Digital Investigation 26 (2018) 40–54

Contents lists available at ScienceDirect

Digital Investigation

journal homepage: www.elsevier.com/locate/diin

Laying foundations for effective machine learning in law enforcement. Majura – A labelling schema for child exploitation materials

Janis Dalins [a, b, *], Yuriy Tyshetskiy [d], Campbell Wilson [a], Mark J. Carman [a], Douglas Boudry [c]

[a] Monash University, Caulfield, VIC, Australia
[b] Australian Federal Police, Melbourne, VIC, Australia
[c] Australian Federal Police, Barton, ACT, Australia
[d] Data61, CSIRO, Eveleigh, NSW, Australia

ARTICLE INFO

Article history:
Received 31 March 2018
Received in revised form
28 May 2018
Accepted 28 May 2018
Available online 31 May 2018

Keywords:
Neural networks
Digital forensics
Child exploitation
Forensic triage
Annotation schema

ABSTRACT

The health impacts of repeated exposure to distressing concepts such as child exploitation materials (CEM, aka 'child pornography') have become a major concern to law enforcement agencies and associated entities. Existing methods for 'flagging' materials largely rely upon prior knowledge, whilst predictive methods are unreliable, particularly when compared with equivalent tools used for detecting 'lawful' pornography. In this paper we detail the design and implementation of a deep-learning based CEM classifier, leveraging existing pornography detection methods to overcome infrastructure and corpora limitations in this field. Specifically, we further existing research through direct access to numerous contemporary, real-world, annotated cases taken from Australian Federal Police holdings, demonstrating the dangers of overfitting due to the influence of individual users' proclivities. We quantify the performance of skin tone analysis in CEM cases, showing it to be of limited use. We assess the performance of our classifier and show it to be sufficient for use in forensic triage and 'early warning' of CEM, but of limited efficacy for categorising against existing scales for measuring child abuse severity.

We identify limitations currently faced by researchers and practitioners in this field, whose restricted access to training material is exacerbated by inconsistent and unsuitable annotation schemas. Whilst adequate for their intended use, we show existing schemas to be unsuitable for training machine learning (ML) models, and introduce a new, flexible, objective, and tested annotation schema specifically designed for cross-jurisdictional collaborative use.

This work, combined with a world-first 'illicit data airlock' project currently under construction, has the potential to bring a 'ground truth' dataset and processing facilities to researchers worldwide without compromising quality, safety, ethics and legality.

Introduction

Reports of increasing workloads, employee 'burn-out' and psychological trauma are common across law enforcement and the judiciary, but the stresses and harms associated with exposure to psychologically harmful and offensive materials (typically child exploitation materials (CEM)[1] and violent imagery associated with online radicalisation) are now regarded as having been underestimated - particularly in instances of regular, lower level exposure. Law enforcement organisations such as the Australian Federal Police (AFP) traditionally employ a combination of regular psychological monitoring and mandatory staff rotations as a mitigating strategy, but these reduce skillsets within relevant teams (further exacerbating the problem), and tend to be reactive to persons already experiencing symptoms of harm.

In this paper we introduce the 'Stonefish' classifier - a machine learning (ML) tool demonstrating the feasibility of automated classifiers for CEM detection, both as triage tools and 'early warning' services for reviewers. This classifier uses supervised learning, an approach requiring high quality training and test data reflective of the 'real world' landscape. We assemble and utilise a collection of

* Corresponding author.
E-mail addresses: janis.dalins@monash.edu, janis.dalins@afp.gov.au (J. Dalins), yuriy.tyshetskiy@data61.csiro.au (Y. Tyshetskiy), campbell.wilson@monash.edu (C. Wilson), mark.carman@monash.edu (M.J. Carman), douglas.boudry@afp.gov.au (D. Boudry).

[1] aka 'Child Pornography', 'Child Abuse Materials', 'Sexually Exploitative Imagery of Children (SEIC)'.

https://doi.org/10.1016/j.diin.2018.05.004
1742-2876/

AFP case data for training, and data from an unrelated case for testing. We detail challenges and safeguards implemented as part of the development process, specifically for practitioner welfare.

Furthermore, in response to practitioner complaints of incompatible tools and data, we introduce the Majura schema, a jurisdictionally independent labelling/annotation schema designed for use in developing ML techniques in the field.

Existing work

Existing work relevant to this paper can be broadly split into multiple categories - the impacts of exposure to CEM (and other offensive materials), the broader challenges in Digital Forensics affecting possible solutions, automated discovery of CEM (both in use and experimental), and the research limitations caused by a lack of relevant datasets.

Exposure to CEM

First-hand exposure to traumatic and offensive events is long documented as psychologically harmful. Surveys of police officers in provincial England and New York state (USA) by Brown et al. (1999) and Violanti and Aron (1995), respectively, indicated comparatively high levels of stress in exposure to traumatic events involving children. Both studies pre-date the mainstream emergence of online child sex abuse, but a key point of note appears to be stress associated with dealing with *victims* of crimes such as rape and child abuse being quite high, with police officers seen as potentially "becoming secondary victims" (Brown et al. (1999)) in such cases.

The absence of studies into the effects of exposure to child exploitation by forensic analysts and other persons involved in the investigation/prosecution process was observed by Edelmann (2010), who noted that employers such as the Metropolitan Police provide mandatory counselling to staff routinely exposed to such imagery.

More recently, Powell et al. (2015) conducted a survey of 32 law enforcement personnel across all Australian jurisdictions, specifically recording the reported impacts of exposure to child exploitation materials[2] within internet child exploitation investigations. Critically, the survey included not only sworn police, but also 'computer analysts' - a role arguably requiring even more regular and in-depth exposure to materials during the course of digital forensic analysis. Interestingly, some respondents indicated an experience akin to the previously mentioned 'secondary victimhood', though contrastingly, some perceived exposure to CEM as less harmful than direct 'interaction with victims of assault'.[3]

Specific factors were listed by survey respondents as increasing a risk of long-term effects from exposure:

- Perceived resemblances between victims and children known to the reviewer (particularly the reviewer's own children);
- 'Unexpected' viewing of child exploitation materials;
- Repeated exposure to specific images or offenders;
- Viewing the progression of an offender from viewer to contact offender[4]; and
- Perhaps unexpectedly, some respondents also reported increased distress from text, as opposed to imagery & multimedia.

An anonymous survey of US law enforcement personnel by Seigfried-Spellar (2017) identified differences in psychological distress between investigators and forensic analysts, with persons conducting both duties in CEM related cases reporting higher levels of traumatic stress than those working single roles. The author hypothesizes this is due to their requirement to both review CEM and interact with victims and offenders, a theory consistent with the "secondary victimhood" identified by Brown et al. (1999). Furthermore, whilst respondents *generally* used healthy coping strategies, those working dual roles "may be more likely to use sedatives …as a coping mechanism."

Powell et al. (2015) note that due to the large number of variables involved, individual investigators' reactions to CEM exposure are impossible to predict. Viewed together with the general reluctance by police to seek assistance, combined with a low (16%) level of mandatory counseling offered by the respondent's agencies, it appears quite feasible that the extent of exposure related stress and harm remains underreported across law enforcement.

As stated by Powell et al. (2015), "purchase of technological strategies for global reduction in exposure to images is therefore warranted".

Challenges in digital forensics

In Powell et al. (2014), the aforementioned study's authors also questioned their respondents about the challenges they personally encounter in the field of Digital Forensics. Identified issues particularly of relevance to this article included:

- Access to "image scanning" software - most likely a reference to CETS (refer Table 1) or another cryptographic digest based content recognition system (refer Section Automated CEM Discovery);
- Inadequate staffing, including a lack of relevant digital forensics experience; and
- The need for "complete" examination - courts requiring every relevant item (image/video) to be reviewed and categorised, rather than accepting a representative sample. A respondent quotes a staff member "going through 500,000 images".

More recently, Franqueira et al. (2017) conducted a targeted survey of Digital Forensic (DF) practitioners worldwide, seeking their comments on challenges in the field of online child exploitation. The survey returned similar results in regard to the stresses and impacts of exposure to such imagery, but the authors' stronger focus on technical specialists[5] resulted in a differing set of reported challenges:

- Emerging technologies such as automatic age estimation are not 'translating' into workable tools for improving practices;
- Stressful working conditions associated with viewing CEM, with recommendations for improving automation to "minimize exposure in the first place"; and
- A need to standardise operations, procedures and legal frameworks globally, necessitating an *"internationally recognised scale of indecency levels and a taxonomy of terms to bridge language and cultural differences"*

The absence of standardisation as a challenge is glaring in Powell et al. (2014), most likely due to the paper's Australian focus. Nine

[2] Referred to as 'internet child exploitation' materials within the paper.
[3] It is unclear if this refers to *sexual* or physical assault, given the context).
[4] The *abuser*, as opposed to viewer of abuse.
[5] The authors use 'DF' in a broad sense, encompassing first responders, consultants and other roles regularly exposed to the crime type.

Table 1
Child Exploitation Tracking System (CETS) scale, with AFP guideline for labelling/annotation of files in child exploitation investigations.

