2016 12th International Conference on Signal-Image Technology & Internet-Based Systems

# All-in-One Mobile Outdoor Augmented Reality Framework for Cutural Heritage Site

Noh-young Park
UVR Lab.
GSCT, KAIST
Daejeon, Rep. of Korea
nypark@kaist.ac.kr

Eunseok Kim
AHRC
GSCT, KAIST
Daejeon, Rep. of Korea
scbgm@kaist.ac.kr

Jongwon Lee
Department of Computer Engineering
CNU
Daejeon, Rep. of Korea
uranos1@cnu.ac.kr

Woontack Woo
UVR Lab.
GSCT, KAIST
Daejeon, Rep. of Korea
wwoo@kaist.ac.kr

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

Figure 1. The processing results of the all-in-one mobile outdoor framework. (a) shows generated 3D keypoints and keyframes from a cultural heritage site, (b) describes keypoints for camera pose estimation and tracking process, (c) shows an example of AR visualization by drawing an axis of local coordinates, and (d) shows a runtime demonstraion of our mobile prototype.

***Abstract*—In this paper, we propose an all-in-one mobile outdoor augmented reality (AR) framework for a cultural heritage site. The framework was designed to incorporate computer vision-based augmented reality technology and ontology-based data-authoring technology. Through this framework, we clearly explain how to create 3D visual data for camera pose estimation and how to connect AR content with a cultural heritage site. In addition, we suggest a multi-threading camera tracking and offer an estimation model for mobile AR application. Finally, we have confirmed the efficiency and reliability of our framework. Through this vision-based AR framework, seamless AR application for a cultural heritage site can be made.**

***Keywords-augmented reality; cultural heritage; outdoor AR; AR framework;***

## I. INTRODUCTION

Nowadays, mobile AR technology is on the rise as a novel way to visualize information in various applications. According to Azuma's definition [1], the most important characteristic of AR is that it is interactive in real time and that it registers the content in 3D space. In order to register a real environment in virtual space, many computer vision algorithms have to be applied to the AR research. In this way, many of computer vision technologies are involved in a mobile AR application. Many computer vision technologies play an important role in mobile AR applications by providing accurate and robust camera detection and tracking algorithms.

Meanwhile, computer vision technology has been adapted into the cultural heritage domain. Among these applications, 3D acquisitions that reconstruct the cultural heritage site as a virtual 3D model is one of the most dominant. Computer vision technology is utilized for 3D replica production and morphological analysis to conserve cultural heritage sites [13, 14]. Studies that focus on this work concentrate on the 3D reconstruction of the object [15-17] and use vision-based algorithms and laser scanners. The focal points of these studies are fidelity (accuracy, resolution) and cost (time, facilities).

In recent trends, those two implementations of computer vision technology have been merged. Mobile AR applications have become a new trend in the cultural heritage domain as mobile devices and computation power evolve. Those applications provide guidance for museums (indoors) and cultural heritage sites (outdoors). After the first outdoor AR for a cultural heritage site was tested on a PC platform [7-9], cultural heritage AR applications moved to a mobile platform [10-11] to be disseminated via smart phones. Recently, AR applications for cultural heritage sites have been tested on wearable platforms [12]. From the point of view of computer vision technology, enabling robust and reliable camera tracking in an outdoor environment is the most significant challenge of addressed in recent research. Early studies tried to solve this problem by detecting known fiducial patterns as markers [2]. Several other researchers have used GPS and motion sensor data from inertial sensors to calculate the location and posture of mobile devices. Recent research has proposed a method that combines feature point detection and the structure-from-motion [3] method to understand the 3D geometry of outdoor environments. This method uses 3D feature points as reference data to estimate camera pose. The latest research attempts to apply both sensor and vision technology in an AR application [4-6] to solve large-scale localization problems.

