Video pornography detection through deep learning techniques and motion information

Mauricio Lisboa Perez
School of Electrical and
Electronic Engineering (EEE)
Nanyang Tech. University (NTU), Singapore
Email: mauricio001@e.ntu.edu.sg

Vanessa Testoni
Samsung Research Institute Brazil
Campinas, Brazil
Email: vanessa.t@samsung.com

Anderson Rocha
Institute of Computing
University of Campinas (UNICAMP)
Campinas, Brazil
Email: anderson@ic.unicamp.br

Abstract—**Recent literature has explored automated pornographic detection — a bold move to replace humans in the tedious task of moderating online content. Unfortunately, on scenes with high skin exposure, such as people sunbathing and wrestling, the state of the art can have many false alarms. This paper is based on the premise that incorporating motion information in the models can alleviate the problem of mapping skin exposure to pornographic content, and advances the bar on automated pornography detection with the use of motion information and deep learning architectures. Deep Learning, especially in the form of Convolutional Neural Networks, has striking results on computer vision, but their potential for pornography detection is yet to be fully explored through the use of motion information. We propose novel ways for combining static (picture) and dynamic (motion) information using optical flow and MPEG motion vectors. We show that both methods provide equivalent accuracies, but that MPEG motion vectors allow a more efficient implementation. The best proposed method yields a classification accuracy of 97.9% — an error reduction of 64.4% wrt. the state of the art — on a dataset of 800 challenging test cases. Finally, we also discuss results on larger and more challenging dataset.**

I. Introduction

Filtering sensitive media (pornographic, violent, gory, etc.) has growing importance, due to the booming consumption of online media by people of all ages; and among sensitive media types, pornography is often the most unwelcome. A range of applications has increased societal interest on the problem, e.g., detecting inappropriate behavior via surveillance cameras; preventing uploading or accessing undesired content for certain demographics (e.g., minors), or environments (e.g., schools, workplace). In addition, law enforcers may use pornography filters as a first sieve when looking for child pornography in the forensic examination of computers, or internet content. The precise definition of pornography is, of course, subjective, but here we will consider "any sexually explicit material with the aim of sexual arousal or fantasy" [1].

A natural approach to pornography detection consists in first trying to detect nudity [2], [3] and then defining appropriate thresholds to further filter the content. Such solutions commonly use human skin features, such as color and texture, and human geometry [4], [5]. Although the motivation for such methods is intuitive, it reveals ultimately naïve. People may show a lot of skin in activities that have nothing to do with sex (e.g., sunbathing, swimming, running, wrestling), leading to a lot of false positives. Conversely, some sexual practices involve very little exposed skin, leading to unacceptable false negatives. Departing from the low-level skin-based methods, in more recent years, several authors have explored other types of solutions for adult content filtering, specially the ones inspired by the bag of words model from text classification [6]–[8]. Clearly, such methods are more robust than skin-based ones, but still suffer from ambiguous cases. Although thus far relatively underestimated for this problem, motion information available in videos would likely help to disambiguate the most difficult cases in pornography classification. Only a few works have exploited spatio-temporal features or motion information in this problem until now.

In spite of the success of deep learning techniques in the computer vision area, their literature on pornography detection is very scarce. Pioneering the trend for pornography detection, Moustafa [9] has explored majority voting classification on a sample of frames classified with off-the-shelf CNNs. However, the authors did not explore the most appropriate network configurations, parameters nor any spatio-temporal or motion information in their solution.

In this work, we design and develop deep learning-based approaches to extract discriminative spatio-temporal characteristics for filtering pornographic content in videos. As far as we know, this is the first time convolutional neural networks (CNNs) — along with motion information — is applied for pornography detection in videos. Although in this work we focus on the pornography modality, the methodology we discuss herein is versatile and its extension to other types of sensitive content is straightforward.

