CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	iCOP: Live forensics to reveal previously unknown criminal media on P2P networks
Authors	Peersman, Claudia;Schulze, Christian;Rashid, Awais;Brennan, Margaret;Fischer, Carl
Publication date	2016-07-16
Original Citation	Peersman, C., Schulze, C., Rashid, A., Brennan, M. and Fischer, C. (2016) 'iCOP: Live forensics to reveal previously unknown criminal media on P2P networks', Digital Investigation, 18, pp. 50-64. (15pp.) DOI: 10.1016/j.diin.2016.07.002
Type of publication	Article (peer-reviewed)
Link to publisher's version	https://www.sciencedirect.com/science/article/pii/S1742287616300779?via%3Dihub - 10.1016/j.diin.2016.07.002
Rights	©2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CCBY license (http://creativecommons.org/licenses/by/4.0/) - http://creativecommons.org/licenses/by/4.0/
Download date	2026-11-03 09:11:16
Item downloaded from	https://hdl.handle.net/10468/8886

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Digital Investigation 18 (2016) 50–64

Contents lists available at ScienceDirect

Digital Investigation

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

iCOP: Live forensics to reveal previously unknown criminal media on P2P networks

Claudia Peersman [a, *], Christian Schulze [b], Awais Rashid [a], Margaret Brennan [c], Carl Fischer [a]

[a] Security Lancaster Research Centre, Lancaster University, Lancaster, UK
[b] German Research Center for Artificial Intelligence (DFKI), Kaiserslautern, Germany
[c] School of Applied Psychology, University College Cork, Cork, Ireland

ARTICLE INFO

Article history:
Received 15 November 2015
Received in revised form 6 June 2016
Accepted 13 July 2016
Available online 16 July 2016

Keywords:
Computer crime
Peer-to-peer computing
Image classification
Text analysis
Forensic triage

ABSTRACT

The increasing levels of criminal media being shared in peer-to-peer (P2P) networks pose a significant challenge to law enforcement agencies. One of the main priorities for P2P investigators is to identify cases where a user is actively engaged in the production of child sexual abuse (CSA) media – they can be indicators of recent or on-going child abuse. Although a number of P2P monitoring tools exist to detect paedophile activity in such networks, they typically rely on hash value databases of known CSA media. As a result, these tools are not able to adequately triage the thousands of results they retrieve, nor can they identify new child abuse media that are being released on to a network. In this paper, we present a new intelligent forensics approach that incorporates the advantages of artificial intelligence and machine learning theory to automatically flag new/previously unseen CSA media to investigators. Additionally, the research was extensively discussed with law enforcement cybercrime specialists from different European countries and Interpol. The approach has been implemented into the iCOP toolkit, a software package that is designed to perform live forensic analysis on a P2P network environment. In addition, the system offers secondary features, such as showing on-line sharers of known CSA files and the ability to see other files shared by the same GUID or other IP addresses used by the same P2P client. Finally, our evaluation on real CSA case data shows high degrees of accuracy, while hands-on trials with law enforcement officers demonstrate the toolkit's complementarity to extant investigative workflows.
© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

Introduction

The proliferation of the Internet and peer-to-peer (P2P) file sharing systems has transformed the distribution of child sexual abuse media (CSA) into a crime without geographical boundaries. While there is scientific debate (Middleton, 2009) on whether the on-line child sex offender is a new type of offender (Quayle et al., 2000) or if those with a pre-disposition to offend are responding to the opportunities afforded by the new forms of social media (Cooper, 1998), empirical evidence points to the problem of Internet-based paedophilia as endemic. The scale of CSA media trafficking in P2P networks has been investigated by a number of studies involving a timespan of several days (e.g. (Hughes et al., 2006, 2008)), weeks (e.g. (Latapy et al., 2013)), months (e.g. (Prichard et al., 2011; Menasche et al., 2009; Stutzbach and Rejaie, 2006; Gummadi et al., 2003)) or even an entire year ((Hurley et al., 2013)). They all showed that, worldwide, hundreds of searches for child abuse images occur each second, resulting in hundreds of

* Corresponding author.
E-mail address: claudiapeersman@hotmail.com (C. Peersman).

http://dx.doi.org/10.1016/j.diin.2016.07.002
1742-2876/© 2016 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

thousands of CSA media being shared each year. Moreover (Wolak et al., 2005), found that 16% of CSA media possessors had committed one or more *contact offences*, i.e. they had directly and physically abused children.

The severity of the problem has already resulted in a number of solutions that can monitor such activity. The Child Protection System (CPS) (Child Protection System) and RoundUp ((Liberatore et al., 2010a, 2010b)) are able to capture data about paedophile activity and identify child abuse media across different P2P protocols. However, these tools rely on matching the files shared on a network against a hash-value database of known CSA media.[1] As a result, they retrieve thousands of files that have been circulating for several months or even years, but they are not able to identify new child abuse media when they are being released on to the network. Nor are they able to detect CSA media that are not on record, that have been altered or embedded into other regular image or video files. In the first part of this paper, we discuss in-depth, semi-structured interviews with a key informant sample of cybercrime investigators from different European countries that highlight the fundamental need for the development of new approaches that support *victim-centric* investigative practices. More specifically, we show that in P2P policing contexts, the distributors of new/previously unknown CSA media are considered high-risk targets, because they could be linked to recent or even on-going child abuse.

To combat the current and future challenges of cybercrime, the authors of (Irons and Lallie, 2014) already argued that there is a need to re-examine standard digital forensic procedures and the use of investigative technology by incorporating *intelligent* technologies, i.e. techniques from artificial intelligence, computational modelling and/or social network analysis. In the second part of this paper, we show that such intelligent techniques can be used to focus P2P network investigations pertaining to child sexual abuse and reduce the amount of time spent looking for digital evidence. More specifically, we present a novel approach that incorporates the advantages of artificial intelligence (AI) and machine learning procedures to automatically identify new/previously unseen CSA media to investigators. The key contributions of our work are as follows:

- We argue that detecting new/previously unknown CSA media requires (semi-)automatic analysis of image and video content. This study presents a new image and video classification module using multiple and – in case of video – multi-modal (visual and audio) feature descriptions, leading to a robust and highly accurate identification of child abuse content.
- Downloading of all candidate files in support of the image and video analysis component we mentioned above is clearly infeasible in a P2P scenario. Hence, such an approach also requires an intermediate step to reduce the number of candidate files to be downloaded for image analysis. In this paper, we present an intelligent solution to this challenge, which adopts automatic text categorisation techniques to determine the likelihood that a candidate file contains CSA content based on its filename. An evaluation on verified CSA media demonstrates the efficiency of the approach when automatically detecting potential candidates for new CSA media out of thousands of non-CSA files that are continuously being shared in P2P networks.
- We propose a triage approach to flag the most pertinent candidates for new/previously unknown child abuse media based on a synthesis of the above two modules.
- We describe the results of evaluating our approach on real CSA filenames and features of CSA media, which show high degrees of accuracy.

The approach has been implemented into the iCOP toolkit, a software package that is designed to perform live forensic analysis on a P2P network environment. Finally, a user evaluation by law enforcement officers highlights the iCOP toolkit's potential to complement and enhance extant investigative workflows pertaining to CSA media.

The rest of this paper is structured as follows. In the next section, we provide an overview of the in-depth interviews, together with the survey of law enforcement user requirements. Next, we discuss background material and related work in Sections Background and Related work. Section Approach describes the key components of our approach, that is, the filename categorisation module and the media classification module. Our approach is evaluated on CSA data in Section Experiments and results. Section The iCOP toolkit outlines the architecture of the iCOP toolkit and discusses how the two analyses are synthesised to perform live triage on P2P data and detect new/previously unknown CSA media. In Section User evaluation we provide insights from our user trial. Finally, Section Conclusion concludes the paper, discusses the limitations of our approach and identifies directions for future research.

Challenges for live forensic analysis of P2P networks

Notwithstanding increased investment in the development of interventions to combat the exchange of CSA media within P2P systems, significant challenges persist for those charged with the policing of these crimes (e.g. (U.S. Department of Justice (2010); Wolak et al., 2012)). In order to understand these constraints, and to identify where capacity for technical solutions to investigative practitioners' problems might exist, we commenced our study with a review. More specifically, our primary objective was to extend the sparse and rather limited knowledge base on the strategies employed by law enforcement in the policing and interception of CSA media exchanges in P2P networks, offence characteristics, investigative objectives, and attendant challenges and requirements in investigative settings. Given the dearth of empirical information in these domains, as well as the inaccessibility of P2P policing and offending processes to the general public, it became evident that this analysis required extensive consultation with specialist P2P investigators given their unique status as "gatekeepers" to these hidden subcultures. Therefore, we conducted both a series of in-depth, semi-structured

[1] Such databases are built through post-hoc forensic analysis of seized computers of offenders.

interviews with a key informant sample of law enforcement investigators expert in the investigation of CSA media exchanges on P2P networks and a broader, web-based survey of domain investigators.

