Browsing Personal Media Archives with Spatial Context Using Panoramas

Brett Adams, Stewart Greenhill, Svetha Venkatesh
Department of Computing
Curtin University of Technology
GPO Box U1987, Perth, 6845, W. Australia
{adamsb,stewartg,svetha}@cs.curtin.edu.au

ABSTRACT

This paper presents novel techniques for using panoramas as spatial context to enhance browsing of personal media archives. This context, scenes where frequent media capture takes place, is present in the disparate photos and videos, but not leveraged by traditional browsing techniques (e.g. thumbnails or zoomable interfaces). Coarse geo-position is often an insufficient index at such media capture hotspots. We experiment with panoramic video, which presents archive video organically blended with panoramas of media capture hotspots; Immersive browsing and filtering with media items projected onto spherical panoramas; and Detection and representation of links between panoramas to enable browsing of situated media in quasi-3D. We present proof-of-concept implementations and observations of their effectiveness, limitations, and open problems. Experiments confirm the intuition that each holds promise for augmenting traditional browsing environments.

Categories and Subject Descriptors

H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing; H.5.1 [**Information Interfaces and Presentation**]: Multimedia Information Systems

General Terms

Algorithms, Human Factors, Experimentation

Keywords

Multimedia browsing, spatial context

1. INTRODUCTION

Personal archives of photos and videos are increasingly digital and voluminous, due to capture device power and portability. A persistent topic of research is applications for managing, searching, and browsing collections, and their enabling technologies. Location-aware media applications have received much interest, particularly following the advent of GPS-enabled cameras and smartphones. Media captured by these devices is able to be indexed by raw global coordinates, and more recently by regions of significance to the user, such as *locations*–a socially loaded term. These are often characterized at the resolution of a building, such as a home, vacation spot, or workplace, due to the specific application in mind or the physical constraint of loss of GPS signal indoors.

For locations at which media is captured infrequently, this coarse geo-position is sufficient to index individual media items. But for locations where much media is captured, such as the home, spatial indices at this resolution are insufficient to filter media items into manageable display sets. In this case we require a finer resolution of socially meaningful location: the *scene*, rooms or distinct living spaces within or around a building. We are motivated by the desire to situate browsed media in this familiar context. We hypothesize that panoramas of frequently used scenes can be used to effectively anchor media, and present three techniques and proof-of-concept implementations for representing this context and integrating it with a browsing environment:

1. Organic blending of new media into flat panoramas of scenes
2. Anchoring and floating filterable sets of new media in spherical panoramas of scenes
3. Linked panoramic spaces of scenes to enable browsing and retrieval of situated media in quasi-3D.

The advantage of this approach is in the utilisation of frequented scenes as the familiar background context and, by situating new media in this context, allowing for natural navigation and browsing. The novelty lies in the use of panoramas as context, and their integration with a media browsing environment as both coarse and fine spatial filters.

2. PANORAMAS FOR PERSONAL MEDIA BROWSING

Representation of room-level physical position of items is a useful index for information-centric, specific item search (e.g. I remember taking the shot of our friends in the dining room). Additionally, physical *proximity* aids two further browsing activities: discovery of related, unknown items, termed situational browsing[7, p. 266] (e.g. A photo of my son covered in dinner at the same meal); and discovery of unrelated items, termed opportunistic browsing[7, p. 270] (e.g. A video of the family unwrapping Christmas presents

MM'06, October 23–27, 2006, Santa Barbara, California, USA.
ACM 1-59593-447-2/06/0010.

