SWAPUGC: Software for Adaptive Playback of Geotagged UGC

Emmanouil Potetsianakis
LTCI, Télécom ParisTech
Université Paris-Saclay
75013, Paris
potetsia@enst.fr

Jean Le Feuvre
LTCI, Télécom ParisTech
Université Paris-Saclay
75013, Paris
lefeuvre@enst.fr

ABSTRACT

Currently on the market there is a plethora of affordable dedicated cameras or smartphones, able to record video and timed geospatial data (device location and orientation). This timed metadata can be used to identify relevant (in time and space) recordings. However, there has not been a platform that allows to exploit this information in order to utilize the relevant recordings in an interactive consumption scenario. In this paper we present *SWAPUGC*, a browser-based platform for building applications that use the accompanying geospatial data to dynamically select the streams for watching an event (or any spatiotemporal reference point). The view selection can be performed either manually, or automatically by a predefined algorithm that switches to the most suitable stream according to the recording characteristics. SWAPUGC is a research tool to test such adaptation algorithms and it is provided as an open-source project, accompanied by an example demo application and references to a compatible dataset and recorder. In this paper, we explain and then demonstrate the capabilities of the platform by an example implementation and examine future prospects and extensions.

CCS CONCEPTS

• **Information systems** → **Multimedia information systems**; **Spatial-temporal systems**; • **Computing methodologies** → Image and video acquisition;

KEYWORDS

User-Generated Video, GPS, Streaming, Mashup, MPEG-DASH

ACM Reference Format:
Emmanouil Potetsianakis and Jean Le Feuvre. 2018. SWAPUGC: Software for Adaptive Playback of Geotagged UGC. In *MMSys'18: 9th ACM Multimedia Systems Conference, June 12–15, 2018, Amsterdam, Netherlands.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3204949.3208142

1 INTRODUCTION

People casually record and upload videos in sharing platforms like facebook, twitter, instagram etc. These platforms amass User-Generated Content (UGC), which can be grouped via either automatic or manual annotation, according to recording location or context. Afterwards, the recordings can browsed by setting geographic and/or temporal criteria.

However, the aforementioned sharing platforms do not take full advantage of the recording capabilities of the devices. For example, common practice for location annotation is to indicate a single recording location, which is usually broad (e.g. in Grenoble, or near Stade des Alpes etc.). This annotation disregards the fact that several location points can be obtained throughout the recording, a feature that can be especially useful in UGC scenarios that the recording devices are often mobile.

On top of the sensors providing the location (i.e. GPS), most of these devices are also equipped with orientation sensors able to provide extra spatial information regarding the camera orientation (facing), like gyroscope, accelerometer and/or magnetic field sensor. This information can be used to identify videos (or parts of a longer video) that comply to more specific queries, like "videos from cameras facing the stage at the philharmonic concert at Stade des Alpes at 19:00 01/01/2010".

A challenge rising from such browsing capabilities is the consumption of the relevant videos. When using the several identified recordings, the final output consists of a series of transitions between the most suitable views. The UGC recordings can be arranged after the end of the event to produce professional-level quality content (e.g. Nine Inch Nails, Las Vegas concert in 2009[1]). But, selecting a suitable view on-the-fly (i.e. with the recordings being uploaded as recorded) can be challenging. Most of the efforts of the research community focus on how to identify relevant views and ranking the optimal selection (reviewed in the following Section 2), and not on techniques on how to switch between those views in a live setup.

Additionally, when serving video on the web, even a single view might be encoded in several qualities, so the client can request the one corresponding to its needs. However, research on adaptive video delivery, that focuses on selecting the most suitable quality for a video, is typically considering networking criteria and is disconnected from such applications.

