A Synchronization Ground Truth for the Jiku Mobile Video Dataset

Mario Guggenberger, Mathias Lux, and Laszlo Böszörmenyi

Institute of Information Technology,
Alpen-Adria-Universität Klagenfurt,
9020 Klagenfurt am Wörthersee, Austria
{mg,mlux,lb}@itec.aau.at

Abstract. This paper introduces and describes a manually generated synchronization ground truth, accurate to the level of the audio sample, for the Jiku Mobile Video Dataset, a dataset containing hundreds of videos recorded by mobile users at different events with drama, dancing and singing performances. It aims at encouraging researchers to evaluate the performance of their audio, video, or multimodal synchronization methods on a publicly available dataset, to facilitate easy benchmarking, and to ease the development of mobile video processing methods like audio and video quality enhancement, analytics and summary generation that depend on an accurately synchronized dataset.

Keywords: Audio, video, multimedia, crowd, events, synchronization, time drift

1 Introduction

With the incredibly fast proliferation of mobile devices capable of video recording, it is now easier than ever for people to quickly record interesting moments at the press of a button. For the research community, this opens up a lot of new and interesting opportunities. As an example, if you have recently been to a concert you might have noticed that people are constantly taking pictures and recording video clips. By the end, a huge dataset distributed over many devices has been generated by the crowd. Supposed there is a way to access this dataset, many interesting post processing methods can be applied to it. To name a few, there is the possibility to detect highlights and key moments by looking at the frequency of concurrent recordings, since people tend to capture what they consider to be most interesting to them or to friends that they want to show the capture. Recordings can be temporally stitched together to get a complete and continuous coverage of the whole event. Even better, vivid videos can be created by switching between different perspectives or showing different shots side-by-side. Quality can be improved by picking the best audio and video tracks from parallel recordings. 3D scenes can be reconstructed from recordings of different angles. It can even help in forensics, e.g. by reconstructing a crime scene and calculating where a gunshot came from.

The key to all these applications is precise automatic synchronization, a topic extensively researched in recent years, aiming to replace the tedious and very time-consuming manual work [16]. While an experienced user can synchronize a pair of recordings in a matter of minutes, it still costs him many hours to synchronize a large dataset. The difficulty of the problem is determined by multiple dimensions and grows with increasing clip amounts, decreasing clip lengths, decreasing perceived clip quality, and wider time frames where the clips are scattered in. To synchronize automatically, algorithms usually look at the audio or video content of the recordings and try to find unique events occurring in multiple recordings, which are then taken as reference points for aligning the recordings on a timeline. There are many published methods and algorithms for automatic synchronization to choose from, but authors usually evaluate them on their own custom datasets. This makes it impossible to compare them in terms of computational complexity, spacial complexity, synchronization rate, and synchronization accuracy.

To mitigate this situation, we contribute an accurate synchronization ground truth for a large publicly available mobile video dataset, and even consider the effect of time drift between the recording devices. It can be used to evaluate current and future synchronization methods, and serve as a foundation for methods that build upon synchronized audio and video tracks.

2 Related Work

There are many methods for audio and video synchronization, and a recent overview of synchronization methods is presented in [10]. Mathematical formulations of the synchronization problem can be found in [10, 16, 19]. There is no publicly available dataset with a precise synchronization ground truth, and individual methods are usually evaluated on custom datasets. Shrestha et al. [17] created a custom dataset captured at two different events by two video cameras, a wedding in a church and a dance event inside a hall, with a total runtime of 3 hours and 45 minutes. In follow-up works, they first extended the dataset with three additional events [18], and later extended it with two concert events [16] covered by 9, respectively 10 cameras. Both extensions consisted of short clips of 20 seconds to 5 minutes length, their total runtime is unknown. Kennedy and Naaman [9] evaluated their work on a reasonably big dataset sourced from YouTube from three big music concerts with about 200 videos each and runtimes between 1 and 10 minutes. Shankar et al. [14] used a custom dataset with videos recorded with mobile and handheld devices at cricket, baseball and football matches, but they did not describe it more detailed. The most recent work was conducted by Casanovas and Cavallaro [10], who again extended the dataset from [16] with additional events. All of these datasets are either too small, not distributable due to copyright restrictions, out-dated and not available any more, or do not capture the real-world characteristics of our use-case. If datasets are too small, they might (un)intentionally mask problems of complexity. If clips are too short or taken from homogeneous sources, they might mask drift. If the perceptual

quality of clips is too high or they are recorded in lab settings, they might mask low robustness.