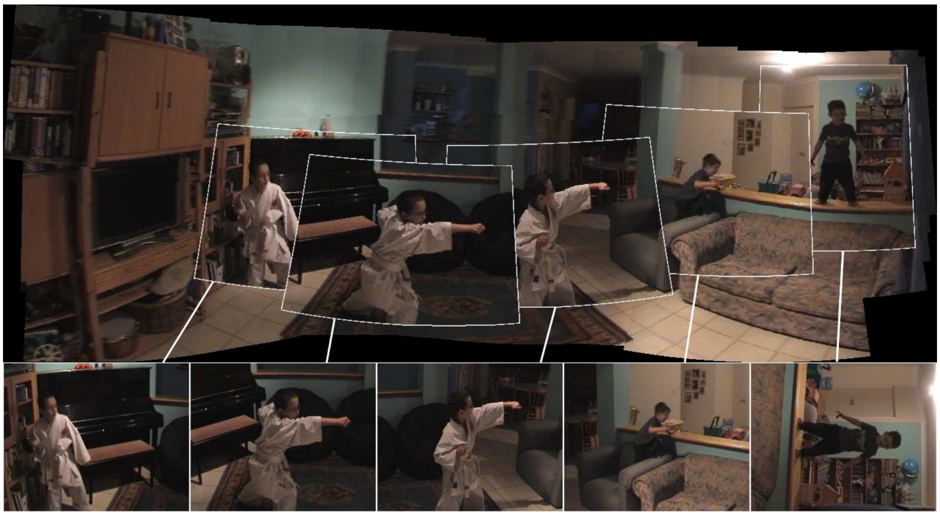


Figure 1: Panoramic video preserves context of video sequence, inherently removing artifacts such as rotation

nearby). Panoramas enable a visual representation of physical proximity analogous to a user's perception.

Panoramas have been used for a variety of applications, such as robot navigation, immersive telepresence[4], video and image summarization[2, 11], object tracking, remote observation[5], rectification of home video[13], and even therapy[8]. A goal common to many of these applications is to impart vicarious presence, the feeling of 'being there,' often operating on the assumption of supplying *unknown* spatial context to the user. For the application of media browsing, panoramas can do the same for a user browsing a collection captured at scene(s) with which they are unfamiliar, such as the case of an interstate relative viewing a shared, online collection. But for the owner of the media, or at least one familiar with its creation context, panoramas provide visual reinforcement of the *known* context of media items, which is lacking in the isolated items themselves.

In order to use panoramas for personal media browsing we must firstly generate them for areas of interest. This is ideally done opportunistically, in view of the typical unwillingness of users to expend effort structuring their collections. An automated process might unsupervisedly mine a collection of photos for panoramas using a panorama stitcher such as Autostitch[1]. Such a process, however, is unlikely to obtain coherent panoramas given the requirement of a roughly fixed viewpoint across multiple, arbitrary photos.[1] Yield for the same process applied to videos in a collection would potentially be higher given that typical camera operations, such as pan and tilt, have by definition a fixed viewpoint, and successive frames have the large regions of overlap required to register images when motion is not extreme. Additionally, candidate image segments for panorama construction would also need to be filtered of those with moving objects, which would result in ghosts (It may even be desirable to detect moving objects against *dynamic* backgrounds [12]). For this work, all panoramas were obtained manually by 'painting' an area of interest with a DV camera. Given the simplicity of the procedure and the high utility of the panorama created–a single panorama of a media hotspot can be valid for months–a user might conceivably expend the effort required.

After panorama creation, media items must be grouped with them and physically located within each. Firstly we might group a photo of a toy truck with the 'back yard,' and secondly, we might be able to locate it 'over by the swing.' We focus on what can be done if the scene *and* position of an item is discovered, while noting that the less constraining problem of simply grouping an item with a panorama could be framed as an unsupervised classification problem, fusing evidence from traditional image features (e.g. colour histograms, texture descriptors) and media capture patterns (e.g. photos clustered in time–i.e. taken in bursts–are likely to be within the same scene).

Given a set of panoramas, and the media contained by them, some of it precisely positioned, how can we use this information to improve browsing efficacy? Below we outline three ideas, each of which is envisaged as being additive within a unified browsing application.

3. PANORAMIC VIDEOS

We first experimented with the usefulness of panoramic home videos. Video is by nature a 'high context' medium. It is able to convey much information through spatial and temporal continuity in both visual and aural modes. Panoramic video can further augment this ability. Specifically, we investigated the construction of flat, panoramic video, where live video is blended organically with an existing panorama. A number of benefits are obtained as a side-effect of panoramic display, all of which are problems of amateur video researched in their own right: camera shake, zoom, and rotation removal, and illumination smoothing.

