure 1.

Important property is mainly uniformity because it assures us that if we choose interval in HSV color space, the colors in this interval will be similar. Therefore we can easily quantize HSV color space to discrete sections containing similar colors.

I have chosen standard HSV quantization:

- Lightness - quantized to 3 values
- Saturation - quantized to 3 values
- Hue - quantized to 18 values

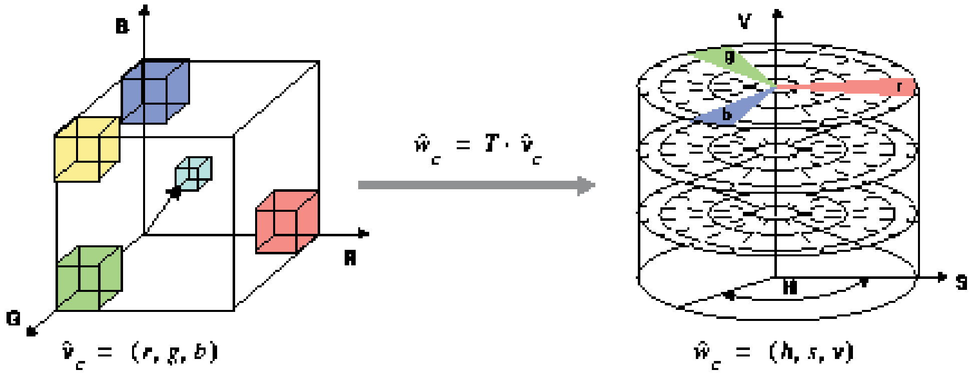

Figure 1: RGB and HSV [1]

Result is $3*3*18 = 162$ different color groups each containing similar colors. Standard quantization also adds 4 types of grey to total 166 color groups, but grey color are not important for this application, so let me ignore it.

2.2 Creating histogram

Every color group has got index number from 0 to 161. Now it is possible to create histogram for every analyzed image. For every color group is necessary to count number of pixels on given image that have color contained in that color group. Result is histogram - ordered vector of 162 numbers, that describes analyzed image.

Next step is histogram normalization. Because images have different size, therefore different total number of pixels, it is necessary to normalize histogram. Let c_i be number of pixels having color from color group i on given image. Histograms were then normalized to hold following:

$$\sum_{i=1}^{162} c_i = 10000 \tag{1}$$

Figure 4 shows histograms of figure 2 and figure 3

2.3 Histogram analysis using k-nearest neighbor method

Resulting histograms can be processed by various methods depending on what kind of content recognition we need. For task of recognition of pornography k-nearest neighbor method gives very good results. We need training set of histograms of pictures that are known to system (pictures are described as pornographic or non-pornographic). When histogram of picture that has to be analyzed is presented, we find k nearest histograms according to Euclidean distance between histograms. Now there are two possible ways how to find category of given unknown histogram.

Simple way is to say that unknown picture is of the same category as *most* of k pictures (f.e: if $k = 7$ and 3 of nearest histograms are pornographic and 4 are non-pornographic then we say that analyzed picture is non-pornographic). Formally:

Figure 2: Example of picture

$$\hat{f}(x) \leftarrow argmax_{(v \in V)} \sum_{i=1}^{k} \delta(v, f(x_i)) \tag{2}$$

Here V is set of categories, for this application $\{0, 1\}$, x is histogram vector, $x_1 \dots x_k$ are nearest k histogram vectors to vector x, function f gives category for histogram from training set (f.e. $f(x_3) = 1$ if histogram x_3 is pornographic). Function δ is defined so:

$$\delta(x, y) = \begin{cases} 1 & \text{if} \quad x = y \\ 0 & \quad else \end{cases} \tag{3}$$

More complicated variation of this method is different: every histogram from group of k nearest histograms has weight and this weight is higher if that histogram is closer to analyzed histogram.

$$\hat{f}(x) \leftarrow argmax_{(v \in V)} \sum_{i=1}^{k} w_i \delta(v, f(x_i)) \tag{4}$$

Where weight w_i of histogram vector x_i is calculated as:

$$w_i = \frac{1}{d(x, x_i)^2} \tag{5}$$

Where d is chosen metric for distance between vectors.

This method of analysis gives reasonable results. Further testing has shown that it is possible get even better results by ignoring black color. Final results of testing this method has shown following effectiveness:

Figure 3: Example of picture

k	% Errors
9	16,5
5	16,5
1	15,5

Surprisingly best results are achieved for $k = 1$. This shows that even most simple variant of k-nearest neighbor method can be effectively used giving reasonable results without need for enormous computing capacity, so very fast.

