Videos from the 2013 Boston Marathon: 
An Event Reconstruction Dataset for 
Synchronization and Localization

Jia Chen, Junwei Liang, Han Lu, Shoou-I Yu, Alexander G. Hauptmann

CMU-LTI-018

Language Technologies Institute
School of Computer Science
Carnegie Mellon University
5000 Forbes Ave., Pittsburgh, PA 15213
www.lti.cs.cmu.edu

© October 1, 2016
Videos from the 2013 Boston Marathon: An Event Reconstruction Dataset for Synchronization and Localization

Jia Chen, Junwei Liang, Han Lu, Shoou-I Yu, Alexander G. Hauptmann
School of Computer Science, Carnegie Mellon University
{jiac, junweil, hlu2, iyu, alex}@cs.cmu.edu

ABSTRACT
Event reconstruction is about reconstructing the event truth from a large amount of videos which capture different moments of the same event at different positions from different perspectives. Up to now, there are no related public datasets. In this paper, we introduce the first real-world event reconstruction dataset to promote research in this field. We focus on synchronization and localization, which are the two basic and essential elements for other tasks in event reconstruction such as person tracking, scene reconstruction, and object retrieval. It covers 347 original videos and 1,066 segmented clips of the real-world event of the explosions at the 2013 Boston Marathon finish line. We provide high precision ground truth labels for localization and two granularity ground truth labels for synchronization on 109 clips. We derive several metrics on video level and frame level to evaluate the two tasks. We also provide auxiliary data including video comments with timestamp, map meta data, environment images, and a 3d point cloud which are helpful for the synchronization and the localization tasks. Finally, we position our dataset as a real-world test dataset, without limiting the usage of extra training data. The dataset is released at http://aladdin1.inf.cs.cmu.edu:8081/boston

1. INTRODUCTION
With the popularity of smartphones equipped with high quality cameras, events around the world can be quickly captured by videos and rapidly shared via social media. When an event happens, different videos capture different moments of the same event at different positions from different perspectives. For example, videos from surveillance cameras usually cover relatively long time spans of an event from a fixed location; videos from television reporters usually cover time spans of medium length of an event in a rather professional perspective at major locations; the videos from passersby usually cover relatively short time spans of an event from a personal perspective at side locations that are often not covered by the news reporters. The above situation is very similar to the story of the “blind men and an elephant” [2]: the event truth corresponds to the elephant in the story, and each single video corresponds to one of the blind men who only touches parts of the elephant, either in the time dimension or in the space dimension. An illustration is given in Figure 1. The goal of event reconstruction is to recover the elephant in its entirety from each of the blind mens’ descriptions.

The input for event reconstruction consists of a set of different videos. The output is task dependent, and the tasks could be, for example, person tracking, scene reconstruction, or object retrieval. Underlying are two task independent elements: the alignment of multiple sensors to the time dimension and the alignment of multiple sensors to the space dimension. Aligning to the time dimension discovers which time period of the event was captured by the video. Aligning to the space dimension reveals which geographical area is covered by the video, and the position from where the video was filmed. We denominate the aligning to the time dimension as the synchronization and the aligning to the space dimension as the localization.

We argue that synchronization and localization are the two basic and essential elements of event reconstruction. To support this argument, we investigate the role of synchronization and localization in three common tasks of event reconstruction: person tracking, scene reconstruction, and
object retrieval. For person tracking, we need to know where and when the same person appears in different videos. For scene reconstruction, we need to know the position of each camera. If the scene changes dynamically, if for example an explosion causes a building to collapse, we need to also know the time span of the video content. For object retrieval, we need to know when and where an object appears or disappears. Synchronization and localization provide the when and where elements of an event, and are, therefore, the basic and essential elements of event reconstruction.