Panoramic videos were built with the following procedure. The room of interest is first filmed in overlapping pans, a task requiring less than a minute. The DV format video is then de-interlaced and converted to a sequence of JPEG images. A subset of all frames are then selected at a fixed temporal interval to be the source images of the panorama.[2]

[1] Indoor environments may offer more opportunity for panorama creation than first imagined, as furniture, field of view, and repetitiveness in functional use of space, can constrain the positions from which media can be captured.

[2] Steps of 5 & 10 frames were used. Finer sampling creates a more stable panorama but requires greater computation.

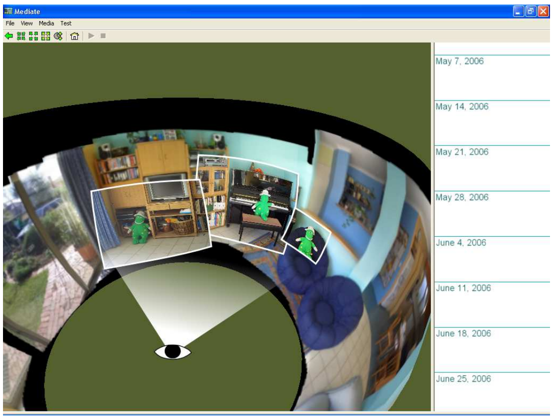


Figure 2: First-person browsing: Media satisfying filter set floats in position, projected onto panorama surface

A video is chosen to be blended with the panorama and is subject to the same procedure, barring only that every frame is kept. Each frame of the video is pooled with the frames obtained from the panoramic sweep, and a panoramic frame created using Autostitch[1, 3] (a SIFT feature based automatic panorama builder). Panoramic frames thus built have constant width but variable height, due to the iterative process of creating panoramas for the original source images plus each new frame of video. A better method would be to first determine the panorama with optimal distortion for *all* video frames and register each video frame against that same, fixed panorama. The last step is to cut them to constant dimensions and render them as a video.[3] Figure 1 depicts an example panoramic video together with frames from the source video.

Projecting live video into a static panorama can produce disconcerting artifacts (e.g. children with missing legs). Extending the user's virtual 'persistence of vision' by continuing to project old frames outside the footprint of the live video can remedy this if the artifact was produced by a zoom-in in the first place. We noted also that zoom is often used purposefully. In this case, the ability for the user to also zoom on the panoramic display, which is of variable resolution, is desirable. We carry this requirement into the second experiment. Finally, while the flat panorama provides more context, it is not immersive.

4. BROWSING WITHIN PANORAMAS

In our second experiment we targeted the browsing activity, of both photos and videos, in a manner that better matches the user's own perception of the living spaces represented by panoramas: the user is provided a first-person perspective with a given spherical panorama wrapped about them. 'Camera' controls are provided allowing full rotation, zoom, and even translation, although it generally is of little use to move from a panorama's central viewpoint.

Each panorama in the browsed collection has media items grouped with it, and a subset of these have information for positioning them within it. Each photo in that subset is floated in place via a matrix that projects it onto the panorama surface.[4] Registration is performed with the full set of fixed source images S^i for each panorama P_i plus the image to be registered I^i_n. Each video V^i_m in P_i has a sequence of projection matrices for each of N frames V^i_{mn} sampled at a fixed step, and interpolation is used to describe a trajectory over the panorama in time. Interpolation inherently deals with the occasional mis-registered frame, as well as jitter caused by registering each frame against a slightly different panorama, a product of the small differences in bundle adjustment for the set $S^i \cup V^i_{mn}$ as each frame is placed. The user's perspective can be locked onto a video, in which case the panorama is seen to gyrate and scale about the fixed video. In addition to grouping media within different panoramas, browsing is performed within a panorama by applying filters to the displayed set. Photos and videos have a timestamp embedded in the EXIF header of images and video thumbnails, which can be filtered on at varying resolutions and offsets using simple zoom and scroll metaphors, respectively. Items can also be filtered by size. This allows search based on implicit categories such as group photos and close-ups, an index uniquely derived from the spatial context provided here by panoramas. A panorama can also have a time window attributed to it indicating its period of validity, beyond which, presumably, differences with the actual space depicted become too great for it to be of use as an orienting tool. Figure 2 depicts an example browsing session with panorama and automatically placed items.

