applied sciences

MDPI

Article

An Enhanced Multimodal Stacking Scheme for Online Pornographic Content Detection

Kwangho Song and Yoo-Sung Kim *

Department of Information and Communication Engineering, Inha University, Incheon 22212, Korea; 22171071@inha.edu

* Correspondence: yskim@inha.ac.kr; Tel.: +82-32-860-7450

Received: 27 February 2020; Accepted: 21 April 2020; Published: 24 April 2020

Abstract: An enhanced multimodal stacking scheme is proposed for quick and accurate online detection of harmful pornographic contents on the Internet. To accurately detect harmful contents, the implicative visual features (auditory features) are extracted using a bi-directional RNN (recurrent neural network) with VGG-16 (a multilayered dilated convolutional network) to implicitly express the signal change patterns over time within each input. Using only the implicative visual and auditory features, a video classifier and an audio classifier are trained, respectively. By using both features together, one fusion classifier is also trained. Then, these three component classifiers are stacked in the enhanced ensemble scheme to reduce the false negative errors in a serial order of the fusion classifier, video classifier, and audio classifier for a quick online detection. The proposed multimodal stacking scheme yields an improved true positive rate of 95.40% and a false negative rate of 4.60%, which are superior values to previous studies. In addition, the proposed stacking scheme can accurately detect harmful contents up to 74.58% and an average rate of 62.16% faster than the previous stacking scheme. Therefore, the proposed enhanced multimodal stacking scheme can be used to quickly and accurately filter out harmful contents in the online environments.

Keywords: online pornography detection; multimodal stacking ensemble; quick detection; low false negative rate; implicative features

1. Introduction

With the recent increase in the number of online distributions of harmful pornographic contents via new types of personal broadcasting services, a system for automatic detection of online pornographic contents is highly being demanded [1,2]. Since much of the recent pornographic contents provided through the online personal broadcasting services have harmful scenes either in the form of visual or auditory manner, fast and accurate detection seems vital. However, most previous studies related to the automatic detection of pornographic contents have mainly focused on single modal detection that extracts and uses either visual or auditory features [3–14]. One of the limitations of existing methods based on single modal detection approach is that the harmful contents without the detectable elements cannot be detected.

In order to reduce these detection errors, several multimodal detection methods have been studied using visual and/or auditory features for detection [15,16]. Previous studies on the multimodal methods have shown better detection performance than previous single modal methods, even when portions of the harmful visual or auditory elements are absent in the content. However, since the methods proposed in [15,16] determine the harmfulness only after the harmful contents have completely played, it is difficult to determine the harmfulness of such contents during the early phase of the play if the contents are provided via streaming way on the online media platform. Therefore, there is a need for a detection method that can quickly and accurately detect harmful contents that are played or

distributed online. In particular, a quick detection method that can minimize the detection omissions of harmful contents is needed.

Recently, to supplement the problems of the existing studies [3–16] and satisfy the requirements of the online harmful content detection, a method that utilizes multiple features extracted from the visual, motion, and auditory elements for the pornographic detections was proposed in [17]. In the detection method proposed in [17], the harmfulness is decided via unit segments that are divided as a particular unit length from the input content. To determine the harmfulness of the online content, the unit segments from an input content are used to classify the harmfulness as quickly as possible. In order to determine the harmfulness of a unit segment, an image descriptor of each video frame, a video segment descriptor of a continuous video frame sequence, a motion descriptor to notate the motion characteristics in the video segment, and an audio descriptor of the unit segment are utilized. These four types of descriptors are extracted from all content segments in the training dataset, and then the independent four component classifiers are developed using each type of the descriptors. This study [17] used the multimodal pornographic detection method with a stacking ensemble approach, where four component classifiers are arranged in a descending order of their performance to improve the recognition performance, especially in terms of the false negative rate, as well as robustness against lack of the visual, motion, or auditory elements in the input content.

However, although the visual and auditory elements of the content have characteristics that change over time, because each descriptor used in [17] uses simple static features by averaging the changes in each unit content segment over time, an accurate reflection of these changes over time can be limited. In addition, since three different types of visual features (image descriptor, video descriptor, and motion descriptor) are extracted from the visual elements of the input content, this method leads to a waste of computation time by engaging in three feature extractions independently. Since the extraction time for the motion features from an input segment via the optical flow is quite long in [17], fast detection tends to be problematic in the online streaming service environments. In addition, when the performance of a specific component classifier is relatively poor, the result of the final decision is dominantly influenced since many false positives and reduced overall accuracy occur due to the unique characteristics of the model stacking method. In particular, since the detection accuracy of the audio component classifier is relatively lower than that of the other component classifiers, the resulting overall detection performance can be lower than expected [17].

In order to solve these weaknesses, we propose an enhanced multimodal stacking scheme that can quickly and accurately detect online harmful contents on new types of personal broadcasting services. In the proposed harmful content detection technique, instead of extracting three descriptors from the visual elements, VGG-16 [18] and bi-directional RNN using LSTM (long short-term memory) [19] are used to extract a single implicative visual descriptor that reflects changes over time [20]. This method improves both the accuracy of the hazard determination and the computational time required for the feature extractions and hazard decisions. In addition, to extract the characteristic that reflects the bidirectional correlations of the neighboring auditory signals, a multilayered dilated convolutional block [21–23] is used to extract and utilize the implicative auditory descriptor to improve the accuracy of the audio component classifier more than that of [17]. As the first detection step, the harmfulness is decided via the fusion component classifier trained by both the visual and auditory features to detect the hazardous contents with high accuracy over a short period of time. When the input content is classified as non-harmful, the content is checked serially with the video component classifier and audio component classifier, each of which is trained based on the visual and auditory features, respectively. By using the proposed multimodal stacking scheme, the harmfulness of the input content can be detected quickly using the first fusion classifier, and any hazardous content missed in the first filtering stage can be detected later by the video classifier or the audio classifier in a serial order.

