

# Immersive Trip Reports

Jan Brejcha<sup>1,2</sup>  
ibrejcha@fit.vutbr.cz

Michal Lukáč<sup>2</sup>  
lukac@adobe.com

Zhili Chen<sup>2</sup>  
zlchen@adobe.com

Stephen DiVerdi<sup>2</sup>  
diverdi@adobe.com

Martin Čadík<sup>1</sup>  
cadik@fit.vutbr.cz

<sup>1</sup>CPhoto@FIT,  
Faculty of Information Technology,  
Brno University of Technology,  
Czech Republic

<sup>2</sup>Adobe Research  
San Jose, CA, USA  
San Francisco, CA, USA

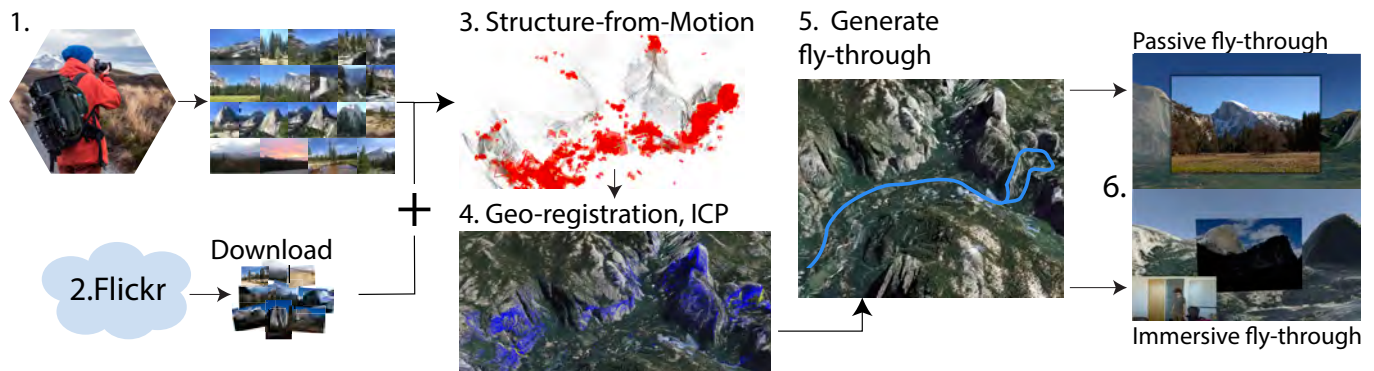


Figure 1. Our virtual trip creation pipeline: 1. User takes photographs during a hike; 2. We augment the input collection with images downloaded from Flickr.com; 3. Camera positions and sparse 3D point cloud reconstruction using *Structure from Motion*; 4. Scene alignment with the terrain using ICP; 5. Fly-through generation from the input photographs from the hike; 6. We export the fly-through to Google Earth or to our virtual reality viewer. Map data © 2018 Google, © Mapbox, © OpenStreetMap.

## ABSTRACT

Since the advent of consumer photography, tourists and hikers have made photo records of their trips to share later. Aside from being kept as memories, photo presentations such as slideshows are also shown to others who have not visited the location to try to convey the experience. However, a slideshow alone is limited in conveying the broader spatial context, and thus the feeling of presence in beautiful natural scenery is lost. We address this by presenting the photographs as part of an immersive experience. We introduce an automated pipeline for aligning photographs with a digital terrain model. From this geographic registration, we produce immersive presentations which are viewed either passively as a video, or interactively in virtual reality. Our experimental evaluation verifies that this new mode of presentation successfully conveys the spatial context of the scene and is enjoyable to users.

## Author Keywords

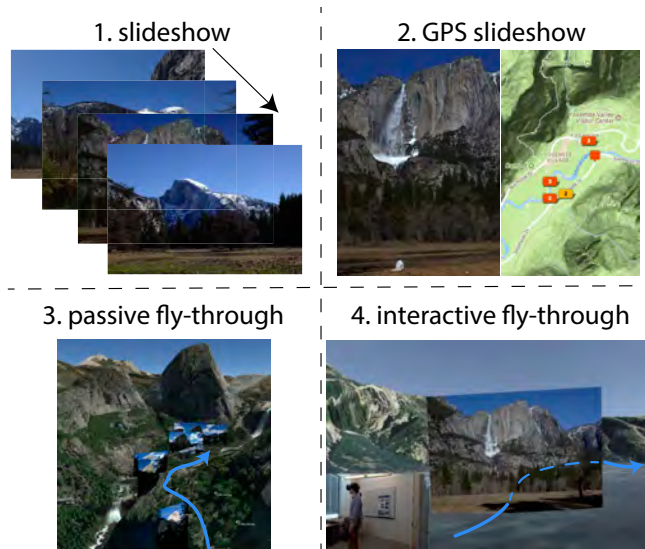
Photography presentation; hike; image geo-localization; terrain model; immersive visualization.

## INTRODUCTION

The human desire to travel is an affectation that goes back to before written history. So does, it would seem, the desire of travelers to share the experiences from their journeys. Travel literature is known to us since antiquity, and was a staple of medieval and early modern writing [44, 27, 5]. More recently as photography became widespread, it started to be widely used to record and share impressions from travels and vacations, indicating a desire to convey these experiences in a more engaging and immersive way.

Previous research has explored putting the photographs in a spatial context by manually registering them to a topographic map represented as a Digital Elevation Model (DEM) through tools such as PhotoOverlay in Google Earth [4]. Photo uncropping methods [37, 60] mine collections of external photographs for visual data with which to extend the field-of-view of the user's own photos. Structure-from-motion (SfM) methods register large collections of photographs of an artifact to create a 3D model, allowing a structured exploration of the photo collection [42, 43, 41, 22]. An extension of this work [50] uses accurate 3D models of urban environments to align the reconstructed scene and photographs with the physical geometry.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).  
UIST 2018, October 14–17, 2018, Berlin, Germany.  
Copyright © 2018 Association of Computing Machinery.  
ACM ISBN 978-1-4503-5948-1/18/10 ...\$15.00.  
<http://dx.doi.org/10.1145/3242587.3242653>



**Figure 2. Visualization of four modes of presentation. 1. slideshow:** photographs are presented sequentially. **2. GPS slideshow:** the slideshow with a map showing the position of currently shown photograph. **3. passive fly-through (ours):** photographs aligned with the terrain are presented in a passive fly-through. **4. interactive fly-through (ours):** the user can freely look around during the fly-through. Map data © 2018 Google, © Mapbox, © OpenStreetMap.

In this paper, we utilize recent advances in computer vision and virtual reality to increase the immersiveness of a photo presentation. Specifically, we have developed a process, illustrated in Figure 1, to extract 3D location and orientation information from collections of photographs taken on hiking trips, which we further use to align the photographs to a virtual representation of the actual terrain. We use this information to enrich the presentation with supplementary geographic data and replay the experience from a first-person perspective. We show that this pipeline works in general landscapes and requires only rough DEM data. By using the recovered information to automatically place the photos in the virtual terrain, we facilitate a rich first-person exploration experience which supplements the aesthetic and informational value of the photographs with contextualized spatial information.

The target audience of our method might be divided into two groups: (1) hikers who wish to share the experience of a hike, and (2) viewers who wish to learn more about hikes at locations they have not yet visited. Users from the second group who enjoy the presentation may then re-create the hike themselves. Therefore, the purpose of sharing travel photographs is not just to enjoy scenery, but to convey the entire experience of visiting the remote location.

Our goal is that our enhanced photo presentation will assist viewers to gain spatial orientation, better understand the scene, and enjoy the viewing experience. To evaluate these effects, we conduct a user study comparing four different modes of presentation (illustrated in Figure 2 and a video in the supplementary material) on four datasets from different locales. The tested modes consist of a traditional slideshow, a slideshow with GPS markers shown on a map (GPS slideshow), and two modes produced by our method. A fly-through from photo

to photo precisely aligned with a virtual terrain model was in one mode viewed passively as a rendered video (passive fly-through), and in the other interactively in virtual reality (interactive fly-through).

In summary, we present a new geo-registration pipeline for outdoor photo albums. We demonstrate this pipeline works in expansive natural outdoor environments, and for photographs captured with ordinary consumer hardware. We apply this geo-registration to automatically generate new modes of immersive first-person presentation for these albums; specifically a passive fly-through, renderable as video and compatible with tools like Google Earth, and an interactive fly-through which presents the trip in virtual reality. We also conducted experiments which demonstrate that these immersive presentation modes help user understand the spatial relations in the region significantly better than a traditional slideshow, and that the interactive VR experience is enjoyable.

