A Motion-based Approach for Real-time Detection of Pornographic Content in Videos

Jhonatan Geremias
jgeremias@ppgia.pucpr.br
Pontifical Catholic University of Parana
Curitiba, Parana, Brazil

Eduardo K. Viegas
eduardo.viegas@ppgia.pucpr.br
Pontifical Catholic University of Parana
Curitiba, Parana, Brazil

Alceu S. Britto Jr.
alceu@ppgia.pucpr.br
Pontifical Catholic University of Parana
Curitiba, Parana, Brazil

Altair O. Santin
santin@ppgia.pucpr.br
Pontifical Catholic University of Parana
Curitiba, Parana, Brazil

ABSTRACT

In recent years, several works have proposed highly accurate CNN-based pornography video detection approaches. However, current techniques are unable to cope with the context-dependent nature of pornography content, wherein the analyzed video frame class may change according to its context, whether it is pornographic related or not. This paper proposes a motion-based approach for fine-grained real-time detection of pornographic content in videos implemented in two phases. First, we extract and classify motion-based descriptors built over adjacent video frames to build a motion image from the analyzed video frame. Second, we jointly evaluate the outcome of each single classified motion descriptor to produce a final video frame classification. Experiments performed in a novel fine-grained dataset built from the manual analysis of over 400 thousand video frames, show that current approaches in the literature are unable to cope with context-dependent pornographic content. In contrast, our proposal can maintain the system accuracy, even in the presence of context-dependent pornographic content, hence, maintaining its reliability. In addition, when the extracted motion-based descriptors are jointly evaluated, our proposal is able to improve the detection accuracy by up to 29%.

CCS CONCEPTS

• **Machine Learning** → **Neural Network**; • **Convolutional Neural Network** → *Video Content Detection*; • **Video Classification** → Pornography Detection;

KEYWORDS

Neural Networks, Video Content Detection, Pornography Detection

SAC'22, April 25 –April 29, 2022, Brno, Czech Republic

ACM ISBN 978-1-4503-8713-2/22/04…$15.00
https://doi.org/10.1145/3477314.3507306

ACM Reference Format:
Jhonatan Geremias, Eduardo K. Viegas, Alceu S. Britto Jr., and Altair O. Santin. 2022. A Motion-based Approach for Real-time Detection of Pornographic Content in Videos. In *The 37th ACM/SIGAPP Symposium on Applied Computing (SAC '22), April 25–29, 2022, Virtual Event, .* ACM, New York, NY, USA, Article 4, 8 pages. https://doi.org/10.1145/3477314.3507306

1 INTRODUCTION

Nowadays, adult websites, those with explicit pornographic content, accounts for 12% of all Internet pages, comprising over 35% of all downloaded media from the Internet [1]. For example, a popular pornographic video website receives, on average, 115 million daily visits [26]. As a result, minors are increasingly exposed to pornographic content, in their majority in video format [1], even without their parents' consent, causing embarrassment or even psychological traumas.

Over the last years, several works have focused on the detection of pornographic content in videos [10]. In such a context, techniques based on Convolutional Neural Network (CNN) architectures have yielded promising reported results, with highly accurate classification models. To achieve such a goal, in general, an image-based CNN model is built according to the extracted video frames from both normal and pornographic videos. Consequently, the built CNN model classifies following a video frame granularity. Therefore, to achieve video classification, in general, proposals deem a video as pornographic when the majority of its frames are not normal [13].

In general, pornographic videos must be classified in real-time. For instance, for the classification of live video streaming platforms, in which pornographic content may occur in real-time transmission. However, traditional CNN-based classification approaches require the prior extraction of all video frames for the video classification; hence, they are unable to be executed in real-time.

In recent years, several works have proposed scene-based classification approaches [22]. In such a case, a video is split in scenes (i.e., a fixed-size sequence of frames), and each video scene is individually classified. However, although such proposals enable real-time classification, they often do not take into account the context-dependent nature of pornographic content.

Pornographic videos are often made of both normal and pornographic video frames. This is because a pornographic video often also contain several scenes with normal content, e.g., clothed actors talking to each other. In contrast, a normal video may also have

adult content, e.g., nudists walking on the beach. Such characteristic of pornographic videos introduces a significant challenge to traditional CNN-based detection approaches, given that similar video frames, with the same actors, in the same environment, video resolution and image quality may be labeled with different classes due to their context, e.g., whether the actors are clothed or not. Consequently, the CNN model must be able to properly differentiate the frame context, even from similar and even adjacent video frames.

Current pornographic datasets widely used in the literature, are labeled in a video granularity [21][2], wherein all pornographic video frames are labeled as pornographic, regardless of the frame context. Thus, proposed techniques built on such datasets are prone to a lack of context, even when scene-based classification is used, as the built CNN model is unable to establish the frame context properly, decreasing its accuracy.