This paper is composed of the following sections. In Section 2, the previous related studies are reviewed. In Section 3, we describe the proposed multimodal stacking scheme about its overall procedure, extracting of the implicative visual and auditory features, developing of the component

classifiers, and stacking ensemble of the component classifiers. In Section 4, the experiments and the analysis results are described. A short discussion is described in Section 5. Finally, in Section 6, the conclusions of the study and future studies are described.

2. Literature Review

Existing multimodal pornographic detection schemes generally use two or more features among the visual features extracted from either a single video frame or a video segment (i.e., a continuous sequence of video frames), the motion-based features, and the acoustic features extracted from the content. Despite the differences in the features used or the combination methods, most multimodal pornography detection methods involve three common steps. The first step extracts the features that each model seeks to use from the corresponding elements of the input content. In the past, low-level features such as skin color, specific female body areas, or distribution of skin pixels were utilized. However, as described in [24], these low-level features do not have sufficient discriminative power to judge the harmful status of the content. Recently, low-level features extracted from the visual and auditory elements have been converted to high-level features by applying the Bag of Word (BoW) or deep learning frameworks [15,16]. The second step creates an overall classification model via training with the selected appropriate machine learning scheme using the extracted multimodal features to recognize pornographic contents. In general, in the mid-level fusion method approach, all multimodal features can be combined a priori into one representative integrated feature set to develop one classifier, whereas in the late fusion method approach, several component classifiers are made using individualized multimodal features. The late fusion method is the most common method in recent studies [15,16] because of its superior performance, as described in [15]. In a previous study [25], simple methods with a pre-determined threshold or simple machine learning model such as a decision tree and a naïve Bayes were used. In the recent studies of [15,16], which mainly used the high-level features, the support vector machine (SVM), neural network, and deep learning architectures (which can clearly reflect the non-linearity of the classification hyperspace) were utilized. The final stage is the output engineering step, where all classification results from the component classifiers are integrated for the late fusion method approach.

However, in the case of [15], since all of the features were extracted from the visual elements, harmful content detection remained difficult if only the auditory elements of the input content were harmful as existing single modal methods use only the visual elements. In addition, in the case of [16], although the disadvantages of [15] were compensated by using the features extracted from both the visual and auditory elements, since the features extracted from the visual elements are comprised of static features extracted from one still image, the method could not detect the harmful contents well. Moreover, since the methods in [15,16] require the entire piece of content to determine the harmfulness, it is difficult to quickly detect the harmful contents that are played or distributed online. There is also a problem of omitting certain types of harmful contents in the detection process.

In order to resolve the disadvantages of previous studies, the authors in [17] proposed a pornographic video detection method that offers a robust detection performance even if some of the elements used for detection are insufficient. The detection process in [17] is also composed of three steps. In the feature extraction step, four types of descriptors are extracted, including an image descriptor that contains the static features of a video frame, a video segment descriptor containing the static features of a video segment, a motion descriptor representing the motion features in a video segment, and an audio descriptor that contains the static features of an audio segment. These descriptors are extracted by dividing the input content into 10-s unit content segments. Each segment is judged for its harmfulness instead of using the entire piece of the content for early detection. The four descriptors extracted through this process are used to train each component classifier via linear SVM. Each component classifier produces a probability value as the decision result for the pornographic status of the input content. Lastly, in order to combine all probability values to make the final decision, one of the model ensemble techniques, the model stacking method, is utilized to improve the final decision accuracy,

especially to improve the true positive rate. In [17], in order to ensure that pornographic videos are found as early as possible, the component classifiers are stacked in descending order according to the accuracy of each classifier—for example, in the respectable order of the video classifier, the image classifier, the motion classifier, and the audio classifier. This method not only provides better performance in detecting typical pornographic scenes with abundant harmful audiovisual elements but also provides good detection performance for scenes lacking some of the necessary elements for reliable pornographic detection.

However, the method in [17] has the following three limitations. First, the video, motion, and audio descriptors used to express the characteristics of the visual or audio elements of the content segment are composed of the averaged static feature values extracted from each segment point. Therefore, the descriptors created through this method cannot reflect the temporal relevance between each signal of the segment. Consequently, the performance of the corresponding classifiers may be degraded since the static properties may not sufficiently reflect the changes over time. Second, although the model stacking technique can increase the true positive rate of the final decision results, the overall classification accuracy decrease because the false positive error rate increases when the performance of some classification modules is poor. Third, the processing time is wasteful since the image descriptor, the video descriptor, and the motion descriptor are extracted from the visual elements, and then the three component classifiers are individually trained and used to decide the harmful status of the content based on such features. Because a great amount of time is required to extract the features and detect the harmfulness, this method could be insufficient in properly detecting harmful contents in the online service environments.

3. Methods

3.1. Overall Procedure of the Enhanced Multimodal Stacking Scheme

In this section, we introduce a newly proposed multimodal stacking ensemble scheme for a fast and accurate online pornographic content detection, which stacks three component classifiers, each of which is trained using only their visual features, auditory features, or both features together, respectively. The proposed multimodal stacking scheme is generated through four steps: pre-processing, feature extraction, training the component classifier, and the multimodal stacking ensemble, as shown in Figure 1.

In the first step, the pre-processing step, the input content is divided into a set of unit segments to quickly decide the harmfulness of the content; each segment is then separated into a video segment and an audio segment. Next, a sample frame for each 1 s of the playing time is randomly sampled in a video segment to reduce the redundant processing time for the continuous changeless video frames. The implicative visual and auditory features are then extracted in the second step, the feature extraction step. The features are called implicative because they implicitly include the correlative relationships with the neighboring signals that change over time. In the third step, the component classifier training step, each video and audio component classifier is trained using only the implicative visual and auditory descriptors, respectively. A fusion component classifier is also trained using both implicative descriptors together. Lastly, in order to ensure that the harmfulness of the online content is quickly and accurately determined, the harmfulness is first investigated using a fusion classifier that considers both the visual and auditory factors of the input content. A video classifier trained only with the visual features is used to judge the content carefully based on its visual factors against the content previously classified as non-hazardous by the first fusion classifier. Then, an audio classifier trained only with the auditory features is used to carefully examine the missing content via previous two classifiers. Thus, three component classifiers are stacked in a serial order of the fusion classifier, the video classifier, and the audio classifier in the enhanced multimodal ensemble scheme.

