Journal of Engineering and Applied Sciences 13 (5): 1174-1182, 2018
ISSN: 1816-949X

Pornographic Video Detection Scheme Using Multimodal Features

Kwang Ho Song and Yoo-Sung Kim
Inha University, Incheon, Korea

Abstract: In this study, we propose the new pornographic video detection scheme using multimodal feature such as image features of each frame using deep learning architecture, image descriptor features of the frame sequence, motion features using optical flow and audio features extracted from video. By using these various features at once we can detect almost all pornographic events without being confused by a specific element of input video. And as the performance evaluation results we can obtain 100% true positive rate and 67.6% overall accuracy. Although, the overall accuracy is little bit low due to high false positive rate we could successfully detect the pornographic videos which are difficult to detect by using only single modal features.

Key words: Pornographic video detection, multimodal features, convolutional neural network, optical flow, spectrogram, support vector machine

INTRODUCTION

As the amount of pornographic video distributed in the online environment increases, the necessity of automatic filtering schemes and national regulations is consistently increasing (Anonymous, 2017). However, because the most of existing researches about automatic filtering pornographic videos are focused on only using the features extracted from the video frame (Moustafa, 2015; Moreira *et al*., 2016; Song and Kim, 2017), they have weaknesses to distinguish some hardly distinguishable scenes such as the situations of baby feeding or sucking sticky things in naked state which are very likely to the pornographic ones or the scenes of masturbation or groaning while wearing clothes which are not easy to differentiate from the pornographic ones.

In order to overcome this drawback in recent years, a method using multimodal features which can be obtained from video contents has been actively researched rather than using only one specific feature (Liu *et al*., 2014; Perez *et al*., 2017; Tablatin *et al*., 2016; Punaiyah and Singh, 2017). These previous research can be classified into two categories according to the combination features as follows: using visual and audio features and using visual and motion features. Liu *et al*. (2014) use the audio features and visual ones to develop the automatic detector of pornographic videos. The scheme uses the generated high level audio feature of each segment which is called Energy Envelope Unit (EEU) using Bag of Word (BoW) framework and the some simple low level video features of key frame which is synchronized with each EEU. In succession, the results of the both classifiers trained by audio and video features are combined into a final classification result through a two-step decision process consisting of weighted fusion and filtering based on the threshold.

On the other hand, Perez *et al*. (2017) proposed a multimodal pornographic video detector using visual features and motion features. The frames of video are transformed into the motion frames using optical flow algorithm (Brox *et al*., 2004) and the high level motion and visual features are extracted from the motion and video frames using deep learning architecture. Based on the combination of these features, the Support Vector Machine (SVM) classifier is trained and used to detect porn video.

These multimodal pornography detection methods have the advantage of detecting the certain type of video that can't be detect by using only visual features. However, the former method, multimodal method based on audio and visual features has a lower False Positive Rate (FPR) performance than the other works. And the latter, multimodal method based on motion and visual features, still cannot detect certain situations such as the scenes with a dressed person who perform sexual actions within static movements or groaning likewise masturbation.

Therefore, we propose the new pornographic video detection scheme based on multimodal features such as image features of each frame using deep learning architecture, features of the frame sequence made by image descriptor, motion features using optical flow and audio features extracted from video based on audio-signal frequency. By using these various features at once, we can detect almost all pornographic events without being confused by a specific element of input video.

Corresponding Author: Kwang Ho Song, Inha University, Incheon, Korea

Literature review: From the past to the present, the almost methods for pornographic video detection used the low level feature related with visual element of video such as color, shape or its general distribution pattern. However, these models have drawbacks to easily generate false positive and false negative on a little bit difficult data and (Adnan and Nawaz, 2016) confirmed that the low-level visual features are insufficient for the pornography detections because their inherent disadvantages. For this reason, researchers have been interested in pornographic videos detection using high-level visual features (Moustafa, 2015; Moreira *et al*., 2016; Song and Kim, 2017 2017) or multimodal features (Liu *et al*., 2014; Perez *et al*., 2017) which can compensate the drawbacks of the former ones.