In this context, the purpose of our research is to suggest a vision-based AR framework for a cultural heritage site. We thus propose an all-in-one outdoor AR framework design for the cultural heritage site. Our framework design includes 1) how to create a 3D visual data for mobile

978-1-5090-5698-9/16 $31.00 © 2016 IEEE
DOI 10.1109/SITIS.2016.81

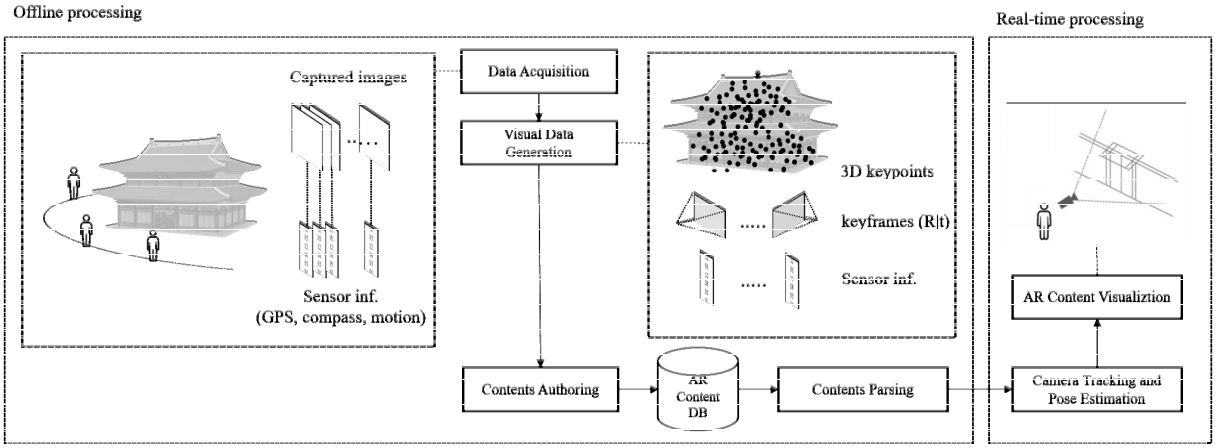



Figure 2. Framework design. The off-line processing part (left) generates visual data for camera tracking and connects AR contents through an authoring process. The online processing part (right) enables AR content visualization through a real-time camera tracking and pose estimation process.

camera tracking, 2) how to connect public information to an AR environment, and 3) how to visualize AR content for the user. Fig. 1 shows the processing results of the proposed framework.

Through the proposed framework, we expect that an outdoor AR guidance application with accurate large-scale localization can be made for cultural heritage sites. In addition, the framework is not specialized for the cultural heritage domain, so it can be implemented for general purposes such as training, education, social commerce, and so on.

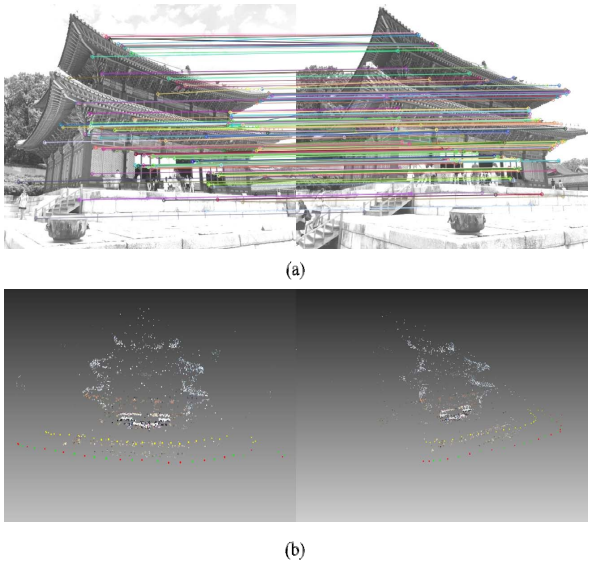
## II. VISION-BASED OUTDOOR AR FARMEWORK DESIGN

The proposed framework is designed to incorporate computer vision-based augmented reality technology and ontology-based data parsing technology to the cultural herniate domain. Our framework consists of three function modules: 1) visual data generation, 2) AR content authoring, and 3) real-time AR content visualization. Each of module is designed to be applied to outdoor cultural heritage site, especially on an ancient building. Fig. 2 describes overall procedure of the framework, and we will explain details of each module in below.

