Augmented Reality based 3D Human Hands Tracking from Monocular True Images Using Convolutional Neural Network

Saif, A F M Saifuddin

IGI Global
2023

Saif , A F M S & Mahayuddin , Z R 2023 , Augmented Reality based 3D Human Hands Tracking from Monocular True Images Using Convolutional Neural Network . in Augmented Reality based 3D Human Hands Tracking from Monocular True Images Using Convolutional Neural Network . IGI Global , pp. 129-137 . https://doi.org/10.4018/978-1-6684-5849-5.ch008

http://hdl.handle.net/10138/354271
https://doi.org/10.4018/978-1-6684-5849-5.ch008

unspecified
acceptedVersion

Augmented Reality based 3D Human Hands Tracking from Monocular True Images Using Convolutional Neural Network

A F M SAIFUDDIN SAIF*[a], ZAINAL RASYID MAHAYUDDIN [b]

[a] *Faculty of Science, University of Helsinki, Finland*
[b] *Faculty of Information Science and Technology, University Kebangsaan Malaysia, Selangor, Malaysia*

Abstract

Precise modeling of hand tracking from monocular moving camera calibration parameters using semantic cues is an active area of research concern for the researchers due to lack of accuracy and computational overheads. In this context, deep learning based framework, i.e. convolutional neural network based human hands tracking as well as recognizing pose of hands in the current camera frame become active research problem. In addition, tracking based on monocular camera needs to be addressed due to updated technology such as Unity3D engine and other related augmented reality plugins. This research aims to track human hands in continuous frame by using the tracked points to draw 3D model of the hands as an overlay in the original tracked image. In the proposed methodology, Unity3D environment was used for localizing hand object in augmented reality (AR). Later, convolutional neural network was used to detect hand palm and hand keypoints based on cropped region of interest (ROI). Proposed method by this research achieved accuracy rate of 99.2% where single monocular true images were used for tracking. Experimental validation shows the efficiency of the proposed methodology.

Keywords: 3D tracking, Augmented Reality, Convolutional Neural Network. 3D Localization

1. INTRODUCTION

Human hands are used to interact with devices as a medium of human computer interaction in a context sensitive way to the user for the intention to superimposes the visual perception. Mobile phone touch screens constrain in a small space of the device where interaction from 2D surface to 3D surface is a fertile research problem in the area of augmented reality. In this context, in order to manipulate virtual objects without previous knowledge interaction of different techniques and knowledge is prime area of investigation to emerge augmented reality research domain. This research aims to broaden the interaction from two dimensional to three-dimensional space in the context of using small space mobile device by proposing vision based three dimensional hand gesture tracking method for augmented reality.

Movement of camera makes the tasks difficult for tracking hand under augmented reality where traditional image processing techniques were previously applied by existing research where computational manipulation was a big concern. In this context, deep learning framework such as convolutional neural network comes in handy to deal with camera movement type of difficult tasks specially for modeling camera calibration parameters. Precise modeling of moving objects and moving camera parameters for analyzing semantic cues with the usage of convolutional neural network can be considered as strong bridge for 3D hands tracking for augmented reality platform. Purpose of this research is to track human hands from monocular true images to interact with virtual objects using augmented reality.

The remainder of this paper is organized as follows. Critical previous research is illustrated in previous study section, details of the proposed methodology is elaborated in proposed method section, details experimental results with analysis for validation is presented in experimental results and analysis section and finally conclusion section presents concluding remarks.

2. PREVIOUS STUDY

Development of own libraries for augmented reality, placement of virtual 3D objects in order to check the suitability in the virtual environment turned lots of active research interest in the area of augmented reality in the broader aspects of computer vision and deep learning specially convolutional neural network (CNN) (Tanzi et al., 2021; Su et al., 2021; Perdpunya et al., 2021; Lan, 2021; Saif and Mahayuddin, 2020). Earlier of this research trends, recognition

of hands and tracking them were done by offline image processing causes huge computational overheads. After that, human hands tracking in real time was progressed using web camera and personal computer. Later, convolutional neural network, features tracking and modeling the network made the overall process easier for human hand tracking. At the current progress of research progress, human hands tracking via mobile phone camera as well as recognizing the pose of the hands in the current camera frame becomes active research problem. This research aims to track human hands pose in continuous frame by using the tracked points to draw a 3D model of the hands as an overlay in the original tracked imaged. In addition, this research also used the points to generate colliders and effectors to interact with virtual objects. Area of previous research methods is shown in Fig.1.

