Research Article

Dynamic Projection Mapping with a Single IR Camera

Naoki Hashimoto, Ryo Koizumi, and Daisuke Kobayashi

School of Informatics and Engineering, The University of Electro-Communications, W9-603, 1-5-1 Chofugaoka, Chofu, Tokyo 182-8585, Japan

Correspondence should be addressed to Naoki Hashimoto; naoki@cs.uec.ac.jp

Received 23 November 2016; Accepted 23 January 2017; Published 12 February 2017

Academic Editor: Hanqiu Sun

Copyright © 2017 Naoki Hashimoto et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

We propose a dynamic projection mapping system with effective machine-learning and high-speed edge-based object tracking using a single IR camera. The machine-learning techniques are used for precise 3D initial posture estimation from 2D IR images, as a detection process. After the detection, we apply an edge-based tracking process for real-time image projection. In this paper, we implement our proposal and actually achieve dynamic projection mapping. In addition, we evaluate the performance of our proposal through the comparison with a Kinect-based tracking system.

1. Introduction

Dynamic spatial augmented reality (SAR) is expected to provide appearance editing by projecting images onto moving targets. Standard SAR for static objects [1], known as projection mapping, has already achieved realistic augmented reality (AR) and provided fantastic entertainment opportunities. A number of events use this projection technique to entertain people by changing the appearances of real objects such as buildings. As an advanced approach to this technique, dynamic SAR has huge potential to drastically extend the projection effects by supporting various projection targets, such as dancing people and their clothes, and interactions with these targets. This technique can also provide novel forms of amusement for near-future interactive AR games.

However, for precise image projection, the dynamic SAR must obtain the 3D position and posture information of its target in real time. Although previous approaches have used a magnetic sensor, an optical motion capture system or a special high-speed vision and projector system, it is still difficult to provide an accessible system that both is cost-effective and provides accurate projection. For example, current depth sensors are very cost-effective and easy to use but have insufficient sensing accuracy for precise

image projection and have a fatal time delay in measurement.

Therefore, in the present study, we propose a dynamic SAR system that uses a common near IR camera. The quality of current image sensors used in 2D cameras is quite high, and high-speed, low-latency models connected via standard USB3.0 or CameraLink are reasonably priced. The biggest problem with the 2D camera is the limitation that the camera can only capture 2D images. Generally, the estimation of 3D position and posture from 2D images is still a difficult problem. Hence, the proposed system introduces machine-learning and edge-based object tracking techniques to solve this problem and achieves dynamic projection mapping.

2. Previous Research

The most reliable way to obtain the 3D position and rotation of a target object is to use a high-end 3D position sensor, such as a 6DOF magnetic sensor [2, 3]. The magnetic sensor, which provides high-accuracy results, is quite sensitive to metal materials in the surrounding environment and is also expensive. The magnetic sensor also has to be embedded in order to obtain effective SAR expressions. In other words, the sensor itself degrades the AR effect of the target, and

FIGURE 1: Process flow of the proposed dynamic SAR with an IR camera.

noncontact measurement, for example, by a camera is suitable for dynamic SAR.

In a previous study [4, 5], a special high-speed vision and projector system was used to track a target and project images on the target. Although high-quality dynamic SAR was achieved, the target object is limited to a number of simple shapes, and the specialized hardware is not fit for practical use at this time.

Some systems [6–8] use a simple RGB-D camera in order to estimate target position and rotation based on depth information obtained with the RGB-D camera. The RGB-D camera is currently a cost-effective device and is widely used in various types of applications. These systems achieved target tracking at approximately 30 to 45 fps. However, the misalignment of projected images on moving targets is still perceived because of the inaccuracy and the time delay of the RGB-D camera. When used in conjunction with other sensors, such as an inertial measurement unit (IMU), for improving the tracking performance, targets and their application fields are limited by their wiring [9]. Motion prediction is also an effective technique to reduce the misalignment with the time delay [10].

On the other hand, stereopaired cameras remain an effective means to obtain accurate 3D shape and are used for precise 3D position tracking in dynamic SAR [11]. However, the tracking speed is too low to provide true interactions with projected targets.

