Journal of Engineering Science and Technology
Vol. 17, No. 3 (2022) 2104 - 2119

A DEEP-LEARNING FRAMEWORK FOR ACCURATE AND ROBUST DETECTION OF ADULT CONTENT

KUSRINI KUSRINI[1,*], ARIEF SETYANTO[1], I. MADE ARTHA AGASTYA[2], HARTATIK HARTATIK[2], KRISHNA CHANDRAMOULI[3], EBROUL IZQUIERDO[3]

[1]Magister of Informatics Engineering, Universitas AMIKOM Yogyakarta, Jl. Ringroad Utara Condong Catur Depok Sleman, Yogyakarta, Indonesia
[2]Computer Science Faculty, Universitas AMIKOM Yogyakarta, Jl. Ringroad Utara Condong Catur Depok Sleman, Yogyakarta, Indonesia
[3]Multimedia and Vision Research Group, School of Electronic Engineering and Computer Science, Queen Mary, University of London, Mile End Road, London, E1 4NS, UK
*Corresponding Author: kusrini@amikom.ac.id

Abstract

Video streaming services has dominated the growth of Internet traffic, which accounts for more than 82% of global network traffic. As videos represent a powerful medium for user engagement, there has been an exponential growth in the different forms of video content being generated and distributed. Such a large diversity of content includes educational, gaming, historical, and entertainment among others. While the positive impact of these video services available through Internet has benefitted a large number of citizens, it is evident that the same platform has been subjected to misuse for the propagation of explicit content not suitable for children. Therefore, it is crucial to develop automated solutions that can successfully filter such content to protect children and vulnerable individuals. Over the last few years, deep learning algorithms have achieved high accuracy in object recognition in comparison to the statistical algorithms trained on hand-crafted features. The deep-learning algorithms has demonstrated the ability to automatically select most representative features to be extracted from visual representation of objects. While this technology has been successfully applied to many important computer vision problems, its use for the classification and filtering of adult content has not been fully explored yet. Addressing this challenge, the research presented in this paper reports a deep-learning framework that exploits spatio-temporal and visual features of video sequences for efficient and effective detection of adult content. The proposed network architecture aims at harvesting information from two important aspects of video content: spatial self-learned features and cues from temporal redundancies and dynamics in videos. First, a Convolutional Neural Network (CNN) architecture based on Inception-v3 is used to model and learn the spatial video features. This CNN architecture builds the basis of the proposed deep-learning framework. Temporal characteristics are then modelled through a long-term short memory approach that correlates information from subsequent frames in the processed video clip. The proposed approach has been validated against an openly available dataset, which includes a large category of adult content and other video sequences that are hard to discriminate, e.g., beach footage, swimming and wrestling. The accuracy of the proposed approach reaches 97.4%, while the recall is 97%, making it highly suitable for practical applications.

Keywords: Adult content, Convolutional neural network, NPDI dataset, Video classification, LSTM network.

1. Introduction

Filtering sensitive media, including pornographic, violent and gory videos distributed through Internet services, has growing importance, due to the pervasive presence of interconnectivity for people of all ages. Among the sensitive media types, pornography is often considered one of the most disturbing forms of content [1]. The pervasiveness of pornographic videos on the Internet is leading to significant societal problems, whose consequences for future generations cannot be underestimated. Indeed, it has been reported that children as young as seventh grade are already accessing adult content, while young teenagers also frequently watch such content [2].

Researchers from psychology and related social science discipline agree that the open availability of adult content is having significant negative consequences in childhood development [3]. According to Short et al. [4], the consumption of adult content reduces the ability of the brain to give an appropriate response to stimuli and this effect is largely augmented in young people. The challenge is further compounded due to the exponential growth of online video streaming services that are particularly targeted at young adults. A range of applications has increased societal interest in the problem, e.g., detecting inappropriate behavior via surveillance cameras; or curtailing the exchange of sexually charged instant messages, also known as "sexting", by minors. In addition, law enforcers may use pornography filters as a first sieve when looking for child pornography in the forensic examination of computers, or internet content.

The critical need for classifying and identifying such content is further reinforced by the release of commercial tools and solutions such as Amazon Rekognition [5], Explicit Content Recognition [6] that offers support for the societal harm that is perpetrated upon the young adults. A report by the ExtremeTech [7] technology site suggests that 30% of all Internet traffic is associated with pornography. Thus, the use of manual curation and/or moderation of increasingly large volume of content has been deemed impossible by Internet service providers.

Early papers addressing this problem can be traced back almost two decades ago. Relevant techniques include the use of IP-based filtering or blocking. Such approaches rely on knowledge about websites previously identified as hosts of adult content [8]. Other early techniques reported in the literature include the analysis of textual analysis and topic models [9], as well as the use of semantic technologies for identifying adult content by exploiting metadata associated with the video of concern [10]. A more detailed review of early techniques for adult content classification and filtering is given in the next section. More recently, computer vision approaches exploiting statistical colour model-based filters have been proposed Yin et al. [11]. Unfortunately, techniques based on statistical models for quantifying skin colour led to several false positives in images containing mostly brown colour as 10%, 23.5%, and 5.2% false-positive occurrences reported in [12, 13].

Conventional machine learning approaches consist of two main processes: feature extraction and classification. Feature extraction is usually handcrafted by engineers according to the target classification problem. Then, the extracted feature vectors are classified as basic k-means or more advanced Support Vector Machines (SVM). In these cases, the accuracy of recognition vastly depends on the engineer's experience and the suitability of exploited features according to a particular problem.

