Towards Multimodal Consumption of Georeferenced Mobile Video Using Shape and Speed

Sérgio Serra
LaSIGE, Faculdade de Ciências
Universidade de Lisboa
1749-016 Lisboa, Portugal
sergioserra99@gmail.com
+351217500087

Teresa Chambel
LaSIGE, Faculdade de Ciências
Universidade de Lisboa
1749-016 Lisboa, Portugal
tc@di.fc.ul.pt
+351217500087

ABSTRACT
An increasing amount of digital video is accessed, captured, and uploaded to the Web everyday, from different platforms and devices, that increasingly can georeference the information they capture and access, allowing to enrich their contextualization. But video search has been limited to keywords, or a set of parameters, providing limited support for temporal and spatial dimensions. We propose novel ways to search and access georeferenced videos, where these dimensions are of central importance, especially by video trajectories shape and speed, using a multimodal interactive mobile interface, involving gestures and movement, with the potential for more natural interactions, increased engagement, sense of presence and immersion.

The preliminary evaluation based on low-fidelity prototypes and encouraging users participation in the design, had positive results. Users found most features quite satisfactory, even fun, and easy to use. Different options and modalities were found interesting and adequate for different use scenarios that could be identified and suggested, and some concerns and challenges were identified to be taken into account in the next design and development phases, towards more flexible and effective interactive content consumption, through more natural interaction with mobile devices on their own or as second screens.

Author Keywords
Georeferenced videos; Gestures; Multimodal; Search; Browsing; Consumption; Space; Speed; Shape; Movement; Trajectories; 360º; Mobile; Second Screen

ACM Classification Keywords
H.5.1 [Information Interfaces and Presentation (I.7)]: Multimedia Information Systems – *video*, *hypertext navigation and maps*; H.5.2 [Information Interfaces and Presentation (I.7)]: User Interfaces – *interaction styles, evaluation*;

INTRODUCTION
Video is becoming a pervasive medium, widely captured, shared and accessed from different platforms and devices. And increasingly videos can be georeferenced, allowing to enrich their contextualization. A large amount of digital video is being uploaded everyday to the Web and is available to search and watch. However the current and most used mechanisms to browse and find videos are keywords and a limited set of parameters such as: keywords, duration, video quality, ignoring the temporal and spatial dimensions. Video has an enormous potential for immersion and mobile devices allow to access information while 'immersed' in reality anywhere. With the proliferation of devices like: smartphones, tablets and more recently wearables, we could take advantage of the multimodal sensors available, to create new ways to find and navigate georeferenced videos through time and space, using more natural interfaces, involving gestures and movement shape and speed, with the potential for increased engagement, sense of presence and immersion when accessing the videos.

In this paper, we describe our work in this direction. The next section presents the background provided by previous work and the vision for the new directions explored in this paper, followed by a section that highlights main challenges and opportunities, and presents most relevant related work. Next, the conceptual model and design options are presented for the multimodal georeferenced mobile video access in space and time, demonstrated in the prototypes and evaluated in the following section. A preliminary user evaluation was conducted with low fidelity prototypes, to find out about perceived usability and acceptance, focusing on usefulness, satisfaction and ease of use, and encouraging users participation in the design [12]. Finally, the paper ends with conclusions and perspectives for future work, also reflecting on questions relevant to effective interactive content consumption.

BACKGROUND AND VISION
This work builds on previous work done in the context of Sight Surfers [6], an interactive web application for sharing, visualizing and navigating georeferenced 360º interactive videos, as hypervideos, including city tours or more extreme activities like kart racing. These can be experienced in increased immersion and isolation, or synchronized with a map while being played. Sight Surfers supports several

2nd International Workshop on Interactive Content Consumption at TVX'14, June 25, 2014, Newcastle, UK.

mechanisms for navigation and orientation. Users can see the location and trajectory of the video in the map and navigate through crossing trajectories, possibly shot by other users, either in the map or as hyperlinks in the current video, and link to movie scenes that take place in that location. Windy Sight Surfers extended the previous system, to run on mobile devices and to empower users in their immersive video experiences [9,10]. It uses geographical and meteorological metadata, sensors and actuators, for increased immersion in terms of video viewing and sensing (visual, auditory and touch), for a more realistic feeling of movement and speed. It can be used on its own or as a second screen, having the video playing in a wide screen free of extraneous information and the mobile providing additional navigation and orientation features, e.g. using a map, for a more immersive experience.