In addition, an optical motion capture system using high-speed cameras is now widely used and obtaining 3D position and rotation is easy. The biggest problem with motion capture is hiding optical markers, which are small balls wrapped with a retroreflective material, from the gallery of the projection. In a previous study [12], a motion capture was cleverly used

to track the target balls and project images onto them during a juggling act by wrapping the balls with retroreflective material and using them as markers.

3. Dynamic SAR with a Single IR Camera

In the present study, we propose a dynamic SAR system that uses a single near IR camera. Current image sensors used in standard cameras have high accuracy, low noise, and low delay. Cost-effective high-speed cameras that have frame rates of over 100 fps are also widely available. These cameras can easily be used as IR image sensors when fitted with appropriate optical filters. Under additional wide-range IR light source, the IR camera can capture target images without projecting images by projectors. Therefore, the use of the IR camera in dynamic SAR is quite suitable and reasonable.

However, the biggest problem with the single IR camera is the lack of 3D-sensing ability. Although the stereopaired IR camera can capture 3D shapes, the computational cost remains huge and unsuitable for use in dynamic SAR.

Hence, we use a machine-learning technique for detecting the initial position and posture of targets from 2D images. The detected target is then quickly and accurately tracked using the IR camera. This architecture, which is extended of [13], can achieve low delay and precise dynamic SAR using only a single IR camera. An overview of the proposed system is shown in Figure 1.

3.1. 3D Position and Posture Detection Using Random Ferns. The proposed machine-learning algorithm for detecting initial position and posture of targets is based on Hough Forests [14], which is a voting-based object-detection method. Input

FIGURE 2: Training flow with multilayered Random Ferns.

2D images for Hough Forests are divided into small patch images and probabilistic votes are cast for possible locations of the center of gravity (COG) of the entire object based on their classified results using pretrained decision trees. In order to extend Hough Forests to the detection of 3D position and posture, 3D depth and rotation information must be trained simultaneously. However, this extension requires a huge amount of training data and leads to a lack of detection performance with Hough Forests.

Therefore, in the present study, we propose a 3D position and posture detection system using multilayered Random Ferns. Random Ferns method is an optimized implementation of Random Forests, which is a well-known decision-tree-based machine-learning method. The proposed method detects a target's 3D position and posture from a 2D image through a voting process, which is similar to Hough Forests. The first Random Ferns layer detects the COG of the target as one of eight rough directional classes, and the second layer detects the precise 3D position and posture based on the results of the first layer.

Figure 2 shows the training flow of the detection process. Training data are first generated by rendering a 3DCG model from various directions, which are then classified into eight directional classes, as shown in Figure 2. The first layer's Random Ferns are trained to classify input images into the empirically defined eight directional classes and, at the same time, the offset vectors to the COG for each input image patch are counted at the end nodes of the decision trees of the Random Ferns. The first layer outputs a directional class and the approximate COG of the target, which can be used for selecting proper Random Ferns for each class and narrowing the range of the input image patches sent to the second layer's Random Ferns. The second layer has eight Random Ferns, which are well trained for the eight directional classes with proper training data. According to the selected class as a result of the first layer, the best set of the decision trees is selected

from the eight Random Ferns. The offset vector to the COG and rotation parameters (3DOF) are sent to the end nodes of the decision trees and are accumulated. The parameters of the decision trees are optimized by minimizing the entropy of the classified data, as is the case with common detection approaches.

Figure 3 shows the detection process of the proposed method. First, an input IR image is divided into a number of patch images, which are applied to the first layer. At the same time, the input IR image is properly scaled for detecting the depth parameter, and the scaled images are also divided and applied to the first layer. After the first layer, the COG is obtained as the maximum of the 3D voting space with the offset vector (2D) and the scale factor (1D). The directional class is also decided based on the class probability accumulated at the end nodes of the decision trees.

