Forensic Detection of Child Exploitation Material using Deep Learning

Mofakharul Islam
Department of Computer Science and Computer Engineering
Latrobe University
Australia.

Abdun Nur Mahmood
Department of Computer Science and Computer Engineering
Latrobe University
Australia.

Paul Watters
Department of Computer Science and Computer Engineering
Latrobe University
Australia.

Mamoun Al Azab
School of Science & Information Technology
Charles Darwin University
Australia.

Keywords: Child, Adult. Face, Features, Exploitation, Deep Learning, Detection

Abstract :- A precursor to successful automatic child exploitation material recognition is the ability to automatically identify pornography (largely solved) involving children (largely unsolved). Identifying children's faces in images previously labelled as pornographic can provide a solution. Automatic child face detection plays an important role in online environments by facilitating Law Enforcing Agencies (LEA) to track online child abuse, bullying, sexual assault, but also can be used to detect cybercriminals who are targeting children to groom up them with a view of molestation later. Previous studies have investigated this problem in an attempt to identify only children faces from a pool of adult faces, which aims to extract information from the basic low- and high-level features i.e., colour, texture, skin tone, shape, facial structures etc. on child and adult faces. Typically, this is a machine learning-based architecture that accomplish a categorization task with the aim of identifying a child face, given a set of child and adult faces using classification technique based on extracted features from the training images. In this paper, we present a deep learning methodology, where machine learns the features straight away from the training images without having any information provided by humans to identify children faces. Compared to the results published in a couple of recent work, our proposed approach yields the highest precision and recall, and overall accuracy in recognition.

INTRODUCTION:

Categorical age estimation by machine offers substantial extent of challenges and difficulty to the computer vision community due to unavailability of complete knowledge on human visual system. Study revealed many factors include among others; individual's lifestyle, health, race, occupation, cultural background has a significant impact on human ageing process especially on face on top of people's gene. Over the last three decades, much interest have been paid adult face detection only but nothing significant has been contributed on its counterpart. An effective and efficient way that attempts to quantify children's face in ways that agree with human intuition needs to have at least some of the similar cues akin to human visual system in making judgment. Defining such cues certainly will put this research to a considerable amount of challenge. The issue is not clear how to extract features that have discriminative ability for age estimation using skin tone or other facial cues visible in naked eyes.

Deep learning, a sub-type of machine learning, is a new era where machine can learn many layers of perception and depiction to create a sense of data like text, audio, and image leading object detection problems in to the next level. Each layer in the network takes in data from the previous layer, transforms it, and passes it on. The network increases the complexity and detail of what it is learning from layer to layer. Here the network learns straight away from the data—no influence from human at all over what salient features are being learned by the network. Deep learning is a data-hungry scheme requires a huge volume of data to produce accurate result while machine learning accuracy is almost plateaus – reach a state of little or no change.

RELATED WORK:

Kwon & Lobo [1] initially employed high resolution facial images to categorise facial images into three age groups – senior adults, younger adults or child. Application on a limited image dataset not allowing this to consider as a robust approach in categorical age estimation though authors claimed 100% accuracy. Later, they suggested another classification technique to accomplish the categorical age estimation where they employed a craniofacial development theory and wrinkle analysis [2]. Once more their approach is disadvantaged by inadequate image, which may result in poor performance while applied on large real-world datasets. The wrinkle analysis employed again by Hayashi et al. [3] combined with geometric relationships between visual structures on a human face for categorical age estimation. Their approach is primarily aimed at classifying age into multiple groups at the five years intervals which in fact, is an age-group wise classification technique - not exactly a categorical age estimation technique.

Lanitis et al. [4], employed Active Appearance Models (AAM) [5] which combined shape and texture parameters to extract using three classifiers – simple quadratic fitting, shortest distance, and Neural Network classifier to estimate the age. Finally, the age estimation accuracies of these classifiers are compared. An algorithm based on uncertainties of the age labels is suggested by Yan et al. [6] to categorise age-group. An approach combining AAM with Support Vector Machine Regression is suggested by Khoa et al. [7] for categorical age estimation. AAM's parameters are employed initially to perform categorical age estimation (adult and children) followed by an age estimation using an age-determination function based on Support Vector Machine Regression. In fact, AAM's parameter-based classification is not an ideal approach to age categorization in the sense that these appearance data are not exclusively able to express enough cues that are necessary to produce an accurate result in classification. Finally, the authors suggested for further expansion in categorical estimation part, in particular, to develop it as a robust categorical age estimation approach.

