Exploratory Multimodal Data Analysis with Standard Multimedia Player

Multimedia Containers: A Feasible Solution to Make Multimodal Research Data Accessible to the Broad Audience

Julius Schöning¹, Anna L. Gert¹, Alper Açık², Tim C. Kietzmann³, Gunther Heidemann¹ and Peter König¹

¹Institute of Cognitive Science, Osnabrück University, Osnabrück, Germany
²Psychology Department, Özyeğin University, Istanbul, Turkey
³Cognition and Brain Sciences Unit, CB2 7EE, Medical Research Council, Cambridge, U.K.

{juschoening, agert, gheidema, pkoenig}@uos.de, alper.acik.81@gmail.com, tim.kietzmann@mrc-cbu.cam.ac.uk

Keywords: Multimodal Data Analysis, Visualization, Sonification, Gaze Data, EEG Data.

Abstract: The analysis of multimodal data comprised of images, videos and additional recordings, such as gaze trajectories, EEG, emotional states, and heart rate is presently only feasible with custom applications. Even exploring such data requires compilation of specific applications that suit a specific dataset only. This need for specific applications arises since all corresponding data are stored in separate files in custom-made distinct data formats. Thus accessing such datasets is cumbersome and time-consuming for experts and virtually impossible for non-experts. To make multimodal research data easily shareable and accessible to a broad audience, like researchers from diverse disciplines and all other interested people, we show how multimedia containers can support the visualization and sonification of scientific data. The use of a container format allows explorative multimodal data analyses with any multimedia player as well as streaming the data via the Internet. We prototyped this approach on two datasets, both with visualization of gaze data and one with additional sonification of EEG data. In a user study, we asked expert and non-expert users about their experience during an explorative investigation of the data. Based on their statements, our prototype implementation, and the datasets, we discuss the benefit of storing multimodal data, including the corresponding videos or images, in a single multimedia container. In conclusion, we summarize what is necessary for having multimedia containers as a standard for storing multimodal data and give an outlook on how artificial networks can be trained on such standardized containers.

1 INTRODUCTION

Multimodal data analysis applications and software are usually tailored to a single analysis task based on a specific dataset. This is because stimulus material (e.g. images and videos), metadata (e.g. object annotations and tags), along with their associated multimodal sensor data (e.g. gaze and EEG) are stored in separate files. Making things even worse, all additional data—gaze trajectories, EEG curves, emotional state descriptors, heart and respiration rates, object annotations etc.—are stored in a diversity of formats, e.g. plain text, XML, MATLAB format, or binary. For the purpose of making these multimodal datasets public, all files are usually compressed into a data archive. The data structures of these archives are also customized and can even be unique. For experts, the use of these data is quite cumbersome and time-consuming, as accessing, visualizing and sonifying them requires special tools. Accordingly, these datasets are very difficult or even impossible to access for the general audience.

Multimedia players can visualize and sonify data. Therefore, it is feasible to encapsulate stimulus material, with multimodal data in a standard multimedia container format, which can be then played back. A similar approach has become common practice for storing text and data, e.g., in the PDF container. State of the art video containers like the open container formats (OGG) (Xiph.org, 2016), MPEG-4 (ISO/IEC, 2003), or MKV (Matroska, 2016) can encapsulate a diversity of data formats, such that they can be interpreted as a single file by standard multimedia players and can be streamed via the Internet. In doing so, exploratory multimodal data analysis with a standard multimedia player is possible. Consequently, multi-
multimodal research data become accessible to the broad audience\(^1\).

Therefore we argue to store the complete spectrum of data:

- **stimuli** like video, images and audio,
- **metadata of the stimuli** like capturing details, object tags, subtitles, and labels,
- **additional object or scene data** like 3D descriptions, online links, scene maps, and object-object or object-scene relations, and
- **sensor data of one or several subjects** like gaze trajectories, heart rate, and EEG curves

in a single multimedia container. Establishing multimedia containers as a standard format for multimodal data will facilitate software applications in various fields, ranging from exploratory analysis tasks, to developing deep neural networks trained on both sensor and annotation data for e.g. bio-empowered face detection or multimedia communication aids for disabled people, like automatic audio scene descriptions based on a virtual viewpoints for blind persons. Furthermore, a standard format will significantly boost accessibility and shareability of these data.