Next, for the second layer, patch images are clipped around the COG detected by the first layer and are finely scaled for precise depth detection. The COG and the scale factor are detected, in the manner described for the first layer, and provide the 3D position (3DOF) of the target. The posture parameter, which is equivalent to the rotation parameter (3DOF), is also decided through the voting process of the rotation parameters accumulated by the trained decision trees, based on the number of classified patches. Thus, the 3D position and rotation of the target are obtained using only the input IR image.

3.2. Edge-Based 3D Tracking. After the initial position and posture detection described in Section 3.1, targets are tracked using the proposed edge-based tracking technique. In this tracking technique, we use a 3D model of the target object. By finding the position and posture for which the edges of the target are finely matched to those of the 3D model, we estimate the present target position and posture in real time.

FIGURE 3: Detecting 3D position and posture.

First, we detect 3D edge points included by the edges of the 3D model, the position and posture of which are estimated using the previous frame. We have to detect the edge points at every frame because the edge points depend on the present position and posture. At the edge of the 3D model, the luminance gradient is generally high. Thus, the edge points are equally sampled from the 3D model based on the luminance gradient. In addition, occluded points are effectively rejected by determining the depth buffer of OpenGL. These processes are shown in Figure 4(a).

Next, we extract 2D edges from an input IR image using the Canny edge detector, as shown in Figure 4(b), and the 3D edge points extracted from the 3D model are projected onto the 2D edge space. The corresponding pairs between the projected 3D edge points and the points on the 2D edges are searched according to the normal direction of each 3D edge, as shown in Figure 5. In order to reduce false detection, target edges are limited to edges having a similar normal.

Next, we calculate the sum of the distances between the corresponded points and use this sum as the error value of the 3D model and the target in the IR image. In order to

FIGURE 4: Detection of 3D edge points and 2D edges.

FIGURE 5: Search for corresponding pairs.

obtain the position and posture of the target, we minimize the error value using the weighted least squares method. If the corresponding pair is correct, the distance of the pair must be quite similar to the distance moved in the nearest frames. Therefore, the weight for each corresponding pair is defined as the similarity ratio between the distance of the corresponding pair and the estimated moving distance of the target in the nearest frames. The updated position and posture of the target, which minimize the error value, are the final result of the estimation using the present IR image, as

shown in Figure 6. If the error value exceeds a predefined threshold, the tracking is stopped and the initial position detection is used to restart the tracking. In addition, this tracking technique is also used for optimizing the result of the initial detection.

4. Experimental Results

We implemented the proposed system and achieved dynamic projection mapping of a real object. As shown in Figure 7, we

FIGURE 6: Edge-based 3D tracking.

FIGURE 7: Experimental environment.

used a mannequin created using a 3D printer. The mannequin was painted with a primer surfacer of plastic models in order to provide a fine reflectance property. The original 3D CAD data for the mannequin were used in the tracking process as a reference 3D model. We used a PC with a CPU (Intel Core i7-4770K, 3.50 Ghz) and a GPU (NVIDIA GeForce GTX Titan Black) for the whole processing. A projector ($1,920 \times 1,080$ pixels, 5,200 lm), an IR camera (640×480 pixels, 337 fps, IR band-pass filter: 850 nm), and squared near IR lights (850 nm, 864 IR LEDs, 34.8 W) for wide-range lighting were placed 1 m from the mannequin. We also used a web camera ($1,920 \times 1,080$ pixels), which was placed near the IR camera, for calibration between the projector and the IR camera. Using the web camera, the intrinsic and extrinsic parameters of the projector and the IR camera system can be well estimated using prior processing.

4.1. Results of Detection. We compared our position and posture detection method with Hough Forests [14]. We used 2,040 positive images and 40 negative images for the

pretraining. The number of decision trees of Hough Forests is 10 and the tree depth is 60. The number of decision trees in the first and second layers is 20. The tree depth of the first layer is 24 and that of the second layer is 23. These parameters were decided based on heuristics.

We generated 100 target images by combining a background image with the 3D mode of the mannequin and the Stanford bunny, the position and posture of which are known, and we applied both detection methods to the target images. The detection rate of the position (x, y, z) allows a difference of within 6 cm from the ground truth. The position indicates the COG of the target. The detection rate of the posture also allows a difference of within 30 degrees for the rotation axis, based on the assumption that the position was detected correctly under the above condition. These loose conditions are applicable because the following tracking process can minimize the remaining differences.

