Video Collections in Panoramic Contexts

James Tompkin[1,5] **Fabrizio Pece**[2] **Rajvi Shah**[1,3] **Shahram Izadi**[2,4] **Jan Kautz**[2] **Christian Theobalt**[1]
MPI für Informatik[1] University College London[2] IIIT Hyderabad[3] Microsoft Research[4] Intel VCI[5]

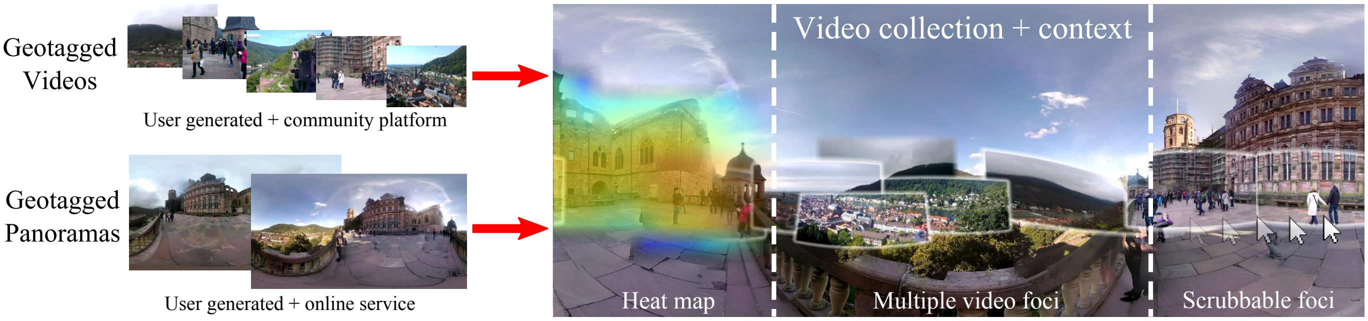
Geotagged Videos
User generated + community platform
Geotagged Panoramas
User generated + online service
Video collection + context
Heat map
Multiple video foci
Scrubbable foci

Figure 1: Panoramas are widely available online, and more and more video content of these places is shared online. With these data, our video-collection+context interface visualizes the dynamic changes within a collection. The right-hand side shows our spatio-temporal index as a heat map (*left*), inlayed video foci (*center*), and fast search with spatial mouse scrubbing (*right*).

ABSTRACT

Video collections of places show contrasts and changes in our world, but current interfaces to video collections make it hard for users to explore these changes. Recent state-of-the-art interfaces attempt to solve this problem for 'outside→in' collections, but cannot connect 'inside→out' collections of the same place which do not visually overlap. We extend the focus+context paradigm to create a video-collections+context interface by embedding videos into a panorama. We build a spatio-temporal index and tools for fast exploration of the space and time of the video collection. We demonstrate the flexibility of our representation with interfaces for desktop and mobile flat displays, and for a spherical display with joypad and tablet controllers. We study with users the effect of our video-collection+context system to spatio-temporal localization tasks, and find significant improvements to accuracy and completion time in visual search tasks compared to existing systems. We measure the usability of our interface with System Usability Scale (SUS) and task-specific questionnaires, and find our system scores higher.

Author Keywords

Video; video collections; panoramas; space-time exploration.

ACM Classification Keywords

H.5.2 User Interfaces: Evaluation/methodology, Interaction styles; H.5.1 Multimedia Information Systems: Artificial, augmented, and virtual realities, Evaluation/methodology.

INTRODUCTION

The abundance of mobile devices with cameras has resulted in an ever increasing number of videos of places around the world. With geotagging, it is very easy to assemble a video collection containing many videos of the same location spanning a period of time. Such a collection can capture both the moment-to-moment dynamics of a location, the comings and goings, and its temporal evolution across days, months, seasons, or years. However, exploring these dynamic changes within places is difficult for users as existing interfaces do not explicitly connect the spatio-temporal content and display it within a unifying context. For example, a virtual tourist wishing to explore the dynamic events taking place over time in a famous square can only see videos in isolation, and has no easy tools to search within the space or time of the place.

Current mapping applications such as Google Maps link videos geographically and provide ways to find videos taken from the same place. However, they do not explicitly relate the changes over space and time into a single view for easy comparison, and users must watch videos in turn. State-of-the-art research systems for video collection browsing, such as Unstructured Video-based Rendering [2] and Videoscapes [31], try to find visual links within videos that all observe the same content either at the same time or across different times. However, often the contents of a geotagged video collection captured from the same place will not visually match because the videos all look out from approximately the same spot — we define these contents as 'inside→out'. For instance, two videos of a touristic vista might take in side-by-side views but never intersect. Further, for many interesting places, it is impossible to 'go around' and we can only 'look around', such as atop the Eiffel Tower. This forbids the application of existing vision-based matching systems which rely on cameras in different positions which converge to a common scene — we define these contents as 'outside→in' because the cam-

UIST'13, October 8–11, 2013, St. Andrews, United Kingdom.
Copyright © 2013 ACM 978-1-4503-2268-3/13/10...$15.00.
http://dx.doi.org/10.1145/2501988.2502013

eras surround the subject. As such, currently it is difficult to structure, relate, and explore 'inside→out' video collections.

To solve this problem, we introduce *Vidicontexts*, a system that embeds videos into the common context of a panoramic frame of reference. Vidicontexts extends the focus+context paradigm and enables the simultaneous visualization of individual videos as multiple foci, and through the context allows the exploration of how videos are spatially and temporally related even though there might be no direct visual match between them. This alleviates the difficulty of spatially and temporally exploring 'inside→out' video collections.

To this end, we align geotagged video from mobile devices to a panoramic context using orientation sensor data (if available) and time stamps. Omnidirectional panoramas exist for many places from online street mapping platforms, and recent work enables accurate pairing of geolocated images and panoramas [16]. Further, panorama stitching is a common easy-to-use application for mobile devices. These sources provide readily available contexts for our video collections.

We explore the application of our interface on different display/input devices: a flat desktop display with a mouse in both perspective and equirectangular projections, a tablet display with perspective projection driven by orientation sensors, and a spherical display with joypad or tablet controls. In a supplemental video, we demonstrate our system and the possible spatio-temporal interactions on various datasets and display types. We verify Vidicontexts utility with user studies that measure the accuracy and time taken of spatio-temporal video location tasks against existing systems. We also investigate the usability of Vidicontexts with two questionnaires, and find it compares favorably to two existing systems.