Category	CETS classification	Guide
1	CEM - No Sexual Activity	Depictions of Children with No Sexual Activity - Nudity, surreptitious images showing underwear, nakedness, sexually suggestive posing, explicit emphasis on genital areas, solo urination.
2	CEM - Solo\Sex Acts between children	Solo masturbation by a child or non penetrative sex acts between children. Includes the use of penetrative sex toys by the victim (if offender is using toy is Cat 4).
3	CEM - Adult Non-Penetrative	Non-Penetrative Sexual Activity between Child (ren) and Adult(s). Mutual masturbation and other non-penetrative sexual activity.
4	CEM - Child\Adult Penetrate	Penetrative Sexual Activity between Child (ren) and Adult(s) - including, but not limited to, intercourse, cunnilingus and fellatio.
5	CEM - Sadism\Bestiality\Child Abuse	Sadism, Bestiality or Humilitation (urination, defecation, vomit, bondage etc) or Child Abuse as per CCA 1995.
6	CEM - Animated or Virtual	Anime, cartoons, comics and drawings depicting children engaged in sexual poses or activity.
7	CEM - Non-illegal \Indicative	Non-illegal child material (believed to form part of a series containing CEM). Includes images of circumcision being performed.
8	Adult Pornography	All pornographic material not considered CEM related.
9	Ignorable	Banners and other non-objectionable graphics useful for establishing proportionality. System files and unrelated images - holiday snaps, landscape, family photos, etc.
0	Unchecked	Material not yet assigned a category.
	If in doubt (about age) make it Category 8 - Adult. If undecided between two categories - make it the lower catgegory.	

jurisdictions are included (6 States, 2 Territories, plus Federal), each with some degree of individual case law and procedures, but de-facto standardisation has occurred - both through the establishment of Joint Anti Child Exploitation Teams (JACETs) in each State/Territory, and the alignment of State legislation and availability of Federal legislation to State Police. Whilst not in blanket use across all prosecutions, the CETS (refer Table 1) scale is used across Australia as a standardised measure of offending, greatly simplifying joint investigations and cross-jurisdictional prosecutions.

Automated CEM discovery

Academic and commercially led research into the development of image classifiers in this field is somewhat limited, largely due to ethical and legal considerations. Firstly, possession of CEM is typically illegal (or at least heavily regulated), whilst the mental health implications of exposure to such materials at a level required for feature selection, training and validation are simply too great for most research and/or commercial organisations.

A dominant academic focus for automated CEM recognition is filename/textual analysis of likely content, particularly in the context of P2P networks - architectures such as LimeWire, BearShare, and BitTorrent allow the collection of metadata *without* downloading actual content, enabling researchers to stop short of crossing local laws and ethical boundaries. Steel (2009) provide a snapshot of the Gnutella network, using tokenised query responses to categorise files as likely child pornography. Whilst most terms are sanitised for ethical reasons, the author provides some indications of common ages and advertised features/actions. Panchenko et al. (2012) identified textual features from filenames of *known* files (provided by law enforcement), providing a level of confidence unachievable from query-based studies. Latapy et al. (2013) also observed specialised vocabulary exclusive to online paedophile activity, a fact supported by the authors of Peersman et al. (2016).

Based on first-hand experience, we can confirm the presence of such 'red flag' terms, though their presence seems to be highly correlated with distribution via P2P networks - we hypothesise this is due to uploaders 'advertising' the files to make them more attractive for download, possibly in order to maintain required upload/download quotas in such systems. A classification system based exclusively upon text analysis is problematic, particularly as filenames (a) are easily obfuscated and/or obscured, and (b) tend to become more descriptive as they are passed through various sharing networks. Arguably, files containing CEM bearing standard device-generated filenames (e.g. *DSC-0001.jpg*) are of far greater interest to law enforcement, possibly indicating a more 'upstream' sharer, or indeed generator, of content.

Content analytics is a more intensive, but arguably also more robust approach. Currently, the dominant method for automatically detecting known materials is via cryptographic hash (e.g. MD5, SHA-1) comparison, recognising identical materials at the *binary* level. Specialist algorithms such as PhotoDNA[6] can measure similarity, allowing the automated recognition of resized, skewed or otherwise slightly altered still images.

Although 'fuzzy' hashes such as PhotoDNA are more robust to changes than cryptographic digests, both approaches are restricted to recognising previously observed and annotated materials, restricting their value to 'downstream' in the sharing process - older files, most likely shared numerous times between production and detection. An obvious help by accelerating analysis, this does little to aid law enforcement in targeting producers and victims.

Approaches capable of detecting 'new' CEM are therefore warranted. Skin color analysis is one such technique, where the proportion of skin/flesh coloured pixels within an image is used as a measure of likely visible nudity. Undoubtedly simple, such measurement is susceptible to noisy results - either by 'innocent' exposed skin (passport photos being one example), and completely unrelated but coincidentally coloured objects such as certain variants of sand building renders. Fig. 1 shows the distribution of skin tone percentages (as a percentage of *all* pixels within each image) for our training and test corpora (refer Section CEM Corpora).

The plots show that unsurprisingly, skin tone alone can't be reliably used as a disambiguator of CEM and adult pornography, nor CEM categories themselves. However, they show that (a) CEM tends to involve lower skin tone percentages than adult pornography, but also (b) consistent with Vitorino et al. (2018), the more extreme categories of CEM (particularly CETS category 5) tend to involve less skin as a proportion of the image, but such feature distributions are largely dictated by the downloading user. Image segmentation (e.g. into foreground and background or animal/human) could be of assistance in such situations, but we are unaware of any research into such an approach for CEM.

The unreliability of skin tone as a sole detection and/or classification technique for CEM is further evidenced by the contrasting

[6] https://news.microsoft.com/features/microsofts_photodna-protecting_children_and_businesses_in_the_cloud/.

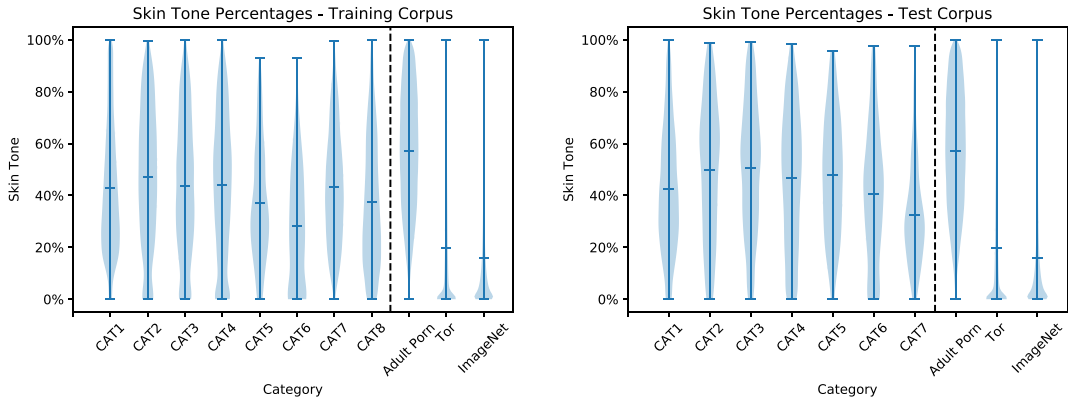

Fig. 1. Skin Tone percentage distributions (with median) calculated using the 'Uniform Daylight Illumination' algorithm by Kovac et al. (2003) per CETS category (refer Table 1) vs adult pornography, Tor and ImageNet - training and test corpora.

results between the training and test corpora. Being based on a single criminal investigation, the test corpus more readily reflects an individual offender's proclivities and perhaps even methodologies (different applications and sources perhaps reflecting their users' biases). It is entirely feasible that relative percentages of skin tone on a per-category basis will change on a case by case basis.

As an aside, the high skin-tone ratios seen within a sample taken from ImageNet (Russakovsky et al. (2015)), a widely cited dataset used extensively in image analysis, come from what appears to be nudity inherent within medical imagery. Whilst we didn't thoroughly review the entire sample, we did not observe any pornographic (i.e. sexually related) materials.

An excellent, in-depth review of colour based approaches for adult pornography detection can be found in Ries and Lienhart (2014). Vitorino et al. (2018) include a comprehensive overview on numerous pornography detection methods and products.

Application of deep learning

The application of deep learning to image classification problems has certainly gained prominence over the last few years, with the emergence of free, open source tools/SDKs such as TensorFlow, Caffe and Caffe2 making such technology publicly accessible. The development of very fast vector processing based hardware such as graphics processing units (GPUs) effectively commodified this approach.

Moustafa (2015) slightly modified and combined the AlexNet and GoogLeNet networks to create a pornographic image classifier, achieving around 94% accuracy.

Of most relevance to our approach is that taken by Vitorino et al. (2018), who created a two tiered Convolutional Neural Network (CNN) for CEM detection - the first step being the less sensitive task of general pornography detection, followed by a second, more sensitive (from the ethical and legal perspectives) step in child detection - limited by access to relevant data.

A dearth of data

As mentioned previously, research in the field of online child exploitation is largely unfeasible within academia and vast swathes of industry due to the ethical and legal implications of accessing and/or possessing such materials.

In their survey of adult pornography detection methodologies, Ries and Lienhart (2014) mention the absence of shared, publicly available databases of *adult* pornography, leading to the conclusion that individual research in most cases *"can't be quantitatively compared"*.

Grajeda et al. (2017) don't report any pornographic (lawful or otherwise) datasets within their survey of digital forensic datasets, with most image/multimedia sets gravitating towards more 'traditional' DF topics such as steganography and device (e.g. camera) forensics.

Avila et al. (2013) create and use a pornography image dataset in order to test the application of their concept detection system ('BossaNova') in pornography detection. Generated by extracting frames from pornographic and non-pornographic movies, the authors extended the corpus' research value by intentionally classifying innocuous content according to the difficulty of disambiguation with pornography. They also focus upon multi-ethnic content across genres. The authors have made this corpus freely available for research purposes (subject to a licensing agreement), and the corpus has since been used in further research by Caetano et al. (2016) and Moustafa (2015). The dataset itself was extended by Moreira et al. (2016), adding further content and overcoming a possible limitation caused by the original version's reliance upon specialised pornography websites for the 'pornographic' content.

Sae-Bae et al. (2014) were forced to use explicit adult images for training and validating elements of their automated child pornography detection system, with a limited (105 image) corpus of 'explicit-like' images of children for validating their overall performance.

