Protecting The Networks from Objectionable Contents

Chang-Yul Kim[1], Oh-Jin Kwon[1]*, Seokrim Choi[1], and Yong-Hwan Lee[2]
[1]Sejong University, Seoul, Korea
changyul.kim@gmail.com, {ojkwon, schoi}@sejong.ac.kr
[2]Far East University, Gamgok, Chungbuk, Korea
hwany1458@empal.com

Abstract

Along with the ever-growing Web, the objectionable videos are easy to spread and have become a serious problem for employers. Besides, they fall in the productivity of employees and may cause even legal problems. In this paper, we present a practical technique to examine the obsceneness of the video that is transmitted through the internal network. This paper does not describe the steps of acquiring a video packet and its decoding process because there are a lot of ways of how to do it. The proposed technique is designed as a multiple procedure to use the least amount of resources in the system. In addition, this technique has an advantage that can block the video playback before obscene scene is exposed because the video is filtered on a frame-by-frame basis. The experimental results show an accuracy of 94% is about the videos of various genres collected at a random.

Keywords: Objectionable, obscene, filtering

1 Introduction

Today, in the company, the network usage of non-business purpose is growing more and more. This undesirable network traffic comes from the site access such as game, stock, entertainment, and adult web page. Among them, adult pages usually have the functions of video streaming and so they occupy a huge amount of network bandwidth. If we do not handle the network connection of the objectionable video, there is the possibility that such network slowdown, bigger network cost, and reduced work productivity. On the other hand, employees share objectionable contents via internal network is a serious concern because most employers have a legal duty to protect their employees from sexual harassments. And sending and receiving them is a security threat because they might also include various types of malware. In this paper, we propose a real-time discrimination technique to identify the obsceneness of the video file that occupied high bandwidth of internal network. Video files through the internal network are usually delivered using the real-time multimedia streaming protocols such as RTSP [9] or HLS [8]. Regardless of which protocol was used, we can obtain the raw data of the video file when the video playback application is decoding the packet data received. There are various techniques that can sniff the packets being sent. However, even if the network packet successfully de-packetized into the video stream, if the stream is DRM encrypted data, we cannot see the contents of it without the corresponding decryption key. We do not describe the packet sniffing techniques of video data because they are beyond the scope of this paper. Previous research efforts on screening obscene images can be found in literatures [2, 3, 11]. Forsyth and Fleck proposed an algorithm for detecting naked people in images by employing skin and geometric filters [2]. Wang et al. proposed a system called wavelet image pornography elimination (WIPE), which uses an algorithm involving color histograms, wavelets, and normalized central moments [11]. Jones

Research Briefs on Informaiton & Communication Technology Evolution (ReBICTE), Vol. 1, Article No. 19 (January 15, 2015)
*Corresponding author: Sejong University, Dept. of Electronics Engineering, 209 Neungdong-ro, Gwangjin-Ku, Seoul, Korea, 143-747, Tel: +82 2-3408-3295

and Rehg proposed a method for skin detection [3]. Jones and Rehg segment skin pixels by estimating the distribution of skin and non-skin colors and extract features for the detection of obscene images. Although these algorithms have presented encouraging results, most of them are for images and are insufficient for obscene video detection due to the lack of exploiting temporal features.

Recently, Lee et al. 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 [6]. In the Early Detection phase, the authors perform the detection using hash signatures by simply extracting the textual data such as file size, video size, frame rate, and etc, from the header of the video files. Then, the authors encrypt the textual data through a hash function where the result is used as a signature that identifies the video. The authors store the signatures of known obscene videos in the database and decide whether a video is obscene if the signatures match. This phase is very fast but it cannot be used for new obscene videos with signatures that are not included in the database. The second phase, Real-time Detection, is performed frame by frame. The authors extract skin color regions and obtain an image that only contains skin regions by removing non-skin regions. Then, the authors resize it to the standard size of 40×40 and use the resulting image as the input feature vector for a support vector machine (SVM). The SVM trained by the sample images in the database decides the obsceneness of the frames. The last phase, Posterior Detection, is performed based on group of frame (GoF) features. They calculate the color histogram of each frame in the hue-saturation-value (HSV) space using 256 bins and use the averaged histogram as the feature vector for the GoF. The SVM is employed for the detection as well.