With the advent of advanced machine learning models, specifically deep learning, an important tool became available to tackle any object classification problem. Consequently, better results can be expected since large amounts of data, i.e., big data, is also available to automatically train deep-learning networks for most classification problems including adult content recognition. To the best of our knowledge, one of the first attempts to exploit deep-learning networks for the detection and filtering of adult content has been reported by Nurhadiyatna et al. [14], while subsequently additional publications have been reported in the literature. This research aims at embedding deep convolutional network models in hand-held devices to block detected adult content. In this case only pornographic still images and keyframes extracted from the videos are processed. Inception v3 network was developed based on convolutional operations, which rely on feature extraction using spatial domain. Therefore, our research focused on the need for the development of visual analysis for the detection of pornographic images and does not consider the temporal correlations between the image sequences. Additionally, the comparative analysis performed on the NDPI dataset, reports on the categorization results using visual features (and not using temporal features)

Exploiting recent developments in machine learning technology, specifically deep learning, a generic network architecture is proposed in this paper for the recognition of adult video content. The aim is to exploit the visual and temporal characteristics of video sequences. The proposed network architecture is based on a classic Convolutional Neural Network (CNN) trained to learn spatial-visual features. On top of this "fundamental" learning network, temporal characteristics are modelled through a Long-Term Short Memory approach, which exploits information from subsequent frames in the processed video clip.

Several state-of-the-art CNN frameworks were initially tested to find the most suitable one for our target application. Compared to the previous versions (Inception v1 and v2), the Inception v3 network structure uses a convolution kernel splitting method to divide large volume integrals into small convolutions. For example, a 3x3 convolution is split into 3x1 and 1x3 convolutions. Through the splitting method, the number of parameters can be reduced; hence, the network training speed can be accelerated while the spatial feature can be extracted more effectively. At the same time, Inception v3 optimizes the Inception network structure module using three different size area grids (35×35, 17×17, and 8×8) [15] As a consequence of this initial analysis, the Inception V3 framework was chosen as the most suitable machine learning model for our classification problem.

Among the several types of deep-learning network architectures, three network models have been widely considered for the extraction of deep-learning features namely VGGNET [16], RESNET [17] and Inception [18]. Following a quantitative assessment of the computational complexity, the use of Inception-v3 networks was adopted for the visual feature extraction. A detailed technical analysis of the Inception-v3 is presented in Section 3 for completeness.

The main contributions of the paper include:

- Summarize the most relevant algorithms reported in the literature for more than two decades for the detection, classification, and filtering of adult content.
- Study of deep-learning feature extraction components for modelling the visual characteristics of explicit and non-explicit content

- Design a framework for computationally deep-learning network for the classification of adult content
- Implement, validate and benchmark the proposed deep-learning framework against the reported techniques in the literature
- Recommend upon the quality of the features and the proposed approach towards improving the classification accuracy for achieving high recall in trade off with precision.

The paper is structured as follows. In Section 2, a survey of the most relevant research works addressing the adult classification problem is presented. In this section a review of deep learning frameworks and relevant architectures are also outlined. In Section 3, the proposed network architecture is presented in detail. Section 4 describes two different approaches for adult content recognition using the same framework but exploiting temporal relationships in two different ways. Section 5 describes selected results from an exhaustive evaluation. The paper closes with conclusions drawn from this research work and potential future extensions in Section 6.

2. Literature Review

One of the earliest publications summarizing the state of explicit video classification contained a review of 46 articles in which the authors noted the definition of adult content as "any sexually explicit material with the aim of triggering sexual arousal or fantasy" [4]. The first attempt at the classification of adult content was conducted by Duan et al. [12] who postulated the hypothesis that explicit content would include nudity and thus to detect skin colour. The presented approach included a colour matching descriptor for the identifying and quantification of the amount of skin that was visible in the image. A similar approach has been adopted by Jeong et al. [19] but the experimental results were performed on a total of 17,400 images mixed with explicit nudity and non-adult content. Subsequently, the use of pixel-based approaches for modelling region of interest (ROI) based on the skin-detection algorithm has been reported in [13, 20]. A range of statistical metrics such as colour moment, histogram, and Gray Level Co-occurrence Matrices (GLCM) have been successfully validated. Furthermore, following the development of machine learning tools and algorithms such as Support Vector Machine (SVM) using Radial Basis Function (RBF) kernel or Neural Network (NN) the research on feature extraction and modelling the characteristics of visual and motion parameters gained popularity.

Following the success of the feature extraction and machine learning techniques resulting the improving accuracy of the classification algorithm, the problem of explicit content classification has been treated as an image classification problem [21] based on the use of local descriptors such as HueSIFT descriptor and Bag of Visual Words (BoVW) feature vectors trained using SVM linear classifier. BossaNova to enrich BoVW has been proposed by Avila et al. [22]. Zhang et al. [9] and Zhuo et al. [23] proposed a Skin ROI detector based on the extracted features containing HueSIFT, Texture and Intensity. Despite the success of the skin-based ROI modelling of the adult content, the use of handcrafted features extracted from a small-scale dataset has been found to be highly correlated to the specific problem and less generalized compared to the deep learning approach. The overfitting of the handcrafted features is a limitation, which has been overcome in the proposed approach.

The image classification changed direction after the successful deep learning algorithm [24] on winning ImageNet Large Scale Visual Recognition Competition (ILSVRC). The deep learning weight that is trained on the ImageNet dataset can be transferred to other domains such as pornographic images [20]. The experiment using VGG-16 architecture in NPDI pornographic video keyframes has shown strong performance with the accuracy reaching 93.8%. An overview of the various techniques reported in the literature, in which the adult content classification was treated as an image classification is presented in Table 1.

Table 1. Explicit image content classification results.

