A Tracking Framework for Augmented Reality Tours on Cultural Heritage Sites

Byung-Kuk Seo* Kangsoo Kim† Jungsik Park‡ Jong-Il Park§

Department of Electronics and Computer Engineering
Hanyang University, Seoul, Korea

Abstract

Visual tracking for augmented reality tours is still challenging for cultural heritage sites because of the great variation of tracking targets and environments on such sites. Even at today's state of the art, it is almost impossible to apply just one tracking method to all the various environments with any hope of success. This paper presents a tracking framework to overcome this problem. It consists of different tracking flows, each efficiently using robust visual cues of the target scene. Analysis of the tracking environment enables more practical tracking at the sites. The reliability of the tracking framework is verified through on-site demonstrations at Gyeongbokgung, the most symbolic cultural heritage site in Korea.

CR Categories: H.5.1 [Information Systems]: Information Interface and Representation—Artificial, augmented, and virtual realities; I.4.8 [Image Processing and Computer Vision]: Scene Analysis—Tracking; I.4.9 [Image Processing and Computer Vision]: Application

Keywords: camera tracking, augmented reality, multimedia tour guides

1 Introduction

Augmented reality (AR)-based on-site tour guides have provided intuitive and immersive experiences to tourists by superimposing virtual contents on real cultural heritage sites. There have been many studies and research projects for AR-based tour guides for cultural tourism [Noh et al. 2009]. For example, augmented reality-based cultural heritage on-site guide (Archeoguide) [Arc] project presented new ways for accessing information at cultural heritage sites such as navigation, visualization of AR reconstruction of ancient life, and multimodal interaction. [Vlahakis et al. 2002] tested and evaluated Archeoguide at the Olympia archaeological site in Greece. Intelligent tourism and cultural information through ubiquitous services (iTacitus) [iTa] is another good example of AR-based tour guides. iTacitus provides 3-D virtual models, with multimedia content such as video and audio, on real sites. The prototype was implemented on Ultra Mobile PCs and more recently, on smartphones. It was demonstrated at Reggia Venaria Reale in Italy and Winchester Castle's Great Hall in the UK [iTa ; Zöellner et al. 2009]. [Tenmoku et al. 2004] presented a navigation system called the Nara Palace Site Navigator, and demonstrated it on Nara palace site. It provides tour guide information with AR contents suited to the user's devices—wearable computer, personal digital assistant (PDA), or cellular phone. [Papagiannakis et al. 2005] developed an AR framework to revive life in ancient fresco paintings in ancient Pompeii and create narrative space. The framework superimposes 3-D virtual characters on the real environment, with simulation of body, speech, facial expression, and cloth.

Tracking is a core technology for AR-based tour guides because it enables the correct presentation of virtual contents to tourists according to their different locations and viewpoints, so providing fully immersive experiences. Tracking technology has been widely studied in the field of computer science, particularly in computer vision and robot navigation. Most sensor devices support tarcking well—e.g., ultrasonic, global positioning system (GPS), inertial, and network sensors. Vision sensors like cameras are commonly used for visual tracking. Visual tracking has been steadily proposed in the literature. A simple and robust approach is to detect visual markers laid on a target scene (ARToolKit [Kato and Billinghurst 1999] is the most popular of these). Because of visual interference of the markers, markerless camera tracking has been recently issued. This extracts natural information of the target scenes such as shape, texture, or geometric primitive. Feature points such as corners or blobs have been widely used for visual cues of the target scenes with benefit of various detectors and descriptors [Tuytelaars and Mikolajczyk 2008], typically scale-invariant feature transform (SIFT) [Lowe 2004]. The use of feature points has mainly been in 2-D planar-based camera tracking, because a camera pose (position and orientation) is easily estimated by homographies between feature points detected on planar scenes or objects. However, it is much more difficult to robustly track 3-D objects with complex geometry or without texture. To overcome this, model-based camera tracking has been proposed, using features of 3-D object models like edges [Drummond and Cipolla 2002] or contours [Rosten and Drummond 2003] have been used for camera tracking.

