Automatic Video Alignment Through Audio Analysis

Mark Koh · mkoh@drexel.edu
Advised by Dr. Youngmoo Kim
Drexel University, Department of Electrical and Computer Engineering

Introduction

With the vast improvement in smartphone video technology in recent years, multiple videos of the same event (e.g.: a music concert) are becoming increasingly more common. If we are able to take a large collection of videos from an event, we could potentially allow people to relive the concert or event. By identifying precisely where videos overlap, we are able to align these videos and combine them to recreate the original experience. While it was previously possible to align videos manually, it took extensive time and effort, thus prompting us to create an automated solution for alignment.

Audio Representation

Below we can see raw audio data taken from videos represented as waveforms.

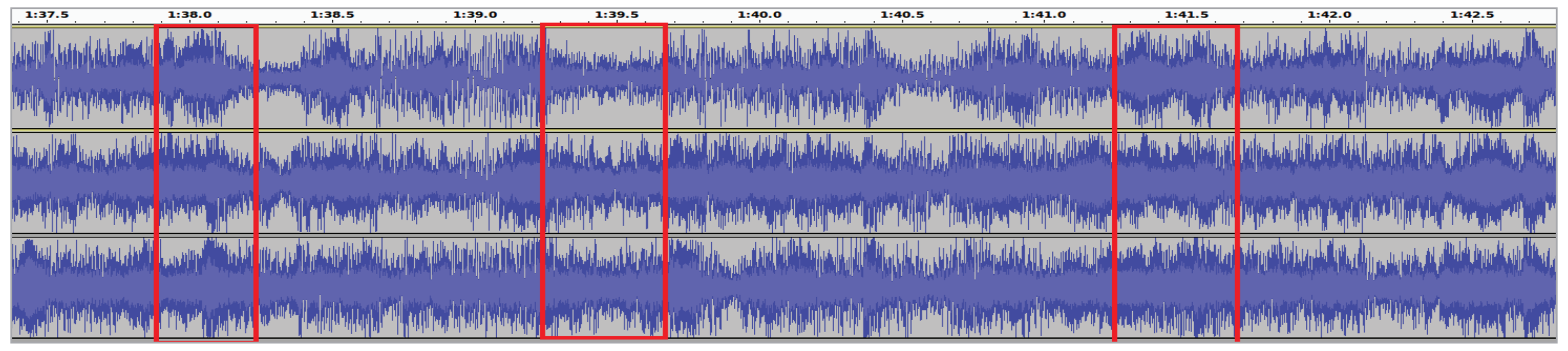


Figure 1: The audio waveforms of three overlapping videos from a concert.

By taking the spectrogram[1] of the audio signal, we can get a visual representation of the change in frequency over time. This is useful for identifying patterns in audio.