## RELATED WORK

Motivating our approach, previous research has found that users' spatial understanding can be facilitated by incorporating animation [2], spatial context [53, 52], interaction [20], and panoramas [6]. We focus on work that exploits spatial information for *presentation*, *processing*, and *management* of photographs, and finally examine other related applications outside the photography domain.

## Photography Presentation

Previous research has explored alternative presentations of photographs. Chippendale et al. [4] summarized possible future applications of geo-localized photography like automatic creation of PhotoOverlays in Google Earth, or photographs augmented with peak names and GPS tracks. Snavely et al. explored presenting photographs in a 3D environment in their PhotoTourism paper [42], which uses SfM to reconstruct 3D point clouds for famous landmarks. They also designed a method for automatic path planning and photo exploration in the reconstructed environment [41]. Subsequent work uses similar techniques for automatic path planning [22] and effective photo acquisition of a site of interest [40]. Hyper-lapse videos [19, 51] yield a similar visual experience by smoothing and stitching egocentric videos. However, 3D point clouds used by PhotoTourism and others [22, 41, 42] are not suitable for visualization of a re-created trip. For example, in natural environments, usually only front facing parts of mountains are reconstructed leading to incomplete point clouds. Since a tour can traverse widely spaced viewpoints, the partial model reconstruction may result in poor visuals between photographs. Our method solves this problem by using the terrain model, which is more suitable for presentation of the whole trip.

Exploring spatially positioned photographs without 3D reconstruction has been proposed as well. Kaneva et al. [17] use image retrieval to find similar images, stitch them together and create a fictitious photorealistic virtual space. Tompkin et al. [46] combine videos with a panoramic image so the user better understands the mutual orientation and temporal relationship of different videos taken from the roughly same place. Veas et al. [49] studied spatial understanding and

navigation in outdoor environments, using video streams from several cameras. Video presentation in a 3D space has also been used to improve medical responses [23].

Visualization of images with geographical information is available commercially via online services such as Flickr and Google Maps. Researchers have explored visualizing photographs in a map online [48], or in virtual reality [25]. Geo-tagged social media enables spatial navigation interfaces for photo albums [47], even composited atop panoramas from Google Street View [10]. Note that these interfaces are not designed to convey a virtual hike experience. VR BBS [25] is for sharing photographs and messages in a virtual environment, with users plotting their own course through a flat map with 3D sprites of photographs. In contrast, our system leads the user automatically through virtual terrain containing the sequence of photographs of a re-created trip. Additionally, these previous works do not precisely align image content with the environment. This is a key ingredient of the seamless in-situ visualization implemented in our method.

The work most similar to ours is Kuchelmeister et al.'s [21] presentation of an immersive visualization of photographs taken by SenseCam jointly with a virtual model of a 3D outdoor scene. The intent of their work is to study the effect of browsing photographs in this virtual environment as a memory-prosthesis for patients suffering from amnesia. In contrast, our work does not use any specific device for collecting photographs, and our experiments are focused towards the orientation of users in the presented space and enjoyment of such a visualization.

In summary, previous methods are not designed to re-create a virtual hike experience. Specifically, we focus on the single-user-multiple-landmarks scenario, whereas PhotoTourism [42] addresses the multiple-user-single-landmark scenario. This has algorithm implications, so e.g. PhotoTourism and VR BBS [25] require much more elaborate capture processes. A key idea of our processing is to download additional imagery to help the reconstruction (see Figure 1), but use only user-generated photographs for the presentation.

### Photography Processing

Automatic immersive presentation of photographs requires precisely estimating camera parameters with respect to a world model. The most similar approach for photograph geo-registration is from Wang et al. [50], who use SfM for scene reconstruction, GPS positions of cameras for initial geo-localization and rigid fine-tuning of the scene with 3D building models using ICP. As they use vanishing points to estimate the reconstructed scene up vector, the method is limited to urban scenes with linear features. In contrast, we demonstrate that for natural environments, the GPS positions of Flickr images provide sufficiently precise initial geo-registration to enable further refinement with the terrain model.

Spatial awareness can be also improved by automatic uncropping of a photograph [37, 60]. These methods differ significantly from our approach in that they do not combine photographs with terrain models, though the concept of down-

loading many similar images from the same place for further processing is related.

### Photography Management and Categorization

Our immersive presentation is related to photo browsing and management systems. The rapidly growing number of photographs being taken has motivated research into effective searching [38] and clustering of photographs [28, 29], which can also be based on space and time [11, 56]. The difficulty of browsing, sorting, and clustering photographs manually has led to novel interfaces such as Photohelix [15]. Rodden and Wood [32] show that users tend to use simple features of photo management software, and also that managing photographs digitally is easier than managing printed photos. Harada et al. [12] designed an automatic searching and browsing tool for photographs on mobile devices. Schoeffmann et al. [33] show that photographs organized into a 3D cylinder or globe help users with faster visual search.

### Related Applications

Researchers have explored narrative storytelling with mobile photos [1] or photo blogs [18], or even writing fictional stories [31], as alternate ways of facilitating user engagement. Chelaramani et al. used photos of a historical site to create a multimedia tour guide for cultural heritage [3]. Other work has combined photos with animations [40] or 3D models [8] for cultural heritage as well. Immersive presentations such as virtual reality [14] and mobile augmented reality [13] have been found to improve appreciation of historical sites [45].

For productivity applications, PhotoScope [56] combines photo albums and building floor plans to aid construction management. Immersive presentations of many video feeds have been used to support video surveillance tasks, with desktop spatial navigation [7], desktop 3D environments [36], or full immersive virtual reality presentation [9]. Taken together, these related applications all support the notion that presenting photo albums of a remote location in virtual reality can improve users' engagement with the presentation and their resulting understanding of the experience.

### METHOD

The pipeline we use in our method is visually summarized in Figure 1; a more detailed flowchart of the pipeline can be found in the supplementary material in Figure 1. Our goal is to reconstruct from photographs a real hike in a virtual model of the real terrain. The input to our method is a collection of photographs taken on a hike  $I_h$ , together with the geo-rectangle designating their rough geographical extent, which can be read from embedded GPS information if available. We take the user photographs  $I_h$  as-is, we do not consider color enhancement as a part of our pipeline. We augment this collection with additional photographs from the same geo-extent  $I_f$ , which can be harvested from online repositories such as Flickr, to improve coverage of the terrain for better reconstruction. We jointly mine the merged photoset for both GPS metadata and visual features, which we use to obtain a rough geo-registration through a Structure from Motion (SfM) pipeline. We align the result of the reconstruction with known DEM terrain data to fine-tune the camera estimation. Finally, we construct a virtual



presentation that shows select photographs and renders fly-throughs from one camera pose to the next as a transition between the consecutive photographs.

### Imageset augmentation

We conducted initial experiments with datasets from the authors' personal collections in a variety of locations. Although these datasets were uncurated (i.e. contained all the photographs taken, including those that would not be selected for presentation), we found that a single user does not usually provide sufficient coverage of the space for a reliable 3D reconstruction. This may be tested by running the matching stage of the SfM reconstruction on the set of user photographs  $I_h$ . If the number of matching images with strong matches is low, we perform *imageset augmentation*. We augment each of the original collections  $I_h$  with images downloaded from Flickr  $I_f$ . This has the additional advantage that the original dataset  $I_h$  need not contain GPS information, since we may use GPS from the downloaded photographs  $I_f$ . However, in absence of any GPS information in user photos  $I_h$ , we need the user to provide the rough extent of the visited area, specified as e.g. center and radius.

We download these images through the Flickr API, querying for the specific geo-extent covering the area of the input photographs, which we additionally restrict to a specific time interval. This ensures that the downloaded photos are taken during roughly the same time of year, improving matching and reconstruction by eliminating seasonal changes.

Some of the images retrieved with a location filter may contain irrelevant data rather than natural outdoor scenes (e.g. indoor images, close-ups of vegetation or portraits of hikers, etc.). We filter them to improve efficiency of our algorithm. To select only relevant images, we apply a scene understanding neural network (ResNet18) trained on Places365 dataset [59] to find images that are most likely both *outdoor* and *natural*. Please note, that we perform this filtering only on downloaded images  $I_f$ , and keep user-generated photographs intact.

Given an input image, the network estimates matching scores for a list of semantic categories defined in the Places365 database. The semantic category is a high-level representation of a place, e.g., *bedroom*, *beach*, or *mountain*. For each semantic category, the Places365 dataset defines whether it is indoor or outdoor. Per-image, we select the semantic categories with the 10 highest scores; if majority of them are indoor, the image is classified as indoor and otherwise outdoor.

