Abstract | The proliferation of child sexual abuse material (CSAM) is outpacing law enforcement's ability to address the problem. In response, investigators are increasingly integrating automated software tools into their investigations. These tools can detect or locate files containing CSAM, and extract information contained within these files to identify both victims and offenders.

Software tools using biometric systems have shown promise in this area but are limited in their utility due to a reliance on a single biometric cue (namely, the face). This research seeks to improve current investigative practices by developing a software prototype that uses both faces and voices to match victims and offenders across CSAM videos. This paper describes the development of this prototype and the results of a performance test conducted on a database of CSAM. Future directions for this research are also discussed.

Child Sexual Abuse Material Reduction Research Program

Developing automated methods to detect and match face and voice biometrics in child sexual abuse videos

Bryce Westlake, Russell Brewer, Thomas Swearingen, Arun Ross, Stephen Patterson, Dana Michalski, Martyn Hole, Katie Logos, Richard Frank, David Bright and Erin Afana

The proliferation of child sexual abuse material (CSAM) online is outpacing law enforcement's ability to manage the problem (National Center for Missing and Exploited Children 2020). These increasing workloads have significant and severe implications for investigators, with recent evidence tying this work to a range of serious psychological harms, including secondary traumatic stress disorder, emotional exhaustion, intrusive thoughts, burnout, and interpersonal and marital problems (Bourke & Craun 2014; Burns et al. 2008; Powell et al. 2015; Seigfried-Spellar 2018). To address these problems, investigators are increasingly integrating automated software tools into their investigatory workflows. These tools can be used to detect or locate files containing CSAM (eg using hash values), as well as extract information from within files (eg biometrics) that can be used to identify both victims and/or offenders (Canadian Centre for Child Protection 2021; Council of Europe 2021; Internet Watch Foundation 2021; Interpol 2022, 2018).

Biometric detection and extraction approaches, in particular, can increase investigatory capacity and enhance CSAM investigations. These tools typically rely on the detection of 'primary' biometric modalities including faces (Macedo, Costa, & dos Santos 2018; Ulges & Stahl 2011) but also use other 'soft' biometric modalities to detect nudity (de Castro Polastro & da Silva Eleuterio 2010; Vitorino et al. 2018), skin tones (Islam, Watters & Yearwood 2011; Sae-Bae et al. 2014; Yaqub, Mohanty & Memon 2018), and subject age (Gangwar et al. 2021; Islam et al. 2019).

However, the reliance on face as the main biometric routinely used in CSAM investigations has several limitations—for example, the lack of distinct facial features appearing in children, the inability to reliably estimate age, instances where the background and the skin tone of the child are similar, the degree of nudity present, and the substantial proportion of CSAM that purposefully shields faces from view (Moser, Rybnicek & Haslinger 2015; Phippen & Bond 2020; Srinivas et al. 2019; Yiallourou, Demetriou & Lanitis 2017). These problems lead to higher than desired rates of false positive and false negative matches, thus reducing task automation and requiring manual intervention (and exposure) by investigators. Recognising these limitations, researchers have advocated for augmenting facial recognition by combining it with additional primary biometrics (eg voice or iris recognition, fingerprints, and vascular patterns) and soft biometric attributes (eg skin/eye/hair colour and tattoos; Bursztein et al. 2019; Sae-Bae et al. 2014).

Trends in the distribution of CSAM online demonstrate that producers and consumers increasingly prefer video files and lucrative on-demand live streams (Brown, Napier & Smith 2020; Dance & Keller 2020; Maxim et al. 2016). In fact, 2019 marked the first year that reports of child sexual abuse (CSA) videos outpaced those of images (National Center for Missing and Exploited Children 2020). This shift highlights the growing importance of video in CSAM investigations, which practitioners flag as an area needing attention, given that video processing capabilities are not yet as mature as those of images (Sanchez et al. 2019). Further investment in such capabilities has the potential to vastly enhance investigators' capacity to identify victims and offenders in CSAM, given the ability to use additional biometric modalities not present in static images, such as voice.

This paper describes the development of software that uses a combination of biometric modalities (both visual and auditory information) contained within a video file to increase the likelihood of matching CSAM victims. This software, entitled the Biometric Analyser and Network Extractor (BANE), is designed to ingest CSAM (particularly videos), extract multiple biometric attributes (currently faces and voices), and match subjects across videos based on these biometric attributes. This paper is presented in four parts. First, the research aims guiding this work are articulated and elaborated upon. Second, a methodological account is provided, detailing the software development process and the challenges encountered. Third, results are presented, illustrating the matching performance of the software using a testing database of CSA videos depicting victims, compiled by Australian law enforcement agencies. Finally, the implications for future research are discussed.

Aims

This research aims to address current limitations associated with using only facial recognition to identify CSAM victims. While facial recognition has proven effective for investigators, it is limited in its ability to derive matches where faces are concealed or the media is of poor quality. Moreover, the proliferation of videos depicting CSA provides new opportunities to augment existing tools to function in new environments. Accordingly, we propose that combining facial recognition with other biometric modalities, namely speaker recognition, can reduce both false positive and false negative matches, and could enhance analytical capabilities during investigations. The following section provides an overview of the software tool, BANE, designed and developed by the research team specifically to analyse CSAM.

Methodology

Software development

The research team developed a unique video processing methodology as a means of overcoming many of the challenges associated with this form of media. BANE is designed to: (1) ingest video files identified for analysis, (2) extract faces, (3) recognise voices and (4) generate inter-video matches based on these extracted biometric characteristics across the media database. Each of these steps is described in turn.

