A VIDEO SUMMARIZATION APPROACH TO SPEED-UP THE ANALYSIS OF CHILD SEXUAL EXPLOITATION MATERIAL

Rubel Biswas[1], Deisy Chaves[1], Laura Fernández-Robles[2], Eduardo Fidalgo[1], Enrique Alegre[1]
[1]Department of Electrical, Systems and Automation, Universidad de León, León, ES
[2]Department of Mechanical, Informatics and Aeroespace Engineering, Universidad de León, León, ES
{rubel.biswas, deisy.chaves, l.fernandez, eduardo.fidalgo, enrique.alegre}@unileon.es

Abstract

Identifying key content from a video is essential for many security applications such as motion/action detection, person re-identification and recognition. Moreover, summarizing the key information from Child Sexual Exploitation Materials, especially videos, which mainly contain distinctive scenes including people's faces is crucial to speed-up the investigation of Law Enforcement Agencies. In this paper, we present a video summarization strategy that combines perceptual hashing and face detection algorithms to keep the most relevant frames of a video containing people's faces that may correspond to victims or offenders. Due to legal constraints to access Child Sexual Abuse datasets, we evaluated the performance of the proposed strategy during the detection of adult pornography content with the NDPI-800 dataset. Also, we assessed the capability of our strategy to create video summaries preserving frames with distinctive faces from the original video using ten additional short videos manually labeled. Results showed that our approach can detect pornography content with an accuracy of 84.15% *at a speed of* 8.05 *ms/frame making this appropriate for real-time applications.*

Keywords: Video summarization, pHash, MTCNN, Child Sexual Exploitation Material, Pornography detection.

1 INTRODUCTION

The use of anonymous tools and private networks by criminals has considerably increased the number of images and videos with illegal content, like Child Sexual Exploitation Material (CSEM), on the Internet [8, 3]. The impact of exposure to this kind of contents is still under investigation [17]. Moreover, since CSEM is illegal in most countries, it has triggered the demand for techniques that enable investigators or Law Enforcement Agencies (LEAs) to browse and detect CSEM content on images and videos quickly.

Literature presents approaches for detecting pornography content or CSEM automatically in images and videos with high accuracy [20, 12, 14, 13, 18]. Nevertheless, they require a significant amount of memory and CPU/GPU capabilities. Besides, most of these approaches [20, 14, 18] have not been assessed in terms of processing time, which limits their use in forensic laboratories where processing times are essential to detect CSEM given the large volume of investigations related to child exploitation.

To overcome these drawbacks, this paper presents a video summarization strategy based on perceptual hashing and face detection algorithms. The strategy generates highly compressed video summaries focused on people's faces, and can perform pornography detection in real-time balancing the trade-off between processing time and accuracy. Existing approaches for video summarization preserve the distinctive scenes of the whole original video [9, 10, 19, 1]. However, our strategy focuses only on the most relevant frames containing faces which may correspond to victims or offenders on CSEM.

Due to the legal constraints to access CSEM and given the sensitive nature of this kind of material, we evaluated the performance of the generated video summaries for the detection of adult pornography on the NDPI-800 dataset [12]. Besides, we created a dataset with ten videos labeled manually with the number of subjects observed, i.e., people with different identity, to analyze the capability of our approach to generate summaries preserving the subjects from the original video.

This work is part of the European project Forensic Against Sexual Exploitation of Children (4NSEEK) and the research lines defined by the Framework agreement between INCIBE (Spanish National Cybersecurity Institute) and the University of León. Therefore, results of this study may help the 4NSEEK tool users to speed-up the detection of victims or offenders on CSEM videos.

The rest of the paper is organized as follows. Closely related work to the one addressed in this paper is presented in Section 2. The proposed video summarization strategy is described in Sec-

https://doi.org/10.17979/spudc.9788497498043.648

tion 3. Experimental evaluation and results are presented in Section 4. Finally, we draw conclusions in Section 5.

