

Joint inter-intra representation learning for pornographic video classification

Dinh-Duy Phan^{1,2}, Quang-Huy Nguyen^{1,2}, Thanh-Thien Nguyen^{1,2}, Hoang-Loc Tran^{1,2}, Duc-Lung Vu^{1,2}

¹Faculty of Computer Engineering, University of Information Technology, Ho Chi Minh City, Vietnam

²Vietnam National University, Ho Chi Minh City, Vietnam

Article Info

Article history:

Received Sep 18, 2021

Revised Dec 14, 2021

Accepted Jan 11, 2022

Keywords:

Inter-intra similarity

Pornographic classification

Video retrieval

Visual similarity

ABSTRACT

This paper addresses video inter-intra similarity retrieval for pornographic classification. The main approaching method is obtaining the internal representation and external similarity between a single unlabeled video and batches of labeled videos, then combining together to determine its label. For the internal representation, we extracted inner features within frames and clustered them to find the representative centroid as the intra-feature. For the external similarity, we utilized a similarity video learning named ViSiL to calculate distance score between two videos using chamfer similarity. With distance scores between input video and batches of pornographic/non-pornographic videos, the inter feature of the input video is obtained. Finally, the inter similarity vector and the intra representation are then concatenated together and fed to a final classifier to identify whether the video is for adults or not. In experiment, our method performs 96.88% accuracy on NPDI-2k, achieved a comparative result comparing to other state-of-the-art methods on the pornographic classification problem.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Thanh-Thien Nguyen

Faculty of Computer Engineering, University of Information Technology

Quarter 6, Linh Trung Ward, Thu Duc District, Ho Chi Minh City, Vietnam

Email: thiennt@uit.edu.vn

1. INTRODUCTION

Based on the popularity of internet-based video sharing services, the quantity and diversity of images or videos on the world wide web have reached unprecedented scales, thus making it difficult to categorize. Moreover, with the increase of sexual websites with no restriction, the demand for filtering and preventing adult content from reaching the youngster becomes essential. In recent years, many efforts have been made to distinguish pornographic visual content from normal ones. Existing studies about pornographic visual recognition and classification can be divided into four categories depending on their approaches, namely skin-based approach, handcrafted feature-based approach, deep learning-based approach, and object-based approach.

The skin-based approach involved recognizing the exposed skin ratio to decide if there is a nude person in that image. To improve performance as well as achieve higher prediction, shape and color features can be combined and determined under mathematic and statistic thresholds. Furthermore, upper features such as facial or body organs localization can be adapted to strengthen the accuracy of skin ratio estimation. Several approaches on skin-based can be adapting various color spaces on different body areas [1], combining skin detection and face localization [2], or utilizing a pre-train discrimination model followed a skin extraction [3]. However, a high ratio of wrong-prediction based on the assumption of exposing skin rather than visual

understanding [4] reduce significantly the performance of these methods.

The handcrafted feature-based approach extracts visual features from images from the same label and maps them into a dictionary codebook. Then, a machine learning classifier is adapted to identify the pornographic elements based on that codebook's features. To describe pornographic features on images, various feature descriptors or extractors are used, such as scale-invariant feature transform (SIFT)/Hue-SIFT [5], BossaNova [6], or temporal robust features [7], which is a space-temporal detector using Fisher Vector feature representation. Despite the effectiveness of identifying pornographic content, the complexity, diversity of pornographic content as well as the omission of spatial relationships make it difficult to determine appropriate features to describe visual pornography.

Deep convolutional neural network (CNN) has been an effective method to tackle pornography recognition problems with state-of-the-art performances [8]-[13]. Rather than selecting appropriate features manually, deep neural network models extract features and refine learning parameters automatically, thus improving the model performance. These studies often use pre-trained deep CNN models on large-scale datasets such as ImageNet, common objects in context (COCO), and fine-tune with a custom small-scale dataset for a specific task, in this case, visual pornography detection. However, the sensitiveness of this research topic prevents these studies to publish their dataset widely, thus making it difficult for training, evaluating, and comparing deep learning models on the same tasks.