II. Deep Learning and spatio-temporal features

The approaches we designed were mainly inspired on the work of Simonyan and Ziesserman [10], in which the motion information is explicitly provided to the CNN, and each type of information (static and motion) is independently processed by the network. Notwithstanding, we explore the motion information differently and incorporate novel sources of motion information in our work. We also propose new ways for combining static and motion for a more effective decision making.

[0] This work relates to a M.Sc. dissertation.

A. Static Information

In the static pipeline we propose (c.f., Fig. 1), we start with a chosen sampling of the video frames and extract their features with a convolutional network. These features are average pooled to form a single description of the whole video. Finally, we feed a classifier with the video description for the final classification.

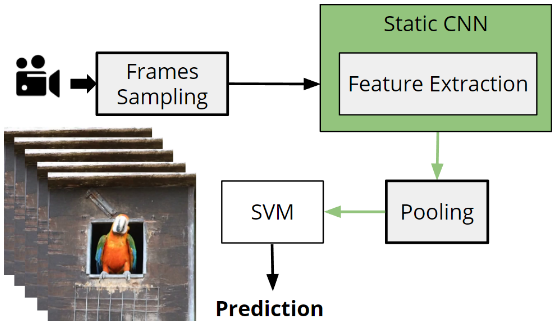


Fig. 1. Pipeline for the static information flow.

For the static CNN, we explored two methodologies. The first one considers a network model trained with natural images obtained with the ImageNet dataset [11] while the second model is custom-tailored (i.e., fine-tuned and properly adapted) to our problem, starting with the weights obtained with the ImageNet samples during a pre-training step rather than using random weights for initializing the network.

B. Motion Information

Initially, we analyze the motion information independently from the static information. The pipeline (Fig. 2) for this type of information is somewhat similar to the static pipeline, with differences in the input and output of the convolutional network.

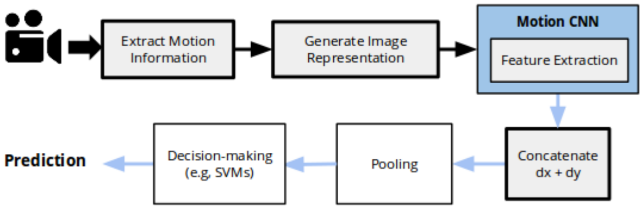
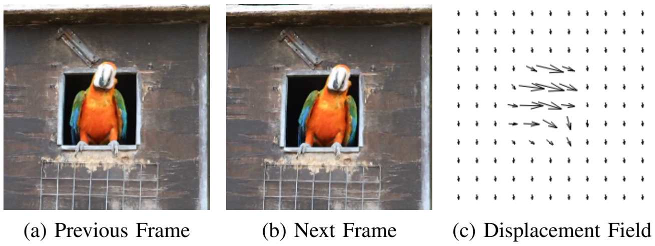

Extract Motion Information
Generate Image Representation
Motion CNN
Feature Extraction
Concatenate dx + dy
Pooling
Decision-making (e.g, SVMs)
Prediction

Fig. 2. Pipeline for the motion information flow.

We evaluate two sources for the motion information: Optical Flow displacement fields and MPEG Motion Vectors. The motion sources follow the pipeline independently, therefore there is a specific motion CNN model and classifier for each. Each source requires an unique form for extracting the motion information.

Our first explored source of motion information is the optical flow displacement fields technique, computed using Brox et al.'s method [12]. Each position of interest provides us with the gradient's magnitude and the direction of the motion. Fig. 3 depicts an example of the output. For a more useful representation, we decompose this information in its horizontal (dx) and vertical (dy) components, generating two motion maps with the magnitude values for each component separately. Fig. 5(b) depicts an example of the optical flow representation, calculated from the generated motion maps.



(a) Previous Frame (b) Next Frame (c) Displacement Field

Fig. 3. Sequential raw frames (left and middle) and the respective Optical-Flow Displacement Field (right) computed from them.