This paper proposes an efficient and fast system for detecting obscene videos. As done in [6]. its aim is to determine if a video is obscene or not whereas the decisions on individual frames are of little interest. Our system focuses on the speed of frame based real time detection corresponding to the Real-time Detection phase in [6]. Therefore, we designed the system to avoid the GoF processing because of its time consuming nature as stated in [6]. Our system consists of five procedures. Frames do not necessarily go through all the procedures. We believe that this multi-step approach enables our system to become fast and efficient. To save the computational cost, our system assumes that title frames, frames at shot boundaries, and frames with global motion may be bypassed with negligible performance degradation. Some frames, such as title frames, have logos or texts in simple colored backgrounds. If a video contains obscene title frames, in most cases, the video has obscene frames in another part of the videos as well. In general, content-based video analysis is performed shot by shot because shot boundaries indicate a semantic change of contents. We may observe that, if the frames at shot boundaries are obscene, the interior frames of the shot are always obscene. We have also observed the relationship between the global motion and the obscene frame. If the frames of global motion are obscene, it is normally observed that the nearby frames without global motion are also obscene. We utilize all these observations for reducing the processing time. In section 2, our system for detecting obscene videos is described. Experimental results and conclusions are given in section 3 and 4, respectively.

2 Proposed System

The structure of proposed system is shown in Figure 1. Solid and dashed arrows mean matched and unmatched passes, respectively. After extracting frames from a video, we detect whether each frame belongs to title frames, shot boundaries, or global motions. For frames that do not belong to these categories, we segment the frames into regions with similar colors and detect skin segments. Next, our system refines the skin segments by analyzing them. If a specific segment is larger than pre-defined size, it is regarded as a body region. The last step performs shape matching by using Zernike moments. The reason why we use Zernike moments is that they are useful for the frames taken from a wide variety of

camera angles. They are recently known as the efficient descriptor invariant to translation, rotation, and scaling of object shape.