2 RELATED WORK

2.1 VIDEO PORNOGRAPHY DETECTION

In the earliest stage of pornography detection, research has been associated with nudity detection [21, 11]. However, these skin-based approaches are not a suitable solution for pornography detection because skin exposure may not always be related to pornographic scenes, e.g., baby breastfeeding, wrestling, and swimming.

Later on, Bags-of-Visual-Words (BoVW) models created robust representations for detecting pornography content using hand-crafted local descriptors [20, 12]. Zhuo et al. [20] introduced a web pornographic image recognition scheme based on the combination of local and global features with a Support Vector Machine classifier. The local features correspond to Oriented FAST and Rotated BRIEF descriptors computed from skin-color regions and represented using BoVW. The global features considered were color descriptors such as Hue, Saturation, Value, or HSV. Moreira et al. [12] proposed an end-to-end BoVW-based framework with a speed of 0.5 sec/frame that integrates temporal information during the detection of pornography videos. Nonetheless, these methods used bags of static hand-crafted features and may disregard significant information to represent the video content.

Recently, Convolutional Neural Network (CNN)-based methods have been used for video pornography detection because of their outstanding performance in comparison to traditional approaches based on hand-crafted features [14, 13]. Perez et al. [14] presented a video pornography detection strategy that combines static and motion-based features for CNN-based video classification. In contrast, Mallmann et al. [13] proposed a CNN architecture, called private Parts Censor (PPCensor), to detect private body parts in real-time environments from pornographic content with a speed of 0.092 sec/frame. PPCensor was implemented as a network proxy server for video streams. Therefore, it can be run through a proxy instead of on the end-user device.

However, few of these approaches [13, 12] assess the processing time, which is crucial during the analysis of CSEM videos due to the large number of cases related to child exploitation in forensic laboratories. In this work, we explore the use of video summaries focused on people's faces to speed-up the analysis of pornography and CSEM videos.

2.1.1 VIDEO SUMMARY

There exist several summarization methods in the literature aiming to generate a short video skim that preserve the most relevant content of the original video [10, 19, 1].

Gygli et al. [10] proposed a supervised approach for learning the importance of global characteristics of a summary. It jointly combines multiple properties including saliency, aesthetics, and presence of people in the frames.

Most recently, Abhimanyu et al. [1] proposed an algorithm for generating multiscale summaries and priority-based ranking of various actions present in egocentric videos, like leading towards the goal (high priority action). Wei et al. [19] introduced a semantic attended video summarization network that comprises a frame selector and a video descriptor to maximize the diversity and representativeness in the summary.

Most of these approaches focused on summarizing videos with general content [9, 10, 19] without focusing on particular details, such as faces that may be relevant for tasks related to facial recognition. Therefore, in this study, we propose a strategy to generate video summaries containing only distinctive scenes with people faces that may be more appropriate for the detection of victims or offenders on CSEM than regular video summaries.

2.2 PERCEPTUAL HASHING

The perceptual hashing approach generates a fixed-length fingerprint, i.e., a hash code based on the perceptual content of the image/video/audio. In the last few years, perceptual hashing has been used in different applications such as tampering detection [15], person re-identification [7], victim identification [4], or illegal Tor domain classification [5].

Sandeep and Prabin [15] proposed a video hashing method to detect malicious video modifications using the three-dimensional radial projection technique and the two-dimensional discrete cosine transform. Fang et al. [7] proposed a multi-statistics on hash feature map descriptor for person re-identification using binarized low-level color and gradient feature maps obtained with perceptual hashing, and regional statistics computed over an image pyramid.

In this work, we used perceptual hashing to identify video frames with similar content during the

https://doi.org/10.17979/spudc.9788497498043.648

video summarization.

3 PROPOSED STRATEGY

We proposed a three-step video summarization strategy focused on the identification of distinctive scenes that contain people's faces, as shown in Figure 1. First, videos are preprocessed to reduce their size and the number of frames per second based on the videos' duration. Second, keyframes are detected using perceptual hashing. Finally, keyframes are filtered to identify the ones containing faces using a deep-learning-based face detection method.

