Deep Learning Features Fusion with Classical Image Features for Image Access

Rehan Ullah Khan
Information Technology Department
Qassim University,
Al-Qassim, KSA

***Abstract*—Depending on the society, the access to the adult content can create social problems. This paper thus proposes a fusion approach for image based adult content filtering. The proposed approach merges the Deep Learning (DL) architecture and classical hand crafted feature extraction approaches. From the DL, we fuse the rich feature extraction capabilities of the Convolutional Neural Networks (CNNs) with the Correlograms features. We optimize the classification by integrating and modifying the Correlograms into skin Correlograms. The results show an increased performance by combining the DL learnt features with the classical hand crafted features. From an evaluation, the proposed approach achieves an Accuracy of 0.93. This work thus motivates the usage of classical hand crafted features to be exploited in the DL architectures for segmentation and detection scenarios.**

Keywords—Deep learning; content based filtering; content analysis; machine learning; support vector machines

I. Introduction

The availability of the explicit adult content on the Internet is increasing rapidly. One of the main problems is that the accessibility of these media resources is becoming easy due to number of available solutions and the internet availability. This opens risks in terms of many social factors, and for many societies, accessibility of such content is crime. The most feared element, however, nowadays is the availability of these media resources to the children. This article thus targets such concerns in the form proposing a solution to media filtering using the fusion of DL and classical image features.

Convolutional Neural Networks (CNNs) have demonstrated its usefulness in the field of Computer Vision (CV) tasks of object detection and classification [1]. A shift has been observed from the hand crafted feature extraction to automated model learning [2], [3] from images. In this article, we investigate media filtering using the DL and classical feature extraction. Classical approaches require hand crafted feature extraction from given set of images. From the proposed approach of media filtering, we find that though DL provides good overall performance, a well-crafted feature set can augment DL approaches and increase detection performance.

We propose an approach for content filtering that flags the image as adult nature or safe image. Our approach exploit two things: Firstly, it uses the DL architecture to learn the rich set of features; secondly, it merges the innovative Deep Learning (DL) feature set and the classical feature extraction methods. We fuse the CNN based features with the Correlograms features. The classical feature set represents the skin color based Correlograms. We optimize the classification by integrating and modifying the correlation grams into skin Correlograms. The results show that it is useful to combine deep learnt features with the classical hand crafted features and achieve good overall performance. We get an increase of 5% detection performance by combining the DL features with the manual feature set. With this setup, we expect an already rich set of features to be exploited in the DL architectures for segmentation and detections.

The efforts to detect and possibly block explicit adult content are not new and there is interesting work regarding adult content detection in the state-of-the-art [4]-[7]. The work in [4] fuses the two DL approaches of AlexNet [3] and GoogLeNet [8] and reports that performance can be increased by fusing the two networks for adult content classification. The work in [6] targets skin locus detection for content filtering using the 24 colors transformations in widely available images and videos. The article [9] presents an evidence combination that includes video sequences, key-shots, and key-frames. The work in [10] is based on the adaptive sampled based analysis showing the usefulness of proposed method in the detection of minor pornographic sequences with 87% detection rate. The approach in [6] employs color based skin filtering and content based filtering based on the skin detection. Lopes et al. [5] uses the text retrieval approach for content filtering and is analyzed and evaluated using datasets, achieving a 93.2% detection rate. The authors in [11] demonstrate a framework to analyze the websites. The framework produces an augmented classification that is independent of the access scenarios. In [12], the authors combine key-frame based approaches with a MPEG-4 statistical analysis of the flow vectors. The work of [13] develops filtering for a Web-based P2P. Authors in [14] use two visual features for media access and filtering. First is the single frame, and the other visual based feature is the decision variable of multiple frames with the Discriminant analysis for optimization. Image filtering is actually similar to content based retrieval. The articles [15]-[18] discuss content image retrieval in detail. The work in [19] uses the Hue-SIFT for nude and explicit content detection achieving better results compared to the SIFT. The approach in [20] takes advantage of the motion flow vectors combining them with the audio features for filtering. Lee et al. [21] propose multi-modal approach comprising of three phases; hashing, real-time detection, and finally group of frames decision. The proposed approach of [22] employs optical flow for filtering and detection, achieving an acceptable 80% detection performance. Authors in [23] have similar approach of the motion estimation to the approach of [22] for media filtering. In [24], the authors present a fusion of audio features and video features based on the SVM classifier. The work in [25] uses spatial features and time-based features for content filtering with an accuracy of 94.4%.

II. Convolution Neural Networks (CNNs)

CNNs have proved to be a very useful and innovative tool of DL to learn image feature set and inherent relationship in low level features to higher level objects in images. The generic architecture of CNN contains interconnected layers and consists of repeated convolutional blocks, Rectified Linear Units (ReLU) and Pooling layers [3]. Convolutional layers perform convolution of the input with set of filters. The filters are then learnt during training. The non-linear behavior in data is modelled by the ReLU layer [3]. The pooling layers samples the input and consolidate image classes.