Another explored source of the motion information is the motion vector data encoded within the MPEG codec. In each vector for a particular frame, it is encoded the position from a given macroblock of pixels in the current frame and its position at the reference frame. Fig. 4 shows an example. In this work, we propose a novel representation for the motion information contained in these vectors. We measure how much the block from each motion vector has moved by computing the distance, in pixel coordinates, from the reference position to the current position in each direction, horizontal and vertical, separately. These distances are analogous to the magnitude of the movement at the region contained in that macroblock, and generate two motion maps, one for each direction, similarly to the optical flow motion extraction. Fig. 5(c) illustrates an example of the generated image representation.

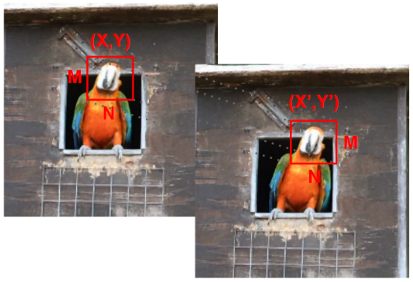
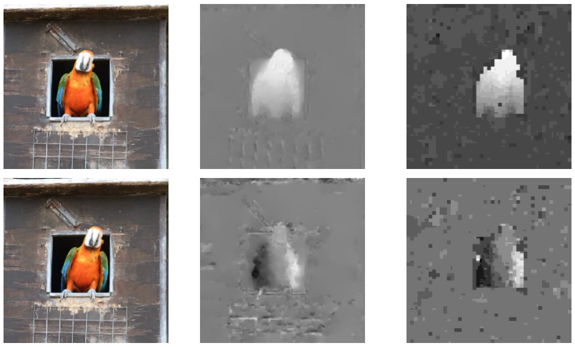


Fig. 4. Example of a macroblock and its respective Motion Vector

After the feature extraction, the descriptions of the components (dx and dy) from the same motion are concatenated to form a single feature vector. The rest of the pipeline is then similar to the static one: the combined descriptions are pooled and fed to a classifier for final decision making.

C. Fusion

The static and motion information can lead to more effective results if their collected evidence (video telltales) are complementary in some sense. Therefore in this section, we explore different forms of combining them (Fig. 6):



(a) Raw frames (b) Optical Flow (c) Motion Vectors

Fig. 5. Sequential raw frames (a) and motion image representations from optical flows (b) and MPEG motion vectors (c). The horizontal (dx) component is on top, and the vertical (dy) one is on the bottom. The regions with more movement in raw frames (e.g., macaw's head and body) appear highlighted (dark or light) in the motion representations, while regions without movement correspond to the neutral middle gray.

1) **Early Fusion:** The static and the motion information are combined at the very beginning of the pipeline, being processed together by a special CNN. This way, the features benefit from both the static and the motion information. This fusion has two variations: 5-channel input, comprising a stack of the three color channels of the frame, along with its respective motion representations, dx and dy; 3-channel input, using the raw frame information in gray scale, instead of color.
2) **Mid-level Fusion:** We concatenate the features extracted from each type of information (static or motion-based), and from each independent CNN, into a single feature vector before feeding a classifier.
3) **Late Fusion:** Each information is processed by a separate decision-making approach (e.g., SVM classifier), generating independent classification scores that can then be combined later on on a single score for the final classification.

D. Architecture Specifications

The convolutional neural network architecture we adopt for the experiments was proposed in [13], and is referred to as GoogLeNet, winner of ImageNet 2014 Challenge [11], achieving a striking 6.67 top-5 error rate in the object classification competition. This architecture was employed for all types of data: Static (raw frames), Motion (optical flow and motion vectors) and Static-Motion (early fusion). For feature extraction, we pick the output from the last layer — fully-connected (FC) — before the final classification. All the network weights, except within Early Fusion, are fine-tuned to the problem of interest herein via backpropagation, initializing the weights with the values learned on the ImageNet 1.2 million images.