In this paper, we introduce a dataset for real-world event reconstruction that can be used for synchronization and localization, and that can be further explored to do person tracking, object retrieval, and other interesting tasks of event reconstruction. Different from event detection datasets [17],[14],[16] which focus on the coverage of different event types, we focus on the data richness within an event, asking if there are enough videos for a successful localization and synchronization. To be specific, an ideal dataset for an event reconstruction should have two properties. First, there should be plenty of videos covering not only different time periods on the event’s timeline for the synchronization but also different geographical areas around the event’s location for the localization. Second, the events in the dataset should enable us to do person tracking, scene reconstruction, or object retrieval, so that they can be used to study event reconstruction tasks in the real world. We introduce the Boston Marathon dataset, which contains 347 original videos and 1,066 segmented clips from these videos covering different time periods and areas of the real-world event of the two explosions at the finish line of the 2013 Boston Marathon. This was a real-world event in which person tracking, scene reconstruction, and object retrieval tasks are all worth to be investigated. The video clips cover three major time periods of the event: the pre-explosion, the two explosions, and the post-explosion. They also cover different geographical areas of the event, including the point of the first explosion, the point of the second explosion, and the point where the evacuation teams gathered, and so on.

In addition to the raw video clip data, we provide high precision ground truth labels for the localization, and two granularity ground truth labels for the synchronization for 100 video clips. We additionally provide auxiliary data such as video comments with timestamp, map meta data, images of the environment, and a 3d point map cloud, all of these are also useful for localization and synchronization. We position our dataset as a real-world test dataset, and we do not limit the usage of extra training data for the following reason: there are multiple clues that could be exploited by having auxiliary data available, e.g. objects, pedestrians, buildings, and so on, and this will improve the localization and synchronization tasks. There are already many large-scale labeled datasets available targeting objects, pedestrians, and buildings separately. So researchers could leverage these datasets for localization and synchronization, and could test their algorithms using our real-world test dataset for event reconstruction. The dataset is released at http://aladdin1.inf.cs.cmu.edu:8081/boston.

2. RELATED WORK

Event videos capture the life in the real world. There is an abundance of research on event videos, most of which is focusing on a more general type of event detection. This research is based on either a class of action events or on activities, which is different from focusing on a specific real-world event which has really happened once in time and will not repeat, such as it is investigated in event reconstruction.

The definition of the MED task [14] is to find all clips that contain a certain event in a video collection, given an event description or event kit. The event kit provides a semantic description in natural language of 20 pre-defined events, for example, “Baking a cake”, “Batting in a run”, or “Assembling a shelter”. Its corresponding dataset IACC contains 34,000 videos. Though the IACC dataset contains a large amount of sample videos for each kind of these pre-defined events, each video usually belongs to one specific event. Specific event labels are not provided.

The definition of the SED task [16] is to detect a set of action events in a surveillance video, including actions such as Embrace, Pointing, ObjectPut, and so on. These action events are fine-grained. However, a complex real-world event is usually composed of a series of such fine-grained action events. For example, a possible suspect may conduct a series of actions, including running, leaving an object, and meeting another suspect, in order to eventually carry out an explosion. The SED task does not provide any major real-life events, which puts a limit on potentially interesting research and analysis that could otherwise be done using these action elements for event reconstruction. The corresponding dataset i-DLS contains 49 videos from only 5 fixed indoor surveillance cameras. Furthermore, these videos do not contain audio tracks and are not synchronized.

For the synchronization task, there are datasets designed with a focus on the audio signal. Bano at al. [11] for example collected such a a dataset biased towards audio signals. 43 out of the total of 48 events are concert events. The remaining 5 events such as the Changing of the Guard and the Olympic Torch Relay also contain strong distinguishable audio signals. On average, there are only 5 videos per event which is a relatively small amount of videos in the context of event detection. By choosing to work with the Boston Marathon dataset, we selected an event based on its real-world impact and on an abundance of related videos. That is to ensure to not having any bias on one signal modality. In fact, we found that also the visual signal is useful in collecting the synchronization ground truth. It is interesting to compare synchronization algorithms of different modalities on an event with large real-world impact.

For the localization task, there is much research on image localization [15],[21],[12],[13],[10]. The Oxford Buildings [15] and the San Francisco Landmarks [12] are single-modality datasets. Both the query photos and the geo-tagged photos come from the web service flickr in high resolution. Furthermore, buildings are the major visible objects in most query images. The Boston Marathon dataset is heterogeneous regarding the localization task: the queries are in the form of videos, and we provide auxiliary geo-tagged images for the localization. Furthermore, most of the time the location smoothly changes within a video which makes the problem more complex than just localizing a single image.