BANE is equipped to ingest and extract biometric information from a variety of common video formats and was tested using .avi, .mp4, .mkv, .mov, .m4v, .mpg, .mpeg, .3gp, .3gpp, .asf, .wmf, .wmv, .divx and .vob extensions. Videos are ingested and fully decoded using the MoviePy video editing and processing Python package (which uses the FFmpeg library), resulting in a complete sequence of frames and corresponding waveform files. BANE has an application programming interface that processes the face and voice information in each video and compiles the results within its database. Once all the videos are ingested into the database, subject faces are extracted and processed through a four-step pipeline (see Figure 1).

Figure 1: Illustration of the face recognition pipeline for a single video
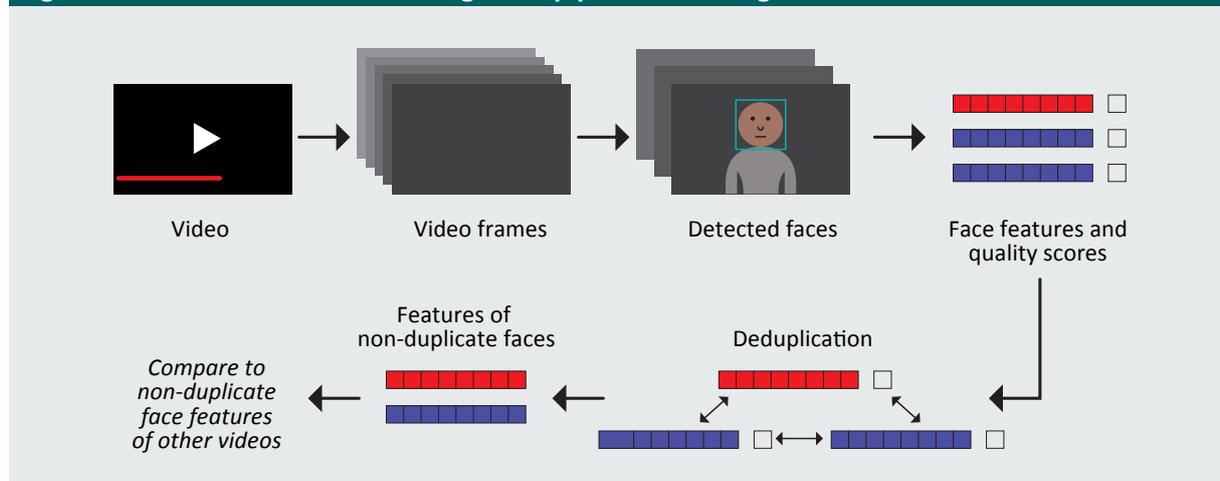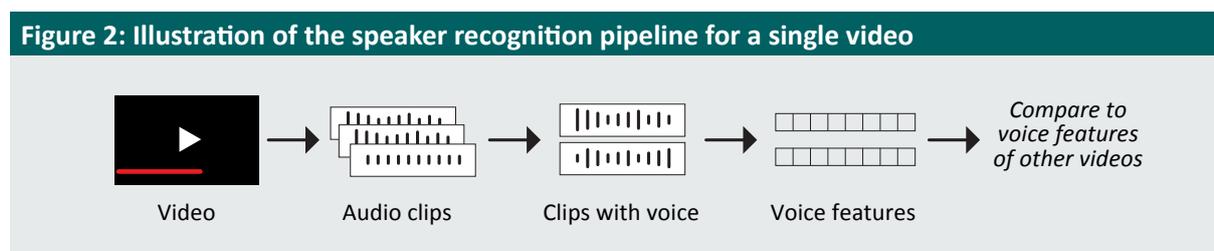
First, every 30th frame in the video (corresponding to approximately one frame every second) is analysed with the face detector to discern how many, if any, faces are present in the frame and their location within the frame. To this end, BANE was developed to allow for different facial recognition algorithms to be implemented or taken out, depending on the context. However, in this work, we use a face recognition algorithm developed by the Australian Government's Defence Science and Technology Group (DSTG), which has been specifically designed to recognise children's faces better than other existing algorithms (Yiu, Malec & Michalski 2021). This algorithm was trained on a diverse range of facial imagery (which included children), and is a convolutional neural network based algorithm, which learns during the training phase which features to use for optimal performance. It does this by maximising the distance between class centres while minimising the distance from samples to their respective class centre.

Second, facial features are extracted from the detected faces using DSTG's algorithm. Once detected, each face is assigned a 'face quality score'—determined by a machine learning system trained to identify face images that lead to higher match scores (Hernandez-Ortega et al. 2019).

Third, the faces within a single video are matched with each other to determine how many faces, if any, are duplicates of the same subject. This step also uses DSTG's algorithm, which was validated using a closed-source dataset, achieving a true match rate of 97.9 percent at a 0.01 percent false match rate. For comparison purposes, the algorithm was evaluated on the open-source dataset of facial images 'Labeled Faces in the Wild' (Huang et al. 2008), achieving a true match rate of 91.9 percent at a 0.01 percent false match rate (acknowledging, however, that this data contains adults only). If the software determines that multiple facial images are of the same person (ie the match score among them is above a specified threshold), then the face with the highest 'face quality' score is marked as the representative face (to be used later for matching) and the rest are marked as duplicates.