Although visual tracking has provided good solutions for AR applications, it is still difficult for it to augment virtual contents accurately on the real sites (as also mentioned in [Zoellner et al. 2008]). In cultural heritage sites, the great variation of tracking targets and environments make it almost possible, even at today's state of the art, to apply just one tracking method to all the vairous environments with any hope of success. This paper presents a tracking framework to overcome this problem. It consists of different tracking flows, each efficiently using robust visual cues of the target scene. Analysis of the tracking environments enables more practical tracking at the sites. The reliability of the tracking framework is verified through on-site demonstrations at Gyeongbokgung, the most symbolic cultural heritage site in Korea.

This paper presents results from an on-going project called mobile augmented reality tour (MART), developing new tour guide services on mobile phones using AR. This paper mainly focuses on visual tracking developed by our research group in the MART project and shows our results.

2 Tracking Framework

The tracking framework of our AR-based on-site tour system is shown in Figure 1. It consists of sensing devices to obtain scene in-

*e-mail: bkseo@mr.hanyang.ac.kr
†e-mail: vistavision@mr.hanyang.ac.kr
‡e-mail: nangsik@mr.hanyang.ac.kr
§e-mail: jipark@hanyang.ac.kr (corresponding author)

VRCAI 2010, Seoul, South Korea, December 12 – 13, 2010.
© 2010 ACM 978-1-4503-0459-7/10/0012 $10.00

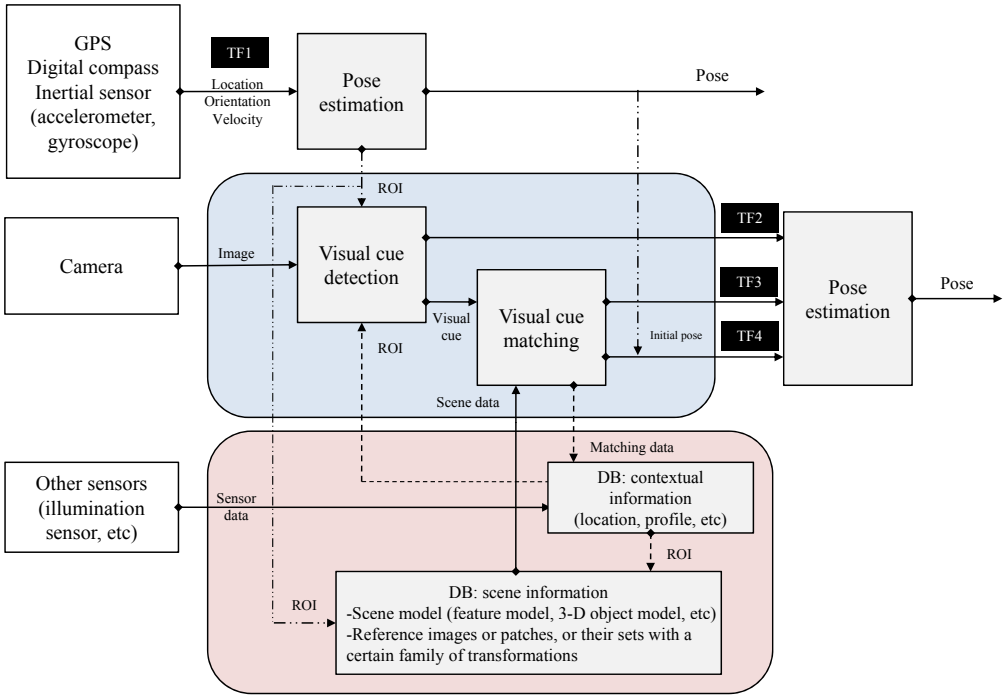


Figure 1: *Tracking framework. (TF1: sensor-based tracking flow, TF2, TF3: vision-based tracking flow, TF4: hybrid tracking flow).*

formation and processing modules to estimate poses of the system. The processing modules also have three tracking flows: sensor-based, vision-based, and hybrid tracking. Each flow is adaptively performed according to tracking environments on real sites.

Sensor-based tracking (TF1) estimates poses of the tour system by measuring the user's location, orientation, and velocity from sensors such as GPS, digital compass, and inertial sensors. GPS is a popular commercialized sensor and commonly provides location-based service (LBS). Most recent smartphones have GPS, allowing many mobile AR applications, such as Layar [Lay] and Wikitude [Wik] to offer LBS using GPS.