3. DATASET COLLECTION

In this section, we first introduce the collection of the raw data and the data cleaning, including event selection, query construction, and video de-duplication. Then we describe the process of labeling the ground truth, and analyze the
precision of the collected ground truth. Finally, we introduce auxiliary data which is useful for the synchronization and localization tasks, including video comments with time, the map meta-data, the environment images, a 3d point cloud map.

3.1 Raw Data Collection and Cleaning

3.1.1 Data Collection

To collect data, we needed to select a suitable event, which has a large impact on the real world, and is also documented by rich video data online. We chose the event “Boston Marathon 2013”[1], in which two consecutive explosions happened on the sidewalk near the finish line of a traditional city marathon in Boston in 2013. Figure 2 gives an illustration of the event. It received widespread international media attention. There are many videos about the event online: 1,180,000 videos are the search result for a search on Google video using the query “boston marathon explosion 2013”, and 415,000 videos are the search result for a search on Youtube using the same query. The dataset collected for this event directly reflects a real-world situation which provides a real-world setting for both research problems and algorithm evaluation.

We constructed the queries “boston marathon 2013 explosion”, “boston marathon 2013 bomb”, “boston marathon 2013 after explosion”, “boston marathon 2013 after bomb” to crawl videos from Youtube[9] and Dailymotion[3], two of the most popular video sharing websites. We crawled the top 500 search results from each query on Youtube, and all the search results from dailymotion.

3.1.2 Data Cleaning

The data cleaning involves cleaning out irrelevant videos and duplicate videos.

Relevance refinement: We defined the relevant videos to only be on-site videos of the “Boston Marathon 2013” event. Videos of other marathons are irrelevant because they are not about the “Boston Marathon 2013”. Videos of the President’s speech about the event are irrelevant because they are not on-site videos. We manually refined the relevance of all the crawled search results by removing irrelevant videos, resulting in 347 relevant videos. Most of these videos were uploaded to the websites in their raw footage version.

Duplication detection: On video sharing sites, popular videos get uploaded multiple times as whole copies, copied parts, and edited copies. Having many duplications in a dataset has three drawbacks: 1. It wastes precious manual effort in ground truth labeling because the same videos are essentially getting labeled multiple times. 2. It disguises the true content diversity in the dataset since many duplications give a wrong impression about the actual size of the diversity. 3. It leads to a biased evaluation as especially popular videos appear multiple times in the test set. To detect the duplicates and duplicated video parts, we segmented the videos into clips, and did duplication detection on clip-level. We used a scene detection tool from ffmpeg[4] to segment 347 relevant videos into 1,066 valid clips. A clip is considered as valid if its length is more than 10 seconds. Both 347 original videos and the 1,066 valid clips are included in the dataset. We adopted a standard hierarchical video duplicate detection pipeline[20] on video clips. In the first stage, we filtered the duplicate pairs out by using their global signature in the color histograms. In the second stage, we refined the remaining pairs by doing local feature matching on keyframes. In the end, we detected 250 duplicate clusters, covering 500 clips in total. Furthermore, we performed manual de-duplication on the remaining clips. In the end, there are now 109 unique clips in the dataset. The clips in the same cluster vary in their resolution, their sharpness, and their audio tracks which could be used to evaluate the algorithms’ robustness to these factors. Thus we provide the duplication cluster lists rather than actually removing the duplicate videos.

3.2 Ground Truth Collecting

Most videos from the sharing websites such as Youtube and Dailymotion do not come with their original meta-data. Thus we had to manually label the ground truth for the synchronization and the localization. The major difficulty we met during manually labeling for these two tasks was to control the precision of the results. We managed to control the precision by finding good reference signals for the synchronization, and good reference materials for the localization labels.

3.2.1 Synchronization Ground Truth

We adopted a hierarchical approach for the labeling process. First, we split the time line of the event into three periods using the signal of the first explosion and the signal of the second explosion, resulting into the periods of the pre-explosion, the explosion, and the post-explosion. Second, within each period, we selected a list of sub-events and labeled their exact time in the clip.