Recent pornography detection studies [14]-[16] focus on detecting sexual objects and organs within image/video frames, then determining the sensitiveness of input visual content. While choosing the appropriate sexual objects depended heavily on the study scale and perspective of recent studies, these studies adapted an existing object detector and fine-tuned it on a custom labeled dataset to be able to identify sexual object recognition properly. Noticeably, Tabone *et al.* [17] proposed seven sexual organs classification on images included buttocks, female breast, female genital, which are divided into two sub-classes: female genital posing and female genital active), male genital, sex toys, and non-porn (normal) class. The authors annotated those objects with five-set labeled points: one center point and four perpendicularly offset for each. However, the biggest challenge of this method is the lacking of data, as there aren't any large-scale visual datasets for sexual organ detection yet for training an effective detection model. While the object-based approaches can ensure the right prediction in most cases, the strong resemblance between sexual objects with common items in some special cases or viewpoints (such as dildo and sausage) makes it difficult to make the right prediction. In our previous studies that focus on identifying sexual objects and organs on object-based approach [18]-[20], we labeled four sexual organs male/female genitals, female breast, and anus with polygon mask for both object detection and instance segmentation tasks. With the labeled dataset, we not only developed a sexual object detector based on mask R-CNN but also utilized the training strategy with two steps learning that helps the detector overcome the false positive prediction on sexual objects, thus enhancing the performance of recognizing and classifying pornography content.

Previous studies about pornographic video recognition mostly experiment on the nucleo de procesamiento digital de imagens (NPDI) pornography datasets [6], [7]. However, these methods predominantly used the extracted key-frames that NPDI's author provided feeding to their model, rather than learning the representation throughout of the video that limited the model's performance. In our previous experiments on pornography videos [18]-[20], we extracted key point frames throughout the whole videos of NPDI instead of using provided key-frames, as we believe it comes with a better result in precision.

In this paper, we proposed a method that calculates and combines the similarity inter and intra features between videos to recognize pornography. We consider this approach to be the first of its kind in this pornographic recognition area. The input video is fed to the anabranch inter-intra feature stream. The 'intra branch' obtains appropriate video inner representative throughout the temporal axis in the frame-level, while the 'inter branch' calculates the similarities of input with a set of videos. Then, we combine features from the two branches and feed them to a classifier to determine the pornographic label. Evaluating our method on the NPDI-2k dataset, we achieved a competitive result of 96.88% accuracy.

2. METHOD

Normally, a video usually defined as a sequence of frames connect together in a temporal dimension. Thus, the basic approach is splitting video into frames and working with them for action recognition [21], searching [22], or information retrieval [23], [24]. For the methodology, we believe if there is a strong

resemblance between an unlabeled video a with a set of videos S sharing the same label l , then there is a high probability that video a have the same label l . However, the similarity, no matter how strong it is, does not fully reflect the true nature of the relationship between a and S , as some internal features of the video does affect that relationship. Therefore, we come up with an approach that leverages both inner features of the video and outer relationships with others to determine the video’s main characteristic, particularly, the eroticism of the video itself.

Figure 1 depicts the overview of our proposed method. The figure portrays the structure of the anabranch river to leverage the advantages of inter-features learning and intra-features learning. Overall, the input video is brought to the ‘inter branch’ for similarity scoring and the ‘intra’ branch for feature extracting. Outputs from these branches are then concatenated to generate the joint representation of input video. Finally, the representation is fed to a classifier for video classification. The description of our approach in detail is described at follows.