3.1 PREPROCESSING

Given a video, we sampled and kept a set of frames depending on the video duration. We also reduced the frames size proportionally to a fixed width and a variable height, to keep the proportions of faces and objects contained on the frame, and speed-up the later summarization process. Table 1 describes the three different scenarios considered to preprocess videos after evaluating different frame sample rates and sizes to balance between speed and accuracy.

Table 1: Preprocessing video conditions.

Video duration	Preprocesing	
(min)	Width resized	Frames to sample
3	480 pixels	5 frames per second
3-30	320 pixels	2 frames per second
+30	320 pixels	3600 frames on the whole video

3.2 DETECTION OF KEYFRAMES

We detected the keyframes in a preprocessed video by obtaining the perceptual hash codes of all the frames using the pHash algorithm [22]. pHash is a perceptual hashing method based on the Discrete Cosine Transform (DCT) that generates hash codes in three-steps. First, the input frame is converted to grayscale and resized to 32×32 pixels. Second, a two-dimensional type-II DCT is applied to obtain the DCT coefficients of the frame. Third, 64 low-frequency DCT coefficients are computed by omitting high-frequency coefficients, to generate the hash code. We selected the low-frequency coefficients because they contain more useful information for identifying keyframes, such as shape and position of the objects in the frame foreground.

After obtaining the hash codes for the video frames, we computed Euclidean distances among all of them. Later on, only the first and those frames which are distinctive to the rest with respect to a threshold, T_{dist}, are kept. This set of frames makes up an initial summary.

3.3 SELECTION OF KEYFRAMES CONTAINING FACES

We selected the keyframes containing at least one face, detected with a confidence score, T, to create the final video summary. As face detector, we chose the Multi-Task Cascade CNN (MTCNN) [23] method due to its high detection speed in comparison to other deep-learning-based detectors [6].

MTCNN detects faces in three steps. First, candidate regions, which may contain faces, are produced through a fast Proposal Network. Second, these candidate regions are refined through a Refinement Network. Third, an Output Network produces the final bounding box of a face.

The obtained video summary is highly compressed and can be used as the initial step to another process, e.g., in a forensic tool to detect CSEM in seized material.

4 EXPERIMENTS AND RESULTS

4.1 EXPERIMENTAL SETUP

Experiments were performed on a GNU/Linux machine box running Ubuntu 18.04, using a CPU Intel Xeon E5-2630 and a GPU NVIDIA Titan XP 12Gb. Besides, we considered two criteria to evaluate the proposed video summarization strategy: (i) the impact of video summaries on the pornography detection performance, and (ii) the quality of video summaries in terms of the number of subjects contained in the summary.

We assessed the performance of our approach with a similarity threshold T_{dist} of 0.2 to detect keyframes and a confidence score T of 0.7 to select the keyframes containing faces, see Section 3.

Moreover, we compared the proposed strategy against two methods. First, we used, as a baseline, video summaries generated by uniformly sampling a video clip every 20 frames from the original video and keeping the frames with faces as described in Section 3.3. Uniform sampling is one of the most common methods for keyframe extraction employed as a baseline for video summarization [16]. Second, to evaluate the performance of the preprocessing stage, in the proposed approach, we used video summaries created by selecting frames that contain faces as detailed in Section 3.3 from a preprocessed video as described in Section 3.1.