Time drift has been mostly ignored in the multimedia community. The problem itself is well known and has been covered in network delay measurements [12] or to identify physical network devices through fingerprinting [15]. In multimedia, [10] is the first paper presenting a synchronization method that, to our knowledge, identifies and acknowledges the time drift problem. We have also already presented a demo application for media synchronization that can semi-automatically handle drift [4], and we described a measurement method in [5].

3 Jiku Mobile Video Dataset

The Jiku Mobile Video Dataset [13] is a collection of crowdsourced videos captured at 5 different events across Singapore by 4 to 15 recording devices in parallel, mostly in HD resolution. The events feature drama, dancing and singing performances. It aims at providing a publicly available collection of videos that (i) captures the unique characteristics of mobile video, (ii) supports researchers in working on solutions instead of spending time gathering test data, and (iii) enables benchmarking by leading to comparability of related methods and algorithms. It is to our knowledge the only currently and publicly available dataset of this kind, and by far the largest (Table 1) and most recent dataset available for event synchronization in general. An additional feature is the complementary metadata of each video recording comprised of compass and accelerometer readings. Potential applications suggested by the authors are (i) *video quality enhancement* by complementing information from multiple concurrent recordings from different viewpoints, (ii) *audio quality enhancement* by improving the audio track of a video with audio data from other concurrent audio tracks, (iii) *virtual directing* by automatically presenting the best shot out of a number of concurrent recordings to the viewer, and switching between them to create vivid multi-camera presentations, (iv) *occlusion detection* to support the selection of recordings that present the intended view of a scene, (v) *video sharing* by simulating events with a multitude of users transmitting their recordings over a network, and (vi) *mobile video analytics* including face detection, tracking, segmentation and de-shaking. Almost all of these suggestions rely on concurrent recordings, which implicates the need of an exact time-based synchronization. The clips are organized by a naming scheme consisting of an event ID, the date of the event, the ID of the recording device, and the recording start timestamp. By looking at the filename, they can be split into the five different event sets, and further divided into subsets by the recording device ID. The timestamps are too inaccurate and cannot be used for synchronization, as described in [17].

4 Methodology

This section describes the process of generating the ground truth. The goal was, for each set of event recordings in the dataset, to (i) lay out all recordings on a

Table 1. Breakdown of the Jiku Mobile Video Dataset. Additional detailed characteristics can be found in the original paper [13].

Event	GT_090912	NAF_160312	NAF_230312	RAF_100812	SAF_290512
Cameras	4	8	15	7	8
Recordings	50	66	117	97	143
Total Length	3h 37m	6h 00m	8h 23m	6h 40m	5h 57m

common timeline and (ii) extract the offset of each recording from the start of the timeline as the synchronization ground truth. The timeline begins at zero which equals to the moment the first recording was started, ends at the moment the last recording was stopped, and covers the whole interval in between. All recordings are placed such that all moments from the real event captured on recordings are placed at the same point on the timeline. We chose to synchronize the recordings by their audio tracks, because (i) it allows higher alignment precision due to the much higher audio sampling rate compared to the video frame rate, (ii) it provides humans a compact overview of the time dimension in the form of audio waveform envelopes which facilitates easy spotting and validation of matching points, and (iii) most currently existing synchronization algorithms work on audio data. The omnidirectionality of audio makes it also much easier to detect overlaps in the time domain than the strict unidirectionality of video, where cameras could be looking at totally different excerpts of the event scene.

