Filtering Of Inappropriate Video Content: A Survey

Mahmoud Taha[1 [0000−0002−9148−6960]], Ahmed Zakey[1 [0000−0002−9148−6960]], and Abdulwahab Al-Sammak[1 [0000−0002−9148−6960]]

Faculty of Engineering at Shoubra, Cairo, Egypt
http://feng.bu.edu.eg/feng/en/

Abstract. With the emergence of screened films, Video content classification has become ubiquitous. Television films and Internet sites films are a big source of violence that may psychologically hurt teenagers. Although recently, Deep learning video classification has been developed quickly, a Comprehensive survey is needed to summarize the previous work done in this field. Therefore, this survey paper shows the common methods used in video classification. We further discuss the importance of filtering sensitive content such as (pornography, violence, gory, etc.) because of the increasing consumption of films by people of all ages. Several real-world verdict cases are similar scenarios to films with many scenes of violence. As deep learning has shown big success in computer vision areas, researchers are giving it a lot of attention.

Keywords: Video filtering · Video analysis · Video classification.

1 Introduction

Videos can be categorized into many categories. It is possible to classify them according to the content, such as educational or entertainment content, or to classify them according to the type of images from which these videos were prepared, such as two-dimensional films, three-dimensional films, and cartoon films. As for the video content, there is much content that should be filtered to protect children, including sexual content, violence and the content that causes autism. We will present some previous attempts aimed at filtering some or all of the inappropriate content from some types of videos.

Pornography has many common definitions in the psychological and scientific fields. One of the most known definitions states that pornography [1] is any commercial product in the form of fictional drama designed to elicit or enhance sexual arousal. Another definition states that it is any printed or visual material containing the explicit description or display of sexual organs or activity, intended to stimulate sexual excitement and whatever you choose from the previous definitions, they all agreed that pornography is all about sexual excitement and show the bounds.

Violence detection is a strenuous problem due to the heterogeneous content and variable quality of videos. Supervised classification is a fundamental task in machine learning. violent scenes are associated with nude colour in video frames and groans and moans in the audio. there are cases like wrestling in which we may have false-negative or false-positive results. Traditional video filtering methods only work on a single dimension of features such as video frames colour analysis or video. When multiple dimensions of features such as (image frames colours, audio content, motion in the frame sequence of video, or emotions of the audience) are used, how can these features be integrated to perform accurate classification? The existence of such features raised the need of multi-feature learning [2], [3] [4].

colour redFor multi-feature classification, it may be required to identify classes of subjects that differ in each of the data views. In the past two decades, video content filtering has attracted more and more attention, so it is necessary to summarize the state of the art and outline open problems and future enhancements. We divide the video filtering methods into model-based approaches and similarity-based approaches. Generative approaches learn the distribution of the features and use generative models to represent video classification. Discriminative approaches optimize a function that tries to keep down the different classes average similarity. Discriminative approaches have many types based on the combination method of the multi-feature information such as common eigenvector matrix or common indicator matrix.

we can also divide previous work done into two categories one of them is the targeted content to be filtered such as (violence only , pornography only or both of them) , the second category can be the targeted media type such as (real videos only , animation cartoons only or both of them), as shown in fig 1.

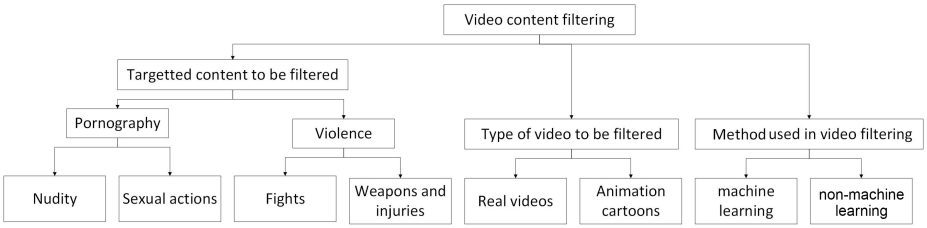

Fig. 1. Types of Inappropriate content filtering

As far as what this paper is concerned, The video filtering and video analysis papers of are published in top machine learning venues like the International Conference on Machine Learning (ICML) [5], Neural Information Processing Systems (NIPS) [6], IEEE International Conference on Computer Vision and

Pattern Recognition (CVPR) [7], International Conference on Computer Vision (ICCV) [8], Association for the Advancement of Artificial Intelligence (AAAI) [9], International Joint Conference on Artificial Intelligence (IJCAI) [10].

Although video filtering has shown considerable success in practice, some open problems limit its advancement. We explain several open problems and hope that readers can have a better version of the automatic video filtering using deep learning. This paper is organized as follows. In section 2, we introduce the existing deep learning methods for video filtering. In section 3, we review video filtering using different approaches other than deep learning. In Section 4, we review previous work done that targets nudity and pornography. Section 5 shows the previous work that targets violence and injuries. In section 6, we introduce some papers that work on animated cartoon classification. Section 7 shows different performance evaluation techniques. In section 8, we introduce a list of data-sets used in video filtering and video content analysis. In Section 9, we show a list of challenges in video filtering research. Finally, we introduce the conclusions.

