applied sciences

MDPI

Article

# Cover the Violence: A Novel Deep-Learning-Based Approach Towards Violence-Detection in Movies

**Samee Ullah Khan [1], Ijaz Ul Haq [1], Seungmin Rho [2], Sung Wook Baik [1] and Mi Young Lee [1,*]**

[1] Intelligent Media Laboratory, Digital Contents Research Institute, Sejong University, Seoul 143-747, Korea; sameek3797@gmail.com (S.U.K.); hijaz3797@gmail.com (I.U.H.); sbaik@sejong.ac.kr (S.W.B.)

[2] Department of Software, Sejong University, Seoul 143-747, Korea; smrho@sejong.edu

* Correspondence: miylee@sejong.ac.kr

Received: 3 October 2019; Accepted: 7 November 2019; Published: 18 November 2019


**Abstract:** Movies have become one of the major sources of entertainment in the current era, which are based on diverse ideas. Action movies have received the most attention in last few years, which contain violent scenes, because it is one of the undesirable features for some individuals that is used to create charm and fantasy. However, these violent scenes have had a negative impact on kids, and they are not comfortable even for mature age people. The best way to stop under aged people from watching violent scenes in movies is to eliminate these scenes. In this paper, we proposed a violence detection scheme for movies that is comprised of three steps. First, the entire movie is segmented into shots, and then a representative frame from each shot is selected based on the level of saliency. Next, these selected frames are passed from a light-weight deep learning model, which is fine-tuned using a transfer learning approach to classify violence and non-violence shots in a movie. Finally, all the non-violence scenes are merged in a sequence to generate a violence-free movie that can be watched by children and as well violence paranoid people. The proposed model is evaluated on three violence benchmark datasets, and it is experimentally proved that the proposed scheme provides a fast and accurate detection of violent scenes in movies compared to the state-of-the-art methods.

**Keywords:** violence-detection; deep learning; video analytics; scene understanding

## 1. Introduction

In the advance era, internet penetration rate, low cost mass storages devices, and higher data transmission rates have significantly increased. Every year, thousands of movies are produced by the movie industry, and everyone can easily watch movies using smart phones and personal computers [1]. Nowadays, the entertainment activity for children is playing video games or watching movies, which mostly contain violent themes. Particularly, those movies which contain violent scenes attract viewers' attention, but they are not suitable for teenagers and children to watch [2,3]. Some of the researchers have suggested that watching violent scenes on TV programs or films tend to make teenage more aggressive with undesirable attitudes [4]. To prevent children from watching violent scenes in movies, it is important to develop a system that automatically removes violent scenes from the movies [5]. Moreover, these systems can also be useful for content providers to assist with children-suitability ratings for movies [6].

The most challenging part in violence detection is the definition of anomalies. It is very difficult to illustrate this high-level concept of violence using mathematical formulation precisely. Many researchers are facing this problem by utilizing their own concept of violence definition. With the identification of violence, the bulk of work is motivated on the low-level features, such as gradients, optical flow, and intensities. In this field, the first effort was done by L. Nayak et al. [7]. They proposed a method for the detection of blood and flames in a video and captured the degree of motion activities in violent

videos. Similarly, L.H. Chen et al. [8] presented a method, which detected fight scenes that have blood on the human body to recognize the violent scenes. C. Clarin et al. [9] developed a system that passed the representative frames of each scenes through Kohonen's self-organizing map to identify the color parts of the blood and skin. Later, B. Zhang et al. [10] proposed a potential violence detector system, which used both high- and low-level features to identify the violence contents in movies. P. Bilinski et al. [11] manually labelled each violent video according to ten subclasses and then trained a support vector machine classifier on multimodal features for violence recognition in video streams. Similarly, E.Y. Fu et al. [12] used a series of features on the basis of motion attributes, containing region, magnitude, and acceleration for violence detection. Lin. J et al. [13] proposed an audio-based violence classifier, which is weakly supervised and combined with an explosion, a motion, and a blood video classifier that separates non-violent and violent prospects in videos. Hassner. T et al. [14] proposed a descriptor that is based on optical flow magnitude changes between two frames. The violent flow descriptor then classifies the behaviors based on the support vector machine. Mabrouk A.B et al. [15] used a new descriptor to recognized violent scenes based on the magnitude and the orientation of the frame of interest. These feature descriptors show better results on the classification of crowded and non-crowded scenes. Khan, M. et al. [16] used scale-invariant feature transform (SIFT) descriptors for the classification of violent and non-violent behaviors especially in social media videos, such as animated cartoons. Nguyen N.T et al. [17] proposed the hierarchical hidden Markov model for violence detection. The main contribution is to handle the shared structures to recognize the indoor activities. Furthermore, for the real time recognition, they proposed a Rao-Blackwellised particle filter that efficiently calculated the distribution filters at a fixed time for each new scene. For the accurate detection of the indoor violent activities, some researchers incorporated video and audio features. Mahadevan, V. et al. [18] developed a violent scenes recognition system through flames and blood with the combination of the motions and the sound degrees. Huang, J.F. et al. [19] analyzed the behavior of a violent crowd. They present a method that only measured the statistical property in video frames, and then they used a support vector machine to discriminate the video frames into two classes, i.e., normal and abnormal. In a surveillance video stream, Zhang, T. et al. [20] identified and localized the abnormal activity that contained violence by developing a Gaussian model on the basis of the optical flow. They used an optical flow histogram orientation for the classification of violent and non-violent classes through a linear support vector machine. After the analysis of previous approaches, Nievas, E.B. et al. [21] proposed a new bag-of-words framework for action recognition in a specific domain of fight detection with descriptors of action, such as space-time interest points (STIP) and motion scale-invariant feature transform (MoSIFT). Gracia I.S et al. [22] also proposed a bag-of-words framework with the help of handcrafted features called motion blobs for the discrimination of fight and non-fight sequences. The spatio-temporal was used for the features extraction and classification purpose. In the temporal dimension of the video, most of the frames are highly correlated with their neighbors, therefore, researcher's attention increased towards the motion information among the contiguous frames. The 3D ConvNet takes successive frames as input, which can capture the appearance as well as short-term motion. Song, W. et al. [23] proposed 3D convolutional neural networks for the detection of violence in videos. They followed two strategies, for the short clips, 3D ConvNet were designed by using the uniform sampling method, while new sampling frame was adopted for longer clips. Ullah, F.U.M. et al. [24] proposed a violence detection framework based on a 3D Convolutional neural network (CNN) model. They classified the scenes into violent and non-violent based on the spatiotemporal features of 3D CNN and improved the existing work. Obviously, to detect a violent scene rich information from the videos can be used, because most violent scenes are related with actions and objects, such as fights, blood, and guns. Beside this, audio tracks also contain some important information especially when visual cues are not reliable therefore Mu, G. et al. [25] presented a system that detects violence and utilized audio features to input into a CNN model. They used the CNNs in two different ways: the first one is used for the deep audio feature extraction, and the other is used as a classifier directly. Furthermore,