At first, the researches on pornographic video detection using high-level visual features are categorized into two classes according to the method used for feature extraction. First method uses Bag of Visual Word (BoVW) framework to extract high-level visual feature (Moreira *et al*., 2016). The BoVW is one of the transformation methods to make high-level visual feature and it can close the semantic gap between the original video data and the high-level target concept, so, it can be used to compensate the shortage of using simply the appearance of the specific body parts to detect pornographic videos. The process of BoVW consists of 3 phases generally. At the first phase, the local descriptor of frame image is created by various description algorithms. Next a codebook is made up through the Gaussian Mixture Model (GMM) or clustering algorithm after generating code word which can represent the video contents globally based on low-level features. At last, making classifier with the training dataset encoded as combination of the probability of each code word. So as a result, this model got a maximum 95.8% accuracy. But this method has some weakness such as the high complexity of model due to using numerous algorithms to create code words and numerous efforts to find appropriate hyper parameter for algorithms.

Therefore, as another method to make high-level visual feature without excessive efforts, the researchers pay attention to classifier re-training technique which uses the well-known pre-trained Convolution Neural Network (CNN) as a high level visual feature extractor (Moustafa, 2015; Song and Kim, 2017). These researches commonly use the original weights or parameters of well-known pre-trained CNN Model to extract high-level visual features and use retraining technique such as fine-tuning or transfer learning to make appropriate classifier for new classification domain. Moustafa (2015), researchers fine-tuned pre-trained deep learning image classifier, AlexNet (Krizhevsky *et al*., 2012) and GoogLeNet (Szegedy *et al*., 2014) to extract high level visual features from image. And based on the features, the image classifier was trained to use for integrated decision making of video domain. As a result, it can get 94.1% as the maximum accuracy. Though this method could help to get high-level image features easily and make fine classification performance without too much effort to implement the entire model at the end of decision making, it is difficult to place full confidence in its performance because it can vary depending on the integration method of each frame's decision results.

Thus, as one of our previous research by Song and Kim (2017) to make a single high level visual feature from videos which called video descriptor without excessive efforts to implement and make robust classifier, we proposed the pornographic video detection scheme using video descriptor. In that previous research, we introduced two kinds of pornographic video detection schemes. And the first one is based on image descriptor and result integration as shown in Fig. 1.

The process of generating image descriptor for pornography detection scheme consists of 4 phases as shown in Fig. 1. The frame image of input video is extracted and reformed in the first phase. In the second phase, feature extraction we use the part of pre-trained VGG-16 (Simonyan and Zisserman, 2014) from first convolution layer to second fully connected layer as a feature extractor to convert frame image into image descriptor.

So, while passing through the five convolution layers of VGG-16, the input images are converted into generic features such as diagonal, vertical, horizontal, circle line, various colors and so on. Subsequently, in the next two fully connected layers, the generic features are combined each other variously and through this process, we can get the high level feature vectors called the image descriptor. Based on the image descriptor of each frame, the classifier is trained in the third phase. And whether each frame may correspond to pornography or not is decided by the previously learned classifier. After then, in the last phase, the decision results are integrated into a single result for the input video by certain integration method such as max pooling, average pooling and so on. The other kind of detection scheme is based on single video descriptor as shown (Fig. 2).

It consists of also 4 phases and also uses the part of pre-trained VGG-16 (Simonyan and Zisserman, 2014) as a feature extractor to generate image descriptor. Therefore, the first and second phases are same to those in the