Paper [11] proposes a discriminative method for event detection using video representation. The focus is to leverage Convolutional Neural Networks (CNNs) to improve event detection.it begins by the frame-level extracting by CNN descriptors then it generates video level representations.

2 Video filtering using machine learning approaches

Among all media types (e.g., texts and audio), images are the most used pornography carrier. Since pornography often has skin exposure, skin detection is commonly used for pornography detection. However, because skin detection is challenging, the Approaches used have a limited generalization ability. Vu Lam and Duy-Dinh Le introduced MediaEval [12] that combined the trajectory-based motion features with SIFT-based and audio features. their results show that the trajectory-based motion features still have very competitive performance. the combination with image features and audio features can improve overall performance for violence detection in videos.

paper [13] states that recently dense trajectories were shown to be an efficient video representation for action recognition and achieved state-of-the-art results on a variety of data-sets. they improve performance by taking into account camera motion to correct them. To estimate camera motion, matching feature points between frames using SURF descriptors and dense optical flow is important.

Human motion is in general different from camera motion and generates inconsistent matches. To improve the estimation, a human detector is employed to remove these matches. Given the estimated camera motion, paper [13] removes

trajectories consistent with it and uses this estimation to cancel out camera motion from the optical flow. This significantly improves motion-based descriptors, such as HOF and MBH. Focusing on filtering real-time videos, Nevenka and Radu produced an effective patent [14] that automatically filters multimedia program content in real-time based on stock and user-specified criteria. It includes one or more multimedia processors, which analyze incoming visual, audio, and textual, content and compares the resultant analysis to specific user-specified or selected criteria.

In paper [15], nude detection is implemented with the use of two algorithms. The first algorithm detects humans from the processed frames and then crop them out. Then the second algorithm is the nude detection algorithm used to determine if the images are nudes or not. the technique used in this paper has many problems. first is that while detecting nudity, it fails to properly classify black and white images as containing nudity or not because this paper is based mainly on skin colour. When a person is wearing a skin-toned dress the paper's algorithm gives a false positive result and it also fails to detect nudity if most of the body portion is not included in the video frames. paper [16] presents a system developed for content-based news video browsing for home users. This system integrates the audio-visual as well as text detection and NLP technique analysis to extract structure and content information of news video and to organize and categorize news stories.

Jin [17] used weighted multiple instances learning to model each image as a bag of regions and train a generic region-based recognition model taking into account the degree of pornography for each region. He shows that the key pornographic contents often are located in local regions in an image, and the background regions may be destructive. He stated that there are two main pornographic contents: the private body parts, and sexual behaviours. We perform bounding box annotation for these key pornographic contents in the training set, and for the sexual behaviour, Using annotations, there are only a small number of annotations (less than 10) required for both types of pornographic contents. each image X is modeled as a bag of n regions $x_i|i = 1, ..., n$. Given a region, the deep CNN extracts layer-wise representations from the first convolutional layer to the last fully connected layer. His CNN architecture is inspired by the GoogLeNet model [18]. The final results for this method are shown in fig 2.

3 Video filtering based on non-machine learning approaches

Due to some obstacles in machine learning such as data acquisition and powerful resources required for training, Some researchers tried to avoid using machine learning and use fixed algorithms instead.

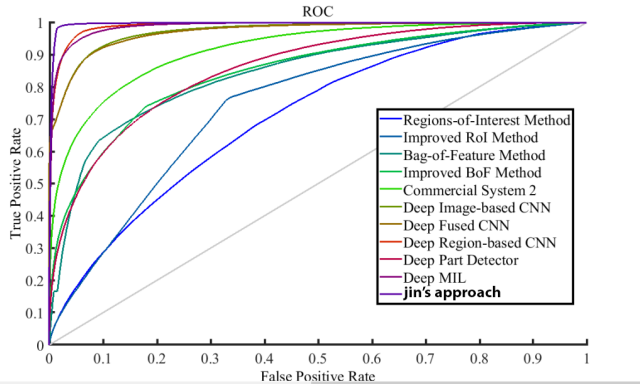

Fig. 2. ROC curve of different methods for pornographic image recognition.Jin's method significantly outperforms traditional methods.In particular,He achieve accuracy of 97.52% TPR at 1% FPR

3.1 Blocking Adult Images Based on Statistical Skin Detection [19]

One of the most important papers worked in images filtering without using machine learning directly to train on images is this paper [19]. Paper [19] shows a method to filter adult images using statistical algorithms. Human skin pixels detection from an image is performed. MaxEnt method used for inferring models from a data set. It works as follows: (1) choose relevant features (2) compute their histograms on the training set (3) write down the maximum entropy model within the ones that have the feature histograms as observed on the training set (4) estimate the parameters of the model (5) use the model for classification. The output of skin detection is a map indicating the probabilities of skin on pixels shown in fig [3]. After that paper [19] computes a sequence of 9 features from this skin map which form a feature vector and uses the fit ellipses to catch the characteristics of skin distribution. Two ellipses are used for each skin mapthe Global Fit Ellipse (GFE) and the Local Fit Ellipse (LFE). A multi-layer perceptron classifier is trained for these features. It is a semi-linear feed-forward net with back-propagation of error. Skin detection is never perfect and analysis of text in images may increase the accuracy of classification.