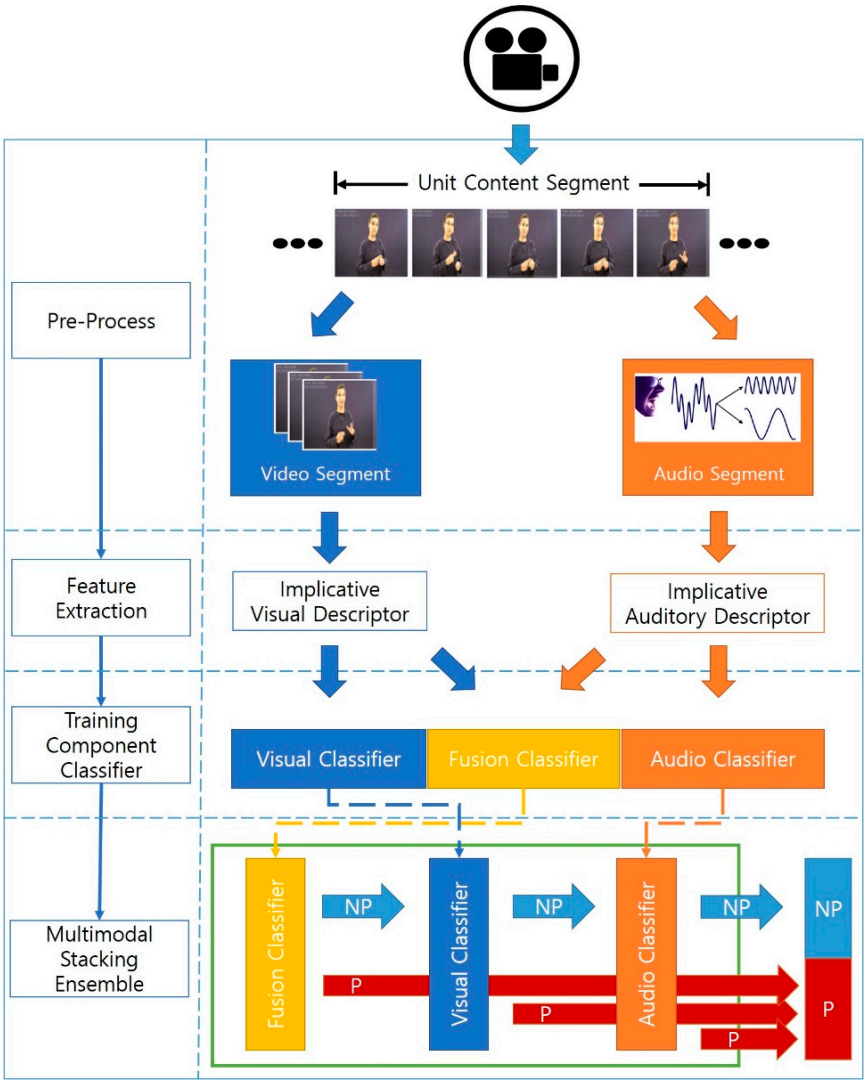

Figure 1. Overall procedure of the proposed multimodal stacking scheme.

3.2. *Extraction of Implicative Features*

In the second feature extraction step of the proposed scheme, an improved process is used to consider the changes in the input segment over time and extract the implicative features that reflect the correlative relationships with the neighboring signals, as shown in Figure 2, while the first step follows the same process as used in the previous study [17]. In order to extract an implicative visual descriptor for each video segment, the visual feature extractor, which consists of the front part of the pre-trained VGG-16 and the bi-directional RNN in [20], is utilized. In the first stage, the front portion of the pre-trained VGG-16 without the upper-level fully connected (FC) layers is used to extract highly expressive high-level features. In general, the inclusion of the upper-level FC layers during transition learning of the VGG-16 leads to a rapid increase in the number of learning parameters. To reduce the number of learning parameters, the upper-level FC layers of the VGG-16 are removed, and two 1*1 convolution layers are applied. In order for the first 1*1 convolution layer to create an output similar in size to the FC layer (which produces 4096 outputs) of the VGG-16, the output feature map of size (7, 7, 84) with (widths, heights, channels) is configured by receiving the feature map of size (7, 7, 512) from the VGG-16. The second 1*1 convolution layer also configures an output feature map of size (7, 7, 21) to create an output similar to the output by the last FC layer of the VGG-16 with a size of 1024. The output vector becomes 1029 in size after the flattening operation. A frame image of (224, 224, 3) is

then entered as the input for the VGG-16 in the feature extractor. Since a vector of 1029 is extracted as a static feature, a sequence of 10 static features is extracted for a single video segment composed of 10 sampled frames. Next, in the second part of the visual feature extractor, instead of using the simple average vector of the static features as in [17], a bi-directional RNN using LSTM [26,27] is used to singularly express the implicative visual feature vector that expressively implicates the bi-directional correlations between the static features that change over time. The Bi-directional RNN extracts the mutual correlative relationships of the forward and reverse directions from a sequential input by using the LSTM cells of each direction. These extractions are combined into a vector and input into the FC layers using the concatenation operation, thereby creating an implicative visual descriptor of size 1024 that infers the visual features of the content segment from the sequence of the static feature vectors.

