# On the Detection of Images Containing Child-Pornographic Material

Emilios Yiallourou, Rafaella Demetriou, Andreas Lanitis
Visual Media Computing Lab
Department of Multimedia and Graphic Arts
Cyprus University of Technology, Limassol, Cyprus
ef.yiallourou@cut.ac.cy, rp.dimitriou@edu.cut.ac.cy, andreas.lanitis@cut.ac.cy

***Abstract*—The vast increase in the use of social networks and other internet-based communication tools contributed to the escalation of the problem of exchanging child pornographic material over the internet. The problem of dissemination of child pornographic material could be addressed using dedicated image detection algorithms capable of rating the inappropriateness level of images exchanged through computer networks so that images with inappropriate content involving children are blocked. However, the complexity of the image detection task coupled with the nonexistence of suitable datasets, inhibit the development of efficient algorithms that can be used for detecting offensive images containing children. To deal with the problem, we propose a methodological approach that can be used for supporting the development of child pornography detectors through the generation of synthetic datasets and through the decomposition of the task into a set of simpler tasks for which training data is available. Preliminary results show the promise of the proposed approach.**

***Keywords—Child pornography detection, synthetic images, face detection, age estimation, expression recognition***

## I. INTRODUCTION

Child pornography [1] and the communication of child pornographic material over the internet, is becoming an important problem in cyber-security. The increased use of social-media also resulted in a vast increase in the circulation of offensive images of children that are often used for cyber-bullying activities. One way to compact the problem is to develop and use automated Child Pornography Detectors (CPD's) in the form of dedicated filters capable of detecting images containing offensive material involving under aged persons. However, the development of highly accurate CPD's is inhibited due to the unavailability of image datasets suitable for the necessary training and testing of CPD's. In the case of child pornographic material, the creation, possession, and dissemination of real data is illegal [2,3] and unethical hence unlike other related cases [4, 5, 6, 7, 8] it is impossible to use real images for experimentation. Furthermore, within-class variability of images containing child pornographic material is huge making the process of machine-based classification a highly challenging task.

Our preliminary work in the area aims to address the two major problems of (1) generating suitable datasets and (2) defining a process that allows the development of efficient CPD's despite the non-existence of suitable training data. Due to the sensitive nature of the topic, our work was carried out in close co-operation with the Local Police Authorities (Cybercrime Division) in order to get advice from cybercrime experts and at the same time to ensure that in all cases images used were legal.

As part of our work several images that possibly contain suspicious content were collected from the internet followed by a content-based analysis of each of those images so that the main features associated with suspicious images are determined. The correlation between inappropriateness level and image features reveals the features associated with images with suspicious content. Once features associated with suspicious content are determined it is feasible to a) create synthetic images displaying possibly suspicious images and b) use such features for detecting images with inappropriate content. Based on the results of a preliminary experimental evaluation the potential of the approach is demonstrated.

In the remainder of the paper a brief literature work is presented followed by the presentation of the work on defining features associated with suspicious images and the generation of synthetic images. In section IV the integration of image analysis algorithms is presented followed by the results of an experimental evaluation. Concluding comments and plans for future work are outlined in section VI.

## II. LITERATURE REVIEW

Traditionally, the problem of child pornography detection is tackled by building a database of websites and images that contain illegal material. For example PhotoDNA [9] is a commercial application that fingerprints images through a unique signature based on the overall visual presence of the picture and therefore should be resistant to various transformations. PhotoDNA enables automatic detection of already known child pornography material, and thus can limit or prevent the dissemination of such material. Notable users of PhotoDNA include Facebook and the United States National Center for Missing and Exploited Children (NCMEC). However, the number of websites and images with illegal content increases every day, making it impossible to keep track of all suspicious websites.

978-1-5386-0643-8/17/$31.00 ©2017 IEEE

Therefore, an approach based on detecting offensive images by real-time analysis of image content, could offer a more effective solution to the problem.

Recent developments in the field of computer vision resulted in methods that could be used for identifying pornographic material, since they can extract useful information from images. These methods are based on the rich visual information extracted from the original images (e.g. skin color, texture, shape, local invariant points). The pioneering work was done by Forsyth [4,5], after creating a mask for the areas covered by skin, combining the properties of texture and color. Subsequently the mask is passed through a filter based on the geometric shape of the body. Several other methods have been proposed in this field [6, 7, 8]. For example, Nian, et al. [6] propose a method for pornographic image detection where a dataset with more than 13 000 pornographic images with adults and over 35 000 normal images was used for training a deep convolutional neural network. It should be noted that all techniques and datasets related to pornography reported in the literature refer to pornographic material displaying adults rather than children.

