A 3D MAP AUGMENTED PHOTO GALLERY APPLICATION ON MOBILE DEVICE

Junsheng Fu[⋆ †] *Lixin Fan*[†] *Kimmo Roimela*[†] *Yu You*[†] *Ville-Veikko Mattila*[†]

[⋆] Tampere University of Technology, Department of Signal Processing, Finland
[†] Nokia Research Center, Media Technologies Lab, Finland
{ext-junsheng.1.fu, lixin.fan, kimmo.roimela, yu.you, ville-veikko.mattila}@nokia.com

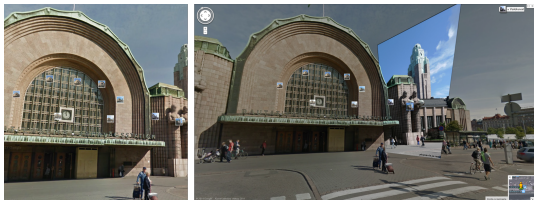
ABSTRACT

This paper proposes a 3D map augmented photo gallery mobile application that allows user to virtually transit from 2D image space to the 3D map space, to expand the field of view to surrounding environments that are not visible in the original image, and to change viewing angles among different global registered images during the image browsing. The processing of images consists of two main steps: in the first step, the client application uploads an image to the GeoImage Engine which extracts the geo-metadata and returns them back to the client; in the second step, the client application requests augmented content from server, and then renders 3D view of images on the screen of mobile devices.

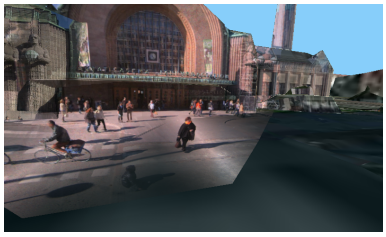
***Index Terms*—** Mobile Image Applications, Augmented Reality, 3D Map

1. INTRODUCTION

Capturing an image with a camera phone and sharing the photo with a friend or on social media has gradually becoming part of our daily activities, because of the increasing penetration rate of mobile phones and popularity of image sharing service. Currently, map-based services also integrate location-based photo sharing functionality. For example shown in Fig. 1 (a), Google Map allows users to view floating thumbnails during street-view navigation, and once a thumbnail is selected, e.g. by mouse clicking, users can change the viewing angle from the street-view image to the 2D images. While it provides an interactive experience, this service is more gear to the augmentation of the street-view navigation, and thus, image browsing is somehow limited to 2D experience. In this paper, we propose a mobile application that provides users with a 3D map augmented image browsing experience by exploiting a back-end server to automatically compute global positions and orientations of images uploaded from the client application. With this mobile application installed, users can see from a mobile device screen where the images were exactly taken in the real word and view the image projection to the 3D building in the map model. One snapshot of the application is shown in Fig. 1 (b).



(a)



(b)

Fig. 1. (a) two screen shoots from Google map street-view service. Floating thumbnails in the street-view image indicates 2D images associated with the current scene. An enlarged image is overlaid on the street view, if the mouse pointer is hovering on a corresponding thumbnail. Note the overlay of thumbnail is based on 2D homograph transform and noticeable artifacts are often observed for non-planar scenes. (b) a screen shoot from our 3D map augmented photo gallery application. One user captured image is shown in a 3D map, and since the building facet and the ground are modelled separately, the mapped photo image gives more immersive 3D experience.

There are two challenging tasks in this kind of augmented reality applications. The first difficulty, for each uploaded photo, lies in the computation of the global 6 degree of freedom camera pose which consists of the position and orientation in a global coordinate system. The second demanding issue is related to the rendering of the user-captured images in the 3D map.

Related works are discussed as follows. With the progress of the structure from motion techniques, image browsing is not only limited in a 2D space and users can interactively move the viewing angle in the 3D space by seamlessly tran-

sitioning between different images. One application is Microsoft Photosynth [1], which can automatically computes each photo's viewpoint, generates a sparse 3D model of the scene, and calculates a smooth path through the camera pose for a set of given photo. With this path, Photosynth provides the experience of moving through a gliding motion and photos are sliced into multi-resolution pyramids for efficient access. While good scalability and impressive rendering effects have been demonstrated, the application relies on the feature matching among the user uploaded images, and calculates the camera poses in a local coordinate system instead of in a global coordinate system. To our best knowledge and surprise, there are only a handful of global camera pose based systems reported in the literature. In a more recent research by Zhang et al. [2] the approach is to match a user generated query image against a database of geo-tagged images with known global 6 degrees of freedom poses. Once a correct image match is made, the point to point correspondence between query and retrieved image is used to compute a homograph transformation which can be used to transfer pixel accurate tag information onto the query image. While good localization accuracy and efficiency are demonstrated, the overall system is more focused on localization applications instead of image browsing. As illustrated by Liu et al. [3], the mobile visual localization system can extract a comprehensive set of geo-context information from a single photo. While high localization accuracy and good scalability are demonstrated, the overall system is more gear to localization applications instead of image browsing. Our earlier work [4] utilizes the associate global camera pose to enrich the video playback experience.