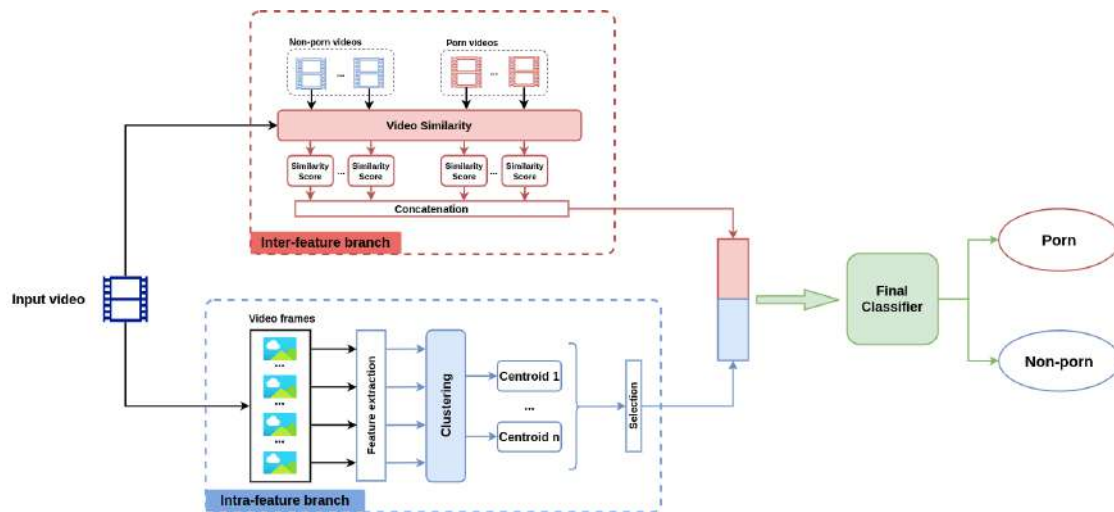


Figure 1. Proposed method’s pipeline. For selecting clusters to represent the video’s intra-feature, we either choosing the largest cluster’s centroid (method 1), or concatenating all the centroids together (method 2)

2.1. Video representation

2.1.1. Intra-feature branch

Initially, input video is fed to the intra-feature branch and split into frames. To extract internal representation of videos, our main approach is to find similar frames across the temporal dimension and cluster them to choose the feature. Although sharing the same idea with [21], rather than training a deep CNN model to learn how to cluster frames based on hamming distance, we utilized an unsupervised clustering algorithm to simplify the computational effort, thus reduce the complexity in time.

More specifically, a feature extractor is adapted to retrieve the representative vector of each frame. With all the extracted vectors, we considered them as data points and adapted the K-means clustering algorithm to cluster them – thus obtained the appropriate representative of the input video. With the divided clusters as well as their centroids, we came up with two methods for obtaining the representation vector v_{inner} , includes: i) select the largest cluster’s centroid and obtained it’s feature vector in (1), ii) obtained all the centroids’ vectors and concatenate them together in (2). The largest cluster is defined as the cluster contains the largest amount of data points (in this case, the frames’ feature vectors), and the cluster’s centroid is the means of all data points within a single cluster.

$$v_{inner} = Centroid_{\max(Cluster_1, \dots, Cluster_n)} \tag{1}$$

$$v_{inner} = concat(Centroid_1, \dots, Centroid_n) \tag{2}$$

2.1.2. Inter-feature branch

In the inter-feature branch, ViSiL [24] is adapted to calculate the spatio-temporal relations between a pair of videos. The main approach of ViSiL is estimating the pairwise frame similarity between videos by apply TensorDot and mean-max filter chamfer similarity (CS) on the region frame feature. After that, the frame-similarity matrix is then feeding to a four-layer CNN followed by CS again to obtain spatial-temporal similarity vector and score between videos. The chamfer similarity, the similarity counterpart of chamfer matching [25], is calculated by averaging similarity of the most similar item in set \mathbf{y} for each item in set \mathbf{x} to determine the closeness score. The differential between the similarity vectors and scores between pairs of relevant and irrelevant videos is presented in Figure 2.

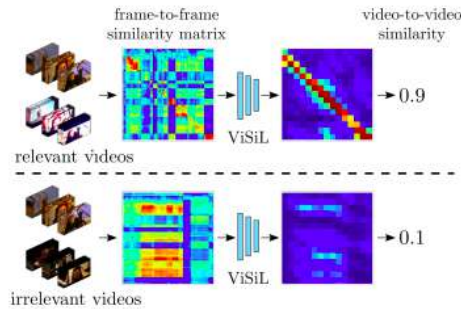


Figure 2. ViSiL spatio-temporal similarity scores [24]

For the frame-to-frame similarity, with two video frames \mathbf{a} , \mathbf{b} , the region feature maps are extracted and decomposed by into region vectors $\mathbf{a}_{i,j}$, $\mathbf{b}_{k,l}$. Then, the CS is adapted to calculate the similarity:

$$CS_{frame}(a, b) = \frac{1}{N^2} \sum_{i,j=1}^N \max_{k,l \in [1,N]} \mathbf{a}_{i,j}^T \mathbf{b}_{k,l} \quad (3)$$

after that, a frame-similarity matrix that comprising pairwise frame similarities is fed to a four-layer CNN. Finally, the element-wise *hard tanh* activation function and CS is applied on the 1D tensor of the CNN output to obtain the similarity score between pair of videos x, y :

$$CS_{video}(x, y) = \frac{1}{X'} \sum_{i=1}^{X'} \max_{j \in [1, Y']} Htanh(S^{x,y}(i, j)) \quad (4)$$

where $S^{x,y} \in \mathbb{R}^{X' \times Y'}$ indicates the output of the CNN network and *Htanh* is the *hard tanh* function.