III. Proposed Deep Architecture

Fig. 1 shows the proposed fusion DL architecture. If fusion is not integrated, then the architecture is similar to the one proposed in [3]. However, we augment this architecture with fusion. The structure of the proposed fusion DL for feature extraction and classification of images is composed of different layers. Our DL architecture contains five Convolution layers. The Convolution layers are followed by three fully connected layers. Each layer uses kernel to filter its two dimension inputs. The coefficients for the kernel are calculated incrementally from the training process. The dot product operation is performed by the fully connected layers between the input and weights vectors. In this connected setup, each neuron is connected to all outputs. For learning, and expediting the learning process, each layer uses the ReLU. The Softmax layer gets its input from the last fully connected layer and thus produces probabilistic distribution for a bi-class problem of "adult" and "non-adult" image. Fig. 1 shows the proposed architecture to enhance and fuse the features from the color based Correlograms. SVM is used for classification after feature from DL and Correlograms are calculated.

IV. Experimental Setup and Results

L-norm distance between the two pixels and corresponding histogram is calculated as:

$$\text{L-norm-hist}= \mathbf{n}^2\boldsymbol{\pi}\mathbf{P}_{(p \in \tilde{\imath})}[\mathbf{p} \in \tilde{\imath}_{ci}] \qquad - \tag{1}$$

$\tilde{\imath}$ is an image with color quantized as $c_1 \ldots c_m$. The π represents the product operation. We define the skin Correlograms as follows:

$$\text{Cor}_{\_\text{Skin}}=\mathbf{P}(p_1 \in \tilde{\imath}p_{c}, p_2 \in \tilde{\imath}p)\ [p_2 \in \tilde{\imath}p_{c}\,|\ |\mathbf{p_r}(p_2>(p_1 \in \text{Æ}(\tilde{\imath}))), |p_1-p_2|=k] \qquad - \tag{2}$$

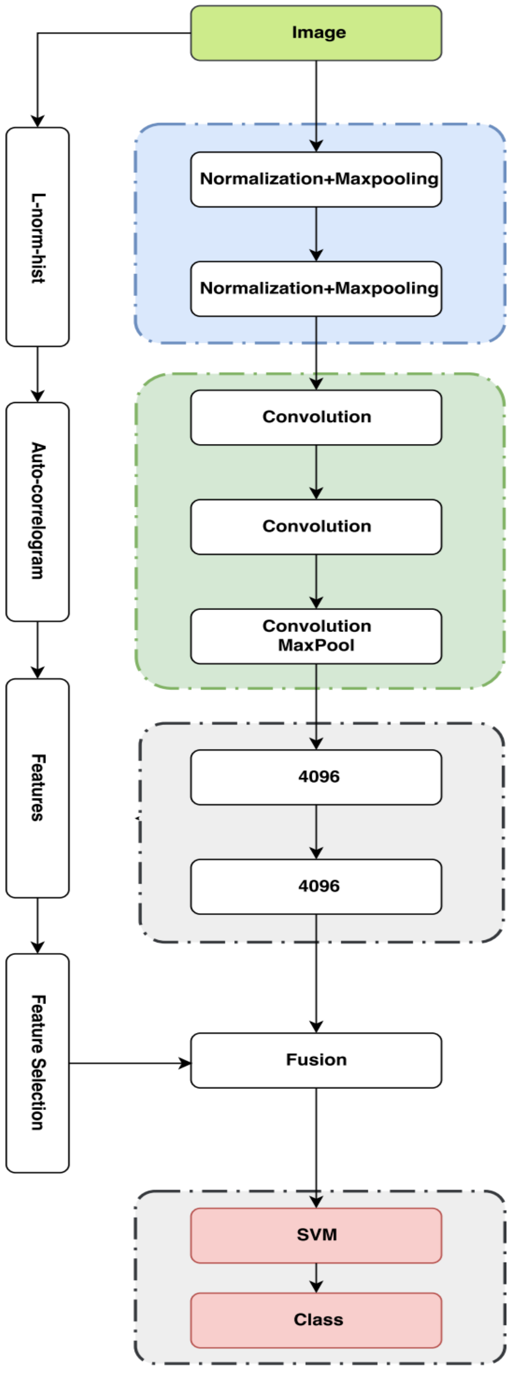

Fig. 1. Proposed Alex Correlogram Network (ACNet) adult image classification. The convolution layers (first five) are followed by two fully connected layers. The output of these layers is represented as DL feature set and merged with the correlograms features. The SVM learns and decides the nature of the test images.