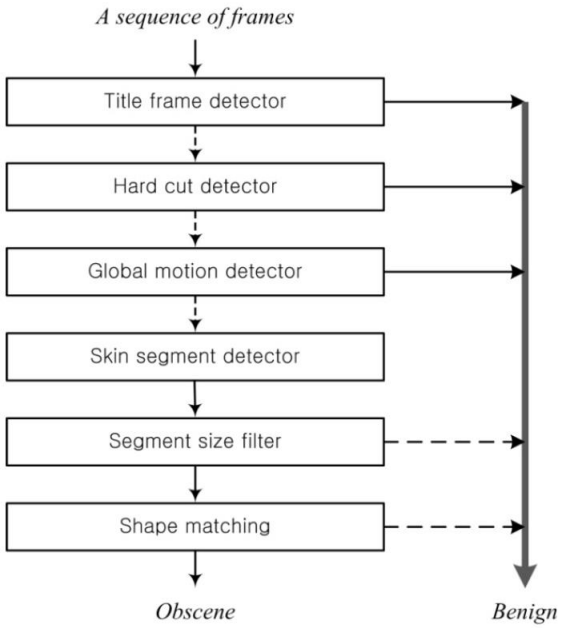


Figure 1: Basic structure of the proposed system

2.1 Title Frame Detector

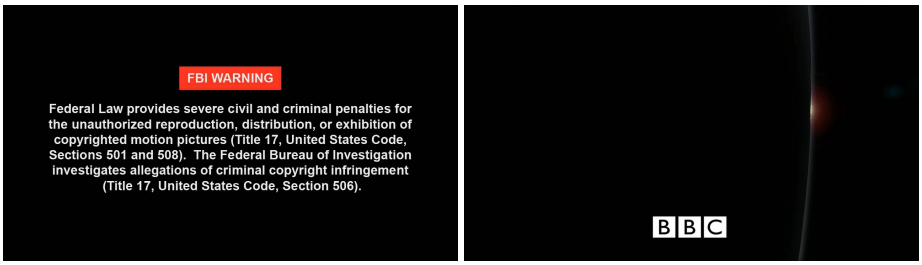
A video file consists of consecutive frames that contain various scenes. Some videos like movie contain the frame the frame that has only text or logo image as foreground on monotonous background. We called this frame as 'Title frame' because these are mostly found at the front part of a video. Occasionally there is also the title frame that is included at the rear part of a video. Typical examples of the title frame are shown below. Figure 2 shows some examples of the title frame that is contained at the front and the rear part of a video, respectively.

For the detection of title frames, we perform a histogram analysis. To speed up the processing time, we quantize each pixel into 512 colors (8 levels for each red, green, and blue component, a total of $8 \times 8 \times 8 = 512$ colors). If the color histogram of a frame shows steep peaks, we decide that the frame is a title frame.

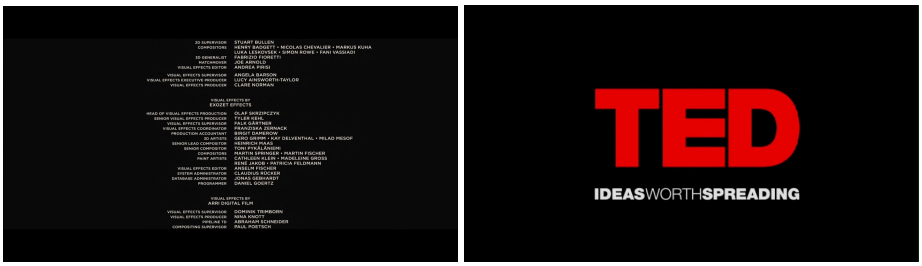
2.2 Hard Cut Detector

Content-based video analysis is normally performed shot by shot because shot boundaries indicate a semantic change of contents. The obscene frame basically include skin color region. Figure 3 shows that the frames inside the shot contain more skin color than the frames at shot boundaries. Thus we exploit a property that the interior frames of a shot have a higher probability of obsceneness.

In general, shot boundaries are classified into three types of edits: hard cut, fade in/out, and dissolve. In our algorithm, the type of shot boundaries is of little importance. To speed up the processing time, we assume that all shot boundaries of fade in/out and dissolve last less than the threshold value of T frames, which represents the maximum shot transition time. Also, we only detect hard cuts between two frames



(a)



(b)

Figure 2: Some examples of the title frame. (a) and (b) are contained at the front part of a video. (c) and (d) are contained at the rear part of a video.

