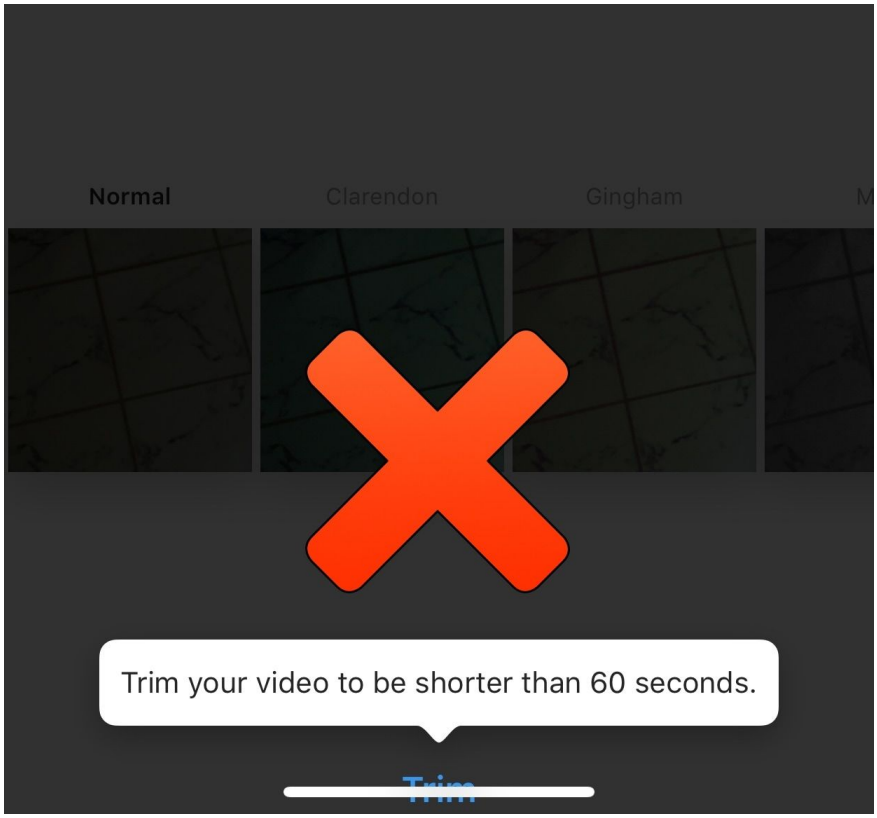
MULTIMEDIA SYSTEMS

MUSIC VIDEO TEASER GENERATION

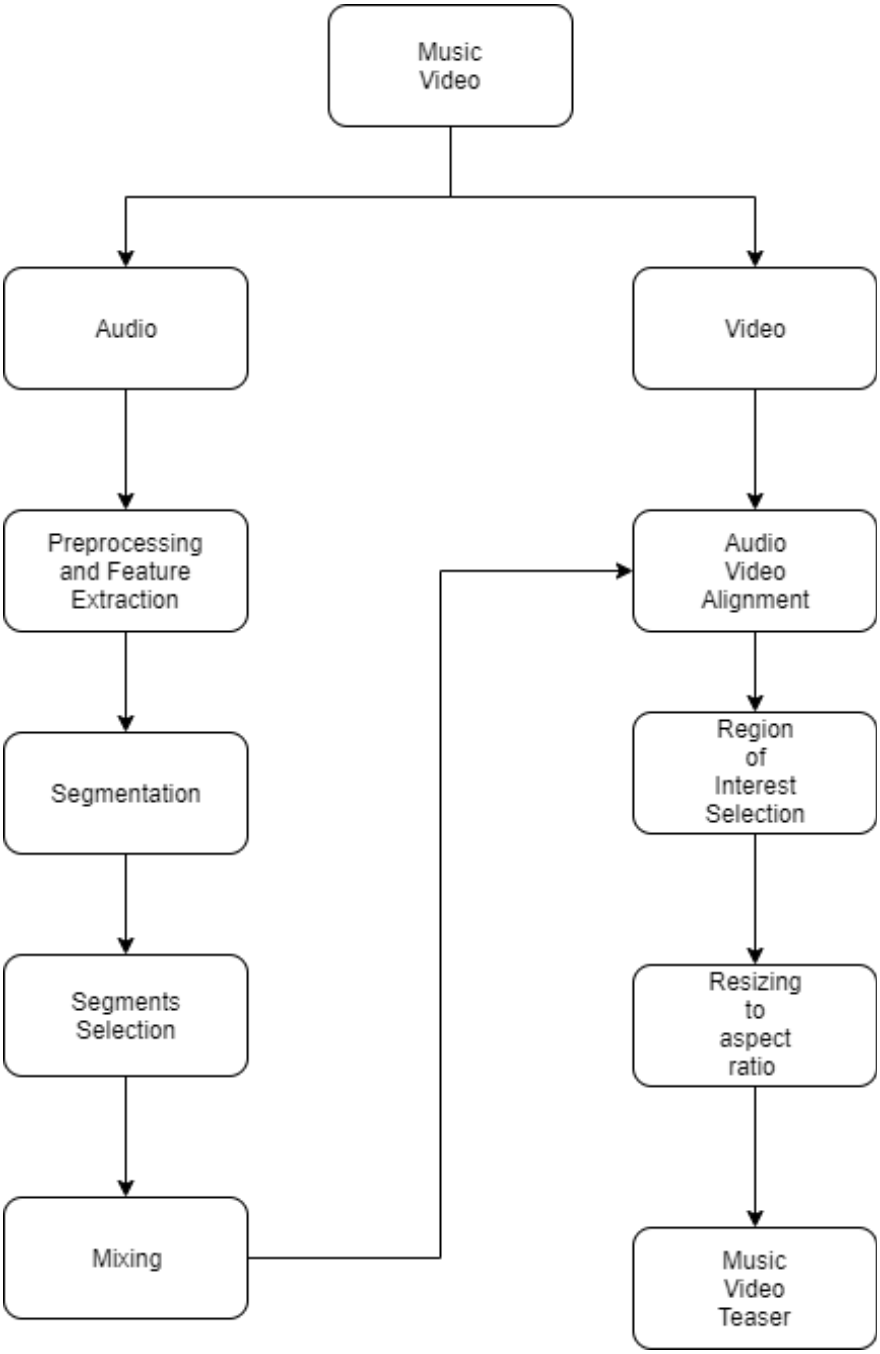
ABHISHEK GOYAL, CHIRAG KHURANA



Introduction

Before the release of any music video by a popular artist, a teaser is first released on social media applications like facebook and instagram to generate excitement for the official video release. This teaser is edited from the original video containing just some specific parts of it. The teaser generated adheres to instagram's maximum 1 minute duration limit and aspect ratio of 1:1. This program aims to automate this teaser generation process given any music video. Given an input music video of any length, the program generates an output video of roughly 1 minute in length, cropped out with relevant features displayed in each frame of the teaser. The audio segments mixed for the final output also are chosen in such a way that same repeating section does not appear in the teaser.

Process Involved



First the music video is split into separate video and audio components. The main goal initially is to determine the audio section of the final teaser clip. First the audio is preprocessed and then