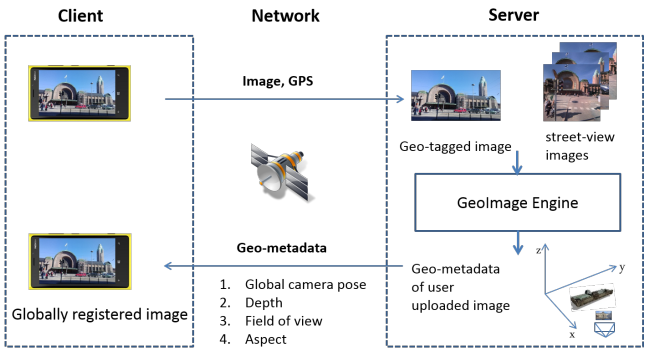
This paper illustrates a novel 3D map augmented photo gallery application that automatically uploads images from mobile clients to the server, extracts images' global camera poses and, consequently, enables augmented and interactive image browsing experiences with the augmentation of a 3D map.

2. SYSTEM FRAMEWORK

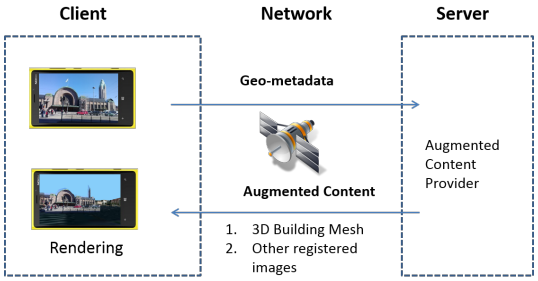
This section gives an overview of the system architecture and important functional modules, as shown in Fig. 2. The system framework consists of two main steps, named *extraction of global geo-metadata* and *client rendering*, and both steps are elaborated in this section.

2.1. Extraction of global geo-metadata

In the first step, the client uploads an image and its GPS to the server, the GeoImage Engine automatically extracts images geo-metadata and returns geo-metadata back to the client, see Step 1 in Fig. 2. The GeoImage Engine is an essential module, and it involves several important components, as shown in Fig. 3.



Step 1: Extraction of the global geo-metadata



Step 2: Client Rendering

Fig. 2. An overview of the system framework, which consists of two main steps.

- Once the image and its GPS are sent to the server, the GeoImage Engine starts to search and download closest 200 street-view images.
- Extract and compare the SIFT [5] features of both uploaded image and the street-view images, and then rank the street-view images according to the similarity to the uploaded one. Top k street-view images together with the uploaded image are used as input for 3D reconstruction.
- The 3D reconstruction module recovers the camera poses within a local coordinate system. This module uses Structure from Motion technique and we refer interested readers to [6, 7] for technique details.
- Based on the 3D reconstruction results, the depth range for the user captured image can be estimated. To find the camera pose in a global coordinate system, such as Earth-Centered Earth-Fixed (ECEF) system, we need to transform the local camera pose. Since registered street-view images have known camera poses in both local and global coordinate systems, a unique rigid transform is recovered [8] so that camera pose of user captured image can be mapped into the ECEF system.
- Finally, the GeoImage engine returns the client appli-

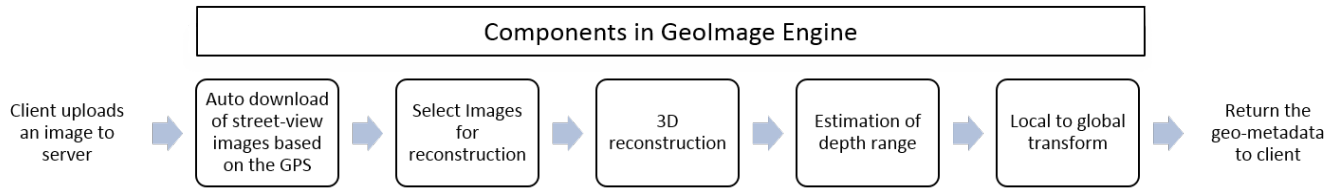


Fig. 3. Data flowchart in GeoImage Engine.

cation with geo-metadata of the user uploaded image, including global camera pose, depth, field of view and aspect ratio.