We considered the challenges of switching between streams (different video and/or different quality), regardless of whether it is due to the content or the infrastructure, and in this paper we present SWAPUGC, a browser-based tool suitable for applications built to dynamically consume sensor-enriched UGC video content. Our platform is based on existing work on the domain and can provide view switching, either by manual selection from the user, or automated based on a selection policy. The selection policy can be defined by multiple criteria (network performance, sensor values etc.) and applied to deployments targeting either offline or

[1] en.wikipedia.org/wiki/Another_Version_of_the_Truth#Las_Vegas

MMSys'18, June 12–15, 2018, Amsterdam, Netherlands

ACM ISBN 978-1-4503-5192-8/18/06…$15.00
https://doi.org/10.1145/3204949.3208142

live content. SWAPUGC aims at providing a common reference for future developments on UGC video consumption.

2 BACKGROUND

The first requirement to be able to use a multi-view selection platform with UGC content, is to identify the relevant recordings. We do not study this topic, since there has already been extensive work, either on the domain of using image features [4], or by using the sensor measurements combined with camera characteristics to estimate the Field-of-View (FoV)[1].

After the relevant recordings have been identified, a pre-processing step correcting any misalignment on the timing information of the streams might be required. A widespread technique applied when recording the same subject, is to synchronize the streams by using audio and/or visual features [5]. This approach is computationally expensive but can be used to synchronize, with sufficient accuracy, streams from any device; alternatively, if the devices are interconnected, clock synchronization protocols (like NTP), can be used during the recording.

All of the aforementioned works are mentioned in this section since they can be used to prepare the data prior to consumption by the client. Regarding the actual consumption of location-tagged media, previous approaches include popular sharing platforms such as instagram[2] and google maps[3]. However, these platforms have only coarse time and location indication (either sensor-recorded or manually annotated) and do not take in account the facing of the camera. For sensor-enriched UGC content, there are currently platforms that support browsing according to the location/orientation like GeoUGV [7] and Spatiotemporal Video Navigation [10]. Even though these platforms are suitable for browsing the streams, and switching from one to another, they do not support any synchronous consumption, nor live or adaptive streaming scenarios.

We built SWAPUGC by extending the Spatiotemporal Video Navigation platform. In the original incarnation, videos are displayed on a map with indications regarding the position and facing of the camera at key points of the recording. The user can select a marker on the map and navigate in the respective video, at the corresponding time. During playback of a video, the current recording location is updated on the map. A screenshot of the platform is shown in Figure 1, with the location marker corresponding to the currently visible frame highlighted. We used the same map/video interface and marker-based interaction paradigm for our demo implementation, which we detail in Section 3.1.

Finally, in order to test our platform we had to find a relevant set of video recordings with accompanying location/orientation measurements. We examined using UGC datasets from spontaneous recordings [7], or from specific events [12], however as of the time of writing of this paper, both repositories were offline.

Instead, we chose to use a dataset that contains recordings from several cameras, both UGC and professionally recorded [2]. Only some of the UGC views have geospatial information recorded from the sensors, which is absent from all of the professional cameras. Therefore, the simultaneous recordings used are few, but the same principles apply regardless of the number of recordings.

[2] instagram.com
[3] maps.google.com

Figure 1: Screenshot of the Spatiotemporal Video Navigation client

3 THE SWAPUGC PLATFORM

The goal of our platform is to provide an environment for researchers to experiment on consumption scenarios for multiple recordings. To foster adaptation, SWAPUGC is browser-based, and in order to encourage collaboration we provide it as an open-source project, hosted in a public repository[4]. This project, consisting of the SWAPUGC platform, an example implementation and accompanying tools and guidelines, aims at bootstrapping development of applications supporting adaptive stream selection for following a spatiotemporal reference (event).

In order to be spatiotemporaly-aware, SWAPUGC uses location and orientation information provided by the recording devices, alongside with the video. The sensor recordings should consist of timestamped location and orientation measurements and with the corresponding video form a *bundle*. Each bundle must have a common timing reference, which, if not inherently present, can be obtained by means like feature extraction and matching, mentioned in the previous section. The timing information is used to identify and synchronize relevant streams.

The input of SWAPUGC for each recording consists of the following files:

- *Descriptor*, with the required synchronization information, the location of the other files and other optional (or application-specific) information.
- *Playlist* of the video segments for the video stream.
- *Location File* with timestamped Latitude/Longitude pairs.
- *Orientation File* with timestamped Yaw/Pitch/Roll[5] triplets.