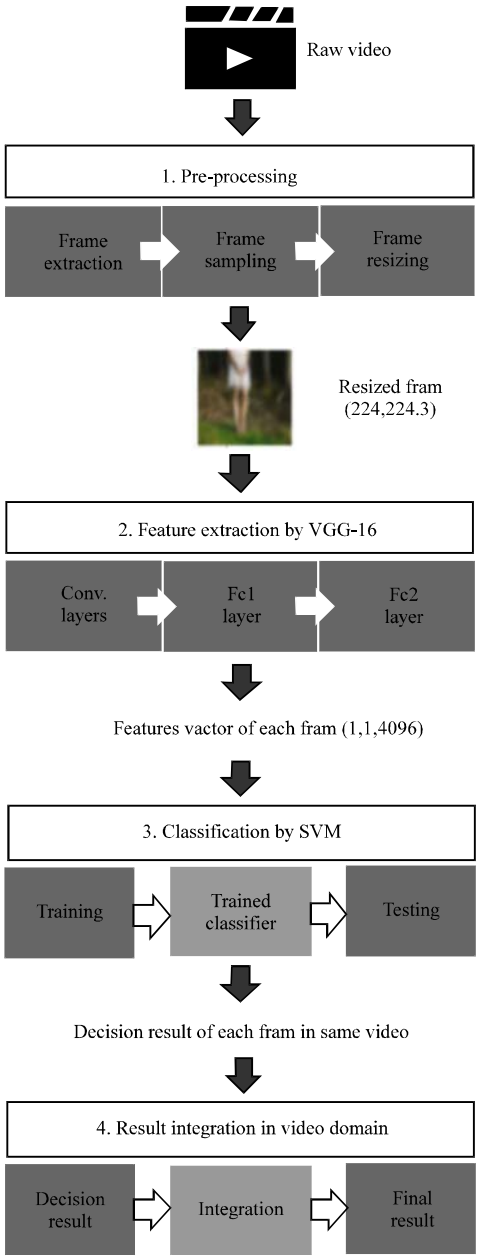

Fig. 1: Process for image descriptor based pornography detector

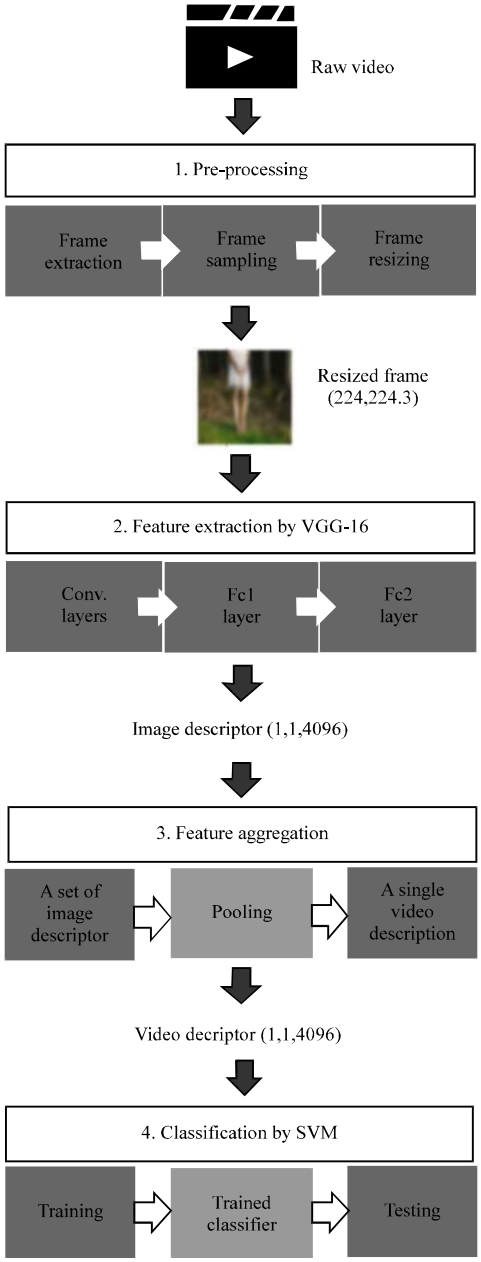

Fig. 2: Process for video descriptor based pornography detector

generation for image descriptor. However, the following phases are differ from the previous ones because the image descriptors extracted from the input video are aggregated into one video descriptor which is used to train classifier in the third step. To make single video descriptor by aggregating the image descriptors, we used the average pooling which is the best performance in the preliminary experiment but other pooling methods can be used. Once the video descriptor has been created, the classifier for pornography detection is trained. Through this method, we can provide more stable performance than previous methods. As the results, we can get 96.6% as maximum accuracy by using image descriptors and max pooling integration while we get 99.3% as the maximum accuracy by using the video descriptor with max pooling scheme. However, this method also has difficulty to detect the video with certain scene such as baby feeding, sucking sticky things in naked state, masturbation or

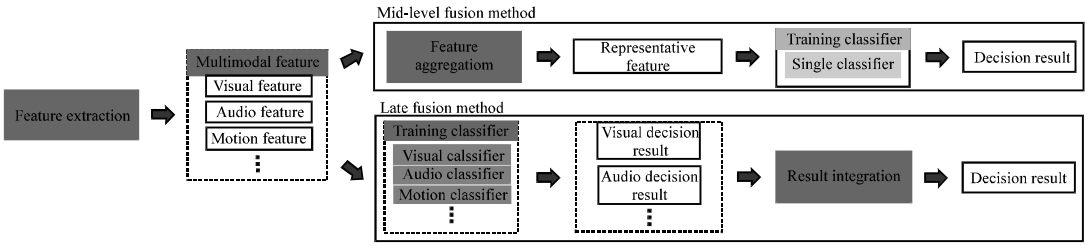

Fig. 3: The multimodel pornography detects

groaning while wearing clothes, since, these scenes are hard to classify by only using visual features extracted from videos.