4 Video Filtering Targeting Pornography

Due to research report [20], The term pornography can be defined as any sexually explicit material that is generally intended to sexually arouse the audience. In this section, we show different papers and researches that target filtering pornography in videos regardless of the used technique for filtering or the type of targeted videos.

Fig. 3. On the right: original color image. On the left: the corresponding skin map output by TFOMure.

paper [21] shows that motion information in videos can help to classify difficult cases of pornography. Only a few works have used motion information in automatic video filtering such as Space-Time Interest Points (STIP)[22] and Temporal Robust Features (TRoF) [23]. As shown, Some researches sought to extract motion information by feeding frames to the CNNs [24], while others opted for feeding this information to the network through a previously computed Optical Flow Displacement Fields image representation [6].

In paper [21], Authors reported a Summary of approaches on skin, nudity or pornography detection. A good survey on this point can be found in this paper.

5 Video Filtering Targeting Violence

In psychology, violent scenes are defined as scenes one would not let an 8-year-old child see because they contain physical violence. This is the subjective definition. The objective definition is physical violence or accident resulting in human injury or pain. In this section, we show different papers and researches that target filtering violence, injuries and fights in videos regardless of the used technique for filtering or the type of targeted videos.

Violence classification started by detecting blood and screams using sound features [25]. In paper [26], A system that identifies violence in each frame by extracting HOG features and classifies it using a random forest classifier. this system shows low accuracy even it didn't require a GPU for computations. Recent researches like [6] and [27] detect violence and fight using deep learning

architectures such as long short-term memory (LSTMs), and two-stream CNN's [6] and convolutional neural networks (CNNs) in [27]. These methods show better accuracy than previous algorithms used.

Author in paper [28] used a pre-trained CNN architecture VGG-19 [29] followed by LSTM. The results from CNN of each frame are grouped and fed to the LSTM. this model has an accuracy of 94.765% which is a good performance. Qichao Xu [30] used the FlowNet 2.0 [31] model for estimating optical flow and the pre-trained SSD-VGG16 for human detection. The localization phase is combined for predictions after a two-stream C3D network [32] is trained on the active regions. The methodology from [33] detects violence using keyframe extraction algorithm and 3D ConvNet achieving 93.5% accuracy with the Crowd violence dataset. Keeli at al. [34] used AlexNet with transfer learning and utilized the Lucas-Kanade method for finding the optical flows of frames. Templates are made from the optical flow and fed as input to AlexNet for feature extraction. After using multiple classifiers, The best results have been obtained from the SVM classifier. In paper [35], Auther used an LSTM network to perform sequence prediction on the vectors resulting from a convolution neural network that extracts the spatial feature as shown in fig[4]. For the spatial feature extraction, The architecture of Xception [36] network has been used with the pre-trained model on the ImageNet dataset [37].

for paper [38], they predict sequences in consecutive frames using both spatial features and temporal or time-related features as well. The Convolutional Neural Network (CNN) used in this paper, is composed of an input convolutional layer followed by three layers of convolution and max pooling. The kernel size for each convolutional layer is 3 3. 64 kernels are used in each convolutional layer. The output from each convolutional layer after passing through relu activation as shown in fig[5].

As soon as the blood is visually present in the image, Additional tags are required to represent the proportion of it in the image. These tags can be: unnoticeable, low, medium, high with the following meanings:

- unnoticeable: there are some blood pixels no more than 5% of the image
- low: pixels represent blood are is in [5% and 25%].
- medium: pixels represent blood are in [25%, 50%].
- high: pixels represent blood are higher than half pixels of the image.

Fights can be annotated in different tags:

- 1 vs 1: fighting between only two people.
- small: for a group of people less than 10 persons.
- large: for a large group of people greater than 10 persons.
- distant attack: somebody is shot.
- Human against animal fights.
- Explosions, Scream and gunshot sounds can facilitate correct classification.