The consultations were carried out between October 2011 and January 2012. Survey and interview respondents were sourced via two channels; (1) the iCOP research teams contact network, as established in preceding P2P-related research projects, and (2) the iCOP law enforcement advisory group. In the first phase of the consultation, twenty-four P2P investigators completed a confidential on-line survey of law enforcement users. Invitations to participate in the survey were issued via email to P2P investigators from child abuse investigation and cybercrime investigation units. The survey contained 21 questions related to the focus of P2P investigations, characteristics of CSA media file sharing behaviour and the data sources employed in typical investigations. The majority of the survey questions were multiple-choice, with the possibility to enter free-form replies in response to four items. Law enforcement representatives from twelve countries participated in the survey, including Germany, the Netherlands, the United Kingdom, Slovenia, Sweden, Ireland, Denmark, Romania, France, Spain, New Zealand and Canada, as well as experts from Interpol.

Upon completion of the survey, ten in-depth key informant interviews were conducted with law enforcement experts from four countries (Australia, Sweden, the Netherlands and the United Kingdom). The interview and survey samples were developed independently, albeit with some overlap between participants (five investigators participated in both consultation activities). The sample frame for the interviews comprised leading law enforcement investigators with expertise in the policing of CSA media offences on P2P. From this sample frame, a number of prospective interviewees were purposively sampled in accordance Tremblays criteria for the selection of key informants (Tremblay, 2003). Prospective interview candidates were first identified as established authorities in P2P network investigation with extensive supporting expertise in the use of P2P monitoring software – in so far as possible the research team sought validation of this status and a personal recommendation from a member of the iCOP law enforcement advisory group. Once identified as key informants, candidates were invited to participate in an interview. The sample was stratified to ensure that an international cross-section of practitioners from countries within and outside of the EU were included in the research. Given that the consultation process was limited to ten candidates, quotas were applied (min. one, max. three candidates per country) to ensure that a broad range of countries was represented in the interview process. Twelve email requests for participation in the research were issued to expert candidates. In all cases, this invitation advised prospective participants of the aims and objectives of the study and key data handling and security provisions to be made by the research team. Of this number, two candidates declined due to competing operational commitments and ten accepted the invitation to participate. The resulting interview transcripts were analysed using combined Thematic Analysis (Braun and Clarke, 2006) and Correspondence Analysis (Benzecri and Bellier, 1973) strategies. The following findings synopsise the operational, legal and technical challenges to live forensic analysis of P2P networks identified in these consultations.

A primary operational challenge identified by our respondents was the actualisation of reliable strategies for identifying child sexual abuse cases (and its offenders) during live forensic analyses of P2P networks. While investigative strategies such as victim identification (e.g. (Interpol, 2014)) maintain child protection and the primacy of the child victim as a central operational concern, in many cases the legal mandate to apprehend offenders has persisted as a principal focus for law enforcement (Taylor and Quayle, 2003). Our respondents' descriptions of live investigations of P2P networks reflected these conflicting imperatives – their accounts were characterised by a tension between the adoption of strategies that supported offender apprehension and those that prioritised the identification and welfare of child victims. However, each P2P investigator in our interview sample clearly expressed locating and safeguarding child victims as a fundamental investigative objective. In this regard, the identification of victims and perpetrators of contact sexual offences was cited by all respondents as the predominant concern for law enforcement investigators, even superordinate to the policing of CSA media offences:

> "If we were to identify, or we believed an on-line identity was committing the abuse of a child, that would be priority number one. It doesn't matter if we think or we can show that they have child pornography, that is all secondary, the reality is that child pornography is secondary to an actual live victim." (Australian Police, Respondent 1). "We need to keep pushing [investigators] toward that person who is abusing children and not just the collector – as bad as that is as well – you want to rescue children." (UK Police, Respondent 1).

Although our respondents' operational motivations and objectives were profoundly and formatively influenced by this victim-centric ethos, several practical challenges to the identification of victims of child sexual abuse (and perpetrators) were reported. For example, the investigators described a parallel legal mandate to enforce the law with regard to broader offences of possession and distribution of CSA media, in which it was often impossible to establish a link between these offences and the actual sexual abuse of a child. Consequently, in some cases investigators stated to afford primacy to legal rather than victim-centric imperatives, selecting a target for investigation on the basis of the duration of the sentence their CSA media-related offending would incur (i.e. based on the number or type of files they made available for distribution), rather than to attempt to identify on-going sexual abuse. Evidently, such approaches are inconsistent with de facto, victim-centric policing strategies that prioritise the identification and welfare of victims of child sexual abuse. However, while state-of-the-art features of P2P monitors support the apprehension of prolific downloaders and distributors of CSA media, our respondents reported that these could not be usefully adapted to victim-centric investigations, because they typically provide little data that readily supports

alternative investigative actions, such as the identification of victims of child sexual abuse in live forensic analysis contexts. The level of automation currently afforded by P2P monitoring solutions has entrenched this prevailing emphasis on offender apprehension in the sense that many investigators, operating in resource-constrained investigative environments, have become reliant upon automated approaches that support large-scale offence detection. In these application scenarios, P2P investigators are required to expend little cognitive effort in the identification of new cases. Rather, they may simply select cases for follow-up from a series of offending targets identified by the monitor, a process a contributor variously described as "shooting fish in a barrel" (UK Police, Respondent 2).

Clearly, in naturalistic decision making environments such as those of P2P investigations, which are characterised by time-constraints, high cognitive load and excessive case load, a scalable tool that automates the detection of offending targets is an essential solution for effective investigation. P2P investigators' adaptation of non-invasive monitoring tools to the identification of offending targets is a beneficial and adaptive investigative strategy that serves a particular value in these settings by decreasing detectives' cognitive load and increasing their capacity to secure charges and convictions against those culpable for significant dissemination of CSA media on P2P networks. However, this adaptation is not without its flaws. While these monitors promote the apprehension of great numbers of P2P offenders with problematic distribution profiles, they do not reliably aggregate offence data, prompt or otherwise enable investigators to develop criminal intelligence in a way that supports the identification of victims of recent or on-going sexual abuse.

Aside from the capacity limitations of P2P monitoring tools in respect of victim identification and other victim-centric endeavours, our respondents' accounts also foregrounded a series of disparities in the availability of broader infrastructural supports that could support victim-centred investigations: (i) the limited availability of human resources, or skilled investigators with requisite expertise in victim identification; (ii) the absence of access to technical infrastructures, such as national and international reference databases of CSA media that can serve (inter alia) as a point for referencing seized imagery and identifying new victimisation; and (iii) the rudimentary implementation of victim-centric, investigative policies and practices. For example, in some countries investigators reported to make systematic referrals of new images seized in P2P investigations to their national databases (and to Interpol's International Child Sexual Exploitation Database), while in other countries similar victim-focused strategies were not apparent. The latter finding highlights a core challenge to the identification of victims and perpetrators of child sexual abuse in P2P investigations – the identification of new or previously unknown CSA media is a fundamental prerequisite of victim identification investigations insofar as these materials comprise the primary evidential artefact upon which image analysis and other victim identification strategies are enacted (Holland, 2005). Hence, a principal concern for each of our respondents was the identification of any type of traded CSA media that suggested the file was *new*, i.e. depicting new victimisation or previously unknown to law enforcement, domestic in origin, or that otherwise suggested some proximate connection between the producers of the material and distributors on a P2P network.

> "Whilst it may not be at the bad end of the scale, I would say that any material, if we identify any material that was being traded that was first-generation material – anything that's been, you know, home movies – if we believe it's been taken by the sender that would be prioritised. Because a first generation image is a clear indicator of abuse." (Australian Police, Respondent 1). "If you run your database across a computer, you are going to know that there is *x* amount of hundred or thousand known images. The ones that aren't known are the ones that the officers have to concentrate and look for." (UK Police, Respondent 1).

The identification of new CSA media was identified over a decade ago as a central operational objective for police and law enforcement concerned with the identification of victims of CSA media; one that was much more difficult, more resource-intensive and much more challenging to operationalise than more traditional investigative activities such as the disruption of CSA media distribution networks (Taylor and Quayle, 2003). However, our respondents reported that significant challenges persist in the identification of new CSA media in P2P investigations. Some 91% of our survey contributors indicated that they maintained no (or very limited) capacity for the identification of new CSA media and its originators in P2P environments and expressed strong support for the development of such a functionality within live forensic analysis contexts. Furthermore, they reported that the significant requirement for manual analysis in the identification of new CSA media, coupled with the sheer volume of CSA media exchanges on P2P, compounds the challenge of enacting victim-centric investigations in this policing domain. Also, law enforcement agencies maintain little capacity to support large-scale downloading of candidate CSA media files in support of content analysis and verification tools.

The review we presented in this section highlights the fundamental need for the development of approaches that enable P2P investigators to identify and prioritise cases where the target is engaged in the sexual abuse of children and/or the production of CSA media. While some preliminary attempts have been made to utilise materials accessed by suspects to assist in prioritising which investigations take place first (e.g. (McManus et al., 2011)), to this day no framework exists that can reliably discriminate high-risk targets in P2P policing contexts – such as those distributing new/previously unknown CSA media that may indicate recent or on-going child abuse. It should be noted, however, that law enforcement requirements for the development of victim-centric approaches in P2P investigation are in some cases not simply met by extending the functionalities of extant P2P monitoring systems to the identification of new CSA media. Law enforcement contributors in several jurisdictions reported that they were unable to deploy non-invasive P2P monitoring tools such as CPS in their P2P investigations due to statutory proscriptions on the enactment of pro-active surveillance

strategies where, for example, suspects may be subjected to continuous monitoring in order to intercept a criminal offence. Yet, they were similarly challenged by the problem of P2P offending and by their inability to consistently identify new CSA media in these investigations.