Geng et al. proposed an approach where ageing patterns are generated for each person in a dataset of facial images presenting every single subject at different ages [8]. The authors introduced a single sample, a collection of temporal face images for each subject, which is finally projected to a low dimensional space. At testing phase, an unseen face is substituted at multiple orientations in a pattern to indicate age of the subject using the process of minimizing the reconstruction error. Here, the authors suggest methods using unique characteristics of ageing like ageing patterns perform better than the standard classification techniques.

Ageing patterns using manifold learning is demonstrated by Fu and Huang [9], where they employed a manifold criterion based discriminative subspace learning to represent the low-dimensional ageing manifold. The authors suggest age estimation improves significantly when Regression is applied on the ageing manifold patterns. Based on this findings Guo et al. [10] applied a Support Vector Machine Regressor (SVR) to learn the relationship between age and coded face representations. Application of a local SVR trained with only ages within a small interval around the initial age estimate utilizing a refined age estimation using a local SVR is the key aspect of this particular work.

Error-Correcting Output Codes (ECOC) on the fused Gabor and LBP features of a face image are employed by Wang et al. [11] to categorise an individual into one of four possible age groups (child, teen, adult and senior adult) . The authors found ECOC combined with AdaBoost or SVM solved the multiclass learning problem like age categorization. FG-NET and Morph datasets are employed to obtain experimental outcomes on its effectiveness and robustness in age categorization and found the algorithm based on the fused features performing better than the one based on Gabor alone or LBP alone.

Our literature review reveals a robust approach for categorical age estimation especially adult and children is still missing – the void which this research is looking to fill.

MODEL IMPLEMENTATION AND TRAINING:

Someone easily can envisage how much challenges are there to capture all the nuances of child face using only low and high level features like colour, textures, and other hand-crafted descriptors alike while, excluding adult faces. In this paper, a new method of a novel child face detection technic is presented that detects features considered distinctive, compared to the basic low and/or high level features (colour, texture, shape, edge etc.) and other hand-crafted features. The proposed approach leverage knowledge from deep convolutional neural networks to acquire discriminative patterns straight away from the available data. However, it involves a high price tag, the requirement for proper training data incorporating the diverse viewpoints and problem point of view itself. This deep learning approach performs more accurately than existing methods, as demonstrated empirically using low and high level features and other hand-crafted features.

A ubiquitous phrase “a picture is worth a thousand words”. That might be true for humans, but can a machine find the same meaning in the images too? Human has got photoreceptor cells in their retina to pick up wavelength of light, but that information doesn’t seem to propagate up to human consciousness as a human can’t express exactly what wavelength of light he/she is picking up. In a similar manner, a camera captures pixels only. Now, the challenge is how we get human level perception from these pixels.

In this vein, in this paper, we propose a novel approach that allow us to take benefits of deep learning concepts while, at the same time, not having adequate data to train a data-driven solution for child face detection from scratch. The basic features like colour, texture, shape etc.) along with other handcrafted features are in fact, descriptors that are built on a set of cue depending upon the specific feature being used to characterise regions of interest in an image. While on the other hand, deep learning paradigm use Convolution Neural Network (CNN), a set of structured layers of neural network, where each layer acts as a feature extractor resulting quite generic and a specific classification task independent.

One of the great advantage of the Convolutional neural networks (CNN) is that a local understanding of any imagery data is good enough to generate expression on that data resulting fewer parameters that greatly improves the data learning efficiency as well as reduces the amount of data required to train the model. While on the other hand, a fully connected layer (FCN) takes weights from each pixel resulting a larger number of parameter. CNN looks at a small patch of the image at a time to learn the weights rather than looking at each pixel in FCN. In fact, CNNs are a clever way to reduce the number of parameters by reusing the same parameters multiple times instead of dealing with FCN. In CNN model, the number of parameters is independent of the size of the original image.

Our proposed CNN model will learn how to classify images to one of the two candidate categories. In effect, "a picture is worth only one word" out of just two possibilities. The proposed method starts with input images to subject them under a series of convolution, pooling, and local response normalization operations as found to be applied in the literatures.

