Retrospective vs. prospective: two approaches to mobile media capture and access

Arttu Perttula*

Tampere University of Technology,
P.O. Box 300,
Pori FI-28101, Finland
E-mail: arttu.perttula@tut.fi
*Corresponding author

Scott Carter and Laurent Denoue

FX Palo Alto Laboratory, Inc.,
3400 Hillview Avenue, Bldg. 4,
Palo Alto, CA 94304, USA
E-mail: carter@fxpal.com
E-mail: denoue@fxpal.com

Abstract: Mobile media applications need to balance user and group goals, attentional constraints and limited screen real estate. In this paper, we describe the iterative development and testing of an application that explores these trade-offs. We developed early prototypes of a retrospective, time-based system as well as a prospective and space-based system. Our experiences with the prototypes led us to focus on the prospective system. We argue that attentional demands dominate and mobile media applications should be lightweight and hands-free as much as possible.

Keywords: mobile; capture and access; prototyping; maps; social computing; multimedia; sharing; collaborative; attention; location-based services.

Reference to this paper should be made as follows: Perttula, A., Carter, S. and Denoue, L. (2011) 'Retrospective vs. prospective: two approaches to mobile media capture and access', *Int. J. Arts and Technology*, Vol. 4, No. 3, pp.249–259.

Biographical notes: Arttu Perttula is a Graduate Student at the Tampere University of Technology in Pori, Finland. His research focus includes mobile social media and mobile games.

Scott Carter holds a PhD from The University of California, Berkeley. He is a Research Scientist at FX Palo Alto Laboratory. His research focus is mobile multimedia.

Laurent Denoue holds a PhD from The University of Savoie, France. He is a Senior Research Scientist at FX Palo Alto Laboratory. His research focus includes HCI, CSCW with an emphasis on mobile computing.

1 Introduction

People are capturing increasing amounts of multimedia data with an increasing diversity of mobile devices. However, tools to organise and synthesise this data are scarce. In some cases, synthesis is not as important and simple streams suffice (e.g. informal sharing via Flickr). For many other tasks, though, it is vital to be able to structure or abstract media. Especially, when data must be synthesised over not only a group of devices but also groups of users this can be so difficult that much media can go completely unaccessed (Klemmer, 2002).

In this work, we show how we evolved a system to capture and access media on mobile devices (see Table 1). We followed the approach pioneered by Trevor et al. (2002) in which two different early-stage prototypes are developed and deployed in order to understand a design space. One prototype uses a time-based visualisation to facilitate *retrospective* review of media captured by individuals and groups. In this approach, the focus of the mobile application is on recording activities for *post hoc* perusal. The other version uses a space-based visualisation to use media to suggest *prospective* activities. In this approach, the focus of the mobile application is on supporting current or immediate tasks rather than access after-the-fact. Importantly, though, these early-stage prototypes are not necessarily 'suggestive of the finished product', but instead serve as a means of exploration (Greenburg and Buxton, 2008). As Shrage (2000) writes, such early-stage prototypes can help 'externalise thought and spark conversation'. Our goal at this stage in our work is to use applications to help to understand the challenges with mobile media capture and access tools.

In the rest of this paper, we describe our two systems: the retrospective prototype that links captured data to other media previously captured in the same place using a temporal representation; and the prospective prototype that uses media to guide users towards more entertaining or useful areas. We also describe case studies and pilot tests for each application and our field evaluation of the final version.