Our contributions are:

1. A system for exploring and manipulating video collections of places within panoramic contexts.
2. A demonstration of the flexibility of our representation with interfaces for different display devices.
3. A study that quantifies the performance benefit of our tools and assesses desirability and usability when compared to existing systems.

RELATED WORK

Panoramic Imaging

Panoramas make an attractive context as they provide wide or omni-directional views of an environment in a single image. There are many established methods to construct panoramas, including using special camera hardware or by stitching individual photographs [3, 5, 30] or videos [1]. Panoramas for hundreds of thousands of places are available through mapping portals such as Google Street View or photography platforms like Panoramio.

Omnidirectional panoramas can be rendered in a variety of ways. Recent work [20] has explored the influence of varying projections on how users are able to locate scene objects. Their work concludes that clear and understandable visualization of the panorama is more important than accurate spatial mapping to enable users to exploit the 360° content.

Spatio-temporal Media Exploration

Exploring large collections of unstructured images depicting the same location can be sometimes difficult or cumbersome. The research community has tried to solve this problem by developing spatio-temporal photo visualization applications. Photo Tourism [29, 28] is one example, as the program aims to arrange and display a set of images in a 3D space so that spatially-confined locations can be interactively navigated. Similarly, the PhotoScope work of Wu et al. [32] extends the standard photo browsing paradigm by visualizing spatial coverage of construction site photos on a 2D map, and by indexing them with a combination of spatial coverage, time, and content specifications.

RealityFlythrough [19] uses videos combined with GPS and orientation data as its input. Videos are situated in a 3D representation of the world, allowing the user to navigate freely while continually transitioning to the most appropriate video for the current view. The system provides the user with some sense of how the videos relate to one another spatially, but no further context is provided and only one video is ever played back at the same time.

Unstructured video-based rendering [2] combines contemporaneous video streams of the same scene or performance, and provides an intuitive 3D-aware interface to these videos. It requires an image-based 3D reconstruction of the scene from photographs beforehand. Tompkin et al. [31] introduce Videoscapes to explore sparse unstructured video collections. They build a graph of videos by visual similarity, exploiting this graph to generate 3D reconstructions at nodes, and then provide various different interfaces to explore this graph with seamless transitions. These works use 'outside→in' assumptions as mentioned earlier, and usually only show the spatio-temporal changes when transitioning between two videos at a time. Furthermore, they require substantial additional data, such as photos to reconstruct a geometric background model or a graph of hundreds of videos. In contrast to our 'inside→out' approach, without an enveloping context they would fail to show videos taken from the same place but with non-overlapping views of the scene.

Dale et al. [7] introduce a system for browsing multiple videos with a common theme, such as the result of a search query on a video sharing website or videos of an event covered by multiple cameras. This browsing companion enhances a primary video by showing thumbnails of other temporally synchronized video clips.

Spatially-enabled exploration of single videos in isolation has also been researched. Hermans et al. [12] visualize a single tripod video as a single augmented panoramic video. Dynamic foreground and background objects are segmented and decoupled to re-time motions in the original video footage. Pongnumkul et al. [26] introduce a map-based storyboard system that presents a single tour video with different coherent shots at different locations pinned to a map.

Focus+context and Video+context Applications.

Focus+context systems show a subset of information in full detail within a wider context of surrounding lower-density de-

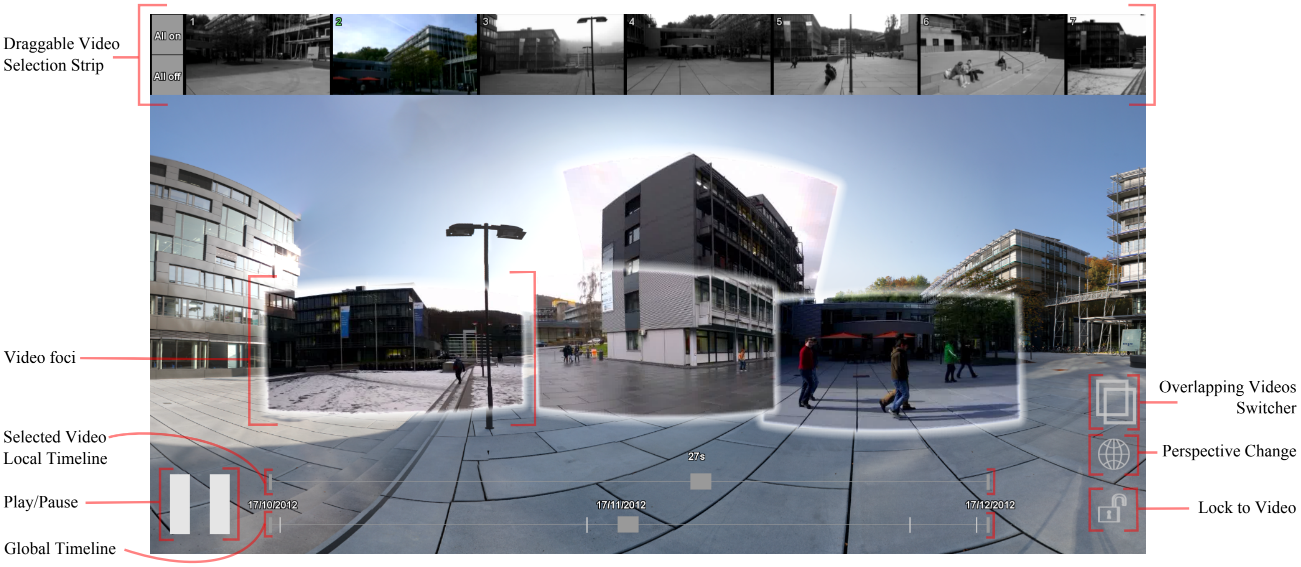

Figure 2: The Vidicontexts interface. Different months of the fall season — rainy October, cloudy skies at twilight in November, snow in December — and dynamic objects are added to the summer scene context.

tail [6]. CamBlend by Norris et al. [23] extends focus+context interfaces to apply to panoramic video collaboration tools. A smaller focus window is moved around within a larger panoramic video to identify objects to viewers of the scene.

Neumann et al. [22] introduced 'live' augmented virtual environments, where video from static surveillance cameras is projected onto geometric models from LIDAR data of a city. Follow up work attempts live painting of the bare geometry environment with texture from video from a mobile observer [13]. de Haan et al. [8] overlay static security video feeds onto geometric models for virtual first-person viewing. Kim et al. [15] propose methods for augmenting aerial visualizations of Earth with dynamic video information. However, the natures of the data (aerial looking down vs. our many inside→out overlapping videos) dictate different and novel interaction tools. Moreover, the interactions in their work are speculative, while we demonstrate significant improvements in studies. These methods were extended to provide automatic camera control for tracking dynamic objects in virtual environments that have been augmented using multiple sparse static video feeds [27]. De Camp et al. [9] map an indoor environment spatially top down, where each room is covered by one omnidirectional camera feed.

