ABSTRACT

We present a novel dataset for multi-view video and spatial audio. An ensemble of ten musicians from the BBC Philharmonic Orchestra performed in the orchestra’s rehearsal studio in Salford, UK, on 25th March 2014. This presented a controlled environment in which to capture a dataset that could be used to simulate a large event, whilst allowing control over the conditions and performance. The dataset consists of hundreds of video and audio clips captured during 18 takes of performances, using a broad range of professional- and consumer-grade equipment, up to 4K video and high-end spatial microphones. In addition to the audiovisual essence, sensor metadata has been captured, and ground truth annotations, in particular for temporal synchronisation and spatial alignment, have been created. A part of the dataset has also been prepared for adaptive content streaming. The dataset is released under a Creative Commons Attribution Non-Commercial Share Alike license and hosted on a specifically adapted content management platform.

Categories and Subject Descriptors

H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems; I.4.1 [Image Processing and Computer Vision]: Digitization and Image Capture

Keywords

Multi-view, spatial, omnidirectional, sensor, mobile

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the Owner/Author.

MMSys ’15 Mar 18-20, 2015, Portland, OR, USA
Copyright is held by the owner/author(s).
ACM 978-1-4503-3351-1/15/03.
http://dx.doi.org/10.1145/2713168.2713191

1http://www.icosole.eu

General Terms

Measurement, Experimentation

1. INTRODUCTION

Many cultural and sports live events do not take place at only a single spot, but are spread out over different stages, halls, cities or even regions, with different actions happening in parallel in each of these places. Examples are music festivals with several stages or tents, city festivals, parades, marathons or bike races. In order to enable immersive coverage of this type of events, the ICoSOLE project1 is developing technologies for live capture and streaming from professional and consumer devices, fusion of audio and video content from heterogeneous devices into a format agnostic representation, and methods for analyzing and filtering streams based on quality and content properties. The approach uses a variety of sensors, ranging from mobile consumer devices over professional broadcast capture equipment to panoramic and/or free-viewpoint video and spatial audio. Methods for streaming live high-quality audiovisual content from the capture devices to the production system and to the consumers are being developed, as well as content acquisition, processing and editing services.

In order to support development of capture, analysis, production and streaming technologies, a dataset has been collected. The event being captured was a studio recording of the BBC Philharmonic Orchestra. The studio setting enables full control, i.e., to repeat capture with different parameters (e.g., spatial setup, light conditions, presence of a small audience), and to capture measurements as ground truth which could not easily be obtained at a real public event with live audience.

This paper presents the publicly available dataset. Section 2 describes the recorded essence (video and audio material) and metadata. Section 3 describes the ground truth annotations and multiple resolutions (for testing adaptive
streaming technologies) that have been created. The platform for managing the dataset and providing public access is described in Section 4. An example of using the provided sensor data and annotations of a user generated video is provided in Section 5, and Section 6 concludes the paper.

2. CONCERT RECORDING DATASET

A series of musical performances has been captured using both professional- and consumer-grade equipment. An ensemble of ten musicians from the BBC Philharmonic Orchestra performed in the orchestra’s rehearsal studio in Salford, UK, on 25th March 2014 (see Figure 1). This presented a controlled environment in which to capture a dataset that could be used to simulate a large event, whilst allowing control over the conditions and performance.

Like the Jiku dataset [11] this dataset covers live events and performances, but differs in several aspects. While Jiku consists only of user generated content, the dataset described in this paper includes both high-end video and audio content, as well as user-generated content. It is restricted to one event, but provides several setups of performances in this location. For all of them a rich set of sensors, as well as spatial and temporal synchronization ground truth is provided.

The musicians performed four pieces of music at different locations within the concert hall. For each piece several takes were captured, both with and without users moving around with capture devices, in order to obtain clean audio recordings. Multiple locations were used so that content may be composited to allow simulation of multiple simultaneous performances, as one might find at a festival, which is relevant to the aims of the ICoSOLE project. In one session the musicians split into two smaller ensembles and played different pieces simultaneously. In another session a small audience was also brought in to watch the performance.

