An Investigation into Content Based Video processing in Cloud Computing Paradigm

Rashed Mustafa $^{1,2,3,4}$ and Dingju Zhu $^{1,2,4,5,*}$

1 Laboratory for Smart Computing and Information Science, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China
2 University of Chinese Academy of Sciences, Beijing, China
3 Department of Computer Science and Engineering, University of Chittagong, Bangladesh
4 Shenzhen Public Platform for Triple-play Video Transcoding Center, Shenzhen, China
5 School of Computer Science, South China normal University, Guangzhou, China

Abstract - Finding Obscenity from videos is a crucial issue due to social and ethical reasons of using online resources. Dating back to two decades the research on this ground has been started. Most of the works are based on image based skin color detection which is not a suitable method for finding obscenity from videos. The reason for this is that, there are many skin like pixels such as beach photos, human skin like animal, skin colored painting that enables false positive (FPR) and false negative rate (FNR). Moreover frame checking time is also a performance factor. In addition all works performed well on some particular set of data. In this paper some aspects of finding obscenity from videos is described delineating strength, weakness and possible extensions of prior works. Introducing some new features and incorporation of multiple classifiers and transfer learning will lead the work more robust. In addition traditional multimedia cloud computing has been investigated in this paper following some extensions.

Keywords: Content Based Video Processing (CBVP), Transfer Learning (TL), Multimedia Cloud Computing (MCC)

1 Introduction

Due to huge development on Information and Communication Technology, there is an enormous number of online resource monitoring cells all over the world. In spite of those security systems, it doesn’t check content based video appropriately. This is a threat for the Internet users while using computers in office or in front of family members or children. Moreover malicious content contradicts social and ethical issues. Hence content based video processing especially for identifying obscenity has now been a challenging research area. It has been almost two decades when Forsyth [1] published the first paper in this issue on “Finding Naked People”. After that a large numbers of works have been accomplished by different scholars all over the world. Text based protection system has been used in early 2000 for screening malicious contents from the Internet [2, 3]. This system is not working for those sites having malicious contents but non objectionable site name. In support of this Drawback, we can find millions of sites that contain objectionable videos.

The remaining parts of this article can be categorized according to the following ways: section 2 briefly describes literature review; different skin detection methods summarized in section 3, in section 4 some open research issues are proposed and finally a comprehensive discussion is presented in section 5.

2 Literature Review

There are extensive literatures on obscenity detection from videos but in this paper some significant works has taken into consideration. Those papers mainly focused on large connected skin regions, erotogenic organs, and feature descriptors following some different classifiers.

The following figure demonstrates number of significant works published since 1996 to till now. It has been observed that in 2010 most of the papers published due to significant improvement of machine learning tools.
According to the method of Qian W et al (2005 [9]; extracting video frames, motion vectors were calculated and employed to segment video frames to frames with global or local motions. For local motion, skin regions were extracted and edge moments were utilized to classify each key frame as an objectionable or a benign frame. The method suffers from using the spatial information of only key frames. The method also needs a general database for matching of moment.

In [10], key technologies for obscene video recognition in the compressed video domain were illustrated. The algorithm extracted skin regions and detected key frames in the compressed domain. In this approach key frames were extracted applying a threshold to the number of skin pixels. Finally a decision tree was utilized to classify key frames.

Lee H et al (2006) [11] proposed two models of features for objectionable video classification. The first model utilized features based on single frame information and the second feature model was based on the group of frames. The features of the two models were classified using two SVM classifiers. The authors extended their work [12] to a multilevel hierarchical system, which utilized very similar features for detecting objectionable videos. The method included three phases, which were executed sequentially. In the first phase, initial detection was performed based on hash signatures prior to a video being downloaded or played. In the second phase, single frame based features were utilized for the detection followed by a third phase where the detection was completed by features based on the group of frames reflecting the overall characteristics of the video. Both algorithms sampled frames periodically to avoid the computational overhead for finding the key frames of a video. This method will not work for the classification of video episodes with different categories in the same video file.

There is another method where motion vectors were extracted from MPEG video stream and smoothed using a median filtering [13]. Then the periodicity of motion vectors was extracted and, based on the motion features, obscene videos were detected. The method uses only motion information for classification. Therefore, the algorithm cannot recognize objectionable videos with global motions or videos with no considerable motion.

Choi et al [14] proposed the X Multimedia Analysis System (XMAS) for the recognition of obscene video frames. The system provided a method for the recognition of obscene videos based on multiple models and multiclass SVM. The system sampled video frames at a rate of 1 frame/second and used MPEG - 7 visual descriptors for feature extraction. The method used only spatial features and its functionality was restricted to MPEG - 7 files. Most of the existing methods for obscene video recognition use only the information of key frames or spatial information for video classification. Additionally, they generally cannot differentiate between

Figure 1 : Published works (1996-2013) on content based pornography detection

To the best of our knowledge, only a few papers for detecting obscene videos have emerged.