Where $\tilde{\imath}p$ is an image after classifier operation represented by Æ. In (2), the $\tilde{\imath}p$ is a probabilistic output of the image representation from the classifier. This setup of feature extraction based on the Correlograms is integrated in the proposed DL architecture in Fig. 1. With the integration of the Correlograms, we represent the proposed DL network of Fig. 1 as the Alex Correlogram Network (ACNet). The convolution layers (first five) are followed by two fully connected layers. The output of these two layers is represented as DL feature set and merged with the Correlograms features. The SVM learns and decides the nature of the test images.

The layered architecture we use contains more than 55 million parameters trained for more than 10000 classes. Rather than training from the scratch, we use the weights of the ConvNet [14] obtained from 1.2 million images. Our input image is fed into the first layer and feature vector is obtained from the seventh layer. The weights for the classification layer are modified based on the labelled data. In the last layer, we are using the SVM. For a test image, the SVM outputs a binary decision “adult” or “non-adult”.

For an evaluation of the proposed architecture, we use the key-frames from the NDPI videos. Further details are available in [26]. Fig. 2 shows some samples.

Fig. 2. Sample images from NDPI [26].

We use the accuracy as an evaluation parameter as it is mostly used in the state of the art for similar problems and applications and is favorable for this evaluation as well. For comparison, we calculate the similar architectures of AlexNet (ANet) of [3], AGNets of [4]. For performance evaluation and comparison, we train 10 ANets, 10 AGNets and 10 ACNets using four of the five folds of NDPI videos. We set the fifth fold for testing. In the evaluation, our objectives are firstly to check the DL architectures with and without external feature fusion. Secondly, we compare the proposed approach to two networks fusion approach of [4]. The setup in [4] combines the AlexaNet and GoogleNet and represent it as the AGNet for performance enhancements. Interestingly, in the evaluation using Accuracy, our proposed approach of external feature fusion in DL architecture outperforms the AGNet. Fig. 3 shows the comparison of the three approaches using the Accuracy. ANet alone achieves an Accuracy of 0.88; the AGNet achieves an Accuracy of 0.90 and the proposed ACNet achieved an Accuracy of 0.93. The proposed approach thus gets 5% increase in detection performance compared to the ANet. The proposed approach achieves 3% increase compared to the AGNet.

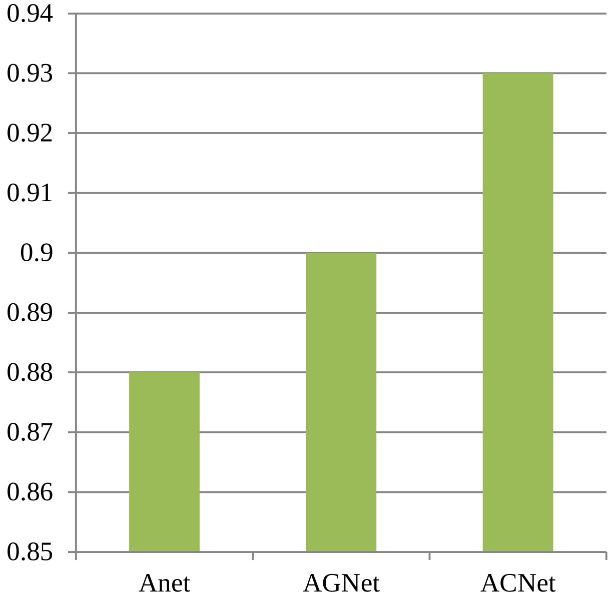

Fig. 3. Comparison based on accuracy between the proposed ACNet, ANet [3], AGNet [4].

The increased performance of our ACNet shows that it is possible to combine Deep learnt features with classical hand crafted features and achieve good overall performance. With this research direction, we expect an already rich set of features extraction paradigm to be exploited and used in conjunction with the DL architectures.

V. Conclusion

We proposed the fusion of the DL feature set with the classical hand crafted feature extraction paradigm. From the DL, we use the CNN feature set with the Correlograms feature set. We achieve 5% enhancement over CNN features alone (ANet) and 3% enhancement with the approach of (AGNet) which combines the fusion of two DL architectures. With this research direction, we expect an already rich set of feature extraction paradigm from last many years of research to be exploited and used in conjunction with the DL architectures achieving combined good overall performance. We hope that the new DL approaches will flourish that will combine multiple cues in supervised and unsupervised methods. Also, the execution of DL will be favored to the available hardware resources. Future work is directed towards semantic integration of skin learning in DL feature generation steps.

Acknowledgments

The work in this article is funded by the Dean of Scientific Research (SRD), Project number: 1335-coc-2016-1-12-S at the Qassim University, Kingdom of Saudi Arabia. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

REFERENCES

[1] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, “From Group to Individual Labels Using Deep Features,” Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. - KDD ’15, pp. 597–606, 2015.