Vision-based tracking (TF2, TF3) estimates poses from images of a target scene captured by a camera. Visual cues such as edge, corner, or feature point are detected in the captured images. Pose estimation is performed along tracking flows which are adaptively selected according to dominant features of the target scene. In TF2, the poses are estimated by using the detected visual cues directly when the characteristics of the target scene are known, e.g., when geometric primitives such as line segments or rectangles are dominant in the target scene. In TF3, reference information of the target scene is predefined as a database. The poses are estimated by matching the detected visual cues to those in the database. The database can contain reference images or patches, or their sets with a certain family of transformations such as rotation, scale, and affine transformation. It can also contain scene models such as feature or object models.

To match the detected visual cues to those in the database, each visual cue (image or patch) is converted to a descriptor and the correspondences are found by comparing their similarity. The poses are then estimated by computing transformations between correspondences, e.g., homographies between planar features. Reference image or patch sets with a certain family of transformations also can be used efficiently because database construction is usually performed offline just once and the matching process is simple. When a target scene has 3-D objects with complex geometry or without texture, scene models that includes 3-D information of the target scene can be used for tracking. They can be a feature or object models. The feature model is built by reconstructing 3-D coordinates of feature points of the target scene. For reconstruction, the structure-from-motion (SfM) method is commonly used, the poses being estimated using known 2D-3D correspondences of the feature points. The object model is obtained by modeling 3-D objects of the target scene. The poses are estimated by matching the detected visual cues of the 3-D object, such as edges or contours, to those of the projected object model.

Hybrid tracking (TF4) combines vision-based tracking with all measurements obtained by sensor-based tracking. The sensor-based tracking can provides information about the target scene that can then be used for efficient vision-based tracking, e.g., to define a region of interest (ROI) and an initial pose of the system. Contextual information of the target scene, such as locations and situations, also offers helpful information to estimate poses.

3 Practical Tracking on Real Sites

The MART project aims at provide mobile AR-based tour services at cultural heritage sites. The Gyeongbokgung area, including the National Palace Museum of Korea, was chosen as the first trial test zone. Gyeongbokgung is a royal palace located in Seoul, Korea and is the most symbolic cultural heritage site in Korea. There are already several tour guide services such as booklets, local guides, and portable audio devices, but these are mainly used to explain historical information. The MART project's AR-based tour system will offer tourists an intuitive and realistic experience by augmenting 3-D virtual contents to reproduce the way life was lived on real sites.

Gyeongbokgung has many palace and court buildings. They mostly have similar external appearances characterized by Korean traditional architecture. The tracking targets and environments are vari-

ous according to their usage and exhibitions. Exhibition zones also are predefined and partially restricted to access because of conservation. Therefore, this paper presents practical solutions under such tracking environments, as discussed in detail below.

3.1 Tracking Environments

Tracking environments of the main spots of a popular tour route in Gyeongbokgung—Geunjeongjeon and its courtyard, Sajeongjeon, Gangnyeongjeon, and Gyotaejeon were analyzed (see Figure 2).

Geunjeongjeon courtyard: A great courtyard streches out in front of Geunjeongjeon. Visual cues for tracking are difficult to detect because it is covered with hewn stones and has several stone markers. Many tiles of the entrance gate's roof and colonnade have repetitive patterns so that feature points detected from them are not unique. Corners detected from the entrance gate and background buildings can be used for visual tracking (see Figure 2(c)), but the courtyard is generally hard to visually track because it is outdoors, with light changing with time of day, season, and weather. In addition, the courtyard has a changing mask of unpredictably moving and standing people, as shown in Figure 2(b).

Geunjeongjeon and Sajeongjeon: These look alike because their royal functions, such as king's affairs, were similar. The royal thrones of both are centered and surrounded by four pillars, and each background has a picture (see Figure 2(d,e)). Geunjeongjeon has plenty of features, but many have complex and repetitive textures. On the other hand, Sajeongjeon has a hanging picture on the upper side of the throne that can be used as a reference for visual tracking. In both sites, 3-D objects such as the royal throne and four pillars can be used as object models, but their edges are not easily extracted from the cluttered background and the modeling is difficult and delicate. There is little lighting change because both sites are indoor environment and one or more doors are opened for tourists.