### *A. Visual Data Generation*

In this module, 3D keypoints and keyframes are automatically generated by the SfM (structure-from-motion) pipeline. As shown in Fig. 2, RGB image and sensor data (GPS, digital compass) that were captured at the same time and location are used as input data. After data acquisition, our system initiates the SfM pipeline.

For the feature extraction and matching process, we use an ORB [23] feature point and descriptor to reduce computational costs for the real-time mobile application. However, even though an ORB feature point is faster than a complex feature point such as SIFT [24] and SURF [25], it is limited to scale invariant conditions and severe viewpoint changes. To compensate for those characteristics of the ORB feature point, we adopt a sensor-fusion approach for the feature matching process. Finally, we reconstruct 3D coordinates of keypoints and keyframes by Bundler [26] method. After the SfM process, the surviving keypoints and keyframes are reconstructed as 3D coordinates. Fig. 3 shows results of each step of the SfM process.



(a)

(b)

Figure 3. (a) ORB keypoint extraction and matching process, (b) the result of visual data generation in which 3D keypoints and keyframes are reconstructed in 3D.

### *B. AR Content Authoring*

On the other hand, the content retrieval process is made to support the AR applications in outdoor environments. To support outdoor AR, the AR content has to be created first. In this authoring process, AR content has to cover not only the virtual content, but also the information for the visual data that was established in the previous step. Next, that AR content is parsed into the user's client or device using an appropriate method according to the conditions. A detailed description of each part of the retrieval process follows.

In the content authoring process, several issues have to be considered. First, there were several raw databases with

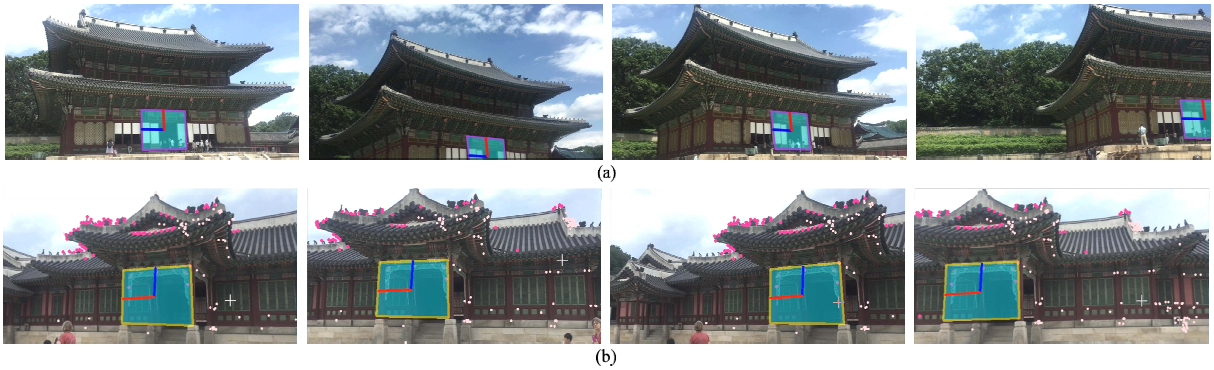


Figure 4. Demonstraion of AR content visualization module. (a) shows the AR content visualization results from different camera viewpoints and locations; (b) shows tracked 3D keypoints and visualizes the 3D axis of the corresponding local coordinates.

heterogeneous structures. To collect the virtual content of several related databases, we have to aggregate the databases.

To aggregate these virtual contents of heterogeneous databases, we adapt ontology as an aggregation method and developed a proper data model [18]. Next, we map the integrated virtual content with the recognition data from previous step. We developed a metadata scheme in our previous study [22] to establish the AR content database for this process. The metadata scheme serves as a blueprint for the AR content database and supports the data-parsing process.

In the data-parsing process, AR content is parsed from the database into the user's client or device. Both virtual content and the recognition data of AR content can be parsed separately, and the parsing process can be made either off-site and on-site. Considering the network bandwidth, memory, and computational power of the device, the virtual content and recognition data can be parsed with the appropriate method. In the case of off-site parsing, it supports off-line work but it requires sufficient device memory and does not guarantee content updates. On the other hand, on-site parsing can guarantee up-to-date content but depends on the network connection. Therefore, according to the environment of the AR application, an appropriate parsing method can be used. As a baseline, the whole parsing process in the prototype is conducted off-line (i.e., already stored in the user's device).