Focusing on the exploratory multimodal data analysis on video and image stimuli with a standard multimedia player, this paper starts with a general section on multimedia data, followed by our proposed methods for storing multimodal data in multimedia container to provide instantaneous visualizations and sonifications in Section 3. Two datasets are introduced in Section 4. The first dataset contains short movies and single frames taken from each movie, together with the corresponding gaze trajectories of multiple observers who viewed those movies and frames in the absence of a specific task. The second data set focuses neurophysiological recordings in a real world environment. It consists of a video along with gaze trajectories and EEG data from one subject. Both datasets were converted to our proposed format. The conversion process, which leads to visualization and sonification, is described in general in Section 5. The statements and the impression from expert and non-expert users, performing explorative multimodal analysis with a multimedia player, are summarized in Section 6. Based on their statements, the converted datasets, and the conversion process, we discuss the benefits of storing multimodal data including the corresponding stimuli in a single multimedia container in Section 7. In the last section, we summarize what is necessary for having multimedia containers as a standard for storing multimodal data.

\(^1\text{cf.: video demonstration https://ikw.uos.de/~7Ecv/publications/VISAPP17}\)

2 **STATE OF THE ART**

Multimedia content used in the domain of entertainment (ISO/IEC, 2003; Bertellini and Reich, 2010), research (Açık et al., 2014; Vernier et al., 2016; Petrovic et al., 2005), or education (Martin et al., 2015; Rackaway, 2010) usually consists of multiple parallel tracks like video tracks, audio tracks, images, subtitle tracks, and metadata related to its content. For storage, distribution, or broadcasting of these multimedia content in the domain of entertainment, these tracks are combined into a single container. To support seeking and synchronized playback of the relevant tracks, multimedia containers have to account for the temporal nature of their payload (cf. Figure 1). In comparison to classical archival file format ZIP (PKWARE Inc., 2016) or TAR (GNU, 2016), the temporal payload handling is a major difference. For converting and packing multimodal data into a single multimedia container which is playable with any standard player, at least the temporal structure, as well as the supported data formats of the container format, must be considered. In brief: Not every payload is suited for every container. Further, various multimedia containers with their encapsulated formats are mentioned.

For DVDs, common multimedia containers, like the formats VOB and EVO, are based on the MPEG-PS (ISO/IEC, 1993, 2015) standard. In MPEG-4 (ISO/IEC 14496, Part 14) (ISO/IEC, 2003), the more modern MP4 format to hold video, audio and timed
text data, is specified. Though MP4 is carefully specified, it handles only the formats introduced in the other parts of this standard, like MPEG-7 (ISO/IEC, 2001), but does not conform to arbitrary video, audio and timed text formats.

The non-profit Xiph.Org (Xiph.org, 2016) foundation initiated the free OGG container format for streaming VORBIS encoded audio files. This container becomes a popular format for streaming multimedia content on the web with the ability to encapsulate Ther eo and Dirac video formats. Thus OGG is nowadays supported by many portable devices. The open the Matroska container format (MKV) (Matroska, 2016) aims at flexibility to allow for an easy inclusion of different types of payload. To establish a standard for multimedia content on the web, a reduced and more restricted version, of MKV serves as a basis for the WEBM (WebM, 2016) format.

3 MULTIMODAL DATA IN MULTIMEDIA CONTAINER

For encapsulating multimodal data which multimedia container format is best? Focusing on the specification only, the best solution is to encode all meta, object, and sensor data in accordance with the elaborate vocabulary of the MPEG-7 standard and encapsulate these encoded data as MP4 containers. Because—unfortunately—no standard media player (like VLC, MediaPlayer, and Windows Media Player) seems to support MPEG-7, an explorative analysis of multimodal data stored in MP4 containers is not possible. To our knowledge, the MPEG-7 support in standard multimedia players is not available—this might be caused by the elaborate specification that requires considerable implementation effort.

However, to provide a solution that makes exploratory multimodal data analysis possible with standard multimedia players, and to highlight the advantages of a single multimedia container file—carrying all multimodal data—a MKV container based approach is proposed here. One advantage is that popular video players, like VLC, support this format well. The flexibility of this multimedia container format allows utilizing a wide selection of data formats, which can be used for the sonification, visualization, and storing of multimodal data. Providing sonifications, one can use common audio formats supported by MKV as WAV or MP3 and encapsulate them as an audio track. By rededicating the subtitle tracks, meaningful visualization of different data streams can be created on top of video content. Other scientific data can be encapsulated into the MKV format, too.