In light of this, this paper proposes a motion-based approach for fine-grained real-time detection of pornographic content in videos. The proposed model is implemented in two phases. First, we extract motion-based features from adjacent video frames to compound a frame motion description. The motion-based features, besides the evaluated frame, is used to describe in a higher-level the behavior of adjacent frames, thus, improving the underlying CNN model input frame features. Consequently, it provides features such as motion direction based on the optical image flow and structural similarity maps computed between adjacent video frames. As a result, our model can provide enriched video frame characteristics to describe the scene context, taking into account that we extract motion-based features from adjacent frames. Second, to jointly evaluate the frame classification outcome from each scene feature, we apply a shallow classifier fed with each individually classified motion-based descriptor. Thus, the shallow classifier enables real-time detection, taking into account that our model only demands the adjacent frame for the extraction of the motion-based features, but also evaluates the frame concerning several motion characteristics. In summary, our work presents the following main contributions:

- A novel fine-grained pornography dataset, namely FPD. The dataset, the first of its kind, made of 476,482 manually labeled video frames, enables the fine-grained evaluation of proposed pornography detection techniques. Each video frame is labeled according to its context, regardless of their video nature, consequently, it can be used for the evaluation of detection techniques concerning the pornography context;
- We show that state-of-the-art image-based CNN pornography detection techniques are unable to provide reasonable accuracy detection rates in FPD. Consequently, proposed techniques are unable to cope with the context-dependent nature of pornographic videos;
- We propose a motion-based representation of video frames to extract the frame context in real-time. Our proposed approach extracts features from adjacent frames such as motion estimation based on the optical flow [18] and structural similarity maps (SSM). The proposed features can preserve the obtained accuracy even in the presence of context-dependent pornographic content;
- We propose a shallow-based classification approach to evaluate the extracted motion-based features from adjacent frames jointly. The proposed approach enables real-time video frame evaluation, taking into account the frame context, while also significantly improving the classification accuracy;

2 BACKGROUND

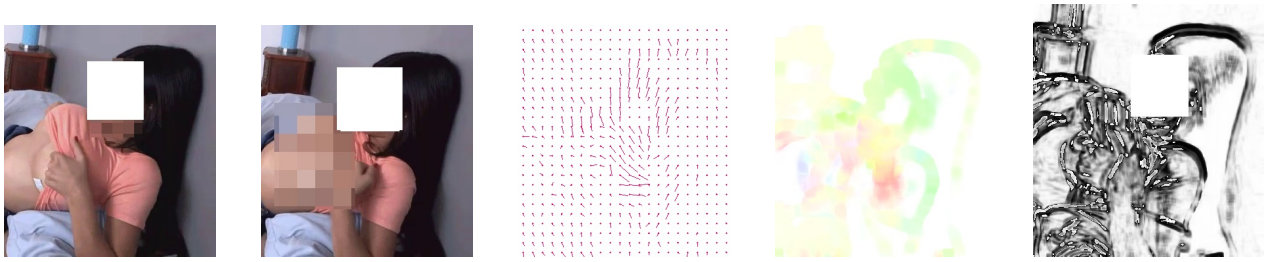
In general, proposed CNN-based techniques for pornography detection in videos are made of three sequential modules [19], namely *Frame Extraction*, *Frame Classification*, and *Alert*. First, the *Frame Extraction* module extracts the video frames for further classification. For instance, often extracting 24 frames/sec from each analyzed video. Then, the *Frame Classification* module classifies each video frame individually, i.g., applying a CNN model for the classification of the input video frame as either normal or pornographic. Finally, the *Alert* module establishes whether the evaluated video is pornographic or normal. In such a case, in general, if the majority of the video frames are classified as pornographic by the *Frame Classification* module, the entire video is deemed as pornographic.

In recent years, several approaches have been proposed for the classification of pornographic video frames, wherein proposed techniques often aim at improving their obtained accuracy in a given dataset [7]. To achieve such a goal, authors often resort to more complex CNN architectures, wherein their underlying CNN models can be made of several Gigabytes of memory, e.g., the InceptionV3 CNN architecture model, which demands up to 12 GB of data for the model execution. Consequently, current proposed techniques can provide significantly highly accuracies in a specific dataset.

However, for feasible deployment of proposed techniques in production environments, one must be able to provide real-time detection while also taking into account the context-dependent nature of pornographic videos. In contrast, current techniques are not evaluated through fine-grained pornographic datasets, i.e., datasets where each video frame is manually labeled according to its context.

An example of a context-dependent pornographic video is shown in Figure 1. In the figure, adjacent video frames – similar frames should be classified into different classes. In the 1st video frame (Fig. 1a), the person is clothed, while in the second video frame is unclothed (Fig. 1b). It is important to note that this characteristic is very challenging to CNN-based classification techniques, taking into account that both video frames are similar, and are likely to be classified to the same class by the underlying CNN model. Therefore, in such a case, the video frame must be evaluated according to its context, as the evaluation of the entire video becomes not feasible (further evaluated in Section IV).

A simplified approach for the extraction of the video frame context resort to the evaluation of adjacent video frames. For instance, consider two successive video frames, namely f^{i} and f^{i-1}. The goal is to extract features that best describe the difference between both video frames f^{i} and f^{i-1}. Over the last years, several descriptors have been proposed for this task, such as the *Structural Similarity Map (SSM)* [30], and techniques based on *Optical fFlow* [18][3] estimators. Figure 1 shows an example of the extracted structural similarity map [30] (Fig. 1e), and two optical flows, the Lucas-Kanade [18] (Fig. 1c), and Pyramidal Lucas-Kanade (PLK) [3] (Fig. 1d) for two adjacent examples of pornographic video frames.



(a) 1st video frame, *Normal* class.

(b) 2nd video frame, *Pornographic* class.

(c) Lucas-Kanade Optical Flow between 1st and 2nd frames.

(d) Pyramidal Lucas-Kanade (PLK) Optical Flow between 1st and 2nd frames.

(e) Structural Similarity Map (SSM) between 2nd and 2nd frames.