they fused both the visual and the audio information, which significantly improves the performance. S. Benini et al. [26] developed a deep neural network for shot scale recognition. They also examined different scale features that quantitatively determine the films mood in terms of energetic arousal, hedonic tone, and tense arousal. J. Yu et al. [27] proposed an algorithm that detects the violent scenes in the videos. They constructed three novel feature approaches, including bag of visual words (BoVW) model, feature pooling technology, and dimensional histograms of gradient orientation (HOG3D). Moreover, they combined all these features based on a kernel extreme learning machine (KELM) for good and generalized abilities.

In the literature of violence recognition, mostly the deep learning-based methods are computationally expensive. There is no automatic system for the users, such as media and guardian services to specify the age category of the movies. In this paper, we proposed an empirically motivated approach to cover the violence in movies to resolve this issue. We fine-tuned a light-weight deep CNN model (MobileNet) with the utilization of pre-trained ImageNet weights in order to converge the model easily and the violence recognition in the datasets. The main contributions of the proposed model are summarized in the following bullet points.

- With the rise of technology and the increase in smart devices, children have access to different entertainment resources that include including movies and video games. Some movies have violent scenes or actions that are inappropriate for children or some sensitive people, which results in the need for automatic techniques to cover up the violent scenes. To prevent individuals from watching violent scenes, we presented a novel framework that incorporates shot segmentation, salient frame extraction, and an efficient CNN architecture to automatically detect violent scenes and cover them in movies.
- The structure of movie data is very complex, and it comprises of different scenes, each having shots. Prior to violence-detection, we take the advantage of some preprocessing mechanisms based on hand-engineered features that are easy to compute and help to structure the movie data before passing it into the next step. Our key contribution is the segmentation of a movie into various shots that have proper structure and provide assistance to the subsequent salient-frame extraction strategy.
- Feeding raw data to a CNN model without any prerequisite filtration mechanism results in wasting resources. Hence, we used a salient-keyframes extraction mechanism to select the salient frames and advance them to the trained CNN model for the final classification of a scene as violent or non-violent. This helped to reduce the time and complexity of our system and ensures efficient detection of violence.

The rest of the paper is organized as follows; Section 2 briefly explains the proposed system. experimental results are discussed in Section 3, which is followed by the conclusion in Section 4.

## 2. Methodology

In this section, the proposed technique is discussed in detail for violence and non-violence scenes detection in movies. Moreover, it also easily adoptable in surveillance videos, because of its high- level of efficiency, which is followed by an effective and light-weight CNN model that is famous for its quick classification predictions. Our method consists of three steps: (1) preprocessing, where we segment the overall movie into shots and select the most salient frame in each shot, (2) fine-tuning of the deep model on violence recognition datasets, and (3) postprocessing for reconstructing a movie without violent scenes. The overall flow of the proposed scheme is illustrated in Figure 1.