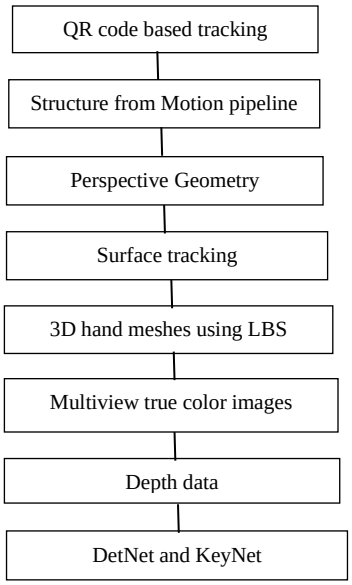

Fig. 1. Hierarchy of previous research methods for hand tracking

In the recent years, object tracking based on camera becomes suitable due to updated technology such as Unity3D engine and Vuforia augmented reality plugin (Nazar et al., 2020). QR code based object tracking is also fascinated by research due to quick and user friendly implementation and validation, however, tracking by this plat is considered old now a days (Auliya and Munasiah, 2019) comparing with Unity3D engine and Vuforia augmented reality plugin. Structure from Motion pipeline (SFM), Perspective Geometry and Surface tracking are also considered as another subset of augmented reality(Nakashima et al., 2015; Samini and Palmerius, 2014; Nakashima et al., 2013; Wang et al.,2005). Structure from Motion pipeline (SFM) was prepared by researchers due to advanced computing systems and ease of the process to construct and visualize point clouds which causes significant improvements in visual quality although computational overheads was a big concern for SFM(Samini and Palmerius, 2014; Huang, T. and Liu, 2019). Perspective geometry by Nakashima et al.(2013) is another option for the researchers, however, due to complex and dependency on external hardware perspective geometry based augmented reality can not be standalone solution for long term research goal. In addition, prominent errors were observed in case of video rendering through augmentation tasks cases huge computation cost. Background and foreground subtraction based surface tracking (Wang et al., 2005) by planar surface was also used by existing researchers which could not provide satisfactory validation for good frame rate per second where lot of improvements were done using deep learning architechures.

Hand shape and pose estimation from depth images was used by researchers convolutional neural networks was implicated to recover 3D hand meshes using LBS(Ge et al., 2019). In this context, true color image sequences were used to estimate hand pose by combining multiview true color images and depth data (Qian et al., 2014). Development of hand gestures datasets were also in consideration to recognize hand shape in the wild where trained model did not perform well when tested on different datasets due to lack of generalization of training data (Zimmermann et al., 2019) which provides the clue that generalization of datasets for hand gesture tracking can be another prime reason for significant improvement in indoors and outdoors environment. In this context, convolutional neural networks can play the vital role for hand pose estimation (Doosti et al., 2019; Yasen and Jusoh, 2019) in lieu with depth map data. Han et al. (2020) performed for hand tracking in virtual reality by using depth-based approach

and DetNet to detect hands followed by KeyNet to predict key-points in hand from the cropped image based on the bounding box provided by DetNet. However, hand scale and distance issue still need robust solution by this research. Zhang et al. (2020) used MediaPipe pipeline for tracking hand and hand gestures by using true color image. However, availability of data imposed by their research needs further trustworthiness to be validated by existing research.

3. PROPOSED RESEARCH METHODOLOGY

Handheld mobile device now a days occupies the place of using heavy hardwares and sensors with the advancement of huge computer processing power, although improvement in terms with achieving high user experience satisfaction was not achieved parallelly. In that context, hand palm tracking is addressed in this research with the usage of augmented reality.

Multiple level of abstractions were set up so that the overall methodology works properly in order to make the proposed method reusable and reproducible. Overall proposed method are categorized into three main steps, i.e. localization of object in augmented reality, palm detection, keypoint detection mentioned in Fig. 2.