Figure 1: Example of a pornographic video comprising both normal and pornographic frames, and the related motion-based description between first and second video frames.

3 RELATED WORKS

Over the last years, several works have proposed highly accurate CNN-based pornography detection techniques. In general, proposed approaches can be divided into either image-based, video-based, or scene-based approaches.

Video-based pornography detection techniques are still in its beginnings. In general, proposed techniques for such goal extract dynamic features from several video frames, such as frame movement, and temporal features [6]. For instance, Caetano *et al.* [5] proposed a video description analysis technique through the analysis of the video frames for video classification purposes. Then, the authors apply a shallow classifier to classify a given video according to the extracted video description. As a result, the authors were able to improve detection accuracy. However, their approach can not be applied in real-time, as it demands the availability of all video frames. Another technique was evaluated by A. Karpathy *et al.* [14] for video classification purposes. In their work, the evaluation shows that CNN models fed with a single video frame per classification round provide the highest classification accuracy. Therefore, videos are classified according to the majority of each video frame class, hence, unable to provide fine-grained classification in context-dependent scenarios.

In recent years, 3D-CNNs, a CNN architecture with several video frame inputs, have also been used for video classification purposes. For instance, S. Ji *et al.* [11] apply a 3D-CNN, with seven input video frames, for human activity recognition. Their proposed technique improves the traditional CNN classification accuracy but introduces more complex CNN architectures. In addition, their technique is also prone to context-dependent scenarios, taking into account that it does not address the video frame similarity challenge.

Motion-based techniques for video classification purposes have shown promising results. M. Gong and Y. Yang [8] extracts disparity and optical flow descriptors to decrease over-fitting. In their work, motion-based descriptors were more robust, with higher detection accuracies when compared to traditional techniques. Similarly, Lin *et al.* [17] propose an object tracking approach through optical flow descriptors to decrease over-fitting. In their work, motion-based descriptors presented higher detection accuracies than traditional techniques. In pornography detection context, M. Perez *et al.* [25] proposes a CNN-based technique through motion-based information. In their work, the authors extract motion vectors from MPEG encoded videos for classification purposes. However, although their technique provides high detection accuracies, they do not take into account the context-dependent nature of pornographic content. A scene-based approach was also proposed by Moreira *et al.* [22]; the authors consider the context-dependent nature of sensitive scenes. To achieve such a goal, the authors split the analyzed video in scenes, each composed by several video frames, then classifies it according to through multimodal features, such as video frames, audio, and motions, which also demands the prior availability of the whole video media.

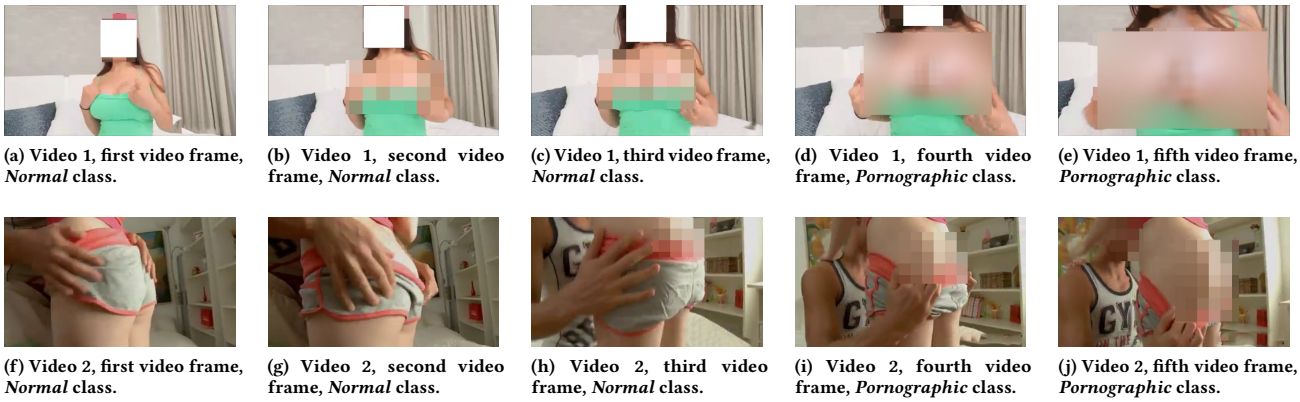
To the best of our knowledge, we are the first work to address both real-time and fine-grained pornography detection taking into account the context-dependent nature of pornographic content. To achieve such a goal, we leverage motion-based descriptors, to extract the video frames structural similarities and motion description, hence, improving the underlying CNN input data, and, consequently, its accuracy.

4 THE CHALLENGE OF CONTEXT-DEPENDENT PORNOGRAPHY DETECTION IN VIDEOS

In this section, we further evaluate the impact of context-dependent pornography content in videos on the classification accuracy of traditional image-based CNN models. More specifically, we first introduce our built dataset, namely fine-grained pornography dataset; then, we evaluate the classification performance of several image-based CNN techniques over our data.

4.1 Fine-grained Pornography Dataset

Current publicly available datasets used for the evaluation of pornography detection schemes, are unable to provide the expected level of granularity in their video frames. In other words, video frames are labeled according to their originating video, regardless of their content, e.g., if the actors are clothed or not. Consequently, proposed techniques built upon such data, although present a high accuracy rate, may operate poorly in production environment conditions, wherein the actual video class may change over time.