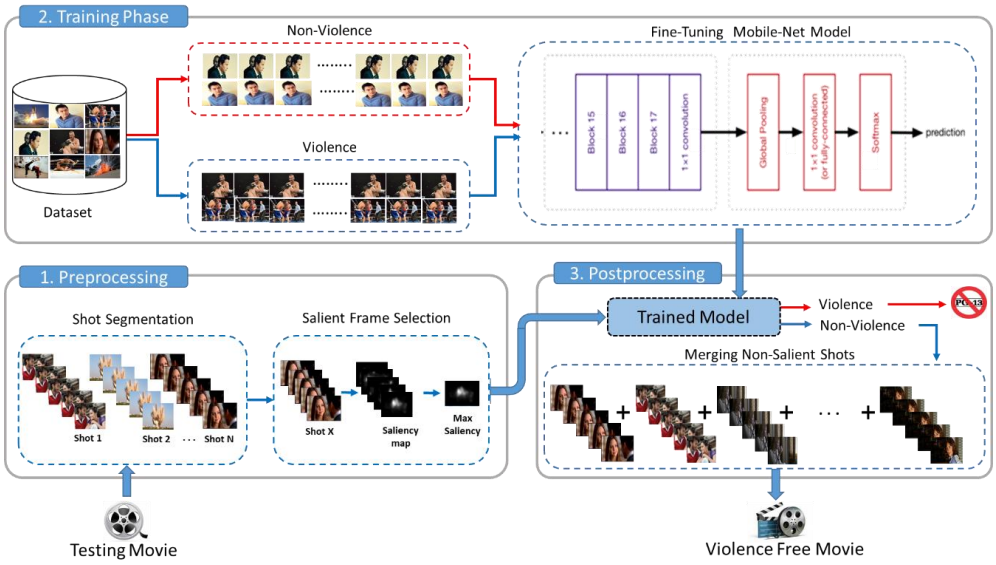


**Figure 1.** Proposed framework for violent scene detection in movies.

### 2.1. Preprocessing

In this step, our goal is to extract the salient frames of each shot, which are then fed into the trained model for violence recognition. Generally, movies can be easily down sampled into shots due to its hierarchical structure compared to other videos, such as surveillance, social media, and medical videos. This hierarchical structure assist to analyze the whole movie at shot-level. Therefore, in our proposed framework, first we segment the movie into shots by adopting a histogram-based method [28]. Next, to select a key-frame based on maximum information and clear objects, a sparse sampling and kernel density-based saliency estimation method with some additional functionality is used [29]. For instance, a saliency map is generated for all the frames of a shot and then compare with each other based on the maximum number of non-zero pixels divided by the total number of pixels in a frame. Equation (1) describes the whole mechanism of the key-frame selection.

$$K_F = Max\left(\sum_{j=1}^{N} \frac{\sum_{k=1}^{n\times m} S_{map(j)}(P_n)}{n \times m}\right) \tag{1}$$

where, $K_F$ is the selected key-frame in a shot having $N$ number of frames, $S_{map\ (j)}$ is the saliency map of the *jth* frame of size n×m, and $P_n$ is the number of non-zero pixels in it. The frame with the maximum number of non-zero pixels will be selected as a key-frame. This mechanism helps us to minimize the time complexity of the proposed method by selecting a frame with maximum information and discarding the frames with low information. Figure 2 represents a fighting shot from the movie *Undisputed II*, where the saliency map of each frame along with their scores are given.


**Figure 2.** Sample frames from movie 'Undisputed II' along with their saliency maps. The frames with red bounding boxes are discarded while the frame in blue is selected as a key-frame with a maximum saliency score of 0.7261.

### 2.2. Fine-Tuning MobileNet Model

Violence detection in computer vision systems is a challenging task especially in movie data due to the occurrence of intensity variations, complex crowd patterns, and various camera views. Therefore, in the case of using traditional methods, the researchers failed to capture effective features due to the occurrence of complex movements in the human body during a violence activity. Deep learning is a subtype of machine learning that allows us to visualize and see like a human [30,31]. In computer vision, a deep learning model requires extensive data for training to solve a problem as compared to machine learning. Its architecture contains three layers, which include convolutional, pooling, and fully connected. The initial layer of the network detects low-level features, such as edges and colors. These layers learn progressively from images as it goes through each layer of the network [32]. One of the main advantages of this model is that it continuously improves the performance as we fed more data. On the other hand, it takes more time in execution which is one of the main challenges. Therefore, we used a pre-trained model called MobileNet with transfer learning approach. Transfer learning is the reusability of pre-trained model weights on a new problem with a related one. This approach is very useful, because we do not have millions of data to train complex models in the real world problems [32]. To manage the time complexity and increase the performance of the model, we used the ImageNet weights. MobileNet is an efficient convolutional neural network, which is especially designed for developing mobile-based vision applications. The main idea behind the MobileNet is the utilization of depth-wise separable convolutions in order to build light-weight deep neural networks. In the case of regular convolutional layers, each layer contains a kernel or filter that applies to all channels of the input images, and each time it gives a sum of the weighted score across all input channels. Let's assume there are three input channels of an image, so when sliding a single convolutional kernel on that image gives only one output channel per pixel. Similarly, we run many convolutional filters, and each one gets its own output. In this manner, it gives a new value no matter how many channels it has, which is shown in Figure 3a. The MobileNet also follows this regular convolution in the very first layer. The rest of the layers use depth-wise separable convolution, which is actually the combination of two different operations, which include a depth-wise and pointwise convolution. In the depth-wise convolution, it does not give the combine result of all the channels. All the convolutional filters applied individually perform on each channel and give its own set of weight as shown in Figure 3b. For example, if the input image contains three channels, then the depth-wise convolutional also creates an image that has three channels. On other hand, the pointwise convolution concept is the same as a regular convolution, but it filters size is $1 \times 1$ as shown in Figure 3c. The main advantage of the model is that they create new features, which are the combination of both the pointwise and the depth-wise convolutions. All the convolutional layers of MobileNet follow the same techniques instead of the standard convolution. The results of both convolutions techniques are pretty similar, because both filter features from the data, but the standard one need more computational power to learn the weights.