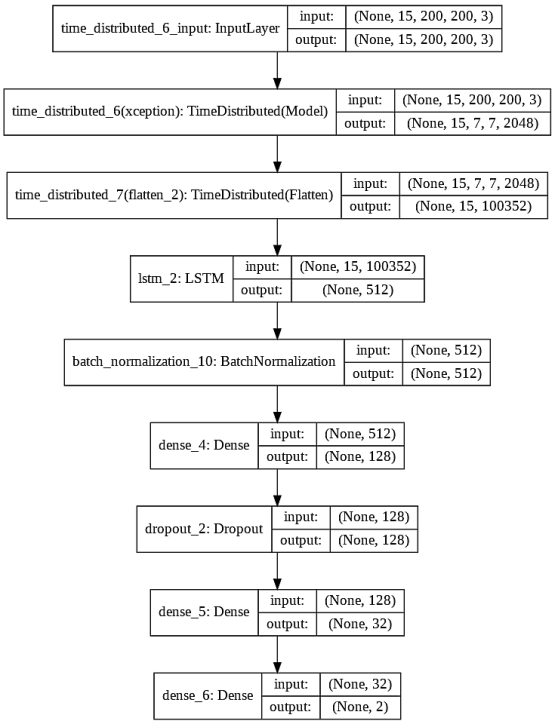

Fig. 4. Model Architecture for Sarthak Sharma Paper.

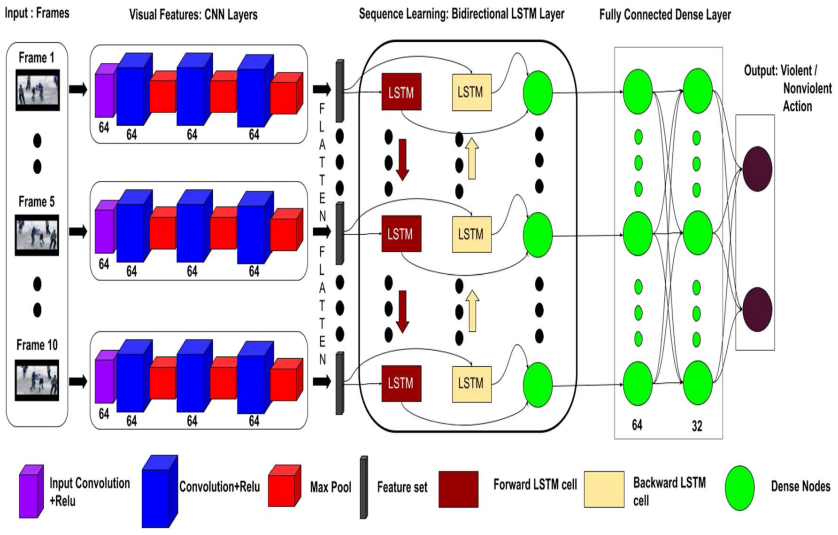

Fig. 5. Model Architecture for Rohit Halder Paper.

6 Filtering Animated Cartoons

Child pornography [39] is an important problem in cyber-security. The increased use of the internet increased the circulation of offensive images of children.

6.1 Unsupervised Discovery Character Dictionaries in Animation Moves [40]

This paper [40] shows a method to create a characters' dictionary for an animation video using unsupervised learning. A set of eight movies was used to evaluate the method suggested in the paper and can be used for labelling animation characters. Character dictionaries in animated cartoons can be the first step for cartoon content analysis. Paper [40] uses a deep neural network natural objects detector to identify some initial characters. These initial characters are pruned using visual object tracking. A character dictionary is then generated for each movie.

One of the best results for the method mentioned is a generalization for animation movies [19] at scale. AS shown in fig[8], there is a high degree of heterogeneity of animated cartoon characters, So Paper [40] can identify both 2-dimensional and 3-dimensional characters with a different collection of character designs, But In digital animations, objects can be fictional so object detection can be more challenging. paper [40] uses a deep neural network called Multi-Box [41] , [42] designed for object detection. Multi-Box was used only to generate a set of character candidates. The performance is quantied by comparing the reference characters with the output dictionaries from the method used.

Due to the success of AlexNet, deep learning has become the most common method for image recognition. A few years ago, R-CNNs were developed to deal with the object detection task but three were many problems with the above networks such as too-long data training and multiple phases training. As a result of these problems, new networks were developed such as YOLO (You Only Look Once) and SSD MultiBox (Single Shot Detector). SSD MultiBox was released in 2016 and reached new records for object detection tasks, scoring over 74% mAP (mean Average Precision) at 59 frames per second on standard datasets such as COCO.

As shown in the fig [6, 7], Its called Single Shot because object localization and classification are achieved in a single forward pass of the network built on the venerable VGG-16 architecture with 1x1 convolutions that helps in dimension reduction but discards the fully connected layers. MultiBoxs loss function also combined two components confidence loss which uses categorical cross-entropy and location Loss which uses L2-Norm. The expression for Loss can be stated as $multibox_loss = confidence_loss + alpha * location_loss$.