Gangnyeongjeon and Gyotaejeon: These buildings offer a glimpse into everyday life in the royal household—Gangnyeongjeon served as the king's living quarters, Gyotaejeon served as the queen's main residence. Rooms in both are opened for tourists so that they can enter and see inside, as shown in Figure 2(f,g). A few feature points can be detected on holding screens in the rooms, but they are not sufficient. As inside exhibitions such as cushion and royal tables (Gangnyeongjeon) and cushions, reading table, and red lacquered mirror stand (Gyotaejeon) have simple shapes, they can be used for object models if their edges can be detected. Doorframes have simple geometric primitives (rectangles) and they can be used as good invariant features for visual tracking. The lighting changes very little. However, the positions of the exhibitions change, so 3-D information of the target scene, used to create scene models, changes too.

3.2 Tracking Using Local Features

As shown above, the tracking environments of Gyeongbokgung do not make it easy to detect and match robust feature points. However, many palace and court buildings consist of geometric primitives such as line segments or rectangles, which are the bases of most man-made structures, so we can use these to track target scenes efficiently. For instance, in Sajeongjeon, the hanging picture on the upper side of the royal throne is rectangular with texture. In Gangnyeongjeon and Gyotaejeon, the doorframes of the main rooms have rectangle structure.

The workflow of rectangle tracking is as follows (also presented in our previous work [Kim et al. 2009]):

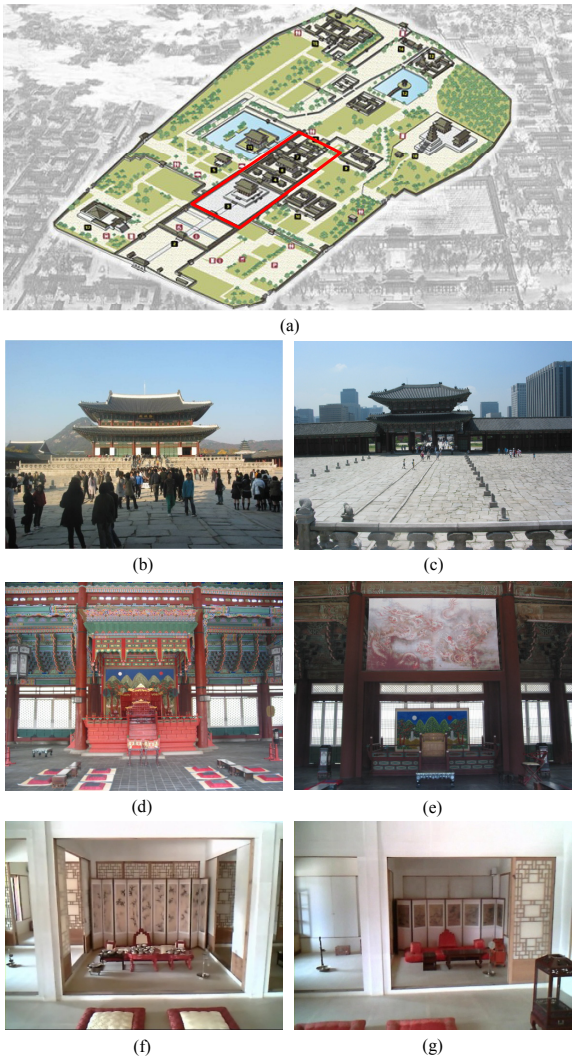


Figure 2: *Tracking environments. (a) tour map of Gyeongbokgung (available at* `http://www.royalpalace.go.kr/`*) (red rectangle indicates the sites covered by the analysis), (b,c) Geunjeongjeon courtyard, (d) Geunjeongjeon, (e) Sajeongjeon, (f) Gangnyeongjeon, (g) Gyotaejeon.*

- Edges of a target scene are detected using the Canny operator.
- Contours which can be candidates for a rectangle in the edges are found. Note that the edges are linked to minimize their discontinuity using dilation.
- The rectangle of the doorframe is extracted by searching the contours with some constraints: convexity, four corners, and the area of the rectangle. When this fails (the rectangle is not extracted), the four corners on the current image can be approximated from the previous ones if the camera's motion is not too fast; thus, we can find the correspondences of the corners obtained from the previous image using optical flow analysis [Bouguet 2000].
- A camera pose is estimated using planar-based pose estimation [Schweighofer and Pinz 2006] with the four corners of the extracted rectangle. Here, we assume that the four corners are coplanar.