In our proposed approach, the input video is fed to ViSiL to compare with all N videos from the training set, including 50% pornographic video and 50% non pornographic ones. All the similarity scores are then concatenated into an external feature \mathbf{v}_{outer} with N dimension:

$$\mathbf{v}_{outer} = concat(S_{(x,1)}, S_{(x,2)}, \dots, S_{(x,N)}) \quad (5)$$

where $S_{(x,i)}$ is the similarity score between input video x and the i^{th} video.

2.1.3. Joint representation

The joint representation vector is calculated by concatenating intra and inter features together in (6), which outcome is a $x + y$ dimensional vector where x and y are the dimension of the features, respectively. The detailed of x and y will be discussed in the experimental section:

$$\mathbf{v}_{joint} = concat(\mathbf{v}_{inner}, \mathbf{v}_{outer}) \quad (6)$$

finally, the concatenated representation between inter and intra vector is then fed to the final classifier to determine if the input video is pornographic or not.

2.2. Final classification model

For the final classification model, we leveraged two classifying models multi-layer perceptron (MLP) and support vector machine (SVM) to discriminate the status of input video with the joint representation. The MLP architecture is a three-layer neural network with each layer dimensions are $(x, 256)$, $(256, 32)$, and $(32, 1)$ respectively, where x is the size of the input representation. The early two layers adapts leaky ReLU as an activation function and is followed by a normalization layer, while the last layer adapts sigmoid for binary classification in Figure 3.



Figure 3. The 3-layer MLP architecture

3. EXPERIMENTAL RESULTS

In the experimental, we trained and evaluated our model using the NPDI-2k dataset [7], which contains 1,000 pornographic videos and 1,000 non pornographic videos. The NPDI-2k videos range from several seconds to thirty minutes approximately, with the frame rates from 15 to 25 FPS. For the evaluation process, the two-fold cross-validation is applied five times, which is similar to the experimental scenarios used in [7], and the final outcome is the mean of five performances. For each cross-validation phase, the number of video is divided equally by two, which makes both training and testing sets contain 1,000 videos (with 500 porn and 500 non-porn videos) for each. To reduce the computational cost, rather than utilizing every single frame per video, we extracted only one frame per second for both similarity calculation and intra-feature extraction.

During the testing process, we utilized ResNet101 and DenseNet121 for feature extraction in the intra branch, with two adapted models are pre-trained on the ImageNet. The obtained feature for each frame has a size of 2,048 or 1,024 respectively. Then, applied the K-mean clustering on these feature vectors, we experimented only with 2 and 3 clusters to ensure the quality of video inner representation. The reason why we selected up to 3 clusters for the experiment not only to maintain the performance and computational cost of our method but also because the minimum amount of frames we could extract from a single video are 3 (equivalent to a three-second video). The inner feature vector shares the same dimension with the corresponded frame feature, depending on the adapted feature extractor. On the other hand, the outer feature created by calculating similarity scores through the inter branch is a 1,000 dimensional vector.

For the final classifier, on the one hand, the MLP model was trained on colab pro with P100 GPU, with the configuration includes 800 epochs, learning rate 0.005, and batch size of 32. On the other hand, the SVM model is utilized with four kernels: linear, polynomial, radial basis function (RBF), and sigmoid. With the outcome prediction is the binary label, we expected that the linear kernel comes with the highest performance among the four.