### C. *AR Content Visualization*

Finally, the content visualization module integrates 3D visual data and content authoring results and visualizes the AR content with the current 6DoF (degree-of-freedom) camera pose of the mobile device. The 3D positions of all AR content are located within the same local coordinate system, which was generated from III-A. Then, visualization module estimates relative 3D position and rotation of each AR content by current camera position. So, it is important to estimate robust camera pose in real-time under stable camera tracking.

To estimate the 6DoF camera pose in real time, we use sensor-fusion feature point matching. First, we extract the ORB feature point and descriptor from the current captured image. Then, we define the nearest keyframe using sensor information. We first filter out keyframes by GPS position and the direction of the digital compass sensor. After finding the nearest keyframe, we match the ORB descriptor between the current input image and the nearest keyframe. From this 3D–2D correspondence we can calculate the 6DoF camera pose by a Perspective-n-Point algorithm [19]. Finally, we properly augment the AR content by considering the current camera position and rotation. Fig. 4 shows examples of visualized AR content using the visualization module.

### D. *Parallel Tracking and Pose Estimation*

In order to stabilize camera detection and tracking, we propose a parallel tracking and pose estimation method. Our implementation process is divided into two different parts through a multi-threading technique. The proposed design consists of a foreground thread and a background thread. The foreground thread estimates the camera pose and the background thread adds the feature points.

The background thread is used in the process of ORB feature extraction and matching. After the background thread transmits matching information to the foreground thread, the foreground thread estimates the camera pose using 3D–2D correspondence of feature points. Then the foreground thread tracks the current ORB keypoints by an optical flow tracking [20] algorithm. Removing the error point is also important, because accumulated error affects the accuracy of the camera pose. In order to maintain an accurate camera pose, the foreground thread removes error points by calculating the re-projection pixel distance from the camera pose. Fig. 5 shows each process through a sequence diagram. Nevertheless, the background thread is several times slower than the foreground thread, the camera tracking module can perform stable tracking in real time.

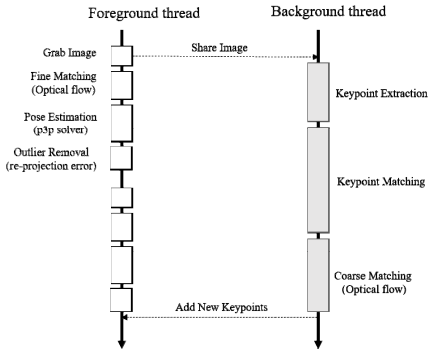
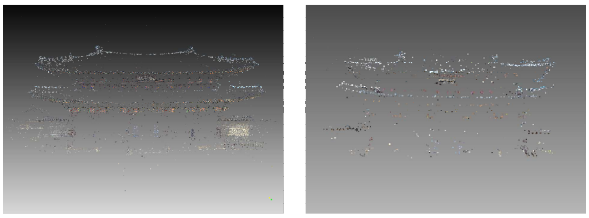



Figure 5. Multi-threading parallel tracking and pose estimation.

(a) (b) (c)

Figure 6. (a) : Dataset 1, (b) Dataset 2, and (c) Dataset 3.



(a) (b)

Figure 7. Result of 3D visual data generation. (a) used the SIFT feature point to compare with our method; in (b) both 3D structures are well reconstructed.

TABLE I. RESULTS OF VISUAL DATA GENERATION PROCESS

| Data | Generated Visual Data | | |
|---|---|---|---|
| | keyframes | 2D keypoints | 3D keypoints |
| Dataset (1) | 22 | 7752 | 2562 |
| Dataset (2) | 29 | 11246 | 3068 |
| Dataset (3) | 22 | 34847 | 12900 |

### *E. Implementation on Mobile Device*

The prototype of the AR content visualization module was also implemented on a mobile device. We successfully applied our prototypes on Android devices through the process of Android NDK cross-compiling. We also accelerated the ORB extraction and matching process using a general purpose GPU with OpenCL [21]. We thus reduced bottleneck processing time. The mobile prototype was tested on two different mobile devices. Fig. 1-(d) shows the AR content visualization module on a mobile device.