Therefore, in order to overcome this drawback, the pornography detection schemes using multimodal features extracted from the input video have been studied (Liu *et al*., 2014; Perez *et al*., 2017). These multimodal pornography detectors have the common process of 3 phases as shown in Fig. 3. At first, the multimodal features which are needed for target models such as audio, image and motion, respectively are extracted from the input video to be used for porn detection. And the next following phases are determined by the fusion method. Meanwhile, the multimodal detector can be also classified into two categories depending on the features used together. The first category is based on motion and visual features. Perez *et al*. (2017) proposed a multimodal pornographic video detector using visual features and motion features. To generate motion information from input video, it uses an optical flow algorithm (Brox *et al*., 2004). Based on this motion information and video frame, this model extracted motion descriptor and video descriptor using GooglLeNet (Szegedy *et al*., 2014) as the feature extractor. And then according to late fusion method, both of classifiers by each feature are trained and integrated result is made up using average pooling. As a result, this model got 97.9% as the maximum accuracy. However, it still have difficult to detect video data that is difficult to detect with only visual features such as the video that is not significantly harmful visually but groaning sound is sexual.

On the other hand as the second category, multimodal detector using audio and visual features is studied recently. Most of studies in this category utilize the audio features as the main feature and treat visual features as secondary ones. Liu *et al*. (2014), the scheme generated the high level audio feature using Bag of Audio Word (BoAW) framework by each audio segment. The process of BoAW is similar to the process for producing BoVW. Therefore, the audio descriptors are made by each Energy Envelope Unit (EEU) in the first phase. And then, the codebook is constructed using k-means clustering in the second phase. At the same time, only one frame is sampled as the key frame which represents the overall frames belonging to the sequence of the same time zone with EEU. Based on these codebook and key frame, SVM classifier is trained using each feature fold and integrated their results using weighted function and filtering with the threshold. As a result, this model gets 94.4% as the maximum True Positive Rate (TPR) and 9.76% as the minimum False Positive Rate (FPR). Compared with the results of other studies, it is not a numerically superior result and also it has the problem by high model complexity same as detectors using BoVW but they showed good aspect that the model can detect the pornographic cases that were difficult to distinguish by other works using only visual features.

Therefore, we propose the new pornographic video detection scheme using multimodal features. The proposed scheme is based on late fusion method. To integrate the results of each classifier using the high level image, video and motion features extracted by deep learning architecture and the low level audio feature based on spectrum of sound frequencies, we use the model stacking method. Through this proposed scheme we can detect almost, every pornographic event which occurred in video even though several elements are not obscene at that moment.

MATERIALS AND METHODS

Pornographic video detection scheme using multi modal features: In this study, we propose the new pornographic video detection scheme using multimodal features which are extracted from frame image, video, motion and audio of the input video. Before explain the proposed scheme we need to define the meaning of 'pornography'. The pornography that, we will detect out is "every single scene, action or sound related to sexual behavior or circumstance". And the process outline of the proposed detection scheme is suggested in Fig. 4. The detection process consists of 2 phases: training and testing. Each phase is composed of successive 4 identical steps. As mentioned earlier because this scheme uses four elements