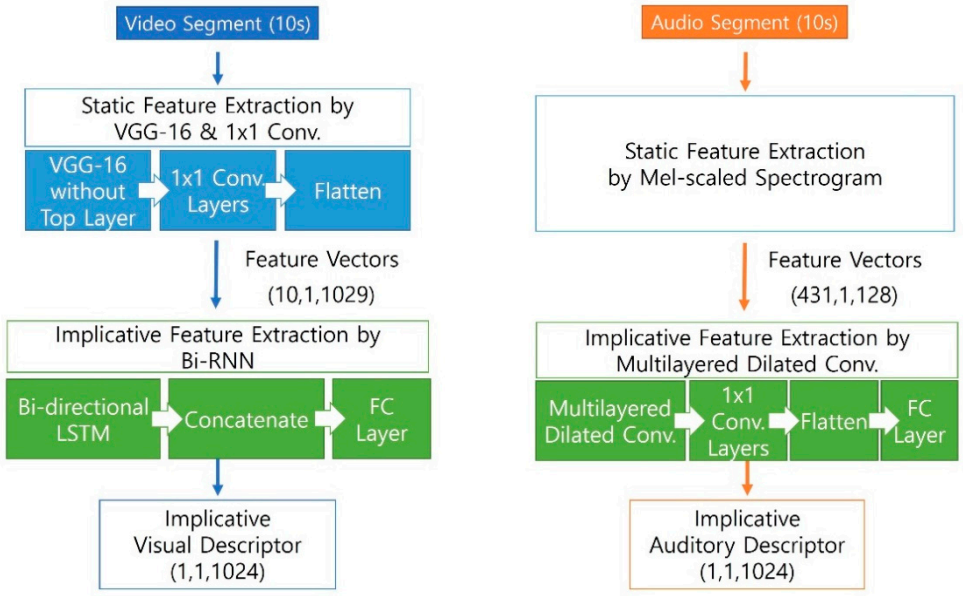

Figure 2. Implicative feature extraction procedure from the video and audio segments.

Moreover, unlike [17], which extracts and uses the static features from the auditory elements of the input content, the auditory feature extractor for extracting an implicative auditory descriptor for each audio segment is designed to use a multilayered dilated convolutional block as in [23]. The auditory feature extractor is composed of two parts. In the first part, a transformation technique that uses a mel-scaled spectrogram [28] is used to extract the static features from an audio segment. When an audio segment is entered as the input, a spectrogram of (431, 1, 128) is extracted as a static feature. The spectrogram used in [28] segments the 0 Hz~8192 Hz band frequency values using 128 filter banks for the unit length determined by the sampling rate of the audio to express the representative energy values. In case the sampling rate of the audio segment is 44 KHz, 431 samples of 128-dimensional static features are extracted per segment. For the second stage, the residual dilated convolution blocks are piled hierarchically in a multi-layered structure, and the output of each layer is connected via a skip connection to implicate the mutual bi-directional correlations between the auditory signals for each point expressed as a spectrogram to represent the signals as a single implicative feature vector. To avoid the problem of a signal from a certain point favorably reflecting the result, dilated rate of the kernel utilized in the calculation of the dilated convolution of each layer of the multilayered dilated convolutional block is extended by three-folds in the manner of 1, 3, 9, and so on. Then, the extraction of the implicative feature map is expedited by the presence of a residual path at each block. In order to convert the feature map into a vector-type implicative auditory descriptor that is similar in size to the implicative visual descriptor, two 1*1 convolution layers are first used to reduce the size. Then, the flattening operation shapes them into a vector, and one FC layer is used with an output sized 1024,

as done in the extraction process step of a visual descriptor. Through this process, the static audio feature vectors are converted into a single 1024-sized implicative auditory descriptor.

3.3. Training Component Classifiers

Each descriptor extracted through the above-mentioned procedure is utilized to train a visual classifier and an audio classifier, respectively, during the component classifier training step. In addition, a fusion classifier that determines the harmfulness of the content with two descriptors together is also trained and used as the first component classifier in the ensemble scheme to quickly and accurately detect the pornographic content with harmful features in its visual or auditory elements. To adequately reflect the non-linearity of the high-dimensional classification hyperspace consisting of high-dimensional descriptor vectors, as a comparatively simple and non-linear learning scheme, a multi-layer perception model consisting of an FC layer of size 1024 with the ReLU activation function and another FC layer of size 2 with the softmax activation function are used (as shown in Figure 3) instead of the linear SVM used in [17]. However, for the fusion classifier, the concatenate operation that connects the two descriptors before the first FC layer is executed to produce a 1024 dimensional output after receiving a 2048 dimensional input by the first FC layer of the fusion classifier marked as a double square (as in Figure 3), unlike for the other component classifiers. In all classifiers, the FC layers (of size 2) then receive 1024 inputs and produce the final 2-dimensional results of either a hazard or a non-hazard.