Table 1 lists the detection rates for both methods and models. These results indicate that the position (3DOF) is well detected by both methods and that the position and posture (6DOF) are difficult to estimate correctly using only

TABLE 1: Detection rate.

Target	Detection condition	Hough Forests	Proposed method
Mannequin	Position (3DOF)	95%	96%
	Position & posture (6DOF)	42%	76%
Stanford bunny	Position (3DOF)	90%	97%
	Position & posture (6DOF)	10%	61%

FIGURE 8: Result of position and posture estimation optimized with tracking.

2D images. Although the detection condition is quite loose, the detection ratio with the Hough Forest is 42% and 10% for the mannequin and the Stanford bunny, respectively. On the other hand, the proposed method can improve the position and posture detection to 76% and 61%, respectively. These results obtained using the proposed method can be fully corrected with the following tracking process, as shown in Figure 8.

We implemented both methods with the CPU, and the required time for detection was approximately 19 seconds with Hough Forests and approximately 2 seconds with the proposed method. The Hough Forest required a much greater calculation time compared with the proposed method because the Hough Forest used deep decision trees in order to extract the best detection performance. On the other hand, the proposed method is faster and effective. The proposed method is used for initial detection and redetection when the target position is lost. Thus, the detection time is sufficient for the proposed system. GPU implementation is expected to provide much faster detection.

4.2. Results of Tracking. After the detection process, the edge-based tracking process is applied to the target objects as shown in Figure 9. In this figure, the mannequin is rotated and moved up and down within 1 second by the user's hand, and, according to the position and posture, a face texture is projected onto the plain mannequin. The update period of the tracking is approximately 12.2 milliseconds (82 fps).

This rate is faster than Kinect-based tracking (33 milliseconds (30 fps)) [6] and we can see that the projection gap between the target and the projected image is well reduced, as shown in Figure 10. In Figure 10, the target is moved faster, and the projection error is emphasized.

4.3. Evaluation of the Tracked Trajectory. We compared the tracked trajectories of the proposed system and the Kinect-based system. We moved the mannequin by the user's hand in a similar manner and recorded the tracked trajectories. At the same time, we also measured the grand truth of the trajectories using an optical position tracker (OptiTrack Flex3). The motion of the mannequin is the combination of simple translation (x -axis (right) and y -axis (top)) and rotation (z -axis (depth)). The tracking speed is uniformed to that of the Kinect-based system (30 fps). The results are shown in Figures 11 and 12. The average and standard deviation of the errors are shown in Figure 13.

In Figure 11, we can find that the results of x -axis and y -axis with the proposed system are better than those with the Kinect-based system. On the other hand, the results of z -axis obtained using the proposed system are unstable and have numerous errors because the changes of z -axis are difficult to capture in the 2D IR images used by the proposed system. Then, small errors in tracking based on the 2D images become large z -axis errors in the 3D real space. In contrast, the Kinect system can measure depth values directly and track the target correctly according to z -axis. However, this

FIGURE 9: Results of dynamic SAR for a moving mannequin.

(a) Kinect-based system

(b) Our proposed system

FIGURE 10: Projection errors with faster movement.

FIGURE 11: Tracked trajectories (position).

FIGURE 12: Tracked trajectories (rotation).

FIGURE 13: Average error of tracked trajectories.

effect is not so remarkable in the actual projection mapping, as shown in Figure 9, because the effect of the depth change on the appearance is also small. Figure 12 shows that the results obtained using the proposed system are better than those of the Kinect-based system.

Based on this evaluation, the proposed system has better tracking performance. In the actual projection, the proposed system can track approximately three times faster than the Kinect-based system. This is an important factor for precise projection mapping because the target is continuously

moving during the update period of the tracking. Therefore, the performance gap in Figures 11 and 12 will be much more notable in the actual projection mapping, as shown in Figure 10.