The proposed network consists of three convolutional layers each of them having a Rectified Linear Units (ReLU) layer, three max pooling layers, one fully connected layer, and one final classification module.

Everything starts in the proposed network with receiving input images and transforming them with a series of convolution, pooling, and local response normalization operations, similarly to many other convolutional neural network models in the [12]. We are primarily aimed at observing the larger objects structures and then gradually looking at the smaller objects. Training images into a convolution layer combined with a Rectified Linear Units (ReLU) layer, where we initially applied a 7 x 7 kernel filter to capture larger structural facial features like nose, lip, mouth, ear, eye, eye brow, chin etc. followed by a max pooling layer. ReLU applies to maintain the non-linearity of the network with the help of a non-saturating activation function, which increases the nonlinear properties of the decision function and of the overall network without affecting the receptive fields of the convolution layer. Max pooling layer applied to perform rescaling or subsampling a convolved output helps reduce the number of parameters that eventually help to not overfit the data. Rescaling or subsampling a convolved output helps reduce the number of parameters, which in turn can help to not overfit the data. The main idea behind this max pooling technique is, it sweeps a window across an image and picks the pixel with the maximum value. Depending on the stride length, the resulting image is a fraction of the size of the original. In our proposed technique we applied a stride length of 2. Another 7 x 7 kernel filter applied on the output (feature map) from the first convolutional layer to capture more details on the larger facial structure.

We applied similar convolution layers, each having a 5 x 5 kernel filter, and a 3 x 3 kernel filter along with a ReLU layer inside and a max pooling layer to incorporate finer details of both the larger and smaller facial objects, each of them followed by a convolution with the same size of kernel filter on the output of the respective convolution. After several convolutional and max pooling layers, the high-level reasoning in the neural network is done via fully connected layers. Neurons in a fully connected layer have connections to all activations in the previous layer, as seen in regular neural networks. Finally, a softmax classification layer is applied to convert our fully connected layers output into probability with the help of a softmax function that takes the final feature vector and transforms it into a vector of real number in range (0,1).

Now our deep neural network model is fully trained to classify face images in to different category.

To address the problem of overfitting, we employed polynomial power of 0.5 alongside a weight decay of 0.01 as a bias term to ensure smooth backpropagation. A bias term and an activation function ReLU are introduced in the convolution layer to maintain its nonlinear behaviour that improves expressiveness. We select a dropout rate of 35%, a learning rate of 0.0005, and a maximum number of 200 epochs while we trained our CNN model. Further, we employ a regularization term in the cost function to combat overfitting. A dropout function is applied to all the layers to minimize overfitting where it takes a randomly selected number of features in that layer while training to make the network redundant and robust to infer output.

EXPERIMENT SETUP, RESULTS, & DISCUSSION:

Our proposed approach primarily aimed at developing a robust categorical age estimation tool that can detect adult and child faces automatically. Initially, we have tested our method on a wide variety of facial image databases freely available on the Internet to check its performance and robustness in a real-world scenario. Finally, a more intensive testing has been conducted for performance evaluation on classification accuracy along with an exhaustive performance comparison with few existing classification techniques. Accuracy and finding the correct category are the criteria based on which effectiveness of the proposed approach is measured. We employed a standard set of criteria and procedure to make a judgement on the detection/ classification accuracy.

One of the most significant challenges in investigation and implementation of such a sensitive detection system is non-availability of appropriate and authenticated facial datasets on children which is considered to be a major constraint in this particular area of research. For validation purpose, we need datasets on facial images on both adult and children having zero or minimum expression. While freely available datasets are extremely limited, a few authenticated paid databases are available from different research communities. In this paper, we utilise the freely available databases for our experimental evaluation, validation and performance comparison.

Further, non-availability of images having zero or minimum expression is another bottleneck as the majority of the databases are predominantly targeted to a classification based on the facial expression – regardless of adult or child.

Generally, face images in these databases are not ready to use straight away and need pre-processing like cropping, scaling, resizing and enhancement. A vast majority of them aren't suitable for use due to poor resolution and other defects of the original image. Apart from that, the individual photo needs to be cropped from group photos to meet our project's requirement. To run our experiment, children, youths, and seniors have taken into consideration. So, we have selected the required number of images from the aforementioned image databases while maintaining the actual requirement of our purpose. Facial images having neutral or minimum expression are carefully selected though these are hard to get. We have taken extra care in choosing the required facial images from these databases as mentioned above. Further, infants are discarded from the children as they are considered not to be vulnerable as reported by a wide range of literature, article, and online surveys in regards to a binary categorization – either adults or children.