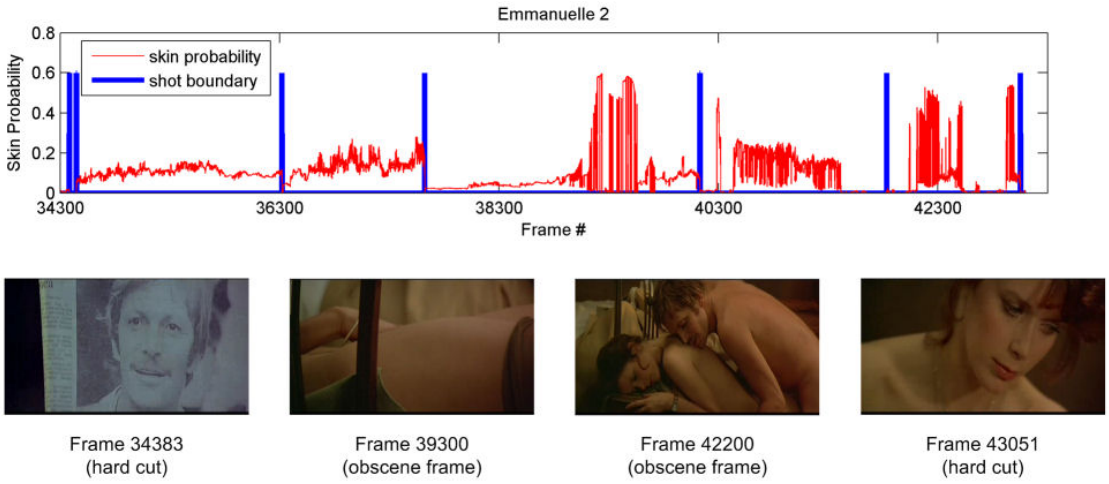


Figure 3: Example of hard cuts and a probability of skin frame

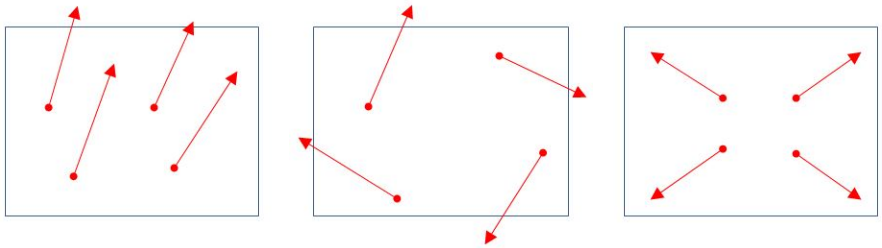
of T distance. The basic idea is that color histogram does not change abruptly within but across shots. The color histogram difference between two frames I_i and $I_i + T$ is given by

$$CHD_i = \frac{1}{N} \sum_{r=0}^{2^B-1} \sum_{g=0}^{2^B-1} \sum_{b=0}^{2^B-1} |p_i(r,g,b) - p_{i+T}(r,g,b)| \tag{1}$$

where $p_i(r,g,b)$ is the number of pixels of color (r,g,b) in the i-th frame I_i of N pixels. B is set to be 3 as done in the title frame detection. A hard cut is detected if the CHD_i exceeds a certain threshold.

2.3 Global Motion Detector

The global motion may be represented by the camera motion of translation, rotation, and zoom in/out as shown in Figure 4. Our purpose does not demand estimating the exact global motion parameters. It is enough to decide whether a frame has a global motion or not. For simplicity, we assume that macro-blocks within a frame show similar motions. We perform the motion estimation by nearest neighbor search using 16×16 sized macro-blocks. When the motion vectors of macro-blocks at four boundaries of a frame are small in magnitude, we decide that the frame does not have a global motion.



(a) (b) (c)

Figure 4: Typical example of global motions. (a) pan and tilt, (b) rotation, and (c) zooming.

2.4 Skin Segment Detector and Size Filter

Skin regions are mostly extracted by pixel-wise color groupings and texture analysis is performed on such regions to remove the small pseudo-skin patches that may appear in the backgrounds such as walls or floors. These approaches are time-consuming. In our segmentation, we use a labeling algorithm based on the union-find structure [10]. In our segmentation, we use a modified labeling algorithm. The original method is designed to operate on binary images and labels the connected regions. This method performs two passes over the image: the first pass to record connections and assign temporary labels and the second pass to replace each temporary label by the label which it finally belongs to. Using Figure 5, we exemplify this method. The sample binary image is shown in Figure 5 (a). The results of the first and the second pass are shown in Figure 5 (b) and 5 (d), respectively. In the first (top-left to bottom-right) pass, if the current pixel is 1 and one of the upper or the left neighboring pixel is 1, the label of a neighboring pixel is assigned to the current pixel, otherwise, a new label is assigned. If both the upper and the left pixel are 1, the lower label is assigned to the current pixel and the information that both labels are connected ones is recorded in the so called PARENT array. Figure 5 (c) shows an example of the PARENT array. In this example, the label 3 and 6 are the ones connected to the label 1 and 5, respectively. In the second pass, the connected labels are merged to the lower labels. We use the HSV color space because this color space is known as the most discriminative space for human skins under arbitrary illuminations and races [1]. We merge two neighboring pixels or segments if the Euclidean distance of their averaged colors is less than the threshold. Otherwise, they are labeled as different segments.