[4] https://github.com/acmmmsys/2018-SWAPUGC
[5] also supports quaternions; other rotation formalisms can be added

First, during the initialization phase, the descriptor file for each bundle is parsed. The synchronization information contained is a required reference point in time, typically indicating the start of the recording. Also, the descriptor contains reference to the other files required, for retrieving the video, location and orientation. Examples of other application-specific information that can be inside the descriptor file are: lens characteristics, used for applications estimating the Field-of-View[1]; or location of files, for example containing mapping-schemes for streaming metadata[11]. After the timing information for each bundle is parsed, the earliest recording is identified and its timeline is used as a reference during the simulation.

Then, the playlist file of each recording is fetched, that contains the available video representations (i.e. the available qualities), and references to the byte-ranges or files of the representation segments. The justification of the architectural design choice to support multiple representations and segmented video files is to simulate a realistic distribution scenario like Adapative Streaming over HTTP (e.g. DASH, HLS), which is the predominant way of distributing video on the web. Then, from the extracted information the video player is initialized, which, since our platform is browser-based, is a `video` HTML element. During playback we update the `video`, by feeding its buffer, using the `Media Source Extensions` (MSE) API.

After the setup of the video player, the Location and Orientation data are fetched, to be used during playback. When the initialization phase is completed, the earliest starting recording is selected as a reference, meaning that it is the initial view and consecutive events are fired on its timeline. The interface outline depends on the implementation, but for the purpose of this paper we consider a default setup with a map having markers indicating the location of the recordings and a video showing the currently active view, as shown in Figure 2 and explained in the following Section 3.1.

When playback commences, the default view is displayed and the respective location/orientation is updated according to the timestamped sensor measurements. As a recording (with the accompanying spatial information) becomes available the user can switch to it. Alternatively, the switching can occur via automated view selection policies. The selection policy can take into account the available spatial information to decide when to switch (e.g. when the camera moves close enough to the subject). The throughput between the server and the client can be throttled to emulate different network conditions, and by designing our platform to support representations of multiple video qualities, quality-adaptation algorithms can be applied. This way the stream selection algorithm can consider both networking metrics and spatial information to decide on the most suitable recording and quality. To the best of our knowledge, SWAPUGC is the first open source player allowing for dynamic view switches in a DASH context. In the following section, we describe an implementation of a SWAPUGC-based application.

3.1 Example Implemented Application

In order to test our platform we created a client application for consumption of simultaneous recordings. Our application supports a manual switching functionality, in which the user clicks on a marker and the main view switches to the recording from that camera. Also, we created a naive view selection policy, in which a 10s round-robin algorithm switches between views.