Pece et al. [24] present a teleconferencing system with smartphone cameras to create a surround representation of meeting places. Live videos from smartphones is inserted into a static panorama using marker- and image-based tracking. However, this result is not used to develop a system for exploring the space and time of a video collection.

Similar to our work, Pirk et al. [25] enhance panoramas with embedded videos to create a new interactive medium. Videos are captured from tripods at the same time as the panorama is captured. Our paper creates and evaluates an interface for spatio-temporal visualization and interaction within video-collection+contexts, using heterogeneous hand-held videos captured at different times than the panorama, whereas their paper focuses on seamlessly blending videos into a panorama using hand-segmented dynamic objects.

VIDICONTEXTS SYSTEM

Vidicontexts[1] (Fig. 2) takes as input a panoramic image and a collection of videos with time stamps, GPS data, and orientation sensor data. We first track and align all videos within the panorama, which yields a sequence of homographies for each video. Next, we build a spatio-temporal video index for exploration. Finally, we provide an interface to explore the collection of videos within their panoramic context. In general, any task that requires spatial or temporal reasoning would benefit from our system. A user might browse a collection of videos to locate object in space/time, follow videos, infer temporal changes, highlight captured regions, filter and isolate video instances that belong to a particular time span or spatial bounds; broadly, relate videos within a collection. Sport, museum, cultural sites, social events, surveillance, and tourist videos could be browsed and analysed.

Capture and Context

Our panoramic contexts can come from online repositories such as Google Street View, panoramic cameras, and DSLR stitches, or from user-assisted tools included in many mobile devices. Although any suitable source could be used, to create the material for the demonstrations in this paper we use Microsoft Photosynth on smartphones and Microsoft Research's ICE for stitching photos from a DSLR. Next, we captured several example video collections ourselves from roughly the same location as the panoramas, returning to the same locations over time. We used Samsung Galaxy II and HTC OneX smartphones to capture both video, GPS location, and orientation data (from integrated accelerometer, gyroscope,

[1] http://gvv.mpi-inf.mpg.de/projects/Vidicontexts/

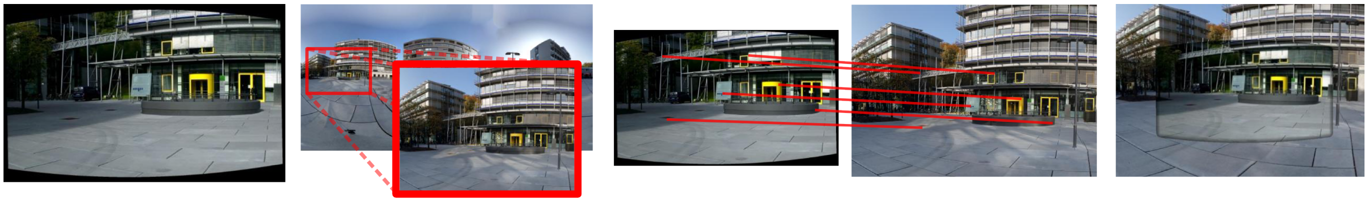
(a) Spherical projection (b) Localization with sensor data (c) Feature-based alignment (d) Panorama composition

Figure 3: (a) Videos a projected into a spherical coordinate system and (b) localized using orientation sensor data. Within this localization, SIFT features are extracted and robustly matched to estimate an alignment (c) for compositing (d).

and magnetometer sensors via the Android API). This camera orientation estimate provides an initial registration to the panoramic context. For online panoramas, pairing geotagged videos to geotagged panoramas can be difficult when GPS data is inaccurate. We assume existing work picks the closest geographical panorama from an online repository [16].

Video Alignment

Orientation data provides only approximate video alignment to the context. Accurate spatial localization is made difficult by *a*) hand-held video capture with jitter; *b*) time changes between context and videos, causing lighting changes, static and dynamic object changes, and broad scene appearance changes from seasonal variation; and *c*) the computational cost of alignment traded-off against the need to handle collections of videos. Despite variations in capture pose, we assume that the spherical panorama is a good proxy geometry for the scene, and we align the video frames to the spherical panorama using sensor- and feature-based image alignment (Fig. 3).

Spherical Projection:

We transform perspective videos into spherical projection with focal length metadata and pitch and roll orientation data from our smartphones [3]. If the focal length, pitch, and roll estimates are accurate, and if there is no parallax, then the spherically transformed video frames would be related to the equirectangular panorama by a translational model; however, due to errors in these estimates, we allow more freedom in the alignment transformation by using a homography model.

Feature Extraction:

We extract SIFT features [18] from the spherically-warped video frames and from the panorama. As feature extraction is a frame-independent task, we parallelize it.

Sensor-data based Localization:

We localize video frames approximately within the panorama using orientation data. Given this, we only match panorama features to video features within a bounding box 20% larger than the approximate localization. This reduces matching time and false matches significantly. For videos with no metadata or sensor readings, we perform an initial robust feature-based match between the panorama and the video to discover approximately the focal length, pitch, and roll angles.

Homography Estimation:

With 4 or more matches between frames and panorama, we can estimate a homography between each video frame and the panorama using the gold standard algorithm [11]. For further refinement, we use the estimated homography to find inliers from the initial set of matches and re-estimate the homography using inliers only [10]. This refinement step is repeated for three iterations in our experiments. As we have a strong expectation for a translation transformation, we perform conservative homography outlier rejection and remove homographies that are not projective or that have a large skew factor.

Estimation of Missing Homographies:

With outlier rejection, it is possible that no good homography is found for short sequences of frames. We approximate these missing homographies: with a neighboring valid homography as a starting point, we accumulate sensor orientation changes until we find a valid homography end point, and then integrate the resulting error over the length of the missing sequence.

Temporal Filtering:

Since frame homographies are estimated independently, some temporal jitter remains due to small but independent alignment errors. We bilaterally filter the frame corner positions over 30 frames in time to reduce temporal jitter. We modulate the contribution of each filter window position (temporal weight) by the image-space Euclidean distance from the center window position (range weight).