In the fourth and final facial recognition step, the set of representative faces from a video are compared to all representative faces from other database videos using the matcher component, and a match score (a similarity score where higher values indicate more similarity) is derived between −1 and 1. By default, BANE alerts the user to a 'match' only for video pairs if the maximum score among all faces in the video pair is above a match score threshold of 0.67 (ie the 'extremum' approach). This threshold corresponds to a 0.01 percent false match rate and a 99.0 percent true match rate.

Once the face matching process is complete, BANE commences the four-step speaker recognition process (see Figure 2). First, the audio from each video is extracted and cut into segments, depending on user-defined parameters (eg a full video or 5-, 10- or 30-second clips).

Figure 2: Illustration of the speaker recognition pipeline for a single video

Video → Audio clips → Clips with voice → Voice features → *Compare to voice features of other videos*

Second, audio segments are processed using a voice activity detector (WebRTC 2017) to ensure that only clips containing voices are entered into the database for subsequent matching.

Third, voice attributes are processed using a speaker recognition algorithm and matcher, specifically designed to address challenging scenarios such as where speech and extraneous noises (eg background music) occur simultaneously. This algorithm, developed and validated by Chowdhury and Ross (2020), extracts features from voice using two commonly deployed features: mel frequency cepstral coefficients (MFCC) and linear predictive coding (LPC). MFCC and LPC represent speech perception features modelled on the human auditory perception system, and speech production features modelled on the human vocal tract. These features are then fused (combined) using a convolutional neural network.

Fourth, using this algorithm, each audio segment from a video is compared to all audio segments from other videos and a match score is given. Two videos are linked as containing the same person if the voice match score is above a certain threshold. As for matching faces, the extremum approach is used to derive match scores (ie the highest match score is used). We use a match score threshold of at least 0.80 as previous analyses of the speaker recognition algorithm on the VoxCeleb database (Chung, Nagrani & Zisserman 2018) identified this as the optimal match threshold in applications such as ours.

Once the face and voice processing and matching is complete, matches can be viewed using a custom-built graphical user interface. In viewing matches, the user can adjust the match score tolerance to better suit the specific operational environment (default: ≥0.67 for face, ≥0.80 for voice).

Compilation of CSA video testing dataset

The BANE software was supplied to Australian law enforcement agencies to evaluate on an annotated database of known CSA videos for the purpose of testing its matching performance. The database contained 70 videos, in their original formats (*.mp4 and *.avi), comprising 21 distinct primary subjects (child victims) with between two and five videos of each. These videos were selected to represent many of the challenging conditions that investigators encounter in real-world deployment. Accordingly, videos contained a variety of forms of CSA, including materials that were both staged and self-produced, and also contained subjects from a range of ages (roughly between 5 and 14 years old). Videos were, on average, 118 seconds in length, with the shortest being 10 seconds and the longest 718 seconds (approximately 12 minutes).

Four criteria governed video selection and inclusion in the dataset. First, each video contained CSAM, as defined under Australian law. Second, each video contained a primary subject (a child) whose face was visible at some point and whose voice was clearly audible (ie they spoke at least a few words). Third, videos containing more than one voice were excluded to permit a standalone assessment of the speaker recognition algorithm. Finally, if any secondary subject (eg offender) appeared in more than one primary subject's videos, those videos were excluded. This ensured that matches identified were of the primary subject and not a secondary (eg background) subject, which would undermine the testing procedure.

Analytical procedure

The entire testing dataset was ingested into BANE and the videos were processed as per the methodology outlined above. Faces were successfully detected in and extracted from 68 of the 70 videos. Voice clips were successfully extracted from 68 of the videos. There was no overlap between the videos in which faces could not be extracted and those for which voices could not be extracted. The research team elected to include the four videos where only a single biometric cue was extracted in the analysis, to simulate real-world deployment, where the quality of files varies. The performance test routine included two additional steps in order to automatically generate results. First, BANE was provided with information pertaining to the 'ground truth' of the dataset. Each video pair was labelled as a genuine pair (where both videos contained the same subject) or an imposter pair (where the videos contained different subjects). Using this data, BANE was programmed to assess its own match performance.

Results and discussion

Given the composition of the database, there were a possible 91 genuine video pairs (ie two videos containing the same subject) and 2,324 imposter video pairs (ie pairs of videos containing different subjects). The performance results for face recognition, speaker recognition, and fusion (face and speaker combined) are presented separately.

Matching performance using only face or speaker recognition

Figures 3 and 4 display the performance for face recognition and speaker recognition in the form of receiver operating characteristic (ROC) curves, which plot the true match rate (% of genuine video pairs correctly classified as a match) against the false match rate (% of imposter video pairs erroneously classified as a match), at many thresholds. Users may select a threshold according to operational requirements. For example, law enforcement may select a lower threshold, to capture the greatest number of genuine matches, but in doing so will need to accept a greater number of imposter pairs being incorrectly classified as matches. Conversely, investigators may select a high threshold, which will reduce the number of imposter matches identified but also reduce the proportion of genuine matches identified.

Figure 3 depicts the ROC curve for face recognition and shows that a false match rate of 1.0 percent corresponds to a true match rate of 72.1 percent. If a higher true match rate is prioritised, accepting a higher false match rate, a false match rate of 5.0 percent corresponds to a true match rate of 91.9 percent. Figure 4 reports the ROC curve for speaker recognition and shows that a false match rate of 1.0 percent corresponds to a true match rate of 65.1 percent. Again, if the priority is identifying as many genuine matches as possible, a higher false match rate of 5.0 percent corresponds to a true match rate of 84.9 percent. These results demonstrate that both face and speaker recognition can effectively be independently used for matching purposes in CSAM contexts, with relative performance being scalable according to investigator needs.

