March 2021

Finding Match Avoidance Attempts At Scale With Video Expansion

Roman Vorushin

Philipp Neubeck

Hanna Maria Pasula

Filip Pavetić

Conrad Burchert

Finding Match Avoidance Attempts At Scale With Video Expansion

ABSTRACT

An important objective of user-generated content platforms such as audio/video hosting or streaming platforms is to ensure that content that is available via their platforms is authorized for use, e.g., is provided by the true owner or with due permission of the true owner. To ensure that unauthorized content is not made available, such platforms match uploaded videos against a repository of reference (original) videos. To avoid video content being matched, content uploaders utilize constantly evolving new content transformation strategies when uploading unauthorized content. This disclosure describes automated techniques that help speed up and scale the collection of training examples of recent techniques of content transformations designed to bypass match detection procedures. These include synthetic generation (automatically generating content examples similar to match avoiding content) and scaled up mining and filtering (which includes performing searches for other content that is similar to match avoiding content on some dimension and filtering such content using high performance matching algorithms) to detect other examples of similar match avoiding content. The corpus of data generated by the described techniques can be used to train and validate a new version of matching procedures that is robust to the recent match avoidance attempts.

KEYWORDS

- User-generated content
- Video expansion
- Video search
- Content transformation
- Content match
- Content fingerprinting
- Match avoidance
- Synthetic training data

BACKGROUND

An important objective of user-generated content platforms such as audio/video hosting or streaming platforms is to ensure that content that is available via their platforms is authorized for use, e.g., is provided by the true owner or with due permission of the true owner. Unauthorized copies of original video content are disallowed. To ensure that unauthorized content is not made available, such platforms match uploaded videos against a repository of reference (original) videos, e.g., from content owners. The matching procedure returns a set of reference videos whose audio and/or video content matches portions of the newly uploaded video. The matching procedure enables content owners to retain control of their content on such platforms.

The matching procedures are engineered to high performance (handle very large numbers of videos) and for high precision and recall. The procedures are robust to various content modifications such as encoding, transcoding, and background noise; frame rate differences; aspect ratio differences; contrast and palette shifts; cropping; overlays; etc. Automated matching techniques typically utilize neural-network (or other artificial intelligence) techniques that create embeddings (vector representations) of video content such that a video and its transformation, although dissimilar in the Euclidean sense, are close in the embedding space. Machine learning models used for matching are trained to detect content matches even in the presence of a variety of content transformations, e.g., videos that are partially obfuscated, cropped, changed in color, blurred, overlaid with text, occupy smaller sections of the screen, etc.

To avoid video content being matched, content uploaders utilize constantly evolving new content transformation strategies when uploading unauthorized content. For example, if one mode of content transformation, e.g., framing unauthorized content in a changing background, is

reliably detected by matching procedures, uploaders use other types of content transformation to fool matching procedures, e.g., sprinkling content with dust particles or firework displays embedded within the video. Automated content matching therefore operates in an adversarial environment.

To train a new version of the matching procedures that is robust to new types of content transformations may require a large number, e.g., thousands of training examples of such new transformations. It can take a long time to collect the required number of training examples to reach high precision/recall. Every unauthorized video that is uploaded in the meantime can negatively impact the genuine content owner and the platform. It is therefore valuable for platform owners to have content matching techniques that can adapt quickly to new types of content transformations.

DESCRIPTION

This disclosure describes techniques to speed up and scale the collection of training examples of recently used content transformations that successfully bypass matching procedures. Content transformations designed to bypass matching procedures can generally be classified as follows.

1. A video with a new *video* transformation is uploaded, such that video matching cannot match it properly, while audio matching only creates short matches that aren't sufficient to automatically classify the video as an unauthorized content upload.

Fig. 1: A video transformation that deliberately attempts to bypass matching procedures: (a) Original video; (b) Transformed video

An example of such a transformation that deliberately attempts to bypass matching procedures is illustrated in Fig. 1. An original video (102) is transformed such that it occupies a smaller section of the video (104) which includes a large background portion of moving patterns (106). The presence of dominant moving patterns may cause content matching techniques to fail to detect that the video is a match for prior content.

2. A video with a new *audio* transformation, e.g., inserted noise or clicks, such that audio matching techniques cannot detect a match, and video matching only creates short matches that aren't enough to claim the video automatically.
3. A video with new *video* and *audio* transformations, such that neither video nor audio matching can establish a reliable match.

In the cases listed above, a match is sometimes manually detected, e.g., by human reviewers and a report is generated. Such a manual detection report includes a content identifier of the uploaded (and suspected unauthorized) video; the content identifier of the corresponding genuine content,

a portion of which was found in the uploaded video; and the start/end times for both videos where the human reviewer detected a match.

When such a manual report is received, it can be used by matching specialists (video reviewers) to manually generate more examples of its content transformation by examining more videos uploaded from the same account; obtaining such examples from other teams or content partners; or by running fine-tuned search queries. As explained before, such manual techniques of training example generation have difficulty keeping pace with content transformations introduced by malicious users.