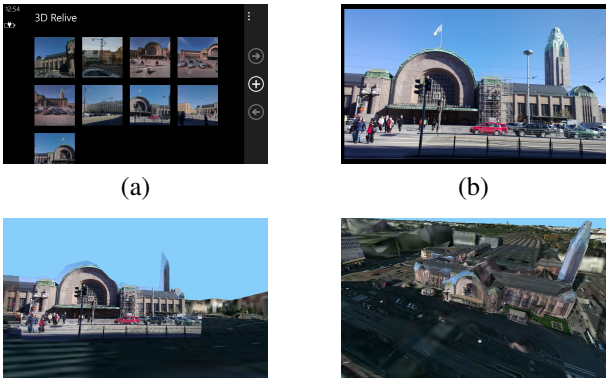
2.2. Client Rendering

Once geo-metadata are returned from the GeoImage Engine, the client application requests nearby augmented content such as 3D map, building mesh, and other globally registered images from the content provider, see Step 2 in Fig. 2. Since the geo-metadata of the image and related augmented data are provided, it is possible to render everything within a unified global coordinate system.

Our rendering algorithm for the images is based on view-dependent texturing [9]. To handle global geo-coordinates, all rendered content is first transformed to eye space, i.e., the local coordinate system of the current rendering camera. From the pose and projection parameters of each image, we calculate a per-image texture matrix that computes projected texture coordinates for image sampling. The texture matrix is passed to a pixel shader that also computes a per-pixel blending factor based on the angles between the original image ray and the current viewing ray, and the image ray and the normal vector of the surface being projected onto. The result RGBA pixels are then blended with the underlying map texture or other projected images.

Consequently, a number of intriguing image browsing experiences are now possible, and we demonstrate the mobile image application in a mobile phone as shown in Fig 4.

- Once the client application starts, users can see the photo gallery as the start screen, see Fig. 4 (a).
- Users can select a photo by clicking the image. Fig. 4 (b) shows one selected photo and the current view is in 2D image space.
- If users pinch in the image, a 3D map will be augmented in the client, and the viewing space would seamlessly transit from 2D image space to 3D map space, as shown in Fig. 4 (c). Moreover, users can see that the selected image is projected to the 3D building mesh based on its global camera pose.



(a) (b) (c) (d)

Fig. 4. Here are four screen shoots from our 3D map augmented mobile image application. (a) start screen. (b) select one image from the photo gallery. (c) switch from the 2D image view to 3D map view when user pinches in the image, and the image is projected to the build mesh. (d) user can arbitrarily change the viewing angles. Currently, user is looking at the projected image from a bird view.

- Users can arbitrarily change the viewing angles, e.g. looking from right, left or even a bird view as shown in Fig.4 (d). If multiple images in the same area are tagged with global camera pose, users can select other images in 3D map space and seamlessly transits from current view to other images' views.

3. EXPERIMENTS

To evaluate the performance of the proposed system, we made two experiments. (1) Test the pipeline with the user captured mobile images in real case. (2) Test the pipeline by using the street-view images with the 6 degree of freedom ground truth camera pose.

In the first experiment, two users captured 147 images in Helsinki downtown area with Nokia Lumia 820 mobile phones. The client application automatically upload the image to the server for processing. A computer with the processor of Intel Core i7 CPU @3.4GHz and the memory of

16 GB, is used for back-end processing. On the server side, based on the image GPS recorded from the mobile device, the most closest 200 street-view images is used for feature matching. Secondly, Bundle Adjustment [7] is used to calculate the camera poses. Thirdly, other metadata, including depth, field of view and image aspect ratio are computed accordingly. Finally, all the metadata are transformed to ECEF coordinate system, and returned to the mobile client.

The experiments results for the 147 user captured images are shown in Table 1, in which about 35% of the user captured images are able to be successfully recovered. Failure modes are mainly due to the lack of reliable features in textureless regions e.g. skies or ground. We are exploring various techniques to still improve the performance of the proposed pipeline. The registration time is on average less than 5 minutes, and we found this registration time is acceptable for our designed use cases, in which the browsing of 3D augmented contents often does not immediately follow the photo taking action. This is especially true when users share a photo with friends through social media networks.

In order to evaluate the accuracy of the registered camera poses, we test the pipeline with Nokia Here street-view images that have ground truth camera pose. In the second experiment, 305 Nokia Here street-view images from Helsinki downtown area are used to test the pipeline. The global camera pose consists of camera's location and orientation, and the camera's orientation difference can be represented as follow:

$$P = |P\_rec - P\_ori|$$

where P_rec means the orientation of the recovered camera pose, P_ori means the orientation of the ground truth pose provided by Nokia Here map, and P indicates the difference between two orientations. In Fig. 5 (a), we use P to represent the orientation difference in degree, and $P \in [0, 180]$ degree. According to Fig. 5 (a), the recovered camera pose for these 305 street-view images have satisfactory accuracy of orientation, and the maximum orientation error among test images is less than 0.18 degree.