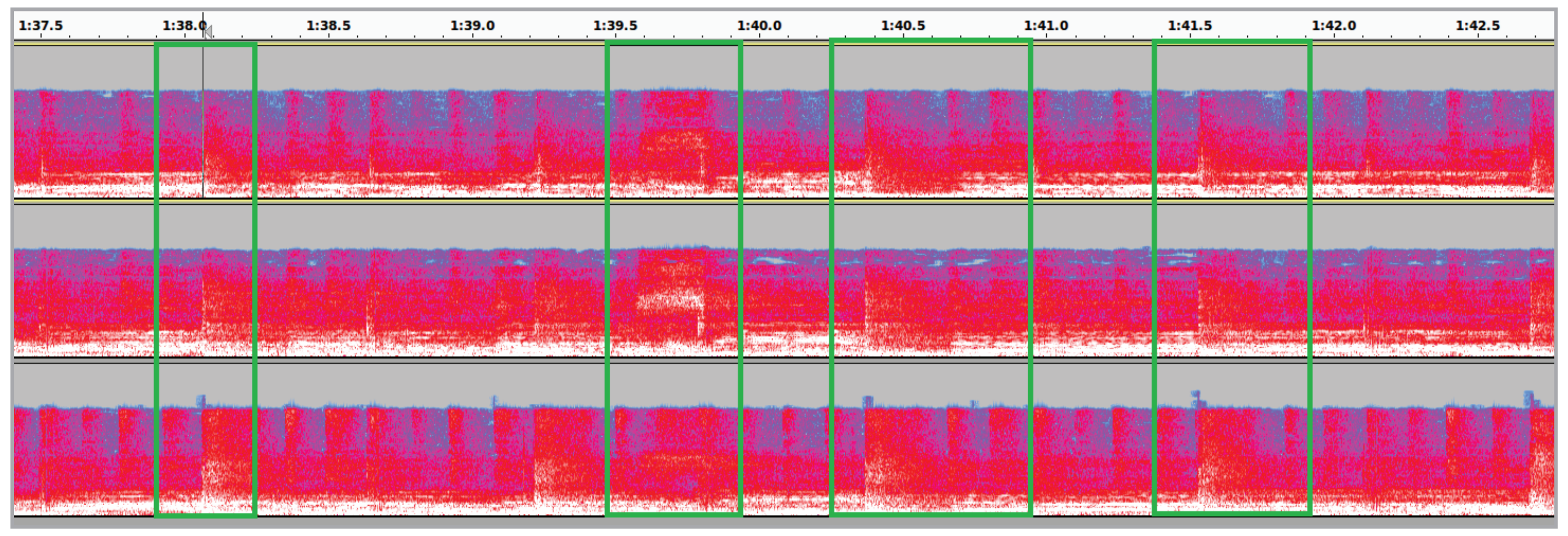
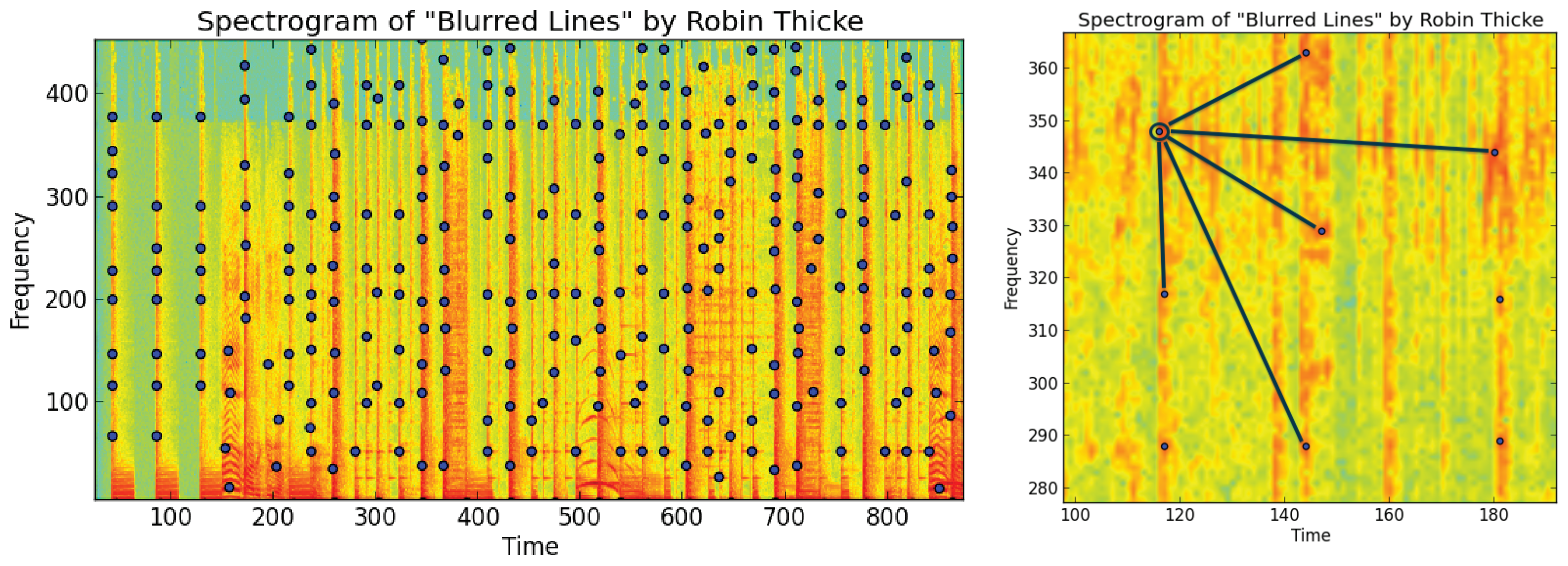


Figure 2: The spectrograms of the three audio signals show above. With the spectrogram, it is much easier to identify similarities between the tracks.

Audio Fingerprinting and Alignment

In our study we utilized an algorithm called Dejavu[1][4] which uses the spectrogram to find frequency peaks in audio creates unique "fingerprints" based off of the distances between these peaks.



Figures 3 & 4: A spectrogram (left) with frequency peaks plotted and one of the fingerprints (right) extracted from the spectrogram.

- Dejavu stores thousands of fingerprints per song in a database. We can query this database with other fingerprints to find matches and determine where songs overlap.
- To determine the alignment of multiple videos of the same event, we ingest all of the video tracks into a "corpus" which keeps track of resolved alignments.