Chatzis et al. (2016) identified the absence of a standard test database when researching facial features (in particular, face to iris ratio) as a means for identifying children within images. In particular, no 'ground truth' system with confirmed ages was found to be available. Instead, a collection of 75 images of publicly available images of persons with known ages was used - a sample of which indicated a strong bias towards images at least capable of portrait-style cropping (i.e. reduction to a passport-style image restricted to the subject's face from approximately directly ahead). Sensing a lack of suitable datasets, the authors of Eidinger et al. (2014) (a paper researching automated face-based age and gender estimation) assembled a corpus of approx. 26,580 age and gender labelled images of 2284 subjects. Critically, the images are "in the wild" - i.e. with unpredictable variations in conditions such as lighting, poses, and background (Wang et al. (2013)). Eight age groups are provided, but unfortunately for CEM identification purposes, one of the labelled age groups is '15–20', making it of limited use in disambiguating near-legal and legal ages (the age of 'adulthood' in terms of CEM within Australia being 18 years).

In their further work on the topic of age and gender classification, Levi and Hassncer (2015) summarise the challenges of data corpora succinctly - gathering a labelled image set of ages and genders either requires access to private information, or sufficient resources to undertake a tedious, time consuming labelling exercise. Assembling a CEM corpus represents a tedious, time consuming and psychologically harmful extension to this challenge.

Realistically, the only organisations *intentionally* gathering *labelling* and *sharing* (to whichever extent) CEM as part of their *core* 'business' are law enforcement agencies, making them an obvious point of contact and collaboration. Caetano et al. (2016) used the Pornography-2K dataset by Moreira et al. (2016) (itself an extension of the dataset produced by Avila et al. (2013)) for training an adult pornography classifier, but were reduced to indirect access to data from one hard drive from one case being conducted by the Brazilian Federal Police for training, testing and validating their CEM classifier. This limitation is entirely understandable and beyond the control of the authors, but we assert that such a tight focus runs the risk (if not guarantee) of overfitting, due largely to (a) suspect/offender procilivities and methodologies, and (b) temporal 'staleness' due to trends and fashions - not only in terms of offending, but fashions and appearances of persons and objects viewed within imagery and multimedia.

Research questions & contributions

In this paper we investigate the following questions:

1. Is it possible to train Deep Network architectures to reliably identify CEM, including distinguishing it from lawful pornography?

   (a) if 'yes', can such an architecture operate reliably across a broad range of use cases?

2. Is it possible to automatically categorise CEM according to severity, as is currently manually done by law enforcement?

   (a) If 'yes', are existing CEM schemas appropriate for such a purpose?

In this paper we demonstrate a proof of concept CEM classifier, based upon a three-module deep learning architecture. More significantly, we identify and address significant, long term challenges for law enforcement working within the machine learning field, particularly in relation to automated identification and analysis of offensive materials. Specifically, we:

- Document the measures undertaken when developing, training and validating a classifier based heavily upon offensive materials, ensuring researcher safety without compromising efficiency and efficacy;
- implement a three stage CEM classifier, designing and implementing two stages without direct trainer access;
- train one stage of the classifier to recognise largely abstract concepts, based upon pre-existing levels of severity;
- assess and measure the classifier's accuracy against materials from multiple unrelated investigations and corpora; and
- design, test and demonstrate the Majura Schema, a pornography taxonomy capable of being mapped to individual jurisdictions' individual requirements. Introduced as a means for overcoming jurisdiction-specific 'language' around CEM, it is also capable of enabling collaboration on corpora development in a traditionally isolated field.

In terms of content, justification and methodology, Vitorino et al. (2018) is closely aligned with our work conducted on this topic - the most significant commonalities being the use of:

- 'Deep learning' with a focus on *content*, rather than metadata such as hashing or filename patterns;
- Adult pornography as a larger, more readily available annotated corpus for training (with subsequent reinforcement using CEM);
- A hierarchical classification approach (though our work extends the structure to include CEM severity); and
- The use of 'real world' case data, in this instance, from the Brazilian Federal Police.

Classifier architecture

Our architecture divides the task of CEM detection into three separate modules:

Module One: is it pornography?

Automated detection of pornographic materials is a well-established commercial enterprise, with myriad products readily available for use in applications such as e-mail filtering. We therefore chose to evaluate an existing product for use in this stage.

OpenNSFW (Mahadeokar et al. (2016)) is an open-source Caffe (Jia et al. (2014)) based classifier for automatically detecting "Not Safe For Work" (NSFW) imagery, and due to its technical similarities with our intended architecture, was selected as the first candidate. A detailed discussion of the classifier's design and training is available at https://github.com/yahoo/open_nsfw. We converted the existing classifier to a tensorflow model using the Caffe to Tensorflow convertor (Dasgupta (2017)).

Fig. 2 summarises the confidences reported by the classifier across the test corpus - the strong performance in disambiguating pornography and CEM from innocent materials making it an ideal first step in culling 'not of interest' materials from the process.

Any imagery identified as pornographic (confidence score $\geqslant 0.8$, as per author advice) is passed to step two.

Module Two: are there children present?

Module two is designed for detecting children within images, including CEM. We trained a binary child detector classifier using $\frac{15}{16}$ (selected using the first character of each image's MD5 digest) of the training corpus for applicable CETS categories, adopting a VGG CNN architecture (Simonyan and Zisserman (2014)). As with the discussion by Chollet (2016), we took a VGG-16 network pre-trained on the ImageNet 1000 class dataset, removed the top stack of fully connected layers, and replaced it with a fresh (untrained) 3-layer fully connected binary classifier. The first two fully connected layers have 512 units each with ReLU activations, and the third layer has two units (one for *isChild* = *True* class, the other for

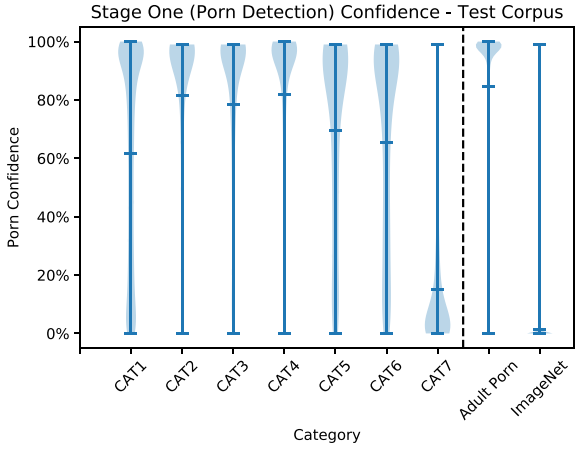

Fig. 2. OpenNSFW pornography confidences - Test Corpus, CETS (refer Table 1).

isChild = *False* class) with a softmax activation. The loss function of the classifier is binary cross-entropy, optimised on a labeled training set of images with and without children (extracted from adult/child porn corpora). A dropout with $p = 0.5$ is applied to the input of the 1st fully connected layer during training (but not during evaluation or image scoring).

Training and fine-tuning of this classifier is described in more detail in Section Training, evaluation, and scoring of images

As with step one, we chose an arbitary 'isChild' confidence score $\geqslant 0.8$, observing its performance to be adequate for our purposes. In a search/triage situation, the workflow can cease here. Otherwise, all images meeting this threshold are subjected to step three.

Module three: what CEM category?

The third module determines the CETS (refer Table 1) category, reflecting the current workflow in use during typical online child exploitation investigations. Given the theoretically unlimited range of styles and representations, CAT6 (Animated/virtual) was regarded as 'out of scope' for training and testing the classifier.

This module has a similar architecture to the previous classifier, with the only difference being in the structure of the fully connected classifier on top of the convolution layer stack. We again use a 3-layer fully connected classifier block, but the first two layers' unit counts are doubled to 1024 each in order to give the classifier more expressive power for learning and distinguishing the largely abstract concepts present across the CETS schema. The third (top) layer has six units, reflecting CETS categories 1–5 and 7 (category 6 being excluded from this experiment), and uses softmax activation. The loss function for this classifier is weighted categorical cross-entropy, allowing to compensate for class imbalance in the training set.

The module is a multi-class classifier. At this time the category with the highest score is treated as the 'winner' regardless of confidence level. An obvious, simple extension may be to recognise confusion by introducing a 'floor' confidence - if no classes cross, the image is deemed 'unclear'.

Although designed and implemented in complete isolation, it appears our design's leveraging of existing pornography detection, combined with novel classifiers around elements of CEM, loosely correlates with that of Vitorino et al. (2018). Unlike their classifier, however, we also make an attempt to disambiguate categories of CEM, with mixed success.

Training, evaluation, and scoring of images

The training of 'isChild' and CETS classifiers was done in two stages: pre-training and fine-tuning, as detailed below.

Pre-training

As detailed in Sections Module two: Are there children present? and Module three: what CEM category?, modules two and three both consist of two stacked parts: a *feature extractor* consisting of several stacked convolutional layer blocks, producing *bottleneck features*, and a *classifier* block of fully connected layers. The weights of the feature extractor are initialised to the weights of the VGG-16 CNN network, pre-trained to classify images from the ImageNet 1000 classes dataset. The weights of the classifier block are initialised randomly.

During the pre-training stage, all convolutional layers in the feature extractor are frozen, with only the fully connected classifier's weights allowed to be updated. The images fed into the model are re-scaled to 224×224 pixels with RGB channel values re-scaled by a factor $1/255$ to be within $[0, 1]$ range. The training images are then augmented via a number of random transformations such as zooming, shearing, flipping horizontal and/or vertical shifting, helping prevent overfitting by increasing variation between images[7]. The rescaled training images are fed in mini-batches of 50, augmented on the fly, and the model is pre-trained for 100 epochs using Adam optimisation with a learning rate of 10^{-3}. We use a validation set to estimate the out-of-sample loss during training, and use early stopping to prevent overfitting to the training set. A snapshot of the model is saved after every 10 epochs. Once training is complete, the snapshot with the best validation loss is kept as the final model.