Background

Digital forensics

After seizing a suspect's computer, police investigators typically create a forensic "image" by copying either the entire disk or a subset from the data (e.g. a disk partition) from the target device to a second hard drive. In some cases, running computers are even unplugged to avoid changes made to the hard disk during the shut-down process. This way, investigators are able to analyse the computer's content without tampering the original evidence. This type of criminal investigation is usually referred to as *traditional* digital forensics (United States Secret Service, 2002; Jones et al., 2006; Vidas et al., 2014). In *live* digital forensics, however, investigators interact with a running computer or system in order to gather intelligence and, based on that intelligence, they determine the following steps in the investigation (Vidas et al., 2014). Furthermore, digital forensics can be applied both *reactively* – as an investigative act in an on-going investigation, and pro-actively – before an incident is officially reported to police investigators.[2] For the purpose of this study, identifying new/previously unknown CSA media on P2P networks is considered pro-active police work using live digital forensics.

Analysing big forensic data

Big Data is defined by (Gartner IT Glossary, 2016) as "high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making". With regard to digital forensics, the increasing volume of digital forensic evidence significantly contributes to lengthy backlogs of cases within computer crime units and forensic laboratories (Casey et al., 2009; Ferraro and Russell, 2004). For a wide range of investigations, including P2P network investigations pertaining to CSA media, there is a need to focus the process of digital forensic data collection and analysis (cf. Section Challenges for live forensic analysis of P2P networks). Recently, a number of methods have been suggested to address these big digital forensic data, including data reduction (Quick and Choo, 2016), data mining (Kantardzic, 2011), artificial intelligence (Marturana and Tacconi, 2013) and triage (Shiaeles et al., 2013; Parsonage, 2016). With respect to digital forensics, triage[3] involves a fast analysis of digital intelligence in order to provide investigative leads that are ranked based on importance or urgency to be acted upon, while maintaining its integrity and preserving it for an extended analysis using a forensic model in the next phase of the investigation (Rogers et al., 2006; Casey et al., 2009; Vidas et al., 2014).

Our approach accumulates insights and techniques from the fields of text categorisation and image classification (both AI) into a forensic triage model, allowing P2P investigators to (1) detect victims at acute risk, (2) assign degrees of importance and urgency to items of evidence in order to assess offenders' potential danger to society and (3) find useful evidence in a timely manner.

Text and image classification

In text/image classification studies, the task consists of automatically sorting documents (e.g. newspaper articles, books, e-mails, etc.) or images and videos into predefined classes or categories (e.g. different genres, topics, etc.) based on their textual/visual content. In current research, the dominant approach to these problems is based on machine learning techniques, in which an inductive process automatically builds a classifier by learning the characteristics of the individual categories from a set of training documents/images. The trained classifier can, then, distinguish between these categories when it is confronted with new texts/images showing similar characteristics.

Automatic text categorisation techniques are currently being used in many different contexts, ranging from spam detection, indexing scientific publications and the population of hierarchical catalogues of Internet resources to finding relationships among biomedical entities.[4] Image classification is mainly used for image retrieval, i.e. detecting (or retrieving) images with a particular visual content in an image dataset.

Related work

Detecting CSA media in P2P networks

As detecting known CSA media is relatively straightforward when a hash-value database is available, initial work in this area mainly focused on the ability to disrupt on-line child exploitation (e.g. (Joffres et al., 2011)), reliability issues regarding mutable identifiers, such as IP addresses and GUIDs (e.g. (Liberatore et al., 2010b; Dai, 2010; Yang et al., 2013)) and the identification of key sharers (e.g. (Westlake et al., 2011)). This has already resulted in a number of forensic tools, such as CPS and RoundUp, that can not only monitor such paedophile activities in P2P networks, but also provide additional features such as geolocation capabilities and centralised databases to assist law enforcement in their international struggle against on-line child exploitation. Moreover, Internet companies such

[2] More specifically, this could mean that the offence has not occurred yet or the offence is still unknown to the police.

[3] This term was adopted from the medical field were it refers to a process for sorting injured people into groups based on their need for immediate medical treatment when only limited medical resources are available.

[4] An overview of text categorisation techniques and applications can be found in, e.g. (Weiss et al., 2010) and (Sebastiani, 2002).

as Google and Microsoft have created software, such as PhotoDNA, that enables law enforcement to detect modified versions of known CSA media[5] (see also (Baluja, 2008; Microsoft, 2009)). However, none of these tools offers support for identifying new/previously unknown child abuse media.

So far, only few attempt have been made to address this issue. The authors of (Edwards and Rashid, 2012) demonstrated that collaborative filtering techniques that are typically used in recommender systems, can be successfully applied to identify new media in P2P networks of a certain category (e.g. pornography, piracy software, popular music). Their method is based on the assumption that file-sharing traffic tends to cluster around interest, especially when it involves illegal content, such as CSA media. Hence, they were able to detect previously unknown examples from these categories without analysing their contents or filenames. Secondly, the MAPAP project (Latapy et al., 2013) specifically targets peer-to-peer file-sharing networks. There, modelling of user activity and identification of CSA-related keywords is utilised to identify child abuse media. However, the first system was not tested on verified CSA data and the latter was not evaluated for the scenario of identifying new/previously unseen CSA media.

Filename categorisation

Although recent work has tackled the problem of detecting deception (Afroz et al., 2012), masquerading behaviour (Rashid et al., 2013) and identifying paedophile grooming activities (Inches and Crestani, 2012) on-line by combining natural language analysis with machine learning, these studies all operated on larger bodies of text (e.g. chat room conversations). Contrary to these studies, our work involved much shorter text fragments, which also included a range of non-standard and/or multi-language forms that are designed by offenders to mask their shared files' illegal content (cf. Section Approach). While a number of studies have focused on textual features that are related to web-accessible images (e.g. (Wang and Kan, 2006; Huang et al., 2003; Munson and Tsymbalenko, 2001)), they do not address any of these additional challenges.

So far, there are only two studies – to our knowledge – that used language analysis techniques to identify CSA media. As we mentioned before, the authors of (Latapy et al., 2013) investigated the feasibility to automatically construct lists of CSA-related keywords. Therefore, in Section Filename categorisation we start by evaluating their keyword-based approach on a new dataset containing verified CSA-related filenames. The second study (Panchenko et al., 2012) examined whether techniques used for SMS normalisation (cf. (Beaufort et al., 2010)) could also be used to circumvent the issue of language variation or noise in CSA filenames. Because their work on pornographic versus non-pornographic filename classification showed very promising results, in Section Filename categorisation we also evaluate our approach on this experimental set-up.

Image and video classification

In recent years, quite some work on detecting pornography in images and videos has been published. Most of these studies focus on the utilisation of skin based features (e.g. (Rowley et al., 2006) (Jones and Rehg, 2002), and (Fleck et al., 1996)) or bag-of-visual-word features (e.g. (Deselaers et al., 2008)). These techniques can also be applied to video data by drawing key-frames from the video stream and extracting image features like colour histograms and skin features (e.g. (Lee et al., 2006)) or skin area shapes (e.g. (Kim et al., 2008)). For video data, prior research also incorporated acoustic features, such as MFCCs (Zuo et al., 2008) or so-called *audio words* (Liu et al., 2011), which entails an analogue approach to the visual words technique, based on vector quantised audio features. Furthermore, motion features, such as motion histograms (Jansohn et al., 2009) and the autocorrelation of motion signals (Xiaofeng et al., 2005), have been considered for pornography detection as well.

However, only few studies have investigated the feasibility to identify CSA data based on visual features (e.g. (Ulges and Stahl, 2011)). The authors of (da Silva Eleuterio and de Castro Polastro, 2012) utilise skin detection in combination with filename analysis and a hash-based detection method. Another quite common approach, especially in forensics, is searching for visually similar image media in an index of already known images as presented in (LTU Engine, 2010) and (Netclean Analyze, 2010). By applying a *query-by-example* technique, visually similar or identical images with respect to the given query can be retrieved. Such functionality has shown to be suitable for finding images that originate from known series or locations. However, similarity retrieval and hash-based approaches appear not directly suitable for detecting new or previously unknown media files, because both techniques require the availability of (similar) media files to be indexed or hashed. This in turn renders these files as already known. Instead, hash techniques can be utilized to filter known CSA media prior to running the detection module, while similarity retrieval allows determining known instances after performing CSA detection.

Approach

In this section, we present the key components of the iCOP toolkit, that is, the filename categorisation module and the media classification module. The architecture of the toolkit itself and its triage model is discussed in Section The iCOP toolkit. For both modules, we designed the following scenarios that were triggered by practical aspects of law enforcement investigations: (1) detection of CSA versus regular media (WORLD) and (2) distinguishing CSA from legal pornographic media (ADULT). As CSA media content can be considered a subclass of pornography, the second scenario was expected to be much more

[5] A comparative analysis of currently used methods for detecting child abuse media on-line can be found in (Westlake et al., 2012).

challenging. We evaluate both classification modules in Section Experiments and results.