2.4 Histogram analysis using rules

Another way how to analyze given unknown histogram is to use of rules. Various rules were tested, all created manually (automatic generating of rules and their evaluation is planned in future). Best results were obtained even with very simple rules like:

IF Picture.Histogram(127)> 242 THEN Picture.Cathegory = 1

This rule simply says that if analyzed histogram vector has number higher than 242 on position 127 (index number for some color group) then histogram is pornographic.

Testing shown that such simple rule has error rate in pornography recognition 13.5%.

3 Analysis of URL

URL (Uniform Resource Locator) is defined in standard RFC1738 [2]. For purpose of this paper only important protocol (schema) defined by this standard is HTTP (Hypertext Transfer Protocol), that has form:

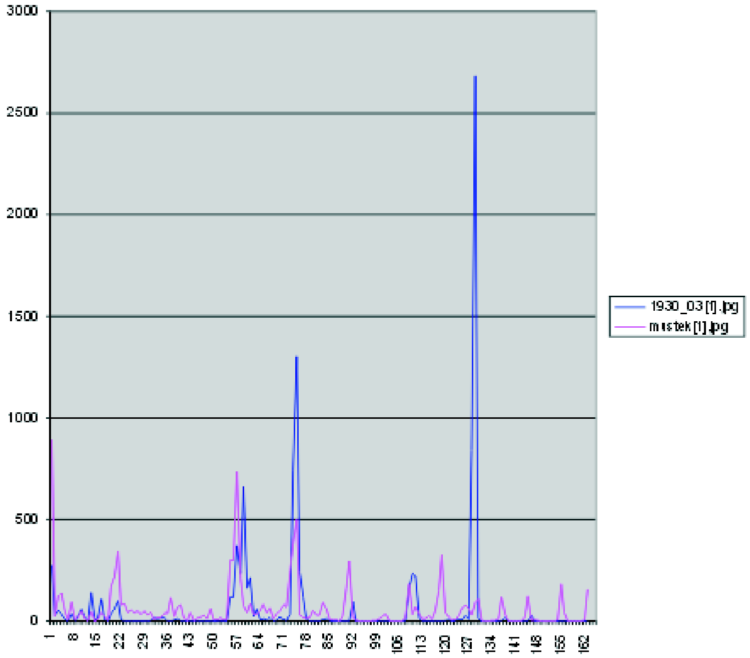

Figure 4: Histogram of previous two pictures

```
http://<host>:<port>/<path>?<searchpath>
```

Aim of this method is identify URLs that contain pornographic images. Examination of pornographic web site has shown that most of them have same special words or phrases in their URLs. Basic principle of this method is to create dictionary of significant text strings, then search every given URL, and finally based on number of found significant string evaluate given URL.

Dictionary of significant text strings looks like following example:

string	*host*	*path*	*searchpath*	*multiplier*
sex	1	0.5	0.4	2
xxx	0.8	0.5	0.4	2
...	...	...	...	...

This table contains list of significant strings. For every string there is its significance for case when it is present in *host* part of URL, *path* part of URL, and *searchpath* of URL. Some strings are not very significant when they are present alone in given URL, but their significance raises when they are found together with other strings. Therefore value *multiplier* is defined for each string, this helps to include this possibility of presence of multiple significant strings in reasoning.

Mechanism for evaluating URLs works formally as follows: Lets say that *URL* is formally set of all substrings that can be derived from given URL.

Wordlist is set of all significant strings in dictionary. Then we define set of relevant strings so:

$$Rel = URL \cap Wordlist \tag{6}$$

We can formally define function $host_{tab}(str)$ that for given significant string returns value of *host* column in dictionary. Now we can define function $host(str)$ that includes possibility that other significant strings were also found in URL.

$$host(str) = host_{tab}(str) * Cooc(str) \tag{7}$$

Function $Cooc(str)$ is based on total number of significant strings in given URL:

$$Cooc(str) = \begin{cases} 1 & \text{if} \quad |Rel| = 1 \\ |Rel| * multiplier(str) & \text{if} \quad |Rel| > 1 \end{cases} \tag{8}$$

Functions $path_{tab}(str)$, $path(str)$, $searchpath_{tab}(str)$, $searchpath(str)$ are defined analogously to functions $host_{tab}(str)$ and $host(str)$.