Thus our solution (cf. Figure 1) provides instantaneous visualization and sonification on the one hand, and all data combined in single files on the other hand. Note that all multimodal data combined in a single file can also be stored by packing them into a data archive, but that such an archive does not provide visualizations, sonifications, and is not streamable via the Internet.

Following the previous discussion (Schönig et al., 2016b, 2017) to reuse, or more precisely, on modifying an existing subtitle format for incorporating visualization of sensor data like gaze trajectories, two kinds of multimedia container prototypes were implemented. The first is based on the Universal Subtitle Format (USF) (Paris et al., 2016) and losslessly encapsulates the complete sensor data for visualization. However, for using this approach a slightly modified version of the VLC media player is needed. The second is based on the Advanced Sub Station Alpha (ASS) (SSA v4.00+, 2016) format and is only able to carry selected sensor data like viewpoints, which can be visualized by most current media players.

3.1 Sensory Data as USF

To use USF for encapsulating sensor data, we analyzed which features of USF are made available in the latest releases of common multimedia players. The current version 3.0.0 of VLC already supports a several of USF attributes (cf. Listing 1), which are text, image, karaoke and comment. The latest USF specification introduces an additional attribute shape. Note, the specification is still marked as under development, although the last modifications happened seven years ago. Since almost every visualization can be created out of simple geometric shapes, like ellipses and polygons, the use of the shape attributes for instantaneous visualization is appropriate.

Since the exact specification of the shape attribute is, as mentioned, not complete, we extended it with respect to rectangles, polygons, and points, as marked in Listing 1. These simple geometric shapes were taken as first components to visualize a multitude of different types of elements. Point-like visualizations are useful to describe locations without area information, e.g., for gaze position analysis in eye tracking studies. Rectangles are most commonly used for bounding box object of interest annotations. In contrast, polygons provide a more accurate, but complex way of describing the contour of an object.

The visualization of USF data is handled by VLC in a codec module. This codec module receives

\[ \text{source code, software tools and datasets can be downloaded from the project homepage cf.} \]
streams of the subtitle data for the current frame from the demuxer of VLC. We extended this module with additional parsing capabilities for our specified shape data\(^2\), which is then drawn into the so-called subpictures and passed on to the actual renderer of VLC. Since the thread will be called for every frame, the implementation is time-critical, and we decided to use the fast rasterization algorithms of Bresenham (1965). Additionally, we added an option to fill the shapes, which is implemented with the scan line algorithm (Wylie et al., 1967). In the course of this project, an open source software\(^3\) is developed which converts time-dependent sensor data files of several subjects to USF files and encapsulate them together with the original video in a single MKV file.

Listing 1: Section of the USF specification (Paris et al., 2016). * marked attributes are added to the specification and implemented in our altered VLC player.

```xml
...<subtitle
  start="hh:mm:ss.mmm" @-start (1)
  stop="hh:mm:ss.mmm" @-stop (0..1)
  duration="hh:mm:ss.mmm" @-duration (0..1)
  type="SubtitleType">+text</text>
</subtitle>
```

structures, and does not account for non-visualizable content like non-frame based elements.

From our USF files of the data, one can generate ASS files using extensible stylesheet language transformations with a simple translation stylesheet\(^2\). After the conversion, a MKV container can be packed in the same manner as with USF. The resulting container with ASS visualizations makes the multimodal data accessible for a broad audience, as many unmodified players can display these visualizations.

### 4 DATASETS

For demonstrating that multimodal analysis can be done with a standard multimedia player, we used one existing dataset and introduced a new, unpublished dataset preview along with this work. All scientific data—here eye tracking raw data, EEG data, stills, video stimuli, etc.—are encoded in our proposed MKV based container format.