Per the techniques described herein, a manual report identifying a video with content transformations that were not detected by the automated content matching procedures triggers the generation of training or validation data in one or more of the following ways:

Synthetic generation

Training videos are manufactured by transforming ordinary video content in a manner similar to new content transformations recently observed. For example, if a new content transformation is found to include large background portions of moving patterns (as illustrated in Fig. 1) or involve peppering videos with floating dust particles, fireworks, etc., then training (or validation) data is manufactured by modifying reference videos to include similar transformations.

Scaled-up mining and automated filtering

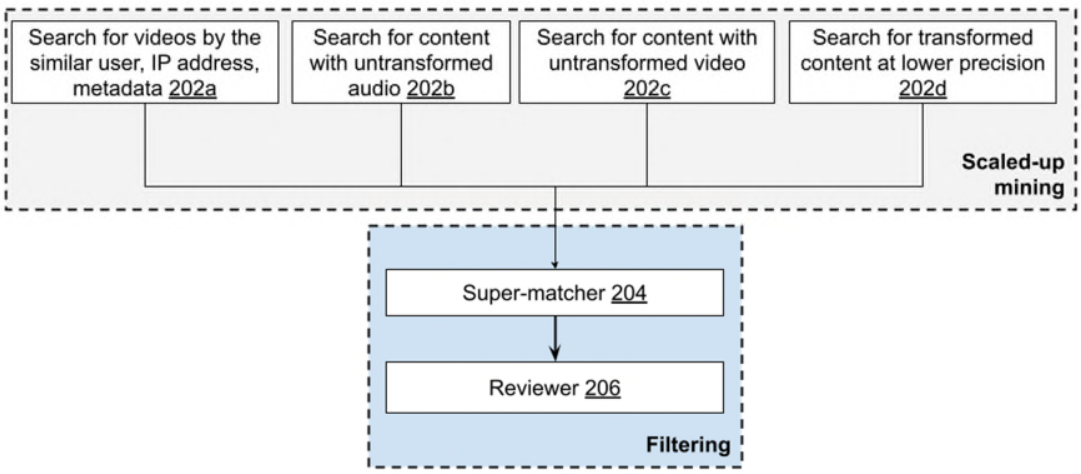

Fig. 2: Scaled up mining and automated filtering

Fig. 2 illustrates scaled-up mining of training data and automated filtering. Videos that share attributes with a video identified in the manual report, e.g., from the same or similar accounts/ IP addresses/ metadata or titles, etc. are searched for and identified (202a).

Scaled-up mining can also leverage the transformation of any one (video or audio) channel of the content but not the other. For example, if the video portion of the content is transformed in a certain manner but not the (entire) audio, then mining for content with the (untransformed) audio can reveal more examples of similarly transformed videos (202b). Conversely, if the audio is transformed in a certain manner but not the (entire) video, then mining for content with the (untransformed) video can reveal more examples of similarly transformed audio (202c).

Scaled-up mining can also be performed by using versions of audio and video matching procedures that have lower precision but high recall or with matching procedures with augmented computational resources (202d). Scaled-up mining in this manner can automatically

find more matches between the uploaded video and other content, but, by its nature, can include benign (untransformed) videos in addition to videos that have been transformed to avoid detection. Matches found by scaled-up mining are referred to as potential (rather than confirmed) matches; confirmed matching is obtained via filtering, further described below.

Content identified by scaled-up mining is flagged, stored in a data store, and forwarded to super-matchers (204), which winnow out match-avoiding (transformed) videos from benign (untransformed) videos. Super-matchers are relatively expensive computational engines that are optimized for video/audio content matching. As opposed to the larger-scale production grade matching procedures (tuned for high throughput and efficiency), super-matchers are designed to run on relatively smaller amounts of content, but more thoroughly. Such increased thoroughness of matching can be achieved, for example, using neural networks of greater breadth and depth, larger feature vectors, recognition parameters, embeddings, resolution, etc. A super-matcher can also, for example, obtain a match by applying a reverse transformation to a suspected content transform. The output of super-matchers, which include examples of recent content-transforms that bypass existing matching procedures, is reviewed by human reviewers (206).

The corpus of data generated by the above-described (synthetic generation and/or scaled-up mining and automated filtering) techniques is utilized to train and validate a new version of matching procedures that is robust to the newly identified match-avoiding content transformations (e.g., the moving patterns of Fig. 1, or other types of transformations). For example, segments of transformed content (generated or mined as described above) and original content can be used to train an edition of matching procedures. The generation of a training data set is repeated for new match avoiding strategies that are detected.

While the foregoing discussion refers to content uploads, the described techniques can be utilized to detect any audio/video content that includes transformations designed to avoid detection by automated matching techniques. For example, detection of violating content in streaming platforms that support live audio/video streaming from cameras, computer games, live audio, etc. can also be improved using the described techniques. User-generated content platforms such as social networks, hosting platforms for podcasts and other audio, streaming content platforms, live-audio platforms, etc. can utilize the techniques.

CONCLUSION

This disclosure describes automated techniques that help speed up and scale the collection of training examples of recent techniques of content transformations designed to bypass match detection procedures on user-generated content platforms. These include synthetic generation (automatically generating content examples similar to match avoiding content) and scaled up mining and filtering (which includes performing searches for other content that is similar to match avoiding content on some dimension and filtering such content using high performance matching algorithms) to detect other examples of similar match avoiding content. The corpus of data generated by the described techniques can be used to train and validate a new version of matching procedures that is robust to the recent match avoidance attempts.