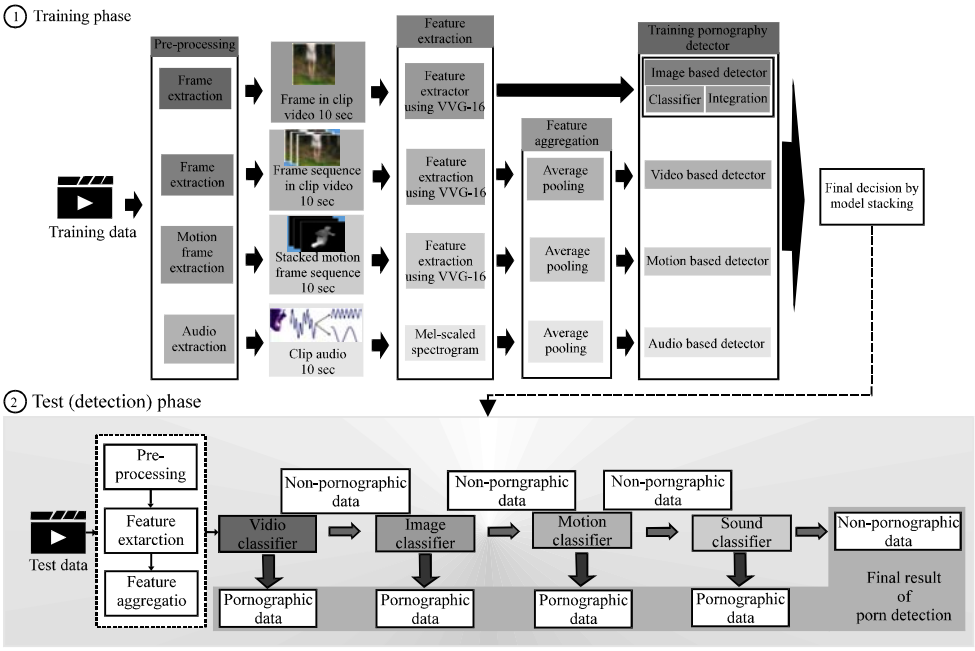

Fig. 4: The proposed defector scheme

at total, the raw video data is converted into appropriate data forms, respectively such as frame image, motion, frame sequence and audio in the first pre-processing step. In the next feature extraction step, features are extracted from the data by applying appropriate feature extraction method according to the type of each data. Each set of the extracted features is used to make detectors separately. Among them, the features extracted from audio, motion, and frame sequence require the additional process for the aggregation of video unit using pooling method in the third step. As mentioned before, these multimodal features are used to make each independent detector in the fourth step. At that time, the image-based detector takes additional sub step, result integration, to make final decision. And finally, the model stacking is used to combine the results of each detector to produce the final decision result.

From now on, further detailed description of each detector is described. However, this model uses the same elementary detectors to the previous the detector using image features and the detector using video features (Song and Kim, 2017). So, only the motion based detecto and audio based detector need to be explained. The detailed description about motion-based detector is described in Fig. 5.

The generation of the detector using motion features consists of 5 phases and also uses the pre-trained VGG-16 as the image descriptor generator as like for the detector using video descriptor features. Therefore, it has a similar process to generate the video feature based detector except for some differences.

The first difference is in the motion frame generation phase. In this phase, the 3-stacked motion frame data of each direction to be used for extracting the features about the motion change over time are generated. So, we use the optical flow algorithm (Farneback, 2003; Lu *et al*., 2016) to make flows of X and Y direction at every pixel in the frame image which used to generate grayscale image about motion of each direction and each fixed time. And in order to extract a change of motion over time as a motion feature, the motion frame of Time T are integrated with the motion frame of Time T-1 and T+1 using channel stacking. As a result, the motion frames of time T to be used for extracting the motion feature have the motion information of the time before and after the Time T.

Another difference is in the feature concatenation step which is one of the substep of feature aggregation step. Unlike the other detectors described earlier, two descriptors, descriptors of X-direction and of Y-direction, respectively are created by video unit after pooling. Therefore, we need to integrate them into a single feature. But, because they independently reflect the movement changes in X and Y directions, respectively and their origins are different, we can't use the pooling method for their integration. Thus, we use concatenation to integrate them into single feature.