In total, 18 takes have been recorded, resulting in a total of 160 video and 1,959 audio clips. The dataset should be useful for research into multimedia systems that integrate professional and user-generated content, as well as research into spatial audio and mobile video systems. The dataset is released under a Creative Commons Attribution Non-Commercial Share Alike license (CC BY-NC-SA 4.0).

2.1 Essence

2.1.1 Video

Video was captured at high definition (HD) and ultra-high definition (UHD, providing 4K resolution) using broadcast quality cameras. Two Canon XF305 cameras were used to capture HD video using a variety of different shots and perspectives. A Sony PMW-F55 camera was used to capture a wide shot of the concert hall in UHD from a static position at the rear of the hall (see Figure 4 for examples). A range of mobile devices was also used to capture video. Table 1 provides a list of devices used and the video formats in which the data is stored.

Figure 1: Concert recording session.

2.1.2 Audio

Full audio multi-track recordings were made with high quality microphones using techniques commonly used in professional surround sound recording, as well as several advanced microphone arrays for capturing spatial audio with height. Figure 1 shows many of these microphones in use during one of the recording sessions. This set of audio recordings can be used to investigate 3D spatial audio production techniques, both for loudspeakers and headphones, as well as interactive audio applications such as scene navigation.

Two 9-channel spaced microphone arrays described in [13] were used, one OCT90 surround array plus 4 super-cardioids pointing upwards and a widely spaced 9-channel A/B setup, which uses omnidirectional microphones. These microphone arrays are both designed to provide low cross-talk between channels, the OCT array using directional microphones and the A/B array using wide microphone spacing. These microphones were positioned approximately 2-5m away from the musicians. Two Hamasaki square [7] arrays were configured at the rear of the concert hall to capture the ambience with minimal direct sound. The arrays were coincident but oriented to achieve orthogonal directivity patterns. A Neumann KU100 dummy head microphone was used to capture a binaural recording at the same position as the OCT array. In addition, a number of “spot” microphones were placed close to the instruments to capture better isolated recordings.

An Eigenmike 32-capsule spherical microphone array was also placed at various positions in the hall through the recording sessions. An array of six Schoeps SuperCMIT hyper-cardioid “shotgun” microphones with digitally enhanced directivity were placed in the middle of the hall, approximately 12m from the musicians, along with a Schoeps Double MS array. The microphones were near coincident in the x-y plane.

All professional audio signals were recorded at a sample rate of 48kHz with 24-bit resolution. The user-grade mobile devices also recorded audio signals, the audio formats in which the data is stored are provided in Table 2.

2.2 Captured Metadata

Together with the content, various types of metadata were captured. This includes positions and properties of the pro-
3. ANNOTATIONS AND DERIVED DATA

3.1 Annotations

Three types of ground truth annotations have been created for the data. The first is temporal synchronization information, which provides information about the offset of each recorded media item wrt. to the studio clock. All professional capture and recording equipment was synchronized to a master clock signal. A clapper board with timecode display was also used at the start and end of each take, to allow synchronization of user-grade recordings. This synchronization information is represented in EBU Core [5].

The second is spatial alignment information, which describes the overlap between the area of the scene captured by mobile devices with the 4K overview shot. The data is thus linked to the 4K videos, and describes the area seen by each mobile device as a region of the 4K frame at time points throughout the recording (see Figure 4 for an example). The times are specified as relative timestamps of the amount of motion of the device. The metadata is represented using the MPEG-7 Audiovisual Description Profile (AVDP) [1], using moving regions to describe the sequence of time indexed polygons for each device.

The third type of metadata is annotations about sensor data correctness (e.g. reliability of orientation changes) and visual quality annotations (e.g., stability, noise). For each user generated video, time segments in which sensor data was found to be unreliable as well as segments with impairments of the visual quality have been annotated. This metadata is also represented using MPEG-7, with the extensions for quality description proposed in [3].