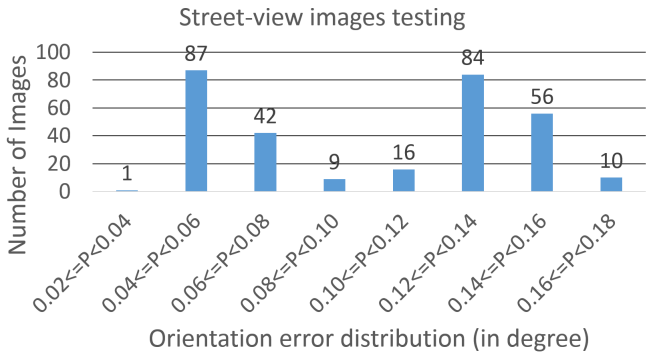
The location distance of the recovered and the ground truth pose can be calculated as follows:

$$d = |l\_rec - l\_ori|$$

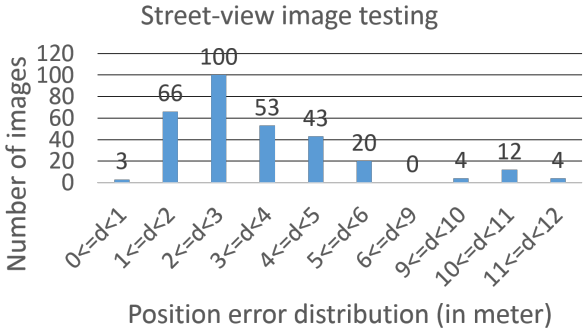
in which l_rec means for location of the recovered camera pose in ECEF coordinate system, l_ori means the ground truth location of camera in ECEF coordinate system, and d is the distance between these two with the unit in meters. Fig. 5 (b) shows the histogram of position distances between the recovered camera pose and the ground truth pose. According to the Fig. 5 (b), around 93.4% of the street-view images have good accuracy and the errors are less than 6 meters. However, there are 5.2% images which has more than 10 meters errors due to wrong feature matches.

Table 1. Testing results for user captured images

Total Images	147
Registered images	52
Registration rate	35.37%
Max processing time (mins per image)	13.2
Min processing time (mins per image)	2.7
Average processing time (mins per image)	4.7



(a)



(b)

Fig. 5. Histogram of orientation errors and the position errors for 305 street-view images.

4. CONCLUSION AND FUTURE WORK

This paper presented a 3D map augmented photo gallery application on mobile devices, which shows that the ordinary image sharing experience can be greatly enriched by leveraging the associated geo-metadata. Compared to existing systems, this mobile application can seamlessly transit from 2D image space to 3D map space, expend the field of view of the image to surrounding environments that are not visible in the original one, and change of the viewing angles. The experiment results demonstrate satisfactory accuracy performance of the pipeline. In the future, we will explore more efficient recovery of camera poses, targeting on real-time application.

5. REFERENCES

[1] Noah Snavely, Steven M. Seitz, and Richard Szeliski, "Photo tourism: Exploring photo collections in 3d," in *SIGGRAPH Conference Proceedings*, New York, NY, USA, 2006, pp. 835–846, ACM Press.

[2] J. Zhang, A. Hallquist, E. Liang, and A. Zakhor, "Location-based image retrieval for urban environments," in *Image Processing (ICIP), 2011 18th IEEE International Conference on*, Sept 2011, pp. 3677–3680.

[3] Heng Liu, Tao Mei, Jiebo Luo, Houqiang Li, and Shipeng Li, "Finding perfect rendezvous on the go: Accurate mobile visual localization and its applications to routing," in *Proceedings of the 20th ACM International Conference on Multimedia*, New York, NY, USA, 2012, MM '12, pp. 9–18, ACM.

[4] Junsheng Fu, Lixin Fan, Yu You, and Kimmo Roimela, "Augmented and interactive video playback based on global camera pose," in *ACM Multimedia*, 2013, pp. 461–462.

[5] David G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comput. Vision*, vol. 60, no. 2, pp. 91–110, Nov. 2004.

[6] Changchang Wu, "Towards linear-time incremental structure from motion," in *3DV-Conference, 2013 International Conference on*, June 2013, pp. 127–134.

[7] Changchang Wu, S. Agarwal, B. Curless, and S.M. Seitz, "Multicore bundle adjustment," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, June 2011, pp. 3057–3064.

[8] W. Kabsch, "A discussion of the solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 34, no. 5, pp. 827–828, Sep 1978.

[9] Paul E. Debevec, Yizhou Yu, and George Borshukov, "Efficient View-Dependent Image-Based Rendering with Projective Texture-Mapping," in *Eurographics Symposium on Rendering/Eurographics Workshop on Rendering Techniques*, 1998, pp. 105–116.