To implement the natural/unnatural classification, we use the image attributes from SUN attribute dataset [26]. Semantically overlapping image attributes describe scenes with fine granularity. We cluster the attributes as either natural (non-urban images) or unnatural (everything else). Examples of natural attributes include *foliage*, *leaves*, or *hiking*; examples of unnatural attributes are *pavement*, *carpet*, or *stressful*. The CNN estimates per-attribute correlations for an input image. We sum all correlations for natural attributes and subtract correlations for the unnatural attributes. If the outcome is greater than zero, then we classify the image as natural.

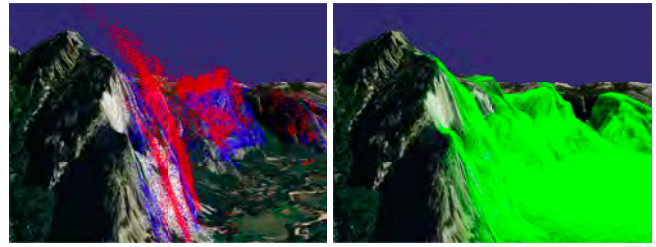


Figure 3. Alignment of input (red) point cloud with the reference (green) point cloud sampled from the terrain using Iterative Closest Points. The blue point cloud is the result. Map data © Mapbox, © OpenStreetMap.

### Scene reconstruction

Next, the mixed collection  $I_m = I_h \cup I_f$  of input hike photos  $I_h$  and Flickr photos  $I_f$  is fed into the reconstruction pipeline. We tested several publicly available Structure from Motion pipelines [55, 54, 24, 43, 34]. For reconstructing our mixed collections  $I_m$  we obtained the best results using the publicly available COLMAP implementation [34]. We found it important to use approximate matching with a vocabulary tree and an enhanced voting strategy for fast spatial verification [35], since exhaustive matching is significantly slower. We use a 256k vocabulary tree provided by the COLMAP authors<sup>1</sup>. Typical reconstruction time of a dataset of 4k photographs was several hours on a desktop PC with NVIDIA 970 GTX GPU.

For geo-registration using GPS from Flickr images, we use a robust Least Median of Squares (LMeds) combined with RANSAC [58] using euclidean distance of the reconstructed camera position and the corresponding GPS position (residual). Instead of minimizing the sum of squared residuals, we minimize their median, which is more robust towards outliers. Using this minimization approach, we estimate a similarity transformation to transform (translate, rotate, and scale) the scene into world coordinates.

### Fine-tuning

Because of uncertainties in camera configuration, GPS location, and other parameters, there is no guarantee that the initial geo-registration actually matches the known terrain. To remedy this, we refine the initial geo-registration by minimizing the euclidean distance between the reconstructed 3D point cloud and the known DEM terrain data. We segment the point cloud into disjoint clusters so that two points in the same cluster are at most 1 km apart from each other. For each cluster, we calculate its bounding box and sample the terrain on a grid with 10 m spacing. We align the reconstructed 3D point cloud and the sampled terrain using Iterative Closest Points (ICP) with the Libpointmatcher library [30] with default parameters. The algorithm first reduces the input and reference point clouds (see Figure 3) by random sampling, keeping 75% of all points. Next, the algorithm iteratively performs a series of steps: 1. each point is matched to its nearest neighbors in euclidean space; 2. points too far from the reference point cloud (outliers) are removed (85% of points with smallest distance are kept); 3. minimization of point-to-plane distance is performed [57]; 4. check if convergence or the maximum number of iterations (40) has been reached.

<sup>1</sup><https://demuc.de/colmap/>

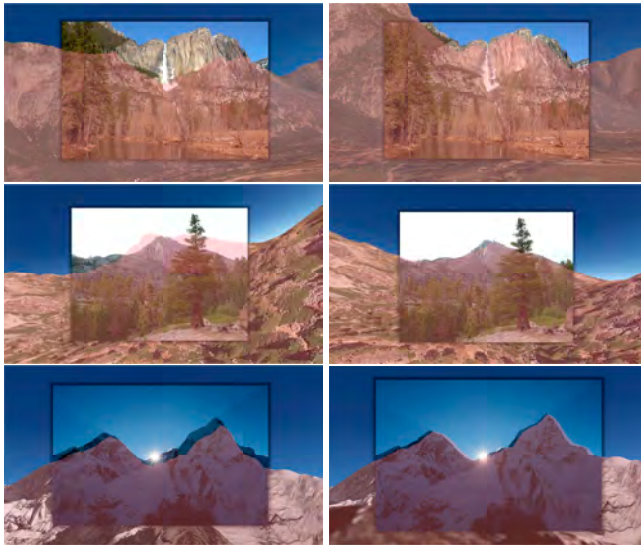


Figure 4. Examples of images before (left) and after point-cloud to terrain alignment using ICP (right). Top row: Yosemite Waterfall, CA, USA, middle row: Jakes Peak at the Lake Tahoe, CA, USA, bottom row: Mount Everest, Nepal. Map data © 2018 Google.

After registering the model, we are often left with mismatches between the photo content and the virtual terrain, most of which are due to bad information about camera configuration (e.g. focal length, exact GPS position, etc.). Furthermore, because of the limited sampling rate of the DEM, some cameras may end up below the virtual terrain after the ICP alignment, which we solve by moving them vertically to the terrain height. However, both of these problems introduce errors in camera orientation parameters.

To correct the registration errors, we leverage our knowledge of the correspondences between 2D points  $o_i$  observed in the photographs and the 3D points  $p_i$  in the virtual terrain. We use these correspondences to optimize the orientation parameters using the Kabsch Algorithm [16]. We project the 2D observations  $o_i$  using camera parameters into 3D points  $\bar{p}_i$  based on the euclidean distance between camera center and the corresponding 3D point. From both sets we subtract their centroids and calculate the rotation matrix using the Kabsch algorithm  $R = K(\bar{p}_i, p_i)$ . The results of the fine-tuning are shown in Figure 4. Table 1 illustrates the matching accuracy of the reconstructed 3D point cloud with the sampled terrain. Because a reconstructed model usually contains a small number of outliers, we report the median euclidean distance between each 3D point  $p_i$  and its closest point on the sampled terrain. To illustrate the accuracy of the reconstruction, we also include the mean reprojection root mean squared error (RMSE) across all cameras in given dataset.

### Fly-through creation

For the fly-through presentation, the user selects a curated subset of photographs  $I_c \subseteq I_h$  based on their aesthetic preference. Although we know the camera pose for each photograph from the registration, we still need to estimate the actual hiking path taken from one camera position to the next. We generate a smooth camera path by constructing a Catmull-Rom spline

dataset	Me $D(p_i)$ [m]	Me $D(p_i)$ -ICP [m]	$\mu r_e$ [px]
Nepal	1624.62	819.98	0.41
Tahoe	2814.24	72.82	0.88
Tatras	2908.59	2410.21	0.47
Yosemite	14041.70	348.33	0.50

Table 1. Median alignment error (Me  $D(p_i)$ ) of the point cloud and the terrain before and after ICP, and mean RMSE of the reprojection ( $\mu r_e$ ). The median alignment error Me  $D(p_i)$  is significantly lower after alignment using ICP.

with the camera positions from  $I_h$  as control points. Alternately, if a full GPS track is available, it may be used as the camera path instead to ensure that the presentation follows the trail between photographs. Please note, that the selection of the curated subset of photographs  $I_c$  affects only *which* photographs will be presented; the reconstructed path is the same for different subsets of  $I_c$ .

We initialize the set of control points  $P_c$  with the positions of the curated photographs  $I_c$ . We add the remaining positions from the reconstructed photos  $I_h$  in a greedy way—a point is added only if it is further than 100 m from all points in  $P_c$ . The control points  $P_c$  are sorted according to the time of capture of the corresponding photograph parsed from EXIF. The Catmull-Rom spline is generated from the selected control points  $P_c$ . In case any point of the spline is located below the terrain, we project it above the terrain height by a fixed margin. We smooth the generated spline using a low-pass box filter.

A part of the spline between consecutive control points is called a *segment*. In passive mode, as the camera moves along a segment, we smoothly interpolate camera parameters. Field-of-view is interpolated linearly between photographs of consecutive control points; the camera orientation is interpolated to look in the direction of the next control point. For transitions from one photo to the next, we use spherical interpolation between the two orientations, with the camera located at the center of the sphere to achieve near-constant angular speed. The speed of the camera is calculated automatically—for more distant control points the camera flies faster, accelerating and decelerating at the start/end of the segment, respectively. In interactive case, the field-of-view and orientation are defined by the output device (e.g. the headset), and the speed of the flight is controlled directly by the user. Also, in interactive mode the user can move in a small neighborhood of the current position on the spline. In passive mode, the position of the camera is restricted to the generated spline.