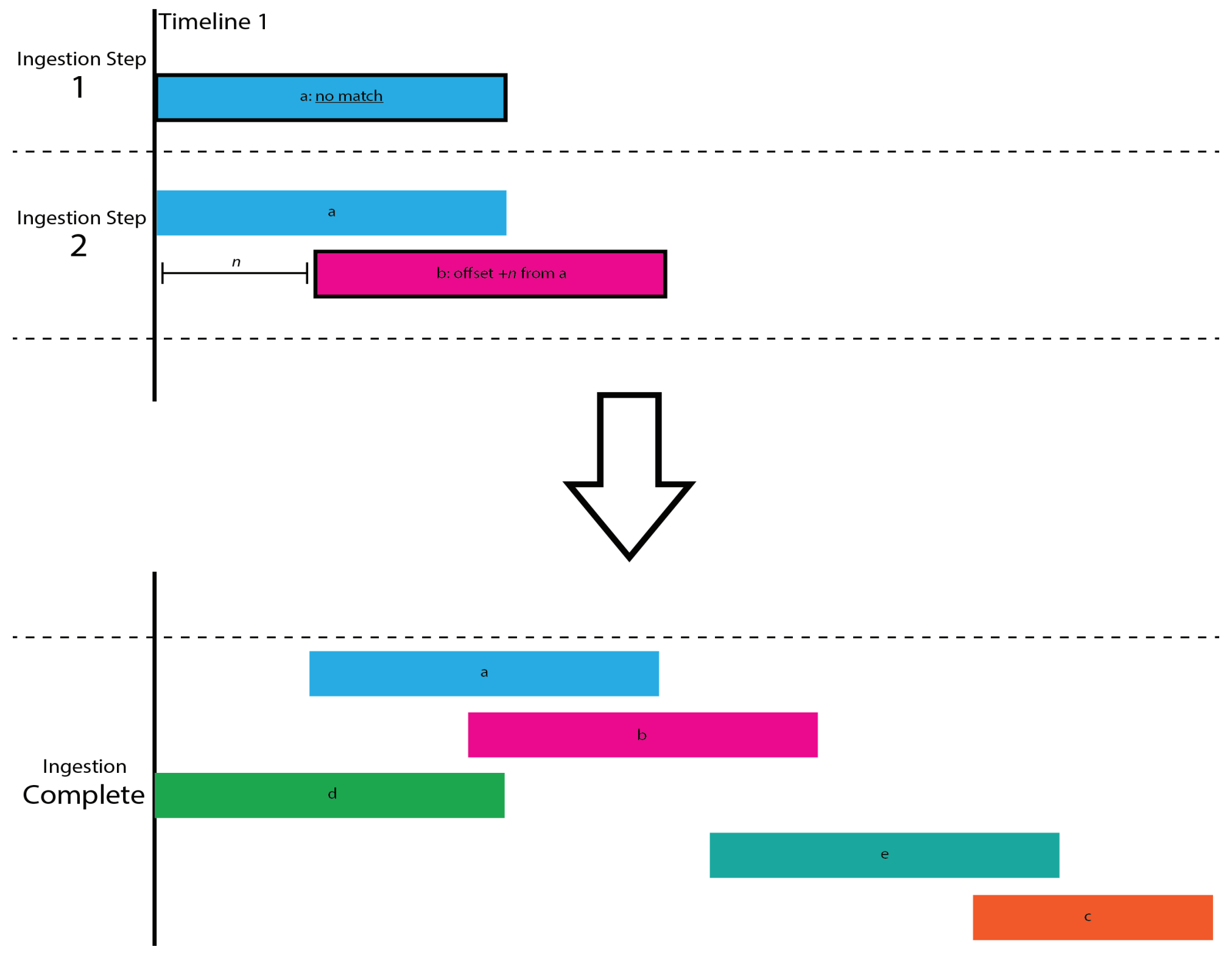


Figure 5: Video tracks are added to the corpus alignment in the order of ingestion.

- When a video does not match any currently ingested videos, it is placed into a new "timeline."
- If a video matches with multiple timelines during ingestion, we are able to merge these timelines together by using the track as a "seam" between the two timelines.