Fine-tuning

After pre-training, we unfreeze the weights of the top convolutional block in the feature extractor, and fine-tune the whole model for 100 more epochs with a reduced learning rate of 10^{-4}. Again, we evaluate validation loss and save a snapshot of the model every 10 epochs, with the snapshot recording the best validation loss kept as the final model.

The validation Receiver Operating Characteristic (ROC) curves of both 'isChild' (module 2) and CETS (module 3) models after pre-training and fine-tuning steps are shown in Fig. 3. We see that pre-training already yields decent classifiers, while fine-tuning results in noticeable further improvement, especially for the binary 'isChild' classifier.

Occlusion maps

The ROC curves of the trained 'isChild' and CETS classifiers indicate good out-of-sample performance, but one needs to make sure the features learned by the classifiers are indeed useful and generalisable to previously unseen images. Indeed, we need a way to tap into what the classifiers actually learned, in order to eliminate a possibility that they learned some accidental features such as the image color palette, or some other superficial peculiarity common to both the training and validation sets.

The risk of such accidental features is significant, particularly when using corpora that can't be adequately inspected due to issues such as size, or as with CEM, sensitivity. Occlusion maps (Zeiler and Fergus (2013)) are one method for gaining an understanding of what a CNN classifier has learned to use when scoring images. These are generated by systematically obscuring (occluding) different parts of an image, observing changes in the classifier's scores. Collating these changes allows distinct areas of the image to be individually assessed for 'value' to the classifier. In this context we are using the term 'occlusion map' specifically to refer to a heat map of classification scores resulting from successively occluding parts of the image from the classifier.

Fig. 4 demonstrates an occlusion map of a benign image featuring both an adult and a child, generated using module 2 ('isChild'). Both faces are clearly visible, but high *isChild* = *True* scores (denoted by red) correspond to the area around the child's face, with the bulk of the adult's face scored not dissimilarly to the neutral background. This indicates that at least in this instance, the classifier has learned to distinguish children from adults using facial features.

Given the nature of materials being classified, we are unable to provide examples of occlusion maps for module three.

Experimental setup

Experiments listed within this paper were conducted using real-world data, including imagery sourced from contemporary AFP investigations. Safety measures (described below) were enforced as a mandatory welfare measure. Only the AFP authors hold permission to review CEM and other offensive materials.

[7] Re-scaling is conducted at time of inference/prediction, but no augmentation is carried out after the pre-training phase.

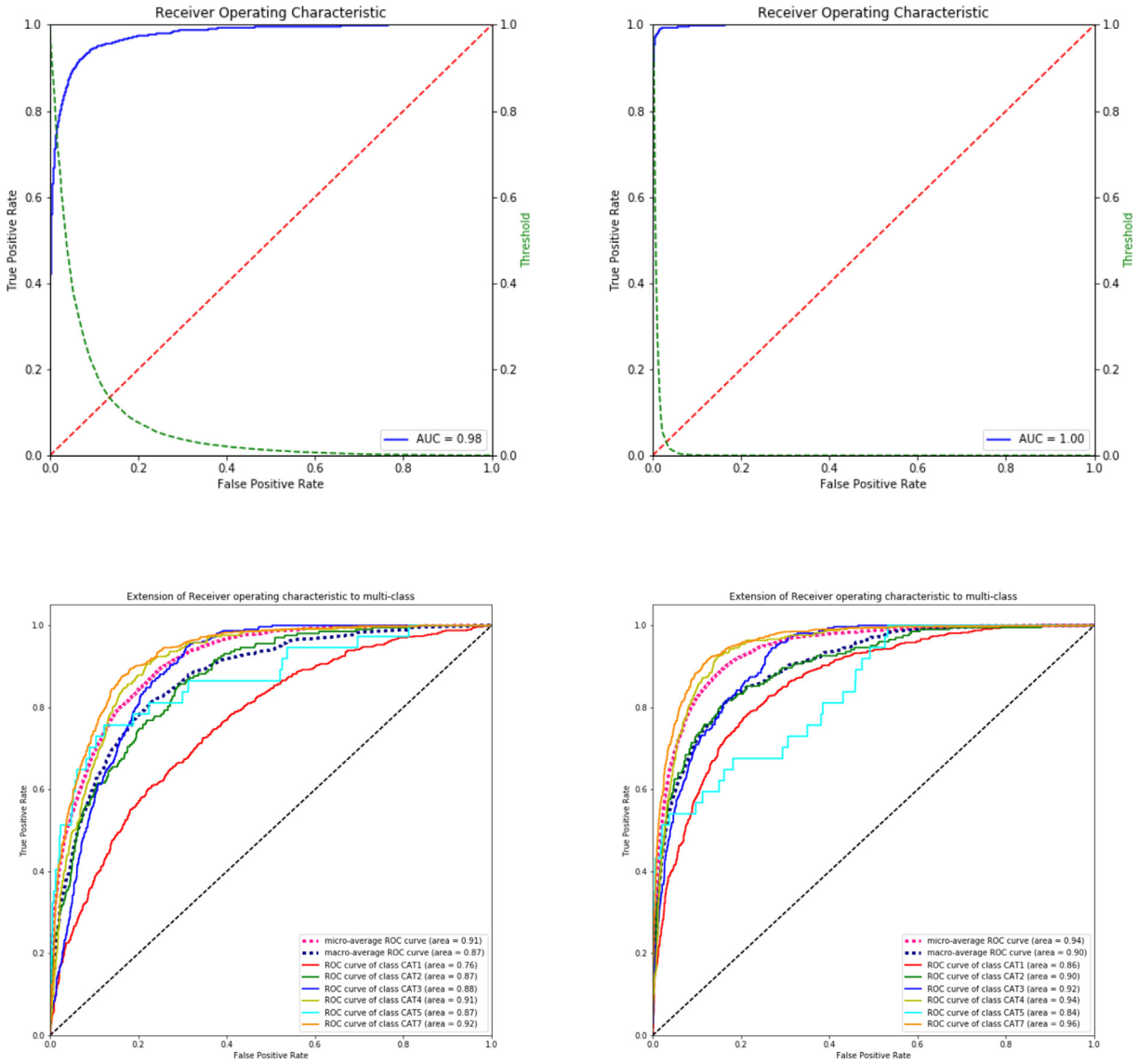

Fig. 3. Validation set Receiver Operating Curves (ROC) for binary (module 2/'isChild') (top row) and module 3/multi-class CETS (bottom row) models, after pre-training (left) and fine-tuning (right) on the corresponding training sets.

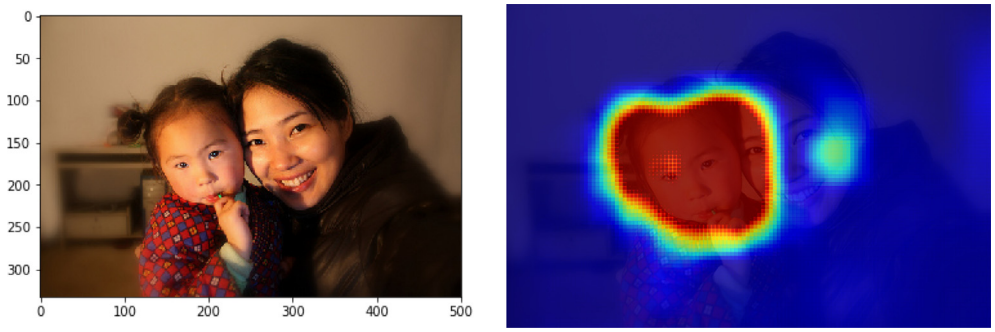

Fig. 4. Original image ($(isChild = True) = 0.980$) and corresponding occlusion map for module 3 ('isChild') classifier. The values of *isChild* = *True* scores are shown in color: red for high, blue for low. (For interpretation of the references to color/colour in this figure legend, the reader is referred to the Web version of this article.)

Ensuring safety

As evidenced by Powell et al. (2015), Powell et al. (2014) and Franqueira et al. (2017), exposure to CEM is a known, acknowledged source of stress with detrimental impacts on reviewer health. All direct interaction with images/movies used within our experiments was conducted by trained law enforcement personnel as part of their normal duties, with data made available once the relevant

cases had been processed - i.e. each image/multimedia file had been manually reviewed and annotated by a qualified investigator or AFP staff member.

The decision was made to minimise (if not completely remove) *any* possibility of inadvertent/intentional access to the source materials and underlying concepts. The following procedures were followed upon receipt of data, prior to upload to the processing server:

1. Filenames were replaced with the MD5 value of the files' contents - many files' names being sufficiently explicit and descriptive to cause concern around distress associated with textual content (refer Section Exposure to CEM); and
2. Files were encrypted at rest, with a decryption module integrated with the training/validation software.

Where unexpected results were observed, individual files were reviewed by an authorised AFP member in isolation from the remainder of the team. Feedback given was restricted to simple 'label is correct/bad', with actual content not discussed.

CEM corpora

With the exception of stage 1 (OpenNSFW classifier already giving good 'Proof of Concept' results), all training and validation of the classifier was carried out using a training corpus constructed from 13 cases held on AFP Digital Forensics systems at the time (February–March 2017) and annotated using the CETS annotation system (refer Table 1). Given the need for annotations/labelling to have been completed by investigators, this tended to correlate with items having been seized during the final quarter of 2016. Whereas cases were drawn from geographically disparate locations (a majority of data coming from the AFP Sydney, Canberra, Melbourne and Perth offices), a risk remains that some matters may have unintentional similarities due to common sources and elements (e.g. two offenders having been members of the same sharing group). This risk was mitigated by (a) the geographical spread of cases used, and (b) the removal of duplicate material during the ingestion process via the renaming of files by MD5 value (refer Section Ensuring safety). The risk of 'identical' (in terms of perception) images with differing MD5 values remains, but given the quantity of images included within the dataset, this risk was perceived as acceptably low.