To generate the actual presentation, we combine the fly-through with the photographs rendered with appropriate camera parameters over the virtual landscape. In the passive case, we cross-fade from the end of a fly-through segment to the photo we wish to display and then cross-fade to the next segment. In the interactive case, the cross fade for leaving the photograph is triggered by the user. In both cases, accurate estimation of camera parameters ensures the transitions are smooth.

dataset	$I_h$	$I_f$	$I_m$	$I_{hr}$	$I_{mr}$
Nepal	1586	815	2401	412	901
Tahoe	302 (36)	0	302	78 (7)	78 (7)
Tatras	0	4146	4146	297	297
Yosemite	543 (117)	4173	4716	167 (33)	2094

**Table 2. Number of photographs in our datasets.**  $I_h$  – input hike photographs captured by user,  $I_f$  – number of downloaded Flickr images,  $I_m$  – number of mixed photographs entering the reconstruction,  $I_{hr}$  – number of hike photographs that were successfully reconstructed,  $I_{mr}$  – number of all reconstructed photographs. Panoramic images are included and denoted by numbers in brackets.

## EXPERIMENTS

The goal of our method is to create an enjoyable presentation which helps the viewer understand the physical layout of the place where the photographs were taken. We conducted a user study that compares four modes of presentation of photographs; two traditional, and two based on our method. We evaluate these methods on viewer enjoyment, sense of presence, and a quantitative task that measures how well the user can localize previously unseen photos from the same space after viewing the presentation. First, measuring enjoyment is important to understand if users want to use our method. Second, we measure the sense of presence to determine how immersed users become on a virtual hiking trip. Third, we assess users’ orienteering capability conditioned on the presentation method to determine if our method measurably impacts users’ spatial understanding of the environment. We use four datasets processed with our pipeline, and from each dataset we select one subset of photographs for presentation and a disjoint subset for evaluation.

### Datasets

Out of the four datasets we used in our experiment, three were captured manually on location at Lake Tahoe, CA, USA, Yosemite Valley, CA, USA, and the Himalaya mountains in Sagarmatha National Park, Nepal. The fourth dataset from the High Tatra mountains in Slovakia was collected from Flickr. Each dataset was captured by a different photographer. The Lake Tahoe dataset was reconstructed directly without any auxiliary photographs, while Yosemite and Nepal were augmented using Flickr images. The statistics on the number of captured photographs  $I_h$ , photographs downloaded from Flickr  $I_f$ , as well as successfully registered user photographs  $I_{hr}$ , and total successfully registered photographs  $I_{mr}$  are shown in Table 2. All four datasets were processed with our geo-registration pipeline and exported to Google Earth through KML for the passive mode, and to our implementation of a VR viewer in Unity with the terrain loaded from Mapbox for the interactive mode.

### Modes and Setup

We compared four modes of presentation, shown in Figure 2 and the supplementary video. The baseline mode *slideshow* is a standard photo slideshow without any additional information. The second mode *GPS slideshow* is a slideshow with camera positions marked on a map as presented in Adobe Lightroom. For each photograph, the user can explore a Google “terrain” map with contour lines in fixed zoom level, where all the photographs in the presentation are localized and the current

dataset	positional error	heading error
Tahoe discr.	0/6	0/6
Tahoe cont.	$353.61 \pm 230.29\text{m}$	$32.05 \pm 28.39^\circ$
Yosemite discr.	0/4	3/4
Yosemite cont.	$1189.33 \pm 748.61\text{m}$	$23.81 \pm 20.44^\circ$
Nepal cont.	$4710.74 \pm 2833.38\text{m}$	$75.14 \pm 53.34^\circ$

**Table 3. Pilot study data.** For discrete version the numbers denote a fraction of wrong answers. For continuous measurements the mean and standard deviation is reported.

one is highlighted. The third mode *passive fly-through* is the passive version of our method: a fly-through in Google Earth generated by our geo-registration pipeline. The user is first shown the path of the tour in a top-down view. The view then transitions to the camera position and orientation of the first photograph, with the photograph drawn over the terrain. As the user presses a button, the view flies to the next camera position and shows the next photograph in the same fashion. Once the fly-through is finished, the presentation returns to the initial top-down view. The final mode *interactive fly-through* is the interactive version of our method, with the fly-through presented in VR. The user is first given an opportunity to familiarize themselves with the region’s terrain from a bird’s-eye view several kilometers up. They are next teleported to the fly-through, which proceeds in a similar fashion to the passive mode, except the user has the opportunity to look around freely and is able to control the speed of movement along the camera path in order to reduce risk of motion sickness.

In all modes, the user sees each photo only once without the option to go back. All modes and datasets were presented on a calibrated<sup>2</sup> 15" MacBook Pro Retina display in native resolution  $2880 \times 1800$  pixels under office lighting, with the exception of the *interactive fly-through* mode, which was presented using an HTC Vive. Each participant tested all four modes, each with a different dataset to avoid learning effects. The mode-dataset pairing and the mode order were randomized for each participant.

### Pilot study

To help design the main study, we performed an initial experiment with one participant. The female participant was a co-author of the *Tahoe* dataset and familiar with the terrain in the *Yosemite* dataset, with extensive experience in using maps for navigation. The purpose of this test was to determine whether the task is better evaluated using discrete or continuous questions. The participant was first shown a presentation of at most 20 photographs and afterwards was asked to complete a task with a selection of photographs from the same dataset, but disjoint from that shown in the presentation.

In the discrete scenario, the participant answered binary questions about camera heading and position. For position, she was shown a query photograph taken chronologically between two consecutive photographs from the presentation, and asked to identify whether the viewpoint of the novel photographs

<sup>2</sup>The calibration was performed by X-Rite GretagMacbeth Eye-One Display colorimeter to D65, 120 cd/m<sup>2</sup>, and colorimetrically characterized by measured ICC profiles.



is closer to that of the earlier photograph or that of the later photograph. For heading, she was shown a query photograph and a reference photograph from the presentation and asked to identify whether the query photograph camera orientation is to the left or to the right of that of the reference photograph.

In the continuous scenario, the participant was asked to mark two points in an online map for each photograph. The first corresponds to the camera viewpoint of the query photograph. For the second, the participant could pick an arbitrary reference point in the query photograph and then select a point on the map that corresponds to the location marked in the photograph (see Figure 5).

Initially, we tested the *Tahoe* dataset in *slideshow* mode, and *Yosemite* dataset in *passive fly-through* with both discrete and continuous variants. For each variant, we tested 4–6 different photographs. Since the participant visited both areas earlier, we added a test on the *Nepal* dataset in *passive fly-through* with a continuous variant. The results are shown in Table 3. The discrete and continuous variants are consistent on the *Yosemite* dataset; the participant is able to estimate heading more accurately than position for both task sets on this dataset. Conversely, even though the participant achieved perfect success on the discrete heading task for both *Tahoe* and *Yosemite*, the continuous heading error is higher on the former. The continuous errors are notably higher on the *Nepal* dataset, suggesting there is a significant difference in difficulty between datasets, possibly related to spatial extent and complexity of terrain. The participant expressed preference for the continuous tasks, describing them as an interesting puzzle, as opposed to the discrete tasks which she tended to answer randomly when in doubt. Another issue in the discrete task is that when the rotation is close to  $180^\circ$  with respect to the reference, it is extremely difficult for the participant to answer correctly as the difference between “left” and “right” is only a few degrees. Based on these observations we selected the continuous task set as the evaluation method for the full user study. We expected it to give us more information with less variance even with a small number of participants, which was limited by the long duration of each test (up to 1 hour for all four modes with each participant). We also expected the continuous task set to be more engaging for the users and thus keep them more focused. Finally, we realized the necessity of normalizing errors per-dataset due to high observed variation in dataset difficulty.