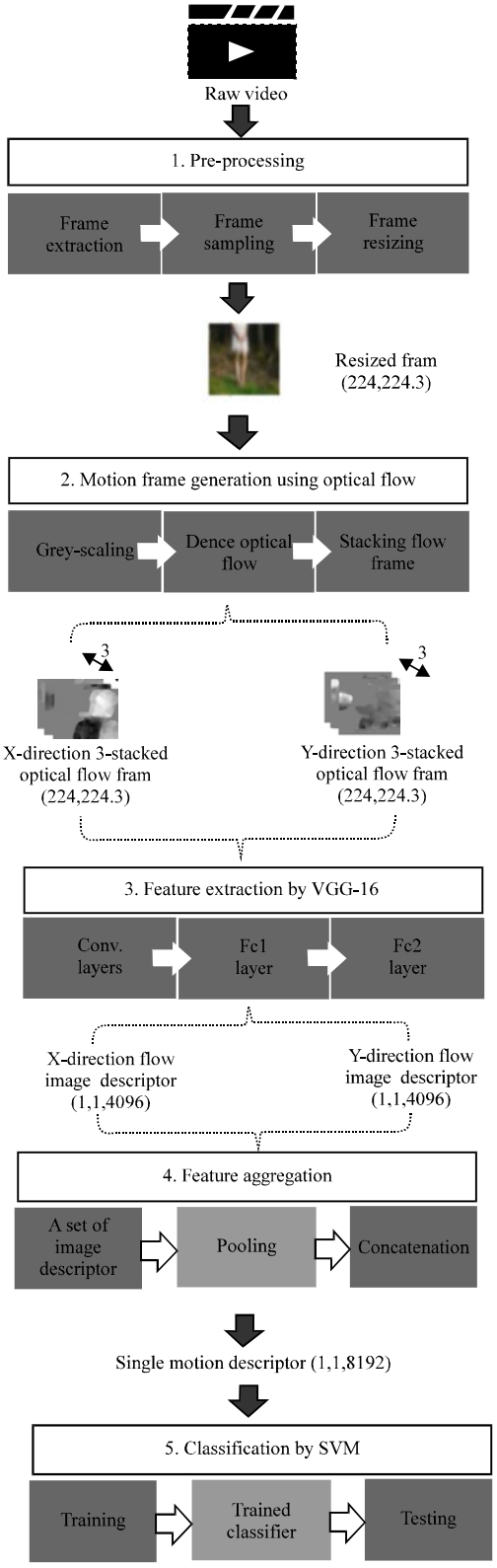

Fig. 5: Process for motion feature based pornography detector

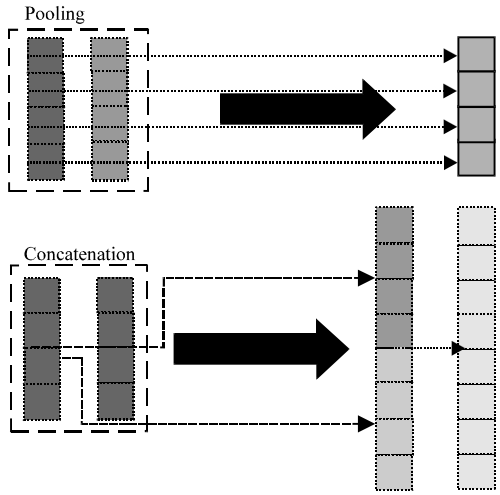

Fig. 6: Pooling and concatenation samples

Figure 6 shows the unlike pooling which collects the same dimension values of original feature vectors and creates a new vector of the same size as the original, concatenation connects the original feature vectors intact. Therefore, the concatenated vector can be integrated without losing the meaning of the value of each dimension of the original. Finally, the motion descriptor created through this processes is used in the learning process to create motion based pornography detector.

In succession, as the last individual feature based pornography detector, the detailed description about audio based detector is described in Fig. 7.

As shown in Fig. 7, the audio feature based pornography detector consists of 4 phases. In the first phase, we separate audio data from the video file and divide them by the unit of 10 sec. And next, we extract audio features of audio frame which constitute clip audio, using Mel-scaled spectrogram (Stevens *et al*., 1937), because the spectrogram is commonly used to capture timbre aspects of sound. After then, the features of each audio frame aggregated into single audio descriptor by pooling method. And based on them, the audio classifier is trained for pornography detection.

Lastly, we construct a multimodal pornography detector that detects pornographic video by combining the four independent detectors using model stacking method. Model stacking is one of the model integration methods to enhance the performances of the individual models. Since, we integrate different 4 models that learned by different types of data each other, stacking the independent models is one of the feasible methods for developing the proposed multimodal pornographic video detector. Thus, as shown in Fig. 4, we stacked each model in order of video feature based model, image feature based model, motion feature based model and audio feature based model according to their performances.