(a) Video 1, first video frame, *Normal* class.
(b) Video 1, second video frame, *Normal* class.
(c) Video 1, third video frame, *Normal* class.
(d) Video 1, fourth video frame, *Pornographic* class.
(e) Video 1, fifth video frame, *Pornographic* class.
(f) Video 2, first video frame, *Normal* class.
(g) Video 2, second video frame, *Normal* class.
(h) Video 2, third video frame, *Normal* class.
(i) Video 2, fourth video frame, *Pornographic* class.
(j) Video 2, fifth video frame, *Pornographic* class.

Figure 2: Sample pornographic videos with both normal and pornographic video frames in *Fine-grained Pornography Dataset*.

In light of this, we present the *Fine-grained Pornography Dataset* (FPD). Our dataset, the first of its kind, is built over both pornography and normal video frames, in which each video frame is manually labeled. To achieve such a goal, for each video frame, we manually label it as either pornography or normal, according to the video frame context. For instance, video frames are only labeled as pornography when showing sexual intercourse, naked actors, among other explicit adult content. In contrast, if a video frame, even when originated from a pornographic video, is showing clothed actors, and no sexual intercourse, the video frame is labeled as normal.

Figure 2 shows sample video frames from the FPD dataset and their related manually assigned labels. The dataset was built through the manual analysis of 14,671 videos, resulting in a total of 476,482 video frames, extracted from 1,351 and 13,320 pornographic and normal videos, respectively. The analyzed videos were gathered from the public domain, such as pornographic websites and public video-sharing platforms. Hence, they include people from different ethnicities, races, and gender performing different activities, either pornographic related or not.

4.2 The Performance of Image-based CNN in Fine-grained Pornography Dataset

To evaluate the impact of context-dependent detection of pornographic content, we apply state-of-the-art image-based CNN architectures in the FPD dataset. The architectures are evaluated according to their video frame accuracy, thus, enabling the evaluation of their accuracy concerning context-dependent pornographic content. Three CNN architectures were evaluated, with and without transfer learning (TL) [23]: the Alexnet [15], Caffenet [12], and Googlenet [28]. The CNNs were built and evaluated in Caffe API, version 1.0. For the transfer learning evaluation, we use the pre-trained weights as obtained from the well-known ImageNet dataset [16].

For training purposes, the FPD dataset was split in train, validation, and test datasets, comprising 60%, 20%, and 20% of the videos that compounds the original FPD dataset, respectively. Each dataset contains unique videos, hence, properly establishing the CNN model generalization capacity. The training dataset is used for the CNN model building, while the validation dataset is used at the training phase, for generalization estimation. The final accuracy is computed through the test dataset. Each evaluated architecture was executed for 1000 epochs, and its learning rate set empirically according to the resulting loss and a momentum weight of 0.9.

Table 1: Traditional image-based CNN architectures accuracy performance in the FPD dataset. Normal and Porn accuracy denotes the ratio of normal and pornographic frames correctly classified as such. Context-dependent (CD) accuracy denotes the ratio of normal frames in pornographic videos correctly classified as normal.

CNN	Video Frame Accuracy (%)		
	Normal	*Porn*	*CD*
Caffenet	94.58	68.52	65.86
Alexnet	97.80	63.14	72.85
Googlenet	95.48	62.07	71.58
TL Caffenet	99.11	79.52	69.73
TL Alexnet	96.92	73.02	65.76
TL Googlenet	95.46	62.07	71.58

Table 1 shows the obtained accuracy in the FPD test dataset for each evaluated CNN architecture. It is possible to note a high detection accuracy for frames labeled as *Normal*. In contrast, video frames extracted from pornographic videos (*Pornographic* and *Context-dependent*, Table 1) present a significantly lower accuracy rate. This is because, in the FPD dataset, a pornographic video may contain both normal and pornographic video frames, significantly increasing the classification difficulty to the underlying CNN model.

4.3 Discussion

The evaluation, through the built dataset, showed that current and widely used pornography detection image-based CNN architectures are unable to detect pornographic content in videos reliably. This is because current proposed schemes in the literature often neglect the context-dependent nature of pornographic videos, in which a pornographic video may also contain normal scenes (see Figure 2). Consequently, proposed techniques that assume that a

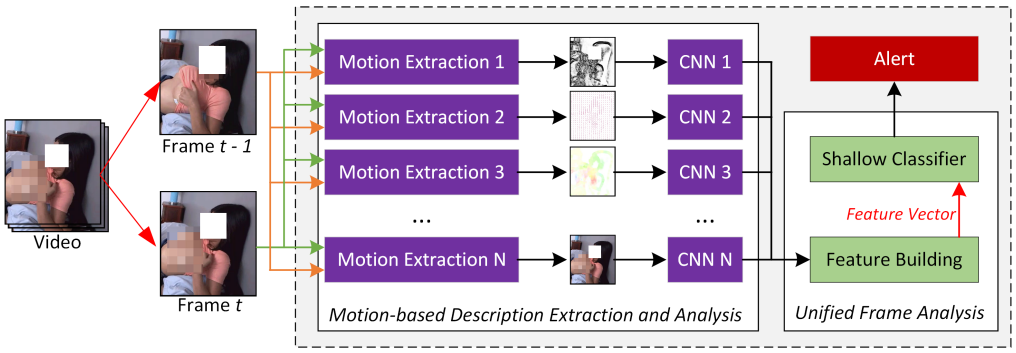


Figure 3: Proposed motion-based pornography detection model for real-time fine-grained detection of pornography in videos.

video will only contain video frames from the same class, either normal or pornographic, will be unreliable for production deployment, decreasing their accuracy under a video context change.