Ref.	Dataset Volume*	Feature Extraction		Visual Vocabulary Clustering	Classifier	Acc. (%)
		Detector	Descriptor			
[12]	760/885	Colour Distribution	Saturation	-	SVM (RBF)	80.7
[19]	4,600/ 13,000	Shape Information	Skin Likelihood	-	SVM (RBF)	96.35
[20]	812/ 16,488	Skin ROIs	Colour Moment, Histogram, GLCM	-	SVM (RBF)	90
[13]	508/482	Skin ROIs	Colour Moment, Histogram, GLCM	-	Adaboost	94.8
[21]	90/90	SIFT blobs	HueSIFT	K-means	SVM (Linier)	84.6
[9]	4,000/ 4,000	Skin ROIs	Colour, Texture, Intensity	K-means	SVM	90.9
[23]	8000/11,000	Skin ROIs	ORB	K-means	RBF	93
[25]	6,387/10,340		VGG-16		FCL	93.8

***Number of pornographic images / Number of non-pornographic images of confusing nature*

Following the increasing availability of large-scale video datasets, the research focus had been shifted from processing images to handle video streams. Addressing the challenge of processing videos and extracting video centric features was reported in a sequence of publications which has been summarized in Table 2.

Table 2. Adult (video) content classification using statistical machine learning algorithms.

Ref.	Dataset Volume*	Feature Extraction		Visual Vocabulary		Classifier	Acc. (%)
		Feature Detector	Feature Descriptor	Feature Clustering	Representation		
[22]	400/400	Regular Grid	HueSIFT	K-means	BOSSA	SVM (χ)	87.1
[26]	400/400	STIP blobs	STIP	Random	-	SVM (Linear)	91.9
[27]	400/400	Colour-STIP Blobs	STIP	Random	-	SVM (Linear)	91.0
[28]	400/400	Regular Grid	HueSIFT	K-means	BossaNova	SVM (χ2)	89.5
[29]	400/400	Regular Grid	Binary descriptors	K-medians	BossaNovaVD	SVM (χ2)	90.9
[30]	400/400	Regular Grid	Binary descriptors	K-medians	BossaNovaVD	SVM (χ2)	92.4
[31]	400/400	3D Hessians Blobs	TroF	GMM	Fisher Vector	SVM (Linear)	95.0

**Number of pornographic images / Number of non-pornographic images of confusing nature*.

Following recent reports on the use of deep-learning networks, there has been several attempts by researchers towards the use of advanced network architectures for improving the video classification models for explicit content. In this regard, one of the earliest Convolutional Neural Network approaches was reported in [32] in which the classification of the explicit content was addressed using AlexNet [24], GoogLeNet [18], and combination of both network architectures. The network structure of the combined architecture has been referred to as AGNet and the reported accuracy for the classification is 94.1%. Similarly, a combination of CNN and RNN was reported in the design of ACORDE-101 [33] network model which has been reported to achieve an overall accuracy of 95.6%. In one of the more recent publications, Perez et al. records an overall accuracy of 97.9 % as their best performance on NPDI dataset [1]. The proposed approach includes two main features namely (i) the use of spatial features extracted from the selected frame and (ii) temporal features extracted from the video based on the specification of MPEG motion vector and optical flow. Finally, the use of 3D CNN modelled to consider both the spatial and temporal data has been reported to deliver an overall accuracy of 95% [34]. A summary of the various deep learning algorithms which has been reported in the literature is presented in Table 3. While the publications report on the overall accuracy of the proposed algorithms, the individual evaluation metrics such as precision, recall and F1 Score have not been reported.

Table 3. Adult content classification using deep-learning network models.

Ref.	Input	Architecture	Accuracy (%)
[32]	Image	AlexNet + GoogLeNet	94.1±2.0
[33]	Image	Resnet101 + LSTM	95.6±1.0
[1]	Image	GoogLeNet	97.0±2.0
[1]	Optical Flow	GoogLeNet	95.8±2.0
[1]	Image + Optical Flow	GoogLeNet	97.9±0.7
[35]	Image	3D CNN	95±1.7

3. Deep Learning for Adult Content Recognition

The recent success of deep learning algorithms for object detection in image repositories and video sequences can be attributed to the automatic extraction of visual features that are able to uniquely distinguish object categories. One of the seminal works presented in 2014, trained the network to classify 1000 different classes as formalized by ILSVRC 2014 dataset [35]. However, the problem of detecting explicit content is further compounded by the lack of generalization within the dataset in which the objective is to detect people performing sexual acts.

To highlight the challenge, an overview of keyframe samples is presented in Fig. 1. The semantic interpretation of the dataset is varied and includes video sequences of non-explicit content which closely correlates to the visual representation of adult content. Considering the appearance of human skin as an indicator there are several examples of non-explicit content such as boxing, swimming, fashion photoshoot and video recording which are to be ignored by the filtering process. Therefore, the objective of the research outcome presented in the paper aims to generalize and propose a neural network architecture that can learn and optimize the weights of neuron interconnections leading to improved classification accuracy of explicit content.

Fig. 1. Overview of the dataset for content classification between explicit vs. non-explicit content.

The proposed framework for the explicit video content classification is presented in Fig. 2. The framework is divided between the training and the testing component. The training pipeline of the framework includes three main steps namely (i) feature extraction; (ii) video signature generation and (iii) model training for binary classification. Similarly, the testing pipeline includes the definition of frame buffer allocation to support the real-time processing of the ingested video streams into the pipeline, followed by the video signature generation. Finally, the trained models stored in the repository are used to swiftly classify the video sequences either as explicit content or not. In the rest of the section each of the processing pipeline components is further elaborated to highlight the research contribution.

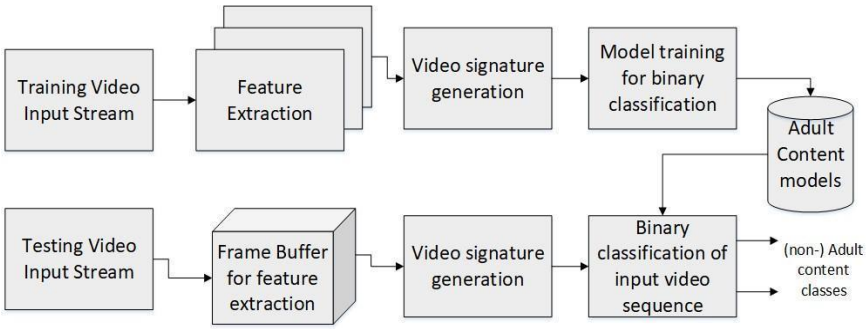

Fig. 2. Proposed architecture for the adult content classification.