Filename categorisation

Building a filename categorisation module that is sufficiently robust so it can be employed by an automatic environment such as the iCOP toolkit is a difficult task for a variety of reasons. First, for a machine learning algorithm to be effective in identifying candidate CSA media based on textual features in their filename, it needs to be trained with both CSA and non-CSA filenames. However, there are no CSA datasets publicly available and crawling for CSA files directly from a P2P network to acquire training data is illegal. Hence, we could only use CSA-related filenames that were provided to us by law enforcement.[6] Secondly, the task typically involves a great number of very short text samples, which inevitably leads to highly sparse data (i.e. with both a huge number of instances and features, but only few features per instance). A third challenge lies within the class imbalance inherent to the task: in a P2P environment, the number of non-CSA media that are being shared highly predominates the number of CSA files. As most machine learning algorithms are designed to optimise the overall accuracy rate, they have been shown to have difficulty identifying documents of the minority class (see e.g. (Seiffert et al., 2014)). Finally, sharers of CSA media tend to create a specialised vocabulary, containing a whole variety of multilingual keywords, abbreviations and acronyms (e.g. "kinderficker", "kdquality", "ptsc") to circumvent detection by law enforcement, while maintaining their availability to other offenders. This poses great difficulties for automated detection techniques – especially because this vocabulary also proved to be dynamic, i.e. it evolves as existing keywords come to the attention of P2P investigators (cf. (Latapy et al., 2013; Panchenko et al., 2012)). Moreover, supporting multiple languages typically requires sophisticated language identification/translation techniques. In this section, we discuss our approach to address these challenges.

Dataset

Prior research on the Gnutella network (Hurley et al., 2013) reported that 1.6 out of every 1000 files matched with known CSA media. Hence, to create a good reflection of reality, for the filename categorisation module, we also adopted a highly skewed data distribution during the learning experiments. More specifically, we matched 10,000 CSA filenames with 1,000,000 regular filenames from the Gnutella network for the WORLD class and 1,000,000 filenames that were linked to legal pornography media that were taken from *PicHunter*, *PornoHub*, *RedTube* and *Xvideos*[7] for the ADULT class. As mentioned earlier, most filenames are very short, containing only 41.6 characters on average ($SD = 23.4$).

[6] These filenames were mainly collected by researchers from evidence in closed court cases.

[7] www.pichunter.com, www.porno-hub.com, www.redtube.com, www.xvideos.com.

Feature types

As we mentioned before, distributors of CSA media tend to use multilingual, specialised vocabulary and include spelling variations together with other noise in their filenames to avoid (automatic) detection of their shared files, while making them widely searchable for other offenders. Because the presence of such "secret keywords" (e.g. "lolita", "childlover", "kdquality", "ptsc") in a filename is highly informative, we first created a dictionary-based filter containing a manually extended version of the CSA-related keyword lists from the MAPAP project (Latapy et al., 2013). We refer to these as our **CSA Keyword features**. We further extended this filter with forms of explicit language use (e.g. "handjob"), expressions relating to children (e.g. "kiddie") and family relations (e.g. "daughter") in English, German, Dutch, French, Italian and Japanese. Together, these three categories, i.e. the *explicit language*, the *child references* and the *family references*, form our **Semantic features**. Hence, a filename without any CSA-related keywords can still become a high-value target with regard to CSA media when it contains, for example, both explicit language use and references to children (e.g. "handjob11yo"). We show an example of the feature construction in Table 1. The presence of the keyword "pt" (*preteen*) results in a hit for the CSA keyword features, while "12yo" (*12 years old*) is identified as a reference to a child.

While prior work (Latapy et al., 2013; Panchenko et al., 2012) mainly focused on automatically identifying and/or normalising typical keywords that are used by Internet child sex offenders to camouflage their files' illegal content, in this study we apply a more comprehensive approach by combining our dictionary-based filter we described above with other linguistic information. More specifically, we first extracted all patterns of two, three and four consecutive characters from the filenames (also called **character *n*-gram features**). As can be seen from the example in Table 1, this approach allowed us to circumvent the issue of alternative keyword spellings: although the actual keyword "lolita" is not present in the example filename, the presence of the "lita" feature could be equally discriminative when training the classifier, because that feature is also present in filenames that do contain the original keyword. Additionally, other potential cues could be picked up by the model, even when they are related to a new/unknown keyword or produced in a language that is not included in our filter.

Media classification

Automatic assessment whether new or previously unknown candidate media files actually contain represen

Table 1
Example of a CSA filename after feature engineering.

Original filename	ptl0lita12yo.jpeg
CSA-rel. keywords	pt CSA_keyword
Semantic feats.	12yo child_ref
2-g feats.	pt tl l0 0l li it ta a1 12 2y yo
3-g feats.	ptl tl0 l0l 0li lit ita ta1 a12 12y 2yo
4-g feats.	ptl0 tl0l l0li 0lit lita ita1 ta12 a12y 12yo

tations of child sexual abuse requires a content-based analysis of images and videos. This aspect arises from the fact that traditional hash-based assessment of media files relies on lists of known media files and, hence, is not suitable for detecting unknown media. Furthermore, file hashes like *MD*5 or *SHA*1 can easily be circumvented by small alterations to the media file, such as cropping, scaling, colour adoptions or format transcoding. A further technique to hide CSA content is to embed them into other regular images or videos. For all these cases, methods utilising hashes can be expected to fail.

Considering the advances in automatic concept detection that have been achieved in recent years, techniques from the field of computer vision can provide the required capability to detect unknown CSA media content. For this, feature representations of media files are presented to a classifier which generates a model of the concept. The classifier model hereby represents an abstraction of the content appearance representing the concept, as described via feature information. Depending on the complexity of concepts learnt, classifiers can detect appearances of the trained concept in previously unseen media files with good accuracies. However, detecting CSA scenes in digital media remains a challenging problem due to its complex appearance and strong variability. Some work on detecting adult pornography has been published in the past, utilising mostly features that describe the visible presence of human skin (see Section Image and video classification). Though it appears reasonable to apply the techniques for adult pornography detection to CSA media as well, the few attempts being made (e.g. (Ulges and Stahl, 2011)) suggest that these known methods do not achieve comparable detection accuracies. This might arise from an insufficient description of CSA content appearance, which possibly can be overcome by using multi-modal feature representations. Additionally, for detecting CSA media in a real world scenario, as imposed by a P2P network, the frequent occurrence of legal pornography increases the challenge even further for the following reason: adult pornography and CSA depict subclasses of a pornography base class that have a very similar appearance. Moreover, borders between these subclasses are often fluent and even experienced human investigators sometimes are unable to decide correctly. As a result, feature descriptions that are discriminative for the pornography base class typically lack discrimination ability when used for differentiating the subclasses. In the following, our approach to the challenging task of CSA detection in real world scenarios is presented.

Dataset

Processing both images and videos, our content classification module contains two input streams: (a) images, being fed into the feature extraction pipeline directly, and (b) video files, that are pre-processed by extracting video frames and a continuous audio stream. All three classes, i.e. the CSA, WORLD and ADULT class, were represented by 20,000 images and 1,000 short videos each. The non-CSA data were collected from various web sources like *flickr.com*, *youtube.com*, *pichunter.com*, *redtube.com*, *and pornhub.com*. Numerical feature representations for instances of CSA media were provided by European law enforcement. For frame extraction, the input videos are split into shots of 100 frames. The centre frame of each of these video segments is taken as a representative keyframe for extraction of visual features. Additionally, audio features are computed for all 4 *s* segments, respectively.

Feature types

So far, the few studies that have investigated the feasibility to identify CSA data were based on visual features (Ulges and Stahl, 2011), RGB based skin detection in conjunction with filename analysis and a hash-based detection method (da Silva Eleuterio and de Castro Polastro, 2012) or searched for visually similar image media in an index of already known images (LTU Engine, 2010; Netclean Analyze, 2010). Contrary to these previous works, our approach combines content describing visual and acoustic features, instead of using them in isolation. This has already shown promising results for pornography detection in e.g. (Ulges et al., 2012). Moreover, instead of using only a single visual modality, we combine a range of various visual and non-visual features for detecting CSA content. More specifically, for describing the visual content of images and video frames we extract: (i) colour-correlograms, (ii) skin features, (iii) visual words and visual pyramids. The audio information of video files, if available, is described by computing (iv) vector quantised MFCC features or Audio Words. A brief description of the utilised feature extractions is presented next:

i **Colour-correlograms** describe the occurrence probability of a colour in a pixel's neighbourhood (see e.g. (Huang et al., 1997; Ojala et al., 2001; Rautiainen and Ojala, 2002; Zha et al., 2008)). Hence, they represent the local spatial correlation of colours in images. Here, we apply a special variant of the colour-correlogram, namely the *auto-colour-correlogram*, which describes the probability of the identical colour c reoccurring within a distance d of the current pixel in image I.

$$\alpha_c^d(I) = \gamma_{c,c}^d(I)$$

For an improved performance, this feature is computed in HSV colour space ((Ojala et al., 2001)).

ii The **Skin-feature** is based on a RGB skin colour model, which was generated using manually segmented images from the COMPAQ database. The presence of skin is indicated via a *skin-probability-map* (SPM) (see also (Jones and Rehg, 2002)).