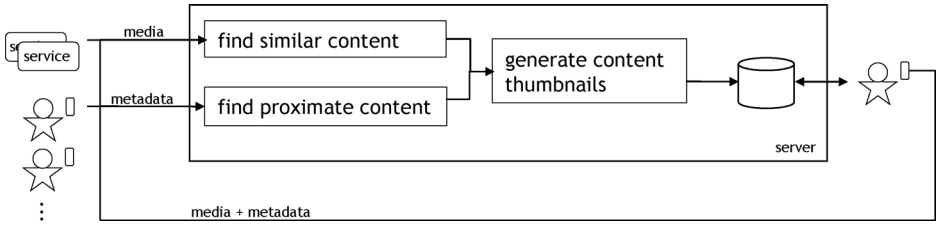
Table 1 Attributes of the retrospective and prospective capture prototypes

Retrospective	*Prospective*
Temporal representation	Spatial representation
Focus on media nearby user	Media where user might want to be
Media captured in the past	Media just captured or potential to be captured

2 Retrospective prototype

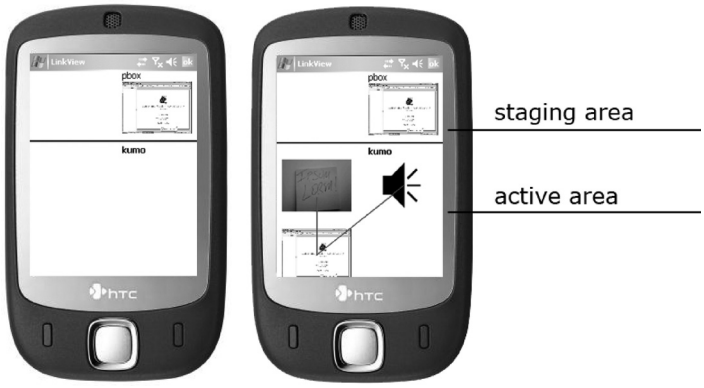
Our retrospective prototype – *Notelinker* – structures recorded media by creating links based on proximity and content. Users can record media with a mobile application that forwards captured content to a server. The server analyses content and metadata to establish links that are forwarded back to the mobile application (Figure 1). The mobile application can capture a wide variety of media (including video, image and audio) as well as metadata (including Bluetooth proximity and device interaction history). The mobile application also includes a pannable interface to organise and annotate media (Figure 2). The server can receive data from the mobile application as well as a variety of different media formats uploaded through a webpage. The server can also receive data from capture services (such as meeting capture systems) running in smart environments. The core novelty of the system is in the links it establishes between different media as well as between annotations of representations of digital documents and the original documents themselves.

Figure 1 Data flow in the Notelinker system



Note: Location-aware services and mobile devices send media and metadata (including location, Bluetooth proximity information and interaction history) to the server. The server creates similarity scores between media using content-based methods and finds nearby media using location and proximity information. The server then generates thumbnail representations of each captured picture or clip and pushes them to relevant users. Users can download the full versions of captured media by dragging a thumbnail from the staging area to the main panel in the mobile application's main UI.

Figure 2 Notelinker in a smart environment



Note: (Left) The system has determined its location ("kumo") automatically from Bluetooth tags in the environment and connected to a slide capture service available in the room ('pbox' in the upper right). (Right) The user has clicked on the slide from the capture service to create a copy in the active area (lower left). The user has also captured an image of notes she made (middle left) and an audio note (middle right) and linked both annotations to the slide.

The server creates links based on proximity and content. To create proximity-based links, the mobile system continuously records audio as well as Bluetooth IDs of nearby devices. When a user makes a recording, this contextual metadata is saved on the server with the original recording. When media from other devices are synchronised with the server, the system automatically connects recordings with nearby Bluetooth IDs. The system also searches for similar audio clips (using a normalised amplitude method) in any video recordings and links to the appropriate segments.

To create content-based links, the system can use image-based features. For pictures, features (such as Grabner et al., 2006) and extracted OCR text are saved as meta-content and linked against other data uploaded to the server. In combination, these features allow

users to connect seamlessly media captured by their device to media captured by non-enabled devices. Furthermore, the system allows users to collect, annotate and organise representations of digital media that will be substituted with their original content when it becomes available.