Then it is possible define auxiliary function $sigf(url, str)$ as follows:

$$sigf(url, str) = \begin{cases} host(str) & \text{if str is contained in host part of url} \\ path(str) & \text{if str is contained in path part of url} \\ searchpath(str) & \text{if str is contained in searchpath part of url} \end{cases} \tag{9}$$

Finally function $signf(url)$ that evaluates total significance for given URL can be defined as:

$$signf(url) = \sum_{str \in Rel} sigf(str, url) \tag{10}$$

3.1 Results of method

For testing of described method simple rule of following form was prepared:

IF Significancy(Picture) > X THEN Positivity(Picture)=1

Testing set and dictionary had following parameters:

- 51 significant strings in dictionary
- 1284 URLs in testing set
- 1021 negative (of nonpornogrphic web page)
- 263 positive (of pornographic web page)

Application of method has given following results in absolute values:

- 63 errors totally
- 11 negative URL were classified as positive
- 52 positive URL were classified as negative

What means in relative values:

- 4,9% errors totally
- 0,8% of all URL were erroneously classified as positive
- 1% of negative URL were erroneously classified as positive
- 4% of all URL were erroneously classified as negative
- 19,7% of positive URL were erroneously classified as negative

By tuning described rule we can lower rate of positive URL erroneously analyzed as negative for cost of raising rate of negative URL erroneously analyzed as positive.

4 HTML analysis

There are two parts of HTML analysis that is presented in this paper.

- Web site structure analysis
- Web page quantitative analysis

Main purpose of first approach of HTML analysis is to identify structure of some web site. If we have some web page, this page is always part of some web site, some wider structure of web pages. If we identify structure of that web site we can also say something about web page that is part of this web site.

Web page quantitative analysis is based on idea, that some classes of web pages can typically contain only little amount of text or there can be made also other assumptions based on relation between quantity of text and quantity of other multimedia data on given web page (images, video, sound).

4.1 Web site structure analysis

If we analyze structure of web pages we need to focus to *anchor* mark contained in HTML document. These hypertext links can point to documents outside current web site:

```
<a href="http://www.seznam.cz">Seznam</a>
```

Or there can be *anchor* marks that point to documents that are in current web site:

```
<A HREF="img/barn.jpg">Picture</A>
```

Most important for web site structure analysis is the second type of marks.

Empirically I found that most pornographic web sites has "gallery" structure. Main page is *selection* page containing thumbnails of images. These thumbnails are also links to documents with picture in real size, so user can just click on thumbnail to open real size picture. Such link has often HTML code like this:

```
<A HREF="barn.jpg"><IMG SRC="sm_barn.jpg" WIDTH="68" HEIGHT="50"
 BORDER="0"><BR><SMALL>barn.jpg</SMALL></A>
```

Typical selection page contains 10-30 links that have such code. These links are all similar, they differ only in name of thumbnail and name of target real size document. Name of target document almost always differs only numerically (f.e. img01.jpb . . . img30.jpg). Target real size document is often .JPG image document. Sometimes is target document HTML web page that contains only real size picture and nothing else. Sometimes selection page and even target document HTML web page contain also advertising. Typical selection web page containing thumbnails and advertising is on figure 5.

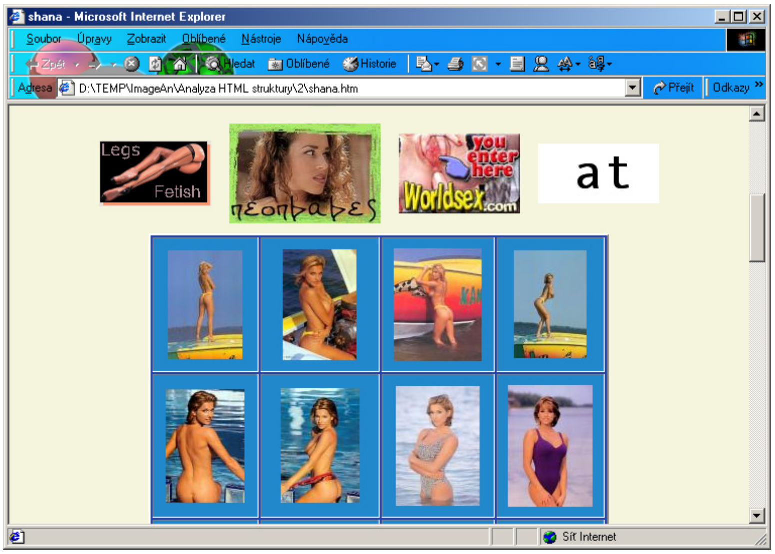

Figure 5: Typical selection page

This page contains at beginning advertising in form of hyperlink pictures:

```
<TABLE cellPadding=8>
    <TBODY>
    <TR>
      <TD><A href="http://www.legs-fetish.com/"><IMG
        src="shana_soubory/0002.gif" border=0></A></TD>
      <TD><A href="http://equis.ya.com/neonbabes/neon/"><IMG
        alt="Neonbabes: daily HQ babe galleries!" src="shana_soubory/nb03.jpg"
        border=0></A></TD>
      <TD><A href="http://www.worldsex.com/"><IMG
        src="shana_soubory/worldsex.jpg" border=0></A></TD>
      <TD><A href="http://www.bloatedgoat.com/"><IMG
        src="shana_soubory/trigger1.gif" border=0></A></TD>
     </TR>
    </TBODY>
</TABLE>
```