The solution we provide to the alignment problem in our system is not the final word as this is a hard problem in its own. Please see our supplemental material for implementation details of our alignment approach.

Spatio-temporal Index

With video alignment, we can construct a spatio-temporal index of where and when each video intersects the context. We iterate through each video and intersect its per-frame bounds against a grid of cells which cover the panorama. The choice of grid resolution depends on the size of the dataset and memory constraints. We set the cell resolution of this index to be $\sim 100 \times 50$ cells, which gives a moderate 40 pixel spatial precision across the panoramic context. Each grid cell stores the spans of frames per video which intersect it.

The spatio-temporal index can be visualized in many ways depending on the application. We choose to render the index with a gradient such as a heat map (Fig. 4). With this, users can see which regions of the context held the most 'attention' among the videos, and our spatio-temporal interaction tools then allow these videos to be found quickly. Selecting individual videos shows a per-video index which defines the spatial extent of the video. Heat maps for specific index

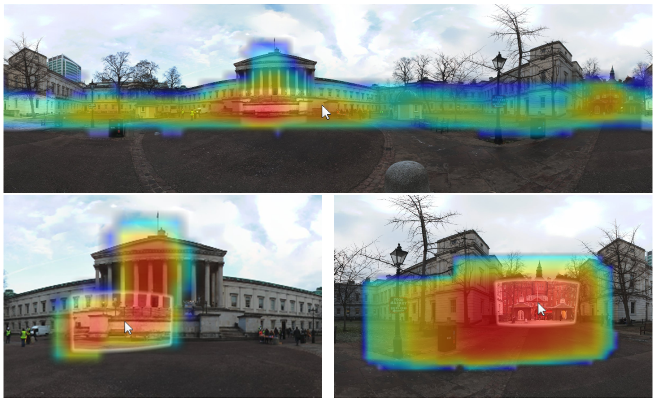

Figure 4: The spatio-temporal index displayed as a heat map to show attention over the context. This index allows quick spatio-temporal search and filtering of the video collection. This is computed globally (*top*) and per-video (*bottom*).

queries can be generated, e.g., video attention for a historic time span. Other visualizations would be possible, such as altering the saturation of the panoramic context locally for when it is important not to overlay further graphics onto the scene, or by displaying the path of the video by rendering a line joining the center-most grid cells along the path. Arrows on the line can show the progression of time, and color with a gradient can show where the video lingers.

Interface and Interaction

The Vidicontexts interface (Fig. 2) presents the context in either look-around perspective projection or as a full equirectangular map projection with an infinitely-rotating canvas. The user is free to pan, zoom, and smoothly switch between perspectives. Videos can be visually followed, or the context can be locked to follow individual videos. We also provide standard video playback controls. Our vidthieo-collection+context interface becomes more interesting as we provide tools for spatio-temporal interactions.

Temporally-driven Interactions

Each video has its own local timeline which appears when the video is selected. Unlike a normal video player, adjusting the timeline affects both the dynamic content within the video and the spatial position of the video in the context, and this provides a quick way to check the spatial extent of a video. This enables new applications: by adjusting the timelines of different videos and by setting A→B loop markers, the user can compose a novel arrangement of the videos within the context to highlight spatio-temporal changes.

As we have timestamps for each video, we also show a global timeline which displays the temporal extent of the video collection. Adjusting the ends of the timeline filters the video collection, for instance, to isolate videos from a particular day or month. The global timeline slider synchronously adjusts the playback of all videos in the collection, and allows the visualization of events which share the same time but otherwise have no visual overlap (Fig. 5). Such relations are difficult to explore when the videos are seen out of context. If multiple panoramas captured from the same position are available and timestamped, then the global slider also switches between them. This shows temporal changes in the context: for instance, in the seasons or in the built environment.

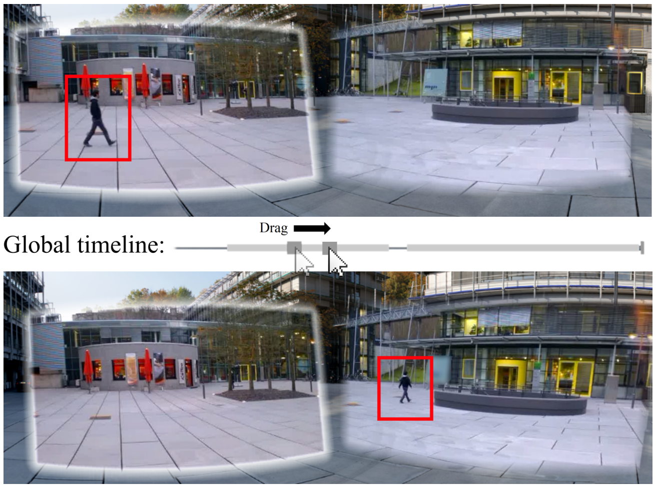

Figure 5: The global timeline allows control over all videos at once and enforces temporal ordering. Here, two contemporaneously captured videos which never intersect are displayed synchronously and in context, allowing easy identification of the movements of a man (red).

Spatially-driven Interactions

Temporal scrubbing has a spatial equivalent: By dragging the mouse over the panoramic context, the user can spatially drag individual videos or all videos at once, providing a fast way to localize many videos at once. As videos aren't guaranteed to visit all locations in space, they scrub to their nearest position. The extents of each video individually and of all videos combined can be shown by visualizing the spatio-temporal index (Fig. 4), and this helps guide spatial exploration.

We also provide area-based spatio-temporal filtering (Fig. 6). By dragging a box over the context to describe a region of interest, the user queries the spatio-temporal index for sequences of frames which intersect the region. This is a very fast way to 'collage' an area of the context with video.

EVALUATION

As this paper describes a time-varying interactive system, we refer the reader to our supplemental video for our interface demonstration. Thus, this section describes performance and timings for the preprocessing and viewer, then describes our quantitative and qualitative interface evaluation to gauge both the improvement over existing interfaces for tasks and the desirability and usability of our interface. For all results shown in both the paper and the supplemental video, we captured our own datasets to jointly capture orientation data.