3.2 Adaptive Streaming Records

Video streaming currently uses multiple technologies and protocols, but most content delivery platforms seem to con-

### Table 1: Captured video formats.

| Device          | Codec          | Bitrate | Resolution | Frame rate |
|-----------------|----------------|---------|------------|------------|
| Canon XF105     | MPEG-2 4:2:2   | 50Mbps  | 1920x1080  | 25fps      |
| Sony PMW-F55    | H.264 High 4:2:2 Intra | VBR (≤ 600Mbps) | 4096x2160 | 50fps     |
| Go Pro          | H.264 High 4:2:0 | 35Mbps  | 1920x1080  | 50fps      |
| iPhone 4S       | H.264 Base 4:2:0 | 21.1Mbps | 1920x1080 | 29.7fps    |
| iPad mini       | H.264 High 4:2:0 | 12.4Mbps | 1920x1080 | 29.7fps    |
| iPad 3          | H.264 Base 4:2:0 | 21.1Mbps | 1920x1080 | 29.7fps    |
| Galaxy S3       | H.264 High 4:2:0 | 17Mbps  | 1920x1080  | 30fps      |
| Galaxy S4       | H.264 High 4:2:0 | 17Mbps  | 1920x1080  | 29.7fps    |
| Galaxy Note     | H.264 High 4:2:0 | 17Mbps  | 1920x1080  | 30fps      |
| Nexus 4         | H.264 Base 4:2:0 | 9.6Mbps | 1920x1080  | 23.97fps   |
| Nexus 5         | H.264 Base 4:2:0 | 12.4Mbps | 1920x1080 | 30.33fps   |
| Nexus 7         | H.264 Base 4:2:0 | 17Mbps  | 1920x1080  | 29.7fps    |
| Nexus 10        | H.264 Base 4:2:0 | 8.3Mbps | 1920x1080  | 20.83fps   |

### Table 2: Captured audio formats.

| Device          | Format                   |
|-----------------|--------------------------|
| Professional microphones | PCM 24-bit 48kHz |
| Go Pro          | Stereo AAC 48Hz 128Kbps  |
| iPad mini       | Mono AAC 44.1kHz 64Kbps  |
| iPad 3          | Mono AAC 44.1kHz 64Kbps  |
| Galaxy S3       | Stereo AAC 48Hz 128Kbps  |
| Galaxy S4       | Stereo AAC 48Hz 128Kbps  |
| Galaxy Note     | Stereo AAC 48Hz 128Kbps  |
| Galaxy Nexus    | Mono AAC 48Hz 96Kbps     |
| Nexus 4         | Mono AAC 48Hz 96Kbps     |
| Nexus 5         | Mono AAC 48Hz 96Kbps     |
| Nexus 7         | Mono AAC 48Hz 96Kbps     |
| Nexus 10        | Mono AAC 44.1kHz 96Kbps  |

### Table 3: Summary of captured sensors.

| Sensor          | ID | Parameters | units |
|-----------------|----|------------|-------|
| Location        | 0  | latitude [deg], longitude [deg], altitude [m], accuracy [m] |
| Accelerometer   | 1  | acceleration x [m s⁻²], acceleration y [m s⁻²], acceleration z [m s⁻²] |
| Gyroscope       | 2  | rotation x [rad s⁻¹], rotation y [rad s⁻¹], rotation z [rad s⁻¹] |
| Magnetic field  | 3  | geomagnetic field x [μT], geomagnetic field y [μT], geomagnetic field z [μT] |
| Orientation     | 4  | azimuth [deg], pitch [deg], roll [deg] |
| Rotation vector | 5  | rotation vector x [deg], rotation vector y [deg], rotation vector z [deg], scalar component [°] |
| Ambient Light   | 6  | ambient light level [lx] |
| Proximity       | 7  | object distance [cm] |
| Pressure        | 8  | air pressure [hPa or mbar] |

3 using magnetic field and accelerometer
4 not provided for all devices
verge towards adaptive streaming solutions, like DASH (Dynamic Adaptive Streaming over HTTP) [12], or HLS as the method to stream videos. Therefore a subset of the concert recording dataset has been edited for adaptive streaming.