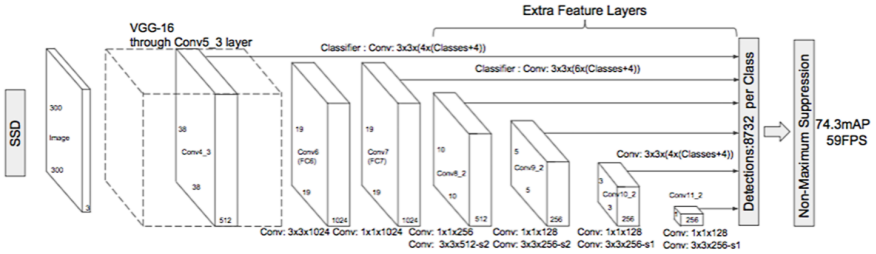

Fig. 6. SSD MultiBox network architecture.

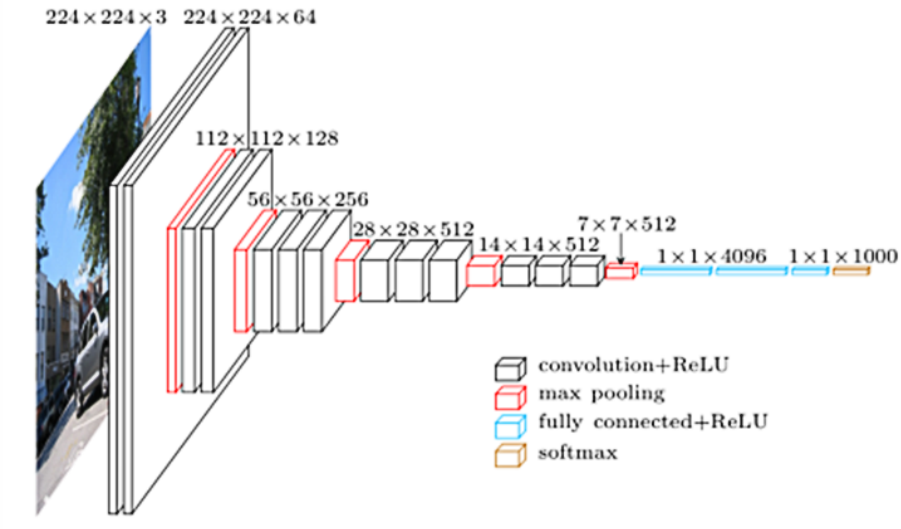

Fig. 7. MultiBox network architecture.

In MultiBox, the researchers created what we call priors, which are pre-computed, fixed-size bounding boxes that closely match the distribution of the original ground truth boxes. In fact, those priors are selected so that their Intersection is greater than 0.5. It is a better strategy than starting the predictions with random coordinates. The resulting architecture contains 11 priors per feature map cell (8x8, 6x6, 4x4, 3x3, 2x2) and only one on the 1x1 feature map, resulting in a total of 1420 priors per image. Assuming we have configured 2 diagonally opposed points (x1, y1) and (x2, y2) for each b default bounding boxes per feature map cell, and c classes to classify, on a given feature map of size $f = m * n$, SSD would compute $f * b * (4 + c)$ values for this feature map. Non-maximum suppression: boxes with a confidence loss threshold less than (e.g. 0.01) and IoU less than (e.g. 0.45) are discarded, and only the top N predictions are kept.

Fig. 8. Examples of the heterogeneity of animated cartoon characters.

Future work One of the drawbacks of paper [40] is using an object detector that was trained with natural human images. As a result, the Authors plan to solve this issue using transfer learning. Authors also suggest using the relevant exemplars and associated cluster as a single unit to facilitate animation video classification.

6.2 Video Analysis for Cartoon-like Special Effects

Paper [43] demonstrates a system for rendering motion within any artistic video. Key to paper [43] work is the analysis of trajectories. it uses additional ghosting lines which indicate the trailing edge of the object as it moves along the streak lines, like in fig [9]. Ghosting lines are usually perpendicular to streak

lines. Deformation is often used to emphasize motion, and a popular technique is a squash and stretch in which a body is stretched tangential to its trajectory. they use a robust motion estimation technique in paper [44] for camera motion and use paper [45] for Harris interest points detection in video frames. they calculate contour motion by a linear conformal affine transform (LCAT) in the image plane. Like camera motion correction, interest points [45] are identified within the tracked object. there are many cases in which point correspondences can not be found such as small feature areas. As a result, coloured markers may be attached to the subject and later be removed automatically. they use tracked features mutual occlusion over time to determine partial depth ordering.

As mentioned in paper [45], The trajectory of any object can be defined by its centroid trajectory. Due to a fixed common basis for each frame, paper can estimate the velocity and acceleration of the centroid of the object. At each instant, they build a curvilinear basis frame using the centroid trajectory.As a result, paper can calculate the point in the world frame with equation $x_t(r, s) = \mu(r) + sn(r)$. but this equation fails on collision points. they have a buffer for saving pixels that occlude the tracked object. occlusion pixels are one that has the Euclidean distance between predicted colour and observed colour and exceeds a threshold.