segmentation is performed on the audio to receive multiple segments ranging from 5-20s from the initial clip. Particular segments are selected from these and then mixed together to create the final audio section of the teaser.

Then the audio is aligned with the corresponding video. In the video section, region of interest for particular frames is detected and then a 1:1 cropping is performed around each of these regions to obtain the final music video teaser.

Segmentation

The segments to be obtained must be distinctive and should ideally correspond to different sections of the song, namely intro, chorus, verse, bridge. This way the teaser is not repetitive as per its audio content.

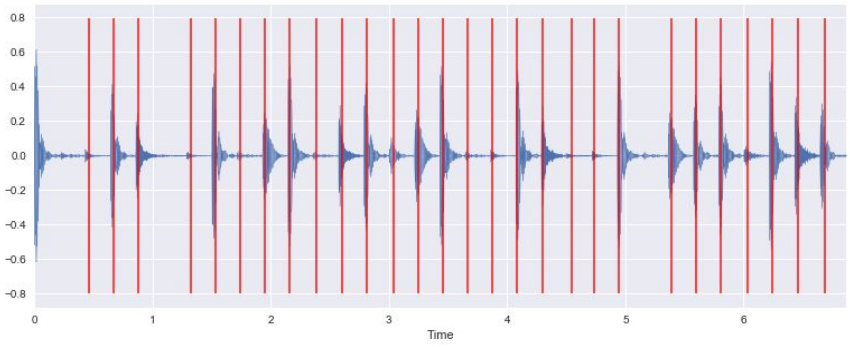
For such segmentation, first some audio features, namely spectral flux, linear predictive cepstral coefficients, zero crossing rate, cepstral flux are extracted from the audio frames. Then these frames are clustered by agglomerative clustering. The clusters obtained should ideally point to different sections of the song. Thus from selective clusters, some audio section (at least 15s in length for the middle sections) is extracted for final audio segments.