Figure 3: *Demonstrations at Sajeongjeon, Gangnyeongjeon, Gyotaejeon, and in Geunjeongjeon courtyard. (a–c) rectangle detection (yellow: rectangle, green: diagonal line of rectangle), (d–f) 3-D virtual character augmentation (king, king with subjects, queen), (g) feature point detection, (h) 3-D virtual character augmentation (royal ritual reproduction), ((d,e,f,h) are snapshot images of animated contents.*

With the estimated poses, 3-D virtual characters are correctly augmented on the target scene as shown in Figure 3—demonstrations at Sajeongjeon, Gangnyeongjeon, and Gyotaejeon. In the demonstrations, the prototype was implemented on a laptop (LG X-NOTE C1, Intel Core2 1.20 GHz) with a USB camera (MS LifeCam NX-6000, resolution 640 by 480, 15 fps).

Geunjeongjeon courtyard is very hard for visual tracking because it is outdoors and local features cannot be detected on the stone-flagged floor. In our approach, we use feature points detected from the entrance gate's roof and its background buildings. As a preprocess, patches centered on feature points are predefined and matched to ones detected from a current image. Camera motion is then estimated using optical flow analysis [Bouguet 2000]. Figure 4 shows a demonstration on Geunjeongjeon courtyard using the hardware devices mentioned above. Note that the demonstration was performed when tourists did not occlude the target scene. The AR service zone was also defined in the service scenario because the target scene covers a very large area of the courtyard; acquisition camera orientation was therefore enough to achieve plausible augmentation.

3.3 Tracking Using Object Models

Visual tracking using feature points is well-established for 2-D planar-based camera tracking. However, it is much limited to robustly track 3-D objects with complex geometry or without texture. There are many exhibitions in the Gyeongbokgung and Angbuilgu (National Treasure No.845) is a good example. It is a kind of sundial, constructed in the shape of a hemisphere to express the shape of the sky. The scale model stands in front of Sajeongjeon as shown in Figure 4.

Figure 4: *On-site exhibition (scale model of Angbuilgu located in front of Sajeongjeon).*

To robustly track a 3-D object, we developed model-based camera tracking that estimates camera poses using both object and feature models. Figure 5 shows the workflow of model-based camera track-

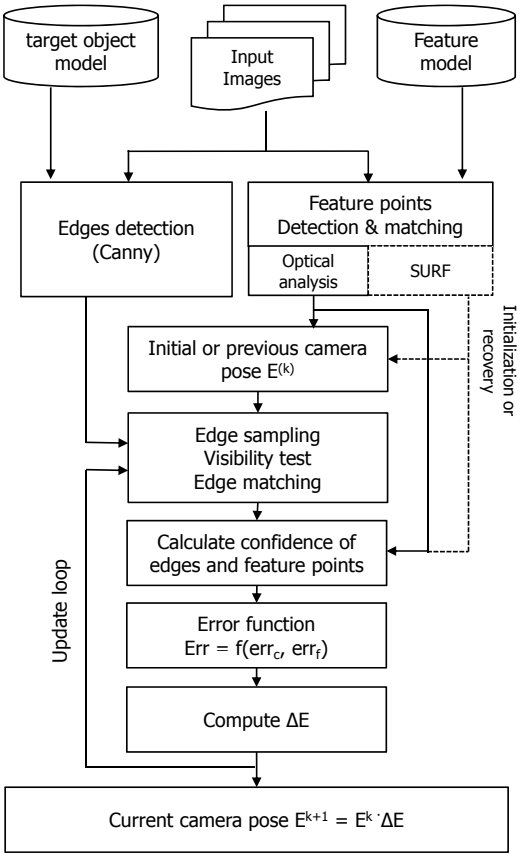


Figure 5: *Workflow of model-based camera tracking.*

ing, but as this is an extended version of our tracking method published in [Park et al. 2010], this paper does not go into full details, just giving a summary.