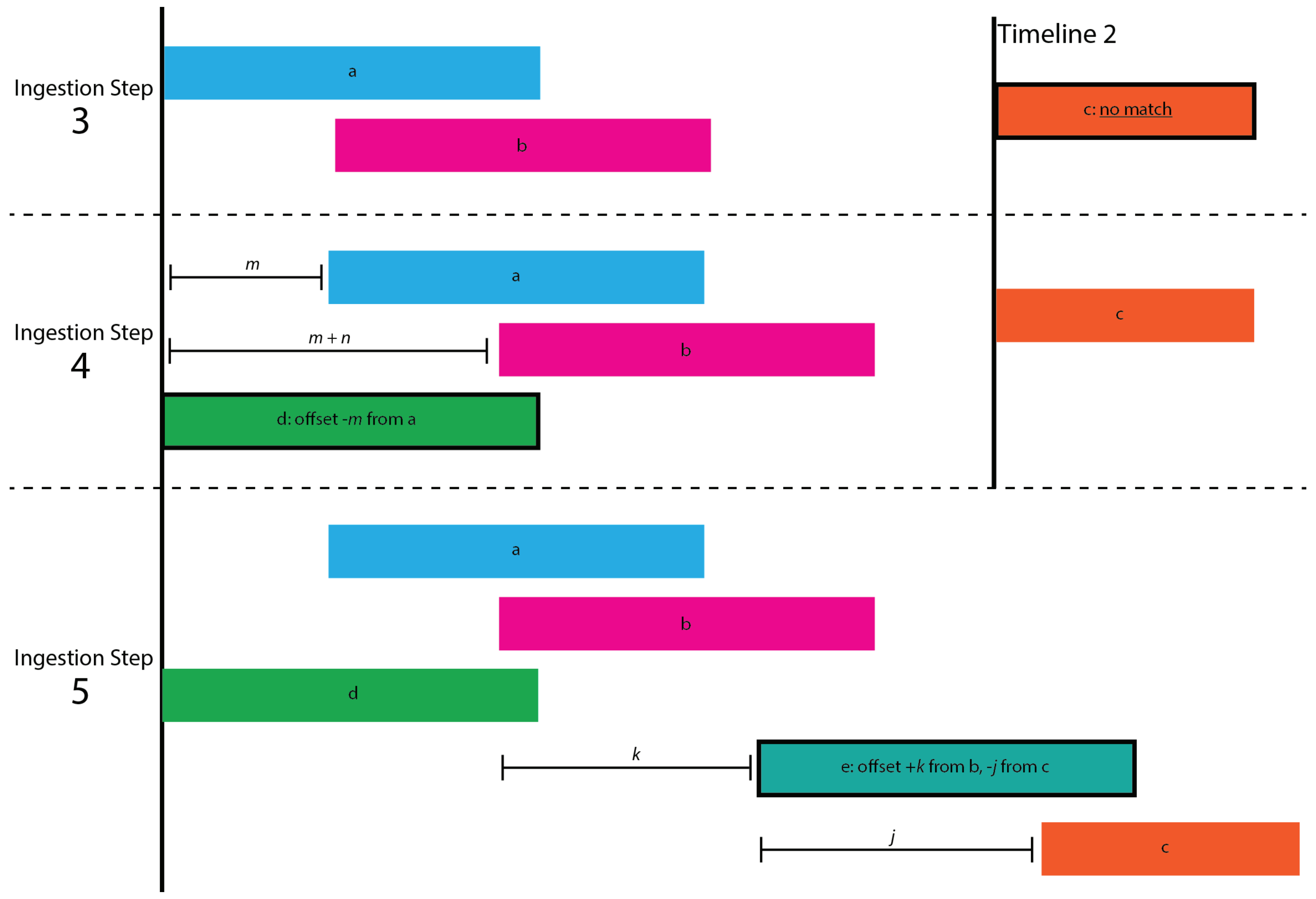


Figure 6: And ingestion step which causes a second timeline to be created (step 3), followed by merging of the new timeline into the original timeline (step 5). Step 4 shows a video with a negative offset shifting an entire timeline.

Future Work

In the future, we would like to replace Dejavu with a fingerprinting technique by LabROSA which is designed for noisy-source and noisy-target fingerprinting [2][3].

References

[1] W. Drevo, 'Audio Fingerprinting with Python and Numpy', Willdrevo.com, 2013. [Online]. Available: http://willdrevo.com/fingerprinting-and-audio-recognition-with-python/. [Accessed: 12- Feb- 2015].

[2] D. Ellis, 'THE 2014 LABROSA AUDIO FINGERPRINT SYSTEM', in ISMIR 2014, Taipei, Taiwan, 2014.

[3] C. Cotton and D. Ellis, 'Audio fingerprinting to identify multiple videos of an event', in Acoustics Speech and Signal Processing (ICASSP), Dallas, TX, 2010.

[4] Shazam Entertainment, Ltd., 'An Industrial-Strength Audio Search Algorithm', Shazam Entertainment, Ltd., Palo Alto, CA, 2003.