When Augment Reality meets Deep Learning: a survey

Haojie He, Lingming Zhang, Longqian Huang

February 2021

1 Introduction

Augmented Reality (AR) is a technology developed since 1960s. It incorporates digital data perceived from users in real world and creates a mixed reality where both real and virtual objects exist. AR exploits the capabilities of desktop and mobile computing systems and allows users to see and interact with digitally generated objects which are projected into the physical environment.[1]

There are lots of research areas in AR technology. According to [2], they can be divided into three parts:

- Tracking Technique
- Interaction Technique
- Display Technique

Tracking technique aims to keep the AR system in track with target objects, thus acquires transformation parameters, enabling cameras and virtual objects to adjust themselves in real-time. Fast Tracking guarantees the interaction and display techniques to perform fluently, leaving the difference between real and virtual objects unnoticed, thus realizing 'augment'.

Interaction technique could endow meanings to virtual objects. With the interaction technique, people could manipulate the virtual object as in reality. *Hand pose estimation* is an indispensable part of interaction technique, which allows AR devices to track and sense user's hand movement so that the AR system could respond to user's hand instructions intelligently.

Display technique is essential to make virtual objects more like a real one. Illumination estimation could give the displayed virtual objects a suitable shadow given a specific light environment so that the object looks more like a real one.

Deep Learning technology utilizes artificial neural networks to extract feature representations from a large amount of data. It empowers the computer to accomplish complicated computer-vision tasks and therefore a great tool for AR. Equipped with deep learning technology, tracking, interaction and display techniques could be improved in both quality and speed, which greatly promote the development of AR.

This survey will focus on the three components of AR, especially when these techniques meet deep learning technology. We will introduce our survey in these research regions. For tracking technology, image feature extraction and matching, as well as camera pose estimation techniques are demonstrated in section 2. For interaction part, we briefly describe some developments of *hand pose estimation* in section 3. For display part, we discuss how deep learning can promote *Illumination Estimation* technology in section 4. We will also discuss some applications of AR briefly in section 5.

2 Tracking Technology

Before 2007, tracking methods could be divided into three parts: sensor-based, vision-based (RGB data) and hybrid. After then, diverse vision-based methods are developed with RGB-D data, simultaneous localization and mapping (SLAM) technology.[3] Since 2016, many deep learning–based methods are proposed, making tracking techniques more robust.

Traditional mainstream methods of tracking are conducted on a two-stage manner: a feature extraction stage and a tracking (feature matching) stage. Based on the two stages, a camera pose is finally estimated. However, these methods often lose track due to motion blur, quick rotations or partial occlusions, etc. Coupled with their complex and expensive pipelines, these approaches are not suitable for many real-world scenarios.

To alleviate these problems, some newly introduced methods used deep learning to refine feature extraction and matching stage, others consider in a new way, which learned end-to-end camera pose using neural networks.

In this section, we will first describe our survey in feature extraction and matching techniques. Then we also discuss how deep learning technology can be utilized for end-to-end camera pose tracking.

2.1 Feature Extraction and Matching

For real-time tracking, feature extraction is often the first step. The part that performs feature extraction function in an AR system is also called a *detector*. AR system needs to know representative feature of the object and keep in track of these features. Robust detector for an AR system is significant since informative and robust feature extraction could make feature matching in the next process easier, so that the system may less likely lose track of the object.

Traditional feature extraction methods are usually based on existing computer vision algorithms, such as SIFT[4] and ORB[5]. However, in some specific environments like scenes with weak texture, these methods tend to fail. With the help of deep learning, quality of the extracted features is improved. In 2016, Akgul *et al.*[6] proposed DeepAR which utilizes CNN to do feature extraction, achieving better performance than ORB. In 2018, DeTone *et al.*[7] designed SuperPoint for feature extraction.

When detector obtained robust features from the object, feature matching techniques will be applied. For AR tracking, speed is an essential aspect for assessing tracking performance and real-time feature matching is needed.

Feature Matching are usually combined with feature extraction algorithms. In 2017, DeTone *et al.* proposed MagicWarp and MagicPoint for feature extraction and matching[8]. In 2020, Sarlin *et al.* proposed SuperGlue, which imported Graph Neural Network(GNN) into feature matching task. They got real-time matching performance and it is promising to be applied in AR tracking.