III. Results

To evaluate the proposed approaches, we adopted two datasets in our experiments: Pornography-800 [8] and Pornography-2k [14]. Pornography-2k is a contribution of this work, it comprise an extension of Pornography-800 (more complete and challenging). Therefore, we opted to report all the experiments with the proposed methods on the Pornography-2k, along with Third-party solutions and the methods we choose as baselines. Finally, we evaluate our best proposed approaches on Pornography-800, for direct comparisons with existing work in the literature.

The reported accuracy is the mean value from accuracy obtained on different splits of the data. In the case of Pornography-2k it was applied a 5×2-fold cross-validation protocol, what resumes to different train and tests folds with 1000 videos each and a balanced number of classes. The protocol followed for Pornography-800 was a more simple 5-fold cross-validation (640 videos for training and 160 for testing on each fold). For training and fine-tuning on each of dataset, the training fold was re-partitioned into actual training and validation, with a proportion of 85%/15% videos in each part.

A. Proposed Approaches

In Table I, we show the obtained video classification accuracy for each approach we have proposed, considering the static and motion information as well as the fusion of different methods. In these experiments, we adopted the Pornography-2k dataset.

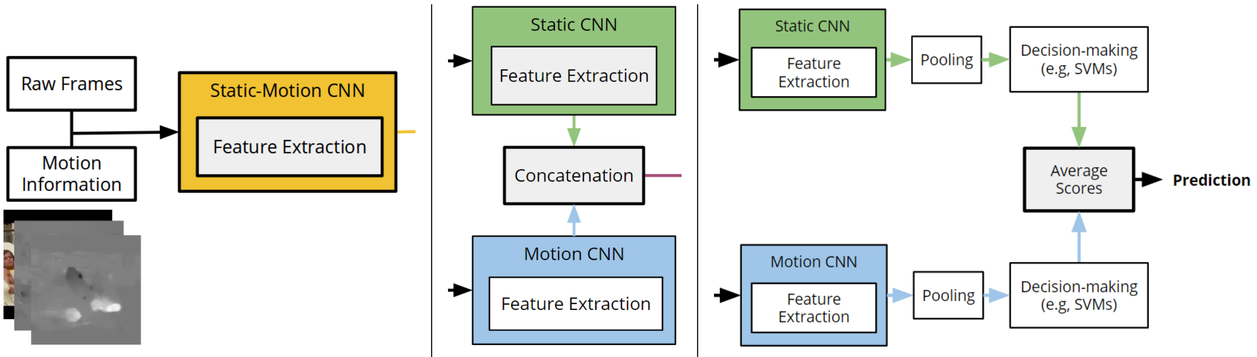
TABLE I

Video classification *accuracy*, averaged over the 5×2 experimental folds, for the proposed approaches on the Pornography-2k dataset. The methods are subdivided in static, motion and fusion modalities. Fusion is performed with the fine-tuned model for static information, and with both motion sources, optical flow (OF) and MPEG motion vectors (MV), except for the early fusion, which, due to its inferior performance with OF, is not employed with MV.

	Proposed Approach		ACC (%)
Static	ImageNet		94.6
	Fine-tuned*		**96.0**
Motion	Optical Flow		**94.4**
	MPEG Motion Vectors		91.0
Fusion	Early Fusion – Gray	OF	95.5
	Early Fusion – Color		90.5
	Mid-level Fusion		96.3
	Late Fusion*		**96.4**
	Mid-level Fusion	MV	96.4
	Late Fusion		96.4

ACC: accuracy — *Fine-tuned and Late Fusion are statistically different (p-values: ACC ≈ 0.03; $F_2 \approx 0.01$) — All standard deviations are smaller than 0.02.

In the static stream, we show that the model relying on the GoogLeNet architecture trained with ImageNet data yields an impressive performance of 94.6% ACC and 95.1% F_2. These results are further improved upon by fine-tuning the network weights with the pornographic data, thus specializing the network to the problem of interest, reaching 96.0% ACC and 96.1% F_2, a 1.5 percentage point improvement in ACC (26% error reduction) and 1 percentage point in F_2.