The pull-based approach of such adaptive streaming technologies is premised on an intelligent client, which selects the most applicable media representation out of a given set, by issuing HTTP requests for individual segments. The selection is based on the users’ context, like device capabilities, personal preferences and current network conditions.

For adaptive streaming preparation a subset of the material recorded in Take 8/Session 2 has been adopted. We have chosen a professional high quality video recording (Sony PMW-F55), as well as user generated content from a smartphone (Galaxy S4) and a tablet (Nexus 7), containing both audio and video. The sequences have lengths between 03:14 and 05:55.

The content of this dataset is provided in different representations, as shown in Table 4. To achieve a maximum platform coverage, we make use of DASH and HLS as streaming service. For both technologies, segments are encoded and multiplexed with four seconds length, based on the evaluation of [8] and [9]. The content has been encoded at 7 different video representations, ranging from 500kbps at 640 × 360 up to 20Mbps at 4096 × 2160. The lower bitrates and resolutions are suitable for mobile devices like smartphones or devices experiencing limited network connectivity, whereas the higher bitrates and resolutions could be used for personal computers, TV sets or set-top boxes in a stable high throughput network environment.

Additionally, two different audio representations are provided with either one or two channels at 96 and 128kbps using a 48kHz sampling rate. The audio and video representations are offered separately enabling the client to choose the appropriate audio and video bitrate independently from each other.

In the context of DASH we make use of DescriptorType elements [2], in the media presentation description (MPD), to describe the annotations (discussed in Section 3.1), shown in Listing 1. The annotations can also be part of an AdaptationSet or Representation, if content from different sources is present within the MPD.

| Resolution | Galaxy S4 | Nexus 7 | Sony PMW-F55 |
|------------|-----------|---------|--------------|
| 640 × 360  | 500kbps   | 500kbps | 500kbps      |
| 960 × 540  | 1Mbps     | 1Mbps   | 1Mbps        |
| 1280 × 720 | 2Mbps     | 2Mbps   | 2Mbps        |
| 1920 × 1080| 3Mbps     | 3Mbps   | 3Mbps        |
| 1920 × 1080| 6Mbps     | 6Mbps   | 6Mbps        |
| 4096 × 2160| -         | -       | 10Mbps       |
| 4096 × 2160| -         | -       | 20Mbps       |

Table 4: Video bitrates and resolutions.

The dataset was produced using bitrate⁵ and can be played with adaptive streaming clients with DASH or HLS support⁶. The segment format is based on the ISOBMFF.

4. THE MAMMIE PLATFORM

The content set is made publicly available for research purposes by the ICoSOLE consortium. It is accessible by means of a web platform called “mammie” for which anyone can request an account. After registering, the user has access to the content and associated ground truth information for which they accepted the terms and conditions during the registration process. After logging into the mammie platform, the user is presented an overview of the content in the platform. This overview is presented as a list of audiovisual items present in the content set.

On the top of the overview window, a search box are shown. The mammie system searches for the inserted query in all metadata streams attached with the content items. On the left hand side of this overview, a list of filters is shown. The user can filter the content list according to the contributor, event, device, session, take, type and license. Upon clicking on a filter, the set of videos shown on the right hand side is reduced to only the items matching the filter.

On the right, every item is shown with a key frame (Figure 2). When this key frame is clicked, a low-res stream of the item starts playing in the browser. On the right of this key frame, the title, license and id is shown. A list of filters applied to this item is displayed underneath the id. Lastly, individual links of different streams associated with this item are shown. When a user clicks such a link, the stream will start downloading.

On the top right, some extra buttons are shown (depending on the access rights given to the user), see Figure 3. A

Listing 1: MPD file with annotations.

```xml
<MPD xmlns="urn:mpeg:dash:schema:mpd..." xmlns:mat="http://www.dash몽기/dash.js"

<ProgramInformation>

<Title>Take8/Session2 UGC</Title>
</ProgramInformation>

<Annotations>
  <schemeIdUri="urn:mpeg:DASH:annotation:2014"
  value="EbuCore"
```

The dataset was produced using bitrate⁵ and can be played with adaptive streaming clients with DASH or HLS support⁶. The segment format is based on the ISOBMFF.