While synchronization on audio tracks automatically leads to synchronized video tracks, they will not be as accurately synchronized due to the difference between the speed of sound and speed of light, and the fact that people in a crowd usually record from different positions with different distances from the target scene. Given the sound traveling at 340 m/s and neglecting the much higher speed of light, a difference of 10 meters distance yields a skew of ≈ 30 ms or ≈ 1 video frame at 30 fps. Luckily, time shifts between video tracks are less likely to be detected by humans, and offsets below the frame rate cannot be detected at all. In contrast, an audio offset of 30 ms is usually very noticeable. According to ITU, subjective research has shown that acceptability thresholds are at about +90 ms to −185 ms [7]. The ATSC found this numbers inadequate and recommends to stay within +15 ms and −45 ms [1]. In either case, switching between video streams that are out of sync will not always go undetected.

To generate the ground truth and lay out all recordings on the timeline, synchronization points between overlapping recordings had to be found, where a synchronization point is a quadruple consisting of two recordings and two time points that specify where the content in one recording equals the content in another recording. Given such a point, one recording can be adjusted to the other on the timeline such that the two time points are placed on the same time instant, which can be seen as a direct synchronization. An indirect synchronization involves intermediate recordings, such that two non-overlapping recordings A and B can be synchronized when recording A overlaps X and X

overlaps B, resulting in a synchronization of A, X, and B. It is not necessary to find synchronization points between all pairs of overlapping recordings, just between as many as are needed for a minimum spanning tree to be built from synchronization points interpreted as edges and recordings as nodes. One such tree then represents a cluster of directly and indirectly overlapping recordings. In the case of coverage gaps where an event is not continuously captured on recordings, multiple unconnected trees are formed. Care must be taken that a synchronization point between two tracks does not automatically lead to the tracks being synchronized over time, it only assures that the content of the two tracks conforms at the exact time points. To synchronize them over time and thus facilitate flawless parallel playback, the drift between the recordings must be detected and eliminated.

4.1 Time Drift Correction

To get our ground truth as precise as possible, we determined the absolute drifts in the Jiku dataset. We did this with the help of the Jiku authors [13] who provided us a mapping of device IDs to recording devices. We gathered devices of the same models and measured their absolute drift at a room temperature of $\approx 25\,^{\circ}\mathrm{C}$ with the same method that we described in [5]. Table 2 lists the recording devices, their dataset IDs and the measured drifts in milliseconds per minute. A positive drift indicates that the real sampling rate of a device is higher than the nominal sampling rate, making the playback time longer than the captured real-time event when played back at the nominal sampling rate. Knowing these drifts, it is now sufficient to synchronize two overlapping recordings at one single point to get them synchronized over their whole overlapping interval. There is still a small fraction of drift error left, resulting from the fact that we did not measure the exact same devices that were used for recording and we do not know the temperatures at which the recordings took place. Series of measurement in our laboratory have shown a standard drift deviation of $\approx 0.1\,\mathrm{ms/min}$ between multiple devices of the same model, and temperature changes between $-20\,^{\circ}\mathrm{C}$ and $+50\,^{\circ}\mathrm{C}$ have shown a variance of $\approx 1\,\mathrm{ms/min}$ [5], which we assume to also be true for the ones used in the dataset. In our opinion, both of these errors left in the measurements do not have a reasonable impact on our ground truth because (i) the temperature difference between our laboratory and the actual air temperature at recording time in Singapore is presumably much lower than between the extreme bounds in our laboratory measurements and (ii) the recordings in the dataset are short enough to minimize its impact. Out of the 481 recordings in the dataset, only 19 are longer than 15 minutes, and more than 75% stay below 5 minutes runtime.

4.2 Manual Synchronization

The manual synchronization was done by an author of this paper who has a lot of experience in multi-track recording and post-production of audio and video

Table 2. Measured absolute drifts in ms/min of the recording devices used to create the Jiku Mobile Video Dataset.