(a) Early (b) Mid-level (c) Late

Fig. 6. Main parts from each of the Fusion methods proposed.

When considering the motion information, optical flow (OF) by itself yielded a performance close to the static model. Meanwhile, the MPEG motion vectors (MV) led to a lower performance, of 91.0% ACC and 92.0% F_2. This difference in performance between these two sources of motion information may be explained by the fact that the MV represents the motion of a macroblock of pixels, which is a much lesser fine-grained description form than OF, which takes into account the motion information for each pixel.

Despite the lower performance of the motion information alone, when we combine it with the static information from the fine-tuned network (pornography-specialized network), by mid-level fusion and late fusion, we improve the ACC and F_2 results. Both early fusion variations, Gray and Color, yielded a lower performance than using the fine-tuned static information by itself. Perhaps it is better to specialize the network to a single type of information, leaving the fusion to a higher level. Another reason might be related to the architecture considered in this work, GoogLeNet. It may not be appropriate for processing five channels or combining static and motion right at the lowest level (e.g., raw data) of the network, demanding some customization such as increasing the number of filters or processing each information independently at the first layers.

We believe that the better performance from the gray variation over color, comes from the fact that we could fine-tune its model using the ImageNet model and that the 3-channel input data is more appropriate for the GoogLeNet architecture. However, we expect that if these issues were overcome (e.g., by training an appropriate architecture with a large collection of samples), the full potential from using all color channels could be reached, outperforming the gray-only variation of this fusion, and perhaps the other fusion approaches, mid-level and late.

Given the low performance of early fusion, and its costly requirements for training, we have opted for not fusing MPEG motion vectors this way.

Mid-level fusion and late fusion, on the other hand, apparently could better combine static and motion information, surpassing the performance of the fine-tuned network alone. Surprisingly, this happened even while combining with MV, showing that, although it had a worse performance when used alone, its complementarity to the static information is still advantageous. In addition, another advantage of using the MVs is that they are readily available during decoding of the video. Still, even that by a small margin, late fusion with OF obtained the best combination of results for ACC.

In fact, our architecture was able to properly learn effective features from the motion data, as our results with middle- and late-fusion approaches showed, which take into account the information provided by the *Static Raw Frames* and *Optical Flows* simultaneously. However, it is possible that using an innate motion-based network could equally produce good results; however such network could be more complex (with more weights) than the one we have extended upon.

B. Comparison using the Pornography-2k Dataset

For a better evaluation of the proposed approaches that obtained the best results in each modality, we compare them with the current state-of-the-art spatio-temporal video description and third-party solutions. Table II shows the respective video classification accuracy of the considered methods. Note that the best proposed methods outperform most of the existing solutions.

TABLE II
RESULTS ON THE PORNOGRAPHY-2K DATASET.

	Solution	ACC (%)
Third-party	NuDetective [15]	72.6
	PornSeer Pro [16]	79.1
BoVW-based	Dense Trajectories [17]	95.8
	TRoF [14]	95.6
Best Proposed Approaches	Static – Fine-tuned	96.0
	Motion – Optical Flow	94.4
	Late Fusion (Optical Flow)	**96.4**
	Late Fusion (Motion Vector)	**96.4**

The third-party solutions, which heavily depend on skin detection and do not take advantage of the space-time information, have shown a poor performance. PornSeer Pro [16] obtained the best ACC among them, with 79.1%, far below the performance using the solutions in the literature and our proposed approaches.

The proposed methods also outperform the Dense Trajectories method [17]. For instance, the spatio-temporal approach, Late Fusion (OF), outperforms Dense Trajectories by a margin of 0.5 percentage point in ACC (14.3% error reduction).