We also performed a field-type experiment where we presented photographs from the *Nepal* dataset in the *slideshow* and the *passive fly-through* presentation modes to a broader audience of approximately 40 people. After the presentation, the audience completed a short questionnaire asking which of the two methods they preferred more and whether the terrain model helped them better understand the positions and orientations of the photographs compared to the slideshow. Out of 40 participants, 22 completed the questionnaire. Regarding the first question, 8 participants replied that they liked the fly-through more than the slideshow, 10 participants liked both roughly the same, and 4 participants liked the slideshow mode more. Responses for the second question were even more pos-



Figure 5. Example task from Lake Tahoe. The participant marks a position on the map (right image #1) of the query photo (left image) and the reference point (left image, red star) and corresponding position of the reference point on the map (right image #2). Map data © Mapy.cz.

itive: 14 participants agreed that the fly-through helped them, 2 participants replied that both modes helped them roughly the same and 4 participants replied that the slideshow helped them more. A final question asked participants to write what they liked or disliked. Participants disliked the abrupt speed of camera rotations during transitions in the fly-through. We identified this as the main reason why 14 out of 22 participants preferred the slideshow or had no preference in the first question. Due to this finding, we adjusted the angular velocity to ensure smooth camera rotations. Furthermore, subsequent experiments were designed based on the experience from this field experiment.

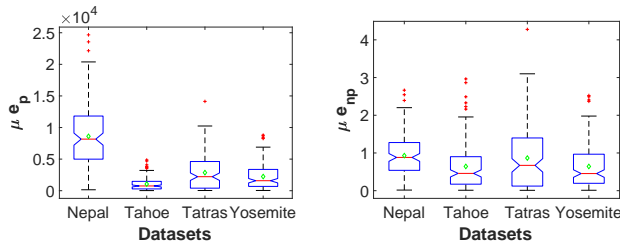
### Evaluation Methodology

Each participant was instructed about the purpose of the experiment and completed a screener questionnaire. Before the first experiment, we explained the task with a dummy example. The procedure was as follows: a presentation of at most 20 photographs was shown to the participant. The participant viewed one picture at a time as determined by the presentation mode. The participant was not allowed to return to previously viewed photographs. After the presentation, the participant viewed 6–7 photographs not present in the presentation, but taken on the same dataset. For each photograph, we performed the continuous task variant, as determined by the pilot study, in which the participant indicated the position and heading of the camera by marking a map. The participant was not allowed to move already placed marks once they continued to the next photo, or to return to a previous photo during the test. The participant was allowed to zoom in to the online map during the task, as well as to move around within the area of the dataset. If they moved out of the area, the moderator would reset the map to the initial view. The initial zoom level was chosen so that the area of the whole dataset would fit inside the window. The digital map featured a top-down view with only the names of points of interest (POI), tourist pathways and contour lines showing elevation.

### User Study

#### Participants

We assembled 21 volunteers, predominantly bachelor and master students of informatics (17) and law (4); 3 women and 18 men. One participant had been to Lake Tahoe, 4 to Yosemite, 13 to High Tatra Mountains, and 1 to Nepal. 14 participants had some experience with virtual reality. Each participant had at least basic knowledge on how to use a map: one



**Figure 6. Differences between datasets before normalization of mean positional error  $\mu e_p$  (left) and after normalization  $\mu e_{np}$  (right). The central red mark indicates the median, the green diamond denotes the mean, the bottom and top edges of the box indicate the 25th and 75th percentiles, respectively. The whiskers extend to the most extreme positional errors not considered outliers, and the outliers are plotted individually using the ‘+’ symbol**

participant used maps several times in his life, 5 participants used maps at least once a year, 10 participants used maps at least once a month, and 5 used maps at least once a week. Where possible, we correct our experimental data for the bias introduced by these factors.

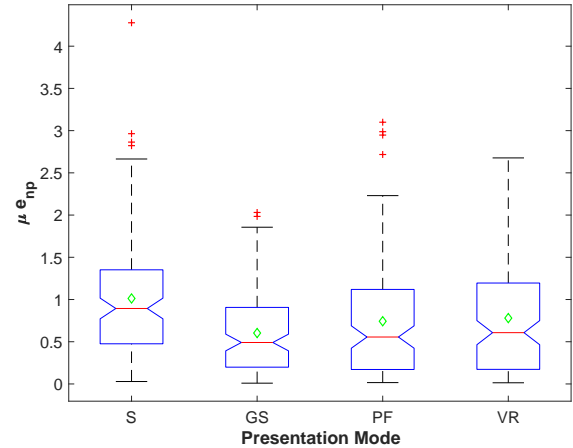
*Error measures*

We report two error measures per test photograph: the positional error  $e_p$ , and the heading error  $e_h$ . Both errors are calculated based on ground truth photograph position  $p_g$  and orientation  $h_g$ . The positional error  $e_p$  is the euclidean distance between the ground truth  $p_g$  and measured  $p_m$  position:  $e_p = ||p_g - p_m||$ . The heading error is the smallest absolute difference between the ground truth heading  $h_g$  and the measured heading  $h_m$  (in degrees):  $e_h = \min(|h_g - h_m|, 360 - |h_g - h_m|)$ .

*Positional Error Model*

We use different datasets for each test to avoid learning effects, but this introduces the possibility that performance may be correlated with dataset difficulty. To compensate for dataset and user differences, we model the positional error as a normal random variable  $e_p \sim \mathcal{N}(sdm, \sigma^2)$ , where  $s$  is a factor of the subject’s ability,  $d$  is a factor of the dataset difficulty,  $m$  is factor of the mode properties, and  $\sigma^2$  models measurement noise. Since we want to compare modes based on the positional error, we need to mitigate the effects of dataset factor  $d$  and subject’s ability factor  $s$ .

We expect that the *Nepal* and *Tatra* are more difficult than the *Tahoe* and *Yosemite* because the trips made in the *Nepal* and *Tatra* are much longer and the terrain profile is more complicated. Figure 6(left) confirms this, but the positional error  $e_p$  has different scale for each dataset due to different geographic extents. One-way ANOVA clearly rejected the null hypothesis ( $F(3, 542) = 149.85, p < 0.001$ ), that the means of positional errors  $e_p$  do not vary significantly across datasets. Further inspection reveals that the *Nepal* dataset has significantly higher positional error than other datasets across all methods. We attempted to normalize the errors by dividing it by the dataset extent. This moved the scale between datasets closer, but the null hypothesis was still clearly rejected ( $F(3, 542) = 10.85, p < 0.001$ ). In this case, the *Tahoe* dataset was shown to have significantly lower mean error than



**Figure 7. Repeated measures scenario comparing differences between normalized positional error  $e_{np}$  on different modes of presentation (S = *slideshow*, GS = *GPS slideshow*, PF = *passive fly-through*, VR = *immersive fly-through*). The mean value for each method is denoted by green diamond.**

other datasets. Instead, we use the baseline mode *slideshow* as a dataset calibration measure. We calculate the normalized positional error  $e_{np}(d)$  for each dataset  $d$  by dividing by the mean of the positional error  $e_p(d, m_s)$  for the *slideshow* mode  $m_s$  and the dataset  $d$ :

$$e_{np}(d) = \frac{N_d e_p(d)}{\sum e_p(d, m_s)}, \tag{1}$$

where  $N_d$  is the number of measurements for dataset  $d$ . This yields the lowest  $F$ -score compared to other normalization methods ( $F(3, 542) = 6.98, p = 0.0001$ ). The null hypothesis is still rejected, due to the fact that the baseline *slideshow* mode has been tested by different users on different datasets. However, the error distributions have almost the same scale, and the result still matches our initial expectations: the *Tahoe* and *Yosemite* datasets exhibit lower mean error than the *Nepal* and *Tatra* (see Figure 6 right).

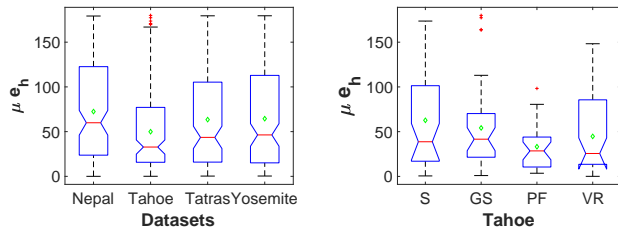
*Subject’s ability factor*

We tested the per-subject mean differences using one-way ANOVA. The test was unable to reject the null hypothesis that the means of positional error  $e_p$  do not vary significantly between users ( $F(20, 525) = 1.13, p = 0.31$ ), which also hold for the normalized positional error  $e_{np}$  ( $F(20, 525) = 1.75, p = 0.23$ ). We further inspected the importance of factors that the subject visited the place before, map proficiency, and map usage frequency. None of them showed significant effect on positional or heading error. In summary, we were not able to prove any significant differences between users in terms of positional and heading error.