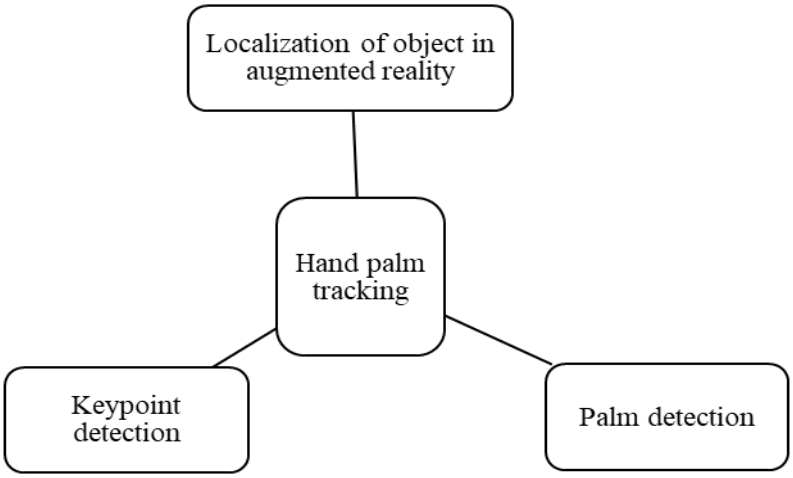

Fig. 2. Key steps for the overall proposed methodology.

Unity3D environment was used for localizing object in augmented reality which is a cross platform. After that, ARCore solution was used to integrate augmented reality for mobile phone in order to make the surface tracking more convenient and to employ the sensor fusion to track planar surfaces which is widely known as Concurrent Odometry and Mapping (COM) which is under US patent. OCM provides image data using mobile phone camera by fusing and correcting them. In addition, ARCore, this research receives details of plane to localize the object, then Unity track the plane on order to render the chosen 3D object on the top of that plane. This research used MediaPipe for detecting hand palm and hand keypoint based on cropped region of interest (ROI). BlazeFace (Bazarevsky et al., 2019) is used for palm detection works on mobile GPU where the major impact of BlazeFace (Bazarevsky et al., 2019) as influential convolutional neural network architechures uses 3x3 kernels based convolution. Based on the increament of kernel size computational cost becomes cheaper comparing with CPU. Feature extraction was performed using 2D convolution uses RGB images as input. Palm detection is followed over the whole results in getting cropped region of the hand palm which is forwarded towards the next step of detecting hand key points from cropped image region.

After that precise localization of hand landmark was performed inside the detected region via regression which consists of three resultant parts, i.e. presence of hand in the input image was indicated by hand flag, left or right handedness classification as binary classification, localization of hand landmarks consisting of x, y and corresponding depth. 3D Coordinate points projection are then supplied in Unity which provides components for physical actions and simulations. Later, collision avoidance was done on the detected hand key points from previous step in order to use them for interaction with 3D object placed in augmented reality.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This research performed experimentation in Intel Core i7 3.17 Ghz CPU having 60 GB RAM. This research used MediaPipe by importing necessary package depending on OpenCV framework which requires that OpenCV should be installed in the machine. Experimentation were done on multiple datasets and results are recorded. Firstly datasets from PoseNet is used to train and test the model. This research used PoseNet because PoseNet uses the same research architechures as this research used for validation. This research also experimented another datasets from Graph Convolutional Neural Network (Ge et al., 2019) where 3D mesh of the same hand pose is created and matched to train the model. This research used various performance metrics to validate the proposed mythology, i.e. detection rate (Saif et al., 2015), map (Saif et al., 2014), f-measure(Saif et al., 2021), CT(Saif & Mahayuddin, 2020), processing speed in frame per second show in Fig. 3.