Also, we can assert that motion feature plays an important role in pornography video detection when comparing the motion-based approaches (Dense Trajectories and proposed approaches) with the third-party solutions. The motion-based approaches remarkably outperform the third-party solutions.

C. Comparison using the Pornography-800 Dataset

In Table III, we compare our best proposed approaches with the literature, using the Pornography-800 dataset.

TABLE III
RESULTS ON THE PORNOGRAPHY-800 DATASET.

	Solution	ACC (%)
BoVW-based	Avila et al. [18]	87.1 ± 2.0
	Valle et al. [19]	91.9 ± NA
	Souza et al. [20]	91.0 ± NA
	Avila et al. [8]	89.5 ± 1.0
	Caetano et al. [21]	90.9 ± 1.0
	Caetano et al. [22]	92.4 ± 2.0
	TRoF [14]	95.0 ± 1.3
CNN	Moustafa [9]	94.1 ± 2.0
Best Proposed Approaches	Static – Fine-tuned	97.0 ± 2.0
	Motion – Optical Flow	95.8 ± 2.0
	Mid-level Fusion (Optical Flow)	**97.9 ± 0.7**
	Late Fusion (Optical Flow)	**97.9 ± 1.5**

The proposed approaches significantly outperform the existing BoVW-based methods [8], [14], [18]–[21], by 3–11 percentage points. The proposed methods also outperform, by almost four percentage points, the results reported in Moustafa [9], which also use Deep Learning. In this case, the error reduction was over 64%. Even though we could not apply Wilcoxon's test, given the large perceptual difference in accuracy between the related works and our best approaches, with smaller standard deviation in some cases, we believe that the results would probably be statistically significant.

Although Moustafa [9] employs the same architecture we use in this work, GoogLeNet, there are critical differences, thus leading to the important difference in performance, we report herein: first of all, he only fine-tuned the network last layer, while in our work we fine-tuned all layers, creating a network model specialized to the problem of interest; second, the network output in that work, for each frame, was used in a majority voting scheme for classifying the video, while, in turn, we have opted for using the network as a feature extractor, pooling the frame descriptions, then feeding them to an classifier for the video classification; finally, that work only considered static information, meanwhile our methods rely upon static and motion information, as well as on effective methods for combining them.

IV. Conclusions and Future Research Directions

The evaluation of our techniques, shows that the association of Deep Learning with the combined use of static and motion information, considerably improves pornography detection. Not only over current scientific state of the art [8], [9], [19], [21], but also over off-the-shelf software solutions [15], [16]. Our solution also proves to be superior to general-purpose action recognition features [17], when applied to pornography detection.

The Deep Learning solution using only static information is already competitive with state-of-the-art action recognition features, Dense Trajectories [17], reaching an error rate of 4%, which is low for such a subjective problem as pornography. For further reducing the error rate, we believe the focus should be on the motion information: by adjusting the CNN, adapting the architecture, boosting the model with more training samples, or improving static-dynamic information fusion.

Besides improving whole-video classification, we are interested in applying our techniques to the harder task of locating in time the pornographic content within the video. To reach that goal, the Pornography-2k video dataset – which is already a great contribution from this work – was annotated at frame level, becoming the first dataset for this problem with this level of annotation. The main motivation for that harder task is filtering pornography in real time, an important goal for video streaming, camera-surveillance systems, or surveillance of video chats for certain publics.

In addition to adapting our current methods for the localization problem (e.g., [23]), another aspect worth exploring is to integrate them to the so called Long Short Term Memory (LSTM) networks. LSTMs are a model of Recurrent Neural Network (RNN) that captures the sequential information of the input data, a highly desirable feature for classification of videos. The LSTM architecture could be used to process the CNN extracted features, using the proposed methods in this work, from a fixed number of frames, improving the real-time classification.

A current on going extension of this work, is related to further specializing the methods proposed here for detecting child pornography in images. In a collaboration with the Brazilian Federal Police, a paper has been submitted and it is on the second round of reviews (refer to Section V). A fact that further reinforce the importance of this work and its contributions to real-world applications.