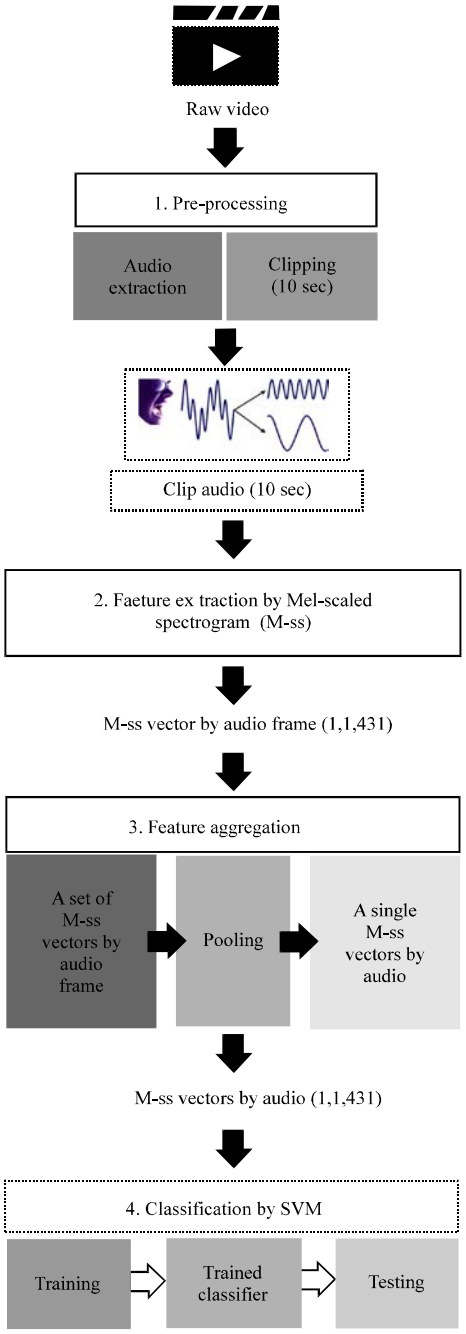

Fig. 7: Process for audio feature based pornography detector

RESULTS AND DISCUSSION

Prior to describing the experiment results, we describe the data. To make the dataset that we used for training and testing SVM we utilized 'pornography-2k' dataset (Moreira *et al*., 2016). The 'pornography-2k' consists of 1000 porn and 1000 non-porn videos in which are variously mixed from sample videos like wrestling, person on beachside, babies and so on. But because each of videos in 'pornography-2k' have unique running time, fps (frames-per-second) and frame size variously we randomly pick out some videos from them and split them into video clips of 10 seconds. As a result, we had created a video dataset with total of 2300 video clips which consists of 1100 non-pornographic ones and 1200 pornographic ones.

Table 1: Performance evaluation of two detectors using motion and audio features

Detector types	Maxi. accuracy (%)	Mini. accuracy (%)	Average accuracy (%)
Motion feature based	93.5	84.8	88.3
Audio feature based	87.0	72.0	80.0

Based on this dataset, we performed some experiments to distinguish the pornographic videos from the non-pornographic ones. The first experiments are performed to evaluate the each performance of motion feature based detector and audio feature based detector. The experiments were conducted on 10 fold cross validation using entire frame data.

As shown in Table 1, the experiment result of motion feature based detector and audio feature based detector does not show better performance compared to the performance of our previous detectors using the image features or the video features, respectively. However, these models were able to distinguish the ambiguous videos which are likely to classify hardly by using only video features such as videos of baby feeding or videos with sexual groan without obscene visual elements. Therefore, it can be seen that complementary elements exist between detectors based on visual related features and detectors based on motion or audio related features.

The second experiment is performed to evaluate the performance of multimodal feature based detector. This experiment was also conducted on 10 fold cross validation using entire frame data.

As shown in Fig. 8, the experiment results of multimodal feature based detector do not show fine performance. Because of the high false positive rate, overall performance was reduced in terms of accuracy. However, in terms of True Positive Rate (TPR), it shows the perfect performance, 100%. It means that the proposed scheme can detect every pornographic video from the test data set including ones to classify hardly.

Lastly, in order to find individual detectors that cause such high false positive rates we conducted additional experiments to examine the trend of false positives that each model creates.

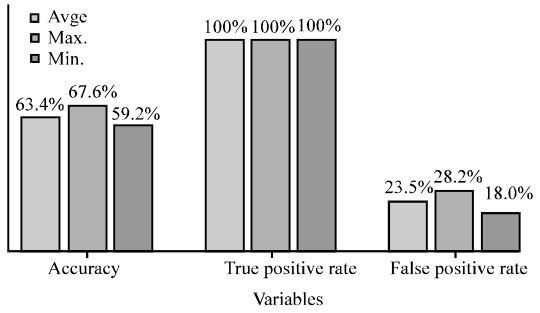

Fig. 8: The experiment results of multi model feature

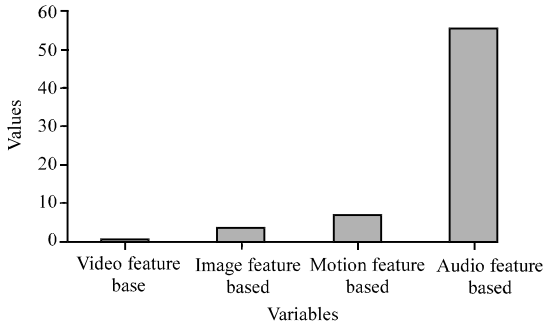

Fig. 9: The average number of false positive

As shown in Fig. 9, the most of the false positive were occurred by audio feature based detector. One of the reasons for this extreme difference is that the audio detectors are based on low level features while the others are based on high level features extracted by deep learning.

And also as the additional reason, not only the timbre element reflected by the spectrogram but also Chroma, amplitude and the other elements need to detect pornographic video accurately based on only audio origin features.