Importantly, the system also includes mechanisms for organising media captured by other nearby users, as well as proximate services, on-the-fly. The mobile application makes available representations of captures as they are recorded, including keyframes for videos, thumbnails of the most recent photo taken and icons representing audio clips. The mobile application also automatically retrieves these representations from nearby users and makes them available to the user in a staging area on the mobile interface. Also, a location resolution system on the mobile application continuously checks Bluetooth IDs recorded by the client against a capture service location database on the server. When the application finds a nearby service, it grabs a representation of the latest capture and places it in the staging area. Once in the staging area, icons can be dragged into the main scene. When this occurs, the system automatically saves the original file to the user's profile i.e. available via the web interface (e.g. if the user selected a keyframe from a video, the video is saved). At this point, the user can select two icons to manually link content, and can add annotations to captured content (Figure 2). In this way, the system not only automatically links content, but also exposes content that otherwise might go unnoticed.

2.1 Scenarios of use

The simplest scenario involves a single user capturing media and linking to media from capture services. Suppose that Bob, a user, wants to make a note during a presentation. If the presentation room has been tagged with a Bluetooth ID and includes a synchronous slide capture service, the system automatically processes the slide stream and makes a keyframe of the current slide available to Bob. He can drag this icon into the main scene and begin annotating it.

Now suppose Bob is in the field and wants to make a text annotation of a segment of video that his friend Marcia is recording on a standard Bluetooth-enabled digital video device. In this case, Bob will necessarily be near Marcia since he is commenting on something that she is recording. Bob can use the system to enter his comments. Behind-the-scenes, the system will automatically send with the comment a clip of 15-sec of audio recorded before and after the comment as well as a snapshot of all of the nearby Bluetooth devices. Later, when Marcia uploads her recorded video, the system will use the audio and Bluetooth data to link Bob's comment to the correct device as well as the correct sequence of video that Bob was annotating. Note that links would have been created for any type of media (e.g. rather than making a text comment Bob could have taken a picture or recorded his own video).

Bob could also take a photo of the same scene that of Marcia is recording just before or after he makes his comment. This action links Bob's comment to a particular keyframe in Marcia's video. Immediately after making the comment, Bob sees on his device the picture he took of the scene with the comment already linked. Later, when Marcia synchronises her video, Bob's picture will become an active link into the source video. Bob can use this method to create collections of media on-the-fly that are combinations of original recordings he has made as well as pointers to recordings others have made. He can organise these clips on his own device immediately – all of the linking will occur *post hoc*.

2.2 Experience

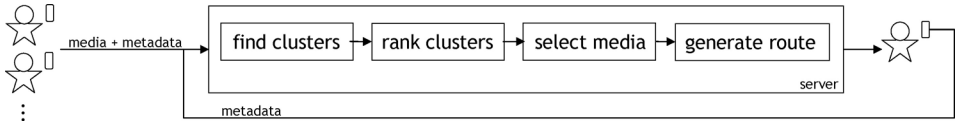
We ran studies to understand how people might use this tool to diary important events (Perttula and Carter, 2008). The studies made use of information exposed from a meeting capture service (Denoue et al., 2005) and an informal public display service (Churchill et al., 2003). These services already include web application programming interfaces (APIs) to expose data, making it easy to connect them to the database. Interviews with participants about their experiences revealed that the system should tie captures more directly not only to other captures, but also their work process. One participant mentioned that she 'want[ed] to be able to make a note on a journal paper I have printed and have the system link in the PDF in the background'. All the participants also agreed that audio was the most useful annotation mechanism.

Importantly, most participants focused on what the tool had linked to that was related to their current task as well as the tool's ability to reveal sources currently available for linking. That is, artefacts were important only if they related to the participant's current task or some task in the immediate future. In addition, participants struggled to imagine the type of data that might be useful to them later, which added a cognitive burden when new media arrived in the staging area. One participant commented that, 'it seems like I have to stop and think about every new piece of information that comes in'. Ultimately, participants expressed a need for a glanceable interface that could help them to launch new tasks.