Performance timings

We process videos independently and, as feature extraction is frame independent, our technique is embarrassingly parallel. Video alignment was computed on an Intel Xeon 8 core 2.40GHz PC; see Table 1 for computation times. All panoramas are 4000×2000 pixels, and all video frames are

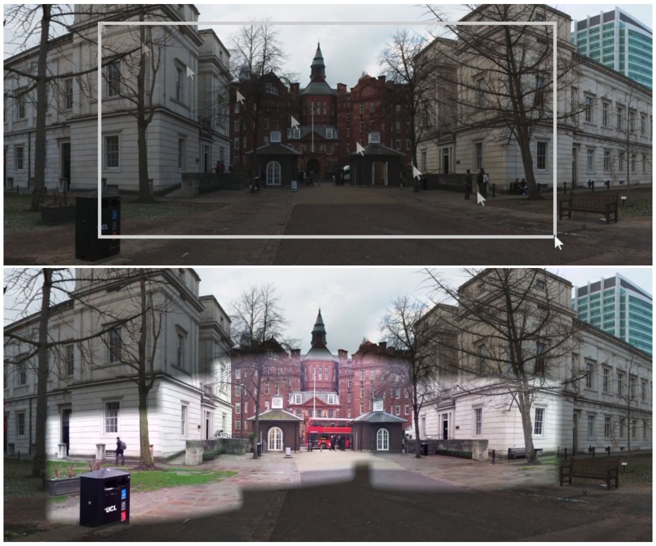

Figure 6: The spatio-temporal index makes searching simple. The user draws a bounding box over the region of interest, and our interface lays out all intersecting videos. The dynamic objects such as the red bus bring the scene to life.

Dataset	# Videos	# Frames	Alignment	Index
College grounds	15	30,426	6hr 10min	40sec
Castle vista	9	17,460	3hr 33min	25sec
New courtyard	11	21,518	4hr 26min	30sec
Neo-classical quad	20	26,635	6hr 16min	55sec
Indoor hallway	6	4,152	36min	16sec

Table 1: Computation times for alignment and spatio-temporal indexing (100×50 cells) for our datasets.

1920×1080 pixels. Our alignment code is written in MATLAB and C++, though GPU-accelerated matching algorithms may speed this up. The computation time for the spatio-temporal indices is also included in this table, and this performance scales linearly with the total number of cells.

The computational performance of our interface is defined by the number of videos visible. The rendering cost is minimal as we need only apply a homography to a pre-warped video and its feathered matte; however, the video decompression cost is large. Our implementation supports approximately 3 1080p HD videos at framerate at once. To cope with more videos, we store a reduced resolution version at a quarter scale, and only switch to full resolution if the user zooms in. While modern CPUs contain hardware to decompress 5+ videos at once, it is difficult to use this as our video format must support fast and exact seeking.

User Study

Design

Vidicontexts facilitates spatio-temporal exploration and comparison within video collections. While this is straightforward to understand and demonstrate, measuring whether our system provides significant benefits over existing video collection interfaces is non-trivial. Therefore, to evaluate Vidicontexts, we conduct a user study with two tasks that require participants to infer spatial and temporal information from a video collection. We compare Vidicontexts with *iMovie*, which offers a chronological browsing window and a resizeable timeline for fast preview, and against iMovie with the panoramic context image available for reference (*iMovie+pano* henceforth). Please see the supplemental document for further explanation of these two interfaces.

Tasks chosen to measure performance should represent general actions performed regularly by users. Common actions while exploring a place include looking for objects/actions in space and in time, following dynamic events within the place, and identifying when changes happen within specific times or areas of the place. As such, we select two tasks that involve counting and tracking events, in our case the comings and goings of people, within several videos. These tasks offer two reliable metrics which a) mimic common tasks performed when browsing a video collection, and b) can be extended to multiple system interfaces for comparison. In addition, we exclude possible tasks which would be trivial with one interface over another (e.g., in our interface, to find all videos which intersect part of the panorama). The resulting tasks are exemplars for real interactions which allow us to assess different systems and validate spatial and temporal understanding.

We wish to assess the accuracy with which participants can correctly obtain a spatial and temporal understanding of a collection of videos. Hence, in both tasks, we measured the completion time and accuracy expressed as errors in the people counts. Following the experiment, participants completed the standard System Usability Scale (SUS) questionnaire [4], which gathers subjective assessments of usability, as well as an additional questionnaire on the task experience (Tab. 3).

Tasks and Datasets

The *people counting* task requires participants to browse 20 videos from the *neo-classical quad* dataset (Fig. 4, 6) and identify the number of different people who sit on a set of benches. Videos differ in length and cover a large area of the environment. As different videos could depict the same person, or show a person sitting near the areas of interest, a participant could potentially make 20 erroneous counts. The maximum number of errors was manually counted.

The *people tracking* task asks participants to review 6 videos from the *new courtyard* dataset (Fig. 5) and track the number of different people who cross between two buildings. Here, the videos never fully track a person and do not overlap, so multiple synchronous videos must be analyzed to obtain the correct result. Videos differ in length, but they all cover a similar area of the environment. A participant could potentially make 12 erroneous counts (again manually counted).

Data Collection and Participants Selection

30 participants from the staff and student population at our university performed both tasks using one system each for a between-subjects design for the system independent condition, and a within-subjects design for the task. While we did not filter the study population for handedness and eyesight, we ensure gender balance was respected. Additionally, the participants were randomly assigned one of the three sys-

Condition	People Counting		People Tracking	
	Error	Time	Error	Time
iMovie vs. +pano	0.958	0.916	0.968	0.898
iMovie vs. Ours	0.040	0.017	0.049	0.014
+pano vs. Ours	0.107	0.023	0.012	0.005

Table 2: Significance (p-values) for each task and condition combination for both error and time to complete. Green values are statistically significant ($\alpha = 0.05$).

tems, within which the order of the two tasks was alternated to minimize the influence of learning effects. All subjects were introduced to their assigned system and to the tasks, and there was no mention of the overarching goal of the study. All participants were familiar with editing in general, and all received training with their system.

Procedure

Each participant performed two different tasks using the same system, with no time limit. Participants could use all features of each system, e.g., in iMovie and iMovie+pano, the built-in video scrubbing and thumbnail expansion. Each participant was given a detailed description of the system's interface and features, and as much time as they liked to familiarize before the task. Each task was conducted in series, with a briefing beforehand to explain the task. Following both tasks, the participant completed two questionnaires.

Hypotheses

We expect accuracy to vary with the sophistication of the spatio-temporal representation, and so we expect Vidicontexts to be more accurate. In turn, we expect iMovie+pano to be more accurate than iMovie alone. For completion time, we expect performance to vary according to the spatio-temporal controls available, and so we expect our system to require the least time. We expect all three conditions to score above average (75%) on the SUS. Finally, we expect Vidicontexts to obtain the highest score for the task-related questionnaire as we believe this is directly related to task performance.