Now, we first outline the datasets and algorithms used in both experiments. We then outline the method for analysing the results.

Face images for training and testing have been sourced from [13], [14], and [15]. The NCMEC is a publicly available database maintained by an US organization working on missing and exploited children and the FG-NET contains scanned face images from newborns to senescence. The LAG facial database contains 3,828 images having images ranging from child/young to adult/old. We have selected 2,000 images from these databases to run our experiment for comparing our results with that of a couple of recent published work. Table 1 summarizes the number of facial images taken from different datasets.

Corrupted value in the imaging data set is one of the major issues as noise are being introduced due to diverse imaging artifacts like low intensity, intensity inhomogeneity, and shadow. Camera electronics and its lens characteristics play a significant role to embed further noise. More noise means more uncertainty. A low-pass filter is applied to the data set to minimize the overall impact of noises resulted from the inherent imaging artefacts. Apart from these difficulties, other difficulties like intensity inhomogeneity, unclear edges of different segments or objects shadows, shading and highlight effects in image dataset are also present in the data set. We apply adaptive histogram equalization to eliminate these difficulties.

TABLE I: FACIAL IMAGES TAKEN FROM DIFFERENT DATABASES.

Method	***Adult***	***Children***	***Total***
FG-NET	300	100	400
NCMEC	0	200	200
LAG	700	700	1400

Out of the three datasets as mentioned above, 1000 images have been chosen from each category (adult or child) for our experimental purpose. In order to generate new training and test images our image dataset has been augmented through flipping horizontally or vertically and scaling up or down resulting in 4 times bigger the size of the original dataset.

We replace softmax classification layer as we used during training of our CNN model with a binary classifier in our testing phase as we are primarily aimed at classifying the images into two groups – adult or child. Non-linear Support Vector Machine (SVM) with RBF kernel [16] is applied here to classify the images using the LIBSVM library. A grid search is performed on the training set to estimate the SVM parameter.

Our test image having a size of 256 x 256 is fed initially into the trained CNN model to classify, where it passed sequentially through a series of CNN layers to extract the features of the input image and finally fed into the SVM classifier following a fully connected layer where extracted features are accumulated in to a feature vector.

Our literature review suggest different authors contributed a good number of face detection techniques already over the decades though not paid too much attention in the area of categorical age estimation, the issue we are dealing with in this particular paper. For quantitative performance comparison, we have to rely on only a couple of literature [7] and

[17] as no other categorical age estimation approach is currently exists. A popular measure, called Precision and Recall (Eq. 1 & 2), based on True Positive (TP), False Positive (FP), and False Negative (FN) is applied on both - our experimental results and the results obtained from [7], to evaluate the classification performance of our novel approach. Apart from the Precision and Recall, we compared accuracy (Eq. 3) of our approach with [7] & [17] .

$$Precision = \frac{TP}{TP+FP} \quad (1)$$

$$Recall = \frac{TP}{TP+FN} \quad (2)$$

$$Accuracy = \frac{TP+TN}{T+N} \quad (3)$$

Out of 8000 face images in our dataset, we have divided the dataset into training and testing sets in to a 70:30 ration each of which representing child and adult face equally. Our CNN model is trained with the training images to classify the test images.

The recognition results were evaluated on a dataset composed of 2400 images (not included in the training set) – 1200 for children, the rest of the 1200 images were adult face images taken from the training set.

TABLE II: COMPARISON OF PERFORMANCE OF OUR PROPOSED METHOD WITH OTHER EXISTING METHODS.

Methods	TP	FP	FN	TN
Khoa et al	937	466	263	734
ICSIPA	1119	88	76	1117
Proposed	1163	36	37	1164

TABLE III: COMPARISON OF ACCURACY OF OUR PROPOSED METHOD WITH OTHER EXISTING METHODS.