3 Prospective prototype

Based on a brainstorming session with 15 participants, we designed the prospective system – *Kartta* – to provide a guide for users in the field based on location-based feedback by other users (Perttula et al., 2009). In this way, the system can be thought of as a Digg for locations. Kartta is composed of a server and a mobile application built using Mobile Python. The mobile application can capture content and context data which are sent to the server automatically. The server uses this information to create a map of the immediate region around a user, highlighting points-of-interest as well as landmarks to help the user to navigate. The mobile application polls the server regularly for a new map and set of landmarks (Figure 3).

Figure 3 Data flow in the Kartta system



Note: Using the mobile application, participants upload media, positive or negative votes for specific locations, and associated metadata to the server. Synchronously or asynchronously, other participants request a map from the server, optionally sending location and orientation information. The server finds and ranks clusters, selects media within each cluster, generates routes and sends the updated information back to the mobile client.

Figure 4 Kartta in a field setting

Note: (Left) The application showing the user's current location as well as votes. (Right) A zoomed-in view showing the user's current location, votes and tags.

The mobile application visualises not only the augmented map returned by the server but also the users' current location (Figure 4). Users can configure the interface either to show media captured by an explicit list of friends, or to show only their own captures. Because, as Fleck et al. (2002) point out, mobile users often feel that capturing media is too distracting, the application includes a simple, one-button interface for recommending that other users visit an area (a positive vote) or avoid it (a negative vote). A positive vote is represented on the map as a green dot and a negative vote a red dot. Overtime, votes implicitly create an interest-map of a place.

Mobile users can also launch media capture applications with another button. The mobile application sends media to the server along with context information (if available) including phone tilt (up, level or down), compass orientation, time-of-day and location. Context data can also be recorded continuously. The application currently senses location information via either embedded or attached GPS devices. While GPS currently affords only a gross estimate of location, we believe that it is sufficient since our application depends on aggregated data (votes). Furthermore, users can add tags to disambiguate areas of interest. Finally, media captured at a location automatically corresponds to a positive vote for that place (similar to the approach in Jaffe et al. (2006)).

The server ingests media and votes sent from mobile devices, finds clusters of media, selects representative captures for each cluster and generates routes from averaged and smoothed GPS traces. The server saves mobile media in a database, and a set of ingest and access services produce multiple downsampled representations of uploaded media: thumbnails of photos, keyframes and lower resolution copies of videos, and copies with lower sample rates for audio clips. The mobile application can use these different representations to provide tiered access to media.

To organise captured media, the server first finds and ranks clusters of interest areas, ranks media within each cluster and then generates route information.

3.1 Scenarios of use

Consider a group of researchers who are attending a conference with a few thousand attendees. Bob sees an interesting poster and wants to notify others. He takes a photo of it and tags the photo. After that others' devices download his photo automatically. Also, all of the group members can see a highly positive vote on the map indicating the place of the poster. Marcia zooms in and sees an interesting tag next to the vote. She presses one key and a photo pops up on the screen. She thinks that she must see that poster and checks out the map. She is already almost in the correct place but she cannot find the poster. So Marcia sends a message to Bob and asks about the poster. Bob can see Marcia on the map and he replies, 'it is behind the corner next to the stairs'. Marcia finds the poster and she thinks it is fascinating. She wants to give a positive vote to the poster and does it just pressing one key. A vote appears on the map and the rest of the group can see that there must be something interesting at that location since there are now two positive votes.

Later, Bob is sitting through an uninteresting talk. He gives it a negative vote, which is sent automatically to the others' maps. Bob also sends a message to everyone about the topic of the talk and writes that there is nothing new. Now everyone can look for something else. Marcia is at a different talk that is more interesting. She gives a positive vote to the talk and takes a photo, tagging it with the topic of the talk. Bob and others decide to move to that talk.