$$P(skin|c) = P_{skin}(c)/(P_{skin}(c) + P_{non-skin}(c))$$

For computing the skin feature, first the SPM is transferred into a *skin-segmentation-mask* (SSM) via morphological operations and adaptive thresholding. Next, the

mean intensities of SPM and SSM are calculated, as well as their centre and variance of skin mass. This yields a 14 dimensional descriptor representing appearance properties of skin.

iii Because colour features – especially skin features – are not very robust towards illumination changes, we also computed **Visual Words and Pyramids** to provide a texture based content representation. Visual words features are computed by scaling the given image to 250 × 250 and extracting patches of 8 × 8 using a regular sampling with a step size of 5 pixels. Next, the DCT coefficients of the YUV transformed patches are computed for each channel. The final feature is represented by 36 low frequency coefficients from the Y-channel and 21 taken from U and V, respectively. Our visual pyramid features are based on the same representation and are structured according to (Lazebnik et al., 2006). Applying 2,000 entry codebooks for vector quantisation yields the final representation of both features.

iv The audio stream of video files is described by extracting **Audio Words**, using the widely used Mel Frequency Cepstral Coefficients (MFCC) (Logan, 2000). We extract MFCCs in steps of 8 *ms*, using a 16 *ms* sliding window. Next, a frequency histogram is computed from the Fourier transform of the signal. For reflecting human acoustic perception, the frequency histogram is weighted by the logarithmic Mel scale. Finally, the weighted histograms are DCT encoded, leading to a 13 dimensional descriptor, which are vector quantised using a 1,000 entry codebook.

Experiments and results

Filename categorisation

To obtain a reliable estimation of the classifier's performance, we applied ten-fold cross validation (cf. (Weiss and Kulikowski, 1991)). In this experimental regime, the available data is randomised and divided into ten equally sized folds or partitions. Subsequently, each partition is used nine times in training and once in test. To enable a comparative analysis between the different scenarios we described in Section Approach, we first compiled a complete dataset containing all 2,010,000 filenames and subsequently created ten training and test partitions. This way, we could vary our training data according to each scenario, but evaluate on the same test data, which still contained all three classes (i.e. CSA, ADULT and WORLD). Next, for each training partition we set up four different learning experiments: (1) CSA VS. WORLD, in which the ADULT filenames were removed; (2) CSA VS. ADULT, where we discarded the world data; (3) CSA VS. MIXED, in which 50% of both non-CSA classes were omitted, and (4) CSA VS. ADULT VS. WORLD, where we reused the data of the third experiment, but set up a three-way classification experiment. Hence, the CSA/non-CSA ratio (i.e. 10,000:1,000,000) in each experiment was maintained. Additionally, we performed these experiments a second time, balancing our dataset in each training partition but maintaining the original skewed datasets in the test partitions. Finally, to enable a valid comparison to previous work in this area (Panchenko et al., 2012), we also set up a balanced learning experiment in which we included only the ADULT and WORLD classes.

For classification, we compared the performance of Support Vector Machines (SVM) to Naive Bayes (NB) and Logistic Regression (LR). During the SVM learning experiments, the *C* parameter was experimentally determined on a development set of each training partition. Because our preliminary experiments showed that a linear kernel was most suitable for dealing with the large, sparse filename dataset, which is in line with (Fan et al., 2008; Hsieh et al., 2008; Yu et al., 2013), we used a linear kernel during the experiments. Additionally, we applied l^2-normalisation on the feature values for faster linear SVM training (cf. (Yu et al., 2012)). Finally, the scores we report are average *precision*, *recall* and *F-score*. These are standard evaluation metrics that can be computed based on the number of true positives (*tp*), true negatives (*tn*), false positives (*fp*) and false negatives (*fn*) in a confusion matrix. The recall score for each class provides information on the number of filenames that were successfully retrieved, while the precision score takes into account all retrieved filenames for each class and evaluates how many of them were actually relevant. The F-score is then the harmonic mean of precision and recall. These measures are defined as follows.

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$F_{score} = 2\,\frac{Precision \cdot Recall}{Precision + Recall} \tag{3}$$

The results in Table 2 show that the Support Vector Machines and the Logistic Regression algorithm both significantly outperform the Naive Bayes classifier. Because the SVM model achieved the best F-score for identifying CSA-related filenames, SVM's were used for the remaining learning experiments. To compare our approach of including additional (noisy) linguistic information to the work of (Panchenko et al., 2012) who attempted to normalise the non-standard language varieties in each filename, we set up a balanced learning experiment in which the classifier was trained to automatically distinguish between the WORLD and the ADULT class. Combining

Table 2
Results of the filename classification experiments using different machine learning algorithms.

Scores (%)	CSA			NON-CSA		
	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.
Naive Bayes	62.3	5.7	10.4	99.2	**99.9**	99.6
Support vector machines	79.7	**43.1**	**55.8**	**99.5**	**99.9**	**99.7**
Logistic regression	**83.0**	37.6	51.7	99.4	99.2	99.3

The best results are printed in bold.

Table 3
Results of the filename classification experiments using different feature types.

Scores (%)	CSA			NON-CSA		
	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.
CSA-rel. keywords	93.6	6.9	12.9	99.2	99.9	99.6
Semantic feats.	25.6	2.4	4.4	99.2	99.8	99.6
Char. *n*-grams	79.2	41.9	54.9	**99.5**	**99.9**	**99.7**
Combined	**79.7**	**43.1**	**55.8**	**99.5**	**99.9**	**99.7**

The best results are printed in bold.

character *n*-gram features with the Semantic features we described in Section Feature types resulted in a slightly higher accuracy score of 99.3%[8] and a 99.4% precision, a 99.1% recall and a 99.3% F-score for the ADULT class.

However, as expected, identifying CSA-related filenames proved to be much more challenging. Although when training on the CSA-related keyword features (cf. Section Feature types) the classifier achieved a very high precision score of 93.6%, it also yielded a very low recall and F-score of 6.9% and 12.9%, respectively. In practice, this would mean that out of 10,000 of our verified CSA-related filenames, 9310 would remain undetected when using the keyword-based approach that was suggested by (Latapy et al., 2013). Therefore, in a next series of experiments we included our Semantic features and character *n*-grams. Combining all features resulted in a significantly higher recall score of 43.1% and a 55.8% F-score, but also led to a decrease of the precision to 79.9%. As a result, out of 5175 predicted CSA-related filenames 932 would be labelled as false positives by law enforcement in the framework of a real-life investigation. The results of these experiments are shown in Table 3.

With regard to the different training set-ups, the best precision score was achieved when including all three categories in training. Reducing the ADULT and WORLD categories to NON-CSA increased the recall to 57.6%, but decreased the precision to 56.9%, which resulted in a slightly higher F-score of 57.2%. This decrease in precision could be explained by the fact that two disparate classes (ADULT and WORLD) were combined into a single category. As can be seen in Table 4, the two other learning experiments both produced high recall scores, but low precision and F-scores. Balancing our dataset in each training partition, while maintaining a skewed dataset in the test partitions, led to a significantly higher recall score of 80,8%, but the precision and F-score both decreased to 13.7% and 23.7%, respectively.

Media classification

During the experiments, we first extracted all feature types we described in Section Feature types from the data, followed by a selection of 1,500 training and 3,000 test samples. Because the filename classifier already filters relevant content, during these experiments, we

Table 4
Results of the filename classification experiments using different training set-ups. The set-ups marked with (*) were balanced in training.

Scores (%)	CSA			NON-CSA		
	Prec.	Rec.	F-sc.	Prec.	Rec.	F-sc.
CSA VS. WORLD	2.0	60.1	3.9	99.5	74.8	85.4
CSA VS. ADULT	18.2	76.6	29.4	99.8	97.0	98.4
CSA VS. MIXED	56.9	57.5	**57.2**	99.6	99.6	99.6
CSA VS. ADULT VS. WORLD	**79.7**	43.1	55.8	99.5	**99.9**	**99.7**
*CSA VS. WORLD	2.6	86.5	5.12	99.8	72.5	84.0
*CSA VS. ADULT	5.5	**87.5**	10.4	**99.9**	87.0	93.0
*CSA VS. MIXED	10.0	85.7	17.9	**99.9**	93.3	96.5
*CSA VS. ADULT VS.WORLD	15.0	80.5	25.3	99.1	94.4	96.7

The best results are printed in bold.

represented positive and negative classes in equal amounts (see also (He and Ma, 2013)). Next, all extracted features f were presented to separate statistical classifiers. For classification, we also used SVM's, as they have shown superior performance compared to other options (e.g. (Ulges et al., 2012)). For estimating the C parameter, we performed a 5-fold cross validation. Once again, our preliminary experiments showed that a linear kernel was most suitable for the task. Finally, the scores of the individual classifiers were combined using a weighted sum *late fusion* scheme, yielding a multi-modal classification score for a 4*s* video segment or image X.

$$P(CSA|X) = \sum_f w_f \cdot P^f(CSA|X)$$

The weights w_f for late fusing the trained classifiers were found by grid searching possible classifier combinations. Averaging the performance of all 5 folds provided the numerical results of the experiments, presented in terms of *average precision* (AP) and *equal error rates* (EER). As can be seen in Tables 5 and 6, our classifiers reach average precision in excess of 92% (image) and 95% (video) when compared with adult pornography.

