Apple's CSAM detection technology

Prof. David Forsyth
Computer Science
University of Illinois at Urbana-Champaign

13 July, 2021

Writer bio: I have been a computer vision researcher for over 35 years. I have a D.Phil from Oxford University and have served as a professor at the University of Iowa, the University of California - Berkeley, and the University of Illinois at Urbana-Champaign. I have written many conference and journal papers about computer vision, and I co-authored a standard textbook in computer vision. I wrote one of the first programs for detecting human nudes in pictures, and served on a NRC panel that produced a report on "Protecting Kids from Pornography and other Inappropriate Material on the Internet". I have served as an expert witness in a number of patent disputes.

The technology: Apple has produced a technology that can compute fingerprints from pictures. These fingerprints are very small compared to pictures. When two fingerprints match, it is very likely that the pictures match. Simple operations like resizing, cropping or compressing a picture will not change its fingerprint.

Application of the technology: This technology can be used to detect pictures showing the sexual exploitation of children (CSAM pictures). For example, the National Center for Missing and Exploited Children (NCMEC) keeps a reference collection of such pictures to support investigations and prosecutions. Apple's technology can be used to match fingerprints from a picture library to fingerprints from known CSAM images. The CSAM fingerprints pass through an irreversible blinding step, so that no-one can recover a CSAM image from the matcher. When a user uploads a picture, the device sends the cloud server a doubly encrypted record. Cryptographic methods ensure that, if the server sees doubly encrypted records from enough pictures that do match known CSAM pictures, it will be able to decrypt them and recover the fingerprints and the visual derivatives for only the matches to known CSAM pictures. The visual derivative summarizes the picture so that law enforcement can tell whether the match is correct. The server can only decrypt if there are sufficient CSAM matches.

Likely impact of the technology: The state of the art of detecting CSAM pictures is not public, but in my judgement this system will likely significantly increase the likelihood that people who own or traffic in such pictures (harmful users) are found; this should help protect children. Harmless users should experience minimal to no loss of privacy, because visual derivatives are revealed only if there are enough matches to CSAM pictures, and only for the images that match known CSAM pictures. The accuracy of the matching system, combined with the threshold, makes it very unlikely that pictures that are not known CSAM pictures will be revealed.

Apple has shown me a body of experimental and conceptual material relating to the practical performance of this system and has described the system to me in detail. Based on this material and my experience and judgement, I believe that:

- **Apple sees visual derivatives only for CSAM matches and only if there are enough matches:** I am persuaded that the cryptographic protocol used will ensure this.
- **No user will get direct feedback from the matching system:** This means that users should never notice the operation of the system. Because harmful users get no feedback, they will not be able to find CSAM images that can safely be uploaded.
- **Visual derivatives from harmless user libraries will be seen by Apple very seldom, if ever:** Apple will see visual derivatives only if there are enough matches. The false positive rate (the rate at which the system thinks that pictures match known CSAM pictures when actually they do not) is low. But a single false positive does not result in an alert. Cryptographic protocols mean that Apple sees visual derivatives only if there is an alert. By choosing an appropriate threshold (the number of matches that cause an alert), Apple has forced the false alert rate to be extremely low.
- **It is highly unlikely that harmless users will be inconvenienced or lose privacy** because the false positive rate is low, and multiple matches are required to expose visual derivatives to Apple. Apple will review these potential reports and notify NCMEC if appropriate. Even if there is a false alert, this review will ensure that harmless users are not exposed to law enforcement actions.
- **CSAM images cannot be recovered from anything the system stores on a device:** The irreversible blinding of CSAM fingerprints ensures this.
- **The system will make possessing collections of CSAM pictures in the Apple environment much more risky** because the false negative rate (the rate at which the system thinks that pictures do not match the CSAM collection when actually they do) is moderate. The system is nearly certain to identify any attempt to store many known CSAM pictures.
- **Apple's approach preserves privacy better than any other I am aware of:** Apple receives an encrypted record from the device for every picture. But cryptographic results guarantee that Apple will be able to see visual derivatives only if the device uploads enough known CSAM pictures, and only for the matching pictures. If there are not enough known CSAM pictures uploaded, Apple will be unable to see anything.