Example: Segments obtained for music video of the song Sham:

```
[(0, 28.258684807256238), (81.98965986394558, 101.05324263038548),
(116.40163265306123, 145.42657596371882), (191.49496598639456, 196.19)]
```

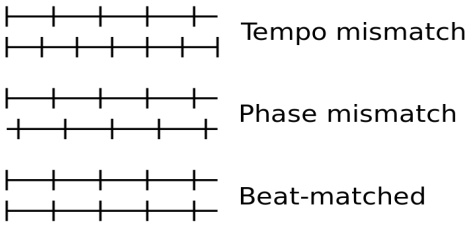
Mixing

After segmentation, the selected audio clips are mixed together in such a way to ensure that the transition is auditorily pleasing. For mixing, first beat tracking is performed to detect the beats in the song.



For detecting beats, the locations of sudden bursts of sound (onset) is detected and to remove false positives, the longest common subsequence of these onsets is calculated to identify the actual beats. This way the tempo of the song (usually corresponding to 4 beats duration) is also calculated.

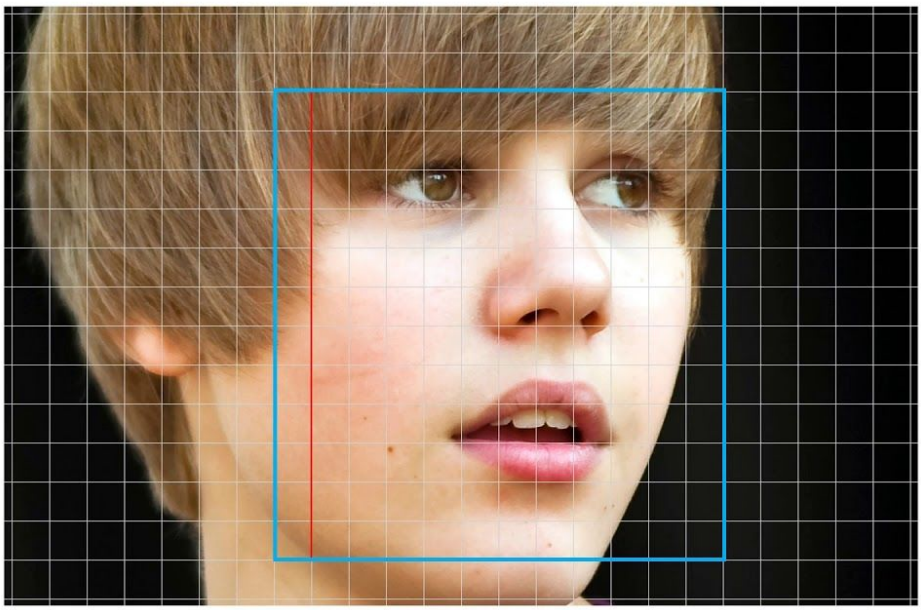
For mixing, the segments earlier are first refined using the tempo of the song. Two segments are then mixed with each other employing a cross fade, with the audio of the first clip fading out and the second fading in with a duration lasting one 4 beat time. Thus the last 4 beats of first clip and first 4 beats of second clip are overlapped with each other to ensure smooth transitions.



This way, each segment ends at a beat multiple of 4 and the next starts at another such multiple with the last 4 beats overlapped between them. As almost all music in the world is produced with this 4 beat structure, the mixing should work perfectly for any such music.

Region of Interest Selection

After aligning the obtained audio section with the video, the region of interest in each frame of the video has to be identified for proper 1:1 cropping. For this ROI selection, face detection is performed at the first frame of each scene (a scene means any original video cut). For face detection, Haar Cascade deep learning algorithm is applied on the video frame.



After detecting all the faces in the given frame, the optimal cropping centre is considered as the centroid of all these face coordinates. This way the cropping box is decided upon which

remains fixed for each scene. If no faces are detected in the frame, then the centre of the video is taken as the cropping box. Thus the cropped video remains focuses on the faces to preserve the video details.

Scene Detection

For optimal cropping, the box is decided at each transition or cut in the video frame. This transition between scenes is detected by using a content-aware detector which compares the current and past frames and determines the scene changes if the difference in the frame contents is above a defined threshold.

Final Video Generation

Finally, according to the cropping boxes, the video is cropped into a 1:1 output. This output also corresponds to just the audio segments generated earlier. This is slightly edited with some video effects for the final teaser.

Output Results

Since the teasers generated can only be qualitatively measured, evaluation is only possible through observation with no set evaluation metric. The program was used to generate a good amount of music videos and it was observed that a very decent amount of them could pass off as an original teaser. Specifically the mixing component worked perfectly with almost zero jarring audio transitions observed. The video cropping led to some tiny visual jinks observed in some cases. It was observed that these could easily pass off as an artistic teaser edit feature and not a technical fault.

The segmentation component did lead to some faults in a small amount of the tests, with too small segments or some arbitrary segments obtained. Otherwise the teasers generated were quite appropriate for social media uploads.