3.1. Feature extraction based on Inception-v3

Video sequences contain highly redundant and closely correlated content among consecutive frames. Considering the entire frames throughout video will not improve the capability of the classifier to recognize the explicit content while overburdening the computing cost as a consequence. To achieve this objective, we have adopted the process of sampling frames within the video sequences. Selecting an adequate number of frames to represent the entire frames is one of the important challenges to consider. In this research, a varied number of frames are selected and evaluated against the recognition rate. Due to the requirements that the classifier needs a uniform number of inputs, we have opted to take periodic samples with a constant number of keyframes samples on each experiment. In order to get the

uniform number of keyframes for every video sample regardless it's length, a skip is calculated for each video in regard to Eq. (1)

$$s = \frac{1}{n} \tag{1}$$

where *s* is the skip (distance between two consecutive keyframes) and *n* is the number of selected keyframes in each video. n values are varied at (2, 4, 8, 16, 32, 64 and 128).

The first component in the development of the explicit content classification algorithm is the feature extraction process. One of the critical requirements for the feature extraction process is to deliver a high degree of distinguishability upon which the machine learning algorithms can be trained for achieving classification of input data sequences. While several research approaches based on the statistical quantification of pixel-based skin modelling techniques have been reported in the literature, the metric is deemed unsuitable for achieving high classification accuracy due to the nature of the dataset complexity. Therefore, following the success of deep-learning models that have been reported to deliver a high-degree of accuracy when applied to multi-class classification, the Inceptionv3 [18] network has been selected for performing feature extraction. The network includes 3 traditional inception modules at the 35×35 with 288 filters each. These modules are further reduced to a 17×17 grid with 768 filters using the grid reduction technique. This is followed by 5 instances of the factorized inception modules. The overall network structure is further reduced to an 8×8×1280 grid with the grid reduction technique. At the coarsest 8×8 level, two Inception modules are designed with a concatenated output filter bank size of 2048 for each tile. The integration of the network within the proposed framework is presented in Fig. 3. One of the key innovations proposed in the paper is to establish the visual correspondence between the keyframes extracted from the video sequence. For each of the keyframes selected, a set of features are extracted using the Inceptionv3 network that results in the final output of 1 x 2048 feature vector for each keyframe from the pretrained Inception-v3 network to be further processed.

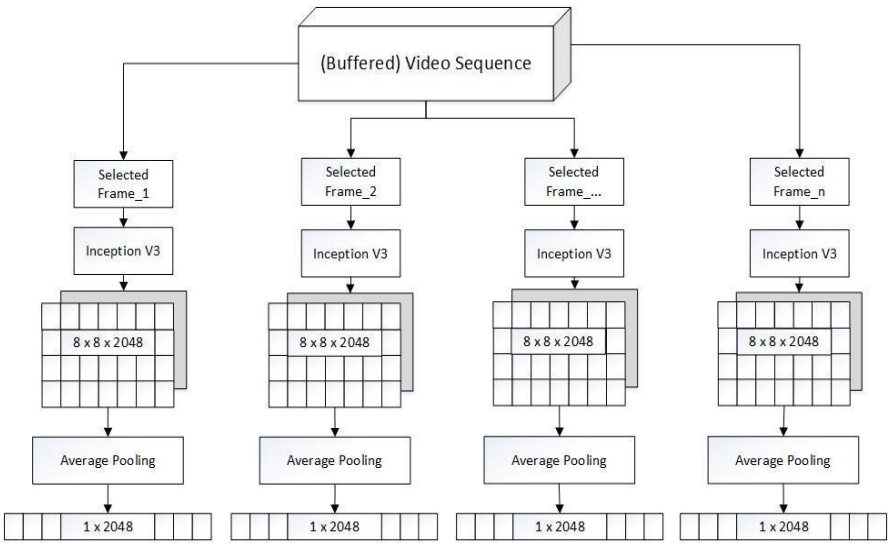

Fig. 3. Inception-v3 feature extraction framework.

The construction of the temporal dependencies within the selected video sequence is modelled using Fully Connected Layer (FCL) and the Long Short Term Memory (LSTM) network architectures and is further detailed in the rest of the section.

3.2. Model training based on FCL, SVM, KNN and RF

The extracted features from the previous step are further processed to train a fully connected 2-layer network for achieving binary classification of explicit content vs non-explicit content. An overview of the network implementation is presented in Fig. 4. The input features considered for training the network includes the generation of the video signature which results from the weighted combination of the features extracted from the keyframes. The dimensionality of the input feature vector size is (1 x (n x 2048)), which is an aggregation of the features from key-frames. Where n is the number of selected keyframes in a single video. In this approach, the features from n keyframes are flatted into a single array to fit the input of FCL classifier. The network is trained to capture the semantics of a short video sequence clip based on the visually correlated features of a short video sequence represented based on the keyframes extracted. The nature of strong association between the visual features that is extracted over a period from the video sequences, is postulated to result in highly robust classification of explicit and non-explicit content. The high dimensionality of the extracted features processed through the FCL satisfies the need for distinguishability between overlapping visual characteristics between the explicit and non-explicit content.

The overall implementation of the FCL network is achieved in TensorFlow using Keras library. The training of the video signature includes the visual profile of the keyframes extracted and thus, the cumulative temporal association of the feature vector of length (n×2048) is used for training the fully connected layer network.