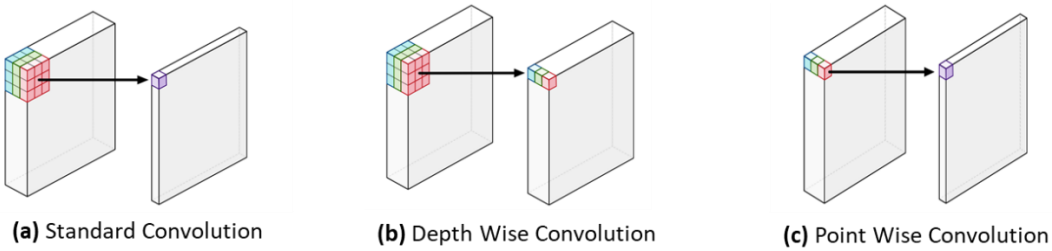


**Figure 3.** Visualization of different kinds convolution operations: (**a**) standard convolution; (**b**) depth-wise convolution; (**c**) point-wise convolution.

In the very first layer of MobileNet, a $3 \times 3$ convolutional is followed by batch normalization, and it has no pooling layers in between. Some of the depth wise layers containing stride 2 which automatically reduce the image dimensions. In each convolutional layer, ReLU is used as an activation

function. At the end, all global average pooling was used followed by a fully connected layer and a softmax. In this article, we fine-tuned the MobileNet model on the video frames of three different violence recognition datasets with hyper parameters, such as learning rate, momentum, batch size, and epochs, which gives good results. The feature map of violence and non-violence scenes are shown in Figure 4, where humans with violent actions are clearly visible in the violent feature maps.

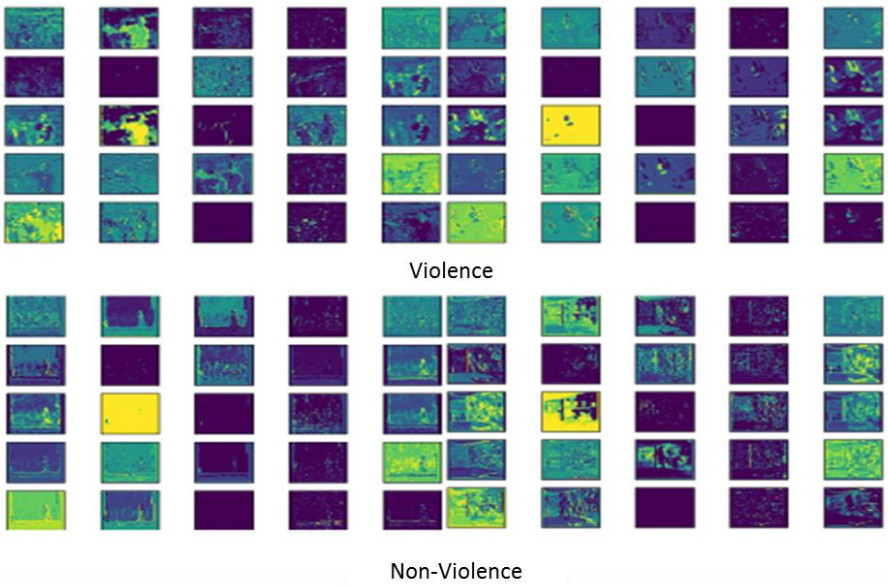


**Figure 4.** Feature maps of violence and non-violence scenes.

### *2.3. Postprocessing*

In this step, all the key frames are forwarded to the fine-tuned model. If the key-frame is recognized as violent, the entire shot is discarded, and the next key-frame is analyzed. In this way, all the shots recognized as violent are discarded, and the non-violent shots are merged together in the same sequence to generate the same movie as violence-free movie, which can be watched by underage children.

## 3. Experimental Results

In this section, the experimental results of the method are discussed in detail. The model was trained and tested on three violence recognition datasets. First, we checked the performance of model on each dataset separately. After that, we tested the overall performance of the model on combined datasets. Moreover, we compared the results of the proposed model with state-of-the-art techniques.

### *3.1. Datasets*

The proposed method is evaluated on various movies datasets, the first and second datasets were used for fight detection. The first dataset, Violence in Movies [21], is comprised of fighting and non-fighting videos. The fight class has 200 videos clips taken from various action movies while non-fight videos are taken from an available dataset for action recognition. The second dataset, which is Hockey Fight [14], contains 50 videos clips of fights and 50 videos clips without fights with a resolutions of 720 × 576 pixels. The third dataset, which is Violent Scene Detection (VSD) [33], is divided into three different subsections, which included web videos, annotations, and features. The web videos contain 86 videos downloaded from YouTube that have a normalized frame rates of 25. This dataset is publicly available as a single compressed file. In this work, each dataset first categorized into training and testing from the perspective of two classes, which include violence and non-violence, separately. Finally, we combined all the violent and non-violent frames in order to become as a whole or one complete dataset. The experimental results of training and testing the method to detect violence in

movie scenes are shown in Figure 5. The confusion matrixes for individuals and the combined datasets are shown in Figure 6.