- A 3-D object and texture information of a target scene are modeled offline.
- With the scene models (object and feature model), edges and feature points are detected in a scene image and matched to correspondences in each model.
  - The edges are detected using the Canny operator. To find their correspondences, the object model is projected, the edges regularly sampled, and the closest edges to their normal direction are defined as the correspondences.
  - The feature points are detected and matched using speeded up robust features (SURF) [Bay et al. 2008]. For computational efficiency, however, they are tracked using optical flow analysis [Bouguet 2000] and SURF is performed for initialization or recovery of a pose.
- To update a camera pose, the camera motion is defined as the displacement of both visual cues and computed by minimizing their reprojection errors. Both measurements are also weighted by their confidences on the target scene.

Tracking results using our model-based camera tracking are shown in Figure 6. In the experiment, a pottery flask was modeled as an object model (wireframe model shown in Figure 6(a)). Given the object model with the feature model, the camera poses were accurately estimated in real-time. Figure 7 shows a demonstration using the scale model of Angbuilgu. The upper part, with a ring shape, was modeled as an object model (wireframe model shown in Figure 7). In the demonstration, the model-based camera tracking was successful even though it was partially occluded by a tourist as shown in Figure 7(e–h).

4 Conclusion

In this paper, we presented a tracking framework for augmented reality-based on-site tours developed from our on-going project. The tracking environments of major sites in Gyeongbokgung were analyzed and more practical visual tracking was successfully applied to the real sites. Promising results were shown through on-site demonstrations.

Currently, we are continuing to test our visual tracking methods and improving their performance as well as implementing the method on the smartphone for mobile augmented reality tour services.

Acknowledgements

This research was supported by Ministry of Culture, Sports and Tourism (MCST) and Korea Creative Content Agency (KOCCA) in the Culture Technology (CT) Research & Development Program 2009 (2nd year).

References

Archeoguide. http://archeoguide.intranet.gr/.

BAY, H., ESS, A., TUYTELAARS, T., AND GOOL, L. V. 2008. SURF: Speeded up robust features. *Computer Vision and Image Understanding 110*, 3, 346–359.

BOUGUET, J.-Y., 2000. Pyramidal implementation of the Lucas Kanade feature tracker description of the algorithm. http://robots.stanford.edu/cs223b04/algo_tracking.pdf.

DRUMMOND, T., AND CIPOLLA, R. 2002. Real-time visual tracking of complex structures. *IEEE Transactions on Pattern Analysis and Machine Intelligence 24*, 7, 932–946.

iTacitus. http://www.itacitus.org/.

KATO, H., AND BILLINGHURST, M. 1999. Marker tracking and HMD calibration for a video-based augmented reality conferencing system. In *International Workshop on Augmented Reality*, 85–94.

KIM, K., SEO, B.-K., HAN, J.-H., AND PARK, J.-I. 2009. Augmented reality tour system for immersive experience of cultural heritage. In *International Conference on Virtual Reality Continuum and its Applications in Industry*, 323–324.

Layar. http://www.layar.com/.

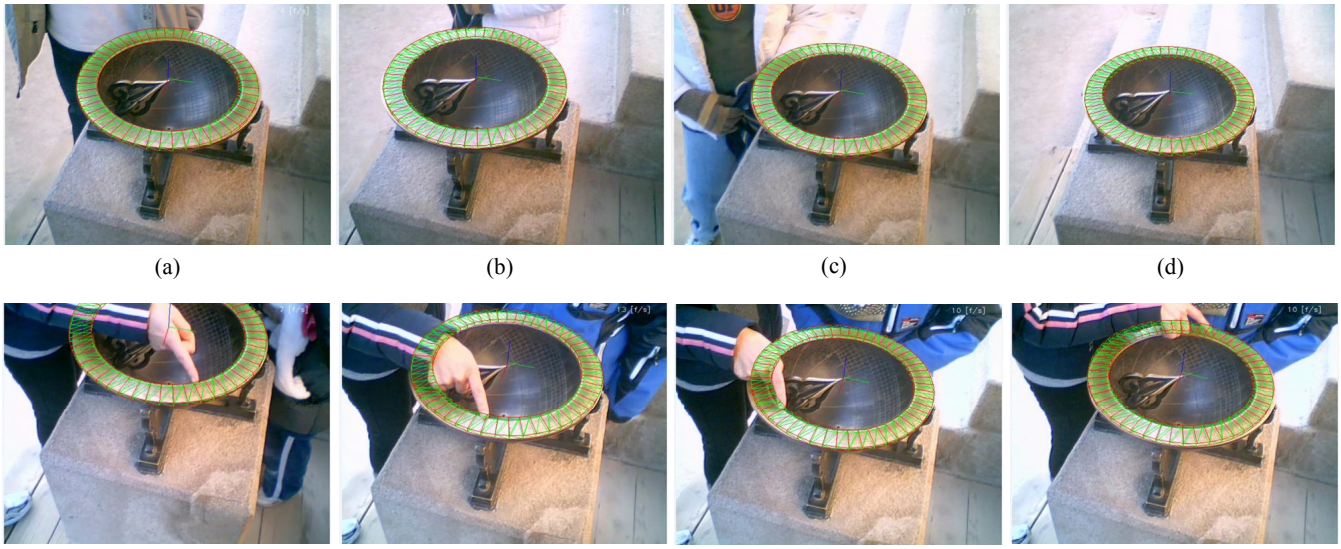
LEPETIT, V., AND FUA, P. 2006. Keypoint recognition using randomized trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 9, 1465–1479.

