Identifying Abusive Videos By Detecting And Removing Video Overlay Content

Johannes Schuler
IDENTIFYING ABUSIVE VIDEOS BY DETECTING AND REMOVING VIDEO OVERLAY CONTENT

ABSTRACT

Disclosed herein is an improved mechanism for identifying abusive videos by detecting and removing video overlay content. The mechanism can begin by detecting whether an overlay is present within multiple frames of a video. For example, the mechanism can detect the presence of an overlay by determining gradients between multiple frames of a video and can estimate that the overlay that has been applied to one or more frames of the video based on a median of the determined gradients. For the frames of the video, the mechanism can then remove the overlay from the frames of the video. For example, the mechanism can estimate a clean frame of the video in a gradient-space by subtracting the estimated overlay from an original frame of the video. In continuing this example, the mechanism can then reconstruct a clean image corresponding to the estimated clean frame of the video in gradient-space. Upon obtaining the reconstructed clean image in which the overlay has been removed, the mechanism can apply a classifier to the reconstructed clean image to determine whether abusive content may be present in the reconstructed clean image.

BACKGROUND

User-generated videos are frequently uploaded to media content sharing services for other users to download or stream. Some content uploaders may upload a video that includes abusive content, such as objectionable content (e.g., pornography, violent content, content with graphic language, etc.) or copyrighted content. Media content sharing services may apply various algorithms to detect videos that include abusive content. However, in some cases, an uploader may circumvent detection of such a video by applying a semi-transparent overlay over the video. Thus, there is a need for an improved mechanism for identifying abusive videos.
DESCRIPTION

The systems and techniques described in this disclosure relate to identifying abusive videos by detecting and removing video overlay content. The system can be implemented on a server, such as a server associated with a media content sharing service. FIG. 1 illustrates an example process for identifying abusive videos. In particular, FIG. 1 illustrates an example process for identifying a video containing abusive content, where at least some portions of the video have been obscured with a semi-transparent overlay.

FIG. 1

102
Determine gradients between multiple frames of a video

104
Estimate an overlay applied to frames of the video by determining a median of the gradients

106
Estimate a clean frame of the video in a gradient-space by subtracting the estimated overlay from the original frame

108
Reconstruct a clean image corresponding to the estimated clean frame in the gradient-space

110
Apply a classifier to the clean image to determine whether abusive content may be included in the clean image
At 102, the system can begin an overlay estimation approach by determining gradients between multiple frames of a video. In some instances, the system can determine gradients between any suitable number of frames of the video (e.g., ten, twenty, one hundred, and/or any other suitable number). In some instances, the video can be any suitable video, such as a video that has been identified as being recently uploaded to a media content sharing service, a video that has received over a particular number of views, and/or any other suitable video.

At 104, the system can estimate an overlay applied to the frames of the video by determining a median of the gradients. Note that, in some instances, a median of gradients of multiple images that each contain a static image, such as an overlay or watermark, may converge to the static image, or overlay, itself. Additionally, note that, in some instances, the estimated overlay can correspond to any suitable spatial portion of a frame of the video. For example, in some instances, the estimated overlay can span an entire portion of a frame of the video.

An example of an equation for estimating an overlay applied to $k$ frames of a video can be represented by:

$$\text{gradient}(\bar{W}(p)) = \text{median}_k(\text{gradient}(J_k(p))).$$

where $W$ is an applied overlay, $J$ is a frame of the video with an applied overlay, and $p$ is a pixel location. Note that, in some instances, the median of the gradients $J_k$ can be computed independently in the x and y directions at every pixel location $p$. Additionally, note that, in some instances, the estimated overlay can be in a gradient-space, as given by the above equation.

At 106, the system can begin an overlay removal approach by estimating a clean frame of the video (that is, a frame of the video that does not include the overlay) in the gradient-space by subtracting the estimated overlay from a frame of the video. That is, in some instances, the system can determine an estimate of gradients of a clean frame of the video.
At 108, the system can then reconstruct a clean image corresponding to the estimated gradient-space clean frame of the video. In some instances, the system can reconstruct the clean image using any suitable technique or combination of techniques. For example, in some instances, the system can use any suitable optimization algorithm to reconstruct the clean image. As a more particular example, in some instances, the optimization to reconstruct the clean image can be solved in Fourier space.

Note that, in instances where the video (and, therefore, the images corresponding to frames of the video) are full-color images, the system can perform any suitable pre-processing prior to reconstructing the clean image. For example, in some instances, the system can split the estimated clean frame of the video linearly into YUV channels prior to reconstructing the clean image. In some such instances, each channel of the YUV channels can be reconstructed separately, and the system can then combine the portions of the reconstructed image corresponding to each of the YUV channels to generate a full reconstructed clean image.

At 110, the system can apply a classifier to the reconstructed clean image to determine whether abusive content may be included in the clean image. In some instances, the classifier can be any suitable classifier, such as a classifier that generates an output that indicates whether an input image is likely to contain particular types of objectionable content (e.g., pornography, violent imagery, and/or any other suitable type of objectionable content) or protected content (e.g., copyrighted content, and/or any other suitable type of protected content). In some instances, such a classifier can use any suitable type of algorithm (e.g., a logistic regression, a neural network, and/or any other suitable type of classification algorithm). Additionally, in some instances, such a classifier can be trained using any suitable training set, such as videos or images that have previously been manually annotated as containing particular types of content.
It should be noted that, in some instances, the classifier can be used to calculate a score of the frames of the video. For example, the classifier can generate an output score that indicates whether the frames of the video are likely to contain particular types of objectionable content (e.g., a pornography score, a violent imagery score, etc.).

Note that, in some instances, the system can perform any suitable action(s) in response to determining that the reconstructed clean image, and therefore, the video, is likely to contain particular types of abusive content. For example, in some instances, the system can cause the video to be removed from a media content sharing service to which the video was uploaded. In a more particular example, in some instances, the system can automatically remove videos from a media content sharing service in which the output score from the classifier (e.g., a pornography score, a violent imagery score, etc.) is greater than a threshold content class score. As another example, in some instances, the system can transmit a message to a creator of the video indicating that the video has been flagged as likely to include abusive content. As yet another example, in some instances the system can transmit a notification to a human evaluator indicating that the video is to be manually reviewed.

Accordingly, a mechanism for identifying abusive videos by detecting and removing video overlay content is provided.