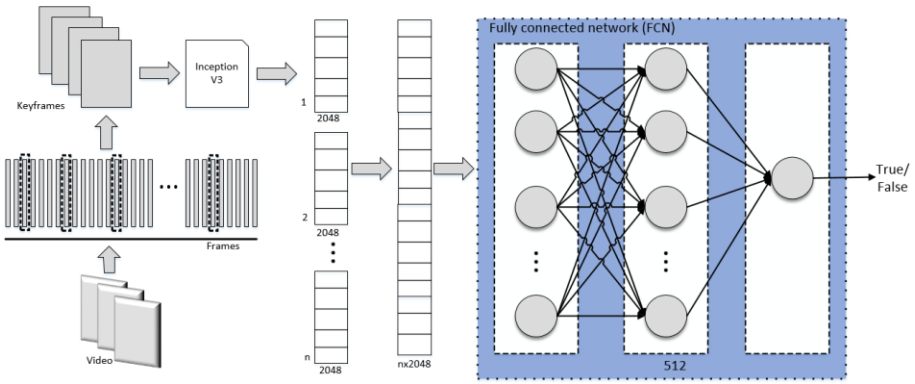

Fig. 4. Fully connected layer model for binary classification.

Figure 4 shows the overall task from keyframe selection at the beginning, feature extraction of each keyframe, flatten the features and classification task throughout the FCL. During training, the weight on the FCL will be adjusted. The training parameters utilized include a learning rate of 1x10-5 and epoch of 30. The loss function has been calculated based on binary cross-entropy. Support Vector Machine (SVM), K nearest neighbor (K-NN) and Random Forest (RF) are three well known classifiers. This research compares those three algorithms to replace

the FCL to finally classify the flattened video features. SVM, KNN and RF replace the blue block of FCL in Fig. 4.

3.3. Model training based on LSTM

The second approach adopted for the construction of video sequence association based on the visual features extracted from the Inception-v3 network utilizes the LSTM network model with two hidden layers and a dense output layer. The input layer uses a 128-sequence formation which is interconnected to a 64-cell LSTM layer, followed by the dense layer with 64 input units and resulting the final output providing the binary classification for the explicit and non-explicit content In comparison to the FCL network, the temporal associations between the keyframes are represented through the LSTM cells in which each of the cells in the input layer are activated by the sigmoid and tanh function, followed by the sequence operation. The overall network architecture as proposed for modelling the video sequence is presented in Fig. 5.

The overall implementation of the LSTM network was carried out using TensorFlow and Keras library. The training of the network was achieved for a range of input keyframes from 2 to 128. However, the LSTM unit has been maintained at 128.

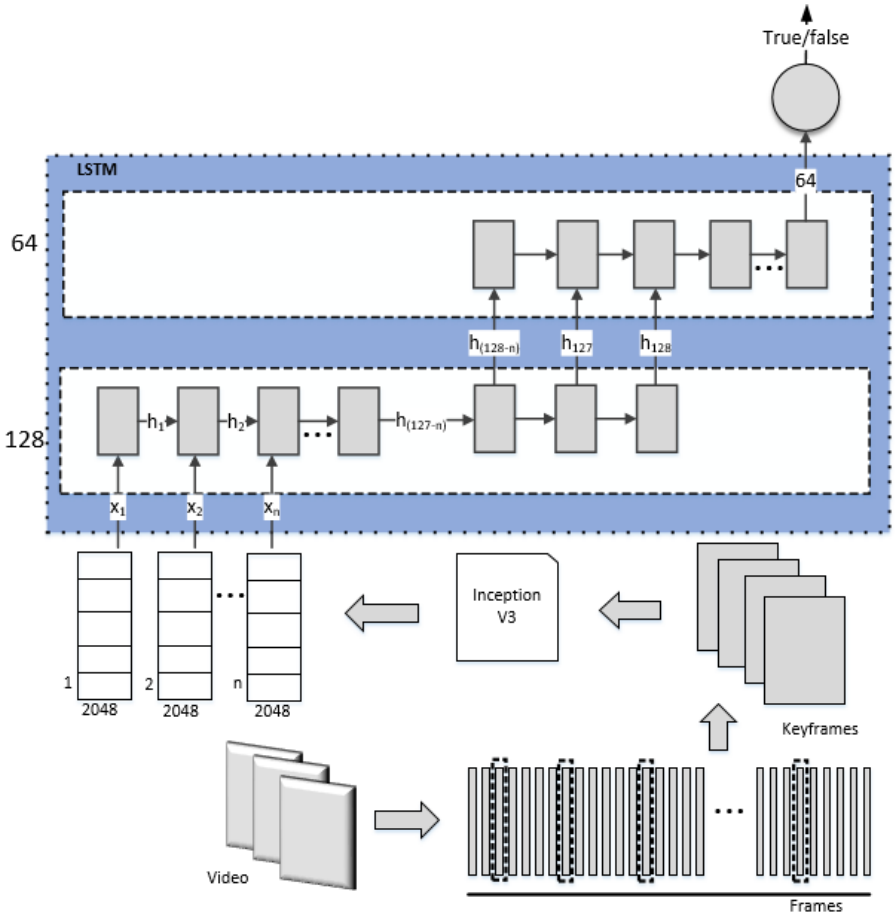

Fig. 5. LSTM layer for binary classification.

Figure 5 shows the overall task from keyframe selection at the beginning, feature extraction of each keyframe using pre-trained Inception V3. The first and second step is identical to that of the FCL approach. In order to maintain the temporal relationship among keyframes, every 2048 features for each keyframe are treated individually as the input of the first layer of the LSTM network. The first layer of LSTM consists of 128 cells and only n (number of keyframes) gets an input. Out of 128 cells, only n to a maximum of 64 cells produce an output to the second layer of LSTM with a smaller data size at 128 bytes. The final output of the second layer of LSTM is 64 bytes passed to the last dense layer to be classified into binary class. During training, the weight on the first and second layer of LSTM as well as the dense layer is adjusted. The training of the overall network was carried out using a learning rate of 1x10-5, training epoch of 30 with loss function computed using binary cross-entropy. The network uses Adam optimizer function for updating the weights of the network.