Users can view around the 360º video by panning the device around, as if holding a window to the surrounding immersive video it is displaying, or to use it as a "wheel" - a second screen that can be rotated to move around the video in the wider TV screen. It may also provide an increased sense of speed and orientation when watching the videos through 3D and a wind interface. But videos are searched mainly by keywords, possibly filtered by regions in the map, time they were shot, duration, and broad categories ranging from fast to slow, in text and check-box based classical interfaces.

To provide a more complete support for the additional spatio-temporal dimension in georeferenced videos, and keeping the purpose of increasing immersion, aligned with the augmented sensorial experience, we want to create richer mechanisms for interactive search, visualization and navigation in more natural modes of interaction. Videos could be searched and accessed by their location as a place (e.g. New York, or the user's current position), as is often possible, but also based on their trajectories by choosing the actual streets, or even by the shape of the trajectories regardless of the actual streets (e.g. motocross or ski), by time (when were shot, and their duration) and by speed. Users could use touch interfaces for this, or use the mobility of their devices or their own movement while walking, running or traveling, on their own or as 2nd screens, to imprint or capture movement shape and speed, in possibly more natural and immersive ways.

RELATED WORK, CHALLENGES AND OPPORTUNITIES

Challenges for this work include providing users with an adequate interactive interface capable of capturing and expressing the temporal and spatial dimensions, allowing to represent speed and trajectories, and at the same time offering an intuitive, simple, effective and natural way to search for, and present resulting videos and navigate them in a mobile environment. It is both a challenge and an opportunity because users are not used to searching and navigating in these dimensions, but technology is allowing to capture movement in mobile devices in ways that hold the potential to support more natural interactions involving shape and speed towards more immersive experiences.

Most video libraries and websites like YouTube or Vimeo are based on keywords and have at most a very limited support to access video based on spatial and temporal dimensions. Rego et al. [11] developed VideoLIB, a digital library that enhances video retrieval by using spatial and temporal operators, based on Dublin Core and MPEG-7 metadata standards. Search criteria include action (what), person (who), time (when) and place (where), and use operators like before, during and after to define time intervals. This allows to make searches like "retrieve Madonnas´s video clips which were produced outside the USA during 1990's". It uses a form and text based interface, without the use of maps, and videos are considered as a whole - trajectories and speed are not taken into account.

There are some approaches to search and browse videos, and mainly photos, using maps. Google Street View is a 360º photo viewer using a spherical image projection and geolocalization, but it does not provide video, nor user generated and alternative views of the places. Panoramio (.com) is a georeferenced photo sharing website accessed as a layer in Google Earth and Google Maps. Users can do text-based search or navigate in the maps, and view photos taken by other users, based on location. The photos are presented along with a map that highlights their location, both as a collection resulting from a query or one by one. There are filters to highlight most popular, recent, famous places and indoor, both on a separate tab with the filtered photos, and by enlarging these photos among those shown on the map. Finsterwald et al. [2] developed The Movie Mashup Application (MOMA), as a public web map-based service for searching movies based on location, combining geotagged resources and text processing, mashing up information from DBpedia, GeoNames and Wikipedia synopses. Through its GUI, it allows to search and browse a data set of movies by director, location in text, by polygonal areas in the map, from locations extracted from movie titles, to compare query distributions and, using a mobile device version, allows to query for movies whose action took place around the user's current location. Although maps are a natural way to represent georeferenced information, and video often involves a trajectory, most solutions only allow users to post or access videos based on a single GPS location (usually the initial position). Seo et al. [14] present user-generated videos that relate to geographic areas in a map interface. They focus on the automatic selection of keyframes to represent the videos, and the determination of the location to place them on the maps. So they emphasize hotspots that are shot in the videos in front of the shooting spot, and not so much on their trajectories.

Concerning spatial and haptic interactive search in a mobile environment, in the last years we noticed a growing popularity of gesture interfaces and second screen applications. Lei & Coulton [3] implemented a gesture controlled application that act as a wand, using mobile sensors. It allows both

proximity and remote search of points of interest (POIs) based on the orientation of the wand, as an interactive spatial 'Flashlight', and the possibility of users to create additional content for a particular POI as photographs tagged with POI's location and the direction from which the photograph was taken. Photographs can then be filtered based on a desired viewing angle in a real world environment. Premraj et al. [8] presented iWalk, a tool that allows multi-media exploration of geo-tagged data through movement, to move through the digital space of a collection, and gesture, for direct data manipulation (e.g. select, go to next, zoom). They experimented with geo-tagged photographs and sound collections, and a non geo-tagged museum collection, where the user defined a mapping between digital and physical spaces. Their approach makes use of computer vision algorithms computed on standard commercial camera inputs and is able to operate in real time.