## III. EVALUATION

We performed a qualitative evaluation and a quantitative experiment to demonstrate the efficiency and reliability of our framework. The experiment consists of three different parts. The first involved proving the accuracy of the generated 3D visual data for AR. In the second part we evaluated the working speed and reliability of the camera tracking module. Lastly, we tested the performance of our mobile prototype version.

We generated a dataset from three buildings at an ancient palace to prove whether our system design could be applied to cultural heritage sites. Fig. 6 shows the appearance of the datasets. Each dataset consists of a set of RGB image and sensor data, as we already explained in section II-A. Table I describes the results of the visual data generation module.

Fig. 7 describes output results of visual data generation using dataset (1). Fig 7-(a) used the SIFT feature, and 7-(b) used our method (ORB feature and sensor information).

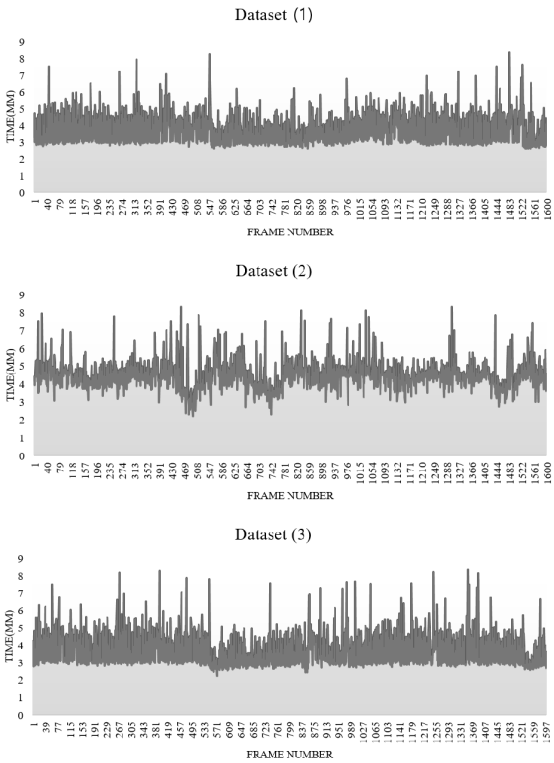



Figure 8. Processing time (milliseconds) for each dataset. The average processing time was less than 4 ms.

Fig 7-(b) shows a sparse distribution compared with 7-(a), but both 3D structures are very similar to the original 3D geometry of the building. Therefore, we can visually verify that our visual data have sufficient accuracy to be used for camera pose estimation and tracking.

In order to verify the efficiency and accuracy of the visualization module, we proceeded with a quantitative experiment. In this experiment, we analyzed both the working speed and the accuracy of the parallel camera tracking module, which we already explained in section II-D. We initially performed our experiment with an Intel i7-6700 CPU with a multi-core processor. Fig. 8 shows the processing speed of the parallel camera tracking module. Each graph shows the processing speed of the foreground thread of camera tracking module using input data from dataset (1), (2), and (3). The average processing time was less than 4 milliseconds per frame. This indicates that our module is sufficient for real-time AR applications, even on mobile devices. Fig. 9 also shows the accuracy of camera pose estimation. We calculated the average Euclidian distance between re-projected 3D keypoints and 2D keypoints using the 6DoF camera pose. In this way we proved the reliability and accuracy of the camera tracking module.