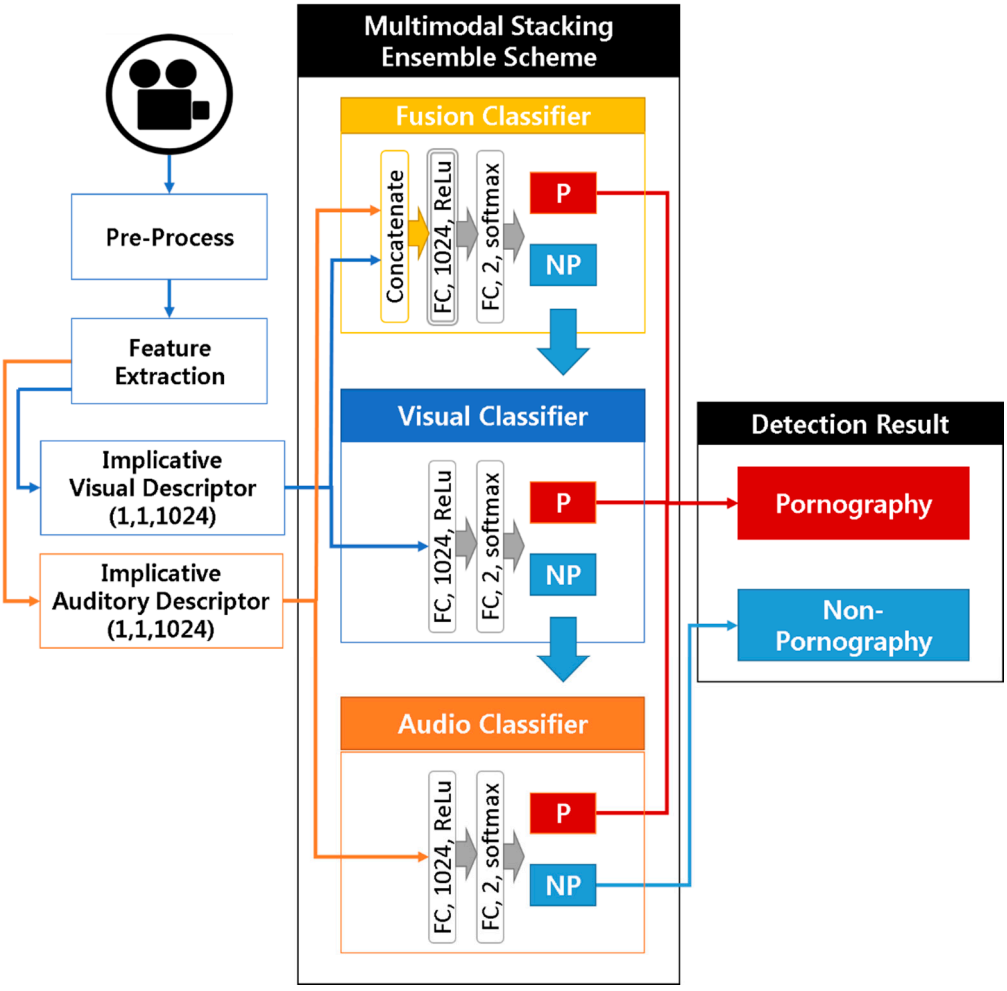

Figure 3. Pornographic detection with the proposed multimodal stacking ensemble scheme.

3.4. Ensembling Component Classifiers

As suggested in Figure 3, the enhanced multimodal ensemble scheme stacks three different component classifiers to make the final decision in order to detect the pornographic content correctly and quickly despite the lack of partial elements in the content's visual and auditory elements and to simultaneously facilitate the maintenance of the low false negative rates. The stacking ensemble arranges the component classifiers sequentially so that the component classifier that is later in the sequence can detect obscene contents not previously detected by previous component classifiers. In other words, if the initial component classifiers determine the input content to be pornographic, then the content is classified as pornographic without further assessments using the subsequent component classifiers. Otherwise, the input content that is classified as non-pornographic by the previous classifiers is checked securely by the upcoming component classifiers to minimize the detection omissions in the final decision result. Therefore, improved performance can be achieved in the true positive rate than when each model is used separately, and the number of false negative errors can be reduced at the same time.

3.5. Implementation and Optimizations

The library of Keras [29], which has been widely used in the field of deep learning, is used to implement our deep learning networks, and the optimizer Adam [30] is used as an optimization tool for learning in this study. In addition, the EarlyStopping callback provided by Keras is used to prevent excessive epochs in training to reduce the over-fitting of the model, and, to optimize training, the learning rate is set be reduced by 1/10 if the reduction of the validation loss does not occur for 5 consecutive epochs through the ReduceLROnPlateau callback.

4. Experiment Results

4.1. Performance Evaluations

This section describes the experiments performed to evaluate the performance of the enhanced multimodal stacking scheme proposed in this paper. First, the "Pornography-2k" [4] dataset, which is widely used to develop harmful content detection techniques, is used for training and testing purposes. The "Pornography-2k" dataset includes 1000 pornographic videos and 1000 non-pornographic videos. Non-pornographic videos include easily distinguishable features from the pornographic videos, such as cartoons, natural scenery, and street scenes, as well as more difficult-to-distinguish contents, such as wresting, people in swimwear, and breast-feeding mothers. Although a majority of the harmful videos are composed of both visual and audio elements that are harmful, some harmful videos are composed of only the visual elements with muted auditory elements or unrelated background music. In order to ensure the detection of additional harmful contents, the harmful auditory contents that are problematic in the recent online environments were additionally collected from the Internet and utilized for this study.

It is important to ensure the objectivity and fairness of labels for contents since the trained model is greatly influenced by the label attached to each datum. Therefore, we sought to ensure the objectivity and impartiality of the data labels by using three human annotators to judge the pornographic nature of each datum independently. After this independent judgment, the results were integrated through a majority vote to ensure fair labelling. Through this process, a total of 8000 cases of segments, each containing 4000 hazardous and non-hazardous content segments, were randomly selected from the labeled content dataset. The hazardous contents in the selected data set are composed of 2106 segments having harmful factors in both the visual and auditory elements, 1114 in the visual element only, and 780 in the auditory element only. Among the selected 8000 segments, 5000 stratified cases were selected for learning. The performance evaluation experiments are then conducted with the remaining 3000 segments consisting of 1500 hazardous and 1500 non-hazardous cases. Since the labels of the harmfulness of each content segment may be different in terms of the visual and auditory elements,

the harmfulness of the corresponding element was used to select the data used for the training of each component classifier.

4.2. Analysis Results for Video Classifier

In the first experiment, as shown in Table 1, to evaluate the video component classifier of this study (which uses the implicative visual features extracted using the VGG-16 and bi-directional RNN), we compare our newly proposed video classifier with the stacking method using three visual classifiers (video, image, and motion) each of which is independently trained by the three visual features extracted through simple average pooling that are stacked in the same way as [17].

Table 1. Detection accuracy comparison based on visual features.

Visual Classifier with Different Visual Feature Extractions	Accuracy
Visual classifier using VGG-16 and Bi-direction RNN	95.33%
Stacking ensemble with only the three visual classifiers used in [17]	94.63%

As a result, the method that extracts the implicative visual features using the VGG-16 and bi-direction RNN in this study showed superior performance to the method used in [17]. Therefore, extracting the implicative visual features using the VGG-16 and bi-direction RNN is better than using a method that integrates the visual elements from a series of frames through simple average pooling. In addition, extracting only one implicative visual feature can save more computational time than extracting the three visual features (the image descriptor, the video descriptor, and the motion descriptor) separately.