3.2 Experience

We first built a simulation tool to tune weights in our ranking algorithm. After running a series of simulations, we ran a pilot field experiment following (Massimi et al., 2007) scavenger hunt approach for evaluating mobile collaborative systems. In our initial pilot, four users recruited from our lab were divided into two teams of two, and we gave every user a device running our system. We used a simplified version of the mobile application that did not include route information, showed one tag per capture rather than several per cluster, and inferred orientation from a GPS trace rather than relying on an external sensor. The system came preloaded with a map of the campus near our building, which includes buildings, several flights of stairs, parking lots, a fountain, etc. The goal of the exercise was to find coloured balls (30 in total) that had been scattered around an office complex. Two researchers followed participants during the study, taking notes and helping if users were completely stuck. Also, we held a focus group session with all participants after the study focusing on potential features.

Participants collectively captured 66 total targets and submitted 45 tags over 1 hr. While the interface was not yet at a stage to judge its usability for groups, but participants provided a host of recommendations in the focus group. Participants, overall, felt that the interface required too much attention, and requested vibratory and auditory alerts. They also suggested that the map may not be necessary at all, and that the visual interface might consist only of hotspots as well as paths to those spots. They also suggested audio tagging, list views of recorded tags and orientation controls and views.

Overall, though, participant interest was high, and they were motivated to use an interface that helped them to explore an environment. Participants were much more engaged because the tool 'helped [them] uncover new opportunities' rather than only documenting their actions. They also appreciated that the glanceability of the vote display required little thought ('I can see immediately where to go').

4 Comparisons of the two prototypes

Our retrospective prototype helped the users to put their media captures in context with other media captured nearby, while our prospective prototype used media to indicate areas of potential interest. Through our pilot testing, we found that users are more likely to do *post hoc* organisation than *in situ* organisation, and that it is more important that the mobile application itself provide information but otherwise not require intense focus, or worse, get in the way of users' real-world tasks. This reinforces previous work that has found that most people would rather pay attention to the events around them than to navigate through a phone's interface (Oulasvirta et al., 2005). Furthermore, *post hoc* organisation could be accomplished using either system since the process of capturing metadata is relatively streamlined. Also, given that mobile users want to focus on *now* and *next*, it is not as useful to provide a full timeline on the mobile application for most capture tasks.

Overall, then, we found that the prospective prototype provided both *in situ* and *post hoc* value, and we decided to expand and test it in a more realistic setting.

5 Follow-up evaluation of Kartta

After improving the usability of the interface, we designed another, more focused experiment to evaluate Kartta. Furthermore, inspired by Baus et al. (2007) finding that audio can be useful for navigation tasks, we wanted to include a variety of media types. In this experiment, six single users recruited from our lab were asked to use Kartta to find and record objects of interest in a semi-familiar, semi-urban environment. We pre-recorded objects using a mixture of media – four photos (two tagged), two audio with no tags, two video with no tags, two tagged negative votes and two tagged positive votes. Participants were given a device with the Kartta application and all but three of the votes and media pre-loaded were asked to take a photo of the object they thought was being recorded or tagged at each location (e.g. one hotspot was linked to an audio clip of the chimes of a clock, of which participants were to take a photo). Media captured by participants was saved to the device locally and uploaded to the server in a separate thread. A few minutes into the study, the application downloaded and displayed the remaining votes (we added a vibrated alert to signal updates). Participants were also asked to avoid areas marked with negative votes.

One researcher shadowed the participant both to record behaviour and to answer questions about the interface in case the participant had difficulties. Since the task took place in an uncontrolled environment we did not incentivise participants to complete the study quickly. Instead, we asked them to signal to the shadower when they believed they had completed the task. At the end, we asked participants several follow-up questions designed to determine which media was most helpful for navigation, the ease of navigating the map and their ability to avoid negative areas.

5.1 Results

Despite there being no incentive, all participants completed the study in the time allotted. However, only one participant successfully captured photos of all objects, and one captured only five. On average, participants captured eight out of ten objects. Media in the set that only appeared after the start of the experiment were those most likely to be left out. All

participants chose photos as either the first or second most useful media for navigation. Four chose tags as the first or second most useful. One participant commented that it was easier to find those media that he could at least 'roughly make out from a distance'. Overall, participants found it useful that the interface reflected their orientation as well as location (four on a five-point scale on average), and they found it easy to navigate the map (also four out of five).

Participants had few questions for the shadower during the task. Though it did not affect participant's ability to complete the task, network reliability was also an issue, and we relied on media retrieved from the device for analysis. In general, participants had fun with the task, one noting that it was 'like virtual geocaching'.

6 Related work

Past work includes a variety of methods to capture data for retrospective exploration. Similar to Notelinker, Fono and Counts (2006) Sandboxes system displays collaboratively captured multimedia on mobile phones. However, this work does not address organisation (i.e. it does not utilise context to structure captures). Furthermore, it does not provide any video recording or other video-related support, such as keyframe generation. Erol and Hull (2003) describe a system to index into a presentation using an image captured with a camera phone. Their access interface displays the original captured slide and the video recording at the time it was presented (a similar system using scanned images appears in Chiu et al. (2000)). Fink et al. (2006) describe a system that senses TV audio to automatically recognise the programme the user is currently watching. They use this technology to support social viewing applications. Fleck et al. (2002) conducted an iterative deployment of a mobile capture tool in a museum setting. They quickly found that the capture application they had designed required an attentional shift away from the activity being captured that was unacceptable to users. In the end, they embedded capture technology into the environment itself and used the mobile device only to initiate an automated capture process (by swiping an RFID over a reader). In most mobile situations, though, it is not possible to instrument the environment in this way. Overall, these systems do not address collaboratively recorded media and are designed primarily for retrieval rather than *in situ* organisation and synthesis.

Some map-based interfaces are used for retrospective purposes, and in many cases they can also be used for long-term planning. One example is EveryTrail (http://www.everytrail.com/) which allows users to upload GPS tracking information as well as GPS-tagged photos. The system automatically plots that information on Google maps and offers a variety of social interactions around uploaded media (comments, ratings, etc.). However, the tool is not designed to collate and redisplay automatically information from multiple users.

Past work in prospective mobile location-based media has focused on creating summaries of collected media, descriptions of static environments or non-spatial visualisations of user-generated content. Jaffe et al. (2006) built a system that uses location and other contextual information to select key photos from a collection. However, their work focuses on summarisation rather than navigation. Grabler et al. (2008) developed a prototype that automatically generates maps based using building textures, road geometry and external landmark information. In this work, the authors focus on static structures and landmarks rather than temporary environments. Studies by Baus et al. (2007) showed that audio landmarks can aid navigation tasks, but their work did not involve user generated content.

GeoNotes (Espinoza et al., 2001) supported user generated place annotation, but was focused on text.

Of course, many applications can be used for both retrospective and prospective purposes. For example, a geonote can archive an experience (retrospective) or remind the user to complete a task (prospective). Still, both our work and past work suggest that users typically want to offload retrospective tasks from the mobile device as much as possible, focusing instead on prospective tasks.

7 Conclusions and future work

We found that attentional demands dominate and mobile media applications should distract as little as possible and provide as much information as possible peripherally or using no visuals at all. To this end, we intend to extend our prospective system further to support glancability and eyes-free notification. In particular, the phone should vibrate when a user is very close to an interesting area or when someone within walking distance votes up an area. We also plan to add audio notifications, such as the name of a user who just made a capture. Finally, we plan to deploy our prospective system to a realistic environment in which people overlap in time while exploring a new space, such as a conference.