https://doi.org/10.17979/spudc.9788497498043.648

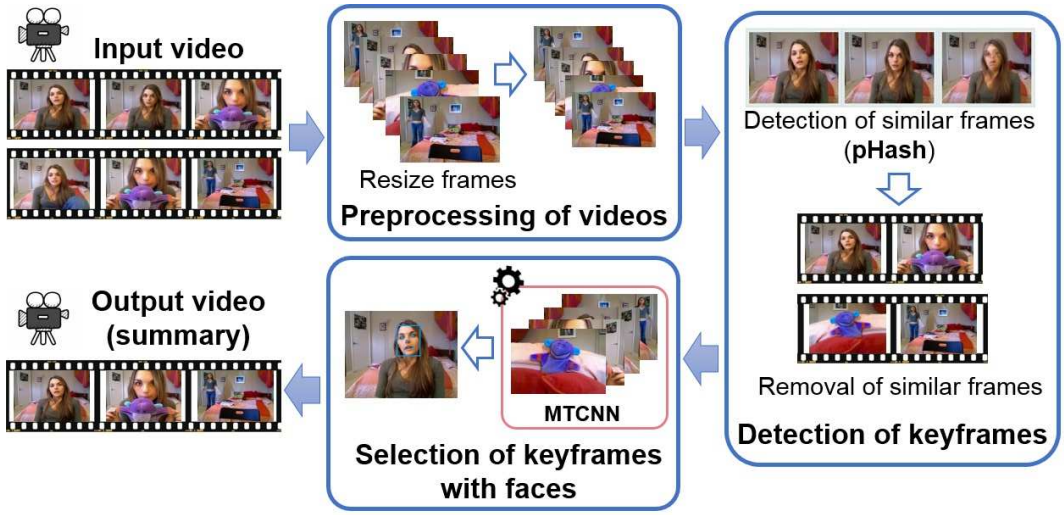

Figure 1: Stages of the proposed video summarization strategy.

To test a video sequence, first, we generated a video summary using one of the evaluated methods. Second, we classified the content of each video summary frame, as pornographic or non-pornographic, using a CNN-based adult pornography detector integrated in the 4NSEEK tool [24]. Finally, the video was labeled based on majority voting.

4.2 EVALUATION OF THE IMPACT OF VIDEO SUMMARIES ON THE PORNOGRAPHY DETECTION PERFORMANCE

We evaluated the pornography detection performance on summarized videos through four metrics: Accuracy (*Acc.*), F2-Score, and total and average processing times per frame required to summarize videos and detect pornography content.

The *Acc.* is measured as the number of correctly classified pornography video summaries out of the total number of evaluated videos. While, the *F2-score* is the weighted harmonic mean of the precision and the recall measures, considering twice the weight of the recall compared to the precision [12].

The evaluation metrics were computed for each of the five folds defined in the NPDI Pornography-800 dataset [2]. This dataset [2] comprises 76.86 hours of 400 pornography and 400 non-pornography videos.

Table 2 presents the averaged results over the five folds. As it can be observed, the proposed video summarization strategy obtained a better performance (average Acc. of 84.15% and *F2-score* of 77.18%) compared to the baseline (average Acc. of 83.39% and F2-Score of 75.06%). Besides, the use of pHash to identify keyframes and remove redundant information generates highly compressed video summaries improving the Acc. and the *F2-score* during pornography detection in comparison to only using the preprocessing and the face selection (MTCNN) steps.

In this case, we observed that the baseline method required less time for summarizing a video and detecting pornography in comparison to our approach. The baseline method generates summaries through uniformly sampling a frame every 20 frames from the original video. In contrast, the proposed approach sampled a variable number of frames initially depending on the duration of the video (see Section 3.1). Despite this, our approach has a processing speed of 8.05 ms/frame, which makes it suitable for real-time applications.

4.3 EVALUATION OF THE QUALITY OF VIDEO SUMMARIES

We evaluated the quality of the generated video summaries considering two metrics: subject detection rate, and percentage of summary frames with people.

The *subject detection rate* is computed as the rate between the total number of subjects present in the summarized video and the total number of subjects in the original video. The *percentage of summary frames with people* is obtained as the total number of frames with faces in the summarized video divided by the total number of frames in the summary.

To conduct the quality evaluation, we created a dataset with 10 short videos with a total of 18.6 minutes (33476 frames). Videos were labeled manually with the number of observed subjects and

https://doi.org/10.17979/spudc.9788497498043.648

Table 2: Average No. of frames per video summary, Accuracy, F2-Score, total time and time per frame for generating video summaries (summary) and detecting pornography (Porn. detection) computed for the five folds of the NDPI dataset.