Table 1. Overall results

Inner representative method ¹	Intra-feature extractor	Classification model	Performance with 2 Clusters ² (Acc)	Performance with 3 Clusters ² (Acc)
Choosing the largest cluster's centroid (method 1)	ResNet101	MLP	96.76	96.76
		SVM ³	95.04	95.68
	DenseNet121	MLP	96.50	96.46
		SVM	95.72	95.22
Concatenating all the clusters' centroids (method 2)	ResNet101	MLP	96.80	96.88
		SVM	95.52	95.56
	DenseNet121	MLP	96.74	96.50
		SVM	95.78	95.62

Acc – Accuracy; MLP – Multi-layer Perceptron; SVM – Support Vector Machine

¹ Method to determine the inner representation after clustering frame-features

² Model performance with corresponded number of inner-feature cluster

³ All the results that use SVM model only utilize the linear kernel

Overall, the frame's feature retrieval with ResNet network achieves better results than their counterpart DenseNet, while the concatenating of all clusters' centroids – after cluster all frame-features using K-means clustering – helps model achieve higher Accuracies than only choosing the largest one (Table 1). Also, dividing into 3 groups for the clustering algorithms comes with higher performances than 2. In the final classification model, our approach's results when using MLP are greater than using SVM to classify videos. As we expected, the Linear kernel comes with the highest performance (Table 2). Eventually, the highest result of our approach is 96.88% Accuracy with Resnet101 feature extraction; 3 groups clustering; all centroids concatenation for intra-representation; and MLP classifier. In the comparison with other methods on the NPDI-2k dataset, our method achieved a competitive result (Table 3) with the second-highest performance in comparison with other methods.

Table 2. SVM's kernels performance comparison

Kernel	Performance (Accuracy)
Linear	95.56
RBF	91.40
Poly	61.00
Sigmoid	94.28

With the detail configuration includes:
Resnet101 feature extraction; SVM classifier;
3 groups clustering; and all centroids concatenation

Table 3. Comparison results on the NPDI-2k dataset

Method*	Performance (Accuracy)
Open-NSFW + Mask R-CNN [18]	90.40
PROS + MFCC + HOG + TRoF [26]	90.75
1-Tiered adult detector [9]	91.50
Space-Time Interest Points [7]**	94.52
VGG-16 + Bi-RNN [27]	95.33
Dense TRoF [7]**	95.58
Dense Trajectories [8]	95.80
Two-stream CNN Late Fusion [8]	96.40
Inter-intra Joint Representation (Our Method)	96.88
AttM-CNN-Porn [10]	97.10

* All experiment results are organized by publication years.

** Results from the paper that proposed the NPDI-2k dataset

4. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel approach to identify pornographic videos that calculating the joint representation of internal and external video's features. While the intra-features of video can be obtained by extracting features in the frame-level with a pre-trained deep learning model and cluster them together, the inter-similarity between pair of videos is calculated using mean-max filter chamfer similarity via a spatio-temporal video similarity architectural named ViSiL. Both inner and outer features are then concatenated together. After that, the joint representation vector is fed to a classifier for video discrimination. Experiments with NPDI-2K dataset, our approach demonstrates a competitive performance with recent results, 96.88 % Accuracy in prediction. We hope our approach could be an intial step in developing a better method for video detection and classification, especially in the pornographic classification manner.

However, there are still works to be done. The computation for inter representation costs a large amount of resources. Therefore, our priority is to reduce the computational cost, while maintaining its performance. Moreover, recent state-of-the-art methods can be used to improve the effectiveness of the final classifier beside the MLP and SVM methods, so that better performance can be achieved.

ACKNOWLEDGEMENT

This research is funded by Vietnam National University Ho Chi Minh City (VNU-HCM) under grant number B2019-26-02.