Acknowledgements

This work is based on an earlier work: 'Kartta: using multimedia and context to navigate unfamiliar environments, in the *Proceedings of the 13th International MindTrek Conference: Everyday Life in the Ubiquitous Era*. Tampere, Finland (September/October 2009). ©ACM, 2009. We would like to thank our pilot users for their patience.

References

Baus, J., Wasinger, R., Aslan, I., Krüger, A., Maier, A. and Schwartz, T. (2007) 'Auditory perceptible landmarks in mobile navigation', *IUI*, pp.302–304.

Chiu, P., Foote, J., Girgensohn, A. and Boreczky, J. (2000) 'Automatically linking multimedia meeting documents by image matching', *ACM Hypertext*, pp.244–245.

Churchill, E., Nelson, L., Denoue, L. and Girgensohn, A. (2003) 'The plasma poster network: posting multimedia content in public places', *Interact*, pp.599–606.

Denoue, L., Hilbert, D., Adcock, J., Billsus, D. and Cooper, M. (2005) 'ProjectorBox: seamless presentation capture for classrooms', *E-Learn*, pp.1986–1991.

Erol, B. and Hull, J.J. (2003) 'Linking presentation documents using image analysis', *Signals, Systems and Computers*, pp.97–101.

Espinoza, F., Persson, P., Sandin, A., Nyström, H., Cacciatore, E., and Bylund, M. (2001) 'GeoNotes: social and navigational aspects of location-based information systems, *Ubicomp*, pp.2–17.

Fink, M., Covell, M. and Baluja, S. (2006) 'Social- and interactive-television applications based on real-time ambient-audio identification', *EuroITV*, pp.138–146.

Fleck, M., Frid, M., Kindberg, T., O'Brien-Strain, E., Rajani, R. and Spasojevic, M. (2002) 'From informing to remembering: ubiquitous systems in interactive museums', *IEEE Pervasive Computing*, Vol. 1, No. 2, pp.13–21.

Fono, D. and Counts, S. (2006) 'Sandboxes: supporting social play through collaborative multimedia composition on mobile phones', *CSCW*, pp.163–166.

Grabler, F., Agrawala, M., Sumner, R.W. and Pauly, M. (2008) 'Automatic generation of tourist maps', *Siggraph*, pp.1–11.

Grabner, M., Grabner, H., Bischof, H. (2006) 'Fast approximated SIFT'. *Asian Conference on Computer Vision*, pp.918–927.

Greenburg, S. and Buxton, B. (2008) 'Usability evaluation considered harmful (some of the time)', *CHI*, pp.111–120.

Jaffe, A., Naaman, M., Tassa, T. and Davis, M. (2006) 'Generating summaries and visualization for large collections of geo-referenced photographs', *Multimedia Information Retrieval*, pp.89–98.

Klemmer, S. (2002) 'A pervasive computing framework supporting collaboration in documentary history projects', *DIS.*

Massimi, M., Ganoe, C.H. and Carroll, J.M. (2007) 'Scavenger hunt: An empirical method for mobile collaborative problem-solving', *IEEE Pervasive Computing*, Vol. 6, No. 1, pp.81–87.

Oulasvirta, A., Tamminen, S., Roto, V. and Kuorelahti, J. (2005) 'Interaction in 4-second bursts: the fragmented nature of attentional resources in mobile HCI', *CHI*, pp.919–928.

Perttula, A. and Carter, S. (2008) 'Retrospective vs. prospective: a comparison of two approaches to mobile media capture and access', *Social Mobile Media Workshop*.

Perttula, A., Carter, S. and Denoue, L. (2009) 'Kartta: using multimedia and context to navigate unfamiliar environments', *MindTrek*, pp.120-123.

Shrage, M. (2000) *Serious Play*. Boston: Harvard Business School Press.

Trevor, J., Hilbert, D. and Schilit, B. (2002) 'Issues in personalizing shared ubiquitous devices', Workshop on user centered evaluations for ubiquitous computing systems: best known methods. *Ubicomp*.