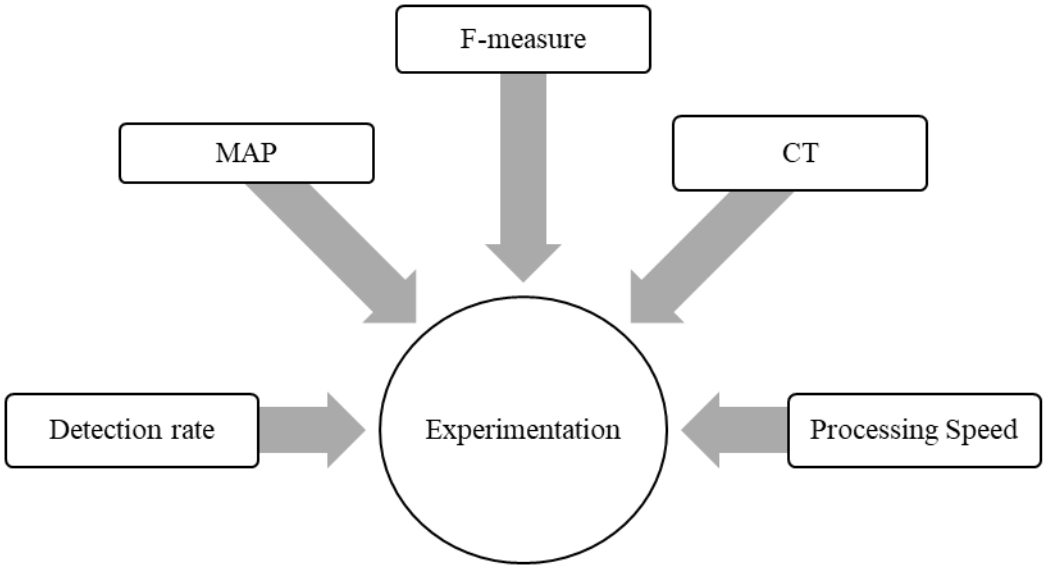

Fig. 3. Performance metrics used for validation.

This research achieved detection rate of nearly 99.2% which is almost closed by Zhang et al. (2020) due to usage of same deep learning platform using MediaPipe where single RGB camera was used for tracking. However, due to sensor fusion for tracking, this research receives lower F-measure of 65.2% comparing Zhang et al. (2020). Zhang et al. (2016) receives lower performance comparing with the proposed research in terms with almost all the performance metric comparing with the proposed method due to stereo vision complexity for camera calibration. Zimmermann et al.(2017) shown lowest results among all the research compared with the proposed method due to high volume of network parameter estimation rather than robust calibration of deep learning architechures. Details results of the proposed method and comparison with existing research results are mentioned in the table 1.

Table 1: Results and comparison with existing research results.

Methods	Detection Rate	MAP	F- Measure	CT(ms)	PS in Frame per Second (Performance Speed)
Proposed Method	>99.2%	~97.2%	~65.2%	3	40+ (CPU)
MediaPipe Hands (Zhang et al., 2020)	>99%	95.7%	94.45%	6.6	200-1000+ (GPU)
Stereo-based Hand Tracking (Zhang et al., 2016))	>80%	82.03%	82.73%	16.1	300 (CPU)
PosePrior network (Zimmermann et al., 2017))	>60%	65.8%	67.8%	36.9	18.5 (CPU)

5. CONSLUSION

This research performed human hand tracking using monocular true images to establish interaction with virtual objects using augmented reality. To achieve the objective, multiple level of abstractions were done, i.e. localization of objects in augmented reality, palm detection followed by keypoints detection in order to make the overall methodology reusable and reproducible. Cross platform Unity3D environment was used to localize object in augmented reality. Later, ARCore solution was used to embed augmented reality for mobile device to make surface tracking more convenient. After that regression was performed for precise localization of hand landmark inside the detected region. Then, collision avoidance was done on the detected hand key points from previous phase to use them for interacting with 3D object placed in augmented reality. Accuracy rate of 99.2% was achieved by this research where single RGB camera was used for tracking. In future, this research aims to integrate advanced CNN frameworks to localize the object more precisely. Performance of the proposed method is expected to contribute significantly for robotics research in the context of fourth industrial revolution.

ACKNOWLEDGEMENTS

The authors would like to thank Universiti Kebangsaan Malaysia for providing financial support under the "Geran Universiti Penyelidikan" research grant, GUP-2020-064.

REFERENCES

Auliya, R., & Munasiah, M. (2019). *Mathematics learning instrument using augmented reality for learning 3D geometry.* Paper presented at the Journal of Physics: Conference Series.

Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., & Grundmann, M. (2019). Blazeface: Sub-millisecond neural face detection on mobile gpus. *arXiv preprint arXiv:1907.05047*.

Doosti, B. (2019). Hand pose estimation: A survey. *arXiv preprint arXiv:1903.01013*.

Ge, L., Ren, Z., Li, Y., Xue, Z., Wang, Y., Cai, J., & Yuan, J. (2019). *3d hand shape and pose estimation from a single rgb image.* Paper presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.

Han, S., Liu, B., Cabezas, R., Twigg, C. D., Zhang, P., Petkau, J., . . . Wang, Z. (2020). MEgATrack: monochrome egocentric articulated hand-tracking for virtual reality. *ACM Trans. Graph., 39*(4), 87.

Huang, T., & Liu, Y. (2019). *3d point cloud geometry compression on deep learning.* Paper presented at the Proceedings of the 27th ACM International Conference on Multimedia.

Lan, E. S. (2021). A Novel Deep ML Architecture by Integrating Visual Simultaneous Localization and Mapping (vSLAM) into Mask R-CNN for Real-time Surgical Video Analysis. *arXiv preprint arXiv:2103.16847*.