#### 4.1 Movie and Frame Dataset

This dataset, presented in Açıkgöz et al. (2014), consists of 216 movie clips featuring a single continuous shot. The clips were taken from two commercial DVDs, Highway One and Belize both from the Colourful Planet collection (Telepool Media GmbH, 2014). The clips had a duration range of 0.8s to 15.4s. Moreover, from each movie clip the median frame was taken to serve as a still image to be presented with a duration equaling the length of the corresponding clip. Some of the frames are displayed in Figure 2. These 532 stimuli (216 movies and 216 frames) were shown in random order to human observers (median age 25) while their eye position was recorded with an Eyelink II eye tracker at a sampling rate of 500Hz. There was
no explicit task, and the sole instruction given was to observe carefully. Here we use representative examples of the movies and frames together with the eye data taken from single observers.

4.2 Real World Visual Processing Dataset

These data are part of a larger dataset, investigating electrophysiological markers during free navigation in a complex, real-world environment. Participants freely moved and inspected objects in the environment. No other task was given to the subject. Together with 128 channel EEG, the recording includes two custom-made, pressure-sensitive foot sensors, eye tracking data, and a scene camera capturing the visual input (cf. Figure 3). EEG data was recorded at 1024Hz, eye tracking data was recorded at around 100Hz, and the world camera recorded at 60Hz. Frames taken from the world camera showing the environment and the complete experimental setup are presented in Figure 3.

5 PROTOTYPES

The datasets, mentioned above, are used for prototyping. For providing instantaneous meaningful visualizations and sonifications, one has to define which attribute of the scientific payload should be used. After that, one can build the multimedia with these data. In the following, we describe an example of how one could create visualizations as well as sonifications, how these are mixed into the MKV and finally how non-experts and experts can use these multimedia containers.

5.1 Visualization of Data

No matter which subtitle format—ASS or USF—one uses for encoding the visualization of sensor data, one has to select which data attributes are to be visualized and in what way. All datasets introduced in Section 4 provide gaze data, which need to be visualized for prototyping. As data attributes to be visualized, we select the subject’s gaze position as viewpoint on the video, their pupil size, and fixations.

Using our open source dataset to USF converter introduced above, we encode the gaze position as point, where the viewpoints on the video correspond to the attribute of point. Using the diameter attribute, we visualize the pupil size relative to the average pupil size of the subject. A visualization of the absolute pupil size will, due to individual size variations between subjects, not be meaningful in our opinion. Fixations are visualized as squares using the rectangle shape. The square’s center corresponds to the point of fixation on the video and width corresponds to the pupil size. Due to the usually higher sampling rates of the eye trackers compared to the frames per second of videos, more than one viewpoint is visualized per frame, as shown in Figure 4.

The original data content is stored as comment within the USF file. Note, as described in Section 3, once the visualization is done in the USF, it can be easily converted to ASS.

5.2 Sonification of Data

To extract the alpha rhythm from the raw EEG data, we chose a parietal-occipital electrode (PO3) and those samples that corresponded to the corresponding frames of the video (61440 samples). Afterward, the raw signal was band-pass filtered to incorporate the respective frequency band (8 – 13Hz). We used a FIR filter with a Hamming window of 1690 samples. To calculate the power of the resulting filtered signal, the absolute of the Hilbert transformed data was...
Figure 4: Visualized gaze data by a reused subtitle track on a video by a standard multimedia player. The gaze position on the frames is visualized by a red circle with its diameter depending on the relative pupil size. Squares visualize fixations. In this example, the sampling rate of the eye tracker is higher than the frame rate of the video. Thus multiple points and squares are shown.

squared (Cohen, 2014). The resulting signal represents the alpha power with respect to the onset of the video. All preprocessing of the EEG data was done in EEGLab (Delorme and Makeig, 2004).

We generated the audio signal using MATLAB (The MathWorks Inc., 2014). To sonify the resulting alpha power, two different approaches were taken. For the frequency modulation, a carrier frequency of 220Hz—corresponding to note a—was modulated in its frequency by the power of the alpha signal. For the volume modulation, the same carrier frequency was modulated in its power with respect to the alpha power, with a louder tone meaning a stronger power. The resulting audio streams were exported as WAV files.

5.3 Creating the MKV Container

Creating, or more precisely muxing, a MKV container is quite easy and can be done with the command line or with the graphical user interface version of mkvmerge (MKVToolNix, 2016). Due to the rededicated subtitle formats for the visualizations, all visualizations are muxed as subtitle tracks, the sonifications (audio tracks) and the attachment (general metadata) carrying the raw data.