Fig. 9. Examples of cues used in animation videos

An animated object can be defined as a distorted version of the original. it is needed to bend an object along the arc of its centroid as we may get unattractive results if we use only squash and stretch tangential to instantaneous motion. The robustness of the algorithm could be evaluated both with ground truth comparisons for measures such as velocity, as well as processing sequences exhibiting distinctly non-planar motion.

Future work paper [43] suggests many aspects that may be improved such as improving occlusion handling. they believe the future work will examine alternatives for Spatio-temporal video analysis.

7 Performance evaluation

Recent efforts show that combining multiple features, such as static appearance, motion features and acoustic features, can perform event detection.for the case [21], the target was to implement the model in a mobile device with limited resources. In this regard, GoogLeNet is superior to VGG as the former has a learned model with only 40 MBs against a learned model of 533 MBs of VGG.in table [1], we show several approaches with the reported results from other methods using the Pornography-800 dataset.Also In figure 10, we show results for paper [17] compared to many different methods.

Table 1. Results on the Pornography-800 dataset reported with the average performance and standard deviations using the dataset's 5-fold evaluation protocol.

Method	Solution	Accuracy
CNN	Moustafa[46]	94.1 ± 2.0
CNN	Mid-level fusion (OF)	97.9 ± 0.7
CNN	Late fusion (OF)	97.9 ± 1.5
CNN	Static - Fine-tuned	97.0 ± 2.0
CNN	Motion - Optical flow	95.8 ± 2.0
BoVW-based	Moreira et al.[23]	95.0 ± 1.3
BoVW-based	Caetano et al.[47]	92.4 ± 2.0
BoVW-based	Valle et al.[48]	$91.9 \pm NA$

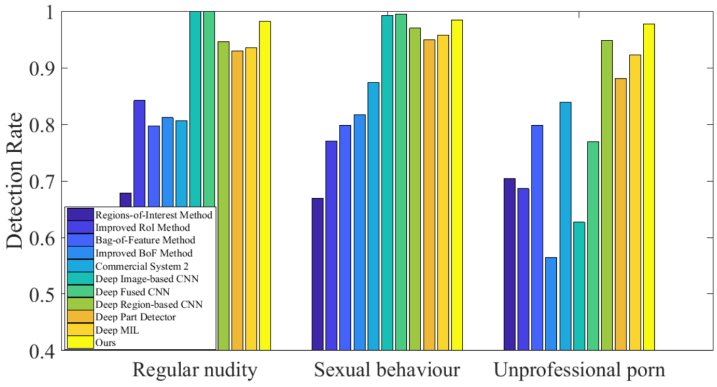

Fig. 10. Comparing results for paper [17] with other different methods.

8 List of data-sets

One important obstacle in this research is the dataset used for DNN learning. Almost all available datasets in the scientific field are for real human videos so it may be needed to search for a good animated cartoons dataset.See table [2].

Table 2. List of data-sets used in Inappropriate content analysis

Name	Datatype	Description
Danbooru2018 [49]	images	A 300 GB Crowd sourced and Tagged 3.33m+ anime images illustration dataset classified into 3 categories: safe, questionable and explicit (pornographic).
Safebooru [50]	images	A 250 MB Crowd sourced and Tagged 2.4m+ anime images with tagged description.
Tagged Anime Illustration [51]	images	A 36 GB Crowd sourced and Tagged anime images with SFW or pornography tagged description.
youtube8m [52]	videos	A big dataset of various types and activities tagged from YouTube.
UCF-101 [53]	videos	A big dataset of various types of human action recognition videos.
kinetics [54]	videos	A dataset of various types of all human action recognition videos.
VSD2014 [55]	videos	A dataset of violent videos or most popular Hollywood films.
Pornography-800 (NPDI-DCC-UFMG) [56]	videos	A dataset of 80 hour videos of 400 pornographic and 400 non-pornographic videos.
Sexualitics [57]	metadata	A dataset of metadata of all video on two big pornographic westies XHamster and Xnxx.

9 Challenges

There are still a lot of challenges in video filtering fields due to the huge variety of types of videos. As an example, one important question to answer is what exactly will we detect in videos, we have a lot of choices (Guns, weapons and alcohols or Blood and injuries or Nudity or Shouting and fighting action recognition or Sex action recognition) or simply detecting all of this. Also, For there are also sub-categories like (Suggestive sex or Explicit Sex or Paedophilia sex). as shown in fig[11], there are different types of cartoon videos and animated graphs like 2-dimensional videos, 3-dimensional videos, stop motion and line drawing. targeted audience age is one of the most important parameters of the filtering process because filtering video for 3-years-old children is different from it for an 18-years-old man watching the same video. the younger the person watching filtered videos, the more content to be filtered in videos.