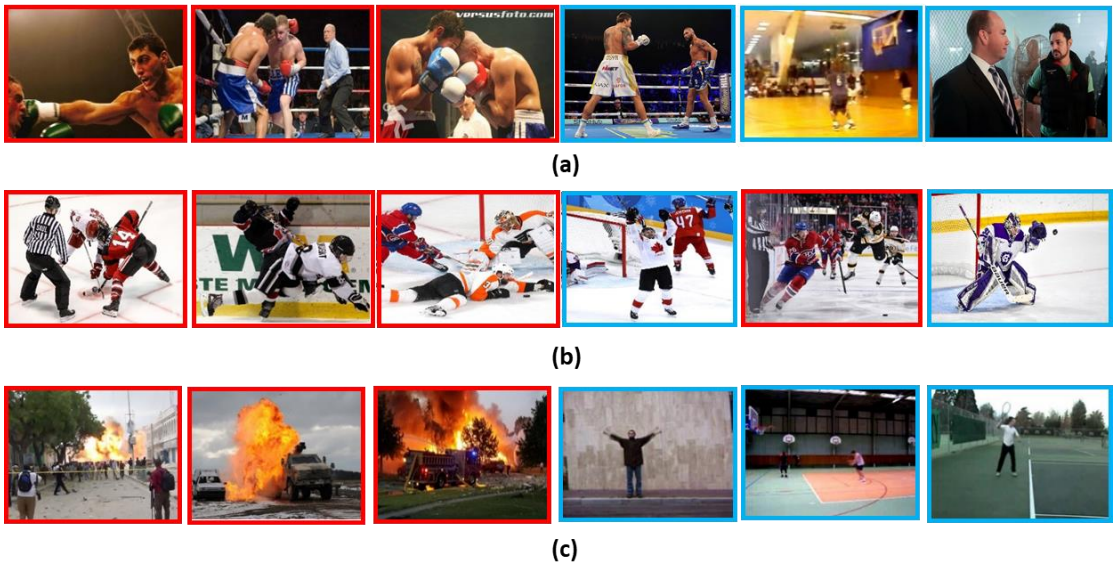


**Figure 5.** Sample frames from the datasets used for evaluation. In each row, the first three samples are from the violent class and the last three samples are from the non-violent class. (**a**) Violence in Movies, (**b**) Hockey Fights, and (**c**) Violence Scenes Detection datasets. The frames in red and blue are detected as violent and non-violent by the proposed method, respectively.

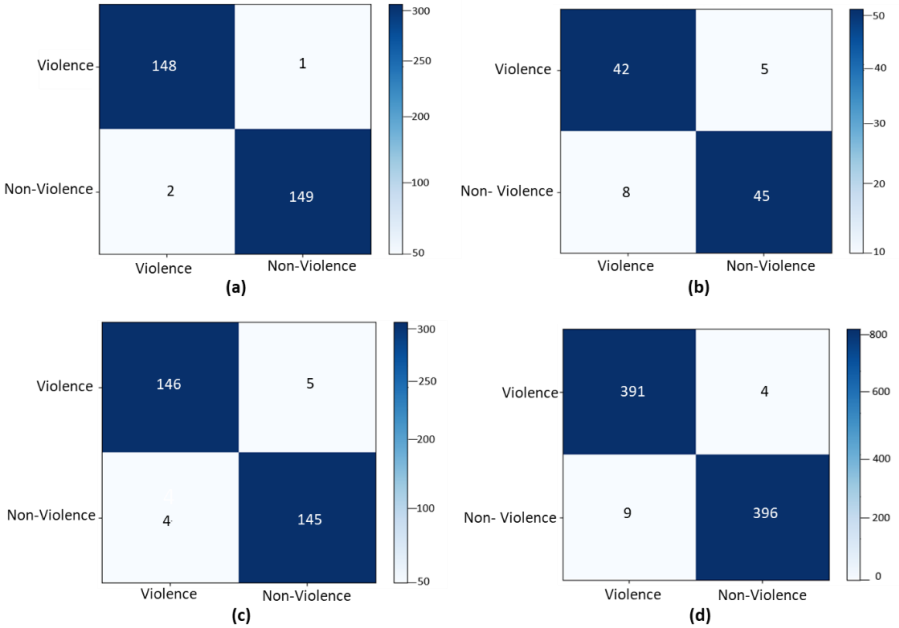


**Figure 6.** Confusion matrices of the (**a**) Violence in Movies, (**b**) Hockey Fight, (**c**) Violence Scenes Detection datasets, and the (**d**) combined dataset.