Finally, we tested the prototype of the AR visualization module, which ran on the mobile device. As we noted in section III-E, we imported the AR visualization module onto an Android device—a Samsung Galaxy S6, which has 4x2.1 GHz Cortex-A57 CPU and Mali-T760MP8 GPU. Table II shows the average processing time on the mobile

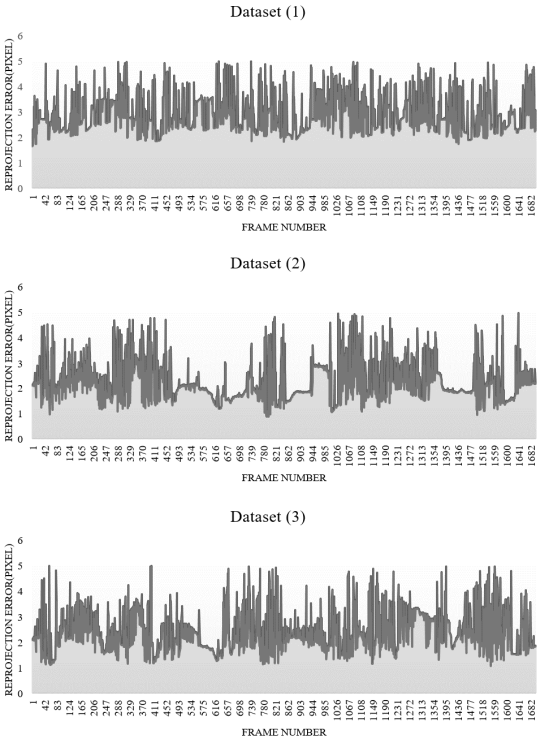



Figure 9. Re-projection error distance (pixels) of each dataset. The average error distance was less than a 3-pixel distance.

TABLE II. PROCESSING TIME ON MOBILE DEVICE

| Image Resolution (w · h) | Processing Time | | |
|---|---|---|---|
| | Processing Time (fps) | Foreground Thread (mm) | Background Thread (mm) |
| 640 · 360 | 30 | 16.338 | 347.904 |
| 640 · 480 | 30 | 17.296 | 341.703 |
| 720 · 480 | 29 | 18.963 | 347.922 |
| 800 · 450 | 27 | 18.403 | 346.781 |
| 960 · 720 | 20 | 25.299 | 517.842 |
| 1280 · 720 | 15 | 32.263 | 646.964 |

device at different resolutions. With these results, we proved that our mobile prototype could work in real-time AR applications. However, we still have an accuracy problem in our prototype. Even though we accelerated the ORB extraction and matching process by mobile GPU, the background thread still performs badly, which reduces tracking accuracy and stability.

## IV. CONCLUSION

In this paper, we proposed an effective design for an outdoor AR framework that can be used especially for cultural heritage sites. The proposed framework was designed to comprise both the generation of visual data for real-time AR applications and ontology-based AR authoring. In addition, we verified that the prototype of the AR application supports the accurate and reliable performance of real-time mobile applications.

An interesting direction for future research would be to extend our method to a large-scale heritage site, also we expect that our AR authoring method can be extended to support heterogeneous data (e.g., 3D models, videos, etc.). Through this work, more comprehensive and seamless AR guidance applications for cultural heritage sites can be developed.

## ACKNOWLEDGMENT

This research is supported by Ministry of Culture, Sports and Tourism(MCST) and Korea Creative Content Agency(KOCCA) in the Culture Technology(CT) Research & Development Program 2014.

## REFERENCES

[1] Azuma, Ronald T. "A survey of augmented reality." Presence: Teleoperators and virtual environments 6.4 (1997): 355-385.

[2] Koller, Dieter, et al. "Real-time vision-based camera tracking for augmented reality applications." Proceedings of the ACM symposium on Virtual reality software and technology. ACM, 1997.

[3] Arth, Clemens, et al. "Real-time self-localization from panoramic images on mobile devices." Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on. IEEE, 2011.

[4] You, Suya, and Ulrich Neumann. "Fusion of vision and gyro tracking for robust augmented reality registration." Virtual Reality, 2001. Proceedings. IEEE. IEEE, 2001.

[5] Schall, Gerhard, et al. "Global pose estimation using multi-sensor fusion for outdoor augmented reality." Mixed and Augmented Reality, 2009. ISMAR 2009. 8th IEEE International Symposium on. IEEE, 2009.

[6] Ribo, Miguel, et al. "Hybrid tracking for outdoor augmented reality applications." IEEE Computer Graphics and Applications 22.6 (2002): 54-63.