Device	IDs	Drift
Samsung GT-i9023 Nexus S	15, 16, 19, 20	−0.37
Samsung GT-i9000 Galaxy S	5	+0.26
Samsung GT-i9100 Galaxy S II	2, 3, 4, 11, 12, 13, 14, 17, 18, 21, 23	+15.95
Samsung GT-i9250 Galaxy Nexus	0, 1, 6, 7, 8, 9, 10	+4.78
Samsung GT-i9300 Galaxy S III	22	+0.34

data and has had the pleasure to synchronize tracks on many occasions. Doing this manually, especially when many tracks need to be synchronized, takes a lot of time and effort. This is why automatic methods are sought after, but both available in the research domain and on the commercial market have not been used by intention since it would contradict the intended purpose of the ground truth. To give the manual process a starting boost, we still applied two automatic approaches to get a rough timeline pre-alignment to start with, but every synchronization point in the final result has been set and verified by hand. The first approach was generating an approximate timeline alignment from the metadata timestamps for all 481 recorded clips. This helped to get a very rough overview of the alignment of recordings and to spot extreme outliers. At this point, almost all recordings were off of their final alignment. The second approach was the application of an audio fingerprinting algorithm [6] that helped to obtain approximate synchronization points for about 50% of all recordings, which specifically helped in those cases where the timestamps were off by a huge amount. The manual work began with the validation and correction of wrong pre-alignments by looking at the waveform amplitude envelopes, trying to find visually matching patterns and listening to the recordings to semantically match them by their content, until all recordings were approximately synchronized. At this stage, the synchronization between recordings was accurate to a few seconds only. Then followed a time consuming manual refinement process, where 397 exact synchronization points were determined by visually looking at the waveforms, aurally listening to the audio data, and fine adjusting their relative offsets until the alignments were as precise as possible, often at sample or even subsample level. It was always followed by a validation step where the overlapping interval was proof-listened. The difficulty of determining a synchronization point varied from easy cases where the signals could be visually matched very clearly to hard cases with extremely distorted signals where only aural matching by repeated careful listening and readjusting was possible. All of this work was done in a custom software specifically developed for synchronization purposes. It took about 20 hours and was approximately cut in half by the automatic pre-alignment. The final result was a list of synchronization points which we transformed into a list of time offsets resulting in the manual ground truth. The timestamps were used as reference to order unconnected clusters of over-

lapping track groups in time, because this information cannot be inferred from the synchronization points alone.

5 Synchronization Ground Truth

The synchronization ground truth contains, for each of the five events, the start times of all recordings ordered on a timeline, the drift correction factors, and all manually generated synchronization points. Laying out all recordings on a timeline with the specified offsets and changing their runtime by the drift factor results in a synchronized event. The start times are relative to the start time of the first recording at the corresponding event, which is assumed w.l.o.g. as zero, and are calculated from the synchronization points. All specified times are given to a fractional seconds precision of 10^{-7} to enable subsample accuracy. Since all synchronization points have been generated and validated manually, they are very precise on one hand, probably more precise than current algorithms are able to achieve, but on the other hand this means that their precision cannot be measured in numbers. It is guaranteed though, that almost all synchronization points are inexact to at most 10 ms, where most are more precise and only a very small part of very hard to determine synchronization points are off by more. These are cases where humans and also computer algorithms probably reach their current limits. All synchronization points are guaranteed to be exact enough for artifacts of nonsynchronous playback, like echoes, to be unperceivable. It is not guaranteed that video frames of concurrent recordings are in sync, because of the already mentioned difference between the speed of sound and speed of light. We had to exclude all recordings from device 5 in the NAF_230312 set because they were not correctly cut, resulting in multiple noncontinuous shots inside its files that rendered them unsynchronizable. The data is available for download on our website[1] in structured XML files.