Satta et al [10] attempt to define a set of functionalities that can be used for detecting child pornographic material through image analysis. They define the processes of face detection, age estimation, gender estimation, object detection and contextual information analysis as key functionalities for child pornography detection. In our work, we built on the suggestions offered by Satta et al [10] to materialize child pornography detectors.

## III. Feature Definition

In this section, we describe a methodological approach used for defining image features associated with inappropriate images containing children followed by the evaluation of such features for use in child pornography detectors

### *A. Defining Features Through Content-Based Analysis*

Several images collected from the internet were shown to volunteers. Each volunteer was asked to determine the inappropriateness level of each image in a scale 1 to 5 where "1" means that the image contents are totally innocent and as such the image could be published on the internet, whereas rating "5" indicates an image with suspicious contents that should not be published on the internet. Volunteers were also asked to determine which features influenced their decisions. A correlation analysis between the weights given to different features against the inappropriateness ratings revealed the image features associated with innocent and non-innocent image content (see Table 1). It should be noted that to be in-line with legislation regarding handling images displaying children, even images that were rated as inappropriate (rating "5") did not contain explicit child pornography scenes. However, the results of the content-based analysis are important for developing child pornography detectors because images with inappropriate content with children (like the ones rated with "5") are usually transmitted before images with explicit pornographic content, hence the ability to detect images with the features shown in Table 1 can be useful for detect early signs of suspicious activity.

### *B. Synthetic Image Generation for Feature Vality Verification*

Based on the image features associated with innocent and suspicious image content, 20 synthetic images incorporating such features were generated. Within this context 3d modelling tools were used for generating images with "innocent" content and subsequently certain image features were modified to convert "innocent" to "inappropriate" images. For example, in figure 1 (left) an image with innocent content was created that despite the contact between the adult shown and a child, the overall setting (i.e. clothing, lighting, genders of persons involved and number of children), indicates an innocent content. In contrast modification of few image features turns the image into a suspicious one (see Figure 1, right).

TABLE I. List of image features corresponding to images with innocent and suspicious content

| ***Image Features*** | ***Innocent*** | ***Suspicious*** |
|---|---|---|
| **Attire of persons shown in image** | jacket, limited skin presence | nude, swimwear, partial nudity |
| **Expressions / feelings of persons shown in image** | joy, smile, laugh | sad, frightened, thoughtful, horrified |
| **Gestures, physical contact and/or actions of persons shown in image** | child holding adult hand | sitting with legs apart in the tub and between the little child<br><br>man holds a girl<br><br>kissing on the mouth |
| **Age difference among persons in image** | small difference in age, youngster / father | significant age difference |
| **Illumination** | bright illumination | dark / dim lighting |
| **Body posture of persons in image** | standing position, holding hands with adult, running/ action / game | tucked, seated with legs apart, lying, high hands, hug |
| **Image scene** | Exterior, schoolyard, playground / park | |
| **Number of people in image** | many people, many children together | small number of children |

Subsequently, the inappropriateness level of synthetic images was rated by 20 volunteers. Per the results the average rating of synthetic images with innocent content was 2.11 whereas the average rating for "inappropriate images" was 3.72. Statistically the difference between the two metrics was found to be significant at a 5% confidence level. The results of the experiment involving synthetic images indicate

that the features listed in Table I can potentially be used as the basis for generating synthetic images that are perceived by human observers as "appropriate" or "inappropriate".

**Fig. 1.** Left: Synthetic image with innocent content. Right: Synthetic image with suspicious content.

## IV. QUANTITATIVE EVALUATION OF IMAGE FEATURES

Based on the definition of image features corresponding to images with suspicious content, a set of quantified features that could potentially be extracted automatically from images are defined. In this section, we outline the full set of quantified features and assess their potential in detecting images with inappropriate content.

### *A. Feature Quantification*

Based on the image features defined in Table I, three groups of quantified characteristics were defined. Group A of features are associated with characteristics of individual persons appearing in images, Group B is associated with the image scene and Group C is associated with face/body gestures and interactions among the persons shown in an image. In total eleven features (F1 to F11) are defined as shown in Table II. In addition to image features the inappropriateness level (IL) of each image is defined in a scale 1 to 5.