Methods	Precision (%)	Recall (%)	Accuracy (%)
Khoa et al.	67	78	70
ICSIPA	93	94	93
Proposed	97	97	97

CONCLUSION:

We have proposed a novel deep learning-based child face model, where low level features works in tandem with some high level features as extracted automatically by a trained CNN

model on facial images to detect child and adult image with maximum accuracy. To the best of our knowledge, this is the first deep learning-based approach which is able to categorize a face image – either a child or adult successfully with maximum accuracy. In doing this, it paves the foundation for research in this particular area. Apart from the categorical age detection, this robust approach will also able to recognise real contents using contextual constraints in images from a wide range of applications like security & surveillance, human body parts identification, dermatological applications and even in marketing – tailoring advertising in shop windows based on the age of shoppers walking past. As a result, an extended avenues and usefulness for research in computer vision domain emerges as an outcome of this particular work. These alternate uses for the outcomes of this particular work provide extended usefulness and avenues for extended research in cybercrime and computer vision domain. This work can also facilitate large-scale analyses of the prevalence of child exploitation material online [18], which is a precursor to identifying strategies to reduce the demand for this material [19].

REFERENCE:

1. Kwon Y.H., Lobo Vitoria N. da (1993) Locating facial features for age classification, In proceedings of SPIE – the International Society for Optical Engineering, vol. 2055, pp.62-72, 1993.
2. Kwon Y.H., Lobo N. (1999) Age classification from facial images. Computer Vision and Image Understanding, 74(1):1–21, 1999.
3. Hayashi J. et al (2001) Method for estimating and modeling age and gender using facial image processing. Seventh International Conference on Virtual Systems and Multimedia, pp. 439–448, 2001.
4. Lanitis A. et al (2004), Comparing different classifiers for automatic age estimation. IEEE Transactions on Systems, Man and Cybernetics, Part B, 34(1):621–628, 2004.
5. Cootes T.F. et al (2001) Active Appearance Models, IEEE Transactions on Pattern Analysis and Machine Intelligence 2001.
6. Yan S. et al (2007) Auto-Structured Regressor from Uncertain Labels. International Conference on Computer Vision, 2007.
7. Khoa L. et al (2009) Age Estimation using Active Appearance Models and Support Vector Machine Regression, Biometrics: Theory, Applications, and Systems, 2009. BTAS '09. IEEE 3rd International Conference, page(s): 1 – 5, Washington, USA, 2009.
8. Geng X. et al (2007) Automatic Age Estimation Based on Facial Aging Patterns, IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 2234-2240, Dec. 2007.
9. Fu Y., and Huang T.S. (2008) Human Age Estimation with Regression on Discriminative Aging Manifold," IEEE Trans. Multimedia, vol. 10, no. 4, pp. 578-584, June 2008.
10. Guo G. et al (2008). Image-Based Human Age Estimation by Manifold Learning and Locally Adjusted Robust Regression, IEEE Trans. Image Processing, vol. 17, no. 7, pp. 1178-1188, July 2008.
11. Yan S. et al (2009) Synchronized Submanifold Embedding for Person Independent Pose Estimation and Beyond, IEEE Trans. Image Processing, vol. 18, no. 1, pp. 202-210, Jan. 2009.
12. Krizhevsky A. (2012) ImageNet classification with deep convolutional neural networks, in Advances in Neural Information Processing Systems, 2012, pp. 1097{1105.

13. NCMEC, National Center for Missing & Exploited Children (NCMEC), http://www.missingkids.com
14. Face and Gesture Recognition Research Network, http://www.fgnet.rsunit.com/
15. Simon B. (2017) Large Age-Gap Face Verification by Feature Injection in Deep Networks, Pattern Recognition Letters, Vol. 90, 2017, DOI 10.1016/j.patrec.2017.03.006.
16. Chang C.C, Lin C.-J (2011) LIBSVM: A library for support vector machines, ACM Transactions on Intelligent Systems and Technology 2 (2011) 1{27, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.
17. Islam M. et al (2011), Child Face Detection Using Age Specific Luminance Invariant Geometric Descriptor, 2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA2011).
18. Watters, P.A. (2018). Investigating malware epidemiology and child exploitation using algorithmic ethnography. *Proceedings of the 51st Hawaii International Conference on Systems Science (HICSS)*, Hawaii
19. Watters, P.A. (2018). Modelling the efficacy of auto-internet warnings to reduce demand for child exploitation material. *Proceedings of the 22nd Asia-Pacific Conference on Knowledge Discovery and Data Mining Workshops*