Results

We provide a summary of our results here and in Figure 7, where box and whisker plots are reported for the completion time, the number of errors committed, and the significance for each task and condition combination; in the supplemental material we provide a full analysis description. We computed Analysis of Variance (ANOVA) using SPSS with the system used as the single factor and completion time/counting error as the dependent variable, with post-hoc Games-Howell tests for pairwise significance tests ($\alpha = 0.05$). Table 2 shows the significances of all compared systems. There was no significant difference between the iMovie and the iMovie+pano cases across all our experiments. When comparing iMovie against Vidicontexts, we see significant error reductions and significant time benefits for both counting and tracking tasks. One anomalous result is that our system is not significantly less eroneous than iMovie+pano in the counting case, though there are large differences between mean and std. dev. values.

Task-related question	iMovie	+pano	Ours
Easy to complete tasks	2.3	2.6	4
Understood video orientation in space	3.5	3.9	4.7
Understood relative video position	3	3.8	4.4
Understood space-time video overlap	2.8	3.8	4.3
Understood temporal order of videos	1.5	2.1	3.4
Environment representation confused	3.2	3.5	1.7
System has enough functions for tasks	3	2.5	4.4
#videos made remembering things hard	3.9	4.2	2.6
Overall mean	2.375	2.62	3.86

Table 3: System mean scores for the task-related questionnaire. The response scale varies between 1 and 5. The scale for negative questions was reversed for mean computation.

From the SUS questionnaire, only Vidicontexts scored above the average ($SUS = 77.5$), followed by the iMovie+pano ($SUS = 62.75$) and iMovie ($SUS = 59.5$) conditions. From this, our system can be classified as a Rank B system [17], whereas both iMovie and iMovie+panorama mode are Rank C systems. From the task-related questionnaire, Vidicontexts performed better than both iMovie and iMovie+pano conditions, scoring a significantly higher mean score of $M = 3.86$. Table 3 presents the mean score for each system and question.

DISCUSSION

User Study

For iMovie and iMovie+pano, user task strategy was to first expand the video thumbnails timeline to obtain an idea of where each video pointed, and then either to use the normal playback tools or to scrub through the videos as thumbnails. For the people tracking task, users played parts of the collection several times before answering. One user in both the iMovie and iMovie+pano conditions struggled to accomplish the task at all, and generally participants from these two conditions struggled more than participants from our condition. For Vidicontexts, most participants used the local video timelines to accelerate video localization in the panorama. The global timeline was frequently used by the participants in the tracking task, but rarely used for the counting task. No users struggled to complete the tasks with Vidicontexts.

In both tasks, Vidicontexts provided greater accuracy, and this agrees with our initial hypothesis. Our spatio-temporal representation combines necessary information to reduce task complexity over iMovie, and this is confirmed by significant error and time reductions. For instance, in the iMovie counting task, users need to spatially locate the video before counting people as only particular regions are of interest. In Vidicontexts, the user need only count people as the video is already spatially located. This reduction in complexity allows the user to perform only the task essential action.

Analysing the user task strategy confirms this explanation: In the counting task, for iMovie and iMovie+pano, users first expanded the video thumbnails timeline to spatially locate each video in turn, and then either used normal playback tools or scrubbed through the videos as thumbnails to count people.

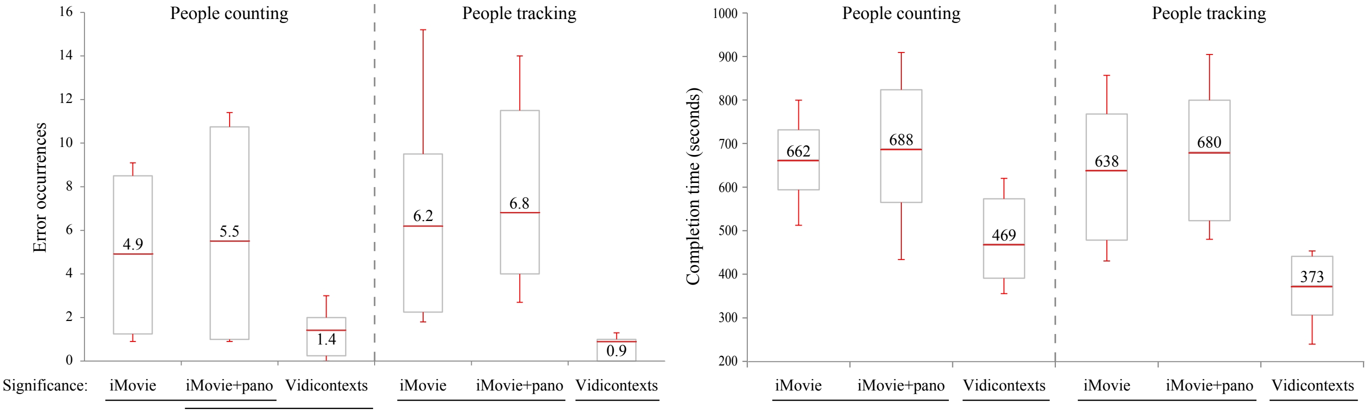

Figure 7: Box and whisker plots for each condition in both tasks. *Left:* Error occurrences. *Right:* Time to complete occurrences. First and third quartiles are reported in the upper and lower boxes, respectively. Any conditions whose name are underlined are considered statistically similar.

For Vidicontexts, participants could exploit the spatial alignment and only needed to search in time. Here, most used the local video timeline tool.

The people tracking task requires temporal and spatial alignment, and this increased cognitive load presented a challenge to users. In both iMovie conditions, users had to replay parts of the collection several times before answering. One user in both iMovie conditions struggled to accomplish the task at all, and generally participants from these two conditions struggled more than participants from our condition. For Vidicontexts, the global timeline maintains temporal alignment, and so this was frequently used by the participants in this task. No users struggled to complete the tasks with our system: the context representation combines necessary information and reduces task complexity.

Users were quick to familiarize with the new Vidicontexts system as video information is presented in a similar way to real life environments. This can help explain why users response to our system was also positive in terms of usability and desirability, as suggested by the much higher questionnaire scores for our interface than for the iMovie conditions. The individual questions reveal that participants considered our system the best tool to convey spatial and temporal information within the video collection, that they perceived our representation as less confusing, and that they thought our tools were more useful for exploration tasks. Additionally, participants agreed that, for tens of videos, our system improved recall; though we must not extrapolate this result to larger numbers of videos.