In recent, Lee et al [4] proposed the hierarchical system for detecting obscene videos which consists of three phases called the Early Detection, the Real-time Detection, and the Posterior Detection. In early detection textual information collected from video frames, in the second phase frame by frame checking using image processing tool. Then, the authors encrypt the textual data through a hash function where the result is used as a signature that identifies the video. The SVM trained by the sample images in the database decides the obscenity of the frames. The last phase, Posterior Detection, is performed based on group of frame (GoF) features. Color histogram of each frame in the hue-saturation-value (HSV) space has been calculated using 256 bins and use the averaged histogram as the feature vector for the GoF. The SVM is employed for the detection as well.

A three step method for identifying obscene videos was employed in [5, 6]. In the first step, tensors and motion vectors were utilized to extract key frames. Then, a cube based colour model was employed for skin detection. Finally suspected videos were recognized by the proposed algorithm.

Rea N et al (2006) [7] proposed a multi stage approach for content based obscene detection in videos. In this method, visual motion information and periodicity in the audio stream were utilized to detect obscene videos.

In [8], an adaptive skin detection algorithm was suggested for content based video classification. The algorithm first detected the face area and calculated the skin parameters using the detected face area. Then, using a statistical approach, skin regions were detected.
normal videos with a high volume of skin region and obscene videos.

It has been observed from extensive reviewing the literature that, there are two fundamental concepts in detecting obscenity. i) Skin detection and ii) Color spaces. It will be delineated shortly in section 3.

3 Skin Detection

Skin detection is a bit challenging task due to large variations in appearance, color and shape, occlusion, intensity, location of light source etc [2, 3]. Noise can appear as speckles of skin like color, and many other objects for example wood, cooper and some clothes that are often confused as skin [15]. The noise can also be occurred by illumination that is the change of light source distribution and the illumination level (indoor, outdoor, highlights, shadows, non-white-lights) produces a change in the color. Illumination for the same person can be differed using different camera. Human skin color can be varied from person to person due to ethnicity. As for example skin color for the people of Asian differs with African, Caucasian and Hispanic groups. Some other factors such as age, sex, body parts, makeup, hairstyle, costumes, background colors, shadows, motion etc also affects the skin color appearance [16]. In general, human skin is characterized by a combination of red and melanin (yellow and brown) and there is somewhat a range of hue for skin and saturation that represent skin-like pixels.


3.1. Color Spaces

Low level image processing is requires for identifying obscenity from video. One of the major preprocessing tasks is selecting color space and model. The choice of color space can be considered as the primary step in skin-color classification. The RGB color space is the default color space for most available image formats. Any other color space can be obtained from a linear or non-linear transformation from RGB. The color space transformation is assumed to decrease the overlap between skin and non-skin pixels thereby aiding skin-pixel classification and to provide robust parameters against varying illumination conditions. It has been observed that skin colors differ more in intensity than in chrominance [22]. Hence, it has been a common practice to drop the luminance component for skin classification. Several color spaces have been proposed and used for skin detection. In this section, we review the most widely used color spaces for skin detection and their properties. The following are widely used color model [22, 23, 24 and 25]:

i) Normalized RGB
ii) HSI
iii) HSV
iv) HSL (Fleck HSV)
v) TSL
vi) YCbCr
vii) YES, YUV, YIQ

3.2 Experimental Results on Skin Detection

In order to proper justification we tested several skin detection methods. For simplicity we used a sample video of 7 minutes duration as input then extracted all key frames. It is to be noted that our system considered 25 Frame per seconds (FPS) and there are total 10,500 frames of which 447 key frames were extracted. Summary of the results are as follows:

<table>
<thead>
<tr>
<th>Method</th>
<th>Identified Key Frames as Objectionable</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>TPR</td>
</tr>
<tr>
<td>Rehg and Jones skin color mixture of Gaussian model [32]</td>
<td>29</td>
<td>96.2%</td>
</tr>
<tr>
<td>Big skin detection for adult image identification [17]</td>
<td>19</td>
<td>84.3%</td>
</tr>
<tr>
<td>Skin Map [33]</td>
<td>27</td>
<td>94.3%</td>
</tr>
</tbody>
</table>

It has been observed that Rehg and Jones [32] method is the best among the three methods but it has quiet large FPR. According to the integrity, Skin map [33] method proved to be the best skin detection method because it has low FPR and...
4 Open Research Issues

In this section some emerging open research issues related to this paper will be discussed.

4.1 Choice of Classifiers and Feature Selection

Video is consists of frames, group of frames can form shots and semantically meaningful shots can be considered as scene. Hence the primary steps of any video processing are video segmentation which can be made either by frame, shot or scene level segmentation. Researchers need to pay attention for temporal behaviour of video content. For example similarities between frames, hard cuts, camera movement, audios etc are the key points of content based video processing. In order to achieve obscenity most of the researchers followed shot and cut detection, shot duration, motion capture, audios, illumination rate etc for identifying obscenity. In addition some image processing can also be utilized for accuracy. There are some problems such as partly nude, non obscene but sexual exposure etc. To reduce false detection researchers are paying more attention on video contents but in that case performance of the system degrades. In this situation cloud based system can make the system more robust. But the problem is to cope video processing into the cloud computing system. Substantial development requires achieving this goal.