8. THE MAMMIE PLATFORM

The content set is made publicly available for research purposes by the ICoSOLE consortium. It is accessible by means of a web platform called “mammie” for which anyone can request an account. After registering, the user has access to the content and associated ground truth information for which they accepted the terms and conditions during the registration process. After logging into the mammie platform, the user is presented an overview of the content in the platform. This overview is presented as a list of audiovisual items present in the content set.

On the top of the overview window, a search box are shown. The mammie system searches for the inserted query in all metadata streams attached with the content items. On the left hand side of this overview, a list of filters is shown. The user can filter the content list according to the contributor, event, device, session, take, type and license. Upon clicking on a filter, the set of videos shown on the right hand side is reduced to only the items matching the filter.

On the right, every item is shown with a key frame (Figure 2). When this key frame is clicked, a low-res stream of the item starts playing in the browser. On the right of this key frame, the title, license and id is shown. A list of filters applied to this item is displayed underneath the id. Lastly, individual links of different streams associated with this item are shown. When a user clicks such a link, the stream will start downloading.

On the top right, some extra buttons are shown (depending on the access rights given to the user), see Figure 3. A
user can request a certificate (public key) to be able to log
in to the platform without the need to enter a password
(instead using the installed public key). Content is also ac-
cessible via a REST service (documentation can be found by
clicking on the button). Users with permission will be able
to add content to the platform by clicking the Upload new
button. The download of all of the current items listed (with
the current filter(s) applied), is possible via the Download all button. The two right buttons clear all filters and log out of the platform.

When a user logs into the platform, a temporary token is
granted. This can be used to access the REST functions.
If the certificate option is available, this can be used to ac-
cess the REST interface as well. This way, the user can
navigate through the content set and download items that
would like to use for their research. The technical de-
tails of the mammie platform are omitted in this paper, but
the interested readers can find the technical description of
the mammie platform in the paper by Braeckman et al. [4].
The mammie platform can be reached through its web URL
https://icosole.lab.vrt.be.

5. EXAMPLE

We show the use of sensor data and ground truth for one
element of the video captured with a mobile phone (Sam-
sgun Galaxy S3) in one of the takes. Using the temporal
synchronization ground truth, we can select both the video
and the captured sensor data for the correct time segment.
Using the accelerometer data, we can remove the amount
caued by gravity and obtain the accelerations (and thus de-
termined the instantaneous speed) for each sample wrt.
the device coordinate system. Together with the rotation vector
data, which provides the rotation of the device wrt. a coor-
dinate system with one axis pointing to the magnetic north
and one upwards perpendicular to the ground plane, we can
transform the speed into the world coordinate system and
use it to obtain relative position updates. We have to resort
to relative position updates, as absolute position metadata
is not precise enough to capture the small movements in the
scene, in particular as the content was recorded indoors. The
orientation vector can be used to determine the view direc-
tion of the device. Figure 5 shows the plot of the device
path and the orientations for each sampled time point.

The spatial alignment ground truth can be used to verify
the sensor information. As it is typically coarse and may
drift due to incremental updates, visual information is likely
to be used for determining captured area of the scene more
precisely. Methods for this problem can be evaluated by
comparing the determined area with the annotated spatial
overlap of the area captured by the user devices and the
4K overview shots. For the example, the area of the 4K
recording captured at four time points throughout take 5 is
visualized in Figure 4.