Then there are thumbnails of pictures that point to real size pictures:

```
<TABLE cellPadding=10 bgColor=#2244aa border=2>
   <TBODY>
   <TR bgColor=#2288cc>
     <TD>
       <CENTER><A
       href="http://equis.ya.com/quicksand/gallery3/shana/shana01.jpg"><IMG
       src="shana_soubory/S_shana01.jpg" border=0></A></CENTER></TD>
     <TD>
       <CENTER><A
       href="http://equis.ya.com/quicksand/gallery3/shana/shana02.jpg"><IMG
       src="shana_soubory/S_shana02.jpg" border=0></A></CENTER></TD>
     <TD>
       <CENTER><A
       href="http://equis.ya.com/quicksand/gallery3/shana/shana03.jpg"><IMG
       src="shana_soubory/S_shana03.jpg" border=0></A></CENTER></TD>
     <TD>
       <CENTER><A
       href="http://equis.ya.com/quicksand/gallery3/shana/shana04.jpg"><IMG
       src="shana_soubory/S_shana04.jpg" border=0></A></CENTER></TD></TR>
   <TR bgColor=#2288cc>
     <TD>
       <CENTER><A
       href="http://equis.ya.com/quicksand/gallery3/shana/shana05.jpg"><IMG
       src="shana_soubory/S_shana05.jpg" border=0></A></CENTER></TD>
...
```

Formally we can say that all hyperlink *anchor* marks on web page (identified by some URL) create set A_{url}. In this set there is important subset of links that are represented by image (thumbnail) on original page - formally $Aimg_{url}$. If there is at least 5 such links on given web page ($|Aimg_{url}| \geq 5$), this page can be selection (gallery) page.

Set $Aimg_{url}$ contains another important subsets - $Aimg^1_{url} \dots Aimg^n_{url}$. These subsets contain links that have same *host* and *pathname* part, differ only in *filename*, formally:

$$\begin{aligned} \forall url_1, url_2 \in Aimg^x_{url} : host(url_1) = host(url_2) \quad & c = 1 \dots n \\ \forall url_1, url_2 \in Aimg^x_{url} : pathname(url_1) = pathname(url_2) \quad & c = 1 \dots n \end{aligned} \tag{11}$$

If there is any subset $Aimg^x_{url}$ such as $|Aimg^n_{url}| \geq 5$ probability that given web page is gallery page is even higher. If target documents are JPG pictures we can be almost sure.

4.2 Quantitative analysis of web page

HTML code of web page contains apart of anchor hyperlink marks also other marks and text. Typical pornographic gallery selection page has minimum amount of text. Nonpornographic gallery pages have often text commentary describing content of every picture.

Formally can be defined set $Words_{url}$ of all words in text part of given web page. Now it is possible to count index based on relation between amount of text and amount of links to images on given web page:

$$P = min\left(\frac{2|Aimg_{url}|}{|Words_{url}|}, 1\right) \tag{12}$$

For cases when $Words_{url} = 0$ is necessary to define P as:

$$\begin{aligned} |Words_{url}| = 0 \wedge |Aimg_{url}| = 0 \Rightarrow P = 0 \\ |Words_{url}| = 0 \wedge |Aimg_{url}| > 0 \Rightarrow P = 1 \end{aligned} \tag{13}$$

P gets value between 0 and 1. For web pages where there is 2 or less words to 1 link to image is P=1. For web pages with more text is P lowering.

If target document of gallery web page is not JPG image file, but HTML document it is possible to analyze this document similarly. Empirical examination shows that most pornographic target web pages, that have HTML form contain no text at all. They contain only real size picture and sometimes also advertising picture with hyperlink to another site. Easy rule can be created that if target picture contains no text at all then likelihood that it is pornographic image is significantly higher.

5 Conclusion

This paper presents effective multiway method for filtering out pornographic web pages on Internet. Such filtering is often one of first steps when working with multimedia data on Internet Testing has shown that combination of presented methods can give results with less than 5% errors. Nature of methods also gives possibility to tune them easily based on error type that is preferred for any given application.

References

[1] Smith,J.R., Integrated Spatial and Feature Image Systems: Retrieval, Analysis and Compression. Graduate School of Arts and Sciences, Columbia University, 1997

[2] Berners-Lee, T., Masinter L., McCahill, M. - RFC1738 - Uniform Resource Locators (URL), Internet http://www.internic.com