5 A MOTION-BASED PORNOGRAPHY DETECTION MODEL

To address the challenge of context-dependent detection of pornographic videos in real-time. We present a motion-based pornography detection model. The goal is to enable the real-time detection of context-dependent pornographic content in videos. That is, without requiring the prior availability of all video frames, or a significant part of it, for the classification process, while maintaining its reliability when a context change in a video occurs. The operation of our model proceeds in two main stages: *Motion-based Description Extraction and Analysis*, and *Unified Frame Motion Analysis*.

The *Motion-based Description Extraction and Analysis* is performed twofold. First, it extracts motion-based descriptors from the analyzed frame. The motion-based descriptors enable the representation of the frame context, concerning its prior frame. Consequently, further analysis can be performed according to the current frame motion, which acts as its context representation. Then, the extracted motion-based descriptors are fed to an image-based CNN model. As a result, the underlying CNN model can analyze the frame context. The *Unified Frame Motion Analysis* is responsible for collecting all the analyzed frame motion descriptors, and the detailed analysis as performed through the CNN model, and output a related frame class.

The module applies a shallow classifier, fed with all used motion-based descriptor analysis output, to produce a unified and final video frame class. Therefore, the model can analyze the video frame context, as obtained from several motion-based analysis, to reliably classify context-dependent pornographic video frames.

The proposal overview is shown in Figure 3. The next subsections describe in detail the proposal stages, including the architecture of the modules that implements the stages and the description of the main components.

5.1 Motion-based Description Extraction and Analysis

Context-dependent pornographic content poses a significant challenge to traditional image-based CNN techniques. Similar frames produce similar input pixels to the underlying CNN model, which in fact might pertain to distinct classes. Therefore, the classification of the input video frames based only on the current analyzed frame may result in a higher error rate in production (evaluated in Section 4.2).

Our model extracts motion-based features to enrich the available input data before a decision can be made regarding the analyzed video frame. However, to enable real-time video analysis, our model extracts motion-based descriptors concerning only two adjacent video frames. Consequently, our proposal only incurs in the requirement of the prior frame from the current analyzed video frame.

Consider an analyzed video frame, namely f^{i}, and a corresponding prior analyzed frame, namely f^{i-1}. In real-time, the module extracts a set of motion-based descriptors from the f^{i}, concerning its prior f^{i-1}. For instance, for each analyzed frame, the module may extract the *structural similarity map* [20], and *optical flow* [27] descriptors (*Motion Extraction*, Figure 3). Then, each of the extracted motion-based descriptors is individually classified by a corresponding image-based CNN (*CNN*, Figure 3).

As a result, each CNN outputs a related video frame label, according to the input motion-based descriptor. Therefore, the proposal can extract several motion-based descriptors, which acts as a context-based measure for pornography detection in our proposal and analyze each descriptor accordingly, through the set of image-based CNN. The final classification outcome, as represented through the set of the CNN classifications, are fed to the *Unified Frame Motion Analysis*.

The proposal insight, through motion-based descriptors, is that context-dependent pornographic content can be detected through image-based CNN models. However, to enable real-time detection, our proposal extracts the motion-based descriptors through the analysis of two adjacent frames. Consequently, decreasing the processing needs while also enabling real-time detection.

5.2 Unified Frame Motion Analysis

Finally, the *Unified Frame Motion Analysis* module is responsible for producing a final video frame classification outcome. The module receives the set of individual CNN classifications, as output by each of the motion-based descriptors. Each classification outcome, comprise the assigned video frame label, and the CNN confidence values for both normal and pornography classes. Consequently,

each used motion-based descriptor produces three values, which are used to compound a unified feature vector. The built feature vector is forwarded to a shallow classifier. The shallow classifier, in turn, produces a final video frame classification, as either normal or pornography. Therefore, the model can jointly evaluate several motion-based descriptors, which act as a representation of the video frame context, to classify pornographic video frames according to their context in real-time.

5.3 Discussion

Fine-grained classification of pornographic video frames introduces several challenges to traditional image-based CNN architectures. The proposed model aims to enable the fine-grained classification of video frames, while taking into account the context-dependent nature of pornographic videos, in a real-time manner. To this end, the proposal is twofold.

First, we extract several motion-based descriptors, which acts as a representation of the video frame context to enrich the available data used for classification purposes. Noteworthy, our proposal only demands two adjacent video frames to extract motion-based descriptors, enabling real-time detection, taking into account that it does not require the prior availability of all video frames for classification purposes.

Second, we jointly evaluate the outcome of the motion-based descriptor analysis by applying a shallow classifier over the individual produced classifications. As a result, the proposed model can jointly evaluate the video frame motion, concerning the set of used motion-based descriptors. Therefore, the proposal can improve the classification accuracy of its underlying models, considering that several motion-based descriptors are used, instead of the evaluation of a single video frame, as made by state-of-the-art techniques.

6 EVALUATION

The present evaluation focuses on answering three research questions: (*Q*1) *Does the proposed motion-based classification enables the detection of pornographic content in videos?* (*Q*2) *Does the proposed unified frame analysis aid in detecting pornographic content in real-time?* (*Q*3) *Can the proposed fine-grained approach be applied to detect pornographic content in a video granularity?*

The next subsections describe how we build the proposed model and how it performs in our dataset.