Examples of our segmentation are shown in Figure 6 (c) and (d). After the segmentation, we detect skin segments because obscene images generally contain large areas of skin colors. For detecting skin segments, we define the following function.

$$SKIN(segment) = \frac{1}{N} \sum_{i=0}^{N-1} skin(y_i, cb_i, cr_i) \tag{2}$$

Function $SKIN(segment)$ indicates the proportion of skin pixels within the given segment. If the

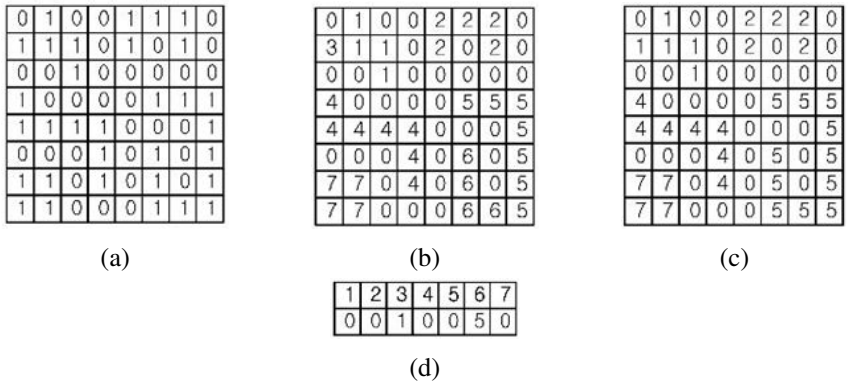


Figure 5: Example of the original labeling method with union-find structure: (a) binary image, (b) result of the first pass, (c) PARENT array, and (d) result of the second pass.

value of this function is above the threshold, the corresponding segment is declared as a skin segment. N is the number of pixels in the given segment and $skin(y_i, cb_i, cr_i)$ indicates whether the i-th pixel with an (y, cb, cr) color value is skin pixel or not. As shown in [5], if the (y, cb, cr) color value of a given pixel satisfies the following equation, we classify it as a skin pixel and $skin(y, cb, cr) = 1$. Otherwise, $skin(y, cb, cr) = 0$.

$$(y, cb, cr) \text{ is classified as skin if:}$$

$$\begin{array}{c} Y > 80 \text{ and} \\ 85 < cb < 125 \text{ and} \\ 135 < cr < 180 \end{array} \tag{3}$$

$$\text{where, } y, cb, cr = [0; 255]$$

When a segment is classified as the skin segment, it passes through the size filtering process, which removes small skin-like patches. Then, only sizable segments of skin color survive as body regions. If a frame does not contain any sizable body region, we regard the frame as a benign frame. Examples of our body regions are shown in Figure 6 (e) and (f).

2.5 Shape Matching

For the last step, we perform shape matching using Zernike moments. Recent researches on Zernike moments show promising performance as affine invariant region-based shape descriptors [7]. Combined with the SVM, they can be used as efficient methods for shape detection. We represent the extracted body region using the Whitening Zernike Moment Shape Descriptor (WZMSD) [7]. The WZMSD performs whitening on the shape pixel data, represents the shape in a canonical form, and extracts Zernike moments. The whitening process of WZMSD provides fast processing because it avoids the rotation operation which is essential and the most complex part in extracting affine invariant shape descriptors [7]. Our system extracts the first 36 Zernike moments by using the WZMSD. Once the moments are computed, we determine the obsceneness of each body region by the SVM. We use the normalized Zernike moments as input vectors to the SVM. The SVM is trained by the body regions in our 100 sample images composed of half obscene and half benign images. If at least one body region in a frame is classified as obscene by the SVM, we declare the frame as an obscene frame.