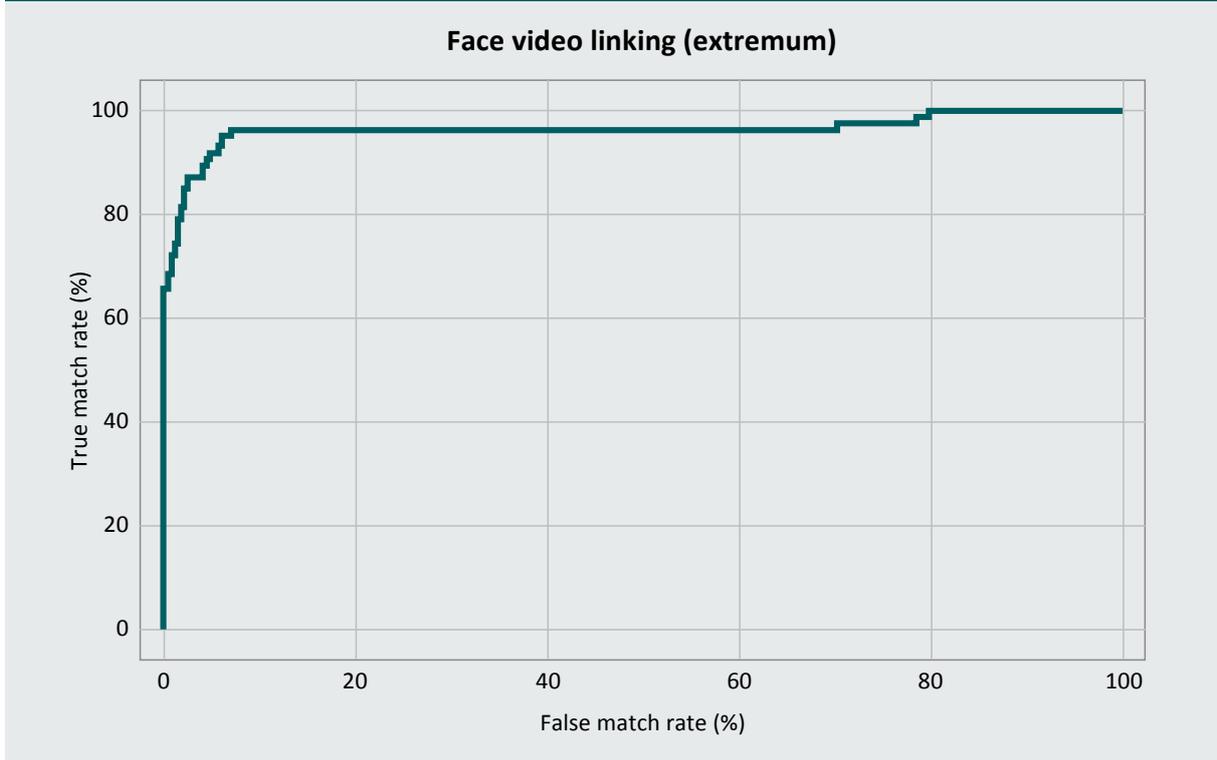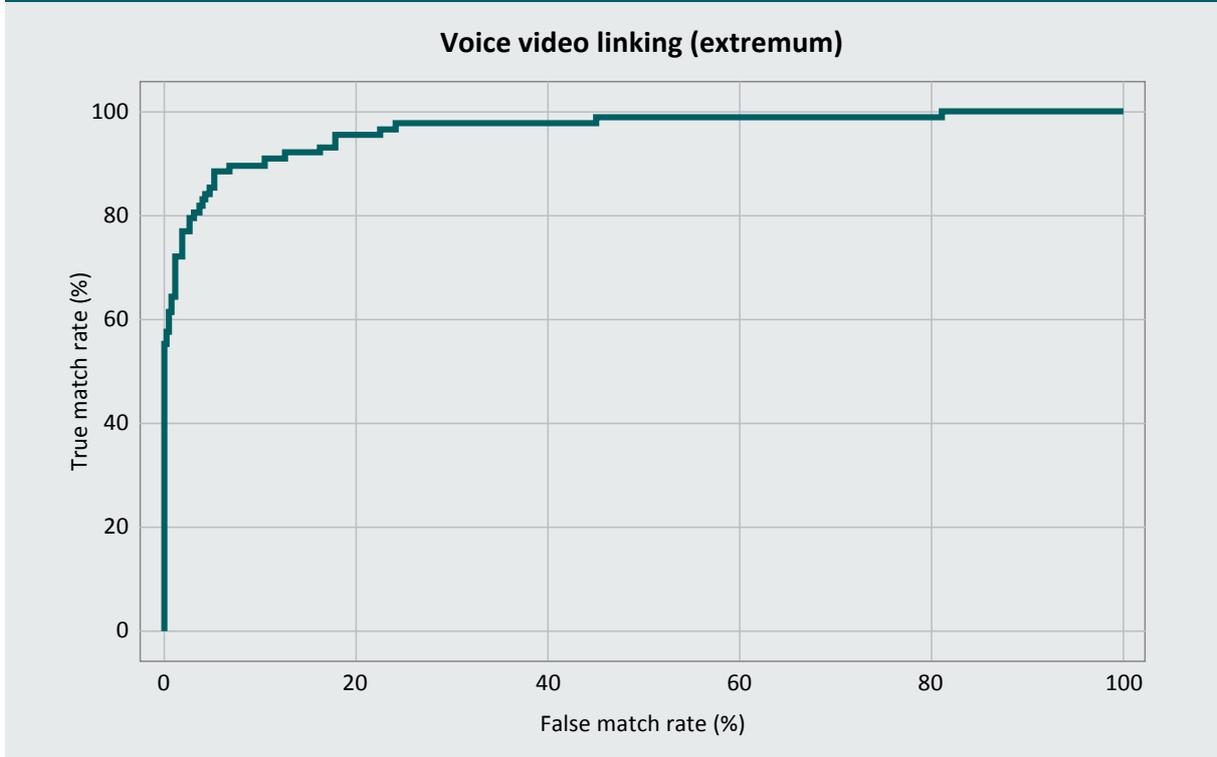**Figure 3: ROC curve for face recognition at all thresholds**