The test corpus is taken from an entirely separate, fully annotated case made available to the authors approximately three months after the initial 'ingestion', containing a relatively similar distribution of CEM categories.

External/'simulated' corpora

The case used as the test corpus did not contain adult (i.e. lawful) pornography. The authors of Caetano et al. (2016) kindly provided us permission to utilise their pornography dataset, but after several reviews, we found the data to be unsuitable for our requirements - for want of a better term, the images depicted within the corpus didn't appear to be 'extreme' enough to act as a proxy for what is being encountered within typical online child exploitation investigations within Australia. As a result, a mix of relevant imagery drawn from discussion fora and commercial websites was used instead. Innocent/ignorable materials typically encountered during investigations were simulated using a subset of the ImageNet corpus (Russakovsky et al. (2015)).

An entirely separate 'Tor' corpus was generated by extracting all images gathered as part of a random walk of Tor (and linked www) sites by Dalins et al. (2018). This is used for testing detection and classification techniques on what in terms of content is a corpus skewed towards illegal and 'of interest' materials.

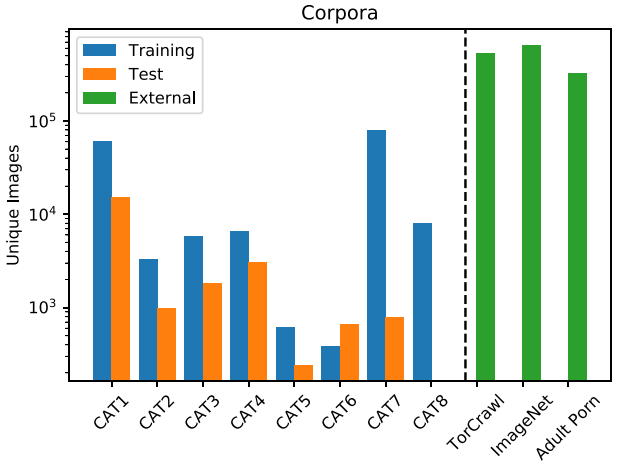

Fig. 5. Training and Test Corpora unique image count (refer Table 1 for category definitions).

Fig. 5 displays the relative counts of each corpus.

Limitations

All corpora used within this paper are based entirely upon still imagery. As with (Vitorino et al. (2018)), we see the automated classification of animated/movie materials as an obvious step for expansion. 'Movie' materials were received, and a process of extraction (based upon every *n*th frame [8]) was utilised for use within training. However, this process was aborted and data not used due to the issue of labelling accuracy on a per *frame* basis - multimedia files being classified/annotated according to the most *extreme* element within the file. Hence, if *n* percent of the movie depicts the illegal, annotated act, $100 - n\%$ of the images extracted will not. Given manual review of bulk materials was strictly out of bounds for this project, quality could therefore not be assured.

By way of example, Fig. 6 shows a typical CAT5 film, with cumulative confidence scores plotted for every second throughout runtime. In this instance, the 'correct' category dominates less than 100 s of runtime (approx $300 \Rightarrow 400$ seconds, or $< 10\%$ of total frames), otherwise fading in with the noise of indistinct categories. Such distinct sampling (typically 'per frame/per *n* seconds') is computationally slow, unreliable (particularly if the sample rate is too low) and wasteful due to the lack of context an information being passed between frames. An approach capable of maintaining knowledge between frames (such as recurrent neural networks) would be better suited for this task, but may require specialised training/test data due to the relatively distinct domain.

In either case, unfortunately what can only be described as a 'wealth' of data was available to researchers, but due to infrastructure limitations (all experiments were carried out using a single Titan GTX GPU), extensive processing times made further experimentation impracticable.

Experimental results

The sheer size of the Tor and ImageNet corpora made complete manual annotation impractical. We therefore split our classifier

[8] An attempt to use keyframes was also made, but failed due to what apppeared to be codec related inconsistencies.

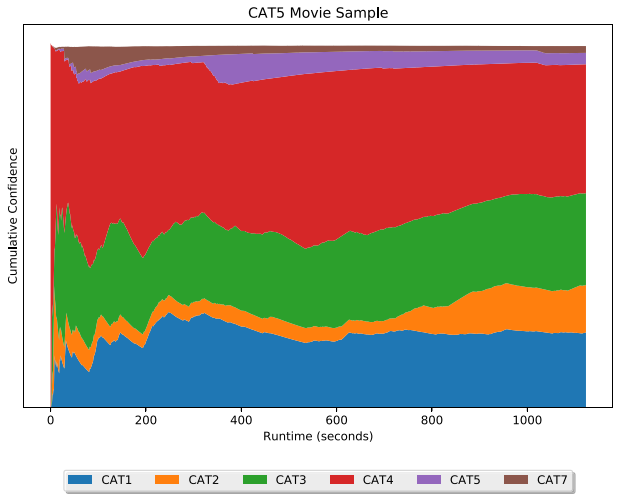

Fig. 6. CAT5 movie example classification - 1 frame per second extraction/classification.

experiments into triage scenarios (where the top results are reviewed for CEM) and a complete scenario. Namely, we classified:

1. the Tor imagery corpus for CEM content, manually reviewing the top 10 results;
2. the ImageNet corpus for CEM content, also reviewing the top 10 results; and
3. the test corpus for CEM content *and* CETS categories, providing the combined results.

Tor imagery

Table 2 shows the top ten results, together with manual review. Where an image is either difficult to view (being a small thumbnail) or taken from an angle impossible to confidently estimate participant age, it is listed as 'difficult' together with the *likely* (according to the reviewer) age.

Of the ten images, all five definite CEM images are correctly identified, with the two likely CEM images also classified accordingly. One adult pornography image is misclassified (at number 10), together with two likely adult pornography images. Of the next ten images (not shown), all were sexually explicit, with one (the thirteenth image in the entire ranked set) obvious CEM.

ImageNet

Table 3 effectively lists what can go wrong when testing classifiers - in this instance, when looking over a dataset that in all likelihood does not contain *any* material of the class (es) sought. No CEM or obviously *pornographic* material was observed throughout our relatively limited review of ImageNet's content, though explicit nudity was observed. We believe these relate to medical imagery, as most such images depicted what appeared to be skin conditions such as rashes or lesions.

Table 2
Tor Triage scenario – Top 10 images, ranked by pornography and child confidence.

Image	Porn	Child	Manual Review	Result
1	0.99	1.0	CEM	✓
2	0.99	1.0	Difficult - Likely CEM	?
3	0.99	1.0	CEM	✓
4	0.99	1.0	CEM	✓
5	0.99	1.0	CEM	✓
6	0.99	1.0	Difficult - likely Adult	?
7	0.99	1.0	Difficult - likely Adult	?
8	0.99	1.0	Difficult - Likely CEM	?
9	0.99	1.0	CEM	✓
10	0.99	1.0	Adult	×

Table 3
'Triage' scenario - Top 10 ImageNet results.

Image	Porn	Child	Manual Review	Result
1	0.99	1.0	Possible human foetus	×
2	0.99	1.0	Male genitalia (age unclear)	?
3	0.99	1.0	Arm with red sores	×
4	0.99	1.0	Maritime organism (flesh coloured)	×
5	0.99	1.0	Female genitalia (age unclear)	?
6	0.99	1.0	Birds in human hands	×
7	0.99	1.0	Human hand with rash	×
8	0.99	1.0	Frog/toad on rock (flesh coloured)	×
9	0.99	1.0	Male genitalia (likely adult)	?
10	0.99	1.0	Sores/rash on neck	×

On this test, we observed seven incorrect results, though only two we would describe as 'blatantly' wrong. The remainder included what is best defined as 'reasonable' mistakes. Fig. 7 displays examples of obviously wrong (Images 1, 6) and reasonably wrong (7, 8), respectively.

On the whole, however, the classifier worked quite well as a filter for non-CEM materials. Table 4 shows that Stage One classifies 0.12% ($\frac{800}{649.357}$) of Please check the legend of Table 1 and correct if necessary.tblimages as pornography, and of these, 71% ($\frac{568}{800}$) as containing a child/children, making a CEM false positive rate of 0.09%.

Test corpus

We ran the classifier over the AFP sourced test corpus, observing the quantities of images correctly identified as 'passing through' the three stages. Fig. 8 (detailed in Table 4) shows the relative results for each category, plus the Adult pornography and ImageNet corpora representing CAT8 and CAT9 (Ignorable), respectively.

Classifier results

In summary, we show that it is possible to train a Deep Network architectures to reliably distinguish CEM. Module one of the classifier performed adequately for triage purposes, identifying around 60%–70% of CATs 2–5 as pornography. Module two performed very strongly, in turn identifying around 80% of such files as containing children. Our selection of 0.8 as the threshold for modules one and two is validated by the results, with the ROC plots in Figs. 9 and 10 showing optimal thresholds near that value, consistent with the validation set results in Fig. 3.

From the other side, the classifier was very effective at filtering out lawful materials. The combination of modules one and two results in around 4% ($\frac{13302}{323383}$) of lawful pornographic materials being misclassified as CEM, and a negligible number of false positives (0.09%, or $\frac{568}{649357}$) arise from ImageNet's 'clean' imagery.

Categorisation of images against CETS is problematic. Outperforming random chance ($\frac{1}{6}$, given six implemented CETS classes), overall performance remains far too poor to be deemed 'acceptable' as a stand-alone reliable classifier in a legal context.