This research will target 2-Dimensional as it is most popular in the animated video industry and still needs a lot of research. Also, we will target children as an audience in our filtering as this is still a challenging problem in video filtering. Watching porn films wrongly for children can cause a delay in the child's developmental stages, and it can also cause anaphylaxis and autism, and give the child a wrong idea about human relationships, and children may commit crimes. Therefore, video filtering for children and cartoon was of special importance to protect society from the risks that may result from this phenomenon.
Another question to answer is what is the main data needed to be classified to give us the right classification of the video. The answer to the previous question can be one or multiple of this data type (frame images of the video, motion sequence between the image frames, voice emotion, Impression of audience and NLP of the speech in the video), whether you choose one or more of the previous categories but this will affect your accuracy and required work and data-set to accomplish classification task. There are many categories of videos that can be classified working on all of them will need more generalized and trained DNN and more generalized data-set. Videos can be real films or animated cartoons and also can be targeting adults or children. Even Cartoon has many sub-categories like 2-dimensional cartoons, 3-dimensional cartoons and stop-motion cartoons.

centring

Fig. 11. Different types of cartoon Animated images.

10 Conclusion

The association of Deep Learning with the combined use of static and motion information considerably improves pornography detection. Not only over the current scientific state of the art but also over off-the The Deep Learning solution using only static information is already competitive with state-of-the-art

action recognition features.

11 Acknowledgment

This research is under the supervision of Dr Ahmed Bayiomy Zaki. We thank our colleagues from Shoubra faculty of engineering - Benha university who provided insight and expertise that greatly assisted the research, they may not agree with all of the interpretations/conclusions of this paper.

References

1. Donald L Mosher. Pornography defined: Sexual involvement theory, narrative context, and goodness-of-fit. *Journal of Psychology & Human Sexuality*, 1(1):67–85, 1988.
2. Chang Xu, Dacheng Tao, and Chao Xu. A survey on multi-view learning. *arXiv preprint arXiv:1304.5634*, 2013.
3. Jing Zhao, Xijiong Xie, Xin Xu, and Shiliang Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.
4. Mary B Short, Lora Black, Angela H Smith, Chad T Wetterneck, and Daryl E Wells. A review of internet pornography use research: Methodology and content from the past 10 years. *Cyberpsychology, Behavior, and Social Networking*, 15(1):13–23, 2012.
5. Nitish Srivastava, Elman Mansimov, and Ruslan Salakhudinov. Unsupervised learning of video representations using lstms. In *International conference on machine learning*, pages 843–852, 2015.
6. Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.
7. Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732, 2014.
8. Chao Li, Jiewei Cao, Zi Huang, Lei Zhu, and Heng Tao Shen. Leveraging weak semantic relevance for complex video event classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3647–3656, 2017.
9. Devendra Singh Sachan, Umesh Tekwani, and Amit Sethi. Sports video classification from multimodal information using deep neural networks. In *2013 AAAI Fall Symposium Series*, 2013.
10. Hehe Fan, Zhongwen Xu, Linchao Zhu, Chenggang Yan, Jianjun Ge, and Yi Yang. Watching a small portion could be as good as watching all: Towards efficient video classification. In *IJCAI*, volume 2, page 6, 2018.
11. Zhongwen Xu, Yi Yang, and Alex G Hauptmann. A discriminative cnn video representation for event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1798–1807, 2015.

12. Claire-Hélène Demarty, Cédric Penet, Markus Schedl, Ionescu Bogdan, Vu Lam Quang, and Yu-Gang Jiang. The mediaeval 2013 affect task: violent scenes detection. In *MediaEval 2013 Working Notes*, page 2, 2013.
13. Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013.
14. Nevenka Dimitrova and Radu Jasinschi. System for parental control in video programs based on multimedia content information, February 3 2015. US Patent 8,949,878.
15. Kamrun Nahar Tofa, Farhana Ahmed, Arif Shakil, et al. *Inappropriate scene detection in a video stream.* PhD thesis, BRAC University, 2017.
16. Wei Qi, Lie Gu, Hao Jiang, Xiang-Rong Chen, and Hong-Jiang Zhang. Integrating visual, audio and text analysis for news video. In *Proceedings 2000 International Conference on Image Processing (Cat. No. 00CH37101)*, volume 3, pages 520–523. IEEE, 2000.
17. Xin Jin, Yuhui Wang, and Xiaoyang Tan. Pornographic image recognition via weighted multiple instance learning. *IEEE transactions on cybernetics*, 2018.
18. Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.
19. Huicheng Zheng and Mohamed Daoudi. Blocking adult images based on statistical skin detection. *ELCVIA Electronic Letters on Computer Vision and Image Analysis*, 4(2):1–14, 2004.
20. Antonia Quadara, Alissar El-Murr, and Joe Latham. *The effects of pornography on children and young people.* Melbourne: Australian Institute of Family Studies, 2017.
21. Mauricio Perez, Sandra Avila, Daniel Moreira, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, 230:279–293, 2017.
22. Ivan Laptev. On space-time interest points. *International journal of computer vision*, 64(2-3):107–123, 2005.
23. Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. Pornography classification: The hidden clues in video space–time. *Forensic science international*, 268:46–61, 2016.
24. Quoc V Le, Will Zou, Serena Yeung, and Andrew Y Ng. Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. 2011.
25. Theodoros Giannakopoulos, Dimitrios Kosmopoulos, Andreas Aristidou, and Sergios Theodoridis. Violence content classification using audio features. In *Hellenic Conference on Artificial Intelligence*, pages 502–507. Springer, 2006.
26. Sunanda Das, Amlan Sarker, and Tareq Mahmud. Violence detection from videos using hog features. In *2019 4th International Conference on Electrical Information and Communication Technology (EICT)*, pages 1–5. IEEE, 2019.
27. G Sakthivinayagam, R Easawarakumar, A Arunachalam, and M Pandi. Violence detection system using convolution neural network. *SSRG Int. J. Electron. Commun. Eng.*, 6:6–9, 2019.