4. Results and Discussion

The experimental evaluation carried out for the validation of the proposed video sequence classification framework for non-explicit and explicit content on the NPDI dataset, which is to the best of our knowledge represents a standard benchmark against which several algorithms have been proposed. The dataset consists of 800 videos in total amounting to approximately 77 hours of the video footage. The content classification in the dataset includes 3 classes namely (i) Porn (also referred in the paper as explicit content); (ii) non-porn (easy) referred to as non-explicit content and (iii) Non-porn (difficult) also referred to as non-explicit content. A summary of the content distribution across these three classes is presented in Table 4.

Table 4. NPDI dataset outline.

Class	No. of Videos	Hours
Porn	400	57
Non-porn ("easy")	200	11.5
Non-porn ("difficult")	200	8.5
All videos	800	

A total of 14 runs were carried out for each of the two proposed network models to evaluate the impact of temporal correlation between the keyframes for successfully classifying the explicit and non-explicit videos. The number of keyframes extracted from each video is one of the hyper parameters that is configured within the feature extraction algorithm component. The experimental setup was carried for keyframes extracted between 2 to 128 from each of the video sequence. The objective of the evaluation is to validate the quality of the network training as presented in Section 3, along with the evaluation of the video signature extracted from the temporal correlation between the keyframes. The two quantitative metrics are evaluated against the NDPI dataset. An overall evaluation summary of both proposed frameworks is presented in Table 5. The results are categorized into precision, recall and F1-score and each of the statistical quantity is analysed using the standard deviation obtained for each run of the experiments carried out.

Following an analysis of the results presented in Table 3 and Fig. 6, it is visible that the classification accuracy improves when the number of selected keyframe

increases from 2 to 16, and inclusion of additional keyframes results in decreased performance of the algorithm. The accuracy results from both proposed architectures deliver high accuracy at 16 selected keyframes. The performance reduction of the proposed framework based on FCL gradually stabilizes at 96% overall accuracy. The performance of the architecture for LSTM based video sequence classification algorithm indicates an improvement of the overall performance at both 16 keyframes and 128 keyframes selection. The overall computational efficiency of the proposed architecture is determined at the selection of 16 keyframes per video sequence. The choice of 16 keyframe-based association for the temporal correlation leads to computationally efficient implementation of the overall architecture, including the amount of memory required to create a temporary buffer for storing the video stream sequences. A total of 2.211 seconds is required for the feature extraction of the 16 keyframes selected and the binary classification is achieved in under 10ms, leading to an overall requirement of 2.212 seconds for classifying the video sequences of length 12.48 seconds. The use of computationally efficient models enables a wide range of application for the proposed framework to be integrated within online video streams and enable faster than real-time computation for the classification of explicit and non-explicit content.

Table 5. Experimental results analysis.

Scenario	Number of key Frames	Accuracy	Precision	Recall	F1-score
Inception-v3 features classified with FCL network, optimized using softmax	**2**	90.9±2.13	90±3.74	92±1.41	91±2.24
	4	94.2±1.21	95.2±3.42	93.6±2.88	94.2±1.1
	8	95.5±1.95	95.8±3.03	95.4±3.51	95.4±2.3
	16	96.6±0.96	96.6±2.07	96.6±1.82	96.4±1.14
	32	96.3±0.46	95.8±1.64	96.6±2.3	96±0.71
	64	96.5±1.58	96.8±2.17	96±3.08	96.4±1.82
	128	96.6±1.38	96±2.74	97.4±1.95	96.6±1.52
Inception-v3 features classified with LSTM network using softmax	**2**	91.4±1.77	92±5.87	91.4±5.18	91.6±1.52
	4	95±0.97	94.8±2.95	95.6±2.19	95±1
	8	97.1±1.03	96.8±2.59	97.6±1.82	97.2±1.3
	16	97.4±1.11	97.8±2.17	97±2	97.4±0.89
	32	97.3±1.28	96.6±2.7	98.2±1.92	97.2±1.3
	64	97±1.03	97±2.83	97.2±2.28	97±1.22
	128	96±0.95	95.2±3.11	97±2.35	96±0.71

The overall performance of the proposed architecture clearly outperforms reported algorithms in the literature whose results are benchmarked against the NDPI dataset. As summarized in Table 3, the reported algorithms utilizing deep learning models have been trained upon the visual data for the classification of explicit and non-explicit content. The highest accuracy reported on the NPDI dataset is 97%. Wehrmann et al. [33] presented an approach based on the use of Resnet101 with LSTM and represents the closest design of the network architecture as presented in the paper and the reported overall accuracy of the architecture is 95.6%. Similarly, Perez et al. [1], have reported an overall accuracy of 97.9% which is 0.4% more than the result presented in this paper. The improved performance reported by Perez et al. [1] is attributed to the use of visual data and optical flow vector for modelling the input video sequences through motion vectors extracted using the MPEG-descriptors. In contrast, the proposed network architectures have effectively exploited the temporal correlation between the keyframes extracted from the video and thus eliminates the need for higher computational resources and

memory requirements. The proposed framework as presented is also suitable for processing real-time video streams and offers an efficient methodology to detect explicit and non-explicit content upon a buffered video stream.