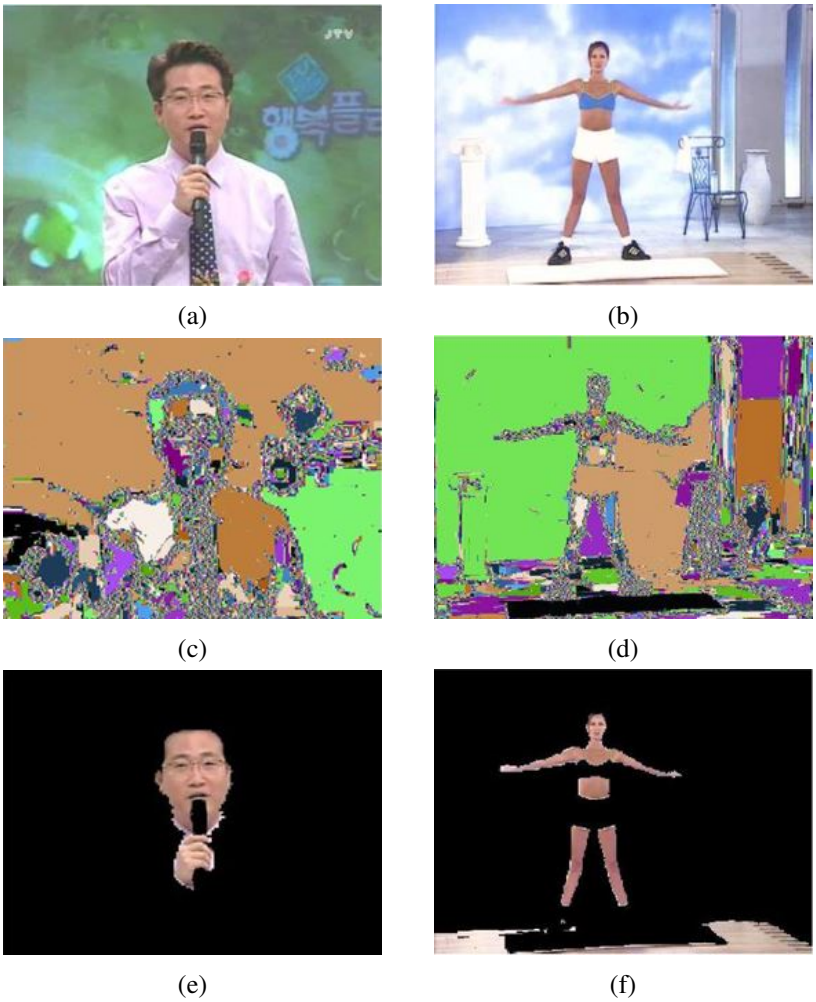


Figure 6: Examples of segmentation: (a) and (b) are input images, (c) and (d) are their segmented images, and (e) and (f) are their body regions, respectively.

2.6 Temporal Reduction of Frames under Inspection

If we apply aforementioned procedures to every frame, it is too time consuming. In this paper, we assume that the transition time of all shot boundaries is less than the maximum shot transition time T and all obscene shots last longer than T. Therefore, if the i-th frame is classified as a benign one, we skip to the $(i+T)$-th frame for the next inspection. If we increase this value, the processing time proportionally decreases whereas the false decision rate for obscene videos increases. For practical usages, it is recommended to set T as the number of frames corresponding to 1 to 5 seconds. We finally decide the video as obscene if we detect more than five successive obscene frames. This enables us to save the processing time quite a lot.

3 Experimental Results

The performance of obscene video detecting system is highly dependent on sample videos used in the experiment. In our experience, we have found that the performance normally varies widely depending on the genre, the running time, and the quality of videos. As shown in our previous paper [4], we did

experiments to determine whether the obscene about the 1,000 video files. Unlike the paper [4], in this paper, the detection of skin segment was performed in the YC_bC_r space, the shape matching stage used a total of 100 training sample. We have randomly collected 100 video files composed of half benign and half obscene videos. As with [4], these sample videos have a variety of motions, resolutions, frame rates, running times, and genres including movies, television shows, news, sports, home videos, etc. Obscene samples also cover a wide range of situations including single and multiple, naked and bikini, and indoor and outdoor scenes. In our experiments, the maximum shot transition time T has been set to the number of frames equivalent to 2 seconds. This means that we inspect only one frame per 60 frames of videos whose frame rate is 30 frames per second.