6.1 Model Building

The same set of CNN architectures used in Section 4.2 were evaluated through our model. Similarly, the same evaluation procedure was adopted, wherein 60%, 20%, and 20% of videos are used for training, validation, and testing purposes.

Three widely used motion-based descriptors in the literature were used. The optical flows obtained with the Lucas-Kanade [31] and PLK algorithms [4], and the structural similarity map between adjacent frames [30]. An example of the extracted motion-based descriptors is shown in Figure 1. As described in Section 5.1, each video frame descriptor was extracted through the analysis of the current and prior one. The motion-based descriptors were implemented in Python programming language using the OpenCV API version 2.4 [24]. For each extracted motion-based descriptor, a CNN model is built and evaluated.

Table 2: Frame granularity accuracy performance of motion-based CNN architectures in FPD dataset considering the following topologies: Caffenet (C); Alexnet (A); Googlenet (G); TL Caffenet (TL-C); TL Alexnet (TL-A); TL Googlenet (TL-G)

Motion Type	CNN	Accuracy (%)		
		Normal	***Porn***	***CD***
Pyramidal Lucas-Kanade (PLK)	C	97.21	43.07	66.65
	A	89.57	65.78	53.52
	G	98.69	24.47	87.74
	TL-C	95.90	61.01	60.27
	TL-A	***93.98***	***65.40***	***58.58***
	TL-G	95.12	42.49	77.03
Lucas-Kanede	C	98.14	67.19	64.89
	A	97.81	62.68	75.47
	G	98.16	65.63	72.73
	TL-C	98.95	77.98	65.86
	TL-A	98.82	77.58	63.42
	TL-G	***99.41***	***74.96***	***76.04***
Structural Similarity Map (SSM)	C	93.75	46.02	62.26
	A	90.61	79.76	40.47
	G	97.45	60.57	65.00
	TL-C	97.83	67.45	64.49
	TL-A	96.07	75.09	58.33
	TL-G	***98.59***	***65.44***	***76.08***

6.2 Motion-based Classification

The first experiment relates to question *Q*1, and aims to evaluate whether each individual motion-based descriptor can be used for classification purposes. We build and evaluate the same set of CNN architectures evaluated in Section 4.2 through each of the extracted motion-based descriptors.

Table 2 shows the obtained video frame accuracy concerning each of the extracted motion-based features. It is possible to note that the used motion-based features presented similar accuracy to those obtained with the image-based CNN. In most cases, the proposal presents more stable detection rates than those obtained from image-based CNN architectures. In other words, the accuracy of CNN architectures, built from motion-based descriptors, is not significantly degraded when facing context-dependent pornographic content. As an example, the most accurate motion-based CNNs (*TL Alexnet* for the PLK-based optical flow, and *TL Googlenet* for the Luca-Kanade and Structural Similarity Map descriptors, Table 2), presented in average only a 6.2% of accuracy difference, for the classification of pornography to context-dependent video frames.

As a result, the proposed motion-based detection scheme can remain reliable to the user, taking into account that it will present similar accuracy rates to the test phase when used in production.

6.3 Unified Frame Analysis

To answer question *Q*2, we select the most accurate CNN models for each of the extracted motion-based descriptors (shown in bold in

Table 3: Proposal final accuracy performance in FPD dataset (by frame).

Shallow Classifier	Accuracy (%)		
	Normal	*Porn*	*CD*
NaiveBayes	95.00	94.07	90.58
RandomForest	***95.78***	***92.56***	***99.98***
MLP	95.66	93.08	90.68
SVM	95.51	93.19	91.57

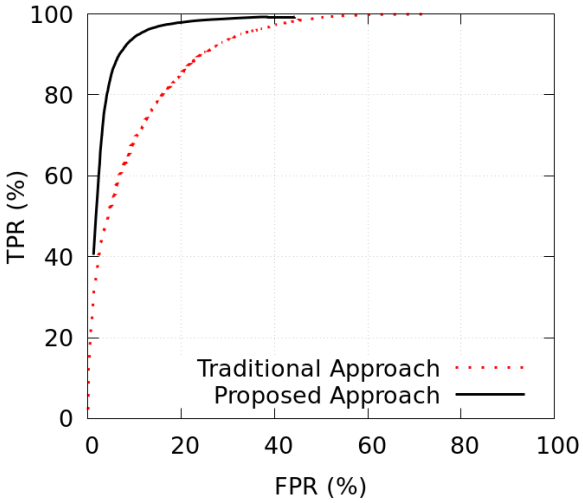


Figure 4: Proposed motion-based pornography detection (RandomForest, Table 3) and traditional image-based CNN detection ROC comparison (TL Caffenet, Table 1).

Table 2), to build the shallow classifier input feature vector (*Unified Frame Analysis*, Figure 3).

We build several well-known shallow classifiers [9, 29] and use them for the video frame classification. The classifiers were built on top of Weka API, version 3.8. The input features for the Naive Bayes were discretized through the supervised discretization algorithm [32]. For the Random Forest classifier, 100 decision trees were used as base-learner. The Multilayer Perceptron (MLP) was implemented with a learning rate of 0.3, and a momentum of 0.2. Finally, the Support Vector Machine (SVM) was executed with the RBFKernel, with a gamma value of 0.01.