Particularly in deep learning literature, the majority of the methods use the same split of 75% and 25% as a training and testing sets [24]. Therefore, we used the same splitting in our datasets. From each dataset, we randomly selected 75% of the data for training, and the remaining 25% of the data were kept for testing. After that, all the three datasets were combined, and again we randomly selected the same percentages of data for training and testing. The training and testing losses decreased as the number of epochs processed increased. At the initial stage of learning, the loss was high, as the

data patterns were not fully learnt, but after few epochs, the model learnt the pattern and gained a certain accuracy. After 60 epochs, the loss became constant, and we achieved the overall accuracies of training (98%) and testing (96.3%) on the 80th epoch. The loss to epochs' ratio is graphically illustrated in Figure 7.

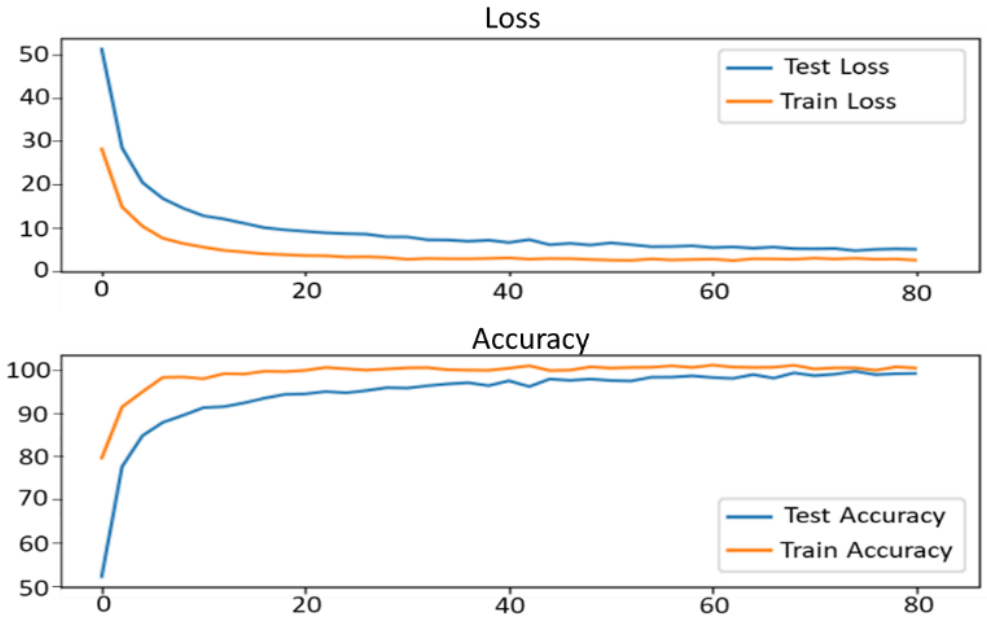


**Figure 7.** Performance evolution of the proposed model.

### *3.2. Comparative Analysis*

In this portion, we evaluate the performance of the proposed method with the state-of-the-art techniques using the same datasets. The existing bag-of-words (BoW) and support vector machine (SVM) algorithm was used for the violent scene detection [21]. Furthermore, they used the space time interest points (STIP) to detect corner points in an image and then extracted histogram oriented gradient using MOSIFT technique, and the bag-of-words approach was adopted for text retrieval. They used the Violence in Movies dataset and achieved 89.5% accuracy, while we achieved 99.5% accuracy on the same dataset. The second method is based on Bag-of-word framework using Violence in movies and the Hockey Fights dataset. They extracted features from the motion blobs in video sequences and trained the Random Forest on these motion features to detect violence, which achieved 96.9% on Violence in Movies and 82.4% accuracy on Hockey Fights. The third article extracted the violent descriptors using a statistical method. They trained the linear SVM on statistical features for violent and non-violent scenes using the hockey fight dataset, they achieved 82.9% accuracy [14]. The fourth comparison method is the content based method using a VSD dataset, which achieved 96.9% accuracy, while the proposed method achieved 97.0% [33]. The last method used for comparison fused different features with a classifier, which performed better than our proposed method in terms of accuracy, but its time complexity was high. They achieved a 99.9% accuracy on Violence in Movies, which is slightly greater than ours, which is 99.5%. On the other hand, it achieved 95.5% accuracy on the Hockey Fight dataset, which is better than ours, because we obtained 87% accuracy on the same dataset. The proposed method achieved better performance on all the datasets with a high accuracy and a lower time complexity. The experimental results demonstrated that our fine-tuned model achieved better results as compared to the state-of-the-art methods. We also verified the compatibility of our framework in different domains, such as surveillance and movies. The rest of the methods in the literature are limited to only a single domain, such as movies or surveillance. The malleability of our framework is due to the usage of effective CNNs for the final output prediction of the input frames. The details are given in Table 1.