How to represent varying levels of information about media items (location, location and scene, location and scene and position) in a unified browsing experience is in general an open problem. We use a layered approach, where the restrictive nature of each layer is communicated to the user.

5. NAVIGATING PANORAMAS IN "3D"

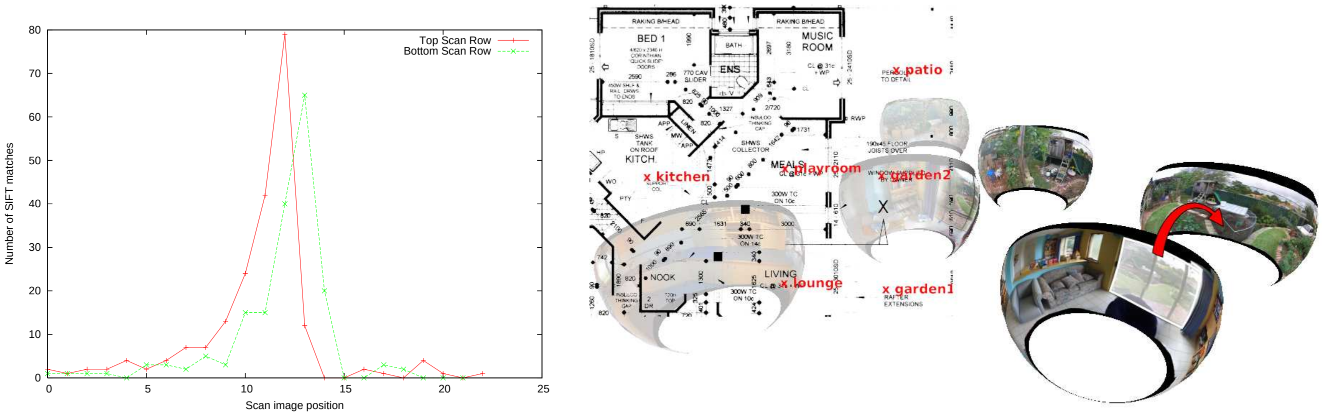
Above we detailed work aimed at more closely approximating the browser's representation of living spaces with the user's own. In addition to this disjoint context, the user also has knowledge of the physical *interconnections* among those living spaces, and it is this aspect that we focussed on discovering and representing in the third experiment. I.e., we discover the graph of links among panoramas and their relative orientations.

The intuition here is that panoramas taken from one viewpoint, say the middle of a room, will include image segments belonging to another room or adjoining space. This will occur where line of sight is unobstructed, such as in outdoor areas, open-plan living spaces, or doorways. Thus images or image segments may be registered into multiple panoramas.

Links between panoramas are detected by performing a brute force search of all pairs of source images S^i_m, S^j_n from panoramas P_i and P_j, where $i \neq j$. If registration occurs, links between an image (region) and a panorama are created in both directions, $L(S^i_m, P_j, S^j_n)$ and $L(S^j_n, P_i, S^i_m)$. Figure 3a. is a plot showing SIFT keypoint matches spiking at angular coordinates corresponding to a common view. A link is represented on the panorama surface by highlighting the image region footprint and as a coloured transparency on mouse-over. Clicking on a link $L(S^i_m, P_j, S^j_n)$ will navigate to the center of the panorama P_j, leaving the user facing

[3] Trimming the unrolled, spherical panoramas results either in removal of a small amount of detail from the bottom left or right corners, or else padding with black.

[4] This matrix is the product of image registration performed with Autostitch, which follows SIFT feature matching and robust pose detection[3].



a.

b.

Figure 3: a. A spike in SIFT keypoint matches indicates angular coordinates of panorama intersection for two rows of images; b. Panorama interconnections and navigation via hotspots

the direction of the image S_n^j. Figure 3b. depicts a graph of interconnections potentially detectable with this process.

Link discovery is worst-case $O(|P||S^i|^2_{max})$, and we note $|P| << |S^i|_{max}$ is likely true for the domain of personal media. This drops to $O(|P||S^i|_{max})$ and $O(|P|^2)$, if registration is performed image to panorama, and panorama to panorama, respectively. Registration directly against panoramas performed poorly as pose consistency checking was not modified to account for transformations applied during bundle adjustment. Incorporation of the warping applied during panorama construction might improve registration against panoramas and hence achieve significant speed-up.