Finally, we observed a general trend in the preferred panorama projection. To complete the tasks, 80% of the population assigned to our system used equirectangular projection. This finding, in accordance with recent work by Mulloni et al. [20], shows that participants thought the 360°-at-once projection conferred more spatio-temporal information than the geometrically-correct perspective projection. We assume that users did not want to be constrained to a limited field-of-view for localization tasks.

Display applications

The video-collection+context representation fits display and interaction devices beyond desktop environments, such as tablets, spherical displays, and head-mounted displays such as Oculus Rift. These devices map the panorama to both virtual and real spatially-located spheres. Mobile devices naturally respect the geometry of 'inside→out' video collections, with a display that can respond to rotation and head movements (similar to [14]). With spherical displays, our context is displayed on a physical sphere in tandem with complementary joypad or tablet controller interfaces (Fig. 8, bottom row). Further, using contexts to join video collections has applications in augmented reality, especially when coupled with wearable computing devices such as Google Glass. Please see our supplemental material for detailed descriptions, and our supplemental video for demonstrations. Different displays provide different real and virtual geometries, and this might impact how users relate to and perform with the panoramic context. We leave this broad study for future work.

Use cases

Vidicontexts has many applications, including virtual tourism, surveillance, and shared browsing of an event. For example, sifting through footage and looking at patterns of behaviors over time within a place for a surveillance review; or, comparing and contrasting past events in a square, such as protesters contrasted against performers and musicians.

One particularly interesting example builds upon Naimark's Time Binoculars idea [21]: similar to the coin-operated binoculars common at vistas, a viewfinder allows users to "look around an actual site and see aligned augmentations of what they see, such as different times of day or different seasons, or historical views, or full-out Hollywood-style reenactments". Vidicontexts removes the need for fixed apparatus for these examples, and instead we imagine a smartphone or tablet app which guides users to spatial hotspots via GPS, from which they can use their mobile device as a viewfinder to explore the 'augmentations', or record new footage to add to the collection. For instance, a heritage site could create content especially for these experiences, such as re-enacted

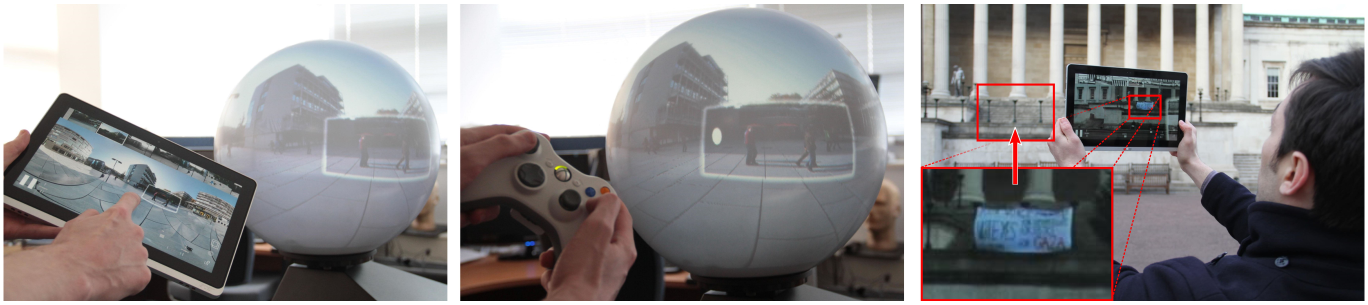

Figure 8: Additional displays and interactions. *Left:* Spherical display with a joypad controlling a cursor. *Center:* A tablet acting as a proxy controller, where the spherical display mirrors the context of the tablet. *Right:* Tablet display in situ, showing a protest that no longer exist in the real environment.

druidic rituals at Stonehenge; or a city could collate videos into a walking tour, say, London in 2012, re-viewing royal wedding, diamond jubilee, and Olympic events.

Limitations

While panoramas are available for many locations in the world, and simple tools on smartphones make panorama capture easy, Vidicontexts still requires a panorama as we register each video individually to the panorama. With only sensor orientation data, videos could still be coarsely aligned within an empty context, though existing videos rarely have embedded orientation data. Future work could explore stitching videos to each other to build a context. Further, even with a panorama, Vidicontexts will fail if large changes have occurred in the environment between the panorama and videos. For instance, many historical videos may only partially match the environment as building development is likely to have occurred. Here, we would have to rely on inter-video homography estimation for times in the video which do not match the panorama, anchored between times which do match. With no visual similarity at all, again we could only rely on captured orientation data.

In this work, we show panoramic contexts and, with in-situ browsing, we show real world contexts; however, other potential contexts exist. One alternative is 3D geometry from laser scans, but these are not readily available. Another alternative is 3D geometry reconstructed from images online. However, often these techniques are brittle with small baselines and are generally less applicable. In contrast, panoramas are widely available and easy to capture, and so our panoramic approach is applicable to more varied places.

Many errors can affect the quality of video alignment to the context, including failures and artifacts in panorama stitching, incorrect or badly synchronized sensor data and camera metadata, large deviations from the proxy geometry assumption, large dynamic objects, and large static changes. The problem of temporally consistent video alignment is difficult even for state-of-the-art vision systems, and improving this is important future work. However, we posit that this improvement would cause a relatively small functional improvement in our interface, and instead we try to show that a useful and wanted system is still possible under these conditions. Further, while orientation sensor data can be bad, it does provide a fallback for cases where visual alignment will have difficulty, and modern smartphones produce fittings from sensor data that are acceptable for many video-collection+context applications (we show one such case in our supplemental video).

With dynamic object segmentation masks, we could build a second spatio-temporal index for manipulating dynamic content directly in a similar way – scrubbing, filtering, and searching – and this would increase the functionality of our interface. Dynamic objects from different videos could be composited into the same context, and index look-ups would create synopses where space-time is compressed to be only those instances and positions where dynamics occur. As segmentation is a hard vision problem, existing work performs this manually for small numbers of short videos [25], but this approach is infeasible for video collections.

Providing context with panoramas is a special case of the larger problem of aligning videos to a 3D virtual world. As such, we impose strict limitations on the source data and so bypass many problems that come with more difficult data. Our examples and experiment do not use real data from community video websites, and many challenges remain to provide context for these varied collections. Our work demonstrates the promise of video-collection+context techniques in general, and produces a system with immediate benefits over existing video collection exploration software for limited subsets of videos. Overall, we feel that our approach is sufficiently practical for the described use cases in online mapping and tourism to be applicable immediately.