**Table 1.** Comparison of the proposed method with the state-of-the-art methods.

| Methods | Datasets Accuracies (%) | | | Domain Adoptability | |
|---|---|---|---|---|---|
| | Violence in Movies [21] | Hockey Fight [14] | Violent Scene Detection (VSD) [33] | Movies | Surveillance |
| **STIP, Bow and SVM [21]** | 89.5 | - | - | ✔ | ✖ |
| **Motion Blobs and Random Forest [22]** | 96.9 | 82.4 | - | ✔ | ✖ |
| **VIF [14]** | - | 82.9 | - | ✖ | ✔ |
| **Content based method [33]** | | - | 96.9 | ✔ | ✖ |
| **HOG3D+KELM [27]** | 99.9 | 95.05 | - | ✔ | ✖ |
| **Proposed Method** | **99.5** | **87.0** | **97.0** | ✔ | ✔ |

## 4. Conclusions and Future Work

Violence scene detection in movies is a challenging problem due to the diverse content and large variations quality. In this paper, a three folded movie analysis scheme is proposed to detect the violent scenes. First, the entire movie is segmented into shots and then a representative frame from each shot is selected based on the level of saliency. Next, a pre-trained MobileNet model is fine-tuned on three benchmark datasets for violence recognition. Finally, the shots that are detected as violent are discarded, and the non-violent shots are merged to generate a violence free-movie that can be watched by children. In the future, we want to boost this work by utilizing sequential learning parameters, such as long short-term memory (LSTM) [34] with CNNs for effective violence detection in complex scenarios. Furthermore, we aim to cover other application domains that are different particularly surveillance [35] in smart cities, and reduce the size of the features by introducing embedded vision [36] technologies for violence-detection.

**Author Contributions:** Conceptualization, S.U.K.; Investigation, I.U.H.; Methodology, S.U.K.; Project administration, M.Y.L.; Supervision, S.W.B. and M.Y.L.; Visualization, S.U.K.; Writing—original draft, S.U.K. and I.U.H.; Writing—review & editing, I.U.H., S.R. and S.W.B.

**Acknowledgments:** This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education (numbers 2018R1D1A1B07043302 and 2016R1D1A1A09919551).

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Elliott, C.; Dastidar, S.G. The Indian Film Industry in a Changing International Market. *J. Cult. Econ.* **2019**, in press.
2. Romer, D.; Jamieson, P.E.; Jamieson, K.H.; Lull, R.; Adebimpe, A. Parental desensitization to gun violence in PG-13 movies. *Pediatrics* **2018**, *141*, e20173491. [CrossRef]
3. Ferguson, C.J.; Markey, P. PG-13 rated movie violence and societal violence: Is there a link? *Psychiatr. Q.* **2019**, *90*, 395–403. [CrossRef]
4. Lam, V.; Phan, S.; Le, D.-D.; Duong, D.A.; Satoh, S.I. Evaluation of multiple features for violent scenes detection. *Multimed. Tools Appl.* **2017**, *76*, 7041–7065. [CrossRef]
5. Hauptmann, A.; Yan, R.; Lin, W.-H.; Christel, M.; Wactlar, H. Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news. *IEEE Trans. Multimed.* **2007**, *9*, 958–966. [CrossRef]
6. Shafaei, M.; Samghabadi, N.S.; Kar, S.; Solorio, T. Rating for Parents: Predicting Children Suitability Rating for Movies Based on Language of the Movies. *arXiv* **2019**, arXiv:1908.07819.
7. Nayak, L. Audio-Visual Content-Based Violent Scene Characterisation. Ph.D. Thesis, National Institute of Technology, Rourkela Odisha, India, 2015.
8. Chen, L.-H.; Hsu, H.-W.; Wang, L.-Y.; Su, C.-W. Violence detection in movies. In Proceedings of the 2011 Eighth International Conference Computer Graphics, Imaging and Visualization, Singapore, 17–19 August 2011; pp. 119–124.