We categorized the clips into the three time periods based on the following standards: a clip is classified to the pre-explosion period if it only reports the moments before the explosion; a clip is classified to the explosion period if it contains either the first explosion or the second explosion; a clip is classified to the post-explosion period if it only reports the moments after the explosions. The clips from the pre-explosion period include the scenes in which the leading runners reached the finish line and the appearance of the suspects. The clips from the post-explosion period clips include the evacuation of the spectators and the arrival of the ambulances. The error range of the exact time labels for the sub-events is controlled within 1 second because the sound of...
Table 1: Hierarchical approach to label the synchronization ground truth

<table>
<thead>
<tr>
<th>period</th>
<th>sub-event</th>
</tr>
</thead>
<tbody>
<tr>
<td>pre-explosion</td>
<td>winner reaching the finish line</td>
</tr>
<tr>
<td>explosion</td>
<td>1st explosion, 2nd explosion</td>
</tr>
<tr>
<td>post-explosion</td>
<td>man in red clothes pulling off the banner, man with red hat in white clothes crossing the fence, military man pulling off the fence</td>
</tr>
</tbody>
</table>

Table 2 summarizes the hierarchical labeling approach for the synchronization ground truth.

### 3.2.2 Localization Ground Truth

Labeling of only one ground truth for one video would not be accurate because there is a lot of content change and camera perspective change during a video. On the other side, the costs for the labeling of the ground truth in every single frame of a video would be too high. To reduce the costs, we chose to label the ground truth only in key frames rather than in every single frame. We can interpolate the ground truth for the rest of the frames using the ground truth from the nearby key frames. As the video content changes only a little between the key frames and the neighboring frames, this strategy incurs a negligible loss in precision for non-key frames but manages to save much manual effort.

The localization ground truth for a video consists two parts: the pose of the camera and the geographical location of the video content. Given an accurate pose of the camera and using a 3D map, the location of the content can be induced automatically. For cases where it is difficult to label the accurate pose of the camera, it is a good alternative to directly label the accurate location of the content. To control the precision on the keyframes, we categorized the videos into two types based on the camera’s position on the z-axis, and designed two different labeling strategies for each category. The videos in the first category are near the ground on the z-axis, and are denoted as the street view videos. They were usually filmed by reporters and spectators, examples are shown in Figure 3(a). The videos in the second category are at least one floor away from the ground on the z-axis, and are denoted as the 45-degree-view videos. They were usually filmed by people in nearby buildings and from the helicopters of news stations, examples are shown in Figure 3(b).

For the street view videos, we leveraged Google Street View[6] as a reference to get the ground truth for the camera pose. To be specific, we asked volunteers to virtually move around in Google Street View to find a view that looked most similar to a given keyframe from a video. We paired the GPS coordinates, the heading direction, and the pitch of the fitting view from the street view service with the keyframe. The annotation interface is shown in Figure 4. The precision of this approach is limited by two constraints: the step size unit in Google Street View and the width of the road. The first constraint is caused by the fact that users cannot move continuously in Google Street View. Therefore, we measured the step size unit in Street View by dividing the length of the street at the finish line (182 meters) by the number of steps (17) in Street View. The error range is [0, 5.3] meters. The second constraint is caused by the fact that the street view is captured by a camera mounted on the roof of a car driving around in the Boston area. As the width of the street in the area is 27 meters[8], the error range is [0, 13.5] meters. Putting both constraints together, we can estimate that the error range is [0, 5.3] meters along the road and [0, 13.5] meters across the road.

For the 45-degree-video, we leveraged Google 3d maps[5] to label the ground truth of a video content. To be specific, we put a marker on the 3D map for the center position of the content of the keyframe based on the reference buildings in the key frame. The labeling interface is shown in Figure 5. The precision of this approach is inversely proportional to the width of the reference buildings. As there are 11 buildings along one side of the street block next to the finish line, we estimated the average width of the buildings in this area as 16.5 meters. That is, the error range is [0, 8.3] meters.

Table 2 summarizes the labeling strategies and error ranges for the localization ground truth.