Nakashima, Y., Sato, T., Uno, Y., Yokoya, N., & Kawai, N. (2013). *Augmented reality image generation with virtualized real objects using view-dependent texture and geometry.* Paper presented at the 2013 IEEE International Symposium on Mixed and Augmented Reality (ISMAR).

Nakashima, Y., Uno, Y., Kawai, N., Sato, T., & Yokoya, N. (2015). AR image generation using view-dependent geometry modification and texture mapping. *Virtual Reality, 19*(2), 83-94.

Nazar, M., Aisyi, R., Rahmayani, R., Hanum, L., Rusman, R., Puspita, K., & Hidayat, M. (2020). *Development of augmented reality application for learning the concept of molecular geometry.* Paper presented at the Journal of Physics: Conference Series.

Perdpunya, T., Nuchitprasitchai, S., & Boonrawd, P. (2021). *Augmented Reality with Mask R-CNN (ARR-CNN) inspection for Intelligent Manufacturing.* 12th International Conference on Advances in Information Technology, pp. 1-7. 2021.

Qian, C., Sun, X., Wei, Y., Tang, X., & Sun, J. (2014). *Realtime and robust hand tracking from depth.* Paper presented at the Proceedings of the IEEE conference on computer vision and pattern recognition.

Saif, A., Prabuwono, A. S., & Mahayuddin, Z. R. (2014). Moving object detection using dynamic motion modelling from UAV aerial images. *The Scientific World Journal, 2014*.

Saif, A. F. M. S., Mahayuddin, Z. R., & Shapi'i, A. (2021). Augmented Reality based Adaptive and Collaborative Learning Methods for Improved Primary Education Towards Fourth Industrial Revolution (IR 4.0). *International Journal of Advanced Computer Science and Applications, 12*(6).

Saif, A. S., & Mahayuddin, Z. R. (2020). Moment Features based Violence Action Detection using Optical Flow. *International Journal of Advanced Computer Science and Applications(IJACSA), 11*(11).

Saif, A. S., & Mahayuddin, Z. R. (2020). Robust Drowsiness Detection for Vehicle Driver using Deep Convolutional Neural Network. *International Journal of Advanced Computer Science and Applications, 11*(10).

Saif, A. S., Prabuwono, A. S., & Mahayuddin, Z. R. (2015). Moment feature based fast feature extraction algorithm for moving object detection using aerial images. *PloS one, 10*(6), e0126212.

Samini, A., & Palmerius, K. L. (2014). *A perspective geometry approach to user-perspective rendering in hand-held video see-through augmented reality.* Paper presented at the Proceedings of the 20th ACM Symposium on Virtual Reality Software and Technology.

Su, M.-C., Chen, J.-H., Azzizi, V. T., Chang, H.-L., & Wei, H.-H. (2021). Smart training: Mask R-CNN oriented approach. *Expert Systems with Applications, 185*, 115595.

Tanzi, L., Piazzolla, P., Porpiglia, F., & Vezzetti, E. (2021). Real-time deep learning semantic segmentation during intra-operative surgery for 3D augmented reality assistance. *International Journal of Computer Assisted Radiology and Surgery*, 1-11.

Wang, H. L., Sengupta, K., Kumar, P., & Sharma, R. (2005). Occlusion handling in augmented reality using background-foreground segmentation and projective geometry. *Presence, 14*(3), 264-277.
Wang, H. L., Sengupta, K., Kumar, P., & Sharma, R. (2005). Occlusion handling in augmented reality using background-foreground segmentation and projective geometry. *Presence, 14*(3), 264-277.
Yasen, M., & Jusoh, S. (2019). A systematic review on hand gesture recognition techniques, challenges and applications. *PeerJ Computer Science, 5*, e218.
Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C.-L., & Grundmann, M. (2020). Mediapipe hands: On-device real-time hand tracking. *arXiv preprint arXiv:2006.10214*.
Zhang, J., Jiao, J., Chen, M., Qu, L., Xu, X., & Yang, Q. (2016). 3d hand pose tracking and estimation using stereo matching. *arXiv preprint arXiv:1610.07214*.
Zimmermann, C., Ceylan, D., Yang, J., Russell, B., Argus, M., & Brox, T. (2019). *Freihand: A dataset for markerless capture of hand pose and shape from single rgb images.* Paper presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision.