Detecting Ads in Video Streams Using Acoustic and Visual Cues

Michele Covell and Shumeet Baluja, Google Research
Michael Fink, The Hebrew University of Jerusalem

A new technique provides accurate, efficient ad detection for online TV rebroadcasts.

As online television grows in popularity, providers are seeking cost-effective ways to replace original advertisements in rebroadcasts with new ads that are not only fresher but that can target individual viewers' interests and preferences. Alternatively, completely removing commercials might be desirable for subscribers who pay for reused content.

Researchers have developed various heuristic techniques to detect and remove ads in video streams. Many of these exploit common differences between advertising and program material including cut rates, soundtrack volume, and surrounding black frames (X-S. Hua, L. Lu, and H-J. Zhang, "Robust Learning-Based TV Commercial Detection," *Proc. IEEE Int'l Conf. Multimedia and Expo,* 2005, http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1521382).

However, such techniques often fail to remove some types of commercials, especially self-advertisements of upcoming programs. In addition, many break down when applied to non-US programming or content that does not adhere to professional standards.

REPETITION-BASED AD DETECTION

John M. Gauch and Abhishek Shivadas developed a repetition-based approach to advertisement detection that first segments video streams approximately at shot boundaries and then looks for visually similar segments ("Identification of New Commercials Using Repeated Video Sequence Detection," *Proc. IEEE Int'l Conf. Image Processing,* 2005; http://ieeexplore.ieee.org/xpl/freeabs_all.jsp?arnumber=1530626).

An advantage of this approach is that the percentage of repeated ads grows with the size of the video collection: Multiple airings of the same ad are more likely to occur in processed video streams that include more stations. Thus, unlike many current ad-detection systems, which are agnostic to the amount of material processed, this technique improves performance with increasingly large amounts of data.

Drawing on Gauch and Shivadas' work, we have developed a technique that uses acoustic as well as visual cues to first efficiently find and then segment repeated material within a large collection of monitored video streams ("Advertisement Detection and Replacement Using Acoustic and Visual Repetition," *Proc. IEEE Int'l Workshop Multimedia Signal Processing,* 2006, www.mangolassi.org/covell/pubs/mmsp06_ads.pdf).

As Figure 1 shows, this process consists of three steps:

- matching repeated audio frames via a robust hashing method;
- verifying matches using visual cues; and
- precise segmenting using dynamic-programming methods, followed by simple heuristics to sort advertisements from repeated program material.

We applied this process to a sample collection of four days' worth of video footage (three days from one TV station, one day from a different station). This footage includes 1,348 minutes of repeated advertisements and 487 minutes of repeated programs. The repeated ads comprise 87.5 percent of all ads in the sample.

MATCHING REPEATED AUDIO FRAMES

The first step in our process is to find repeated audio frames within a monitored video stream. As Figure 1a shows, our approach accomplishes this by simply chopping the stream up into short nonoverlapping chunks, with no content analysis. This avoids the computational overhead of presegmentation. We use a chunk length of 5 seconds, which is half of the shortest duration expected for a commercial. This content-independent segmentation ensures that at least one probe will lie fully within each ad.

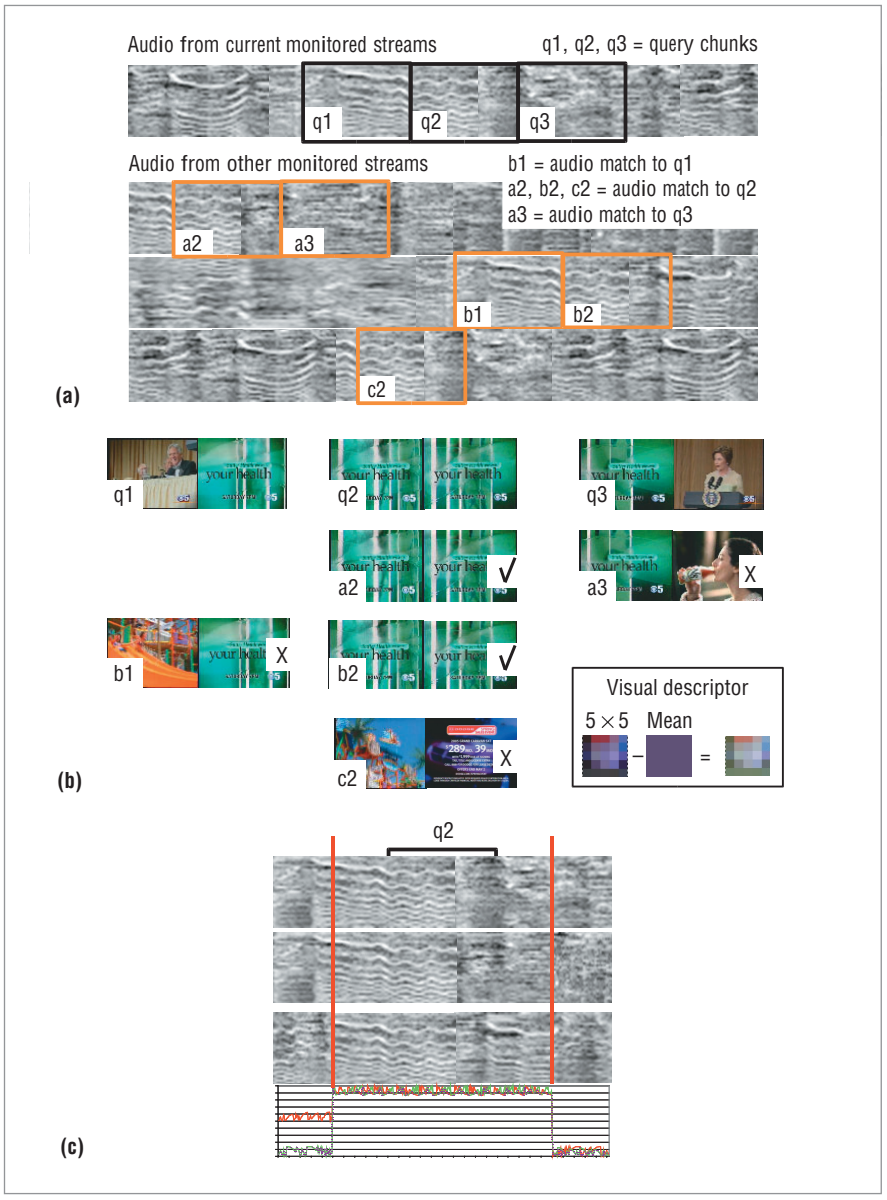

Figure 1. Ad detection, verification, and segmentation process: (a) match 5-second audio chunks within/across monitored video streams; (b) validate candidate matches using video-frame fingerprints; (c) refine temporal segmentation to 11-ms resolution.

To create an initial list of candidate matches based on the audio signal, we use a music retrieval system developed by Yan Ke, Derek Hoiem, and Rahul Sukthankar that identifies compact, discriminative audio features via an AdaBoost-based machine-learning method (www.cs.cmu.edu/~yke/musicretrieval). The system extracts these features from the outputs of 32 filters that operate on the spectrogram as if it were an image, thresholding outputs to retain only one bit per filter at each 11-ms time step.

Each 32-bit descriptor serves as an independent hash key into a video-collection database's audio component. The system also indexes nearby neighbors using all variations of each hash key within a Hamming distance of two. This retrieval generates a list of temporal channel offsets between the 5-second query chunk and other parts of the collection. The final list passed on by step 1 includes those match offsets supported by retrievals from many 11-ms slices within the current 5-second chunk.

Using nonoverlapping 5-second queries on our four-day video collection, we obtained 92 percent recall and 87 percent precision rates. These results are less accurate than those reported by Ke and colleagues for music identification. Because we made no effort to prealign query boundaries with content boundaries, about one-sixth of the queries straddled match-segment boundaries, and many of the false positives and false negatives (27 percent and 42 percent, respectively) were on these boundary cases.

The third step in our process corrects both overmatching (matching beyond the true boundary of the repeated ad) and undermatching errors.

VERIFYING MATCHES WITH VISUAL CUES

Many audio-matching mistakes occur on segments that contain stock music without voice-overs. Step 2 of our procedure removes these coincidental matches as well as some boundary-straddling matches, using visual verification as shown in Figure 1b. We found that, when the audio tracks serendipitously match, the visual channels are easily distinguishable.

The database uses simple 5 × 5 24-bit RGB thumbnails, taken on each 10th frame (approximately three per second), to represent the candidate sequence. To help eliminate systematic transmitter/receiver distortions, we subtract the color-image mean.

The system creates visual thumbnails for the probe snippet on the fly to correctly align its sampling with that imposed by the database thumbnail sampling, as seen through the match-alignment offset that the music-retrieval system provides. Given these normalized thumbnails, the system simply compares the pixels and thresholds the differences; those below the difference threshold proceed to the next step.

Our system hypothesizes that matches passing both acoustic and visual consistency checks are parts of commercials. When applied to our four-day video collection, the visual-verification step improved precision to

92 percent—a 40 percent improvement relative to acoustic matching only—but reduced recall to 93 percent, a 10 percent relative decrease.

REFINING PRECISION THROUGH SEGMENTATION

At this point in the process, the match boundaries only coarsely locate the advertisement boundary due to the 5-second-chunk granularity, leading to occasional over- and undermatching. Step 3 of our procedure corrects both of these shortcomings by combining fine-grained acoustic match confidences across all matching pairs and then detecting end points on these temporal profiles, as Figure 1c shows.

For each set of match candidates, the system collects a list of all the times and channels to which it matched, both acoustically and visually. It forces this multiway match to share the same start and end points, as measured from the center of the query probe and its matches.

The system uses the minimum match similarity of each 11-ms slice within the list to create a single profile of fine-grained match scores for the full list. This increases segmentation accuracy when transitions to or from some of the airings of the ad are low-volume or silent and is yet another aspect of our approach that benefits from processing increasingly large amounts of video.

To find the end point of the ad segment, we use forced Viterbi alignment (B. Gold and N. Morgan, *Speech and Audio Signal Processing: Processing and Perception of Speech and Music,* Wiley & Sons, 1999) starting from the center of the snippet match and running forward in time. We also used this alignment starting from the center of the snippet match and running backward in time to locate the segment's beginning. In each case, we use a two-state first-order Markov model to locate the optimal transition point from "matching" to "not matching," given the minimum-similarity profile. The Viterbi decoder runs forward or backward from the match center for 120 seconds. If the full match profile, from the detected starting point to the detected end point, extends for 2 minutes or more, it is most likely a repeated program.

Since the system is unlikely to be matching advertisements over such a long period, we can safely remove that overlong match from consideration. Otherwise, the system uses the location indicated by the decoding variable as the transition point, ensuring that it is using the optimal starting/end point for the segments. Finally, if the duration given by combining the optimal starting and end points is too short (less than 8 seconds), we also discard the match list as being simple coincidences.

Using segmentation to recover the fine-grained advertising boundaries, we achieved a precision rate well above 99 percent and a 95 percent recall rate. The higher precision was due to the use of the minimum similarity profiles to determine repetition, while the increased recall rate was attributable to the match profile from neighboring matches correctly extending across previously missed matches on straddled segment boundaries (ad/ad or ad/program). The latter recovers the loss in recall that the visual-verification step introduces and even improves on the original acoustic-matching results.

Our results far exceed those of heuristic-based approaches and are directly applicable to non-US programming. Another benefit of our technique is that it automatically identifies commercials without the need for advertisers or content publishers to insert special signals, add extraneous annotations, or otherwise modify their production process. Ultimately, our goal is to create a database of known advertisements and to detect these ads in future broadcasts.

Michele Covell *is a staff research scientist at Google Research. Contact her at covell@google.com.*

Shumeet Balija *is a senior staff research scientist at Google Research. Contact him at shumeet@google.com.*

Michael Fink *is a PhD student at the Interdisciplinary Center for Neural Computation, the Hebrew University of Jerusalem. Contact him at fink@huji.ac.il.*

Editor: Bill Schilit, schilit@computer.com