Face video linking (extremum)

Figure 4: ROC curve for speaker recognition at all thresholds

Voice video linking (extremum)

Fusion (combining face and voice)

The respective match scores reported for both face and voice can also be combined to create a fusion match score. Biometric fusion combines match scores from multiple sources to render a single, combined match score (Ross, Nandakumar & Jain 2006). This can be used to improve the error rates observed in Figures 3 and 4.

The first step of fusion involves standardising the face and voice scores, so that each exists in the range from 0 to 1. The final fused scores result from summing corresponding face and voice scores. This has the potential to improve match scores for one biometric attribute by supplementing it with another. Figure 5 demonstrates the utility of this approach, achieving a true match rate of 93.8 percent with a false match rate of 1.0 percent, and a 98.8 percent true match rate with a false match rate of 5.0 percent. Put another way, any observed deficiencies with face recognition (Figure 3) and speaker recognition (Figure 4) are overcome by combining these biometric attributes for each subject.

Figure 5: ROC curve for fusion (face and voice) at all thresholds

Fused (face + voice) video linking (extremum)
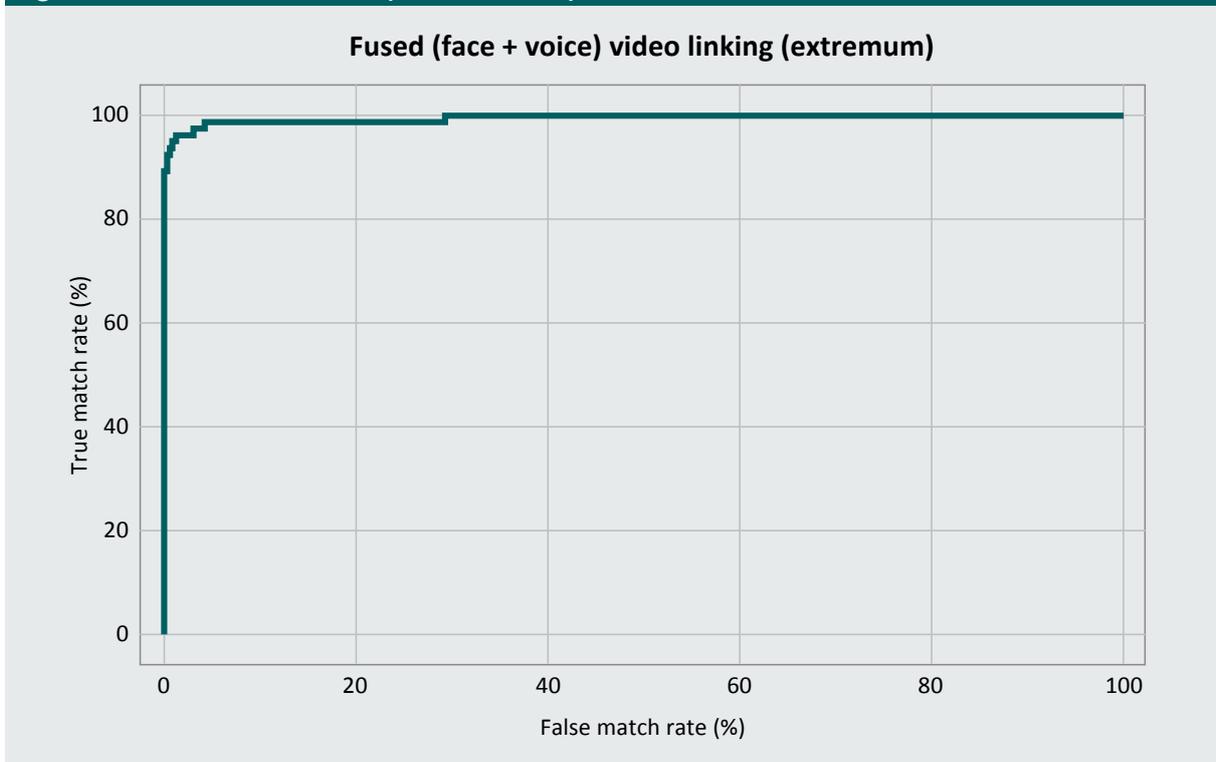
The results underscore the importance of using (fusing) multiple primary biometric attributes for recognition (Chowdhury et al. 2018) and reinforce Bursztein et al.'s (2019) and Sae-Bae et al.'s (2014) call for the use of classification or clustering techniques when matching biometric attributes. This is particularly important in an environment where video quality can be poor or features obscured, thereby limiting single attribute matching performance. Devising automated methods to derive matches using multiple biometric cues has the potential to greatly enhance the depth and scale of investigations beyond the use of hash values (particularly at a time when the volume of CSAM requiring investigators' attention continues to proliferate). Moreover, using automated technologies such as BANE has the added benefit of reducing investigators' direct exposure to CSAM (ie removing the need to manually watch and catalogue media) and can help address the excessive workloads and high degree of burnout frequently experienced and reported in the literature (see Foley, Louise & Massey 2020).