	Number of frames under inspection	Number of frames filtered by			
		Title frame detector	Hard cut detector	Global motion detector	Skin segment size filter
Obscene videos	186,523	3,639	1,721	6,982	11,298
Benign videos	124,525	4,989	2,845	15,239	76,427

Table 1: Number of filtered frames

Firstly, we measure the overall video classification performance of the system. In this paper, we use the detection rate (DR) and the false positive rate (FPR) as performance measures. The DR and the FPR are defined as the true decision rates for obscene videos and the false decision rates for benign videos, respectively [6]. The DR fo the proposed system is 95.4% and the FPR of it is 7.6%.

Secondly, we compare our performance of frame classification with the WIPE system, which is designed for detecting obscene images [11]. We chose the WIPE because it is known for its reliability and is the most referenced by other papers for filtering obscene images. We perform this experiment by substituting the procedures of section 2.4 and 2.5 in our system by the WIPE. The DRs of our system and WIPE are 95.4% and 82.4%, respectively. The FPRs of our system and WIPE are 7.6% and 10.1%, respectively. These results prove that our system outperforms the WIPE when directly applied to the video frames. We have found that the accuracy difference between two systems is mainly due to the characteristics of video samples. The WIPE has been designed mainly for still images. Compared to the images for which it aimed, the frames of our video samples tend to show relatively more distortions such as blocking artifacts, low contrast, interlacing, etc. It has been known that the skin detector used in the WIPE is more sensitive to the quality of images.

We also compared the processing times of the two systems. Our video samples show various frame sizes. In order to make the results independent of frame size, we have collected the videos whose frame size is most popular on web sites. 50 video files whose frame size is 640×480 have been chosen and the processing time per frame on our i7 2.4GHz quad-core platform has been calculated and averaged. The time required to discriminate one frame each took 16.5 milliseconds and 36.2 milliseconds in our system and the WIPE. The wavelet transform for texture analysis and shape matching of the WIPE system has demanded more processing time than our spatial domain process.

We have also inspected the number of frames filtered by each procedure for obscene and benign videos and the results are shown in Table 1. The difference between obscene and benign videos is due to the fact that benign videos have relatively more hard cuts and global motions than obscene videos. In practical usages, most videos for inspection are benign. Therefore, it is noteworthy that only 20% frames of benign videos have survived up to the final decision step. Moreover, about 61.4% of the frames have been filtered by the skin segment size filter.

4 Conclusions

This paper presents multi-step framework for detecting obscene videos. All the design parameters are optimized for reducing processing time. We have described several assumptions for bypassing frames under inspection. It has been observed that these assumptions decrease both the processing time and the false positive rate. We have implemented the fast body region extracting algorithm based on the shape matching using Zernike moments. The magnitude of the Zernike moments is rotation invariant feature. Therefore, this feature can be used for object recognition with another feature descriptor. Experimental results have shown that our algorithm is very fast with high detection precision and suggests a practical tool for detecting obscene videos. In the future, our research will be focused on finding additional methods useful for reducing processing time. It is thought that additional assumptions for bypassing frames with negligible performance degradation and adjusting the maximum shot transition time T adaptively based on the shot characteristics may improve the performance further.

Acknowledgements

This work was supported by the ICT R&D Program of MSIP/IITP [2014(I5501-14-1007), 3D Smart Media / Augmented Reality Technology, KCJR Cooperation International Standardization].