### 3.3 Auxiliary Data Collection

We collected four kinds of auxiliary data: video comments with timestamp, map meta-data, environment images, and a 3D point cloud of the environment. We additionally provide these auxiliary data in the dataset because they contain
Table 2: Labeling strategies and error ranges for localization ground truth

<table>
<thead>
<tr>
<th>z-position</th>
<th>location of video content</th>
<th>camera pose</th>
</tr>
</thead>
<tbody>
<tr>
<td>street view</td>
<td>induced from camera pose</td>
<td>labeling by using Google Street View as reference ([0, 13.5])</td>
</tr>
<tr>
<td>45-degree-view</td>
<td>labeling by using buildings on the 3d map as references ([0, 8.3])</td>
<td>NA</td>
</tr>
</tbody>
</table>

Table 3: Example of video comments with timestamp

<table>
<thead>
<tr>
<th>Time</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>0:08</td>
<td>pause it at 0:08, there’s nobody injured in the middle of the cross walks, but skip ahead to 3:08 and we see someone being helped on the ground. Wtf?</td>
</tr>
<tr>
<td>1:25</td>
<td>did anyone else see at 1:25 there was a national guards man on his phone my thoughts if I were him oh there were bombs in the marathon nah they don’t need me I’ll just post a selfie on instagram</td>
</tr>
<tr>
<td>0:50</td>
<td>The lady in green, seems to throw something back to the old man that falls to his knee’s, pretty strange. Oh and btw there is an Illuminati card called the jogger, with a picture of a lady in green, also blonde, also wearing a head band, lol.</td>
</tr>
</tbody>
</table>

background and reference information for synchronization and localization. To be specific, the video comments with timestamp provides description of content at a particular time in video; the map meta-data provides the contour of the buildings and the location of roads; the environment images provide street view images and satellite images with GPS coordinates; the 3d point cloud provides a 3d model of the surroundings.

Video comments with timestamp: Youtube provides timestamp in comment service to its users. We crawl the comments for the videos from Youtube and parse the timestamp from comment. Altogether we collect 61 comments with timestamp for 16 videos. Table 3 shows some examples of the video comments with timestamp

Map meta-data: From Openstreetmap[7], we collected the map meta-data from a rectangular area covering the Boston Marathon finish line. To be specific, in latitude it ranges from 42.347306 to 42.350676 and in longitude it ranges from −71.087848 to −71.075212. The collected meta data is visualized in Figure 6: buildings are highlighted in red, green and blue colors; roads are highlighted in yellow; crossing points are highlighted in black. In total, the map meta-data consists of 562 buildings, 39 roads, and 39 crossing points.

Environment images: We collected two kinds of environment images: street view images and 45-degree-view images. To collect the street view images, we uniformly sampled GPS coordinates along the roads at a step size of 8 meters, resulting in 683 GPS coordinates. At each GPS coordinate, we sampled 36 headings (direction in the xy-plane) and 3 pitches (angle along the z-axis) to get 108 images of different views. Figure 7 illustrates a schema for the collection of the street view environment images. This kind of environment images are useful for locating street view clips. To collect the 45-degree-view images, we uniformly sampled GPS coordinates on a grid with a step size of 10 meters, resulting in 440 GPS coordinates. At each GPS coordinate, we crawled 45-degree-view images in 4 directions. Figure 8 illustrates a schema for the collection of the the 45-degree-view environment images. This kind of environment images are useful for locating 45-degree-view clips. Both street view images and 45-degree-view images were crawled from Google maps[5].

3d point cloud of the environment: We used visualSFM[18][19] to reconstruct the 3d point cloud of the environment from the street view environment images. The 3d point cloud is useful in several ways, including estimating the camera pose of a clip, rendering a reconstructed event, and providing a 3d map for tracking. Figure 9 shows three views of the 3d point cloud.

4. DATASET STATISTICS AND ANALYSIS

4.1 General Statistics and Diversity
The dataset includes 347 related videos of 5 hours 29 minutes after manual cleaning. The mean video length is 2 minutes 8 seconds. Sorting videos by length, we have the 10% percentile at 30 seconds and the 90% percentile at 3 minutes 57 seconds. The videos in the dataset are recorded by different devices, including the spectators’ cell phones, first-person views from the runners’ Go-pro cameras, the reporters’ professional cameras, and from cameras in news stations’ helicopters, as shown in Figure 10. The video content is also diversified. The audio data includes the sounds of explosion, human voice, ambulance alarm, and so on. The vision data includes buildings, smoke, people, and so on. Furthermore, the filmed people can be classified into several groups: the runners, the spectators, the policemen, the volunteers, the reporters, and so on.