5.1 Accuracy

To evaluate the accuracy of our manually generated synchronization points, we chose to cross-correlate short intervals of audio samples that surround the points. The idea was that a low cross-correlation offset with a high correlation coefficient would confirm a synchronization point valid, while a high offset would be an indicator that the manually set synchronization point is inaccurate and can even be improved by the offset. Cross-correlation in general is a computationally expensive operation, but a 1-second interval sufficed because we knew for sure that all potential manual synchronization errors are much smaller, since e.g. an error of 50 ms would stand out heavily and cannot go undetected during validation. It turned out that the correlation results could not be used to automatically classify the manual synchronization points into true and false positives because we were unable to set a reasonable threshold. A problem is that

[1] `http://www-itec.aau.at/~maguggen/jikusync/`

we do not know the maximum achievable correlation coefficient between pairs of recordings, due to noise, the different frequency pickup patterns of the recording devices, and the time drift error. Upon inspection of the results, we found a lot of cases where the correlation offset was rated with a high coefficient but was actually too far off the optimal synchronization point, leading to audible echoes when listened to carefully. In contrast, we had many cases of valid offsets with much lower coefficients. Experiments with different interval lengths, sampling rates, and frequency filtering did not have any significant impact on the results. We could still learn a lot about the ground truth by manually analyzing the results. Looking at Figure 1, we can see that 200 of the 397 synchronization points result in a cross-correlation offset within ± 5 ms, and 274 are within ± 10 ms. This means that in all these cases, our manually generated synchronization points correlate highly with those calculated by the cross-correlation, confirming the accuracy of our manually generated data. All other cross-correlation results were manually double-checked and found to be more inaccurate compared to the manually identified points. The extreme cases where the cross-correlation offsets lied within the three-digit range happed in very noisy audio tracks where the correlation series are flat and the maximum correlation coefficients not located at distinct peaks, leading to ambiguous results.

5.2 Comparison

To show that the timestamps of the dataset are not reliable enough to be used for synchronization, we compared our ground truth with the timestamps. We measured the time difference of each recording as the error between the ideal position in the event timeline from the ground truth and the position from the timestamp-based synchronization. The distribution of the offsets is shown in Figure 2, which clearly indicates that a timestamp-based synchronization approach is not suitable to be taken as a ground truth because even half a second offset between two concurrent recordings causes a heavily noticeable lag in the audio and video tracks, and larger lags make it often even impossible to perceive two recordings as concurrent. The majority of offsets is greater than one second, and the manually generated ground truth is therefore essential for the development and evaluation of synchronization-dependent methods. A few clips had enormous offsets because the clocks of the recording devices were not set correctly, resulting in timestamps years behind (around January 2000).

5.3 Evaluation

To demonstrate the usefulness of our ground truth, we chose to evaluate the synchronization performance of the well known audio fingerprinting algorithm by Haitsma and Kalker [6] by measuring the preciseness of the calculated synchronization points. This method has been shown to be a promising method for media synchronization in [16] and [3], and we had it already implemented in our own synchronization tool. We applied it with the default parameters as described in the original paper on each of the five events in the Jiku dataset, which yielded

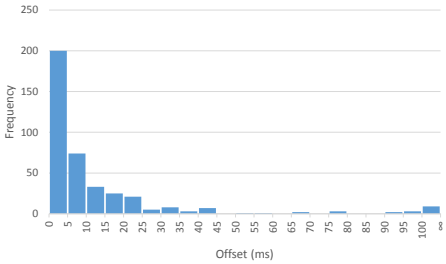

Fig. 1. Distribution of the calculated cross-correlation offsets from the manually generated synchronization points.

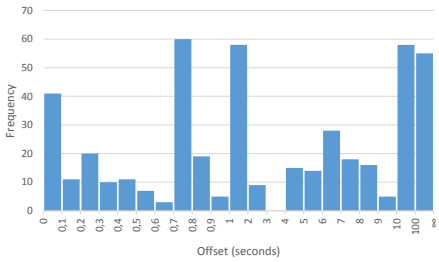

Fig. 2. Distribution of the error offsets between the timestamp synchronization and the synchronization ground truth.