Fig. 7. Images 1, 6, 7 and 8 of the ImageNet 'top 10' (refer Table 3).

Table 4
Test Corpus Classifier Results (Percentages shown are cumulative, not per-stage).

CAT	Total	Classified as Porn	Classified as CEM	Classified as CAT	% Porn	% CEM	% CAT
CAT1	15,124	6929	5769	3098	45.8%	38.1%	20.5%
CAT2	976	719	635	199	73.7%	65.1%	20.4%
CAT3	1805	1217	1147	473	67.4%	63.5%	26.2%
CAT4	3029	2248	2149	1339	74.2%	70.9%	44.2%
CAT5	241	134	124	20	55.6%	51.5%	8.3%
CAT6	657	330	232	0	50.2%	35.3%	0.0%
CAT7	790	20	12	2	2.5%	1.5%	0.3%
Adult Porn	323,383	258,186	13,302	0	79.8%	4.1%	0.0%
ImageNet	649,357	800	568	0	0.1%	0.1%	0.0%

Table 5
Majura schema - pornography.

Pornography	
Is the image/material pornographic?	
Pornographic	Depicts nudity or other sexual concepts
Suggestive[13]	*Depicts activity alluding to or 'teasing' sexual concepts, without explicit display*
Not Pornographic	Does not depict nudity or any other sexual/adult concepts

Table 6
Majura schema - nudity.

Nudity	
What are the levels of nudity visible within this image?	
Nudity	Complete and/or partial nudity are visible. In this instance, 'nudity' is consistent with Western, corporate standards i.e. visible genitalia and/or buttocks. Visible nipples are regarded as 'nudity' when on breasts.
Suggestive	Clothing, revealing and/or posing of a suggestive nature. Examples include 'side boob', revealing cleavage, nudity with genitalia behind improvised coverage, lingerie pictures - in simple terms, NSFW.
Nil	No nudity visible.

Performance with CATs 1 and 7 in particular proved disappointing across modules one and three, with the hierarchical nature of the classifier resulting in errors propagating through all stages. Our suspected reasons for under-performance within specific categories are detailed further in Section Building an effective classification schema.

Contrastingly, we observed the classifier outperforming on CAT4 imagery, well in excess of performance elsewhere. Significantly, this is arguably the most strictly defined CETS category, requiring only the presence of sexual penetration - a clearly definable concept. We are obviously unable to provide examples, but occlusion maps (refer Section Occlusion maps) of CAT4 images showed extremely strong focus levels by the classifier in comparison with other categories.

Table 7
Majura schema - penetration.

Penetration	
Is penetration visible? Any form of penetration can be included (the nature of the item/limb performing the penetration is irrelevant)	
Oral	Oral penetration- for example: penis to mouth, sex toys, props etc.
Vaginal	Vaginal penetration: penis to vagina, cunilingus, sex toys/props
Anal	Anal penetration: penis to anus, anilingus, sex toys/props
Other	Penetration of other human/animal orifices, both 'natural' and 'manufactured' - for example, nostrils, wounds.
None	No penetration visible within image.

Table 8
Majura schema - BDSM.

BDSM	
Violent, aggressive, derogatory or otherwise physically painful/submissive behaviour for gratification.	
Bondage	The use of restraints (including weighing down of limbs) to maintain physical control of participants.
Domination	The overpowering or other control over participants, without the use of restraints. Often includes physically aggressive sexual interaction.
Sadism	The infliction of physical pain upon others for apparent sexual gratification.
Masochism	The infliction of physical pain for the recipient's apparent sexual gratification.
None	No BDSM (or similar) present

Table 9
Majura schema - props.

Props	
Are props (i.e. mechanical or inanimate items) depicted being used in a sexual or suggestive manner.	
Sex Toy	Item(s) appearing to be commercially manufactured and designed to be used in a sexual manner
Other	Items appearing to have been improvised for use as sex toys. For example: vegetables, gloves.
None	No props observed

Table 10
Majura schema - virtual.

Virtual	
Is the image/video animated, CGI or otherwise 'simulated'?	
Yes	The entire image (or the main focus) is CGI or otherwise animated. This does NOT include backgrounds (e.g. 'green screens') or cutaways.
No	The entire image (or the main focus) isn't animated/simulated.

Table 11
Majura schema - bodily fluids.

Bodily Fluids	
Are bodily fluids (e.g. blood, semen, spit, urine) visible?	
Yes- self/non interactive	Bodily fluids are visible, but are clearly not in contact with participants, or are only present on the generating person(s).
Yes - interactive	Bodily fluids are visible either present on 'receiving' participants, or clearly en route. For example, 'facials', 'money shot', 'Bukkake'
No	No bodily fluids visible within the image.

Building an effective classification schema

Note: Given the particularly offensive (if not harrowing) nature of these concepts, we have limited in-depth discussion of CETS categories.

Whilst difficult to quantify, our hypothesis is that the largely abstract nature of CETS categories greatly increases the complexity of training effective machine learning tools. We argue that CAT4, being reliant solely upon sexual penetration, is the most objective illegal CETS category. Whilst impossible to measure, we posit that the classifier's results (refer Fig. 8) loosely reflect a bell curve of objectivity - CAT4 being the most 'objective', and CATS 1 and 7 being the least. CAT7 can include individually lawful images occurring as part of series containing CEM, making it impossible to accurately categorise without broader context.

CAT1 is less broad, but can still include 'sexualised' or suggestive imagery - examples of which may appear socially acceptable when depicting adults, but offensive (if not outright illegal) with children.

Whilst not as broad as CAT7, the severity of offending inherent in CAT5 makes its rather broad remit particularly challenging. Beyond the presence of children, there really aren't any 'common' visual elements across the category.

CAT6, being focused solely on virtual/animated materials, effectively forces an overlap with other categories. Not a concern at time of CETS' inception, this has potential to become an issue as the quality and realism of some CGI renderings improves. Whilst not present in material quantities within the test corpus, the AFP authors of this article have observed some CGI CEM of a quality sufficient to be mistaken as real-world at first glance.

This move towards visual ambiguity will eventually result in CAT6 largely becoming redundant. Whereas the reason for differentiation is understandable ('real' vs 'simulated' victims), this distinction will be further muddied by the emergence of 'deepfake' materials - deep learning based software used to 'learn' a target's face and use it to replace existing actor(s) in real footage. A particular reported use is that of creating simulated celebrity pornography, such as that shown in Fig. 11.

The need for agreed standards

Franqueira et al. (2017) recognised the absence of a "recognised scale of indecency levels and a taxonomy of terms" as a challenge in the investigation of online child exploitation. We wholeheartedly agree - building training/validation/test corpora around individual jurisdictions' definitions is wasteful, with the unfortunate side-effect of actively *discouraging* collaboration. An alignment of international jursidictions' definition of child exploitation is unlikely within the foreseeable future, but an agreeable *taxonomy* seems readily achievable - we may not be able to standardise prosecution, but we can at least help standardise our tools.

Defining child exploitation imagery

A key challenge in establishing a *lingua Franca* of child exploitation is its reliance upon defining legislation. For example, Commonwealth (Australia) legislation[9] defines "child pornography material" as material(s) depicting, appearing to depict:

- person(s) under the age of 18 involved in sexual poses or activity (with or without other persons;
- person(s) under the age of 18 in the presence of a person involved in sexual poses or activity; and/or

[9] Criminal Code Act (Cth) 1995, 473.1 — Definitions.

Table 12
Majura schema: Participants.

Participants	
Describe the participants and their interactions. Select all that apply. Interactions needn't be penetrative (this is recorded in another topic).	
Male_Female	Male(s) and female(s) visibly interacting/in contact with one another.
Female_Female	Multiple females visibly interacting/in contact with one another.
Male_Male	Male(s) visibly interacting/in contact with one another.
Animal_Male	Animal(s) and male(s) visibly interacting/in contact with one another.
Animal_Female	Animal(s) and female(s) visibly interacting/in contact with one another.
Animal_Transgender	Animal(s) and transgender person(s) (visibly inconsistent genital configuration/appearance) visibly interacting/in contact with one another.
Male_Transgender	Male(s) and transgender person(s) (visibly inconsistent genital configuration/appearance) visibly interacting/in contact with one another.
Female_Transgender	Female(s) and transgender persons (visibly inconsistent genital configuration/appearance) visibly interacting/in contact with one another.
Female	Female(s) not visibly interacting with other persons.
Male	Male(s) not visibly interacting with other persons.
Transgender	Transgender person(s) not visibly interacting with other people. NOTE: Use the appearance of visibly inconsistent genital configuration/appearance as a guide.
None	No people are visible within this image.

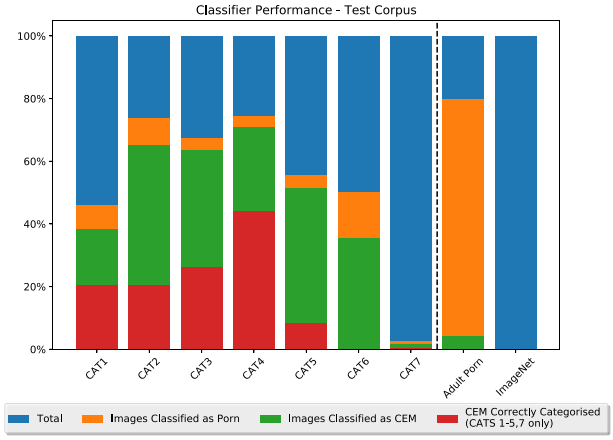

Fig. 8. Classifier performance on test corpus. Note 'CAT6' category not implemented.