*Position Evaluation*

Having normalized results for dataset difficulty, we can formulate the comparison of presentation modes as one-way repeated measures ANOVA with the presentation mode as a within-subject variable with four conditions. This way, the test can account for performance differences between subjects. As the





**Figure 8.** Left: comparison of dataset difficulty with respect to heading error  $e_h$ . Right: comparison of heading errors achieved by presentation modes (S = *slideshow*, GS = *GPS slideshow*, PF = *passive fly-through*, VR = *immersive fly-through*) on the easiest *Tahoe* dataset.

numbers of photographs differ between datasets, we first calculate mean per-subject and method. This way, we have one measurement per subject and method. We formulate the null hypothesis that means of normalized positional error  $e_{np}$  do not differ significantly between methods. The null hypothesis was clearly rejected ( $F(3, 375) = 8.13, p < 0.001$ ). Post-hoc analysis reveals that the baseline presentation mode has significantly larger mean normalized positional error  $e_{np}$ , than *GPS slideshow* ( $p < 0.001$ ), *interactive fly-through* ( $p = 0.034$ ), and *passive fly-through* modes ( $p = 0.009$ , see Figure 7). There is no significant difference between the *GPS slideshow*, *passive fly-through* and *interactive fly-through* according to our data and this test ( $p \geq 0.434$ ) for all remaining combinations). In summary, it seems the positional information contained in *GPS slideshow*, *interactive fly-through*, and *passive fly-through* modes helps users with location estimation.

#### Heading Evaluation

We were not able to find any significant differences between presentation modes for heading error  $e_h$ . Significant difference was found between the *Tahoe* and *Nepal* datasets using one-way ANOVA ( $F(3, 542) = 4.23, p = 0.0057$ ), supporting our expectation that the *Tahoe* dataset is easier than *Nepal* (and according to Figure 8 left probably the easiest among all datasets). Our data suggest it is fairly difficult to understand what the camera is looking at in a photo and then mark it on a map. The only dataset where the orientation exhibits some tendency is the easiest *Tahoe* dataset, where the *passive fly-through* has the lowest mean heading error and *interactive fly-through* has second lowest (see Figure 8 right), however, these differences are not statistically significant. Other datasets seemed to be too difficult for heading estimation as all the methods exhibited similar variance and mean across the remaining datasets, probably due to large random error. In summary, on the easiest *Tahoe* dataset, the *passive fly-through* and the *interactive fly-through* seem to have marginally lower orientation error than remaining two modes of presentation.

#### Presence Evaluation

We included a presence questionnaire to evaluate how successfully the user is immersed by each presentation mode. To reduce the time of the experiment, we tested just two modes of presentation—the *GPS slideshow*, and the *interactive fly-through* on a randomly selected half of our participants. For this evaluation, we use the SUS presence questionnaire [39], because of its relative compactness. As a first measure, we

calculate number of high responses (6, 7) for each presentation mode—7 for *interactive fly-through*, and 6 for *GPS slideshow* (higher is better). We also calculate mean and standard deviation of scores for both methods: the *interactive fly-through* is  $3.94 \pm 1.40$ , and *GPS slideshow* is  $3.29 \pm 1.68$ . We can see that the *interactive fly-through* is better than the *GPS slideshow*, however, one-way ANOVA does not find significance. In summary, the *interactive fly-through* seems to exhibit slightly better scores in terms of presence compared to the *GPS slideshow*.

In the post-test questionnaire, we asked users whether they think that the terrain model (*passive fly-through* or *interactive fly-through* modes) helped them to create a better idea about the area of the dataset. The terrain model was helpful for 7 participants, 8 participants thought that the terrain model helped them roughly the same as the *GPS slideshow*, and 6 participants replied that the *GPS slideshow* helped them more.

#### Enjoyment

We asked users to identify which method was the most enjoyable. 17 participants preferred the *interactive fly-through* the most. They liked being able to look in the direction they were interested, and that they could control the speed of flight using the controller. Two participants preferred *passive fly-through* the most. The reason was that the VR did not suite their taste and they felt a little bit disorientated after the task in VR, but they liked the possibility of seeing the pictures aligned in the virtual terrain model. Two respondents preferred *GPS slideshow* the most, since they felt it has been the most helpful to fulfill their task. In summary, the *interactive fly-through* is the most enjoyable mode of presentation according to our evaluation.

#### Discussion

We have measured the subjects' ability to estimate camera position and orientation of a previously unseen photograph based on what they learned from the presentation. We further evaluated the subjects' enjoyment of different presentation modes and the sense of presence they confer. The use of four datasets of different difficulty posed a challenge in the evaluation, since we needed to normalize positional errors in order to compare the differences between the presentation modes.

The results suggest that the *GPS slideshow* is likely the best mode for the position estimation task. We suspect that this is because the mode of presentation—markers on a map—is so close to the evaluation task that the effect of recall may dominate that of the actual sense of spatial orientation. The use of the same modality then leads to marginally better results over *passive fly-through* and *interactive fly-through*.

According to our measurements, it seems that the length of the fly-through and terrain complexity affect the learning effect of the *interactive fly-through*. For a short and easy trip, such as in the *Tahoe* dataset, the *interactive fly-through* scored slightly better than *GPS slideshow* in terms of position and heading, but on more complicated datasets, such as the *Nepal*, the

*GPS slideshow* performed better. This suggests that users get confused when watching large, complicated presentations.

In terms of enjoyment, the *interactive fly-through* mode was preferred by 17 out of 21 participants. The main listed reason was the possibility to freely look around. The *presence* evaluation also suggest that the users feel more immersion in this mode than in the *GPS slideshow*.

Based on these results we believe that while the *GPS slideshow* is somewhat better for the quantitative tasks, as it can directly display the queried information, the immersive modes convey an experience closer to that of actually doing the hike in the real-world space. In fact, we suspect that if we had included a real-world hike as a mode of presentation, the users would face similar issues in the evaluation as they did with the *interactive fly-through*, as the sense of spatial proprioception acquired by first-hand experience may not necessarily map to an accurate knowledge of spatial layout. It would be possible to verify this analogy with an experiment where we would have participants view an *immersive fly-through* and then ask them to retrace the same path in real-life without the use of navigation aids, but an experiment such as this would be difficult to perform and ethically problematic.

## CONCLUSION

We present a complete geo-registration pipeline for collections of images from hiking trips, which is capable of performing 3D reconstruction and scene geo-registration for expansive outdoor scenes without requiring speciality capture hardware. From this geo-registration, we create an immersive photo fly-through presentation where the images are overlaid on a virtual model of the terrain. We produce these presentations for four datasets from different geographical areas, with both a passive variant based on viewing these images in Google Earth and an interactive variant in a VR viewer.

Further improvements of our pipeline—e.g. optimization of the photo augmentation step by estimating how many and which photographs to download—could be interesting future work. Moreover, projecting the photograph texture onto the terrain during the fly-through is another direction worth exploring.

We compared our immersive presentation modes with two more traditional ones—a slideshow and a slideshow accompanied by a map—in a user study, where we measured user enjoyment, feeling of presence in the outdoor space, and the ability to understand the location and orientation of images in space. We found that in terms of spatial understanding, our modes performed significantly better than a pure *slideshow* and on par with the *GPS slideshow*, while the VR-based *interactive fly-through* conveyed a superior sense of presence and was preferred as the most enjoyable by the majority of users.

We hope our immersive trip reports can be useful both in private settings, to simply share the experience of a trip, as well as in public where they could be used to share e.g. trip instructions from users familiar with the area to the users who have yet to visit.

## ACKNOWLEDGEMENTS

This work was supported by V3C – “Visual Computing Competence Center” by Technology Agency of the Czech Republic, project no. TE01020415; by The Ministry of Education, Youth and Sports from the “National Programme of Sustainability (NPU II) project IT4Innovations excellence in science - LQ1602”; and by the IT4Innovations infrastructure which is supported from the Large Infrastructures for Research, Experimental Development and Innovations project “IT4Innovations National Supercomputing Center - LM2015070”.