2020 synchronization points in total. Figure 3 shows a histogram distribution of their offsets from the ground truth, binned in steps of 5 milliseconds. Most of the synchronization points are within the range of ±50 ms; 140 are outside the 100 ms range of which most are false positives that are off by many minutes and connect completely unrelated clips. The 95% confidence interval of the mean is between 21.2 ms and 22.8 ms. To test our hypothesis that cross-correlation might improve synchronization results, we applied it on all synchronization points by correlating 1-second audio signal excerpts centered around the positions they point to. This post-processing step improved the fingerprinting results significantly by shifting them towards smaller offsets and almost tripling the synchronization points in the range of ±5 ms. The 95% confidence interval of the mean moved down to 10.2 ms-11.4 ms. The improved results are also shown in Figure 3 for comparison.

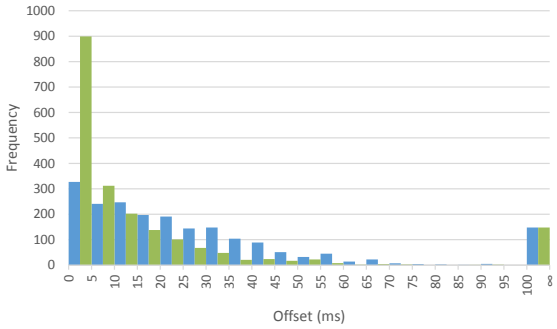

Fig. 3. Histogram distribution of the offsets to the ground truth of all synchronization points as found by the fingerprinting approach (blue), and additionally post-processed by cross-correlation (green).

The overall synchronization rate of the algorithm, which is the number of clips that are covered by the calculated synchronization points, can also be determined with the help of the ground truth. For this, we compared the optimal minimum spanning trees of the overlapping event recordings generated from the ground truth with the minimum spanning trees generated from the computed synchronization points. Table 3 contains for each dataset the number of edges in the optimal MST, the number of determined MST edges by fingerprinting, and the resulting synchronization rate. It shows that this fingerprinting method does not yield satisfying results, owed to the real-life characteristics of the dataset that place high demands on the robustness of synchronization methods due to the uncontrolled environment and heterogeneous sources. There are many heavily distorted audio tracks due to background noise, heavy compression, and poor built-in microphones or analog-to-digital converters that cannot cope with high sound pressure levels like they usually occur at such live events.

Just like we demonstrated the determination of the overall synchronization rate and the individual improvements gained by cross-correlation, our ground truth can be used for the evaluation and comparison of all methods presented in Section 2, where some are expected to perform better. For the fingerprinting method that we evaluated, there are also a few iterative improvements proposed in [8], [2] and [11], which could also be objectively evaluated.

Table 3. Synchronization rate of the fingerprinting method on the Jiku events showing the optimal number of MST edges in the ground truth (MST_{GT}), the achieved number through fingerprinting (MST_{FP}), and the rate in percent.

Event	GT_090912	NAF_160312	NAF_230312	RAF_100812	SAF_290512
MST_{GT}	44	63	106	82	102
MST_{FP}	23	54	73	15	78
Rate	52%	86%	69%	18%	76%

6 Conclusion

This paper presents an audio based manually generated and validated synchronization ground truth for the Jiku Mobile Video Dataset. It cleans the dataset from time drift and extends the timestamps in the dataset to a much higher precision. It aims at researchers who want to evaluate or benchmark synchronization algorithms, researchers who develop methods that rely on a synchronized dataset, and demonstrates through an exemplary evaluation experiment how helpful the ground truth can be.

To further improve the dataset, interesting future work could be the determination of the audio to video track offsets to make audio and video data perfectly synchronized at the same time. User studies to determine detectability and acceptability thresholds of offsets between parallel audio tracks are needed to

assess the maximum acceptable error offset. Other interesting future work could include the evaluation of different synchronization algorithms on this ground truth to determine the best fit for the evergrowing use-case of crowd sourced mobile video.

Acknowledgments. This work was supported by Lakeside Labs GmbH, Klagenfurt, Austria, and funding from the European Regional Development Fund (ERDF) and the Carinthian Economic Promotion Fund (KWF) under grant 20214/22573/33955. Special thanks go to the authors of the Jiku Mobile Video Dataset for creating and providing it to the community.