REFERENCES

- [1] D. C. Moreira and J. M. Fechine, "A machine learning-based forensic discriminator of pornographic and bikini images," In *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, pp. 1–8, doi: 10.1109/IJCNN.2018.8489100.
- [2] R. Balamurali and A. Chandrasekar, "Multiple parameter algorithm approach for adult image identification," *Cluster Computing*, vol. 22, no. 5, pp. 11909–11917, 2019, doi: 10.1007/s10586-017-1510-3.
- [3] K. Zhou, L. Zhuo, Z. Geng, J. Zhang, and X. G. Li, "Convolutional neural networks based pornographic image classification," In *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, 2016, pp. 206–209, doi: 10.1109/BigMM.2016.29.
- [4] A. Zaidan, H. A. Karim, N. Ahmad, B. Zaidan, and A. Sali, "An automated anti pornography system using a skin detector based on artificial intelligence: A review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 27, no. 04, p. 1350012, 2013, doi: 10.1142/S0218001413500122.
- [5] A. P. Lopes, S. E. de Avila, A. N. Peixoto, R. S. Oliveira, and A. d. A. Araújo, "A bag-of-features approach based on hue-sift descriptor for nude detection," In *2009 17th European Signal Processing Conference*, 2009, pp. 1552–1556.
- [6] S. Avila, N. Thome, M. Cord, E. Valle, and A. D. A. Araújo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013, doi: 10.1016/j.cviu.2012.09.007.
- [7] D. Moreira *et al.*, "Pornography classification: The hidden clues in video space-time," *Forensic science international*, vol. 268, pp. 46–61, 2016, doi: 10.1016/j.forsciint.2016.09.010.
- [8] M. Perez *et al.*, "Video pornography detection through deep learning techniques and motion information," *Neurocomputing*, vol. 230, pp. 279–293, 2017, doi: 10.1016/j.neucom.2016.12.017.
- [9] P. Vitorino, S. Avila, M. Perez, and A. Rocha, "Leveraging deep neural networks to fight child pornography in the age of social media," *Journal of Visual Communication and Image Representation*, vol. 50, pp. 303–313, 2018, doi: 10.1016/j.jvcir.2017.12.005.
- [10] A. Gangwar, V. González-Castro, E. Alegre, and E. Fidalgo, "Attn-cnn: Attention and metric learning based CNN for pornography, age and child sexual abuse (CSA) detection in images," *Neurocomputing*, vol. 445, pp. 81–104, 2021, doi: 10.1016/j.neucom.2021.02.056.
- [11] N. Kamaruddin, A. Wahab, and Y. Rozaidi, "Neuro-physiological porn addiction detection using machine learning approach," *Indonesian Journal of Electrical Engineering and Computer Science (IJECS)*, vol. 16, no. 2, pp. 964–971, 2019, doi: 10.11591/ijeecs.v16.i2.pp964-971.
- [12] F. Nian, T. Li, Y. Wang, M. Xu, and J. Wu, "Pornographic image detection utilizing deep convolutional neural networks," *Neuro-computing*, vol. 210, pp. 283–293, 2016, doi: 10.1016/j.neucom.2015.09.135.
- [13] J. Mahadeokar and G. Pesavento, "Open sourcing a deep learning solution for detecting nsfw images." yahoeng.tumblr.com. <https://yahoeng.tumblr.com/post/151148689421/open-sourcing-a-deep-learning-solution-for> (accessed May 1, 2021).
- [14] H. A. Nugroho, D. Hardiyanto, and T. B. Adji, "Nipple detection to identify negative content on digital images," In *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pp. 43–48, 2016, doi: 10.1109/ISITIA.2016.7828631.
- [15] Y. Wang, X. Jin, and X. Tan, "Pornographic image recognition by strongly-supervised deep multiple instance learning," In *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 4418–4422, doi: 10.1109/ICIP.2016.7533195.
- [16] C. Tian, X. Zhang, W. Wei, and X. Gao, "Color pornographic image detection based on color-saliency preserved mixture deformable part model," *Multimedia Tools and Applications*, vol. 77, no. 6, pp. 6629–6645, 2018, doi: 10.1007/s11042-017-4576-2.
- [17] A. Tabone, A. Bonnici, S. Cristina, R. A. Farrugia, and K. P. Camilleri, "Private body part detection using deep learning," In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods - ICPRAM*, 2020, pp. 205–211, doi: 10.5220/0009101502050211.
- [18] Q.-H. Nguyen, K.-N.-K. Nguyen, H.-L. Tran, T.-T. Nguyen, D.-D. Phan, and D.-L. Vu, "Multi-level detector for pornographic content using cnn models," In *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020, pp. 1–5, doi: 10.1109/RIVF48685.2020.9140734.
- [19] H. L. Tran, Q. H. Nguyen, D. D. Phan, T. T. Nguyen, K. N. K. Nguyen, and D. L. Vu, "Additional learning on object detection: A novel approach in pornography classification," In *International Conference on Future Data and Security Engineering*, pp. 311–324, 2020, doi: 10.1007/978-981-33-4370-2_22.
- [20] D.-D. Phan, T.-T. Nguyen, Q.-H. Nguyen, H.-L. Tran, K.-N.-K. Nguyen, and D.-L. Vu, "A novel pornographic visual content classifier based on sensitive object detection," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 5, 2021, doi: 10.14569/IJACSA.2021.0120591.
- [21] X. Liu, S. L. Pintea, F. K. Nejedasl, O. Booi, and J. C. van Gemert, "No frame left behind: Full video action recognition," In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 14892–14901, doi: 10.1109/CVPR46437.2021.01465.
- [22] D. Mohammad, I. Aljarrah, and M. Jarrah, "Searching surveillance video contents using convolutional neural network," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 11, no. 2, pp. 1656–1665, 2021, doi: 10.11591/ijece.v11i2.pp1656-1665.
- [23] W. Widiarto, M. Hariadi, and E. M. Yuniarno, "Keyframe selection of frame similarity to generate scene segmentation based on point operation," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, no. 5, pp. 2839–2846, 2018, doi: 10.11591/ijece.v8i5.pp2839-2846.
- [24] G. Kordopatis-Zilos, S. Papadopoulos, I. Patras, and I. Kompatsiaris, "Visil: Fine-grained spatio-temporal video similarity learning," In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6351–6360, doi: 10.1109/ICCV.2019.00645.
- [25] H. Barrow, J. Tenenbaum, R. Bolles, and H. Wolf, "Parametric correspondence and chamfer matching: Two new techniques for image matching," In *Proceedings of the 5th international joint conference on Artificial intelligence*, 1977, pp. 659–663.