4.3. Analysis Results for Audio Classifier

In the second experiment, the methods to extract the implicative auditory features from an audio segment are compared, as shown in Table 2. That is, we compare our method that extracts the implicative features and connotes the bi-directional correlative relationships between the auditory signals using a multilayered dilated convolutional block and the method that uses the static features with a mel-scaled spectrogram, as employed in [17]. As a result, the method that extracts the implicative auditory features through a multilayered residual dilated convolutional block showed better performance in the harmful content detection than the method that uses the static features, as shown in [17]. Thus, the use of the implicative auditory features that implicitly express the correlation between neighboring signals over time is evaluated to be more accurate in detecting harmful contents than utilizing the static features, such as when determining the harmfulness based on the auditory elements of the content.

Table 2. Detection accuracy comparison based on auditory features.

Audio Classifier with Different Auditory Feature Extractions	Accuracy
Audio classifier using a multilayered dilated convolution network	89.16%
Audio classifier in [17]	61.83%

4.4. Analysis Results for Multimodal Ensemble

In the third experiment, the accuracy and the false negative rates of the different multimodal ensemble schemes for the pornographic content detection are compared, as shown in Table 3. The enhanced stacking scheme proposed in this study, the previous multimodal stacking scheme used in [17], and (as a representative ensemble method) the majority-voting ensemble scheme are compared. The enhanced multimodal stacking scheme suggested in this study is shown to outperform both of the other ensemble schemes in terms of its accuracy and false negative rate.

Table 3. Detection performance comparison of multimodal ensemble schemes.

Multimodal Ensemble Scheme	Accuracy	False Negative Rate
Enhanced multimodal stacking	92.33%	4.60%
Previous multimodal stacking used in [17]	88.17%	5.67%
Majority-voting ensemble	84.30%	27.33%

As described in the introduction section, in order to use the scheme in the online service environments, it is necessary to minimize the number of detection omissions that cannot be detected until the final detection stage. The number of harmful contents detected (true positives) and the number of harmful contents that could not be detected (false negatives) by both the enhanced multimodal stacking scheme and the previous scheme used in [17] are analyzed, as shown in Table 4. Using the proposed enhanced multimodal stacking scheme showed that, among the 1500 total harmful contents in the testing data set, 1398 harmful ones (93.20% of the input) were detected by the first fusion classifier. Out of 102 harmful contents that were not detected by the first classifier, 29 (28.43% of the input) were detected by the second video classifier. Among the 73 undetected by both the first and second classifiers, 4 (5.47% of the input) were detected during the third detection stage. In summary, a total of 1431 (95.40%) of the 1500 harmful contents were detected, and 69 (4.60%) were ultimately not detected. Meanwhile, as the results of the experiment with the previous stacking scheme in [17] used the same data, the first stage video classifier detected 1162 (77.47%) harmful contents, the second stage image classifier detected 97 (28.70%) out of the remaining 338 harmful contents, the motion classifier detected 85 (35.3%) of the 241 remaining harmful contents during the third stage, and the audio classifier detected 71 (45.5%) of the 156 harmful contents in the fourth stage. In summary, 1415 (94.3%) out of the 1500 harmful contents were successfully detected, and 85 (5.67%) harmful contents failed to be detected by the previous multimodal stacking method. Since the number of harmful contents that cannot be detected in the online service environments must be minimized, the enhanced multimodal stacking scheme proposed in this paper produces superior performance to the previous multimodal stacking scheme used in [17], with better accuracy, fewer false negatives, and fewer necessary component classifiers.

Table 4. Comparison of detection effects of the component classifiers.

	True Positives					False Negatives
	1st Classifier	2nd Classifier	3rd Classifier	4th Classifier	Sum	
Enhanced stacking	Fusion classifier	Video classifier	Audio classifier	None	1431/1500 (95.40%)	69/1500 (4.60%)
	1398/1500 (93.20%)	29/102 (28.43%)	4/73 (5.47%)			
Previous stacking [17]	Video classifier	Image classifier	Motion classifier	Audio classifier	1415/1500 (94.33%)	85/1500 (5.67%)
	1162/1500 (77.47%)	97/338 (28.70%)	85/241 (35.27%)	71/156 (45.51%)		

4.5. Analysis Results for Detection Time

In addition, since it is important to rapidly detect harmful contents in a short time period for practical application in the online streaming environments, we compare the detection time and average content duration time of the component classifiers from the enhanced multimodal stacking scheme with those of the previous scheme used in [17], as shown in Table 5. In the proposed enhanced multimodal stacking scheme, the processing time of the first classifier, which is the fusion classifier, takes 0.13 s, the second video classifier takes 0.13 s (with a cumulative time of 0.26 s), and, finally, the audio classifier takes 0.04 s. A cumulative total time of 0.30 s is, therefore, required. However, for the method in [17],

the processing time of the first classifier, which is the video classifier, takes 0.10 s, the second image classifier takes 0.55 s (for a cumulative time of 0.65 s), and the third motion classifier takes 0.50 s (for a cumulative time of 1.15 s). Finally, the processing time for the audio classifier takes 0.03 s, yielding a total detection time of 1.18 s. Here, since the image classifier first evaluates the harmfulness of each of the ten sampled frame images that comprise the segment and then determines the harmfulness by averaging the ten results, it requires more computation time. A long detection time duration is needed for the motion classifier because a great deal of preprocessing time is spent for extracting the horizontal and vertical movements of the video frames. If the weighted average processing time required for each content is calculated in conjunction with the number of harmful content detected by each component classifier shown in Table 4, the results are as shown in the rightmost column of Table 5. Since the detection time must be minimized in the online streaming environments, the enhanced multimodal stacking scheme proposed in this paper is comparatively superior in its performance because it detects harmful contents up to 0.88 s faster than the previous multimodal stacking scheme used in [17], thereby improving the detection time by up to 74.58%, and shortening the average detection time by 0.23 s at a rate of 62.16%.