Summarization approach	No. frames summary	Accuracy (%)	F2-Score (%)	Total time (ms.)		Avg. time per frame (ms.)	
				Summary	Porn. detection	Summary	Porn. detection
Uniform sample 5% + MTCNN (baseline)	**16896.60±760.39**	83.39 ± 3.32	75.06 ± 5.87	**6671310**	**357820**	**4.24**	**0.23**
Preprocesing + MTCNN	27739.80 ± 1109.38	84.03 ± 2.97	76.58 ± 4.46	12651700	689900	8.05	0.44
Preprocesing + pHash + MTCNN (ours)	17451.00 ± 1253.25	**84.15±2.10**	**77.18±3.36**	12572020	406160	8.05	0.26

Table 3: No. of frames, No. of subjects, subject detection rate, and percentage of summary frames with people in video summaries.

Summarization approach	No. frames summary	No. subjects summary	Subject detection rate (%)	Summary frames with people (%)
Uniform sample 5% + MTCNN (baseline)	501	102	85.00	97.21
Preprocessing + MTCNN	1673	**109**	**90.83**	**97.49**
Preprocessing + pHash + MTCNN (ours)	**139**	76	63.33	94.96

contain a total of 120 subjects. Table 3 shows the results of the evaluation metrics on this dataset.

The proposed approach has a low value for the *subject detection rate* metric (63.33%) in comparison to the values achieved by the baseline method (85.00%), and the summaries generated only using the preprocessing and the face selection (MTCNN) steps (90.83%). This behavior may be caused at the selection of keyframes and faces in our approach since some selected keyframes with pHash may correspond to low-intensity frames and frames with faces in different orientations where the MTCNN face detector failed. Nonetheless, the proposed approach proved to generate summaries with relevant information that yield a better performance in terms of accuracy and F2-Score for pornography detection purposes (see Section 4.2).

We noticed (see Table 3) that the baseline method generated summaries with 501 frames whereas our approach generated summaries with only 139 frames (highly compressed). Hence, the proposed approach can be used as an initial step for other time-consuming tasks such as face recognition or age estimation.

5 CONCLUSIONS

In this work, we presented a video summarization approach that combines pHash and the MTCNN face detector to create video summaries comprised of keyframes that contain people's faces. The aim is to support LEAs by speeding-up the detection of CSEM in videos.

Experimental results showed that our approach obtained the best trade-off between processing time (8.05 ms/frame) and pornography detection accuracy (84.15%) on the NDPI-800 dataset. Therefore, it can be suitable to support the detection of pornography and CSEM in videos in real-time.

As future work, robust face detectors will be evaluated to preserve a higher number of subjects in the video summaries.

ACKNOWLEDGEMENTS

This work was supported by the framework agreement between the Universidad de León and INCIBE (Spanish National Cybersecurity Institute) under Addendum 01. Also, this research has been funded with support from the European Commission under the 4NSEEK project with Grant Agreement 821966. This publication reflects the views only of the authors, and the European Commission cannot be held responsible for any use which may be made of the information contained therein. Finally, we acknowledge the NVIDIA Corporation for the donation of the TITAN Xp GPU.

References

[1] Abhimanyu, S., Chowdhury, A. S. (2020) "Multiscale Summarization and Action Ranking in Egocentric Videos", *IEEE Multimedia*, 133, pp 256-263.

[2] Avila, S., Thome, N., Cord, M., Valle, E., Araújo, A.A. (2013) "Pooling in image representation: the visual codeword point of view", *Computer Vision Image Understanding*, 117, pp 453-465.

[3] Al-Nabki, W., Fidalgo, E., Alegre, E. and Fernández-Robles, L. (2019) "ToRank: Identifying the most influential suspicious domains in the Tor network", *Expert Systems with Applications*, 123, pp 212 - 226.

https://doi.org/10.17979/spudc.9788497498043.648

[4] Biswas, R., González-Castro, V., Fidalgo, E.,Chaves, D. (2019) "Boosting child abuse victim identification in Forensic Tools with hashing techniques", *V Jornadas Nacionales de Investigación en Ciberseguridad (JNIC)*, 272, pp 344-345.

[5] Biswas, R., González-Castro, V., Fidalgo, E. and Alegre, E. (2020) "Perceptual image hashing based on frequency dominant neighborhood structure applied to Tor domains recognition", *Neurocomputing*, 27, pp 778–790.