Table 3 shows the obtained classification accuracy for each of the evaluated shallow classifiers. In such a case, it is possi ble to note a significant accuracy improvement when compared to the traditional image-based CNNs. In other words, the proposed approach was able to improve the detection accuracy of context-dependent pornographic content significantly. For instance, regarding the most accurate shallow classifier, RandomForest, the proposed approach was able to improve detection accuracy by 13% and 30% for pornographic and context-dependent video frames, respectively - with a tradeoff of only 3% for the classification of normal video frames.

In Figure 4 we further investigate the accuracy improvement provided by our proposal when compared to the state-of-the-art.

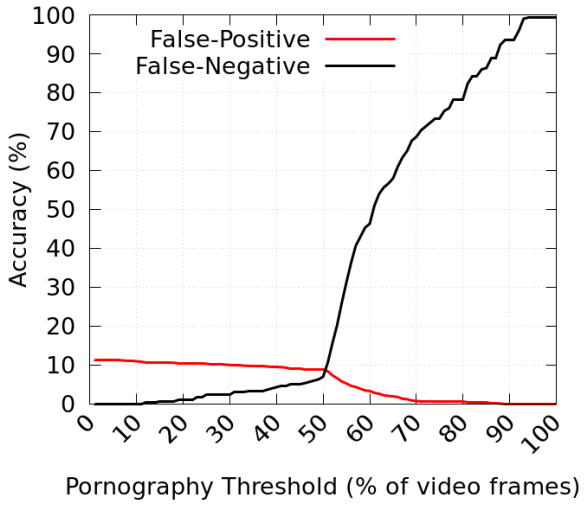


Figure 5: Proposal, implemented through RandomForest, FP and FN relation when used for video classification purposes.

The proposed scheme significantly improves AUC, providing more accurate operation points when compared to the traditional approach.

6.4 Video Classification

Finally, to answer question $Q3$, we apply the proposed detection scheme for video classification purposes. We evaluate all video frames and deem a video as pornography according to a selected pornographic threshold. The pornographic threshold establishes the ratio of video frames that must be classified as pornographic to classify a given video as pornographic related.

Figure 5 shows the relation between the pornographic threshold and the obtained video accuracy. It is possible to note that our proposal is able to reach an FN rate of 8% and an FP rate of 9% when using a pornographic threshold of 50%. Nonetheless, when varying the used pornographic threshold, one is able to further decrease the FP rate according to his discretion.

Besides being applicable for real-time fine-grained (video frame granularity) classification of pornographic videos, our proposal can also be applied for offline classification of pornographic videos, with high accuracy rates for both cases.

7 CONCLUSION

Current approaches for pornography detection in videos are unable to provide real-time and fine-grained detection of pornographic content while taking into account the context-dependent nature of such kind of media. We proposed a motion-based classification scheme able to significantly improve detection accuracy in real-time for fine-grained context-dependent pornographic content classification in videos.

The proposal insight leverages motion-based descriptors to increase the video frame features to be used by the underlying CNN model. In addition, to enable the usage of several motion-based descriptors, we apply a shallow classifier to evaluate the final video

frame classification outcome jointly. As a result, our proposal was able to improve state-of-the-art pornography detection approaches accuracies significantly while providing a real-time and fine-grained detection scheme.

For future works, we are going to pursue a lightweight detection approach for resource-constrained devices, and evaluate our proposal in further context-dependent fields, such as violent scenes and human activity recognition.

ACKNOWLEDGMENTS

The authors thank Brazilian National Council for Scientific and Technological Development (CNPq) for partial financial support (grants 430972/2018-0, and 306684/2018-2) and URCOP (Unidade de Repressão aos Crimes de Ódio e a Pornografia Infantil), a special division for pornography crackdown of Federal Police of Brazil for the support to the project.

REFERENCES

[1] 2019. Internet Pornography by the Numbers; A Significant Threat to Society. Pornhub Insights. https://www.webroot.com/us/en/resources/tips-articles/internet-pornography-by-the-numbers Accessed: January 21, 2020.

[2] Sandra Avila, Nicolas Thome, Matthieu Cord, Eduardo Valle, and Arnaldo de A. Araújo. 2013. Pooling in image representation: The visual codeword point of view. *Computer Vision and Image Understanding* 117, 5 (May 2013), 453–465. https://doi.org/10.1016/j.cviu.2012.09.007

[3] J. Bouguet. 1999. Pyramidal Implementation of the Affine Lucas Kanade Feature Tracker. *Intel Corporation, Microprocessor Research Labs.*

[4] T Brox and J Malik. 2011. Large Displacement Optical Flow: Descriptor Matching in Variational Motion Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33, 3 (March 2011), 500–513. https://doi.org/10.1109/tpami.2010.143

[5] Carlos Caetano, Sandra Avila, William Robson Schwartz, Silvio Jamil F. Guimarães, and Arnaldo de A. Araújo. 2016. A mid-level video representation based on binary descriptors: A case study for pornography detection. *Neurocomputing* 213 (Nov. 2016), 102–114. https://doi.org/10.1016/j.neucom.2016.03.099

[6] Tadilo Endeshaw, Johan Garcia, and Andreas Jakobsson. 2008. Classification of indecent videos by low complexity repetitive motion detection. In *2008 37th IEEE Applied Imagery Pattern Recognition Workshop*. IEEE. https://doi.org/10.1109/aipr.2008.4906438

[7] Jhonatan Geremias, Altair O. Santin, Eduardo K. Viegas, and Alceu S. Britto. 2020. Towards Real-time Video Content Detection in Resource Constrained Devices. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE. https://doi.org/10.1109/ijcnn48605.2020.9207125