Mobile devices may also act as second screens [1] to complement and interact with larger screens like TV or even public displays. MobiToss, an application created by Scheible et al [13], allows for mobile multimedia art sharing and creation. By using a mobile device with built-in accelerometer sensors, users can take a photo or video and "throw" it onto a large public display, with a gesture, for viewing and manipulation, through tilting. The users-created clips are augmented by the system with items like music or brand names and sent back to their phones as personal artifacts of the event. The preliminary user evaluation showed that capturing and throwing mobile content onto a large screen and manipulating it with gesture control into an art piece was perceived as an intuitive and fun activity. They enjoyed and engaged in the experience and appreciated getting something out of it, especially something artistic. But it requires improvements, by applying a more balanced set of video effects, adding group interaction and a more intuitive UI, to accommodate different movements users do to throw, and increase the perception of what is going on. This work explores natural gestures with a mobile device to manipulate photos or videos as a second screen, but in doing so, it does not explore spatial and temporal dimensions in videos. And none of the related work found addresses speed and trajectories in video as we propose to do.

VIDEO ACCESS BY SPACE, TIME AND… SPEED

Space and time dimensions are taken into account primarily in videos' locations, trajectories shapes and speed. To explore the interactive interfaces and user experience with these dimensions in georeferenced videos, a set of sketches and low-fidelity prototypes with different variations were designed for different use scenarios. The high-fidelity prototypes will explore the use of sensors and location services. The early low-fi prototyping and evaluation - using paper based screens on top of a real smart phone, users gestures and imagination - allow to test different design alternatives to tackle each goal in a faster and more efficient way, addressing challenges, sorting out and refining options, since an early stage. Next, we present the design rationale behind the main options for searches and accesses based on trajectory speed (Fig.1) and shape (Fig.2) in different modalities.

Through Touch – with finger

This is the more conventional interface, that allows to draw the query shapes by touching the screen (Fig.2b), or even on a touch pad of a laptop or desktop. The speed of this drawing may also be captured to query by speed only, or by both shape and speed. This kind of interaction may be more familiar and provide for better accuracy than the following, especially when the user has a hand free for this interaction

Through Gesture – with mobile

This modality can be used by moving the mobile in a gesture, to draw a shape (Fig.2a) or demonstrate a speed level (Fig.1a). This can be done with the hand that is holding the device, even if the other one is not free, and has the potential for a more natural or immersive modality to select the desired videos, to watch on the mobile or on a wider screen like a TV. In this context, viewers are used to interacting with a control in one hand, and keeping focused on the video on the wide screen.

Through Traveling – on the move

When on the move, users are often travelling by car, train, subway, planes or even walking or running. In addition or as an alternative to use your current location to access videos shot in the same location, it can be interesting to take the chance, especially when not driving, to watch videos that were shot at a similar speed, and be able to enjoy the viewing experience in a more immersive way, by matching the speed of what you are seeing with the speed that you are feeling or experiencing in reality. This might have a special impact in high speed videos in more extreme activities. As in the previous cases, both speed and trajectory shape can be captured this way (Fig.1b). Speed and even location might have more potential for immersion in the immediate video viewing, to get related videos on the spot. But capturing a trajectory can also be interesting, to search for other videos that have similar paths even if in different places, e.g. another similar Kart race elsewhere in the world. While gestures could rely on sensors, travel features would rely on location services like GPS.

Where - on the Map or Current Location

Since Sight Surfers videos are georeferenced, search can be made dependent on their location. Users can use a map to select locations (Fig.2b), or the current location can be captured, for the queries (besides the possibility to specify a set of places, like cities, as in more traditional interfaces).

Anywhere

Videos may also be searched independent of their location. In this way, a map is not used and only speed (Fig.1a) and shape are drawn on the screen or in the air (Fig.2a), or captured on travel (Fig.1b), without a geographical reference.

Figure 1. Search by Speed, through: a) gesture; b) traveling speed on a car; Results in video List with: c) Colored timelines with 3 colors; d) Gray-scale timelines; e) Color Highlight timelines with green color for the searched speed.