Results for CSA media detection

The predominant method that investigators use to discover CSA content in P2P networks is a matching of candidate files with known material based on file hashes.[9] Additionally, 70% of our law enforcement experts claimed to use lists of CSA-related keywords and abbreviations in their investigations.[10] Other information sources, like the image content, were less common.

In this section, we showed that it is feasible to design an intelligent filtering module that can automatically distinguish between CSA-related filenames and other P2P material (including adult pornography) while maintaining the complex conditions of a P2P scenario – a large, highly skewed, sparse dataset. Although this approach significantly outperforms the standard keyword-based approach,

[8] The authors of (Panchenko et al., 2012) reported a best accuracy score of 97.7% for detecting adult pornography filenames.

[9] 96% of our survey participants claimed to use this method.

[10] In 53% of cases, these were official lists distributed by organisations like InHope, Interpol, IWF, FBI, ICE, CPS and in the other cases self-created lists.

Table 5
Late fusion weights w_f and classification results (single and fused) for CSA detection in images.

Feature	CSA VS WORLD			CSA VS ADULT		
	w_f	AVP	EER	w_f	AVP	EER
Correlogram	0.6	**92.9**	**14.0**	0.7	**91.1**	**16.8**
Vispyramids	0.4	91.4	16.1	0.4	87.4	20.6
Skin segment	0.1	81.3	26.4	0.0	74.2	33.6
Fused		**94.7**	**11.7**		**92.1**	**15.5**

The best results are printed in bold.

Table 6
Late fusion weights w_f and classification results (single and fused) for CSA detection in videos.

Feature	CSA VS WORLD			CSA VS ADULT		
	w_f	AVP	EER	w_f	AVP	EER
Audio words	0.4	88.3	16.2	0.5	**90.2**	**15.3**
Correlogram	0.3	90.2	14.0	0.4	86.1	16.1
Vispyramids	0.2	89.9	14.1	0.2	82.0	19.4
Viswords	0.1	**90.4**	**13.8**	0.0	79.4	20.5
Fused		**97.3**	**7.5**		**95.7**	**8.2**

The best results are printed in bold.

a false positive rate of 20.3% indicates that a decision from the filename classification module is still insufficient to label a candidate file as CSA-related media. Hence, a highly precise image classification module is required as a second step in the analysis. When combined, our system was able to further reduce this false positive rate to 7.9% for images and 4.3% for videos.

In the next two sections, we discuss the iCOP toolkit's architecture and we present the results of a user trial.

The iCOP toolkit

The filename and image classification approaches are synthesised in the iCOP toolkit to identify and prioritise new/previously unknown CSA media. As shown in Fig. 1, the toolkit has two major components: the P2P Engine and the iCOP Analysis Engine.

The P2P engine provides functionality to monitor public traffic on Gnutella, but other monitors can be plugged into the engine as well. The monitor extracts information such as IP addresses, filenames and hash values of files, together with meta data, such as when a particular peer was last seen sharing a file. The latter is essential to identify the originator of a file after it has been labelled by the toolkit as containing new/previously unknown CSA content. This information is passed on to the iCOP analysis engine, which undertakes the following steps:

1. It compares the hash values of files to a list of known hashes. As we mentioned above, such hash value lists are established by law enforcement when CSA media are seized. This filtering mechanism ensures that the system disregards known CSA media. Although the user interface does indicate when a peer is sharing known CSA media, the toolkit does not download or process the files given the focus on identifying new/previously unknown CSA media. This significantly reduces both storage and computation requirements. We currently use a file of SHA1 hashes in base-32 (one hash per line), because this is the most common format in which law enforcement store hash values for CSA media. As a result, the design enables law enforcement officers using the toolkit to plug in their own hash value lists without substantial effort to import them into a specific format or database.
2. The names of files that do not occur in the known hash list are then passed on to the filename classifier for identifying their likelihood of containing CSA media. This is the first step of the automatic CSA media analysis. Filenames that are deemed to be non-CSA media are discarded.

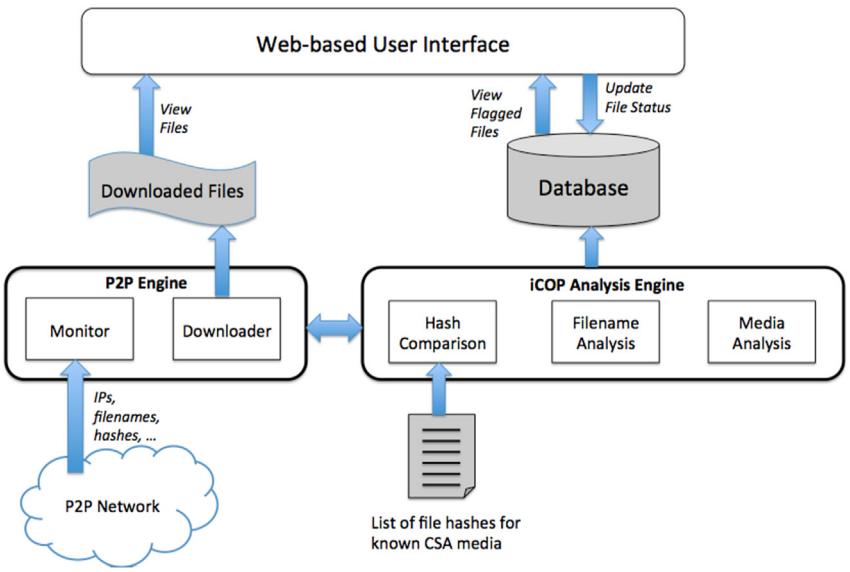


Fig. 1. Overview of the iCOP toolkit.

3. Files that are flagged by the filename classifier as potentially containing CSA media are passed back to the P2P engine for downloading. The downloaded files are then piped back to the iCOP analysis engine and are analysed by the media classifier – the second step of the automatic CSA media analysis– to determine if the content indeed contains child abuse images.

The results of the analysis are stored in real-time in the database. An investigator can login to the GUI to access the iCOP "dashboard", which automatically triages the results to flag the most pertinent candidates for CSA media as the highest priority. More specifically, the main table displays details of the connections sharing the greatest number of suspected files based on the results of the image analysis module. The table can also be sorted according to total number of shared files, number of suspicious filenames, or number of shared files known to be CSA media. Additionally, the user can view thumbnails as well as the full media files to verify whether the flagged items are indeed CSA media. If so, these items can be marked "confirmed" by the user and are fed back into the hash database so that they are considered to be known child abuse media in future searches.

Furthermore, the toolkit GUI is designed around a list of connections, which maps closely to the way P2P software works. A connection is defined as:

$$connection = IP\ address + Port + GUID$$

Each connection is assumed to be a single user sharing a given set of files from a specific location. This is in contrast with an IP address alone, which could potentially be shared by multiple users (e.g. several machines in a home) or a GUID alone, which could potentially be used from different locations (e.g. work, home, travel). The toolkit can display files shared by a particular IP or a particular GUID. Hence, an investigator can easily view which connections are related via a common IP address or GUID. As mentioned above, the most pertinent candidates are flagged to the user as high priority via the dashboard. Given legal constraints governing law enforcement, the toolkit's modules can also be integrated separately into extant investigative workflows and/or configured to focus on particular geolocations (e.g. a particular country or region). Additionally, the system provides a demo mode to allow for testing and debugging it using dummy P2P network data and legal pornography. This is because any monitoring and downloading of CSA media can only take place at suitable law enforcement premises.

Finally, in order to accommodate the requirements for the development of victim-centric approaches in jurisdictions that are restricted by statutory proscriptions on the enactment of pro-active surveillance strategies (cf. Section Challenges for live forensic analysis of P2P networks), the toolkit was developed in accordance with a modular design that permitted flexibility in any operational application of the iCOP toolkit. This configuration enabled domain law enforcement to deploy iCOP as an integrated solution alongside existing P2P tools in proactive monitoring contexts, or to adapt its media or filename classifiers to the identification of new CSA media during reactive investigative activities (e.g. incorporating iCOP's component classifiers into triage investigation procedures during post hoc forensic examinations of seized hardware).

Table 7
Survey results.

Part.	Q1	Q2	Q3	Q4	Q5	Q6
1	6.0	6.0	6.0	5.0	5.0	6.0
2	6.0	6.0	6.0	5.0	5.0	6.0
3	4.5	6.0	6.0	4.0	7.0	7.0
4	4.0	4.0	3.0	6.0	5.0	4.0
5	6.0	7.0	7.0	5.0	6.0	7.0
6	5.0	5.0	5.0	4.0	5.0	7.0
7	6.0	7.0	6.0	7.0	5.0	7.0
8	1.0	1.0	2.0	2.0	1.0	1.0
9	3.0	5.0	3.0	4.0	3.0	5.0

User evaluation

To acquire more insight into the toolkit's usability, we conducted a live testing workshop with 9 law enforcement officers engaged in CSA investigations on P2P networks from 7 different law enforcement agencies across Europe. Given the legal (as well as ethical) issues pertaining to such a live exercise, the two day workshop was conducted on law enforcement premises. The participants were provided background information on the toolkit as well as details of how the analysis is performed by the backend. They were provided training in the use of the user interface followed by actual use of the toolkit on live data. A lot of usability feedback was gathered in focus-group style discussions and used to improve the functionality and the user interface subsequently. At the end of the workshop, a questionnaire was completed by the participants. Each question was answered according to a 7-point Likert scale (1 – strongly disagree … 7 – strongly agree) and had room for comments. The questions were as follows:

1. The toolkit has all the capabilities I need to prioritise investigations.
2. I believe the toolkit can facilitate my investigations.
3. I believe the toolkit will allow me to more efficiently carry out investigations.
4. I believe the toolkit will assist me in efficiently analysing the large number of files shared on P2P networks.
5. I would frequently use this toolkit as part of my investigations.
6. Overall, I believe the toolkit to be a valuable aid to law enforcement.