Hybrid classifiers (SVM and Adaboost) can be use for better performance. Sometimes censored videos may also be useful for education purposes or may be categorized according to age and culture. These things will be done perfectly by incorporation of knowledge base [34 and 35]. Necessary steps could be done, training the stored database and auto updateable database. In addition unsupervised transfer learning [26] can be applicable instead of traditional supervised classification because it is impossible to classify specific nude picture using a predefined set of dataset. Since there are more and more different kinds of objectionable pictures, the traditional machine learning algorithm maybe perform inefficient to find new type of nude picture based on the old training datasets. In this situation transfer learning can be applicable to assist the discovery procedure.

4.2 Multimedia Cloud Computing

Multimedia cloud computing becoming a popular research topic due to wide spread information sharing and up gradation of network bandwidth. There are many works has been accomplished in cloud computing handling multimedia (Audio, video, image) but very little works in processing content based multimedia. [27, 28, 29 and 30]. Still now there are only three methods to process contents of images in cloud computing. i) HIPI (Hadoop Image Processing Interface [30]) ii) Hadoop SEQUENCE files [30] and iii) BIGS [28]. Sequence files perform better than standard applications for small files, but must be read serially and take a very long time to generate [30]. HIPI Image Bundle have similar speeds to Sequence files, do not have to be read serially, and can be generated with a MapReduce program [30]. Additionally, HIPI Image Bundles are more customizable and are mutable, unlike Sequence files.

For instance, HIPI has the ability to only read the header of an image file using HIPI Image Bundles, which would be considerably more difficult with other file types. Those three methods of handling images in Hadoop can process images based on some ground truth and image statistics. No work efficiently process video according to appropriate machine learning scenario. In [31] the use of parallel SVM in Hadoop has been elucidated but still there in need rigorous analysis. In this perspective, scholars can think about how to fit machine learning strategies in Cloud Computing architecture.

5 Discussions and Conclusions

Most of the existing works are based on skin color region detection which can’t perfectly recognize whether a video contain obscenity in general sense. The reason for this is that, there are many skins like pixels such as beach photos, human skin like animal’s fur, skin colored painting etc which enable false positive rate. In addition all existing works perform well on particular set of video dataset. It has been observed that, in spite of successfully applying skin color segmentation and geometrical structure of human body, eventually it was failed to detect naked body perfectly due to absence of machine learning tools [1]. On the other hand, using skin color segmentation and machine learning tools improve detection rate. Here choice of appropriate color models, skin detection algorithms and classifiers are the factors of performance. The performance of pornography detection has been dramatically changed using human body parts specifically erotogenic parts. In this case choice of proper color model, skin segmentation algorithms, classifiers and human erotogenic body parts are the factors.

In all cases a specific dataset has been applied for test images and hence the detection could not satisfy for all types of arbitrary nude images. Because there are huge amount of nude images available in their different pose, angle, illumination condition, partial occlusion, highly and partially exposed form. In this case unsupervised transfer learning could be a solution because it creates new data sets from already learned old datasets and thus would perform well on random unlabelled nude picture identification.

There are some common trends in every detection algorithms taking consideration of accuracy and performance. These two things are inversely proportional. To minimize this challenges parallel and distributed systems can be applicable. If we consider accuracy of obscenity detection then should pay attention for substantial improvement of existing image based obscenity detection or if we consider a suitable existing method and want to improve the performance of the system then we should pay attention for applying parallel and
distributed system. A cloud computing system can be utilized in this perspective. Content based image processing in the cloud is still an open research issue. The reason for this is that in cloud computing paradigm only text based data can be recognizable and there is no such built in tool to handle byte oriented image data [Mohamed H. Almeer (2012)]. Some scholars indicated to handle this problem using Hadoop SEQUENCE file [Hadoop (2013)]. But this technique also not had been proved yet. Recently there are two tools has been deployed to process large-scale image such as HIPI [30] and BIGS [28]. These two techniques devoted to process some special images for example remote sensing and medical images. It is not sure whether it can work on objectionable images or not. There is another problem of processing images in distributed environment. Image or frame can’t be split during the processing phase because it affects the quality of the original video frame during merging [27]. If those issues could be resolved then the performance of the CBPD (Content based pornography detection) will increase significantly.

Acknowledgements
This research was supported in part by Shenzhen Technical Project (grant no. HLE201104220082A) and National Natural Science Foundation of China (grant no. 61105133) and Shenzhen Public Technical Platform (grant no. CXC201005260003A)

6 References