CONCLUSION

Vidicontexts is a system for interacting with an ‘inside→out’ video collection within a panoramic context. We capture panoramas and videos with handheld smartphones, and visually embed the videos into the panorama. With a spatio-temporal index of the scene, we create novel interface tools to quickly search and filter the video collection. We extend our system for tablet devices with physical rotation and zoom, in situ with real world contexts, and to spherical displays. We conduct a task-based study to compare with existing systems: we find that Vidicontexts provides significant benefits to accuracy and time taken in localization tasks, that it is preferred both by SUS and for our tasks, and that it would be used frequently and recommended to friends/colleagues if deployed.

Acknowledgements

Anthony Steed, Mike Terry, Pan Ye, Antti Oulasvirta, Kwang In Kim, Will Steptoe, Peter Grindrod, and all our participants.

REFERENCES

1. Agarwala, A., Zheng, K., Pal, C., Agrawala, M., Cohen, M., Curless, B., Salesin, D., and Szeliski, R. Panoramic Video Textures. *ACM Trans. Graph. (TOG) 24*, 3 (2005), 821–827.

2. Ballan, L., Brostow, G. J., Puwein, J., and Pollefeys, M. Unstructured Video-based Rendering: Interactive Exploration of Casually Captured Videos. *ACM Trans. Graph. 29*, 4 (2010).

3. Benosman, R., and Kang, S. B. *Panoramic Vision: Sensors, Theory, and Applications*. Springer, 2001.

4. Brooke, J. SUS: A Quick and Dirty Usability Scale. In *Usability Evaluation in Industry*. Taylor & Francis, 1996.

5. Brown, M., and Lowe, D. G. Automatic Panoramic Image Stitching using Invariant Features. *International Journal of Computer Vision 74*, 1 (2006).

6. Cockburn, A., Karlson, A., and Bederson, B. B. A Review of Overview+detail, Zooming, and Focus+context Interfaces. *ACM Comput. Surv. 41*, 1 (2009), 2:1–2:31.

7. Dale, K., Shechtman, E., Avidan, S., and Pfister, H. Multi-video browsing and summarization. In *CVPR Workshops* (2012).

8. de Haan, G., Scheuer, J., de Vries, R., and Post, F. H. Egocentric Navigation for Video Surveillance in 3D Virtual Environments. *2009 IEEE Symposium on 3D User Interfaces* (2009).

9. DeCamp, P., Shaw, G., Kubat, R., and Roy, D. An Immersive System for Browsing and Visualizing Surveillance Video. *Proc. MM '10* (2010).

10. Farin, D. S. *Automatic Video Segmentation employing Object/Camera Modeling Techniques*. PhD thesis, Technische Universiteit Eindhoven, 2005.

11. Hartley, R., and Zisserman, A. *Multiple View Geometry in Computer Vision*, 2 ed. Cambridge University Press, New York, NY, USA, 2003.

12. Hermans, C., Vanaken, C., Mertens, T., Van Reeth, F., and Bekaert, P. Augmented Panoramic Video. *Computer Graphics Forum 27*, 2 (2008).

13. Hu, J., You, S., and Neumann, U. Texture Painting from Video. *Journal of WSCG* (2005).

14. Joshi, N., Kar, A., and Cohen, M. Looking at You: Fused Gyro and Face Tracking for Viewing Large Imagery on Mobile Devices. In *Proc. SIGCHI '12*, ACM (New York, NY, USA, 2012), 2211–2220.

15. Kim, K., Oh, S., Lee, J., and Essa, I. Augmenting Aerial Earth Maps with Dynamic Information from Videos. *Virtual Reality* (2011), 1–16.

16. Kroepfl, M., Wexler, Y., and Ofek, E. Efficiently Locating Photographs in Many Panoramas. In *Proc. SIGSPATIAL '10*, ACM (New York, NY, USA, 2010).

17. Lewis, J., and Sauro, J. The Factor Structure of the System Usability Scale. In *Proc. Human Centered Design*, Springer (2009), 94–103.

18. Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision 60*, 2 (2004).

19. McCurdy, N. *RealityFlythrough: A System for Ubiquitous Video*. Ph.d., University of California, San Diego, 2007.

20. Mulloni, A., Seichter, H., Dünser, A., Baudisch, P., and Schmalstieg, D. 360 degrees: Panoramic Overviews for Location-Based Services. *Proc. SIGCHI '12*.

21. Naimark, M. Time Binoculars. `http://www.naimark.net/projects/pending/timebinoculars.htm`, 2010.

22. Neumann, U., You, S., and Hu, J. Augmented Virtual Environments (AVE): Dynamic Fusion of Imagery and 3D Models. *IEEE Virtual Reality '03* (2003), 3–9.

23. Norris, J., Schnädelbach, H., and Qiu, G. CamBlend: An Object Focused Collaboration Tool. In *Proc. SIGCHI '12*, ACM (New York, NY, USA, 2012).

24. Pece, F., Steptoe, W., Wanner, F., Julier, S., Weyrich, T., Kautz, J., and Steed, A. Panoinserts: Practical spatial teleconferencing. In *Proc. SIGCHI '13*, ACM (New York, NY, USA, 2013).

25. Pirk, S., Cohen, M. F., Deussen, O., Uyttendaele, M., and Kopf, J. Video Enhanced Gigapixel Panoramas. *SIGGRAPH Asia 2012 Technical Briefs* (2012).

26. Pongnumkul, S., Wang, J., and Cohen, M. Creating Map-based Storyboards for Browsing Tour Videos. *Proc. UIST '08* (2008).

27. Silva, J. R., Santos, T. T., and Morimoto, C. H. Automatic Camera Control in Virtual Environments augmented using Multiple Sparse Videos. *Computers & Graphics 35*, 2 (2011).

28. Snavely, N., Garg, R., Seitz, S. M., and Szeliski, R. Finding Paths Through the World's Photos. *ACM Trans. Graph. 27*, 3 (2008).

29. Snavely, N., Seitz, S., and Szeliski, R. Photo Tourism: Exploring Photo Collections in 3D. *ACM Trans. Graph. 25*, 3 (2006).

30. Szeliski, R. *Computer Vision: Algorithms and Applications*, 1st ed. Springer-Verlag, 2010.

31. Tompkin, J., Kim, K. I., Kautz, J., and Theobalt, C. Videoscapes: Exploring Sparse, Unstructured Video Collections. *ACM Trans. Graph. 31*, 4 (2012).

32. Wu, F., and Tory, M. PhotoScope: Visualizing Spatiotemporal Coverage of Photos for Construction Management. *Proc. SIGCHI '09*.