Table 5. Comparison of the detection time (accumulated time) of the component classifiers (sec/content).

	1st Classifier	2nd Classifier	3rd Classifier	4th Classifier	Weighted Average Detection Time
Enhanced stacking	0.13	0.13 (0.26)	0.04 (0.30)	None	0.14
Previous stacking [17]	0.10	0.55 (0.65)	0.50 (1.15)	0.03 (1.18)	0.37

5. Discussion

The main objective of this study is to develop an enhanced multimodal stacking scheme that can be used in real-time streaming environments by reducing the computation time for feature extractions and judging the harmful status of the input content. To accurately detect the harmful contents, the implicative visual and auditory features are extracted by a bi-directional RNN with VGG-16 and by a multilayered dilated convolutional network, respectively. Moreover, three component classifiers are trained, respectively, by using only the implicative visual features (for video classifier), only the implicative auditory features (for audio classifier), and by using both features together (for fusion classifier). Here, to reduce the detection time, we decreased the number of component classifiers to be stacked in the ensemble scheme to three from four as in the previous scheme proposed in [17]. Then, these three component classifiers are stacked in the enhanced ensemble scheme to reduce the false negative errors in a serial order of the fusion classifier, video classifier, and audio classifier for quick online detections. According to the analysis of the experimental results, the performance rates of the proposed scheme are 95.40%, 92.33%, and 4.60% for the true positive rate, accuracy, and false negative rate, respectively.

In recent years, many studies have reported of high performance in the harmful content detections using various deep learning approaches [6–11,13,14]. Among the studies, some use video frame image or video clips [7–11], motion analysis [6], or age prediction from facial images [14] as the visual element of input content to determine the harmfulness. When comparing the performance results of these approaches, the approach of [6] with the accuracy rate of 95.1% and the approach of [7] with the true positive rates of 97.52% are showed better performance than the enhanced multimodal stacking scheme suggested in this study. However, since the techniques used in [6–11,13,14] cannot properly detect the harmful contents based on the acoustic elements, the proposed scheme in this paper, which includes an auditory element detection, can be evaluated as more advanced.

In addition, we investigated as many previous studies as possible that utilize both the visual and auditory elements simultaneously for the detection of harmful content as in this study. We confirm that the performance of the proposed method in this study is more superior to the true positive rate of 94.44% for the current state-of-the-art technology in this field [16]. However, because it is difficult to

use the same data for performance comparisons because of the nature of the research field, it is difficult to determine relative superiority by simply comparing the numerical values published in each paper.

In order to provide a meaningful performance comparison, the enhanced stacking scheme proposed in this study is compared to the multimodal stacking scheme of the previous study [17] conducted by our team using the same test data set. According to the experiment analysis results, the performances of the enhanced multimodal stacking scheme are analyzed to have the improved true positive rate of 95.40% and the false negative rate of 4.60% than 94.33% and 5.67% of the previous study, respectively. In addition, it is analyzed that the proposed scheme can detect the harmful contents up to 74.58% and an average of 62.16% faster than the previous scheme.

As to our best knowledge, this study is the first study to present the detection time required to determine harmfulness using the multimodal stacking ensemble technique to suggest the online pornographic content detection scheme for the online streaming environments. Therefore, higher accuracy and lower false negative rates with faster detection times are observed, showing this method's greater harmful content filtering performance in the online environments. However, because of the incomplete performance of the component classifiers, especially the audio classifier, and the false negative rate of 4.6% demands an improvement. Since each element classifier needs to be trained separately, a great amount of time is still required to train all the classifiers. Additional efforts are needed to develop an optimized integrated model capable of the end-to-end learning in the future.

6. Conclusions

In this paper, a multimodal stacking ensemble scheme for the online pornography content detections is proposed. In the stacking ensemble scheme, three component classifiers are trained using only the implicative visual features, implicative auditory features, and both implicative visual and implicative auditory features, arranged serially. In order to detect the harmful content quickly, the input content is divided into the unit content segments to use them as the harmful detection units. We also propose an extraction process for the implicative visual features and auditory features that express signal pattern changes over time implicatively within the input unit content segment to detect the harmful contents more accurately. The two extracted features are independently utilized to train the video classifier and the audio classifier, and then both features are used together to train the fusion classifier to use the trained classifiers as the component classifiers. In addition, we apply a stacking ensemble scheme that orderly stacks the fusion classifier, video classifier, and audio classifier for early detection and to avoid the omissions of any harmful content. According to the analysis of the experimental results, the performance rates of the proposed scheme were 95.40% and 92.33% for the true positive rate and accuracy, respectively. However, the false negative rate was about 4.60% because of the incomplete performance of the component classifiers, especially the audio classifier. Therefore, in the future, studies should focus on improving the performance of the audio component classifier.

Author Contributions: Conceptualization, K.S.; methodology, K.S.; software, K.S.; validation, K.S. and Y.-S.K.; investigation, K.S.; resources, K.S. and Y.-S.K.; writing, original draft preparation, K.S.; writing, review and editing, K.S. and Y.-S.K.; visualization, K.S.; supervision, Y.-S.K.; project administration, Y.-S.K. All authors read and agreed to the published version of the manuscript.

Funding: This research was funded by Inha University grant number 58284.

Conflicts of Interest: The authors declare no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

1. KOREA Communications Commission. They Try to Block the Distribution Source of Pornographic Content. Available online: http://it.chosun.com/news/article.html?no=2843689 (accessed on 20 December 2019).
2. Agarwal, S.; Farid, H.; Gu, Y.; He, M.; Nagano, K.; Li, H. Protecting world leaders against deep fakes. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, Long Beach, CA, USA, 15–21 June 2019; pp. 38–45.