References

1. ATSC. *Relative Timing of Sound and Vision for Broadcast Operations (IS-191)*. Advanced Television Systems Committee, June 2003.
2. S. Baluja and M. Covell. Content fingerprinting using wavelets. In *Visual Media Production, 2006. CVMP 2006. 3rd European Conference on*, pages 198–207, Nov 2006.
3. N. Duong, C. Howson, and Y. Legallais. Fast second screen tv synchronization combining audio fingerprint technique and generalized cross correlation. In *Consumer Electronics - Berlin (ICCE-Berlin), 2012 IEEE International Conference on*, pages 241–244, Sept 2012.
4. M. Guggenberger, M. Lux, and L. Boszormenyi. Audioalign - synchronization of A/V-streams based on audio data. In *Multimedia (ISM), 2012 IEEE International Symposium on*, pages 382–383, Dec 2012.
5. M. Guggenberger, M. Lux, and L. Böszörmenyi. An analysis of time-drift in handheld recording devices. In *MultiMedia Modeling*, Lecture Notes in Computer Science. Springer International Publishing, 2015.
6. J. Haitsma and T. Kalker. A highly robust audio fingerprinting system. In *Proceedings of the 3rd International Conference on Music Information Retrieval (ISMIR)*, Paris, France, 2002.
7. ITU. *Relative timing of sound and vision for broadcasting (ITU-R BT.1359-1)*. International Telecommunication Union, Nov. 1998.
8. Y. Ke, D. Hoiem, and R. Sukthankar. Computer vision for music identification. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 597–604 vol. 1, June 2005.
9. L. Kennedy and M. Naaman. Less talk, more rock: Automated organization of community-contributed collections of concert videos. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, pages 311–320, New York, NY, USA, 2009. ACM.
10. A. Llagostera Casanovas and A. Cavallaro. Audio-visual events for multi-camera synchronization. *Multimedia Tools and Applications*, pages 1–24, 2014.
11. P. Mansoo, K. Hoi-Rin, Y. M. Ro, and K. Munchurl. Frequency filtering for a highly robust audio fingerprinting scheme in a real-noise environment. *IEICE transactions on information and systems*, 89(7):2324–2327, 2006.
12. S. Moon, P. Skelly, and D. Towsley. Estimation and removal of clock skew from network delay measurements. In *INFOCOM '99. Eighteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, volume 1, pages 227–234 vol.1, Mar 1999.

13. M. Saini, S. P. Venkatagiri, W. T. Ooi, and M. C. Chan. The jiku mobile video dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*, MMSys '13, pages 108–113, New York, NY, USA, 2013. ACM.
14. S. Shankar, J. Lasenby, and A. Kokaram. Warping trajectories for video synchronization. In *Proceedings of the 4th ACM/IEEE International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Stream*, ARTEMIS '13, pages 41–48, New York, NY, USA, 2013. ACM.
15. S. Sharma, A. Hussain, and H. Saran. Experience with heterogenous clock-skew based device fingerprinting. In *Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results*, LASER '12, pages 9–18, New York, NY, USA, 2012. ACM.
16. P. Shrestha, M. Barbieri, H. Weda, and D. Sekulovski. Synchronization of multiple camera videos using audio-visual features. *Multimedia, IEEE Transactions on*, 12(1):79–92, 2010.
17. P. Shrestha, H. Weda, M. Barbieri, and D. Sekulovski. Synchronization of multiple video recordings based on still camera flashes. In *Proceedings of the 14th Annual ACM International Conference on Multimedia*, MULTIMEDIA '06, pages 137–140, New York, NY, USA, 2006. ACM.
18. P. Shrstha, M. Barbieri, and H. Weda. Synchronization of multi-camera video recordings based on audio. In *Proceedings of the 15th International Conference on Multimedia*, MULTIMEDIA '07, pages 545–548, New York, NY, USA, 2007. ACM.
19. A. Whitehead, R. Laganiere, and P. Bose. Temporal synchronization of video sequences in theory and in practice. In *Application of Computer Vision, 2005. WACV/MOTIONS '05 Volume 1. Seventh IEEE Workshops on*, volume 2, pages 132–137, Jan 2005.