5. Conclusion and Future Research

We proposed dynamic SAR with the detection of multilayered Random Ferns and edge-based tracking using an IR camera. The proposed method can provide fine dynamic projection mapping for a target object being moved by the user's hand. Evaluations revealed that the proposed system exhibited better performance than the Kinect-based system. In the near-future, we intend to evaluate the proposed system using various targets subjected to various motions. We also intend to incorporate a sophisticated motion prediction method with the proposed precise, high-speed tracking method in order to reduce the potential delay of projectors, which will enable high-quality dynamic SAR using off-the-shelf, cost-effective projectors and will extend the applicability of dynamic SAR.

Competing Interests

The authors declare that they have no competing interests.

Acknowledgments

The present study was supported by JSPS KAKENHI Grant no. 16K00267.

References

- [1] R. Raskar, G. Welch, M. Cutts, A. Lake, L. Stesin, and H. Fuchs, "The office of the future: a unified approach to image-based modeling and spatially immersive displays," in *Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '98)*, pp. 179–188, Orlando, Fla, USA, July 1998.
- [2] D. Bandyopadhyay, R. Raskar, and H. Fuchs, "Dynamic shader lamps: painting on movable objects," in *Proceedings of the IEEE and ACM International Symposium on Augmented Reality (ISAR '01)*, pp. 207–216, New York, NY, USA, October 2001.
- [3] D. Kondo and R. Kijima, "Proposal of a free form projection display using the principle of duality rendering," in *Proceedings of the 8th International Conference on VSM, pp. 346–352*, September 2002.
- [4] K. Okumura, H. Oku, and M. Ishikawa, "Active projection AR using high-speed optical axis control and appearance estimation algorithm," in *Proceedings of the IEEE International Conference on Multimedia and Expo (ICME '13)*, pp. 1–6, IEEE, San Jose, Calif, USA, July 2013.
- [5] T. Senoo, Y. Yamakawa, Y. Watanabe, H. Oku, and M. Ishikawa, "High-speed vision and its application systems," *Journal of Robotics and Mechatronics*, vol. 26, no. 3, pp. 287–301, 2014.
- [6] K. Tsuboi, Y. Oyamada, M. Sugimoto, and H. Saito, "3D object surface tracking using partial shape templates trained from a depth camera for spatial augmented reality environments," in *Proceedings of the 14th Australasian User Interface Conference (AUI '13)*, vol. 139, pp. 125–127, Adelaide, Australia, 2013.
- [7] D. Kobayashi and N. Hashimoto, "Spatial augmented reality by using depth-based object tracking," in *Proceedings of the ACM Special Interest Group on Computer Graphics and Interactive Techniques Conference (SIGGRAPH '14)*, Vancouver, Canada, August 2014.
- [8] R. Koizumi, D. Kobayashi, and N. Hashimoto, "Acceleration of dynamic spatial augmented reality system with a depth camera," in *Proceedings of the International Conference on Cyberworlds (CW '15)*, pp. 50–53, October 2015.
- [9] Y. Zhou, S. Xiao, N. Tang, Z. Wei, and X. Chen, "Pmomo: projection mapping on movable 3D object," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, pp. 781–790, Santa Clara, Calif, USA, May 2016.
- [10] J. Knibbe, H. Benko, and A. D. Wilson, "Juggling the effects of latency: software approaches to minimizing latency in dynamic projector-camera systems," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology (UIST '15)*, pp. 93–94, ACM, Daegu, South Korea, November 2015.
- [11] C. Resch, P. Keitler, and G. Klinker, "Sticky projections—a model-based approach to interactive shader lamps tracking," *IEEE Transactions on Visualization and Computer Graphics*, vol. 22, no. 3, pp. 1291–1301, 2016.
- [12] S. Miyafuji, M. Sugasaka, and H. Koike, "Ballumiere: real-time tracking and projection for high-speed moving balls," in *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pp. 101–102, Tokyo, Japan, October 2016.
- [13] N. Hashimoto and D. Kobayashi, "Dynamic spatial augmented reality with a single IR camera," in *Proceedings of the ACM SIGGRAPH*, Anaheim, Calif, USA, July 2016.
- [14] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 11, pp. 2188–2202, 2011.