- depictions/representations of sexual organs, anal region, or breasts (female only) of person(s) under the age of 18, with the *dominant* characteristic of being for a sexual purpose;

in a way that *reasonable persons would regard as being, in all circumstances, offensive.*

Such 'vagueness' is in direct response to the unpredictable nature of offender proclivities and methodologies - codifying specific behaviours runs the risk of unintentionally creating loopholes. Typically, law enforcement agencies use varying scales to quantify materials identified and their respective severity. Table 1 displays the CETS scale,

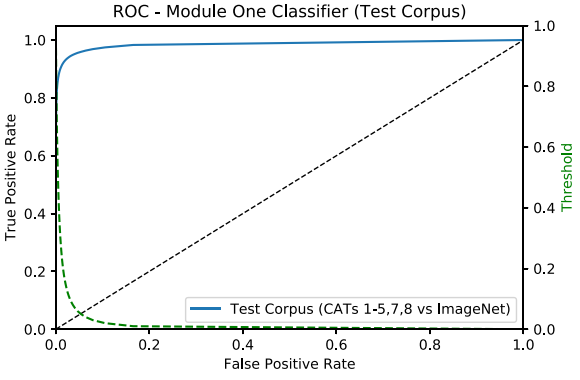

Fig. 9. Module One ROC - Pornography/sexual material detection.

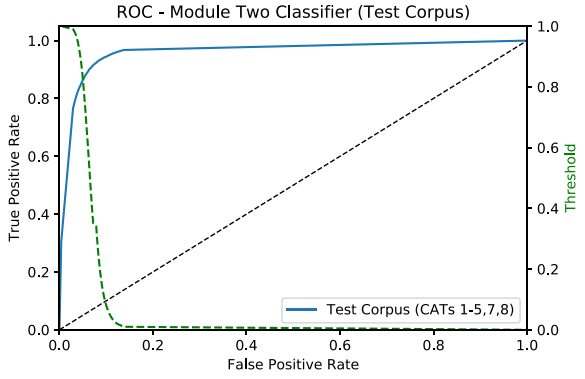

Fig. 10. Module Two ROC - CEM and pornography disambiguation.

Fig. 11. Still image taken from 'Deepfake' Katy Perry pornography video (Quach (2018)). *Image redacted by authors for ethical reasons.*

as used by the AFP in online child exploitation investigations - of note, most categories are quite broad in terms of activities *capable* of being depicted, with the exception of category 4 - penetration being narrow, definitive concept when compared with, say, 'suggestive' posing.

We observed the classifier's performance to largely reflect this fact. Occlusion maps of category 4 images indicated a heavy (if not complete) focus upon regions depicting sexual penetration, whilst other categories would perhaps score genitalia, breasts (or lack thereof) as a relevant characteristic.

Building a taxonomy

A key issue in developing the classifier was the generation of an adequately sized, representative corpus of materials. Considering the only difference between 'adult' and 'child' pornography is participant

Fig. 12. Male, nude. Is it pornography?.

age, we made the conscious decision to build our taxonomy around adult pornography, both for safety and practicality purposes - after all, there is a *lot* of adult pornography available online.

A digital forensic practitioner was tasked with downloading several thousand adult pornography images which he deemed approximately representative of what is typically encountered within investigations. CEM was not used at this stage, due to the inclusion of non law-enforcement reviewers - adult pornography with 'similar' posing, scenarios etc was used instead.

Four reviewers (three law enforcement, one academia) then assembled and reviewed a random selection of several hundred of the aforementioned images. A 'roundtable' was then conducted about each image, with attributes deemed relevant for law enforcement recorded and subsequently arranged into broad categories.

Of particular concern, the schema needed to meet three separate requirements:

1. The ability to be mapped into established CEM scales such as CETS (refer Table 1);
2. Simplicity to a level allowing reliable use without *reasonable* conflicts between labellers' annotations (i.e. different answers both being 'right'); and
3. A capability to record visually disparate participant attributes, such as race, ethnicity and gender.

We emphasise that capability 3 is a quality assurance measure - recording attributes such as gender, race and ethnicity is not for direct use by any subsequent classifiers, but rather as a quality-assurance measure to help avoid inadvertent 'whitewashing'[10] or the gender equivalent. The authors of this paper and the AFP view innate characteristics such as gender and race as irrelevant for establishing criminality, or severity of criminality, and would refuse to provide cooperation for any parties intending to use these attributes for such a purpose.

The classification of race/ancestry proved challenging, particularly when considering the range of possible options. Whilst organisations such as the US census bureau record five possible races (United States Census Bureau (2018))[11], the Australian Bureau of Statistics takes a more granular approach in their census preparations, with 320 possible responses for 'ancestry' (Australian Bureau of Statistics of Statistics (2016)) ('race' not being recorded *per se*). Obviously, such delineations are made more in response to statistical need rather than visual representation, and each approach presents dangers in over-simplification and unnecessary complexity, respectively. In the end, unreliability of recorded data (refer Section Testing the schema) resulted in race/ethnicity being dropped from the schema.

After numerous iterations, version 1.0 of the Majura[12] schema was agreed, with provisions for recording pornography, nudity, penetration, BDSM, props, virtual/animation, bodily fluids, and participants (See Tables 5–12).

Creating a test corpus

We assembled a corpus of several hundred thousand lawful[14] pornographic images, crawling numerous free pornography websites and pornography-themed discussion boards (the latter being a particularly rich source of 'unconventional' but lawful pornography).

An initial crawl of several sites was observed to be biased towards males' tastes (both heterosexual and homosexual). Two female volunteers subsequently gathered materials they deemed representative of their tastes, and these were integrated with the remaining corpus.

Testing the schema

Six labellers (three male, three female) were asked to annotate a selection of 49 images taken from the aforementioned corpus. The selection represented a broad spectrum of lawfully available materials, including 'traditional' hard-core and soft-core, bestiality, 'extreme' pornography, parody materials and innocent/non-sexual imagery. The images were printed in a large thumbnail format next to a table detailing the annotation schema, with reviewers invited to circle the attributes relevant to each. Some inconsistencies were observed regarding perceptions of 'pornography', but significantly, recordings of race (refer Section Race/Ancestry) were found to be extremely inconsistent across labellers.

The labelling schema (with aforementioned changes) was then ported to a browser based application, and fourteen digital forensic practitioners were invited to annotate a selection from the full adult pornography corpus. The labellers were allowed to work within the same office and discuss images (if required), but the actual images shown to each person were randomised to avoid collaboration and 'groupthink'. 3438 unique images were annotated - 3420 by individual users, 15 by two users, and one by three.

In particular, several inadvertently vague and/or difficult questions became readily apparent:

Race/ancestry

As mentioned previously, the schema also included the option to record race/ancestry - not for direct use by any subsequent classifiers, but rather as a quality-assurance measure to counter 'whitewashing'. This was dropped relatively early in the process, due to labellers'

[10] Inherent bias/recognition of Caucasian participants, most likely due to skewed datasets and tuning/development.

[11] White, Black or African American, American Indian or Alaska Native, Asian, Native Hawaiian or Other Pacific Islander.

[12] Named after the AFP's National Forensics facility.

[13] Added after initial testing.

[14] Within the Australian Capital Territory.

reported difficulties identifying such characteristics, particularly in participants only partially depicted in low-quality imagery.

Pornography

An immediate issue was identified in the identification of 'pornography' - the definition thereof extremely difficult to objectively and consistently apply. We observed a cluster of images consistent with sexuality or suggestiveness, but not (in the labellers' opinion) constituting pornography - of the 49 images originally assessed, 14 involved labeller disagreement - 5 with 5 vs 1, 9 with 4 vs 2 split of viewer opinion. For example, Fig. 12 is certainly suggestive, but the absence of visible genitalia or sexual posing led to strong disagreement over its interpretation as 'pornography', particularly as the person shown is an adult male. We originally considered broadening the definition of 'pornography' to include NSFW [15], but reviewer feedback indicated this would swing too far towards *inclusion* of borderline materials.

We instead added a 'suggestive' attribute within the pornography section, akin to the 'Racy' attribute offered by MS Azure Computer Vision API[16]. *Prima facie* this has provided annotators with a more comfortable middle ground, with the happy side-effect of allowing *context* to be introduced into otherwise abstract concepts - 'racy' involving adults being of little interest to law enforcement, but of great concern when children are involved.

'Suggestive' nudity remains in the schema for this version - on the whole, labeller feedback indicated a preference for the additional granularity.

'Virtual' imagery

Probably the most unanticipated 'inconsistency' with the schema's results concerned animated and/or virtual imagery. During the second, more thorough labelling session, users reported difficulty in declaring an image 'virtual' - particularly in cases where images contained (for example) captions, and in one unanticipated series of images, 'thought' and 'speech' bubbles. Animated/virtual characters were observed interacting with 'real world' participants, in a manner not dissimilar to 'augmented reality' scenarios.

Originally, the schema demanded an 'all or nothing' approach to animation, with an image assumed to either be animated, or not. This was unable to accurately record such 'half–half' images. As a result, the schema was updated to require the main focus (as defined by the viewer) to be animated/CG for the image to be labelled 'virtual'.

'Negative' labels

The schema includes 'negative' labels for all categories, in an attempt to ensure disambiguation between items not present vs. questions not asked/answers not recorded. Strong feedback was received from users regarding these labels, with their recommendations tending to be either:

1. Remove the negative, due to confusion and unnecessary labeller workload; or
2. The relevant question's default result is 'None/No' - i.e., if the question is asked and the user *doesn't* select any options.

We regard this is a reasonable request, but should be regarded as an implementation requirement rather than an integral design feature.

[15] Not Safe For Work - best defined as if the viewer (and colleagues) would feel comfortable observing the imagery within a professional work environment.

[16] https://azure.microsoft.com/en_us/services/cognitive_services/computer_vision/.

Practitioner feedback regarding the schema's simplicity and completeness was positive, and the data could be comfortably annotated.

Conclusions

In the first half of this paper, we demonstrate the limitations of current research into automated detection of CEM. Long considered unreliable (with a tendency to under-report 'extreme' categories (Vitorino et al. (2018))), we quantify skin tone detection/measurement's limitations in disambiguating CEM categories, CEM from adult pornography, but also the high degrees of variability between users (most likely reflecting individual tastes).

We introduce a three-stage classifier trained and validated on data from multiple, isolated, 'real world' criminal cases, and report its performance on multiple, thematically distinct corpora including a completely separate case, Tor imagery, ImageNet and adult pornography. We observed performance adequate for triage purposes across all test corpora, particularly in CEM detection. We did, however, observe poorer performance than we had observed during training and validation - a typical characteristic of overfitting.

The three-stage architecture introduces cascading errors, largely due to the dropping of false negatives from processing at each stage. This makes the first stage (pornography detection) particularly important, and whilst OpenNSFW worked extremely well considering its 'off the shelf' nature, we identified room for improvement - specifically around 'extreme' pornography.

Critically, we observed limitations of the CETS scale's suitability for machine learning, resulting in a corresponding under-performance in categorisation by module 3. We believe this is due largely to the abstract, context-heavy nature of most categories. Category 7 ('indicative' materials) is probably impossible to implement *without* providing a classifier access to co-located materials and relevant case data, both of which our currently 'out of scope' for this research.

We introduce and test the Majura schema, a pornography/CEM annotation schema specifically designed to support collaborative development of ML tools and techniques within a traditionally under-researched area. By providing a jurisdictionally independent 'lingua Franca' for annotation, we provide a convenient means for researchers and law enforcement to share 'prior work'.

This research represents the beginnings of long-term foundations for improved data exploitation and information retrieval within law enforcement. We regard the Majura schema as a living document, with revisions anticipated to cope with scenarios not anticipated at time of writing. Our work in this field is ongoing, and we welcome comment and encourage potential collaboration.

Acknowledgements

This research was conducted as part of Project **Stonefish**, an Australian Federal Police initiative for law enforcement, academic and commercial cooperation in the field of automated offensive material detection.

The authors wish to thank The Honourable Michael Keenan, the then (Commonwealth) Minister for Justice during the conduct of these experiments and the (Australian) Attorney-General's Department, Cybercrime Division for their endorsement and/or approval of this research.

Design and implementation of the classifier detailed within this paper was conducted in 2017 as part of a six month APS 'Data Champions' Data Fellowship undertaken by author Dalins with Data61, administered by the Department of Prime Minister and Cabinet (now managed by the Data Transformation Agency).

The authors also acknowledge and thank the Collier Foundation for their financial support in the provision of storage and computing equipment used to host elements of these experiments.

Finally, the authors also thank the anonymous reviewers of this article for their extensive comments and suggestions relating not only to this work, but also wider issues within the field.

Future work

A short-term priority for this work is the generation of an adult pornography dataset for training classifiers based upon the Majura schema, with a second 'child' identifier used to disambiguate CEM from adult pornography. This will form part of a long-term initiative to improve automated detection methodologies without unreasonable exposure to CEM and other offensive materials.

The AFP and Monash University are currently exploring options around the development of a data 'airlock', a storage and processing solution allowing indirect access to CEM and other offensive data for the purposes of developing and testing classifiers and other automated detection tools. By giving indirect access to the data (with processing hosted internally), we aim to enable and improve research by helping researchers avoid ethical and legal restrictions typically associated with this field.

References

Avila, S., Thome, N., Cord, M., Valle, E., Araújo, A. de A., 2013. Pooling in image representation: the visual codeword point of view. Comput. Vis. Image Understand. 117 (5), 453–465.

Brown, J., Fielding, J., Grover, J., 1999. Distinguishing Traumatic, Vicarious and Routine Operational Stressor Exposure and Attendant Adverse Consequences in a Sample of Police Officers, vol. 13.

United States Census Bureau, January 2018. Race. https://www.census.gov/topics/population/race/about.html.

Caetano, C., Avila, S., Schwartz, W.R., Guimarães, S.J.F., Araújo, de A., 2016. A mid-level video representation based on binary descriptors: a case study for pornography detection. Neurocomputing 213, 102–114 (binary Representation Learning in Computer Vision).

Chatzis, V., Panagiotopoulos, F., Mardiris, V., July 2016. Face to iris area ratio as a feature for children detection in digital forensics applications. In: 2016 Digital Media Industry Academic Forum (DMIAF), pp. 121–124.

Chollet, F., June 2016. Building Powerful Image Classification Models Using Very Little Data. https://blog.keras.io/building-powerful-image-classification-models-using-very-little-data.html.

Dalins, J., Wilson, C., Carman, M., March 2018. Criminal motivation on the dark web: a categorisation model for law enforcement. Digit. Invest. 24.

Dasgupta, S., January 2017. Caffe to Tensorflow. https://github.com/ethereon/caffe-tensorflow.

Edelmann, R.J., 2010. Exposure to child abuse images as part of one's work: possible psychological implications. J. Forensic Psychiatr. Psychol. 21 (4), 481–489.

Eidinger, E., Enbar, R., Hassner, T., Dec 2014. Age and gender estimation of unfiltered faces. IEEE Trans. Inf. Forensics Secur. 9 (12), 2170–2179.

Franqueira, V.N., Bryce, J., Mutawa, N.A., Marrington, A., March 2018. Investigation of indecent images of children cases: challenges and suggestions collected from the trenches. Dig. Invest. 24.

Grajeda, C., Breitinger, F., Baggili, I., 2017. Availability of datasets for digital forensics – and what is missing. Digit. Invest. 22, S94–S105.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T., 2014. Caffe: Convolutional Architecture for Fast Feature Embedding arXiv preprint arXiv:1408.5093.

Kovac, J., Peer, P., Solina, F., 2003. Human skin color clustering for face detection. In: The IEEE Region 8 EUROCON 2003. Computer as a Tool, vol. 2, pp. 144–148.

Latapy, M., Magnien, C., Fournier, R., 2013. Quantifying paedophile activity in a large p2p system. Inf. Process. Manag. 49 (1), 248–263.

Levi, G., Hassncer, T., June 2015. Age and gender classification using convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 34–42.

Mahadeokar, J., Farfade, S., Kamat, A.R., Kappeler, A., October 2016. Open Nsfw Model. https://github.com/yahoo/open_nsfw.

Moreira, D., Avila, S., Perez, M., Moraes, D., Testoni, V., Valle, E., Goldenstein, S., Rocha, A., 2016. Pornography classification: the hidden clues in video space–time. Forensic Sci. Int. 268, 46–61.

Moustafa, M., 2015. Applying Deep Learning to Classify Pornographic Images and Videos. CoRR abs/1511.08899.

Australian Bureau of Statistics of Statistics, August 2016. Ancestry 1st Response/ancestry 2nd Response/ancestry Multi Response (2901.0-census Dictionary, 2011). http://www.abs.gov.au/ausstats/abs@.nsf/Lookup/2901.0Chapter602011.

Panchenko, A., Beaufort, R., Fairon, C., 2012. Detection of child sexual abuse media on p2p networks: normalization and classification of associated filenames. In: Proceedings of the LREC Workshop on Language Resources for Public Security Applications.

Peersman, C., Schulze, C., Rashid, A., Brennan, M., Fischer, C., 2016. icop: live forensics to reveal previously unknown criminal media on p2p networks. Digit. Invest. 18, 50–64.

Powell, M., Cassematis, P., Benson, M., Smallbone, S., Wortley, R., Jun 2015. Police officers' perceptions of their reactions to viewing internet child exploitation material. J. Police Crim. Psychol. 30 (2), 103–111.

Powell, M.B., Cassematis, P., Benson, M.S., Smallbone, S., Wortley, R., 2014. Police officers' perceptions of the challenges involved in internet child exploitation investigation. Policing: Int. J. 37 (3), 543–557.

Quach, K., January 2018. Fyi: There's Now an Ai App that Generates Convincing Fake Smut Vids Using Celebs' Faces. https://www.theregister.co.uk/2018/01/25/ai_fake_skin_flicks/.

Ries, C.X., Lienhart, R., Apr 2014. A survey on visual adult image recognition. Multimed. Tool. Appl. 69 (3), 661–688.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L., 2015. ImageNet large scale visual recognition challenge. Int. J. Comput. Vis. 115 (3), 211–252.

Sae-Bae, N., Sun, X., Sencar, H.T., Memon, N.D., Oct 2014. Towards automatic detection of child pornography. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 5332–5336.

Seigfried-Spellar, K.C., Dec 2017. Assessing the psychological well-being and coping mechanisms of law enforcement investigators vs. digital forensic examiners of child pornography investigations. J. Police Crim. Psychol.

Simonyan, K., Zisserman, A., 2014. Very Deep Convolutional Networks for Large-scale Image Recognition. CoRR abs/1409.1556.

Steel, C.M., 2009. Child pornography in peer-to-peer networks. Child Abuse Neglect 33 (8), 560–568.

Violanti, J.M., Aron, F., 1995. Police stressors: variations in perception among police personnel. J. Crim. Justice 23 (3), 287–294.

Vitorino, P., Avila, S., Perez, M., Rocha, A., 2018. Leveraging deep neural networks to fight child pornography in the age of social media. J. Vis. Commun. Image Represent. 50, 303–313.

Wang, H., Kang, B., Kim, D., Nov 2013. Pfw: a face database in the wild for studying face identification and verification in uncontrolled environment. In: 2013 2nd IAPR Asian Conference on Pattern Recognition, pp. 356–360.

Zeiler, M.D., Fergus, R., 2013. Visualizing and Understanding Convolutional Networks. CoRR abs/1311.2901.