These videos are further split into 1,066 clips. Most of these clips contain only one raw footage. We manually removed clips that are not on-site footages, e.g. clips of a news studio. After an automatic duplication detection, we got 109 unique clips. On the timeline, 3 of them were categorized to the period of the pre-explosion, 26 of them were categorized to the period of the explosions, and 51 were categorized to the period of the post-explosion. 29 of them remained uncategorized due to insufficient clues that a human labeler can detect. For 68 street view clips, we collected the ground truth of the camera pose on 508 frames in total. For 41 45-degree-view clips, we collected the ground truth of the content location on 504 frames in total.

Considering that there were only 13 seconds between the two explosions, the content of our dataset is very rich for this period since we have 26 unique clips of the explosions. At the z-position, 41 of them are categorized as 45-degree-view and 68 of them are categorized as street view. On the xy-plane, we plotted the heat map of 109 unique clips on the map based on the labeled localization ground truth. As shown in Figure 11, the clips are distributed along Boylston Street. Most of them are close to the two explosion points, especially to the first explosion point which was almost at the finish line.

4.2 Localization Ground Truth Analysis

For each street view clip, we analyzed the moving distance accumulated by the cameras, the heading rotation, and the pitch rotation. Figure 12 shows the distribution of the accumulated moving distance, and the rotation angles, respectively. On average, the accumulated moving distance within a clip is 16 meters and the accumulated rotation angles within a clip is at 173 degrees for heading and at 39 degrees for pitch.

To illustrate the meaning of the accumulated moving distance, we give an example in Figure 13. It is a clip with a large accumulated moving distance of 320 meters. The first two rows show the sampled key frames along the time line. The third row shows the heat map of the camera position in the clip. At the beginning, the camera resided at the finish line to capture runners reaching the end. Please refer to
the frames 1, 2 for an illustration. Then the explosion happened and the camera moved forward to the first explosion point. Please refer to the frames 3, 4, 5, 6, 7 for an illustration. Meanwhile, the policemen and the volunteers arrived to rescue the victims. The camera man left the explosion point to record the evacuation and assistance. Please refer to the frames 8 for an illustration. Finally, the camera man returned to the first explosion point to record the evacuation and assistance. Please refer to the frames 9, 10 for an illustration.

To illustrate the meaning of accumulated heading and pitch rotation, we give two examples in Figure 14. One clip contains a large accumulated rotation and one clip contains a small accumulated rotation.

In the first clip, the camera initially aimed at the marathon race. Then the first explosion happened and the camera turned left to capture the explosion. 13 seconds later, the second explosion happened and the camera turned right to capture this explosion. Finally the camera man seemed to panic and to figure out what was happening, and turned the camera up and down, left and right for several times.

The second clip was filmed by a Go-pro camera mounted at the forehead of a runner. Thus, the runner keeps looking ahead with little rotation while running. The clip ends when the explosion happened.

4.3 Synchronization Ground Truth Analysis

Our hierarchical labeling approach leads to the result that the ground truth of different clips varies in granularity: the event period and the exact time. We first analyze the ground truth of the exact time. Figure 15 lists several clips synchronized using the exact time as ground truth. It shows that the ground truth of the exact time is very accurate, and could be used in an evaluation of automatic synchronization. However, it is very difficult to manually label the exact time in all the clips. Figure 16 shows 4 clips that do not provide enough clues to enable volunteers to label the exact time. As we only collect the ground truth for the exact time in parts of the data, additional manual verification is required to evaluate the time algorithm’s output in an unbiased way. Furthermore, such additional manual verification could be used to augment the ground truth of the exact time which would be difficult to be labeled exhaustively.

5. EVALUATION METRICS

We propose the following evaluation metrics for two granularities: for the clip level and for the frame level.

**Clip level:** This is a coarse granularity. For the synchronization, the clip level metric requires the algorithm to output those clip pairs that have an intersection on the timeline. No specific offset is required. For the localization, the clip level metric requires the algorithm to output one GPS coordinate for the content of each clip. Neither a camera location nor a frame level localization is required. The clip level output is useful for data organization: for grouping the clips by time and by location.