The results are summarised in Table 7 where Q is the question and P the participant. Participant 8 was a consistent outlier in terms of low scores.[11] Participant 9 emphasised that he found problems in the user interface rather than the available features. Particularly noteworthy is the positive feedback for facilitation of investigations (Q2), and value for law enforcement (Q6).

[11] This participant chose to give low scores because s/he did not find any files that prompted him/her to launch an investigation during the short workshop.

Conclusion

The increasing amount of CSA media being shared across borders and with apparent impunity leads to new children being found on-line every day. Each of these children, often from within the family circle of the offender, is a victim of child sexual abuse. Whether charged with enforcing the law in respect of broader offences of possession and distribution, or with the apprehension of producers of child abuse media, the identification of contact sexual abuse and abuse victims were cited as paramount concerns for P2P investigators. This finding resonates with earlier observations that a primary goal of P2P investigations is to catch child abusers and help children that are being sexually victimised, rather than simply detecting and confiscating images in the context of possession offences (Liberatore et al., 2010a, 2010b). However noble, these objectives are difficult, nigh impossible to realise using state-of-the-art tools such as CPS and RoundUp. Such tools, which identify suspects involved in the exchange of known CSA files, yield many potential targets for law enforcement but offer little support for the identification and prioritisation of high-risk targets – such as those distributing new/previously unknown CSA media that may indicate recent or on-going child abuse. In this paper, we presented the iCOP toolkit: a forensic software package that is designed to highlight sharers of new/previously unknown child sexual abuse media in P2P networks. Additionally, it offers secondary features, such as showing sharers of known CSA files and allowing police investigators to see other files shared by the same computer or other IP addresses used by the same P2P client. Hence, the software allows P2P investigators to more rapidly locate the producers of such content and the victims therein. Moreover, its modular design enables law enforcement agencies to integrate (parts of) the toolkit into their extant investigative workflows or to add new extensions for other types of P2P clients and networks. The software is currently being made available to law enforcement. Interested parties should contact the authors for more information.

Although the current realisation of both the filename and the image classification modules already provide very high results, they could be further optimised. First, the classification of images is still limited by operating only on low-level visual features. Future research can potentially address this in multiple ways. While CSA video classification can be significantly improved by additionally using other modalities, i.e. audio or motion information, image classification could be extended by utilising high-level visual features. For example, the novel SentiBank feature (Borth et al., 2013), which consists 1, 200 classifier scores indicating the presence of pre-trained concepts, could achieve some orthogonality towards low-level descriptions in feature space, because they build up on different information sources. Another challenge for future research is the age verification of individuals appearing in questioned images and videos. Though evaluations have been conducted during the development of the media classification module, current approaches to determine the age of persons for supporting the classification decision cannot provide the robustness that is needed in an automated environment such as the iCOP toolkit. Also, the filename classification module could be further enhanced by analysing and retraining on previously unknown filenames that were identified by the toolkit and labelled as true positives by police investigators.

Furthermore, the same techniques for monitoring and analysing filenames and file content we propose for the Gnutella network could also be applied to other file sharing systems, such as eDonkey and Bittorrent. Monitoring these networks, however, will require different software libraries for each protocol and may yield different types of information that require slightly different database structures. In addition, some protocols rely on central servers and will need to be manually configured to monitor the servers of interest. Finally, quite a few networks are designed to provide anonymity and prevent freeloading. These issues will provide a further venue for our future work.

Acknowledgement

This work was funded by the European Commission Safer Internet Programme project (SI-2010-TP-2601002), *iCOP*: *Identifying and Catching Originators in Peer-to-Peer Networks*, and by the Antwerp University, *DAPHNE*: *Defending Against Paedophiles in Heterogeneous Network Environments*. The authors would also like to thank all law enforcement agencies that have contributed to this project. Finally, special thanks go out to Interpol. Without their efforts, our research would have been impossible.

References

Afroz S, Brennan M, Greenstadt R. Detecting hoaxes, frauds, and deception in writing style online. In: Proceedings of the IEEE Symposium on Security and Privacy; 2012. p. 461–75.

Baluja S. Building Software Tools to find Child Victims. 2008. http://googleblog.blogspot.co.uk/2008/04/building-software-tools-to-find-child.html. last accessed in March, 2014.

Beaufort R, Roekhaut S, Cougnon L, Fairon C. A hybrid rule/model-based finite-state framework for normalizing SMS messages. In: Proceedings of ACL; 2010. p. 770–9.

Benzecri J, Bellier L. L'analyse des donnees. La taxinomie, vol. 1; 1973.

Borth D, Ji R, Chen T, Breuel T, Chang S. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In: ACM Multimedia; 2013.

Braun V, Clarke V. Using thematic analysis in psychology. Qual Res Psychol 2006;3(2):77–101.

Casey E, Ferraro M, Nguyen L. Investigation delayed is justice denied: proposals for expediting forensic examinations of digital evidence. J Forensic Sci 2009;54:1353–64.

Child Protection System. P2P Monitoring software developed at TLO, http://www.tlo.com/, USA.

Cooper A. Sexuality and the internet: surfing into the new millenium. Cyberpsychology Behav 1998;1(2):187–93.

Dai L. An Ising-based Approach for Tracking Illegal P2P Content Distributors. Master Thesis. USA: Iowa State University; 2010. http://archives.ece.iastate.edu/archive/00000600/01/thesis.pdf. last accessed in March, 2014.

da Silva Eleuterio P, de Castro Polastro M. An adaptive sampling strategy for automatic detection of child pornographic videos. Int. Conf. on Forensic Comput Sci 2012:12–9.

Deselaers T, Pimenidis L, Ney H. Bag-of-visual-words Models for Adult Image Classification and Filtering. ICPR; 2008. p. 1–4.

Edwards M, Rashid A. Collaborative filtering as an investigative tool for peer-to-peer filesharing networks. SCIENCE 2012;1(2):67–79.

Fan R, Chang K, Hsieh C, Wang X, Lin C. LIBLINEAR: a library for large linear classification. J Mach Learn Res 2008;9:1871–4.