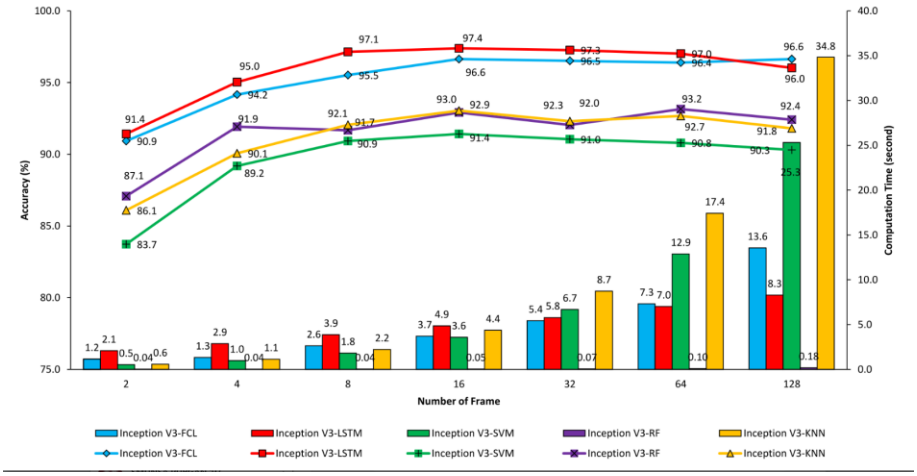

Fig. 6. Accuracy and testing time map against number of frames selected.

Figure 6 shows the result of our experiments, as can be seen the Inception 3 with LSTM generally achieves the best accuracy. The rest of our experiments using KNN, SVM, RF and FCL as the classifier attain lower accuracy in all numbers of keyframes. RF shows superior performance at 64 keyframes compared to SVM and KNN as 93.2% accuracy achieved. The rest of the experiments shows LSTM consistently over performs compared to FCL, SVM, RF and KNN. The fact supports our argument that temporal relationships play an important role in video classification. Figure 6 also shows the fact that the best accuracy achieved at 16 keyframes. It is evident that the experiment with more keyframes did not lead to better recognition rate. Since the visual content of the consecutive frames are highly redundant, adding more frames is meaningless and leads to an increase of processing time without accuracy improvement.

Computation time is an important aspect of classification performance. Figure 6 also presents a comparison of testing time among all experiment settings. The testing set consists of 160 videos, we take total execution time for the entire testing process. Therefore, the execution time to classify a single video at the specified time in Fig. 6 needs to be divided by 160. The best accuracy achieved by InceptionV3-LSTM with 16 selected keyframes at 97,4% needs 4.9 seconds for 160 videos. The execution time for a single video would be at 0.03 second on averages, allowing real time implementation.

5. Conclusions

In this paper, the authors present two approaches for the detection of pornographic content based on inception V3 network. The use of FCL and LSTM networks has been presented in the paper for categorizing the video sequences as containing explicit and non-explicit content. The proposed approaches have been evaluated against the NDPI dataset, which has been widely used in the literature. The

experimental results indicate that, Inception v3-LSTM network outperforms the rest of the algorithms when 8 to 64 keyframes are selected. In contrast, the Inception v3-FCL network delivers higher performance when 128 keyframes have been selected. Both approaches outperform the results presented in the literature. Additionally, the experimental analysis using SVM, RF and KNN also indicate the higher performance of the proposed approach.

The research presented in the paper paves the way forward in the design and development of the pornographic video sequence classification. Despite the high performance of the algorithms presented in the paper, there still exist gaps for improvement which include (i) an extensive archive of content to be used in training; (ii) addressing the notion of content generalization which can be further extended to include a degree of exposure; (iii) parameterized categorization based on the audience age group. Finally, there are research investigations in the field of action recognition, which has resulted in a high degree of success for actions such as running, walking, etc. Such research outcomes could be further extended to analyse visual sequences of repeated actions commonly encountered in the pornographic sequences.

Abbreviations

BoVW	Bag of Visual Words
CNN	Convolutional Neural Network
FCL	Fully Connected Layer
GLCM	Grey Level Co-occurrence Matrices
KNN	K-Nearest Neighbour
LSTM	Long Short Term Memory
RBF	Radial Basis Function
RF	Random Forest
SVM	Support Vector Machin

References

1. Perez, M.; Avila, S.; Moreira, D.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; and Rocha, A. (2017). Video pornography detection through deep learning techniques and motion information. *Neurocomputing*, 230, 279-293.
2. MacLaughlin, K. (2017). The detrimental effects of pornography on small children. Retrieved July 26, 2021, from https://www.netnanny.com/blog/the-detrimental-effects-of-pornography-on-small-children/
3. Quadra, A.; El-Murr, A.; and Latham, J. (2017). *The effects of pornography on children and young people*. Research Report. Australian Institute of Family Studies, Melbourne.
4. Short, M.B.; Black, L.L.; Smith, A.H.; Wetterneck, C.; and Wells, D.E. (2012). A review of internet pornography use research: methodology and content from the past 10 years. *Cyberpsychology, Behavior, and Social Networking*, 15(1), 13-23.
5. AWS, (2017). Detect explicit or suggestive adult content using Amazon Rekognition. Retrieved Sep 20, 2019, from https://aws.amazon.com/about-aws/whats-new/2017/04/detect-explicit-or-suggestive-adult-using-amazon-rekognition/.
6. Google Cloud (2019). Detecting explicit content in videos. Retrieved Sep 20, 2019, from https://cloud.google.com/video-intelligence/docs/analyze-safesearch.