- [26] D. Moreira *et al.*, "Multi-modal data fusion for sensitive scene localization," *Information Fusion*, vol. 45, pp. 307–323, 2019, doi: 10.1016/j.inffus.2018.03.001.
- [27] K. Song and Y.-S. Kim, "An enhanced multimodal stacking scheme for online pornographic content detection," *Applied Sciences*, vol. 10, no. 8, p. 2943, 2020, doi: 10.3390/app10082943.

BIOGRAPHIES OF AUTHORS






Dinh-Duy Phan    received B. Eng. degree in computer engineering in 2011 and an M. Sc. degree in computer science in 2014 from the University of Information Technology, Vietnam National University Ho Chi Minh City (VNU-HCM), Ho Chi Minh City, Vietnam. In 2011, he became lecture of computer engineering faculty, University of Information Technology, VNU-HCM. His research interests include embedded system design, IC design, machine learning, computer vision and their applications on PC and embedded system. He can be contacted at email: duydp@uit.edu.vn






Quang-Huy Nguyen    received his B. Eng. in computer engineering from the University of Information Technology, Vietnam National University Ho Chi Minh City (VNU-HCMC), Ho Chi Minh City, Vietnam, in 2020. He is currently working as a post-baccalaureate research assistant for the faculty of computer engineering, University of Information Technology, VNU-HCMC, while aiming to pursue Ph. D. in computer science. His research interests is working with trend computer vision approaches such as object detection, instance segmentation, few-shot learning, vision transformer, and multi-modal learning. He can be contacted at email: 15520306@gm.uit.edu.vn






Thanh-Thien Nguyen    received B. S. degree in information technology in 2013 and M. Sc. degree in computer science in 2018 from the University of Science, VNU-HCM, HCM City. From 2013 to 2016, he was a researcher in the Software Engineering Laboratory at the University of Science. From 2016 to 2018, he had been a teaching assistant at the embedded systems and robotics department, faculty of computer engineering, University of Information Technology, and became a lecturer in 2018. His research interests include machine learning, computer vision and their applications. He can be contacted at email: thiennt@uit.edu.vn.



Hoang-Loc Tran    received a B. Eng. degree in computer engineering from the University of Information Technology, Vietnam National University Ho Chi Minh City (UIT VNU-HCMC) in 2018. In 2021, he received his M. Sc. degree in Computer Science from UIT VNU-HCMC. From 2018 he started working as a researcher and teaching assistant at the faculty of computer engineering, UIT VNU-HCMC. His research interests include machine learning, computer vision, edge computing, and embedded system. He can be contacted at email: locth@uit.edu.vn.



Duc-Lung Vu    received B. S. and M. Sc. degrees in computer engineering from the Peter the Great St.Petersburg Polytechnic University in 1998 and 2000, respectively. He got his Ph. D. in computer science from Saint Petersburg Electrotechnical University in 2006. He has been working at the University of Information Technology, Vietnam National University Ho Chi Minh City, as an associate professor since 2015 and chancellor of the school since 2020. His research interests include machine leaning, human-computer interaction, embedded systems and digital system design on FPGA. He can be contacted at email: lungvd@uit.edu.vn