[7] Vlahakis, Vassilios, et al. "Archeoguide: first results of an augmented reality, mobile computing system in cultural heritage sites." Virtual Reality, Archeology, and Cultural Heritage. 2001.

[8] Vlahakis, Vassilios, et al. "3D interactive, on-site visualization of ancient Olympia." 3D Data Processing Visualization and Transmission, 2002. Proceedings. First International Symposium on. IEEE, 2002.

[9] Caggianese, Giuseppe, Pietro Neroni, and Luigi Gallo. "Natural interaction and wearable augmented reality for the enjoyment of the cultural heritage in outdoor conditions." International Conference on Augmented and Virtual Reality. Springer International Publishing, 2014.

[10] Ancona, Massimo, et al. "Mobile vision and cultural heritage: the Agamemnon project." Proceedings of 1st International Workshop on Mobile Vision, Graz, Austria. 2006.

[11] Angelopoulou, Anastassia, et al. "Mobile augmented reality for cultural heritage." International Conference on Mobile Wireless Middleware, Operating Systems, and Applications. Springer Berlin Heidelberg, 2011.

[12] Caggianese, Giuseppe, Pietro Neroni, and Luigi Gallo. "Natural interaction and wearable augmented reality for the enjoyment of the cultural heritage in outdoor conditions." International Conference on Augmented and Virtual Reality. Springer International Publishing, 2014.

[13] Pieraccini, Massimiliano, Gabriele Guidi, and Carlo Atzeni. "3D digitizing of cultural heritage." Journal of Cultural Heritage 2.1 (2001): 63-70.

[14] Gomes, Leonardo, Olga Regina Pereira Bellon, and Luciano Silva. "3D reconstruction methods for digital preservation of cultural heritage: A survey." Pattern Recognition Letters 50 (2014): 3-14.

[15] Guidi, Gabriele, J-A. Beraldin, and Carlo Atzeni. "High-accuracy 3D modeling of cultural heritage: the digitizing of Donatello's" Maddalena"." IEEE Transactions on image processing 13.3 (2004): 370-380.

[16] Wenzel, Konrad, et al. "High-resolution surface reconstruction from imagery for close range cultural Heritage applications." ISPRS Arch 39 (2012): B5.

[17] Bevan, Andrew, et al. "Computer vision, archaeological classification and China's terracotta warriors." Journal of Archaeological Science 49 (2014): 249-254.

[18] Kim, Sunhyuck, et al. "Towards a semantic data infrastructure for heterogeneous Cultural Heritage data-challenges of Korean Cultural Heritage Data Model (KCHDM)." 2015 Digital Heritage. Vol. 2. IEEE, 2015.

[19] Quan, Long, and Zhongdan Lan. "Linear n-point camera pose determination." IEEE Transactions on pattern analysis and machine intelligence 21.8 (1999): 774-780.

[20] Bouguet, Jean-Yves. "Pyramidal implementation of the affine lucas kanade feature tracker description of the algorithm." Intel Corporation 5.1-10 (2001): 4.

[21] Stone, John E., David Gohara, and Guochun Shi. "OpenCL: A parallel programming standard for heterogeneous computing systems." Computing in science & engineering 12.1-3 (2010): 66-73.

[22] E. Kim, J. Kim, W. Woo, "5W1H-Based Metadata Schema for Context-Aware Augmented-Reality Application in Cultural Heritage Domain", Digital Heritage Conference, 2015

[23] Rublee, Ethan, et al. "ORB: An efficient alternative to SIFT or SURF." 2011 International conference on computer vision. IEEE, 2011.

[24] Lowe, David G. "Distinctive image features from scale-invariant keypoints." International journal of computer vision 60.2 (2004): 91-110.

[25] Bay, Herbert, et al. "Speeded-up robust features (SURF)." Computer vision and image understanding 110.3 (2008): 346-359.

[26] Snavely, Noah. "Bundler: Structure from motion (SFM) for unordered image collections." Available online: phototour. cs. washington. edu/bundler/(accessed on 12 July 2013) (2010).