References

[1] T. Dzmitry, H. Mohamed, and C. Liming. Face detection in video using combined data-mining and histogram based skin-color model. In *Proceedings of the 3rd International Symposium on Image and Signal Processing and Analysis*, volume 1, pages 500–503, 2003.

[2] D. A. Forsyth and M. M. Fleck. Automatic detection of human nudes. *International Journal of Computer Vision*, 32:63–77, 1999.

[3] M. J. Jones and J. M. Rehg. Statistical color models with application to skin detection. *International Journal of Computer Vision*, 46:81–96, 2002.

[4] C.-Y. Kim, O.-J. Kwon, and S. Choi. A practical system for detecting obscene videos. *IEEE Transactions on Consumer Electronics*, 57:646–650, 2011.

[5] G. Kukharev and A. Nowosielski. Visitor identification - elaborating real time face recognition system. In *Proceedings of WSCG (Short Papers) '04*, pages 157–164, 2004.

[6] S. Lee, W. Shim, and S. Kim. Hierarchical system for objectionable video detection. *IEEE Transactions on Consumer Electronics*, 55:677–684, 2009.

[7] Y. Mei and D. Androutsos. Robust affine invariant region-based shape descriptors: the ICA Zernike moment shape descriptor and the whitening Zernike moment shape descriptor. *IEEE Signal Processing Letters*, 16:877–880, 2009.

[8] R. Pantos and W. May. Http live streaming. `http://tools.ietf.org/html/draft-pantos-http-live-streaming-13`, last viewed October 2014, 2014.

[9] H. Schulzrinne, A. Rao, R. Lanphier, M. Westerlund, and M. Stiemerling. Real time streaming protocol 2.0. `http://tools.ietf.org/html/draft-ietf-mmusic-rfc2326bis-36`, last viewed October 2014, 2013.

[10] L. G. Shapiro and G. C. Stockman. *Computer Vision*. Prentice Hall, 2001.

[11] J. Z. Wang, J. Li, G. Wiederhold, and O. Firschein. System for screening objectionable images. *Computer Communications Journal*, 21:1355–1360, 1998.

Author Biography

Chang-Yul Kim received B.S. and M.S. degrees in computer science from Sejong University, Seoul, Korea, in 1998 and 2000, respectively. Currently he is working at SK planet, at the same time is in the doctorial course in electronics engineering at Sejong University. His research interests are image and video coding, object segmentation, and object recognition.

Oh-Jin Kwon received B.S. degree from Hanyang University, Seoul, Korea, in 1984, the M.S. degree from the University of Southern California, Los Angeles, 1991, and the Ph.D. degree from the University of Maryland, College Park, in 1994, all in electrical engineering. From 1984 to 1989, he was a research staff member at the Agency for Defense Development, Korea, and from 1995 to 1999, he was the head of Media Lab in Samsung SDS Co., Ltd., Seoul. Since 1999, he has been a faculty member with Sejong University, Seoul, Korea, where he is currently an Associate Professor. His research interests are image and video coding, watermarking, analyzing, and processing.

Seokrim Choi received B.S. degree from Hanyang University, Seoul, Korea, in 1984, the M.S. degree from the University of Southern California, Los Angeles, 1991, and the Ph.D. degree from the University of Maryland, College Park, in 1994, all in electrical engineering. From 1984 to 1989, he was a research staff member at the Agency for Defense Development, Korea, and from 1995 to 1999, he was the head of Media Lab in Samsung SDS Co., Ltd., Seoul. Since 1999, he has been a faculty member with Sejong University, Seoul, Korea, where he is currently an Associate Professor. His research interests are image and video coding, watermarking, analyzing, and processing.

Yong-Hwan Lee received the M.S. degree in Computer Science and the Ph.D. in Electronics and Computer Engineering from Dankook University, Korea, in 1995 and 2007, respectively. Currently, he is an assistant professor at the Department of Smart Mobile, Far East University, Korea. His research areas include Image/Video Representation and Retrieval, Image Coding, Face Recognition, Augmented Reality, Mobile Programming and Multimedia Communication.