7. ExtremeTech, (2019). Just how big are porn sites? Retrieved Sep 20, 2019, from https://www.extremetech.com/computing/123929-just-how-big-are-porn-sites.
8. Hammami, M.; Chahir, Y; and Chen, L. (2003). WebGuard: Web based adult content detection and filtering system. *Proceedings IEEE/WIC International Conference on Web Intelligence* (*WI* 2003). Halifax, NS, Canada, 574-578
9. Zhang, J.; Sui, L.; Zhuo, L.; Li, Z.; and Yang, Y. (2013). An approach of bag-of-words based on visual attention model for pornographic images recognition in compressed domain. *Neurocomputing*, 110, 145-152.
10. Ali, F.; Khan, P.; Riaz, K.; Kwak, D.; Abuhmed, T.; Park, D.; and Kwak, K.S. (2017). A fuzzy ontology and SVM-based web content classification system. *IEEE Access*, 5, 25781-25797.
11. Yin, L.; Dong, M.; Deng, W.; Guo, J.; and Zhang, B. (2012). Statistical color model based adult video filter. 2012 *IEEE International Conference on Multimedia and Expo Workshops*. Melbourne, VIC, Australia, 349- 353.
12. Duan, L.; Cui, G.; Gao, W.; and Zhang, H. (2002). Adult image detection method base-on skin color model and support vector machine. *ACCV*2002*: The* 5*th Asian Conference on Computer Vision.* Melbourne, Australia, 1-4.
13. Lee, J.-S.; Kuo, Y.-M.; Chung, P.-C.; and Chen, E.-L. (2007). Naked image detection based on adaptive and extensible skin color model. *Pattern Recognition*, 40(8), 2261- 2270.
14. Nurhadiyatna, A.; Cahyadi, S.; Damatraseta, F.; and Rianto, Y. (2017). Adult content classification through deep convolution neural network. 2017 *International Conference on Computer, Control, Informatics and its Applications* (*IC3INA*). Jakarta, Indonesia, 106-110.
15. Dong, N.; Zhao, L.; Wu, C.H.; and Chang, J.F. (2020). Inception v3 based cervical cell classification combined with artificially extracted features. *Applied Soft Computing*, 93, 106311.
16. Simonyan, K.; and Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. Proceedings of the *International Conference on Learning Representations* (*ICLR*). San Diego, California, 1-14.
17. He, K.; Zhang, X.; Ren, S.; and Sun, J. (2016). Deep residual learning for image recognition. 2016 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Las Vegas, NV, USA, 770-778.
18. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D; Vanhoucke, V.; and Rabinovich, A. (2015). Going deeper with convolutions. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*). Boston, MA, USA, 1-9.
19. Jeong, C.-Y.; Kim, J.-S.; and Hong, K.-S. (2004). Appearance-based nude image detection. *Proceedings of the* 17*th International Conference on Pattern Recognition*, 2004. *ICPR* 2004. Cambridge, UK, 467-470.
20. Rowley, H.A.; Jing, Y; and Baluja, S. (2006). Large scale image-based adult-content filtering. *Proceedings of the First International Conference on Computer Vision Theory and Applications - Volume* 1*: VISAPP*. Sebutal, Portugal, 290-296.
21. Lopes, A.P.B.; de Avila, S.E.F.; Peixoto, A.N.; Oliveira, R.S.; and Araújo, A.d.A. (2009). A bag-of-features approach based on Hue-SIFT descriptor for nude detection. 2009 17*th European Signal Processing Conference*. Glasgow, Scotland, 1552-1556.

22. Avila, S.; Thome, N.; Cord, M.; Valle, E.; and A.d.A. (2011). BOSSA: Extended bow formalism for image classification. 2011 18*th IEEE International Conference on Image Processing*. Brussels, Belgium, 2909-2912.
23. Zhou, K.; Zhuo, L.; Geng, Z.; Zhang, J.; and Li, X.G. (2016). Convolutional neural networks based pornographic image classification. 2016 *IEEE Second International Conference on Multimedia Big Data* (*BigMM*). Taipei, Taiwan, 206-209.
24. Krizhevsky, A.; Sutskever, I.; and Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6), 84-90.
25. Agastya, I.M.A.; Setyanto, A.; Kusrini; and Handayani, D.O.D. (2018). Convolutional neural network for pornographic images classification. 2018 *Fourth International Conference on Advances in Computing, Communication & Automation* (*ICACCA*). Subang Jaya, Malaysia, 1-5.
26. Valle, E.; Avila, S.; da Luz, A.; de Souza, F.D.M.; Coelho, M.; and Araújo, A.d.A. (2012). Content-based filtering for video sharing social networks. *Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais*. 625-638.
27. Souza, F.; Valle, E.; Camara-Chavez, G.; and Araujo, A.d.A. (2012). An evaluation on color invariant based local spatiotemporal features for action recognition. *Proceedings of the* 25*th Conference on Graphics, Patterns and Images*. Ouro Preto, Brazil, 1-6.
28. Avila, S.; Thome, N.; Cord, M.; Valle, E.; and Araújo, A.d.A. (2013). Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding*, 117(5), 453-465.
29. Caetano, C.; Avila, S.; Guimar, S.J.F.; and Ara, A.d.A. (2014). Pornography detection using BossaNova video descriptor. 2014 22*nd European Signal Processing Conference* (*EUSIPCO*). Lisbon, Portugal, 1681-1685.
30. Caetano, C.; Avila, S.; Schwartz, W.R.; Guimarães, S.J.F.; and Araújo, A.d.A. (2016). A mid-level video representation based on binary descriptors: A case study for pornography detection. *Neurocomputing*, 213, 102-114.
31. Moreira, D.; Avila, S.; Perez, M.; Moraes, D.; Testoni, V.; Valle, E.; Goldenstein, S.; and Rocha, A; (2016). Pornography classification: The hidden clues in video space-time. *Forensic Science International*, 268, 46-61.
32. Moustafa, M.N. (2015). Applying deep learning to classify pornographic images and videos. *Proceedings of the* 7*th Pacific‐Rim Symposium on Image and Video Technology* (*PSIVT* 2015), Auckland, New Zealand, 1-10.
33. Wehrmann, J.; Simões, G.S.; Barros, R.C.; and Cavalcante, V.F. (2017). Adult content detection in videos with convolutional and recurrent neural networks. *Neurocomputing*, 272, 432-438.
34. Da Silva, M.V.; and Marana, A.N. (2019). Spatiotemporal CNNs for pornography detection in videos. *Lecture Notes in Computer Science*, 11401, 547-555.
35. Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M.; Berg, A.C.; and Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, 115, 211-252.