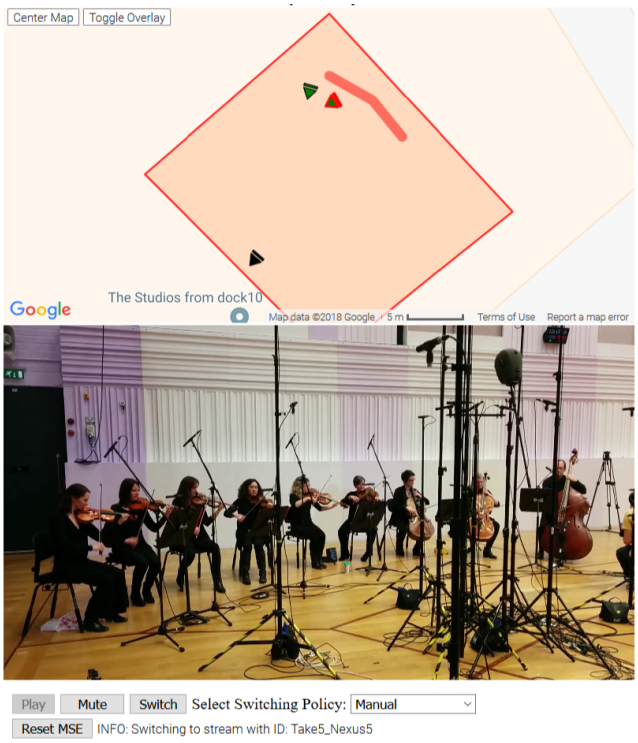

Figure 2: Screenshot of the implemented application

The interface of the application consists of a map and a video element. After loading the timing characteristics, SWAPUGC parses the location measurements and the initial markers are placed on the map indicating the recording location corresponding to the beginning of each -currently available- video. Also, the respective orientation value is parsed, in order to rotate the marker, to emulate the facing of the recording device, as shown in Figure 2. Thus the markers *indicate* the field of view (they are *not* simulating it), and it is the same technique as previously shown in Figure 1, with the difference that instead of having one marker for a specific timestamp, we have one marker for a specific recording.

When the playback starts, the placed markers are updated, as new orientation / location measurements are parsed, corresponding to the timeline. If a recording becomes available at any time during playback a new marker is placed on the map, or removed if it becomes unavailable. When the user clicks on a marker, the video switches to the selected view.

On top of the click-to-activate view, which is based on the user, the platform can operate in an automated fashion. Since all the characteristics of the recordings are available, a view selection policy can be defined that will switch to the better matching view. Events can be fired to indicate when a view has become available/unavailable.

To test this application, we used the dataset mentioned in Section 2, containing studio recordings of the BBC philharmonic[6]. More specifically, we used 3 recordings in total (from "Take #5"):

- `A002C001_140325E3` fixed studio camera - no sensor data
- `20140325_121238` roaming handheld camera - sparse orientation data
- `Take5_Nexus5` roaming handheld camera - orientation data

Because the recording session was indoors, the location data was severely inaccurate and we did manual annotation. Also, the recording with id `20140325_121238` had very coarse and sparse orientation data, thus were recreated, using the open-source Spatiotemporal Video Navigation recorder[7]. A screenshot of the application running with the three recordings is shown in Figure 2. Finally, we used GPAC[6], to create segmented video files so they can be distributed in a manner than would resemble a live scenario (potentially with multiple qualities for adaptive streaming). The segment playlist is in `mpd` format, which is natively supported by SWAPUGC, in order to be compatible with the MPEG-DASH standard. All of the aforementioned files and the source code are available in the SWAPUGC repository, under a Creative Commons Attribution Non-Commercial Share Alike license (CC BY-NC-SA 4.0). Also, a website hosting a demo of the application is available online[8].

4 CONCLUSIONS AND FUTURE WORK

In this paper we presented SWAPUGC, an open platform for dynamically consuming geotagged UGC content. The novelty of our approach is the consideration of timing when switching between streams and the capability to utilize the respective device location/orientation for visualization or algorithm input. This way applications can consider both system metrics (latency, throughput etc.) and spatial characteristics (location, facing etc.) for stream selection. Also, because segmented video of multiple qualities is supported, on top of selecting the most suitable video, the application can also select the most suitable quality of this video.

For our demo application we implemented a naive round-robin view selection algorithm and a map-based interface for the user to switch to a desired stream according to its spacial characteristics, but SWAPUGC can by utilized to accommodate other stream switching policies. More specifically, it can be used to test various approaches regarding the criteria of the selected streams [13][8][3]. These criteria can be either subjective / cinematic (e.g. camera positioning, stability / shake), or objective / system like video resolution, connection history etc.

As an example of mixed criteria adaptation policy, the application can build an index of the available bitrates per recording and a ranking of the recordings according to their content by using cinematic criteria like distance, shaking etc. Then, the highest ranked recording is being consumed, but if at some point the connection quality deteriorates and the lowest bitrate of this recording is still too high for the available throughput, the main view switches to the next recording on the ranking that has a suitable bitrate. For similar scenarios, the ranking can be built by also considering the image quality degradation at lower bitrates; for example, a high-bitrate stream of a lower-relevance video can be prioritized over a stream that is slightly more relevant but with significantly higher fidelity.

We are currently planning to conduct user studies in order to evaluate stream adaptation policies, to identify the most relevant weighted criteria for video source selection. Also, we are in the process of creating a dataset with recordings accompanied by spatial information and network metrics. The combination of network metrics and location can be used to identify novel adaptation policies, predicting the deterioration of the connection quality and/or GPS measurement accuracy[9]. Thus the dataset will have the potential to be used as common reference when evaluating different policies for a broad spectrum of applications.

[6] Available on the repository is also the parser used to extract timing information and format the XML-based data of the dataset, to the SWAPUGC JSON-based format.
[7] https://github.com/emmanouil/Spatiotemporal-Navigation-Recorder
[8] https://acmmmsys.github.io/2018-SWAPUGC/

REFERENCES

[1] Sakire Arslan Ay, Roger Zimmermann, and Seon Ho Kim. 2008. Viewable scene modeling for geospatial video search. In *Proceedings of the 16th ACM international conference on Multimedia*. ACM, 309–318.

[2] Werner Bailer, Chris Pike, Rik Bauwens, Reinhard Grandl, Mike Matton, and Marcus Thaler. 2015. Multi-sensor concert recording dataset including professional and user-generated content. In *Proceedings of the 6th ACM Multimedia Systems Conference*. ACM, 201–206.

[3] Sophia Bano and Andrea Cavallaro. 2016. ViComp: composition of user-generated videos. *Multimedia tools and applications* 75, 12 (2016), 7187–7210.

[4] Axel Carlier, Lilian Calvet, Duong Trung Dung Nguyen, Wei Tsang Ooi, Pierre Gurdjos, and Vincent Charvillat. 2014. 3D interest maps from simultaneous video recordings. In *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 577–586.

[5] Anna Llagostera Casanovas and Andrea Cavallaro. 2015. Audio-visual events for multi-camera synchronization. *Multimedia Tools and Applications* 74, 4 (2015), 1317–1340.

[6] Jean Le Feuvre, Cyril Concolato, and Jean-Claude Moissinac. 2007. GPAC: Open Source Multimedia Framework. In *Proceedings of the 15th ACM International Conference on Multimedia (MM '07)*. ACM, New York, NY, USA, 1009–1012. https://doi.org/10.1145/1291233.1291452

[7] Ying Lu, Hien To, Abdullah Alfarrarjeh, Seon Ho Kim, Yifang Yin, Roger Zimmermann, and Cyrus Shahabi. 2016. GeoUGV: User-generated mobile video dataset with fine granularity spatial metadata. In *Proceedings of the 7th International Conference on Multimedia Systems*. ACM, 43.

[8] Sujeet Mate and Igor DD Curcio. 2017. Automatic Video Remixing Systems. *IEEE Communications Magazine* 55, 1 (2017), 180–187.

[9] Sami Mekki, Theodoros Karagkioules, and Stefan Valentin. 2017. HTTP adaptive streaming with indoors-outdoors detection in mobile networks. *arXiv preprint arXiv:1705.08809* (2017).

[10] Emmanouil Potetsianakis. 2017. Streaming and Presentation Architectures for Extended Video Streams. In *Adjunct Publication of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video (TVX '17 Adjunct)*. ACM, New York, NY, USA, 129–132. https://doi.org/10.1145/3084289.3084298

[11] Emmanouil Potetsianakis and Jean Le Feuvre. 2016. Streaming of Kinect Data for Interactive In-Browser Audio Performances. In *Proceedings of the Audio Mostly 2016 (AM '16)*. ACM, New York, NY, USA, 258–265. https://doi.org/10.1145/2986416.2986438

[12] Mukesh Saini, Seshadri Padmanabha Venkatagiri, Wei Tsang Ooi, and Mun Choon Chan. 2013. The jiku mobile video dataset. In *Proceedings of the 4th ACM Multimedia Systems Conference*. ACM, 108–113.

[13] Prarthana Shrestha, Hans Weda, Mauro Barbieri, et al. 2010. Video quality analysis for concert video mashup generation. In *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 1–12.