6. CONCLUSION

In this paper we have presented a comprehensive event
recording dataset, using a broad range of professional and
consumer devices, and different technologies for spatial au-
dio capture. Due to the fact of recording in a controlled
studio environment, the dataset enables researchers to work
comparatively with content that differs only in one charac-
teristic, such as the setup of the musicians, capture condi-
tions or the equipment and technologies used. For many
devices, in particular the mobile consumer devices, a range
of sensor data has been captured and is available with the
essence. This dataset provides a good basis for the evalua-
tion of technologies for improving localization, fusing au-
diovisual content from multiple sources or creating virtual
scenes from spatial audiovisual content. The sensor data and
the spatial alignment information will be used to assess the
reliability of different sensor types, and to evaluate methods
for image-based localization. In addition, the sensor data
and the quality annotations will be used to evaluate auto-
matic quality analysis on mobile devices from both sensor
data and content analysis. The dataset will also be used to
investigate immersive audio production and rendering, in-
cluding audio aspects of spatial scene navigation.

To the best of our knowledge, we have created the first
adaptive streaming dataset including annotations about lo-
cation, time and audiovisual parameters. It also comprises
two different streaming technologies, namely DASH and HLS,
to achieve a wide platform coverage, in up to 4K resolu-
tion. This dataset has been especially designed for compar-
isons and evaluations and consists of user generated con-
tent as well as professional video recordings. Based on this
dataset we plan to evaluate different possibilities in con-
text of MPEG-DASH, like spatial navigation for immersive
content exploration, leveraging viewpoints in DASH playout
scenarios. We will also evaluate stream switching algorithms
in different network scenarios (cf. [10] and [6]).
Acknowledgement

The authors thank the musicians and staff of the BBC Philharmonic Orchestra, and Elisabeth Höldrich for providing the spatial alignment and sensor quality annotations. The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 610370, ICoSOLE (“Immersive Coverage of Spatially Outspread Live Events”, http://www.icosole.eu).

7. REFERENCES

[1] Information technology – Multimedia content description interface – Part 9: Profiles and levels, Amendment 1: Extensions to profiles and levels. ISO/IEC 15938-9:2005/Amd1:2012, 2012.

[2] Information technology – Dynamic adaptive streaming over HTTP (DASH) – Part 1: Media presentation description and segment formats. ISO/IEC 23009-1:2014, 2014.

[3] W. Bailer and P. Schallauer. An MPEG-7 extension for describing visual impairments. In Proc. of International Workshop on Image Analysis for Multimedia Interactive Services, Klagenfurt, AT, 2008.

[4] K. Braeckman, R. De Sutter, M. Matton, and T. Blomme. A media sharing platform built with open source software. In Proc. of International Conference on Distributed Multimedia Systems, Illinois, USA, October 2010.

[5] EBU Core Metadata Set (EBU Core), v. 1.5. EBU Tech 3293, Apr. 2014.

[6] R. Grandl, K. Su, and C. Westphal. On the interaction of adaptive video streaming with content-centric networking. In Proc. of International Packet Video Workshop, 2013.

[7] K. Hamasaki. Multichannel Recording Techniques for Reproducing Adequate Spatial Impression. In Proceedings of AES 24th International Conference on Multichannel Audio, 2003.

[8] S. Lederer, C. Mueller, and C. Timmerer. Dynamic adaptive streaming over HTTP dataset. In Proc. of ACM Multimedia Systems Conference, 2012.

[9] S. Lederer, C. Mueller, C. Timmerer, C. Concolato, J. Le Feuvre, and K. Fliegel. Distributed DASH dataset. In Proc. of ACM Multimedia Systems Conference, 2013.

[10] C. Mueller, S. Lederer, and C. Timmerer. A proxy effect analysis and fair adatpation algorithm for multiple competing Dynamic Adaptive Streaming over HTTP clients. In Proc. Visual Communications and Image Processing, 2012.

[11] M. Saini, S. P. Venkatagiri, W. T. Ooi, and M. C. Chan. The Jiku Mobile Video Dataset. In Proc. of ACM Multimedia Systems Conference, 2013.

[12] T. Stockhammer. Dynamic Adaptive Streaming over HTTP - Design Principles and Standards. In Proc. of ACM Conference on Multimedia Systems, 2012.

[13] G. Theile and H. Wittek. Principles in Surround Recordings with Height. In Proc. of International Conference on Spatial Audio, 2011.