V. Publications, Awards and Relevant Production

1) Perez, M., Avila, S., Moreira, D., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2017). *Video pornography detection through deep learning techniques and motion information*. Elsevier Neurocomputing, 230:279–293. (**Impact Factor:** 2.392 Journal Citation Reports)
2) Avila, S., Vitorino, P., Perez, M., and Rocha, A. (2017). *Leveraging Deep Neural Networks to Fight Child Pornography in the Age of Social Media*. Submitted to Elsevier

Journal of Visual Communication and Image Representation. (**Impact Factor:** 2.164 Journal Citation Reports)

3) Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2015). *RECOD at MediaEval 2015: Affective Impact of Movies*. Proc. of the 2015 MediaEval Workshop.
4) **2nd place at Violent Scenes Detection: Generalization Task.** Moreira, D., Avila, S., Perez, M., Moraes, D., Cota, I., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2014). *RECOD at MediaEval 2014: Violent Scenes Detection Task*. Working Notes Proceedings of the MediaEval 2014 Workshop.
5) **Patent.** Avila, S., Moreira, D., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., and Rocha, A. (2016). *Multimodal and Real-time Method for Sensitive Media Content Filtering*. United States. Register number: 15198626, Registration institution: United States Patent and Trademark Office. Deposit: 30/06/2016

VI. Impact Outside Academia

The results obtained on this dissertation, specially the high accuracy achieved, have been the topic of news at different venues. Including articles in nationwide news portals and TV news. The following list contains some examples:

1) Fapesp: ”Novo método identifica cerca de 97% da pornografia em telas de celulares e computador” – http://agencia.fapesp.br/novo_metodo_identifica_cerca_de_97_da_pornografia_em_telas_de_celulares_e_computador/25023/
2) UOL: ”Novo sistema consegue barrar quase 100% da pornografia de sua TV e celular” – https://tecnologia.uol.com.br/noticias/redacao/2017/04/04/novo-sistema-consegue-barrar-quase-100-da-pornografia-de-sua-tv-e-celular.htm
3) Jornal da Cultura 1^a Edição — 03/04/2017 – https://youtu.be/eoUe5MftAgI?t=14m50s
4) Band Campinas: ”Inteligência artificial é capaz de filtrar até 90% do conteúdo pornográfico em redes sociais” – https://www.facebook.com/bandcampinas/videos/1245325798914831/

Acknowledgment

Part of the results presented in this paper were obtained through the project “Sensitive Media Analysis”, sponsored by Samsung Eletrônica da Amazônia Ltda., in the framework of law No. 8,248/91. We also thank the financial support from CNPq, FAPESP (Grant #2015/19222-9) through the DéjàVu project, and CAPES, through the DeepEyes project. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Tesla K40 GPU used for this research.

References

[1] M. Short, L. Black, A. Smith, C. Wetterneck, and D. Wells, “A review of internet pornography use research: Methodology and content from the past 10 years,” *Cyberpsychology, Behavior, and Social Networking*, vol. 15, no. 1, pp. 13–23, 2012.

[2] M. Fleck, D. Forsyth, and C. Bregler, “Finding naked people,” in *European Conference on Computer Vision (ECCV)*, vol. 1065, 1996, pp. 593–602.

[3] H. Zheng, M. Daoudi, and B. Jedynak, “Blocking Adult Images Based on Statistical Skin Detection,” *Electronic Letters on Computer Vision and Image Analysis (ELCVIA)*, pp. 1–14, 2004.

[4] M. Jones and J. Rehg, “Statistical color models with application to skin detection,” *International Journal of Computer Vision (IJCV)*, vol. 46, no. 1, pp. 81–96, 2002.