## REFERENCES

1. M. Balabanović, L. I. Chu, and G. J. Wolff. 2000. Storytelling with digital photographs. *Proceedings of the SIGCHI conference on Human Factors in Computing Systems* 2, 1 (2000), 564–571. DOI: <http://dx.doi.org/10.1145/332040.332505>
2. B. B. Bederson and A. Boltman. 1999. Does animation help users build mental maps of spatial information?. In *Information Visualization, 1999. (Info Vis '99) Proceedings. 1999 IEEE Symposium on*. 28–35. DOI: <http://dx.doi.org/10.1109/INFVIS.1999.801854>
3. Sahil Chelaramani, Vamsidhar Muthireddy, and C. V. Jawahar. 2017. An Interactive Tour Guide for a Heritage Site. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*. 2943–2952. DOI: <http://dx.doi.org/10.1109/ICCVW.2017.347>
4. P. Chippendale, M. Zanin, and C. Andreatta. 2009. Collective photography. In *CVMP 2009 - The 6th European Conference for Visual Media Production*. 188–194. DOI: <http://dx.doi.org/10.1109/CVMP.2009.30>
5. Rustichello da Pisa and Marco Polo. c. 1300. *Livres des Merveilles du Monde*. (c. 1300).
6. A. Dalvandi, B. E. Riecke, and T. Calvert. 2011. Panoramic Video Techniques for Improving Presence in Virtual Environments. In *Proceedings of the 17th Eurographics Conference on Virtual Environments & Third Joint Virtual Reality (EGVE - JVRC'11)*. Eurographics Association, Aire-la-Ville, Switzerland, Switzerland, 103–110. DOI: <http://dx.doi.org/10.2312/EGVE/JVRC11/103-110>
7. G. de Haan, H. Piguillet, and F. Post. 2010. Spatial Navigation for Context-Aware Video Surveillance. *IEEE Computer Graphics and Applications* 30, 5 (Sept 2010), 20–31. DOI: <http://dx.doi.org/10.1109/MCG.2010.64>
8. Kit Devine. 2007. Place and Time : Creating Contextualised Presence for Virtual Heritage. In *Proc 13th Intl Conference on Virtual Systems and Multimedia*. 65–72.
9. Ruofei Du, Sujal Bista, and Amitabh Varshney. 2016. Video Fields: Fusing Multiple Surveillance Videos into a Dynamic Virtual Environment. In *Proceedings of the 21st International Conference on Web3D Technology (Web3D '16)*. ACM, New York, NY, USA, 165–172. DOI: <http://dx.doi.org/10.1145/2945292.2945299>

10. Ruofei Du and Amitabh Varshney. 2016. Social Street View: Blending Immersive Street Views with Geo-tagged Social Media. In *Proceedings of the 21st International Conference on Web3D Technology (Web3D '16)*. ACM, New York, NY, USA, 77–85. DOI : <http://dx.doi.org/10.1145/2945292.2945294>
11. Adrian Graham, Hector Garcia-Molina, Andreas Paepcke, and Terry Winograd. 2002. Time as essence for photo browsing through personal digital libraries. In *Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries - JCDL '02*. 326. DOI : <http://dx.doi.org/10.1145/544220.544301>
12. Susumu Harada, Mor Naaman, Yee Jiun Song, QianYing Wang, and Andreas Paepcke. 2004. Lost in memories: interacting with photo collections on PDAs. In *Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*. 325. DOI : <http://dx.doi.org/10.1109/JCDL.2004.1336143>
13. A. C. Haugstvedt and J. Krogstie. 2012. Mobile augmented reality for cultural heritage: A technology acceptance study. In *2012 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*. 247–255. DOI : <http://dx.doi.org/10.1109/ISMAR.2012.6402563>
14. Mahmoud Haydar, David Roussel, Madjid Maïdi, Samir Otmane, and Malik Mallem. 2011. Virtual and augmented reality for cultural computing and heritage: a case study of virtual exploration of underwater archaeological sites (preprint). *Virtual Reality* 15, 4 (01 Nov 2011), 311–327. DOI : <http://dx.doi.org/10.1007/s10055-010-0176-4>
15. Otmar Hilliges, Dominikus Baur, and Andreas Butz. 2007. Photohelix: Browsing, sorting and sharing digital photo collections. In *Tabletop 2007 - 2nd Annual IEEE International Workshop on Horizontal Interactive Human-Computer Systems*. 87–94. DOI : <http://dx.doi.org/10.1109/TABLETOP.2007.20>
16. W. Kabsch. 1976. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A* 32, 5 (1976), 922–923. DOI : <http://dx.doi.org/10.1107/S0567739476001873>
17. Biliana Kaneva, Josef Sivic, Antonio Torralba, Shai Avidan, and William T. Freeman. 2010. Infinite images: Creating and exploring a large photorealistic virtual space. In *Proceedings of the IEEE*, Vol. 98. 1391–1407. DOI : <http://dx.doi.org/10.1109/JPROC.2009.2031133>
18. Gunhee Kim, Seungwhan Moon, and Leonid Sigal. 2015. Joint photo stream and blog post summarization and exploration. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 3081–3089. DOI : <http://dx.doi.org/10.1109/CVPR.2015.7298927>
19. Johannes Kopf, Michael F. Cohen, and Richard Szeliski. 2014. First-person Hyper-lapse Videos. *ACM Trans. Graph.* 33, 4, Article 78 (July 2014), 10 pages. DOI : <http://dx.doi.org/10.1145/2601097.2601195>
20. Nermina Kraljic. 2008. Interactive Video Virtual Tours. In *Proceedings of the Central European Seminar on Computer Graphics (CESCG)*.
21. Volker Kuchelmeister and Jill Bennet. 2014. The Amnesia Atlas. An immersive SenseCam interface as memory-prosthesis. In *Proceedings of the 2014 International Conference on Virtual Systems and Multimedia, VSMM 2014*. 217–222. DOI : <http://dx.doi.org/10.1109/VSMM.2014.7136663>
22. Avanish Kushal, Ben Self, Yasutaka Furukawa, David Gallup, Carlos Hernandez, Brian Curless, and Steven M. Seitz. 2012. Photo tours. In *Proceedings - 2nd Joint 3DIM/3DPVT Conference: 3D Imaging, Modeling, Processing, Visualization and Transmission, 3DIMPVT 2012*. 57–64. DOI : <http://dx.doi.org/10.1109/3DIMPVT.2012.62>
23. Neil J McCurdy, William G Griswold, and Leslie A Lenert. 2005. RealityFlythrough: enhancing situational awareness for medical response to disasters using ubiquitous video. *AMIA Annual Symposium proceedings (AIMA) (2005)*, 510–4. <http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1560501>
24. P. Moulon, P. Monasse, and R. Marlet. 2013. Global Fusion of Relative Motions for Robust, Accurate and Scalable Structure from Motion. In *2013 IEEE International Conference on Computer Vision*. 3248–3255. DOI : <http://dx.doi.org/10.1109/ICCV.2013.403>
25. Satoshi Oonuki and Tetsuro Ogi. 2008. VR BBS Using Immersive Virtual Environment. In *22nd International Conference on Advanced Information Networking and Applications - Workshops (aina workshops 2008)*. 1006–1011. DOI : <http://dx.doi.org/10.1109/WAINA.2008.119>
26. Genevieve Patterson and James Hays. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. 2751–2758. DOI : <http://dx.doi.org/10.1109/CVPR.2012.6247998>
27. Pausanias. c. 150. Graecae descriptio. (c. 150).
28. J. C. Platt. 2000. AutoAlbum: Clustering digital photographs using probabilistic model merging. In *Proceedings - IEEE Workshop on Content-Based Access of Image and Video Libraries, CBAIVL 2000*. 96–100. DOI : <http://dx.doi.org/10.1109/IVL.2000.853847>
29. John C. Platt, Mary Czerwinski, and Brent A. Field. 2003. PhotoTOC: Automatic clustering for browsing personal photographs. In *ICICS-PCM 2003 - Proceedings of the 2003 Joint Conference of the 4th International Conference on Information, Communications and Signal Processing and 4th Pacific-Rim Conference on Multimedia*, Vol. 1. 6–10. DOI : <http://dx.doi.org/10.1109/ICICS.2003.1292402>