28. Al-Maamoon R Abdali and Rana F Al-Tuma. Robust real-time violence detection in video using cnn and lstm. In *2019 2nd Scientific Conference of Computer Sciences (SCCS)*, pages 104–108. IEEE, 2019.
29. Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.
30. Qichao Xu, John See, and Weiyao Lin. Localization guided fight action detection in surveillance videos. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 568–573. IEEE, 2019.
31. Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. Flownet 2.0: Evolution of optical flow estimation with deep networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2462–2470, 2017.
32. Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015.
33. Wei Song, Dongliang Zhang, Xiaobing Zhao, Jing Yu, Rui Zheng, and Antai Wang. A novel violent video detection scheme based on modified 3d convolutional neural networks. *IEEE Access*, 7:39172–39179, 2019.
34. AS Keçeli and AYDIN Kaya. Violent activity detection with transfer learning method. *Electronics Letters*, 53(15):1047–1048, 2017.
35. Sarthak Sharma, B Sudharsan, Saamaja Naraharisetti, Vimarsh Trehan, and Kayalvizhi Jayavel. A fully integrated violence detection system using cnn and lstm. *International Journal of Electrical & Computer Engineering (2088-8708)*, 11(4), 2021.
36. François Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.
37. Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25:1097–1105, 2012.
38. Rohit Halder and Rajdeep Chatterjee. Cnn-bilstm model for violence detection in smart surveillance. *SN Computer science*, 1(4):1–9, 2020.
39. Tim Tate. *Child pornography: An investigation*. Trafalgar Square, 1990.
40. Krishna Somandepalli, Naveen Kumar, Tanaya Guha, and Shrikanth S Narayanan. Unsupervised discovery of character dictionaries in animation movies. *IEEE Transactions on Multimedia*, 20(3):539–551, 2018.
41. Dumitru Erhan, Christian Szegedy, Alexander Toshev, and Dragomir Anguelov. Scalable object detection using deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2147–2154, 2014.
42. Christian Szegedy, Scott Reed, Dumitru Erhan, Dragomir Anguelov, and Sergey Ioffe. Scalable, high-quality object detection. *arXiv preprint arXiv:1412.1441*, 2014.
43. John P Collomosse, David Rowntree, and Peter M Hall. Video analysis for cartoon-like special effects. In *BMVC*, pages 1–10, 2003.
44. Philip Hilaire Sean Torr. *Motion segmentation and outlier detection*. PhD thesis, University of Oxford England, 1995.
45. Chris Harris. Stephens. a combined corner and edge detector. In *Alvey Vision Conference*, pages 147–152, 1988.
46. Mohamed Moustafa. Applying deep learning to classify pornographic images and videos. *arXiv preprint arXiv:1511.08899*, 2015.

47. Carlos Caetano, Sandra Avila, William Robson Schwartz, Silvio Jamil F Guimarães, and Arnaldo de A Araújo. A mid-level video representation based on binary descriptors: A case study for pornography detection. *Neurocomputing*, 213:102–114, 2016.
48. Eduardo Valle, Sandra de Avila, Antonio da Luz Jr, Fillipe de Souza, Marcelo Coelho, and Arnaldo Araújo. Content-based filtering for video sharing social networks. *arXiv preprint arXiv:1101.2427*, 2011.
49. Danbooru2018, 2018. `https://www.gwern.net/Danbooru2018`.
50. Safebooru. `https://www.kaggle.com/phryxia/safebooru-2018`.
51. Tagged anime illustration dataset. `https://www.kaggle.com/mylesoneill/tagged-anime-illustrations`.
52. youtube8m dataset. `https://research.google.com/youtube8m/`.
53. Ucf-101 dataset. `https://www.crcv.ucf.edu/research/data-sets/human-actions/ucf101/`.
54. kinetics dataset. `https://deepmind.com/research/open-source/open-source-datasets/kinetics/`.
55. Vsd2014 dataset, 2014. `https://www.technicolor.com/dream/research-innovation/violent-scenes-dataset`.
56. `https://sites.google.com/site/pornographydatabase/`.
57. Sexualitics dataset, 2014. `http://sexualitics.github.io/`.