Limitations and directions for future research

BANE represents a first attempt at integrating the automated extraction and matching of multiple biometric attributes across CSA videos. The software development and testing process was challenging, and a number of limitations emerged which should be addressed in future research.

The software requires a high level of computational intensity, which has been accounted for in the software development roadmap. The software was developed to be flexible with respect to its deployment environment, and has been tested on local desktop machines using the central processing unit (CPU), but also in high-performance clusters using graphics processing units (GPUs) for enhanced performance. Future versions of the software will further prioritise performance (including through software optimisation, as well as transitioning to a cloud-based infrastructure) and facilitate large-scale deployment.

The testing database contained selected videos that were specifically included because they contained both a face and voice that were clearly discernible and verifiable. It is acknowledged that videos encountered by practitioners in operational settings may not always contain good quality facial images and/or voice recordings, which may result in degraded performance. Moreover, it is expected that CSAM encountered 'in the wild' will often contain additional variability beyond that portrayed in the videos included in the testing dataset. For example, videos may contain faces and voices of subjects from a wider array of ages (ie from infants and toddlers to teenagers). In addition, videos may also contain multiple speakers or a tremendous range of vocal sounds, including whispers, screams, crying, laughing or singing. Therefore, the creation of larger training and testing CSAM datasets by law enforcement will provide important insights into these contexts and drive the development of new algorithms capable of effectively operating under such conditions.

Some innovative work has already been undertaken with regard to facial recognition, with new algorithms being trained, such as the DSTG's algorithm, included in this study, which was specifically developed to recognise both children's and adults' faces. New methods are also emerging with respect to speaker recognition, permitting speaker disambiguation using novel clustering and deep learning techniques (see Park et al. 2022). However, further development of these and other approaches will likely be required in order to achieve the accuracy necessary to support full-scale operational deployment.

We acknowledge the potential for algorithmic bias in the results presented. The facial recognition algorithm was tested for gender and age bias and found to perform better with children than other algorithms (Yiu, Malec & Michalski 2021). Similarly, the speaker recognition algorithm was tested for gender and language bias; however, this was done with adults only (Chowdhury, Cozzo & Ross 2020). Given the nature of CSAM, we were unable to ensure that various ethnicities and genders were represented equally in our test data. Therefore, it is possible that certain biases could exist. Future research should attempt to test these, and other incorporated algorithms, for such biases.

While videos were required to contain a face and voice to be included in our testing dataset, a proportion of CSA videos being distributed online contain neither a face nor a voice. This suggests a need to extend the software's extraction and matching capabilities to include additional soft and primary biometric attributes, such as vascular patterns, age, gait, gender, hair colour and ethnicity (eg Macedo, Costa & dos Santos 2018; Moser, Rybnicek & Haslinger 2015; Sae-Bae et al. 2014; Yiallourou, Demetriou & Lanitis 2017). Such algorithms can be integrated into future iterations of BANE and may further enhance matching performance (for individual attributes and combinations of attributes).

Beyond the development and integration of additional algorithms, future research can also target deeper integration of BANE with other data collection technologies, such as the web crawler previously developed by the research team, 'The Dark Crawler' (Monk, Allsup & Frank 2015; Westlake, Bouchard & Frank 2017). These technologies, designed to crawl the surface web and darknet, can be coupled with BANE to permit widespread automated detection, collection and analysis of CSAM as it becomes available online. To accomplish this task, software like The Dark Crawler will require further development to enhance video detection and extraction capabilities to account for evolving CSAM preferences and trends. For example, new video formats (such as live streaming) and website implementations have become increasingly commonplace (Cubitt, Napier & Brown 2021; Internet Watch Foundation 2021), and need to be identified and specifically accounted for in future iterations of the software. In developing and deploying these capabilities, researchers and practitioners must also acknowledge their ethical and legal obligations and strictly adhere to best practice (see Brewer et al. 2021; Jain, Klare & Ross 2015; Tanwar et al. 2019).

References

URLs correct as at February 2022

Bourke ML & Craun SW 2014. Secondary traumatic stress among Internet Crimes Against Children Task Force personnel. *Sexual Abuse: A Journal of Research and Treatment* 26(6): 586–609. https://doi.org/10.1177/1079063213509411

Brewer R, Westlake B, Hart T & Arauza O 2021. The ethics of web crawling and web scraping in criminological research: Navigating issues of consent, privacy and other potential harms associated with automated data collection. In A Lavorgna & T Holt (eds), *Researching cybercrimes*. Cham: Palgrave: 435–456. https://doi.org/10.1007/978-3-030-74837-1_22

Brown R, Napier S & Smith RG 2020. Australians who view live streaming of child sexual abuse: An analysis of financial transactions. *Trends & issues in crime and criminal justice* no. 589. Canberra: Australian Institute of Criminology. https://doi.org/10.52922/ti04336