### *B. Assessing Representation Power of Quantified Features*

Three volunteers were asked to determine manually the values of each of the 11 features (F1 to F11) for a total of 50 images consisting of 30 real images and 20 synthetic images. For each of those images the inappropriateness level (in the scale 1 to 5) was also defined and the average values obtained from the three volunteers was considered as the ground truth.

Based on the numerical values of all features among an image set, a linear regression model was trained to learn the relationship between the variable IL and the 11 image features. The Leave-one-out cross-validation (LOOCV) method was used for evaluating the applicability of the regression models in the case that the regression model was built using data from the real images, the synthetic images and the combined training set consisting of both real and synthetic images. Moreover, the goodness of the regression model was evaluated when the model was trained on synthetic images and tested on real images. The results of the experiment (see Table III) indicate that the Mean Absolute Error (MAE) between actual and predicted inappropriateness level is less than 0.5, hence the potential of using the feature set defined in Table II as a basis of estimating the inappropriateness level of images is justified. In our preliminary investigation, the synthetic images used did not display realistic expressions hence it was expected to obtain better results when using just real images. However, it is worth mentioning that even in the case that the regression model was trained using features defined from synthetic images and tested on real images, the MAE is less than 0.5 indicating that it is feasible to use synthetic images for training child pornography detectors. Figure 2 shows examples of synthetic images and the corresponding errors between real and estimated IL.

TABLE II. LIST OF QUANTIFIED IMAGE FEATURES

| ***Group*** | ***Feature*** | ***Description*** |
|---|---|---|
| **A** | F1: Child presence | Number of persons with age <18 |
| | F2: Number of persons | Total number of persons in image |
| | F3: Attire | Statistics related to percentage of skin exposure among all persons in image |
| | F4: Age diversity | The standard deviation among the ages of all persons in the image |
| | F5: Gender distribution | Indicates presence of persons of different gender in the image |
| **B** | F6: Image setting | Boolean feature to indicate outdoors or indoors scene |
| | F7: Image Lighting | The brightness of the scene measured with three options: Low, Medium, and Strong lighting. |
| | F8: Image objects | Boolean feature to indicate presence of suspicious objects in the scene |
| **C** | F9: Expressions of persons in image | Measured as the proportion of positive expressions (joy, smile and laugh), the proportion of neutral expressions and the proportion of negative expressions (sad, afraid, thoughtful) |
| | F10: Person Interaction | Indicates physical contact and suspicious interactions between persons. |
| | F11: Body posture | Categorizes postures in Suspicious / Neutral / Non-suspicious images. |
| **N/A** | IL: Image content inappropriateness Level | Defined in a scale 1 to 5 (1: innocent content, image could be published on the web and 5: image with offensive content that should not be published on the web). |

TABLE III. REGRESSION ANALYSIS RESULTS USING FEATURES F1-F11

| ***Training/Test Set*** | ***Mean Absolute Error*** |
|---|---|
| Synthetic LOOCV | 0.4396 |
| Real LOOCV | 0.3294 |
| Combined Syntetic/Real Images LOOCV | 0.4064 |
| Synthetic (train), Real (test) | 0.4989 |

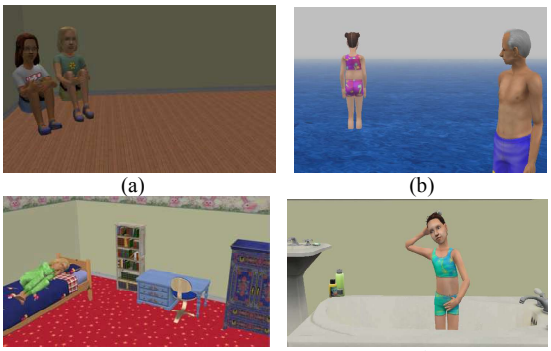

**Fig. 2.** Typical synthetic images used for training/testing the regression models and the corresponding IL errors obtained. Image (a) Actual = 2.95, Prediction = 2.951, Error = 0.001), Image (b) Actual = 3.36, Prediction = 2.043, Error = 1.317, Image (c) Actual= 3.05, Prediction = 2.99, Error = 0.06, Image (d) Actual = 3.9, Prediction = 5.077, Error = 1.177.