[8] Minglun Gong and Yee-Hong Yang. 2006. Disparity Flow Estimation using Orthogonal Reliability-based Dynamic Programming. In *18th International Conference on Pattern Recognition (ICPR'06)*. IEEE. https://doi.org/10.1109/icpr.2006.456

[9] Pedro Horchulhack, Eduardo K. Viegas, and Altair O. Santin. 2022. Toward feasible machine learning model updates in network-based intrusion detection. *Computer Networks* 202 (Jan. 2022), 108618. https://doi.org/10.1016/j.comnet.2021.108618

[10] Christian Jansohn, Adrian Ulges, and Thomas M. Breuel. 2009. Detecting pornographic video content by combining image features with motion information. In *Proceedings of the seventeen ACM international conference on Multimedia - MM '09*. ACM Press. https://doi.org/10.1145/1631272.1631366

[11] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2013. 3D Convolutional Neural Networks for Human Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 35, 1 (Jan. 2013), 221–231. https://doi.org/10.1109/tpami.2012.59

[12] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. 2014. Caffe. In *Proceedings of the ACM International Conference on Multimedia - MM '14*. ACM Press. https://doi.org/10.1145/2647868.2654889

[13] Xin Jin, Yuhui Wang, and Xiaoyang Tan. 2019. Pornographic Image Recognition via Weighted Multiple Instance Learning. *IEEE Transactions on Cybernetics* 49, 12 (Dec. 2019), 4412–4420. https://doi.org/10.1109/tcyb.2018.2864870

[14] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-Scale Video Classification with Convolutional Neural Networks. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE. https://doi.org/10.1109/cvpr.2014.223

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems 25*, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger (Eds.). Curran Associates, Inc., 1097–1105.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (May 2017), 84–90. https://doi.org/10.1145/3065386

[17] Luyue Lin, Bo Liu, and Yanshan Xiao. 2017. An object tracking method based on CNN and optical flow. In *2017 13th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery (ICNC-FSKD)*. IEEE. https://doi.org/10.1109/fskd.2017.8393149

[18] Bruce Lucas and Takeo Kanade. 1981. An Iterative Image Registration Technique with an Application to Stereo Vision (IJCAI). *[No source information available]* 81.

[19] Jackson Mallmann, Altair Olivo Santin, Eduardo Kugler Viegas, Roger Robson dos Santos, and Jhonatan Geremias. 2020. PPCensor: Architecture for real-time pornography detection in video streaming. *Future Generation Computer Systems* 112 (Nov. 2020), 945–955. https://doi.org/10.1016/j.future.2020.06.017

[20] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2016. A Large Dataset to Train Convolutional Networks for Disparity, Optical Flow, and Scene Flow Estimation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[21] D. Moreira, S. Avila, M. Perez, D. Moraes, E. Testoni, V. Valle, S. Goldenstein, and A. Rocha. 2016. Pornography classification: The hidden clues in video space-time. In *Proc. Forensic Science Int.* 46–61.

[22] Daniel Moreira, Sandra Avila, Mauricio Perez, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. 2019. Multimodal data fusion for sensitive scene localization. *Information Fusion* 45 (Jan. 2019), 307–323. https://doi.org/10.1016/j.inffus.2018.03.001

[23] A. Moujahid. 2019. A Practical Introduction to Deep Learning with Caffe - Python. http://adilmoujahid.com/posts/2016/06/introduction-deep-learning-python-caffe/ Accessed: January 21, 2020.

[24] Opencv. 2020. Opencv - Open Source Computer Vision Library. In *Opencv*. https://opencv.org/

[25] Mauricio Perez, Sandra Avila, Daniel Moreira, Daniel Moraes, Vanessa Testoni, Eduardo Valle, Siome Goldenstein, and Anderson Rocha. 2017. Video pornography detection through deep learning techniques and motion information. *Neurocomputing* 230 (March 2017), 279–293. https://doi.org/10.1016/j.neucom.2016.12.017

[26] Pornhub. 2019. The 2019 Year in Review. Pornhub Insights. https://www.pornhub.com/insights/2019-year-in-review Accessed: January 21, 2020.

[27] Nusrat Sharmin and Remus Brad. 2012. Optimal Filter Estimation for Lucas-Kanade Optical Flow. *Sensors* 12, 9 (Sept. 2012), 12694–12709. https://doi.org/10.3390/s120912694

[28] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going Deeper With Convolutions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[29] Eduardo Viegas, Altair Olivo Santin, and Vilmar Abreu Jr. 2021. Machine Learning Intrusion Detection in Big Data Era: A Multi-Objective Approach for Longer Model Lifespans. *IEEE Transactions on Network Science and Engineering* 8, 1 (Jan. 2021), 366–376. https://doi.org/10.1109/tnse.2020.3038618

[30] Z. Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (April 2004), 600–612. https://doi.org/10.1109/tip.2003.819861

[31] Zhen Wang and Xiaojun Yang. 2018. Moving Target Detection and Tracking Based on Pyramid Lucas-Kanade Optical Flow. In *2018 IEEE 3rd International Conference on Image, Vision and Computing (ICIVC)*. IEEE. https://doi.org/10.1109/icivc.2018.8492786

[32] H. Zhang. 2004. The Optimality of Naive Bayes. In *American Association for Artificial Intelligence*. FLAIRS Conference.