Identifying illegal websites in photos

September 1 2015, by Inderscience

European computer scientists have developed a way to "read" web addresses in images that could improve filters for blocking pornographic, gambling and other sites. They provide details in the new issue of the International Journal of Reasoning-based Intelligent Systems.

Internet marketers of all shades might add a website address, a URL, to a graphic or photo that might then be found through an image search engine. The user finding such an image may be interested in visiting said site, but will have to type out the URL into their browser's address bar to do so. Conversely, the URL might point to illicit content - pornography, gambling sites, illegal drugs, terrorist propaganda. In that content, those in authority, whether parents and guardians of children or law enforcement, may wish to automatically blacklist such URLs.

Now, Nikolay Neshov of the Technical University of Sofia, Bulgaria and colleagues at the University of Karlstad, Sweden, and the University of Belgrade, Serbia, have developed a computer algorithm that can detect the presence of text overlaid on to an image or a still from a video, extract the text and convert it into an active URL for accessing or blocking a website.

Simple optical character recognition (OCR) does not work well with text overlaid on images as the background is usually complex, the text is likely to be of lower resolution and lower intensity and contrast than that seen in a scanned document or page, for instance. The new approach uses an identification extraction technique that finds anomalies in an

image that would be present if text is overlaid. It then removes the details surrounding those anomalies leaving just the area occupied by any text - the team calls this the binarisation process. This isolated text image can then be fed into an (OCR) system to convert the image of the text into actual text in the computer.

The team has successfully tested their algorithm on thousands of images with overlaid URLs. They were able to identify 619 URLs from a random selection of 1000 test images at a rate of three per second using their approach. Conventional OCR was faster but only found 83 URLs in the same 1000 images, an improvement from about 8% to more than 60%.

The researchers' initial motivation was to assist computer forensic investigations in which tens of thousands of illegal and illicit photos must be scanned and any associated websites identified quickly in an investigation. This is critical in investigations of child pornography and child sexual abuse, the team reports, but such work is often stymied by the vast numbers of images involved.

Given that internet search companies and other service providers are involved in various initiatives to identify and block illegal material on the internet, this new approach to URL extraction from images could be added to their arsenal of techniques for detecting such content as well as being useful in criminal investigations surrounding said content.

More information: "Finding URLs in images by text extraction in DCT domain, recognition and matching in dictionary." *Int. J. Reasoning-based Intelligent Systems*, Vol. 7, Nos. 1/2, pp.78–92. DOI: 10.1504/IJRIS.2015.070916

Provided by Inderscience