## V. AUTOMATIC EXTRACTION OF IMAGE FEATURES

In this section, we assess the viability of estimating the image inappropriateness level using automatically extracted image features rather than using manually annotated features. In our preliminary investigation, we focused our attention on a subset of the following five image features: F1: Child presence, F2: Number of people, F4: Age diversity, F5: Gender Distribution and F7: Lighting. For the calculation of these features algorithms from Intel's "Open Source Computer Vision Library" (OpenCV) [11] were used. A description of the actual algorithms used is provided in the following section.

### *A. Algorithms Used For Extracting Image Features*

**Face Detection:**

Face detection [12] is an essential part of the entire process of feature estimation, since the results of face detection are required for implementing age and gender estimation algorithms. In addition, face detection can provide the value for feature F2 (Number of people in image). For face detection, we used the cascade classifier based on Haar features proposed by Viola and Jones [12] and developed by Lienhart et al. [13]. However, the face detector encounters difficulties when the persons depicted are not in frontal posture, or when dealing with occluded faces.

**Age and Gender Estimation:**

Age [14] and gender estimation [14] is also an important part of the process since from this method it is possible to estimate values for three features (F1: Child Presence, F4: Age diversity, and F5: Gender distribution). For age and gender estimation the approach suggested by Eidinger et al [15] was used along with the dataset provided for training. The dataset consists of 4550 cropped and aligned faces found in nature (wild) with age group and gender labels. The seven age groups given, are defined as follows: 0-2, 3-7, 8-12, 13-19, 20-36, 37-65, 66+. The classification process relies on a representation based on local binary patterns (LBP) and Four-Patch LBP codes (FPLBP) followed by a single linear SVM classifier.

**Lighting Intensity Estimation:**

To calculate the lighting intensity of an image a formula based on the HSP color system proposed by Finley [16] was used. Based on this formulation the perceived brightness in an image (P) is calculated using equation 1.

$$P = R * 0.299 + G * 0.587 + B * 0.114$$

Equation 1 results in values ranging from 0 to 1, where values close to 1 represent high brightness. In our formulation P values in the range [0, 0.4], [0.4,0.55] and [0.55,1] are considered as Low, Medium and Strong illumination respectively.

## VI. EXPERIMENTAL EVALUATION

In this section, we describe experiments where we aim to estimate the inappropriateness level of previously unseen images based on an automatically extracted subset of features F1, F2, F4, F5 and F7. During the training stage, a linear regression models was trained, using features values extracted from the training set of combined real-synthetic images introduced in section III.

Testing was carried out using 200 previously unseen images downloaded from the internet. Three volunteers rated the inappropriateness level of test images and the average ratings among the three observers was considered as the ground truth. The algorithms outlined in section V were used for estimating the values of features F1, F2, F4, F5 and F7 for all 200 test images, so that it was possible to obtain estimates for the inappropriateness level of each test image based on the automatically extracted features and the regression model defined during the training process.

For better analysis of the results we present the results for all images (200 images) and separately for images where at least one face was detected (100 images). This separation was done as in the images where no person is detected, the calculation of the inappropriateness level was based only on the value of F7 (image brightness). Per the results (see Table IV) MAE for images with detected faces was 0.9519 whereas the MAE for all images was 0.9974.

TABLE IV. RESULTS OF INAPPROPRIANETES LEVEL ESTIMATION ON UNSEEN IMAGES

| | ***Mean Absolute Error*** | ***Max Error*** | ***Min Error*** |
|---|---|---|---|
| All images | 0.9974 | 3.3048 | 0.0018 |
| Images with detected Face | 0.9519 | 2.9002 | 0.0129 |

Figure 3 shows the percentage of images in relation to the error produced. Errors obtained were divided into five groups (0-0.5, 0.5-1, 1-1.5, 1.5-2, 2+). According to the results for more than half of the images in both sets an error less than 1 was recorded. In both cases for 75% of the images the recorded error was less than 1.5.


**Fig. 3.** Graph of the proportion of images (X-axis) with respect to the error showed (Y axis).

Table V shows the confusion matrix obtained when all test images are separated in the group of images appropriate for publishing (IL values 1 to 2.5), neutral (IL values 2.5 to 3.5), and inappropriate for publishing (IL values 3.5 to 5). As shown in Table V, almost half the photos (96) have been classified in the correct group. 67 of the 88 inappropriate images (76.14%) have been considered correctly inappropriate. Most importantly only four inappropriate images (2%) were classified as appropriate.