Burns CM, Morley J, Bradshaw R & Domene J 2008. The emotional impact on coping strategies employed by police teams investigating internet child exploitation. *Traumatology* 14(2): 20–31. https://doi.org/10.1177/1534765608319082

Bursztein E, Clarke E, DeLaune M, Elifff DM, Hsu N, Olson L, Shehan J, Thakur M, Thomas K & Bright T 2019. *Rethinking the detection of child sexual abuse imagery on the internet*. World Wide Web Conference, 13 May, pp 2601–2607. https://doi.org/10.1145/3308558.3313482

Canadian Centre for Child Protection 2021. Project Arachnid. https://projectarachnid.ca/en/

Chowdhury A, Atoum Y, Truan L, Liu X & Ross A 2018. *MSU-AVIS dataset: Fusing face and voice modalities for biometric recognition in indoor surveillance videos*. Proceedings of the 24th IAPR International Conference on Pattern Recognition (ICPR), Beijing, pp 3567–3573. https://doi.org/10.1109/ICPR.2018.8545260

Chowdhury A, Cozzo A & Ross A 2020. *JukeBox: A multilingual singer recognition dataset*. Proceedings of Interspeech Conference 2020, Shanghai, China. https://doi.org/10.21437/Interspeech.2020-2972

Chowdhury A & Ross A 2020. Fusing MFCC and LPC features using 1D Triplet CNN for speaker recognition in severely degraded audio signals. *IEEE Transactions on Information Forensics and Security* 15: 1616–1629. https://doi.org/10.1109/TIFS.2019.2941773

Chung JS, Nagrani A & Zisserman A 2018. VoxCeleb2: *Deep speaker recognition*. Proceedings of Interspeech Conference 2018. https://doi.org/10.21437/Interspeech.2018-1929

Council of Europe 2021. *Automated detection of child sexual abuse materials*. Octopus Conference 2021. https://www.coe.int/en/web/cybercrime/workshop-automated-detection-of-child-sexual-abuse-materials

Cubitt T, Napier S & Brown R 2021. Predicting prolific live streaming of child sexual abuse. *Trends & issues in crime and criminal justice* no. 634. Canberra: Australian Institute of Criminology. https://doi.org/10.52922/ti78320

Dance GJX & Keller MH 2020. Tech companies detect a surge in online videos of child sexual abuse. *New York Times*, 20 February. https://www.nytimes.com/2020/02/07/us/online-child-sexual-abuse.html

de Castro Polastro M & da Silva Eleuterio PM 2010. *NuDetective: A forensic tool to help combat child pornography through automatic nudity detection*. 2010 Workshops on Database and Expert Systems Applications, pp 349–353. https://doi.org/10.1109/DEXA.2010.74

Foley J, Louise K & Massey D 2020. The 'cost' of caring in policing: From burnout to PTSD in police officers in England and Wales. *The Police Journal: Theory, Practice, and Principles* 94(3): 298–315. https://doi.org/10.1177/0032258X20917442

Gangwar A, González-Castro V, Alegre E & Fidalgo E 2021. AttM-CNN: Attention and metric learning based CNN for pornography, age and child sexual abuse (CSA) detection in images. *Neurocomputing* 445: 81–104. https://doi.org/10.1016/j.neucom.2021.02.056

Hernandez-Ortega J, Galbally J, Fierrez J, Haraksim R & Beslay L 2019. *FaceQnet: Quality assessment for face recognition based on deep learning*. Proceedings of the 12th International Conference on Biometrics, Crete. https://doi.org/10.1109/ICB45273.2019.8987255

Huang GB, Mattar M, Berg T & Learned-Miller E 2008. *Labeled Faces in the Wild: A database for studying face recognition in unconstrained environments*. Workshop on Faces in 'Real-Life' Images: Detection, Alignment, and Recognition. https://hal.inria.fr/inria-00321923

Internet Watch Foundation 2021. The annual report 2020. https://annualreport2020.iwf.org.uk/

Interpol 2022. Victim identification. https://www.interpol.int/en/Crimes/Crimes-against-children/Victim-identification

Interpol 2018*. Towards a global indicator on unidentified victims in child sexual exploitation material: Technical report*. https://ecpat.org/resource/technical-report-towards-a-global-indicator-on-unidentified-victims-in-child-sexual-exploitation-material/

Islam M, Watters PA & Yearwood J 2011. Real-time detection of children's skin on social networking sites using markov random field modelling. *Information Security Technical Report* 16(2): 51–81. https://doi.org/10.1016/j.istr.2011.09.004

Islam M, Watters P, Mahmood AN & Alazab M 2019. Toward detection of child exploitation material: A forensic approach. In M Alazab & M Tang (eds), *Deep learning applications for cyber security*. Cham: Springer: 221–246. https://doi.org/10.1007/978-3-030-13057-2_11

Jain AK, Klare B & Ross A 2015. *Guidelines for best practices in biometric research*. Proceedings of the 8th IAPR International Conference on Biometrics (ICB), Phuket, Thailand, pp 541–545. https://doi.org/10.1109/ICB.2015.7139116

Macedo J, Costa F & dos Santos JA 2018. *A benchmark methodology for child pornography detection*. 2018 31st SIBGRAPI Conference on Graphics, Patterns and Images (SIBGRAPI), Paraná, Brazil, pp 455–462. https://doi.org/10.1109/SIBGRAPI.2018.00065