30. François Pomerleau, Francis Colas, Roland Siegwart, and Stéphane Magnenat. 2013. Comparing ICP Variants on Real-World Data Sets. *Autonomous Robots* 34, 3 (Feb. 2013), 133–148. DOI: <http://dx.doi.org/10.1007/s10514-013-9327-2>
31. O. Radiano, Y. Graber, M. Mahler, L. Sigal, and A. Shamir. 2017. Story Albums: Creating Fictional Stories From Personal Photograph Sets. *Computer Graphics Forum* 37, 1 (2017), 19–31. DOI: <http://dx.doi.org/10.1111/cgfm.13099>
32. Kerry Rodden and Kenneth Wood. 2003. How do people manage their digital photographs? *CHI '03: Proceedings of the SIGCHI conference on Human factors in computing systems* 5, 1 (2003), 409–416. DOI: <http://dx.doi.org/10.1145/642611.642682>
33. Klaus Schoeffmann, David Ahlstrom, and Marco A. Hudelist. 2014. 3-D interfaces to improve the performance of visual known-item search. *IEEE Transactions on Multimedia* 16, 7 (2014), 1942–1951. DOI: <http://dx.doi.org/10.1109/TMM.2014.2333666>
34. Johannes Lutz Schönberger and Jan-Michael Frahm. 2016. Structure-from-Motion Revisited. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4104–4113. DOI: <http://dx.doi.org/10.1109/CVPR.2016.445>
35. Johannes L. Schönberger, True Price, Torsten Sattler, Jan Michael Frahm, and Marc Pollefeys. 2017. A vote-and-verify strategy for fast spatial verification in image retrieval. In *Computer Vision - ACCV 2016 - 13th Asian Conference on Computer Vision, Revised Selected Papers (Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics))*, Vol. 10111 LNCS. Springer Verlag, Germany, 321–337. DOI: [http://dx.doi.org/10.1007/978-3-319-54181-5\\_21](http://dx.doi.org/10.1007/978-3-319-54181-5_21)
36. Ismail Oner Sebe, Jinhui Hu, Suya You, and Ulrich Neumann. 2003. 3D Video Surveillance with Augmented Virtual Environments. In *First ACM SIGMM International Workshop on Video Surveillance (IWVS '03)*. ACM, New York, NY, USA, 107–112. DOI: <http://dx.doi.org/10.1145/982452.982466>
37. Qi Shan, Brian Curless, Yasutaka Furukawa, Carlos Hernandez, and Steven M. Seitz. 2014. Photo uncrop. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Vol. 8694 LNCS. 16–31. DOI: [http://dx.doi.org/10.1007/978-3-319-10599-4\\_2](http://dx.doi.org/10.1007/978-3-319-10599-4_2)
38. B. Shneiderman. 2000. Visualization methods for personal photo collections: browsing and searching in the PhotoFinder. In *2000 IEEE International Conference on Multimedia and Expo. ICME2000. Proceedings. Latest Advances in the Fast Changing World of Multimedia (Cat. No.00TH8532)*, Vol. 3. 1539–1542. DOI: <http://dx.doi.org/10.1109/ICME.2000.871061>
39. Mel Slater, Martin Usoh, and Anthony Steed. 1994. Depth of Presence in Virtual Environments. *Presence: Teleoperators and Virtual Environments* 3, 2 (1994), 130–144. DOI: <http://dx.doi.org/10.1162/pres.1994.3.2.130>
40. Scott L. Smith. 2015. Stopmotion Photowalk Animation for Spatial Immersion in a Remote Cultural Heritage Site. In *Proceedings of the Conference on Electronic Visualisation and the Arts (EVA '15)*. BCS Learning & Development Ltd., Swindon, UK, 298–305. DOI: <http://dx.doi.org/10.14236/ewic/eva2015.33>
41. Noah Snavely, Rahul Garg, Steven M. Seitz, and Richard Szeliski. 2008. Finding paths through the world's photos. *ACM Transactions on Graphics* 27, 3 (2008), 1. DOI: <http://dx.doi.org/10.1145/1360612.1360614>
42. Noah Snavely, Steven M Seitz, and Richard Szeliski. 2006. Photo tourism: Exploring Photo Collections in 3D. *ACM Transactions on Graphics* 25, 3 (2006), 835–846. DOI: <http://dx.doi.org/10.1145/1141911.1141964>
43. Noah Snavely, Steven M. Seitz, and Richard Szeliski. 2008. Modeling the world from Internet photo collections. *International Journal of Computer Vision* 80, 2 (2008), 189–210. DOI: <http://dx.doi.org/10.1007/s11263-007-0107-3>
44. Strabo. 7. Geographica. (7).
45. M. Claudia tom Dieck and Timothy Hyungsoo Jung. 2017. Value of augmented reality at cultural heritage sites: A stakeholder approach. *Journal of Destination Marketing & Management* 6, 2 (2017), 110 – 117. DOI: <http://dx.doi.org/10.1016/j.jdmm.2017.03.002> Special edition on Digital Destinations.
46. James Tompkin, Fabrizio Pece, Rajvi Shah, Shahram Izadi, Jan Kautz, and Christian Theobalt. 2013. Video collections in panoramic contexts. In *Proceedings of the 26th annual ACM symposium on User interface software and technology - UIST '13*. 131–140. DOI: <http://dx.doi.org/10.1145/2501988.2502013>
47. Carlo Torniai, Steve Battle, and Steve Cayzer. 2007. *Sharing, Discovering and Browsing Geotagged Pictures on the World Wide Web*. Springer London, London, 159–170. DOI: [http://dx.doi.org/10.1007/978-1-84628-827-2\\_15](http://dx.doi.org/10.1007/978-1-84628-827-2_15)
48. Severi Uusitalo, Peter Eskolin, Yu You, and Petros Belimpasakis. 2010. An extensible mirror world from user-generated content. In *Proceedings - IEEE Virtual Reality*. 311–312. DOI: <http://dx.doi.org/10.1109/VR.2010.5444751>
49. Eduardo Veas, Alessandro Mulloni, Ernst Kruijff, Dieter Schmalstieg, Holger Regenbrecht, and Dieter Schmalstieg. 2010. Techniques for view transition in multi-camera outdoor environments. In *Proceedings of Graphics Interface*. Canadian Information Processing Society, Toronto, Ont., Canada, Canada, 193–200. DOI: <http://dl.acm.org/citation.cfm?id=1839214.1839248>

50. Chun-Po Wang, Kyle Wilson, and Noah Snavely. 2013. Accurate georegistration of point clouds using geographic data. *3DTV-Conference, 2013 International Conference on* (2013), 33–40. DOI : <http://dx.doi.org/10.1109/3DV.2013.13>
51. M. Wang, J. B. Liang, S. H. Zhang, S. P. Lu, A. Shamir, and S. M. Hu. 2018. Hyper-Lapse From Multiple Spatially-Overlapping Videos. *IEEE Transactions on Image Processing* 27, 4 (April 2018), 1735–1747. DOI : <http://dx.doi.org/10.1109/TIP.2017.2749143>
52. Y. Wang, D. Bowman, D. Krum, E. Coalho, T. Smith-Jackson, D. Bailey, S. Peck, S. Anand, T. Kennedy, and Y. Abdrakov. 2008. Effects of Video Placement and Spatial Context Presentation on Path Reconstruction Tasks with Contextualized Videos. *IEEE Transactions on Visualization and Computer Graphics* 14, 6 (Nov 2008), 1755–1762. DOI : <http://dx.doi.org/10.1109/TVCG.2008.126>
53. Yi Wang, David M. Krum, Enylton M. Coelho, and Doug A. Bowman. 2007. Contextualized Videos: Combining Videos with Environment Models to Support Situational Understanding. *IEEE Transactions on Visualization and Computer Graphics* 13, 6 (Nov. 2007), 1568–1575. DOI : <http://dx.doi.org/10.1109/TVCG.2007.70544>
54. C. Wu. 2013. Towards Linear-Time Incremental Structure from Motion. In *2013 International Conference on 3D Vision - 3DV 2013*. 127–134. DOI : <http://dx.doi.org/10.1109/3DV.2013.25>
55. C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. 2011. Multicore bundle adjustment. In *CVPR 2011*. 3057–3064. DOI : <http://dx.doi.org/10.1109/CVPR.2011.5995552>
56. Fuqu Wu and Melanie Tory. 2009. PhotoScope: Visualizing Spatiotemporal Coverage of Photos for Construction Management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '09)*. ACM, New York, NY, USA, 1103–1112. DOI : <http://dx.doi.org/10.1145/1518701.1518869>
57. Chen Yang and Gérard Medioni. 1992. Object modelling by registration of multiple range images. *Image and Vision Computing* 10, 3 (1992), 145–155. DOI : [http://dx.doi.org/10.1016/0262-8856\(92\)90066-c](http://dx.doi.org/10.1016/0262-8856(92)90066-c)
58. Zhengyou Zhang. 1998. Determining the Epipolar Geometry and its Uncertainty: A Review. *International Journal of Computer Vision* 27, 2 (1998), 161–195. DOI : <http://dx.doi.org/10.1023/a:1007941100561>
59. Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2018. Places: A 10 million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* PP, 99 (2018), 1–1. DOI : <http://dx.doi.org/10.1109/TPAMI.2017.2723009>
60. Zhe Zhu, Hao Zhi Huang, Zhi Peng Tan, Kun Xu, and Shi Min Hu. 2016. Faithful Completion of Images of Scenic Landmarks Using Internet Images. *IEEE Transactions on Visualization and Computer Graphics* 22, 8 (2016), 1945–1958. DOI : <http://dx.doi.org/10.1109/TVCG.2015.2480081>