Manual link creation would be a simple exercise for the motivated user. Link hotspots could be placed by hand and target panoramas indicated by drag and drop. Alternatively, a map interface could be provided upon which the user indicates scene locations. Relative scene geometry could then be determined automatically.

6. CONCLUSION

We have presented a novel media browsing environment using panoramas, with the chief motivation being the visual reinforcement of the fine context of captured media–something present in user's mind, and in the isolated media to a degree, but largely missing from traditional browsing environments. We conducted informal experiments into panoramic video, panoramically positioned browsing, and inter-panorama browsing in quasi-3D. Each was found to hold promise as another arrow in the quiver of traditional personal media browsing technologies.

Further opportunities for exploration abound. Generation of coarse depth maps combined with view interpolation would help reduce the feeling of claustrophobia generated by lack of parallax and other depth cues[10]. Statistically significant experimentation of the failure modes of image registration for the 'genre' of personal media (e.g. family photos), a subset of generic images, are required. E.g., SIFT keypoints are theoretically invariant under 30°of affine shift; How limiting is this for typical collections? For video-rich collections, composite environment maps consisting of panorama types appropriate to different camera operations–spherical for pan and tilt, cylindrical for dolly–might better approximate a user's experience of a familiar space and make for compelling browsing. Verisimilitude and hence immersion would be enhanced by dynamic elements, such as video textures[9] or dynamosaics[6].

7. REFERENCES

[1] Autostitch. www.autostitch.net, 2006.

[2] A. Aner and J.R. Kender. Video Summaries through Mosaic-Based Shot and Scene Clustering. In *Proc. of the European Conference on Computer Vision*, Denmark, 2002.

[3] M. Brown and D. G. Lowe. Recognising panoramas. In *Proc. of the 9th IEEE Int. Conference on Computer Vision*, page 1218, Washington, DC, USA, 2003.

[4] A. Majumder, W. Seales, M. Gopi, and H. Fuchs. Immersive teleconferencing: a new algorithm to generate seamless panoramic video imagery. In *Proc. of the 7th ACM Int. Conference on Multimedia*, pages 169–178, New York, NY, USA, 1999.

[5] N. Qin, D. Song, and K. Goldberg. Aligning windows of live video from an imprecise pan-tilt-zoom robotic camera into a remote panoramic display. In *IEEE Int. Conference on Robotics and Automation*, May 2006.

[6] A. Rav-Acha, Y. Pritch, D. Lischinski, and S. Peleg. Dynamosaics: Video mosaics with non-chronological time. In *Proc. of the IEEE CS Conf. on Computer Vision and Pattern Recognition*, pages 58–65, Wash., DC, USA, 2005.

[7] R. Rice, M. McCreadie, and S.-J. Chang. *Accessing and Browsing Information and Communication.* MIT Press, 2001.

[8] A. Rizzo, L. Pryor, R. Matheis, M. Schultheis, K. Ghahremani, and A. Sey. Memory assessment using graphics-based and panoramic video virtual environments. In *Proc. 5th Intl Conf. Disability, Virtual Reality & Assoc. Tech.*, 2004.

[9] A. Schodl, R. Szeliski, D. Salesin, and I. Essa. Video textures. In *Proc. of the 27th annual conference on Computer graphics and interactive techniques*, pages 489–498, New York, NY, USA, 2000.

[10] R. Szeliski. Video mosaics for virtual environments. *Computer Graphics and Applications, IEEE*, 16(2):22–30, March 1996.

[11] X.-S.Hua, S. Li, and H.-J. Zhang. Camera notes. In *Multimedia and Expo, 2005. ICME 2005*, July 2005.

[12] J. Zhong and S. Sclaroff. Segmenting foreground objects from a dynamic textured background via a robust Kalman filter. In *Proc. of the 9th IEEE Int. Conference on Computer Vision*, pages 44–50, 2003.

[13] W.-Q. Yan and M. Kankanhalli. Detection and removal of lighting & shaking artifacts in home videos. In *Proc. of the 10th ACM Int Conference on Multimedia*, pages 107–116, New York, NY, USA, 2002.