3. Moustaf, N.M. Applying deep learning to classify pornographic images and videos. *arXiv* **2015**, arXiv:1511.08899.
4. Moreira, D.; Avila, S.; Perez, M.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; Rocha, A. Pornography Classification: The Hidden Clues in Video Space-Time. *Forensic Sci. Int.* **2016**, *268*, 46–61. [CrossRef] [PubMed]
5. Song, K.; Kim, Y. Pornographic Video Detection Scheme using Video Descriptor based on Deep Learning Architecture. In Proceedings of the 4th International Conference on Emerging Trends in Academic Research, Bali, Indonesia, 27–28 November 2017; pp. 59–65.
6. Varges, M.; Marana, A.N. Spatiotemporal CNNs for Pornography Detection in Videos. In *Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications*; Springer: Cham, Switzerland, 2019; pp. 547–555.
7. Jin, X.; Wang, Y.; Tan, X. Pornographic Image Recognition via Weighted Multiple Instance Learning. *IEEE Trans. Cybern.* **2019**, *49*, 4412–4420. [CrossRef] [PubMed]
8. Xiao, X.; Xu, Y.; Zhang, C.; Li, X.; Zhang, B.; Bian, Z. A new method for pornographic video detection with the integration of visual and textual information. In Proceedings of the IEEE 3rd Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC), Chongqing, China, 11–13 October 2019; pp. 1600–1604.
9. Farooq, M.S.; Khan, M.A.; Abbas, S.; Athar, A.; Ali, N.; Hassan, A. Skin Detection based Pornography Filtering using Adaptive Back Propagation Neural Network. In Proceedings of the 8th International Conference on Information and Communication Technologies (ICICT), Karachi, Pakistan, 16–17 November 2019; pp. 106–112.
10. Kusrini, H.A.; Fatta, S.P.; Widiyanto, W.W. Prototype of Pornographic Image Detection with YCbCr and Color Space (RGB) Methods of Computer Vision. In Proceedings of the 2019 International Conference on Information and Communications Technology (ICOIACT), Yogyakarta, Indonesia, 24–25 July 2019; pp. 117–122.
11. Ashan, B.; Cho, H.; Liu, Q. Performance Evaluation of Transfer Learning for Pornographic Detection. In Proceedings of the International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD 2019), Kunming, China, 20–22 July 2019; pp. 403–414.
12. Gangwar, A.; Fidalgo, E.; Alegre, E.; González-Castro, V. Pornography and child sexual abuse detection in image and video: A comparative evaluation. In Proceedings of the 8th International Conference on Imaging for Crime Detection and Prevention (ICDP 2017), Madrid, Spain, 13–15 December 2017; pp. 37–42.
13. He, Y.; Shi, J.; Wang, C.; Huang, H.; Liu, J.; Li, G.; Liu, R.; Wang, J. Semi-supervised Skin Detection by Network with Mutual Guidance. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 2111–2120.
14. Macedo, J.; Costa, F.; dos Santos, J.A. A benchmark methodology for child pornography detection. In Proceedings of the 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Parana, Brazil, 29 October–1 November 2018; pp. 455–462.
15. Perez, M.; Avila, S.E.; Moreira, D.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; Rocha, A. Video pornography detection through deep learning techniques and motion information. *Neurocomputing* **2017**, *230*, 279–293. [CrossRef]
16. Liu, Y.; Yang, Y.; Xie, H.; Tang, S. Fusing audio vocabulary with visual features for pornographic video detection. *Future Gener. Comput. Syst.* **2014**, *31*, 69–76. [CrossRef]
17. Song, K.; Kim, Y. Pornographic Video Detection Scheme Using Multimodal Features. *J. Eng. Appl. Sci.* **2018**, *13*, 1174–1182.
18. Liu, Q.; Feng, C.; Song, Z.; Louis, J.; Zhou, J. Deep Learning Model Comparison for Vision-Based Classification of Full/Empty-Load Trucks in Earthmoving Operations. *Appl. Sci.* **2019**, *9*, 4871. [CrossRef]
19. Zeyer, A.; Doetsch, P.; Voigtlaender, P.; Schlüter, R.; Ney, H. A comprehensive study of deep bidirectional LSTM RNNS for acoustic modeling in speech recognition. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2462–2466.
20. Song, K.; Kim, Y. A Fusion Architecture of CNN and Bi-directional RNN for Pornographic Video Detection. In Proceedings of the 7th International Conference on Big Data Applications and Services, Jeju, Korea, 21–24 August 2019; pp. 102–114.
21. Google Deepmind. Wavenet: A generative model for raw audio. *arXiv* **2016**, arXiv:1609.03499v2.

22. Farha, Y.A.; Gall, J. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 3575–3584.
23. Song, K.; Kim, Y. Pornographic Contents Detection Scheme using Bi-Directional Relationships in Audio Signal. *J. Korea Contents Assoc.* **2020**, accepted.
24. Adnan, A.; Nawaz, M. RGB and Hue Color in Pornography Detection. *Inf. Technol. New Gener.* **2016**, *448*, 1041–1050.
25. Jang, S.; Huh, M. Human Body Part Detection Representing Harmfulness of Images. *J. Korean Inst. Inf. Technol.* **2013**, *11*, 51–58. [CrossRef]
26. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef] [PubMed]
27. Sak, H.; Senior, A.W.; Beaufays, F. Long short-term memory recurrent neural network architectures for large scale acoustic modeling. In Proceedings of the 15th Annual Conference of the International Speech Communication Association, Singapore, 14–18 September 2014.
28. Zhang, X.; Yao, J.; He, Q. Research of STRAIGHT Spectrogram and Difference Subspace Algorithm for Speech Recognition. In Proceedings of the 2nd International Congress on Image and Signal Processing, Tianjin, China, 17–19 October 2009; pp. 1–4.
29. Keras: Deep Learning Library for Theano and Tensorflow. Available online: https://keras.io (accessed on 13 April 2020).
30. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2017**, arXiv:1412.6980v9.