Ferraro M, Russell A. Current issues confronting well-established computer-assisted child exploitation and computer crime task forces. Digit Investig 2004;1(1):715.
Fleck M, Forsyth D, Bregler C. Finding Naked People. ECCV; 1996. p. 593–602.
Gartner IT Glossary: Big Data. http://www.gartner.com/it-glossary/big-data/. Accessed 16 May 2016.
Gummadi K, Dunn R, Saroiu S, Gribble S, Levy H, Zahorjan J. Measurement, modeling and analysis of a peer-to-peer file-sharing workload. SIGOPS Oper Syst Rev 2003;37:314–29.
He H, Ma Y. Imbalanced Learning: Foundations, Algorithms, and Applications. John Wiley & Sons; 2013.
Holland G. Identifying victims of child abuse images: an analysis of successful identifications. In: Quayle E, Taylor M, editors. Viewing Child Pornography on the Internet: Understanding the Offence, Managing the Offender, Helping the Victims. Russel House Publishing; 2005.
Hsieh C, Chang K, Lin C, Keerthi S, Sundararajan S. A dual coordinate descent method for large-scale linear SVM. In Proceedings of the twenty fifth international conference on machine learning (ICML). 2008. p. 408–415.
Huang J, Kumar S, Mitra M, Zhu W, Zabih R. Image indexing using color correlograms. CVPR 1997:762–8.
Huang W, Huang W, Tan C, Leow W. Model-based chart image recognition. In: Proceedings of the International Workshop on Graphics Recognition; 2003. p. 87–99.
Hughes D, Walkerdine J, Coulson G, Gibson S. Is deviant behaviour the norm on p2p file-sharing networks? IEEE Distrib Syst Online 2006;7(2).
Hughes D, Rayson P, Walkerdine J, Lee K, Greenwood P, Rashid A, et al. Supporting law enforcement in digital communities through natural language analysis. In: Proceedings of the International Workshop on Computational Forensics; 2008.
Hurley R, Prusty S, Soroush H, Walls R. Measurement and analysis of child pornography trafficking on P2P networks. In Proceedings of the international world wide web conference, Brazil, 2013.
Inches G, Crestani F. Overview of the international sexual predator identification competition at PAN-2012. In: Forner P, Karlgren J, Womser-Hacker C, editors. CLEF 2012 Evaluation Labs and Workshop Working Notes Papers; 2012.
Interpol. Victim Identification. 2014. http://www.interpol.int/Crime-areas/Crimes-against-children/Victim-identification. Last accessed August, 2014.
Irons A, Lallie H. Digital forensics to intelligent forensics. Future Internet 2014;6:584–96.
Jansohn C, Ulges A, Breuel T. Detecting pornographic video content by combining image features with motion information. In: ACM Multimedia; 2009.
Joffres K, Bouchard M, Frank R, Westlake B. Strategies to disrupt online child pornography networks. Paper presented at the European intelligence and security informatics conference, Athens, Greece, 2011.
Jones M, Rehg J. Statistical color models with application to skin detection. Int J Comput Vis 2002;46(1):81–96.
Jones K, Bejtlich R, Rose C. Real Digital Forensics: computer Security and Incident Response. Addison-Wesley; 2006.
Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms. New York: Wiley; 2011.
Kim C, Kwon O, Kim W, Choi S. Automatic system for filtering obscene video. In Proceedings of the 10th international conference on advanced communication technology. 2008. p. 1435–1438.
Latapy M, Magnien C, Fournier R. Quantifying paedophile activity in a large P2P system. Inf Process Manag 2013;49(1):248–63.
Lazebnik S, Schmid C, Ponce J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. IEEE computer society conference on computer vision and pattern recognition, vol. 2. 2006. pp 2169–2178.
Lee H, Lee S, Nam T. Implementation of high performance objectionable video classification system. ICACT 2006:959–62.
Liberatore M, Erdely R, Kerle T, Levine B, Shields C. Forensic investigation of peer-to-peer file sharing networks. In Proc DFRWS Annu Digital Forensics Res Conf, 2010.
Liberatore M, Levine B, Shields C. Strengthening forensic investigations of child pornography on P2P networks. In Proc ACM Conf Emerg Netw Exp Technol (CoNEXT), 2010.
Liu Y, Wang X, Zhang Y, Tang S. Fusing audio-words with visual features for pornographic video detection. In: Proceedings of TRUSTCOM'11; 2011. p. 1488–93.
Logan B. Mel frequency cepstral coefficients for music modeling. In: Int. Symposium on Music Inf. Retrieval; 2000.
LTU Engine, available from http://www.ltutech.com/en/products/ltu-engine-2 (retrieved: June 2010).
Marturana F, Tacconi S. A machine learning-based triage methodology for automated categorization of digital media. Digital Investigation Int J Digital Forensics Incident Response 2013;10(2):193–204.
McManus M, Long M, Alison L. Child pornography offenders: towards an evidence-based approach to prioritizing the investigation of indecent image offences. In: Alison L, Rainbow L, editors. Professionalizing offender profiling: forensic and investigative psychology in practice; 2011. p. 178–88.
Menasche D, Rocha A, Li B, Towsley D, Venkataramani A. Content availability and bundling in swarming systems. In: Proc ACM CoNext; 2009. p. 121–32.
Microsoft. New Technology Fights Child Porn by Tracking its "PhotoDNA". 2009. http://www.microsoft.com/en-us/news/features/2009/dec09/12-15photodna.mspx. last accessed in March, 2014.
Middleton D. "Internet Sex Offenders", Ch. 12 in Assessment and Treatment of Sex Offenders: a Handbook. 2009.
Munson E, Tsymbalenko Y. To search for images on the web, look at the text, then look at the images. In: Proceedings of the 1st International Workshop on Web Document Analysis; 2001. http://www.csc.liv.ac.uk/wda2001.
Netclean Analyze, available from http://www.netclean.com/(retrieved: June 2010).
Ojala T, Rautiainen M, Matinmikko E, Aittola M. Semantic image retrieval with HSV correlograms. In Proceedings of the 12th scandinavian conference on image analysis. 2001. p. 621–627.
Panchenko A, Beaufort R, Fairon C. Detection of child sexual abuse media on p2p networks: normalization and classification of associated filenames. In Proc Workshop Lang Resour Public Secur Appl 8th Int Conf Lang Resour Eval (LREC), 2012.
Parsonage H. Computer forensics case assessment and triagesome ideas for discussion. http://docplayer.net/10051888-Computer-forensics-case-assessment-and-triage-some-ideas-for-discussion.html, [accessed 16.05.16].
Prichard J, Watters P, Spiranovic C. Internet subcultures and pathways to the use of child pornography. Comput Law Secur Rev 2011;27(6):585–600.
Quayle E, Holland G, Linehan C, Taylor M. The internet and offending behaviour: a case study. J Sex Aggress 2000;6:78–96.
Quick D, Choo K. Big forensic data reduction: digital forensic images and electronic evidence. Clust Comput 2016:1–18.
Rashid A, Baron A, Rayson P, May-Chahal C, Greenwood P, Walkerdine J. Who am I? Analyzing digital personas in cybercrime investigations. IEEE Comput 2013;46(4,):54–61.
Rautiainen M, Ojala T. Color correlograms in image and video retrieval. In Proceedings of the 10th finnish artificial intelligence conference. 2002. pp. 203–212.
Rogers M, Goldman J, Mislan R, Wedge T, Debrota S. Computer forensics field triage process model. In Proceedings of the 2006 conference on digital forensics security and law. 2006. p. 2740.
Rowley H, Jing Y, Baluja S. Large scale image-based adult-content filtering. Int Conf Comp Vis Theory Appl 2006:290–6.
Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv 2002;34:1–47.
Seiffert C, Khoshgoftaar T, Van Hulse J, Folleco A. An empirical study of the classification performance of learners on imbalanced and noisy software quality data. Inf Sci 2014;259:571–95.
Shiaeles S, Chryssanthou A, Katos V. On-scene triage open source forensic tool chests: are they effective? Digit Investig 2013;10(2):99115.
Stutzbach D, Rejaie R. Understanding churn in peer-to-peer networks. In: Proc ACM IMC; 2006. p. 189–202.
Taylor M, Quayle E. Child pornography: an internet crime. Brunner-Routledge; 2003.
Tremblay M. The key informant technique: a non-ethnographic application. In: Burgess R, editor. Field Research: a Sourcebook and Field Manual. Routledge; 2003.
Ulges A, Schulze C, Borth D, Stahl A. Pornography detection in video benefits (a lot) from a multi-modal approach. In: Workshop on Audio and multimedia Methods for Large-Scale Video Analysis. ACM; 2012.
A. Ulges, A. Stahl, "Automatic Detection of Child Pornography using Color Visual Words", In Proc. of the Int. Conf. Multimedia and Expo, 2011.
United States Secret Service. Best Practices for Seizing Electronic Evidence. ed. 2 2002.
U.S. Department of Justice. The National Strategy for Child Exploitation Prevention and Interdiction: a Report to Congress. Washington, USA: U.S. Department of Justice; 2010.
Vidas T, Kaplana B, Geigerb M. OpenLV: empowering investigators and first-responders in the digital forensics process. Digit Investig 2014;11(suppl. 1):S45–53.

Wang F, Kan M. "NPIC: hierarchical synthetic image classification using image search and generic features", In Proceedings of the Conference on Image and Video Retrieval, pp. 473–482, 2006.

Weiss S, Kulikowski C. Computer Systems that Learn: classification and Prediction Methods from Statistics, Neural Nets, Machine Learning, And Expert Systems. San Mateo, USA: Morgan Kaufmann; 1991.

Weiss S, Indurkhya N, Zhang T, Damerau F. Text Mining: Predictive Methods for Analyzing Unstructured Information. Springer; 2010.

Westlake B, Bouchard M, Frank R. Finding the key players in online child exploitation networks. Policy Internet 2011;3(2):1–32.

Westlake B, Bouchard M, Frank R. "Comparing methods for detecting child exploitation content online", In European Intelligence and Security Informatics Conference (EISIC), pp. 156–163, 2012.

Wolak J, Finkelhor D, Mitchell K. Child-Pornography Possessors Arrested in Internet-Related Crimes: Findings from the NJOV Study. Technical report. National Center for Missing and Exploited Children; 2005.

Wolak J, Finkelhor D, Mitchell K. Trends in Arrests for Child Pornography Possession: the Third National Juvenile Online Victimization Study (NJOV-3). Technical Report. Durham, USA: Crimes against Children Research Center, University of New Hampshire; 2012.

Xiaofeng T, Duan L, Xu C, Tian Q, Hanqing L, Wang J, Jin J. Periodicity detection of local motion. IEEE International Conference on Multimedia and Expo. 2005. pp. 650–653.

Yang S, Kurose J, Levine B. Disambiguation of residential wired and wireless access in a forensic setting. In: Proceedings of INFOCOM; 2013. p. 360–4.

Yu H, Ho C, Arunachalam P, Somaiya M, Lin C. Product Title Classification versus Text Classification. 2012. Technical report.

Yu H, Ho C, Juan Y, Lin C. LibShortText: A Library for Short-text Classification and Analysis. 2013. Technical Report, http://www.csie.ntu.edu.tw/cjlin/papers/libshorttext.pdf.

Zha Z, Liu Y, Mei T, Hua X. Video Concept Detection using Support Vector Machines – Trecvid 2007 Evaluations. 2008. Technical report.

Zuo H, Wu O, Hu W, Xu B. Recognition of blue movies by fusion of audio and video. In: Proc. of ICME; 2008. p. 37–40.