TABLE V. CONFUSION MATRIX OF CLASSIFICATION RESULTS

| | **Classified Appropriate** | **Classified Neutral** | **Classified Inappropriate** |
|---|---|---|---|
| **Appropriate** | 7 | 8 | 22 |
| **Neutral** | 10 | 22 | 43 |
| **Inappropriate** | 4 | 17 | 67 |

## VII. CONCLUSIONS

A methodological approach that can be used for supporting the development of child pornography detectors through the generation of synthetic datasets and through the decomposition of the task into a set of simpler tasks for which training data and suitable algorithms are available, was presented. Preliminary results show the promise of the proposed approach. It is worth mentioning that false negative error obtained (i.e. inappropriate images classified as appropriate) was very low, suggesting that a filter like the one developed could be efficient in blocking offensive material. The results of the experiment also indicate that the use of synthetic images for training CPD's is feasible.

The experimental results reported make use of only five features. In the future, we plan to use machine vision algorithms to extract and use in the classification all 11 features. Given the success of deep learning algorithms [17] in computer vision applications, the use of deep learning will be incorporated in the process. Also, we plan to test the method on extended datasets. The product of this work will be incorporated as a plug-in in internet browsers so that offensive images containing children will automatically be blocked during the uploading or downloading process, providing in that way a powerful tool for protecting under-aged internet users.

## ACKNOWLEGEMENTS

Authors acknowledge help from the Cyprus Police Authorities (Cybercrime Division) and seed funding from the European Union's Horizon 2020 Project ENCASE (http://encase.socialcomputing.eu/)

## REFERENCES

[1] T. Tate, Child pornography: An investigation., London: Methuen, 1990.

[2] K. S. Williams, "Controlling internet child pornography and protecting the child", Information & Communication Technology Law, vol. 12, no. 1, pp. 3-24, 2003.

[3] S. Edwards, "Prosecuting child pornography: Possession and taking of indecent photographs of children." The Journal of Social Welfare & Family Law 22, no. 1, 1-21, 2000.

[4] M. Fleck, D. Forsyth, and C. Bregler, "Finding Naked People", European Conference on Computer Vision, vol. III, pp. 592-602, 1996.

[5] D. Forsyth and M. Fleck, "Identifying nude pictures", IEEE Workshop on the Applications of Computer Vision, pp. 103-108, 1996.

[6] F. Nian, T. Li, Y. Wang, M. Xu, J. Wu, "Pornographic image detection utilizing deep convolutional neural networks", Neurocomputing, p. 283–293, 19 October 2016.

[7] QF. Zheng, W. Zeng, G. Wen, WQ. Wang, "Shape-based Adult Images Detection", International Journal of Image and Graphics, vol. 6, no. 01, pp. 115-124, 2006.

[8] C. Jansohn, A. Ulges, and T. M. Breue, "Detecting pornographic video content by combining image features with motion information", 17th ACM international conference on Multimedia, Beijing, 2009.

[9] "PhotoDNA Cloud Service", November 2016. [Online]. Available: https://www.microsoft.com/en-us/PhotoDNA.

[10] R. Satta, J. Galbally, L. Beslay, "State-of-the-Art review: Video Analytics for Fight against on-line Child Abuse", Publications Office of the European Union, Luxembourg, 2013.

[11] G. Bradski, Dr. Dobb's Journal of Software Tools, 2000.

[12] P. Viola and M. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features", in IEEE Computer Sociefy Conference on Computer Vision and Pattern Recognition, 2001.

[13] R. Lienhart and J. Maydt, "An Extended Set of Haar-like Features for Rapid Object Detection", IEEE ICIP, 2002.

[14] G. Panis and A. Lanitis. "An overview of research activities in facial age estimation using the FG-NET aging database." In European Conference on Computer Vision, pp. 737-750, 2014.

[15] E. Eidinger, R. Enbar, and T. Hassner, "Age and Gender Estimation of Unfiltered Faces", IEEE Transactions on Information Forensics and Security (IEEE-TIFS), special issue on Facial Biometrics in the Wild, vol. 9, no. 12, pp. 2170-2179, 2014.

[16] D. R. Finley, "HSP Color Model — Alternative to HSV (HSB) and HS", 2006. [Online]. Available: http://alienryderflex.com/hsp.html.

[17] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning." Nature 521, no. 7553: 436-444, 2015.