[6] Chaves, D., Fidalgo, E., Alegre, E., Jáñez-Martino, F., Alaiz-Rodríguez, R., Azzopardi, G. (2020) "Assessment and Estimation of Face Detection Performance Based on Deep Learning for Forensic Applications", *Sensors*, 20 (16).

[7] Fang, W., Hu, H.-M., Hu, Z., Liao, S., Li, B. (2018) "Perceptual hash-based feature description for person re-identification", *Perceptual hash-based feature description for person re-identification*, 272, pp 520–531.

[8] Gangwar, A., Fidalgo, E., Alegre, E., González-Castro, V. (2017) "Pornography and Child Sexual Abuse Detection in Image and Video: A Comparative Evaluation", *8th International Conference on Imaging for Crime Detection and Prevention (ICDP)*, pp 37-42.

[9] Gong, Y., Liu, X. (2001) "Video summarization with minimal visual content redundancies", *In Processing of the IEEE International Conference on Image Processing (ICIP)*, pp 362 - 365.

[10] Gygli, M., Grabner, H., Gool, L. V. (2015) "Video summarization by learning submodular mixtures of objectives", *In Computer Vision and Pattern Recognition (CVPR)*, pp 3090 - 3098.

[11] Lee, J.-S., Kuo, Y.-M., Chung, P.-C., Chen, E.-L. (2007) "Naked image detection based on adaptive and extensible skin color model", *Pattern recognition*, 40, pp 2261 - 2270.

[12] 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, pp 46-61.

[13] Mallmann, J. and Others (2020) "PPCensor: Architecture for Real-Time Pornography Detection in Video Streaming", *Future Generation Computer Systems*, 112, pp 945 - 55.

[14] Perez, M. and Others (2017) "Video Pornography Detection through Deep Learning Techniques and Motion Information", *Neurocomputing*, 230, pp 279 - 293.

[15] Sandeep, R., Prabin, K.B. (2020) "Detection of Malicious Video Modifications using Perceptual Video Hashing", *5th International Conference on Computing, Communication and Security (ICCCS)*, pp 1-5.

[16] Shruti, J., Jasim, M. (2020) "Unsupervised video summarization framework using keyframe extraction and video skimming", *IEEE 5th International Conference on Computing Communication and Automation (ICCCA)*, pp 140-145.

[17] Tyson, G. and Elkhatib, Y. and Sastry, N. and Uhlig, S. (2015) "Are people really social in porn 2.0?", *The International AAAI Conference on Web and Social Media*, pp 436-444.

[18] Wehrmann, J., Simoes, G.S., Barros, R.C., Cavalcante, V.F. (2018) "Adult content detection in videos with convolutional and recurrent neural networks", *Neurocomputing*, 272, pp 432-438.

[19] Wei, H., Ni, B., Yan, Y., Yu, H., Yang, X., Yao, C. (2018) "Video Summarization via Semantic Attended Networks", *In AAAI Conference on Artificial Intelligence*, pp 216-223.

[20] Zhuo, L., Geng, Z., Zhang, J., Li, X.G. (2016) "ORB feature based web pornographic image recognition", *Neurocomputing*, 173, pp 511-517.

[21] Zheng, H., Daoudi, M., Jedynak, B. (2004) "Blocking adult images based on statistical skin detection", *Electronic Letter Computer Vision Image Anal.*, 41 (1), pp 256-263.

[22] Zauner, C. (2010) "Implementation and Benchmarking of Perceptual Image Hash Functions", University of Applied Sciences, University of Applied Sciences Hagenberg, Austria.

[23] Zhang, K., Zhang, Z., Li, Z. and Qiao, Y. (2016) "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks", *IEEE Signal Processing Letters*, 23(10), pp 1499-1503.

[24] Zauner, C. (2016) "Open Not Suitable/safe For Work (NSFW) model", https://yahooeng.tumblr.com/post/15114868

https://doi.org/10.17979/spudc.9788497498043.648

9421/ open-sourcing-a-deep-learning-solution-for.

https://doi.org/10.17979/spudc.9788497498043.648