LOWE, D. 2004. Distinctive image feature from scale-invariant keypoints. *International Journal of Computer Vision 60*, 2, 91–110.

MATAS, J., CHUM, O., URBAN, M., AND PAJDLA, T. 2002. Robust wide baseline stereo from maximally stable extremal regions. In *British Machine Vision Conference*, 384–393.

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

Figure 6: *Model-based camera tracking using a 3-D object (pottery). (a) wireframe model (yellow) overlaid on the real model, (b–c) 3-D virtual character augmentation (some parts of video sequences).*



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

(e) (f) (g) (h)

Figure 7: *Demonstration using an on-site exhibition (scale model of Angbuilgu). Tracking results in (a–d) normal case and (e–h) occlusion case (some parts of video sequences).*

NOH, Z., SUNAR, M. S., AND PAN, Z. 2009. A review on augmented reality for virtual heritage system. *Lecture Note on Computer Science 5670*, 50–61.

PAPAGIANNAKIS, G., SCHERTENLEIB, S., O'KENNEDY, B., AREVALO-POIZAT, M., MAGNENAT-THALMANN, N., STODDART, A., AND THALMANN, D. 2005. Mixing virtual and real scenes in the site of ancient Pompeii. *Computer Animation and Virtual Worlds 16*, 1, 11–24.

PARK, H., OH, J., SEO, B.-K., AND PARK, J.-I. 2010. Automatic confidence adjustment of visual cues in model-based camera tracking. *Computer Animation and Virtual Worlds 21*, 2, 69–79.

ROSTEN, E., AND DRUMMOND, T. 2003. Rapid rendering of apparent contours of implicit surfaces for realtime tracking. In *British Machine Vision Conference*, 719–728.

SCHWEIGHOFER, G., AND PINZ, A. 2006. Robust pose estimation from a planar target. *IEEE Transactions on Pattern Analysis and Machine Intelligence 28*, 12, 2024–2030.

TENMOKU, R., NAKAZATO, Y., ANABUKI, A., KANBARA, M., AND YOKOYA, N. 2004. Nara palace site navigator: Device-independent human navigation using a networked shared database. In *International Conference on Virtual Systems and Multimedia*, 1234–1242.

TUYTELAARS, T., AND MIKOLAJCZYK, K. 2008. Local invariant feature detectors: A survey. *Foundations and Trends in Computer Graphics and Vision 3*, 3, 177–280.

VLAHAKIS, V., IOANNIDIS, N., KARIGIANNIS, J., TSOTROS, M., GOUNARIS, M., STRICKER, D., GLEUE, T., DAEHNE, P., AND ALMEIDA, L. 2002. Archeoguide: An augmented reality guide for archaeological sites. *IEEE Computer Graphics and Applications 22*, 5, 52–60.

Wikitude. `http://www.wikitude.org/`.

ZOELLNER, M., PAGANI, A., PASTARMOV, Y., WUEST, H., AND STRICKER, D. 2008. Reality filtering: A visual time machine in augmented reality. In *International Symposium on Virtual Reality, Archaeology and Cultural Heritage*, 71–77.

ZÖELLNER, M., KEIL, J., DREVENSEK, T., AND WUEST, H. 2009. Cultural heritage layers: Integrating historic media in augmented reality. In *International Conference on Virtual Systems and Multimedia*, 193–196.