[5] H. Bouirouga, S. El Fkihi, A. Jilbab, and D. Aboutajdine, “Skin detection in pornographic videos using threshold technique,” *Journal of Theoretical and Applied Information Technology*, vol. 35, no. 1, pp. 7–19, 2012.

[6] T. Deselaers, L. Pimenidis, and H. Ney, “Bag-of-visual-words models for adult image classification and filtering,” in *International Conference on Pattern Recognition (ICPR)*, 2008, pp. 1–4.

[7] A. Ulges and A. Stahl, “Automatic detection of child pornography using color visual words,” in *IEEE International Conference on Multimedia and Expo (ICME)*, 2011, pp. 1–6.

[8] S. Avila, N. Thome, M. Cord, E. Valle, and A. Araújo, “Pooling in image representation: The visual codeword point of view,” *Elsevier Computer Vision and Image Understanding (CVIU)*, vol. 117, no. 5, pp. 453–465, 2013.

[9] M. Moustafa, “Applying deep learning to classify pornographic images and videos,” in *7th Pacific-Rim Symposium on Image and Video Technology (PSIVT)*, 2015.

[10] K. Simonyan and A. Zisserman, “Two-stream convolutional networks for action recognition in videos,” in *Advances in Neural Information Processing Systems (NIPS)*, 2014, pp. 568–576.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, “ImageNet Large Scale Visual Recognition Challenge,” *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[12] T. Brox, A. Bruhn, N. Papenberg, and J. Weickert, “High Accuracy Optical Flow Estimation Based on a Theory for Warping,” in *European Conference on Computer Vision (ECCV)*, 2004, pp. 25–36.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 1–9.

[14] D. Moreira, S. Avila, M. Perez, D. Moraes, V. Testoni, E. Valle, S. Goldenstein, and A. Rocha, “Pornography classification: The hidden clues in video space-time,” *Elsevier Forensic Science International (FSI)*, vol. 268, pp. 46–61, 2016.

[15] M. Polastro and P. Eleuterio, “Nudetective: A forensic tool to help combat child pornography through automatic nudity detection,” in *IEEE Database and Expert Systems Applications (DEXA)*, 2010, pp. 349–353.

[16] “PornSeer Pro,” http://www.yangsky.com/products/dshowseer/porndetection/PornSeePro.

[17] H. Wang and C. Schmid, “Action Recognition with Improved Trajectories,” in *IEEE International Conference on Computer Vision (ICCV)*, 2013, pp. 3551–3558.

[18] S. Avila, N. Thome, M. Cord, E. Valle, and A. Araújo, “BOSSA: Extended bow formalism for image classification,” in *IEEE International Conference on Image Processing (ICIP)*, 2011, pp. 2909–2912.

[19] E. Valle, S. Avila, A. da Luz Jr., F. Souza, M. Coelho, and A. Araújo, “Content-based filtering for video sharing social networks,” in *Brazilian Symposium on Information and Computer System Security (SBSeg)*, 2012, pp. 625–638.

[20] F. Souza, E. Valle, G. Cámara-Chávez, and A. Araújo, “An evaluation on color invariant based local spatiotemporal features for action recognition,” in *Conference on Graphics, Patterns and Images (SIBGRAPI)*, 2012, pp. 31–36.

[21] C. Caetano, S. Avila, S. Guimarães, and A. Araújo, “Pornography detection using bossanova video descriptor,” in *European Signal Processing Conference (EUSIPCO)*, 2014, pp. 1681–1685.

[22] C. Caetano, S. Avila, W. R. Schwartz, S. J. F. Guimarães, and A. Araújo, “A mid-level video representation based on binary descriptors: A case study for pornography detection,” *Elsevier Neurocomputing*, vol. 213, pp. 102–114, 2016.

[23] X. Chang, Y. Yang, E. P. Xing, and Y.-l. Yu, “Complex event detection using semantic saliency and nearly-isotonic SVM,” in *ACM International Conference on Machine Learning (ICML)*, 2015, pp. 1348–1357.