5.4 Using the MKV Container

The usage of multimodal data presented in a multimedia container is quite intuitive, as it uses the same user interface metaphors known from entertainment content. Hence, the user can change the visualiza-

5.3 Creating the MKV Container

Creating, or more precisely muxing, a MKV container is quite easy and can be done with the command line or with the graphical user interface version of mkvmerge (MKVToolNix, 2016). Due to the rededicated subtitle formats for the visualizations, all visualizations are muxed as subtitle tracks, the sonifications (audio tracks) and the attachment (general metadata) carrying the raw data.

5.4 Using the MKV Container

The usage of multimodal data presented in a multimedia container is quite intuitive, as it uses the same user interface metaphors known from entertainment content. Hence, the user can change the visualiza-

6 USER STUDY

Is multimodal analysis possible with a standard multimedia player by the use of our multimedia containers? To answer this, we asked one expert and two non-experts in the field of multimodal analysis to exploratively investigate these datasets. The major points of their feedback are mentioned below and discussed in the next section.

6.1 Professionals

The first response of the expert to this multimodal data in a multimedia container was “It’s a beautiful approach because one gets a first expression of the data without the need to run any scripts or install a special player”. The expert considers the exploration of unprocessed data highly useful for several reasons. First, it allows a plausibility check of the data. Experimental setups are highly complex and mistakes, for example in alignment of reference frames or synchronization, might otherwise go undetected. Second, standard statistical testing often concentrates on low dimensional subspaces with strong assumptions regarding the underlying statistical properties of com-
plex data. Here the visualization and sonification of high-dimensional data sets is an important tool. Third, presently the investigation of human sensorimotor interaction under realistic conditions in natural environments is largely exploratory (Einhäuser and König, 2010; Einhäuser et al., 2007). For these circumstances, the multimodal analysis for the complex experimental settings is an invaluable tool and guides the development of explicit hypotheses.

6.2 General Audience

The non-expert users highly appreciated the easy and interactive demonstration of multimodal data. They were less interested in complex statistical evaluation of the data set, but in the generation of a qualitative look and feel for the data. Furthermore, it “brought to live” the data and improved understanding of a quantitative evaluation. Finally, it served as a helpful backdrop and means of communication for discussion and the exchange of ideas. Here, the interactive nature of data exploration was instrumental. In summary, the ease of use lowered thresholds to get in close contact with the data and fostered fruitful discussions.

7 DISCUSSION

Multimodal data, in general, is hard to analyze but an instantaneous visualization makes selected data comprehensible to the broad audience, which is in our opinion, the main advantage of our proposed approach of storing them in multimedia containers. Experts might argue that the broad audience is not capable of formulating the correct assumption from the data. We, in contrast, think that the broad audience will understand and agree on assumptions easily if they can explore the data by themselves. Further, only selected data and attributes can be visualized or sonified. Thus, a careful selection by experts must be made to ensure an objective representation of the multimodal dataset.

8 CONCLUSIONS

The importance of multimodal datasets is hard to analyze but an instantaneous visualization makes selected data comprehensible to the broad audience, which is in our opinion, the main advantage of our proposed approach of storing them in multimedia containers. Experts might argue that the broad audience is not capable of formulating the correct assumption from the data. We, in contrast, think that the broad audience will understand and agree on assumptions easily if they can explore the data by themselves. Further, only selected data and attributes can be visualized or sonified. Thus, a careful selection by experts must be made to ensure an objective representation of the multimodal dataset.

In conclusion, we believe that the datasets pro-

Figure 6: Exploratory analysis with VLC. (a) one can change the visualizations, here the eye tracking, like subtitles; (b) one can change the sonification like audio languages; (c) example of sonification. The operator can hear the alpha wave, here visualized by VLC’s spectrometer.
vided, shared, visualized and sonified in such a way will facilitate, besides analysis tasks, applications in various fields, ranging from sensor improved computer vision (Schöning et al., 2016a), over communication aids in movies for disabled people, to neural networks with human-like multimodal sensor input.

ACKNOWLEDGEMENTS

This research was supported by H2020-FETPROACT-2014, 641321 socSMCs.

REFERENCES