Maxim D, Orlando S, Skinner K & Broadhurst R 2016. *Online child exploitation material: Trends and emerging issues*. Canberra: Australian National University Cybercrime Observatory and Office of the Children's eSafety Commissioner. https://doi.org/10.2139/ssrn.2861644

Monk B, Allsup R & Frank R 2015. *LECENing places to hide: Geo-mapping child exploitation material*. 2015 IEEE International Conference on Intelligence and Security Informatics (ISI), Baltimore, pp 73–78. https://doi.org/10.1109/ISI.2015.7165942

Moser A, Rybnicek M & Haslinger D 2015. *Challenges and limitations concerning automatic child pornography classification*. Proceedings of the 10th International Conference on Computer Vision Theory and Applications, Berlin, pp 492–497. https://doi.org/10.5220/0005344904920497

National Center for Missing & Exploited Children 2020. CyberTipline. https://www.missingkids.org/gethelpnow/cybertipline

Park TJ, Kanda N, Dimitriadis D, Han KJ, Watanabe S & Narayanan S 2022. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language* 72. https://doi.org/10.1016/j.csl.2021.101317

Phippen A & Bond E 2020. Image recognition in child sexual exploitation material: Capabilities, ethics and rights. In H Jahankhani, B Akhgar, P Cochrane & M Dastbaz (eds), *Policing in the era of AI and smart societies*. Advanced Sciences and Technologies for Security Applications. Cham: Springer: 179–198. https://doi.org/10.1007/978-3-030-50613-1_8

Powell M, Cassematis P, Benson M, Smallbone S & Wortley R 2015. Police officers' perceptions of their reactions to viewing internet child exploitation material. *Journal of Police and Criminal Psychology* 30(2): 103–111. https://doi.org/10.1007/s11896-014-9148-z

Ross A, Nandakumar K & Jain AK 2006. *Handbook of multibiometrics* vol. 6. Springer. https://doi.org/10.1007/0-387-33123-9

Sae-Bae N, Sun X, Sencar HT & Memon ND 2014. *Towards automatic detection of child pornography*. 2014 IEEE International Conference on Image Processing, Paris, pp 5332–5336. https://doi.org/10.1109/ICIP.2014.7026079

Sanchez L, Grajeda C, Baggili I & Hall C 2019. A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). *Digital Investigation* 29: S124–S142. https://doi.org/10.1016/j.diin.2019.04.005

Seigfried-Spellar KC 2018. Assessing the psychological well-being and coping mechanisms of law enforcement investigators vs. digital forensic examiners of child pornography investigations. *Journal of Police and Criminal Psychology* 33(3): 215–226. https://doi.org/10.1007/s11896-017-9248-7

Srinivas N, Ricanek K, Michalski D, Bolme DS & King M 2019. *Face recognition algorithm bias: Performance differences on images of children and adults*. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, pp 2269–2277. https://doi.org/10.1109/CVPRW.2019.00280

Tanwar S, Sudhanshu T, Kumar N & Obaidat MS 2019. Online signature-based biometric recognition. In MS Obaidat, I Traore I & I Woungang (eds), *Biometric-based physical and cybersecurity systems*. Cham: Springer: 535–570. https://doi.org/10.1007/978-3-319-98734-7_10

Ulges A & Stahl A 2011. *Automatic detection of child pornography using color visual words*. 2011 IEEE International Conference on Multimedia and Expo, Barcelona, pp 1–6. https://doi.org/10.1109/ICME.2011.6011977

Vitorino P, Avila S, Perez M & Rocha A 2018. Leveraging deep neural networks to fight child pornography in the age of social media. *Journal of Visual Communication and Image Representation* 50: 303–313. https://doi.org/10.1016/j.jvcir.2017.12.005

WebRTC 2017. Real-time communication for the web. https://webrtc.org/

Westlake BG, Bouchard M & Frank R 2017. Assessing the validity of automated webcrawlers as data collection tools to investigate online child sexual exploitation. *Sexual Abuse: A Journal of Research and Treatment* 29(7): 685–708. https://doi.org/10.1177/1079063215616818

Yaqub W, Mohanty M & Memon N 2018. *Encrypted domain skin tone detection for pornographic image filtering*. 2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance, Auckland, pp 1–5. https://doi.org/10.1109/AVSS.2018.8639350

Yiallourou E, Demetriou R & Lanitis A 2017. *On the detection of images containing child-pornographic material*. 24th International Conference on Telecommunications, Cyprus. https://doi.org/10.1109/ICT.2017.7998260

Yiu SY, Malec C & Michalski D 2021. *Performance of facial recognition algorithms for the 5RD combating child exploitation network*. DSTG-CR-2021-0160. Edinburgh, Australia: Defence Science and Technology Group

Bryce Westlake is an Associate Professor at San Jose State University.

Russell Brewer is an Associate Professor at the University of Adelaide.

Thomas Swearingen is a doctoral candidate at Michigan State University.

Arun Ross is a Professor at Michigan State University.

Stephen Patterson is a Detective Sergeant for the Joint Anti-Child Exploitation Team, South Australia Police.

Dana Michalski is a Defence Scientist at the Defence Science and Technology Group.

Martyn Hole is a Software Developer at the Defence Science and Technology Group.

Katie Logos is a Postdoctoral Research Fellow at the University of Adelaide.

Richard Frank is an Associate Professor at Simon Fraser University.

David Bright is a Professor at Deakin University.

Erin Afana is a Researcher at San Jose State University.