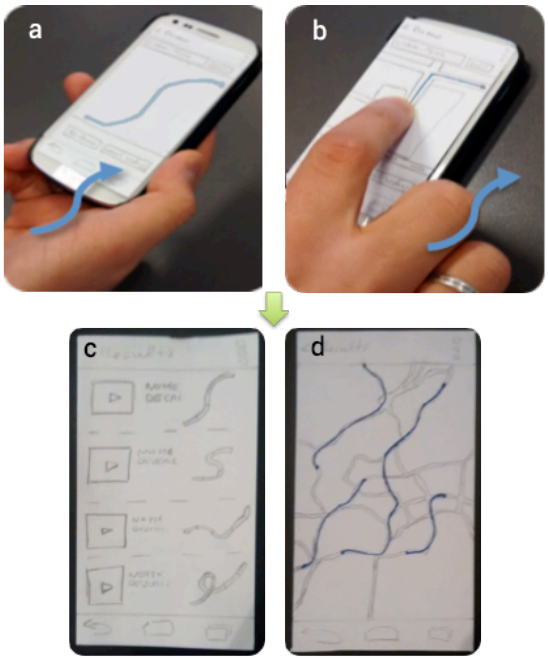

Figure 2. Search by Shape, through: a) gesture; b) geo-referenced shape with touch; c) Results on a List with video shaped timelines; d) Results on a Map showing videos trajectories.

Results in Maps or Lists

The resulting videos can be presented as trajectories on a map (Fig.2d), where each trajectory can also be seen as the video timeline synchronized with the video as in Sight Surfers [6]. And this is the default when search is based on a location. But results may also be presented independent of their location, e.g. in a list, where the speed (Fig. 1c-e) and or the shape (Fig.2c) can be emphasized in each video timeline. Also note that users could switch between map and list views and select what to show, for the same results.

Search results are presented to the user in different design alternatives, each one offering different visual cues and information about the content retrieved, in terms of shape and speed, both in maps or lists.

Speed Awareness

Speed can vary along a video, so the results would present first the videos that keep the desired speed (with a tolerance) for longer, but still, it could be interesting to be aware of the segments where the speed is as queried for. Fig.1c-e shows three alternative designs for presenting speed in the video *timelines* in: 1) Color: green for the searched speed, red for faster and blue for slower; 2) Gray-Scale: mid tone for searched speed, darker for faster and lighter for slower; and 3) Color Highlight: green for searched speed and two gray tones for faster an slower as in 2), allowing for higher contrast of the searched for speed.

Shape Awareness

Shape is shown by default when on a map, but can also be presented in the list view, as in Fig.2c where each video timeline takes the shape of the corresponding trajectory. Whereas speed awareness is optional in all the timelines: map, and list with or without shape.

PRELIMINARY USER EVALUATION

We conducted a user study to evaluate the features designed and to investigate about preferred alternatives and users' perception about usability and user experience, and their application in real use scenarios.

Method

We performed a task-oriented evaluation based mainly on Observation and semi-structured Interviews, after explaining the purpose of the evaluation and the concept behind the Sight Surfers application context and the new features being evaluated in the low-fidelity prototypes, using a Wizard of Oz approach for the interaction feedback. At the end of each task and at the end as overall, users provided a 1-5 USE (Usefulness, Satisfaction, and Ease of use) rating [4] about the tested interactive features, and were encouraged to make comments and suggestions, that at the current stage had the potential for contributions in a participatory design perspective [12].

Participants

There were 10 participants aged 21-52 (24 on average, 4 F, 6 M). All users had at least finished high school, 3 from computer science, the rest a mix of backgrounds, all had a

smart phone used on a daily basis to access info, and 9 often search for and watch videos but mainly on PCs, sometimes on tablets and seldom on mobiles.

Results

Main results are summarized by mean values for USE and most significant comments, for each of the categories of features. Std deviation was 0.5-1.0 (in a 1-5 scale) reflecting some different opinions that are commented.

Search by speed through gestures with the mobile was considered quite fun and very easy to use, and found useful by some (U:2.6;S:3.1;E:4.1), e.g. *"I can imagine being a nice thing to have when you practice some sort of motorsport or if you like to watch sports like Formula 1, guys that like speed would probably like this",* but others had doubts about its usefulness in real life, on a daily basis. Some concerns about the device precision to capture the wanted speed were also raised, especially by the computer science students. Travel speed was considered more useful in a real world scenario than moving the phone to capture speed (3.2;3.2;4.0). *"It is more useful for example for someone that practices motorsports", "I prefer it, I can use it to know the speed I'm going and look for videos with it"*.

Search by shape with touch was considered quite fun and easy to use, but users were not sure about its usefulness (U:2.5;S:3.1;E:3.8): *"It could be fun sometimes, but normally we want to search for things"*. On the other hand, participants appreciated having the ability to georeference the shape, finding several use scenarios where it could be used. *"I think georeferenced video search is much more useful, I can use it to see videos from a particular street, city, whatever", "It's better than normal shapes, I'm an athlete and I can use it to find and see running tracks"*. Some users pointed out that it is not easy to draw the trajectory *"In Google maps I can click the points I want and it draws the trajectory automatically"*. Searching for videos with a specific shape by using the phone to draw was found easier for free shapes (3.6;3.3;4.2), and not so much for geo-referenced shapes (2.8;2.4;3.0), but less precise. Users are not used to using the phone for this kind of action, preferring the finger to do it. *"it is not practicable, imagine walking and doing it on the street", "ok with a phone, but maybe not with a larger device like a tablet"*.

In general, users preferred the more familiar way of showing the videos in a **list view** (U:4.3;S:4.0;E:4.6), by being easy and simple, *"less fancy but does the work"*. But some fancied the **map** idea better (3.2;3.2;3.2) for the additional information, and most said it was useful when looking for georeferenced shapes, videos in specific places, to be aware of the video locations and trajectories length. *"As a geographical engineer, this could be useful for my work"*. Main concerns referred to awareness of the amount of videos retrieved and the representation in the presence of a huge amount. Filtering of results [6] was not in the scope of this test, but aligned with their concerns. A user suggested to have a mix of list and map where one could spot the videos on a map while hovering through a list of the selected ones, and on the other way around, hovering a trajectory could show the info that is presented in a list entry (with speed, duration, video image, etc.) in a popup with more detailed information.

Users found the **timelines** useful, satisfactory and easy to use for shape and speed awareness. Regarding speed and the different designs: *"are a nice idea to have because they easily indicate the speed of the video",* although a couple of users mentioned they would only possibly needed to identify the searched speed and not distinguish higher and lower speeds. Most users preferred the color highlight (U:4.0; S:4.0; E:4.6) in the search speed with the other speeds made less noticeable in gray for being easier to use, *"It's less confusing and very simple to use, easier than the 3 colors"*; and then the color version (4.3;4.1;3.9), which they found useful and satisfactory although more difficult to use. The gray version was found more difficult to use and less useful, satisfactory and even fun (1.9;1.7;2.4).

Time and Space Revisited

Space was addressed in georeferencing and represented in maps and in trajectories' shapes. Time is inherent to the video and represented in the timelines, and combined with space can be represented as trajectories in maps that can be presented in synchrony with the videos [6] and was now also addressed in speed. For its more natural mapping, the interactions that involved space, and especially shape and speed, received more attention in the conception of natural interfaces so far.

In the near future, we intend to explore further the temporal dimension in search and navigation, both inside each video and among videos that were shot in different periods in time. Although we have some ideas about the design, we wanted to learn from the users about their visions for interactions involving the temporal dimension in a more participatory way. Almost all participants immediately associated time with a timeline, like the ones we already have in the prototype in the different designs. A similar concept could be extended to refer to the time when the videos where shot, allowing to select temporal moments or intervals, and the possibility to enter specific dates in a text field was also mentioned. They acknowledged the importance of this dimension and the way it was already addressed. It was not so easy for them to move from what was already familiar, although they were curious and open about the possibility of also using different modalities.

CONCLUSIONS AND FUTURE WORK

We presented the motivation and design options for georeferenced mobile video access in space and time, developed in a context of user generated content, with a special focus on shape and speed in different modalities based on touch, gestures and movement. The preliminary evaluation based on low-fidelity prototypes had encouraging results. Users found most features quite satisfactory, even fun, and easy to use, and different options and modalities were found

interesting and adequate for different use scenarios that could be identified. But although some users found these interesting uses, others were skeptical about the usefulness of using trajectories and speed for video access in real life, something they are not used to having. There were also some concerns about sensor accuracy in gesture modalities.

It is important to highlight that evaluating this type of features on low-fidelity prototypes does not offer the same experience than a high fidelity application running on the smartphone, even more when using sensors and viewing videos – both providing dynamic information - are central aspects in the system. Also some users tended to more familiar ground and conservative options, in this first contact with these features, especially the ones with less technical background, or not used to accessing georeferenced media. Still, the evaluations, comments and suggestions allowed to identify upfront main strengths and concerns, and aspects to improve and reinforce in the design, implementation and reevaluation as a high fidelity prototype, that is already in progress as the next iteration. As innovative interaction modalities, based on sensors and location services, there are technological challenges in terms of accuracy and smooth integration of modalities that are being taken into account for the effectiveness of the designed interactions.

The temporal dimension in search and navigation will also be further explored, taking the feedback received into account, along with the enrichment of navigation when viewing the video. The adoption of these more natural interactions also in this context, and building on our previous work on immersive video [9,10] has the potential to increase the sense of presence and immersion when navigating around, crossing trajectories, changing viewing speed, or even throwing the video [1] to view on a wider screen like a TV, while possibly keeping the mobile as a second screen, or capturing the video playing on the TV to the mobile, to go on watching it on the move.

Closing remarks: The interactive forms of video access when the user is more active and often in a more reflective cognitive mode [5] are interweaved with moments of more passive consumption when the user views the video in a more experiential cognitive mode. We believe that more natural and immersive forms of viewing and interacting with video influence the quality of the user experience, mainly in experiential modes, and that a good design can contribute to foster and support the reflective modes in complementing and harmonious ways, so important in learning [7]. We are following previous research work [7,6,9,10] with the goal to contribute in this direction towards richer, more effective and satisfactory consumption of video-based media in different scenarios.

ACKNOWLEDGMENTS

This work is partially supported by FCT through funding of LaSIGE Strategic Project, ref. PEst-OE/EEI/UI0408/2014 and the ImTV research project ref. UTA-Est/MAI/0010/ 2009.

REFERENCES

1. Courtois, C., D'heer, E. Second screen applications and tablet users: constellation, awareness, experience, and interest. Proc. EuroiTV'12, ACM Press (2012),153-156.
2. Finsterwald, J. M., Grefenstette, G., Law-To, J., Bouchard, H., and Mezaour, A.D. The Movie Mashup Application MoMa: Geolocalizing and Finding Movies. Proc. GeoMM'12, ACM Press (2012), 15-18.
3. Lei, Z., and Coulton, P. A. Mobile Geo-wand Enabling Gesture Based POI Search an User Generated Directional POI Photography. Proc. ACE'09, the Int. Conf. on Advances in Computer Entertainment Technology, ACM Press (2009), 392-395.
4. Lund, A. M. Measuring usability with the USE questionnaire. Usability and User Experience, 8(2), (2001).
5. Norman, D. Things that Make us Smart. Addison Wesley Publishing Company (1993).
6. Noronha, G., Álvares, C., and Chambel, T. Sight Surfers: 360º Videos and Maps Navigation. Proc. GeoMM'12, ACM Press (2012), 19-22.
7. Prata, A., and Chambel, T. Going Beyond iTV: Designing Flexible Video-Based Crossmedia Interactive Services as Informal Learning Contexts. Proc. EuroITV'2011, ACM Press (2011), 65-74.
8. Premraj, V., Schedel, M., and Berg, T.L. iWalk, A Tool for Interacting with Geo-Located Data Through Movement and Gesture. Proc. ACM MM'10, ACM Press (2010), 1059-1062.
9. Ramalho, J., and Chambel, T. "Immersive 360º Mobile Video with an Emotional Perspective". Proc. ImmersiveMe'2013, ACM Press (2013), 35-40.
10. Ramalho, J., and Chambel, T. "Windy Sight Surfers: Sensing and Awareness of 360º Immersive Videos on the Move". Proc. EuroiTV'2013, ACM Press (2013), 107-115.
11. Rego, A., Baptista, C., Silva, E., Schiel, U., and Figueirêdo, H. VideoLib: a Video Digital Library with Support to Spatial and Temporal Dimensions. Proc. SAC'07, ACM Press (2007), 1074-1078.
12. Sanders, E. From User-Centered to Participatory Design Approaches. In Design and the Social Sciences. J. Frascara (Ed), Taylor & Francis Books Limited (2002).
13. Scheible, J., Ojala, T., and Coulton, P. MobiToss: A novel gesture based interface for creating and sharing mobile multimedia art on large public displays. Proc. ACM MM'08, ACM Press (2008), 957-960.
14. Seo, B., Hao, J., and Wang, G. Sensor-rich Video Exploration on a Map Interface. Proc. ACM MM'11, ACM Press (2011), 1013-1016.