[2] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning Hierarchical Features for Scene Labeling,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1. Curran Associates Inc., pp. 1097–1105, 2012.

[4] M. Moustafa, “Applying deep learning to classify pornographic images and videos,” Nov. 2015.

[5] A. P. B. Lopes, S. E. F. de Avila, A. N. A. Peixoto, R. S. Oliveira, M. de M. Coelho, and A. de A. Araújo, “Nude Detection in Video Using Bag-of-Visual-Features,” in 2009 XXII Brazilian Symposium on Computer Graphics and Image Processing, 2009, pp. 224–231.

[6] A. Abadpour and S. Kasaei, “Pixel-Based Skin Detection for Pornography Filtering,” Iran. J. Electr. Electron. Eng., vol. 1, no. 3, pp. 21–41, 2005.

[7] R. Ullah and A. Alkhalifah, “Media Content Access: Image-based Filtering,” Int. J. Adv. Comput. Sci. Appl., vol. 9, no. 3, 2018.

[8] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going Deeper with Convolutions,” Sep. 2014.

[9] E. Valle, S. Avila, F. Souza, M. Coelho, and A. de A. Araujo, “Content-Based Filtering for Video Sharing Social Networks,” in XII Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais—SBSeg 2012, 2011, p. 28.

[10] P. Monteiro, S. Eleuterio, M. De, and C. Polastro, “An adaptive sampling strategy for automatic detection of child pornographic videos.”

[11] N. Agarwal, H. Liu, and J. Zhang, “Blocking objectionable web content by leveraging multiple information sources,” ACM SIGKDD Explor. Newsl., vol. 8, no. 1, pp. 17–26, Jun. 2006.

[12] C. Jansohn, A. Ulges, and T. M. Breuel, “Detecting pornographic video content by combining image features with motion information,” in Proceedings of the seventeen ACM international conference on Multimedia - MM ’09, 2009, p. 601.

[13] J.-H. Wang, H.-C. Chang, M.-J. Lee, and Y.-M. Shaw, “Classifying Peer-to-Peer File Transfers for Objectionable Content Filtering Using a Web-based Approach.”

[14] Hogyun Lee, Seungmin Lee, and Taekyong Nam, “Implementation of high performance objectionable video classification system,” in 2006 8th International Conference Advanced Communication Technology, 2006, p. 4 pp.-pp.962.

[15] D. Liu, X.-S. Hua, M. Wang, and H. Zhang, “Boost search relevance for tag-based social image retrieval,” in 2009 IEEE International Conference on Multimedia and Expo, 2009, pp. 1636–1639.

[16] J. A. Da, S. Júnior, R. E. Marçal, and M. A. Batista, “Image Retrieval: Importance and Applications.”

[17] S. Badghaiya and A. Bharve, “Image Classification using Tag and Segmentation based Retrieval,” Int. J. Comput. Appl., vol. 103, no. 15, pp. 20–23, Oct. 2014.

[18] A. N. Bhute and B. B. Meshram, “Text Based Approach For Indexing And Retrieval Of Image And Video: A Review,” Apr. 2014.

[19] A. P. B. Lopes, A. P. B. Lopes, R. E. F. De Avila, A. N. A. Peixoto, R. S. Oliveira, and A. De A. Araújo, “A Bag-of-Features Approach based on Hue-Sift Descriptor for Nude Detection.”

[20] N. Rea, G. Lacey, R. Dahyot, and C. Lambe, “Multimodal periodicity analysis for illicit content detection in videos,” in 3rd European Conference on Visual Media Production (CVMP 2006). Part of the 2nd Multimedia Conference 2006, 2006, pp. 106–114.

[21] S. Lee, W. Shim, and S. Kim, “Hierarchical system for objectionable video detection,” IEEE Trans. Consum. Electron., vol. 55, no. 2, pp. 677–684, May 2009.

[22] Y. S. L Li, “Objectionable videos detection algorithm based on optical flow,” Comput. Eng., vol. 12, p. 77, 2007.

[23] Y. Qu, ZY; Liu, YM; Liu, Y; Jiu, K; Chen, “A Method for Reciprocating Motion Detection in Porn Video based on Motion Features,” in IEEE International Conference on Broadband Network and Multimedia Technology, 2009, pp. 183–187.

[24] Z. ZHAO, “Combining SVM and CHMM classifiers for porno video recognition,” J. China Univ. Posts Telecommun., vol. 19, no. 3, pp. 100–106, Jun. 2012.

[25] V. M. T. Ochoa, S. Y. Yayilgan, and F. A. Cheikh, “Adult Video Content Detection Using Machine Learning Techniques,” in 2012 Eighth International Conference on Signal Image Technology and Internet Based Systems, 2012, pp. 967–974.

[26] “Pornography Database.” [Online]. Available: https://sites.google.com/site/pornographydatabase/. [Accessed: 09-Nov-2017].