**Frame level:** This is a finer granularity. For the synchronization, the frame level metric requires the algorithm to output the offset between the aligned clip pairs in seconds. For the localization, the frame level metric requires the algorithm to output the camera location for each frame. The frame level output provides precise meta-data for a detailed analysis, e.g. for the reconstruction of the entire event ground truth from the synchronized clips and the camera locations.

5.1 Metrics for Evaluating Synchronization

Both clip level and frame level metrics are defined on clip pairs. Recall that on the clip level, one pair is considered as correct if the clips have an intersection on the timeline. For the metric for clip level synchronization, we calculate the mean average precision as a function of the number of output pairs.

\[
p@K = \frac{\text{#correct pairs}}{K}\]

\[
\text{map@K} = \sum_{k=1}^{K} p@k
\]

A helpful property of this metric is that different algorithms can be compared for different numbers of output pairs.

Since the provided labels do not cover the complete ground truth, it is a little bit tricky to calculate this metric in practice. First, we encourage researchers to manually verify the top output pairs in addition to the provided labels to make the metric value more accurate. Meanwhile, additional labels can be added to expand the ground truth, which will benefit future research. Second, it is not necessary to increase the number of output pairs to all possible pairwise numbers for two reasons. The chance of missing ground truth increases as the number of output pairs grows, and the manual verification effort increases dramatically. Our
Figure 12: Distribution of accumulated moving distance, heading rotation, and pitch rotation

Figure 13: Illustration of accumulated moving distance
Figure 14: Illustration of accumulated rotation in heading and pitch

(a) 2, 107° in heading, 126° in pitch

(b) 6° in heading, 3° in pitch
suggestion is to stop increasing the output pair number $K$ when the precision drops to a certain value, e.g., 10%. The highest output pair number $K_{p=0.1}$ is proportional to the recall at a precision of 10%. This can be used as a proxy for the real recall because the real recall value cannot be calculated until all ground truth pairs are labeled.

In summary, the synchronization algorithms are compared on clip level by plotting the map curve (MAP-CURVE): $(1, \text{map@1}), (2, \text{map@2}), \ldots, (K_{p=0.1}, \text{map@}K_{p=0.1})$ and by limiting the output pair number (STOP-NUMBER): $K_{p=0.1}$.

For the synchronization metric on frame level, we calculate the RMSE of the offset only for the ground truth pairs.

$$RMSE = \left( \sum_{p \in \text{correct pairs}} (\text{offset} - \widehat{\text{offset}})^2 \right)^{1/2}$$

where offset is the ground truth offset, and $\widehat{\text{offset}}$ is the predicted offset.

### 5.2 Metrics for Evaluating Localization

For the clip level, we use the mean GPS coordinates of the clip as the ground truth, and calculate the distance between the predicted GPS coordinates and the ground truth in meters for each clip. Different applications have different precision requirements for the ground truth of the localization. For example, a 50 meter-precision is acceptable for a rural area while in a city center a 20 meter-precision is preferred. Considering that the error range of the ground truth itself is at around 10 meters, we derive a correctness radius $r = 20$, a predicted location is considered correct if its distance to the ground truth is within 20 meters. We calculate the precision for each correctness radius $r$: $Pr=r=20, Pr=r=30, Pr=r=40, Pr=r=50$.

For the frame level, we calculate the distance between the predicted camera position and the ground truth in meters, and the angle between the predicted heading/rotation and the ground truth in degrees.

### 6. CONCLUSION AND FUTURE WORK

We introduce the first real-world dataset for event reconstruction, the event of the 2013 Boston Marathon finish line explosions, for synchronization and localization to promote research in this field. We show that the videos included in the dataset are quite diverse as they cover different moments of the event at different positions from different perspectives. We provide high precision ground truth labels for localization and two granularity ground truth labels for synchronization. We derive a number of metrics on video level and frame level to evaluate the two tasks. In the future, we will provide baseline performances on these metrics. Furthermore, we will expand the ground truth data based on the output of automatic algorithms.

### 7. ACKNOWLEDGMENTS

This project was conducted in partnership with Carnegie Mellon’s Center for Human Rights Science (http://www.cmu.edu/chrs). The authors would like to thank the MacArthur Foundation, Oak Foundation, and Humanity United for their generous support of this collaboration. This material is based in part upon work supported by the National Science Foundation under Grant Number IIS-1638429. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

### 8. REFERENCES