9. Clarin, C.; Dionisio, J.; Echavez, M.; Naval, P. DOVE: Detection of movie violence using motion intensity analysis on skin and blood. *PCSC* **2005**, *6*, 150–156.
10. Zhang, B.; Yi, Y.; Wang, H.; Yu, J. MIC-TJU at MediaEval Violent Scenes Detection (VSD) 2014. In Proceedings of the MediaEval 2014 Workshop, Barcelona, Spain, 16–17 October 2014.
11. Bilinski, P.; Bremond, F. Human violence recognition and detection in surveillance videos. In Proceedings of the 2016 13th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS), Colorado Springs, CO, USA, 23–26 August 2016; pp. 30–36.
12. Fu, E.Y.; Leong, H.V.; Ngai, G.; Chan, S.C. Automatic fight detection in surveillance videos. *Int. J. Pervasive Comput. Commun.* **2017**, *13*, 130–156. [CrossRef]
13. Lin, J.; Wang, W. Weakly-supervised violence detection in movies with audio and video based co-training. In *Pacific-Rim Conference on Multimedia*; Springer: Berlin/Heidelberg, Germany, 2009; pp. 930–935.
14. Hassner, T.; Itcher, Y.; Kliper-Gross, O. Violent flows: Real-time detection of violent crowd behavior. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 1–6.
15. Mabrouk, A.B.; Zagrouba, E. Spatio-temporal feature using optical flow based distribution for violence detection. *Pattern Recognit. Lett.* **2017**, *92*, 62–67. [CrossRef]
16. Khan, M.; Tahir, M.A.; Ahmed, Z. Detection of violent content in cartoon videos using multimedia content detection techniques. In Proceedings of the 2018 IEEE 21st International Multi-Topic Conference (INMIC), Karachi, Pakistan, 1–2 November 2018; pp. 1–5.
17. Nguyen, N.T.; Phung, D.Q.; Venkatesh, S.; Bui, H. Learning and detecting activities from movement trajectories using the hierarchical hidden Markov model. In Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05), San Diego, CA, USA, 20–25 June 2005; pp. 955–960.
18. Mahadevan, V.; Li, W.; Bhalodia, V.; Vasconcelos, N. Anomaly detection in crowded scenes. In Proceedings of the 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 1975–1981.
19. Huang, J.-F.; Chen, S.-L. Detection of violent crowd behavior based on statistical characteristics of the optical flow. In Proceedings of the 2014 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, China, 19–21 August 2014; pp. 565–569.
20. Zhang, T.; Yang, Z.; Jia, W.; Yang, B.; Yang, J.; He, X. A new method for violence detection in surveillance scenes. *Multimed. Tools Appl.* **2016**, *75*, 7327–7349. [CrossRef]
21. Nievas, E.B.; Suarez, O.D.; García, G.B.; Sukthankar, R. Violence detection in video using computer vision techniques. In *International conference on Computer Analysis of Images and Patterns*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 332–339.
22. Gracia, I.S.; Suarez, O.D.; Garcia, G.B.; Kim, T.-K. Fast fight detection. *PLoS ONE* **2015**, *10*, e0120448.
23. Song, W.; Zhang, D.; Zhao, X.; Yu, J.; Zheng, R.; Wang, A. A Novel Violent Video Detection Scheme Based on Modified 3D Convolutional Neural Networks. *IEEE Access* **2019**, *7*, 39172–39179. [CrossRef]
24. Ullah, F.U.M.; Ullah, A.; Muhammad, K.; Haq, I.U.; Baik, S.W. Violence detection using spatiotemporal features with 3D convolutional neural network. *Sensors* **2019**, *19*, 2472. [CrossRef] [PubMed]
25. Mu, G.; Cao, H.; Jin, Q. Violent scene detection using convolutional neural networks and deep audio features. In *Chinese Conference on Pattern Recognition*; Springer: Singapore, 2016; pp. 451–463.
26. Benini, S.; Savardi, M.; Bálint, K.; Kovács, A.B.; Signoroni, A. On the influence of shot scale on film mood and narrative engagement in film viewers. *IEEE Trans. Affect. Comput.* **2019**. [CrossRef]
27. Yu, J.; Song, W.; Zhou, G.; Hou, J.-J. Violent scene detection algorithm based on kernel extreme learning machine and three-dimensional histograms of gradient orientation. *Multimed. Tools Appl.* **2019**, *78*, 8497–8512. [CrossRef]
28. Li, Z.; Liu, X.; Zhang, S. Shot boundary detection based on multilevel difference of colour histograms. In Proceedings of the 2016 First International Conference on Multimedia and Image Processing (ICMIP), Bandar Seri Begawan, Brunei, 1–3 June 2016; pp. 15–22.
29. Tavakoli, H.R.; Rahtu, E.; Heikkilä, J. Fast and efficient saliency detection using sparse sampling and kernel density estimation. In *Scandinavian Conference on Image Analysis*; Springer: Berlin/Heidelberg, Germany, 2011; pp. 666–675.

30. Zhang, X.; Yao, L.; Wang, X.; Monaghan, J.; Mcalpine, D. A Survey on Deep Learning based Brain Computer Interface: Recent Advances and New Frontiers. *arXiv* **2019**, arXiv:1905.04149.
31. Liu, N.; Wan, L.; Zhang, Y.; Zhou, T.; Huo, H.; Fang, T. Exploiting convolutional neural networks with deeply local description for remote sensing image classification. *IEEE Access* **2018**, *6*, 11215–11228. [CrossRef]
32. Haq, I.U.; Ullah, A.; Muhammad, K.; Lee, M.Y.; Baik, S.W. Personalized Movie Summarization Using Deep CNN-Assisted Facial Expression Recognition. *Complexity* **2019**, *2019*, 10.
33. Demarty, C.-H.; Penet, C.; Soleymani, M.; Gravier, G. VSD, a public dataset for the detection of violent scenes in movies: Design, annotation, analysis and evaluation. *Multimed. Tools Appl.* **2015**, *74*, 7379–7404. [CrossRef]
34. Hussain, T.; Muhammad, K.; Ullah, A.; Cao, Z.; Baik, S.W.; de Albuquerque, V.H.C. Cloud-assisted multi-view video summarization using CNN and bi-directional LSTM. *IEEE Trans. Ind. Inform.* **2019**, in press. [CrossRef]
35. Muhammad, K.; Hussain, T.; Baik, S.W. Efficient CNN based summarization of surveillance videos for resource-constrained devices. *Pattern Recognit. Lett.* **2018**, in press. [CrossRef]
36. Hussain, T.; Muhammad, K.; Khan, S.; Ullah, A.; Lee, M.Y.; Baik, S.W. Intelligent Baby Behavior Monitoring using Embedded Vision in IoT for Smart Healthcare Centers. Journal of Artificial Intelligence and Systems. *J. Artif. Intell. Syst.* **2019**, *1*, 15.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).