CONCLUSION

In this study, we proposed the pornographic video detection scheme using multimodal features each of which is comprehensive of image, video, motion and audio, respectively. Most of features are generated by re-trained VGG-16 except only audio feature which is generated by spectrogram. In succession, to combine the individual detection models we use model stacking method. And we progress experiment to evaluate the performance of the proposed scheme. As a result, we can obtain 100% true positive rate and 67.6% average accuracy. Although, the overall accuracy is low due to high false positive rate we could successfully detect the pornographic video including videos which are difficult to detect.

RECOMMENDATIONS

In the future, we try to reinforce the audio feature based pornography detector through applying the high level audio feature extraction method to compensate the shortcomings of proposed scheme and make more accurate decision making.

ACKNOWLEDGEMENT

This research was supported by Institute for Information and Communications Technology Promotion (IITP) grant funded by the Korea Government (MSIT) (2017-0-00318, Development of Framework to Create and Control Trusted Media based on Social IoM)

REFERENCES

Adnan, A. and M. Nawaz, 2016. RGB and Hue Color in Pornography Detection. In: Information Technology: New Generations, Latifi, S. (Ed.). Springer, Switzerland, ISBN:978-3-319-32466-1, pp: 1041-1050.

Anonymous, 2017. [The Korea Communications Commission (KCC) will block the distribution of illegal harmful information]. Chosun Media Co., Ltd., Seoul, South Korea. (In Korean) http://it.chosun.com/news/article.html?no=2843689.

Brox, T., A. Bruhn, N. Papenberg and J. Weickert, 2004. High Accuracy Optical Flow Estimation based on a Theory for Warping. In: Computer Vision, Pajdla, T. and J. Matas (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-21981-1, pp: 25-36.

Farneback, G., 2003. Two-Frame Motion Estimation Based on Polynomial Expansion. In: Image Analysis, Bigun, J. and T. Gustavsson (Eds.). Springer, Berlin, Germany, ISBN:978-3-540-40601-3, pp: 363.

Krizhevsky, A., I. Sutskever and G.E. Hinton, 2012. Imagenet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems, Leen, T.K, G.D. Thomas and T. Volker (Eds.). MIT Press, Cambridge, Massachusetts, ISBN: 978-0-262-12241-3, pp: 1097-1105.

Liu, Y., Y. Yang, H. Xie and S. Tang, 2014. Fusing audio vocabulary with visual features for pornographic video detection. Future Gen. Comput. Syst., 31: 69-76.

Lu, T., K. Sung and B. Jiang, 2016. Markerless motion capture for entrance guard systems. Intl. J. Technol. Eng. Stud., 2: 172-179.

Moreira, D., S. Avila, M. Perez, D. Moraes and V. Testoni *et al*., 2016. Pornography classification: The hidden clues in video space-time. Forensic Sci. Intl., 268: 46-61.

Moustafa, M., 2015. Applying deep learning to classify pornographic images and videos. Comput. Vision Pattern Recognit., 1: 1-10.

Perez, M., S. Avila, D. Moreira, D. Moraes and V. Testoni *et al*., 2017. Video pornography detection through deep learning techniques and motion information. Neurocomput., 230: 279-293.

Punaiyah, K. and H. Singh, 2017. Biped robot for walking and turning motion using raspberry Pi and Arduino. Biped robot for walking and turning motion using raspberry Pi and Arduino. 3: 49-58.

Simonyan, K. and A. Zisserman, 2014. Very deep convolutional networks for large-scale image recognition. Master Thesis, Cornell University, Ithaca, New York.

Song, K. and Y.S. Kim, 2017. Pornographic video detection scheme using video descriptor based on deep learning architecture. Proceeding of 4th International Conference on Emerging Trends in Academic Research (ETAR'17), September 14-15, 2017, Golden Tulip Galaxy Banjarmasin, Banjarmasin, Indonesia, pp: 59-65.

Stevens, S.S., J. Volkmann and E.B. Newman, 1937. A scale for the measurement of the psychological magnitude pitch. J. Acoust. Soc. Am., 8: 185-190.

Szegedy, C., W. Liu, Y. Jia, P. Sermanet and S. Reed *et al*., 2014. Going deeper with convolutions. Comput. Vision Pattern Recognit., 1: 1-12.

Tablatin, C.L.S., F.F. Patacsil and P.V. Cenas, 2016. Design and development of an information technology fundamentals multimedia courseware for dynamic learning environment. J. Adv. Technol. Eng. Stud., 2: 202-210.