2.2 End-to-end Camera Pose Estimation

Camera pose estimation methods (or, vision localization technologies) could be divided into two classes: *Topological Localization* and *Metric Localization*. Topological localization models the problem as image retrieval and metric localization utilizes the feature extraction and matching paradigm. CNN based end-to-end camera pose estimation methods, in essence, belongs to the first category [9].

In 2016, PoseNet[10] is first proposed to directly regress the 6-DoF camera pose from a single RGB image with GoogLeNet. Taking advantage of transfer learning, it is lightweight, scalable and robust. But it is less accurate than traditional image retrieval based methods. On the basis of PoseNet, Kendall *et al.* [11, 12] make the weight between camera translation and rotation loss self-adaptive and introduce the geometric reprojection loss. Melekhov *et al.* [13] improved PoseNet architecture by introducing skip connections with ResNet34 architecture, whereas Walch *et al.* and Clark *et al.* extended PoseNet using LSTM cells to better exploit the spatial [14] and temporal [15] information in an image flow (e.g. videos). To make better use of datasets, Sattler [16] explored how the Field-of-View, data augmentation and LSTM cells can benefit the network performances.

To incorporate the cheap and ubiquitous sensory inputs into the end-to-end camera pose estimation system, Brahmbhatt [17] *et al.* proposed MapNet in 2018. This work formulated the sensory inputs as loss terms in both training and inference stages to allow the network to update in a self-supervised manner. Observing that the network predictions are locally noisy but drift-free while the Visual Odometry(VO) information from sensory inputs is locally smooth but drifty, MapNet fuses these two into a moving window fashion with pose graph optimization(PGO) [18, 19, 20] to exploit the complementary noise characteristics. In addition, MapNet introduced the logarithm of unit quaternion as a parameterization for rotation, which seems to be better suited for deep-learning based camera pose regression.

Adversarial networks are introduced into camera pose estimation area in 2019 by Bui *et al.* [21]. Besides the pose regression network, a discriminator is trained to distinguish between regressed and ground truth poses given the embedding of an input image. Bui *et al.* further refines the regressed pose during an inference stage by making use of the trained discriminator, showing the discriminator actually learns a meaningful representation of the camera pose and image space. This architecture achieved a slightly better performance than MapNet on the 7-Scenes dataset given that only RGB images are provided.

Radwan *et al.* [22] developd a deep multitask learning framework for semantic visual localization and odometry. The framework, which is called VLocNet++, consists of a localization network, an odometry network and a semantic segmentation network. Some layers of these networks are shared and fused in a self-adaptive manner to synergistically utilize information from multiple datasets. Temporal information was exploited by warping and fusing specific layer weights of the network when the previous frame is taken into the current one as network input. Benefited from the multitask learning framework, VLocNet++ achieved the state-of-the-art on both deep learning–based visual localization and semantic segmentation at the time.

Sattler *et al.* [9] pointed out that CNN based pose regression networks mathematically follow the paradigm of image retrieval methods. As long as the last layer of the network is linear, the outputs of the network is a linear combination of the parameters of the last layer. The parameters of the last layer thus form base of the output space. Then the layers before the last layer can be regarded as computing the similarity of the input image and some implicitly encoded base images. In that view, neural networks encode the training set and serve as the function to calculate image similarity. Therefore neural network–based methods, in general, could be seen as an upgraded version of the image retrieval method. Some specially designed experiments are carried out to prove that MapNet and PoseNet cannot consistently outperform traditional image retrieval methods. Thus they show that deep learning–based end-to-end camera pose estimation methods have a long way to go.

3 Hand Pose Estimation in Interaction Technique

Hand tracking is a major problem in interaction technique development. Hand pose need to be estimated when performing hand tracking. Here we only introduce some development of vision-based hand pose estimation techniques. For basic hand pose estimation problem formulation, refer to [23] for an overview. Researchers tend to explore different type of images and use deep learning to make full use of them. In [24], single RGB images are put into a CNN and obtain a 3D hand pose. In [25], depth images were used to estimate hand pose, which were processed to generate 3D point cloud and put into a neural network which outputs a hand pose. In [26], combined utilization of depth image and point cloud were explored. They were put into a 2D CNN and a 3D CNN respectively to make hand pose estimation.

However, blurred depth image caused by fast hand movement is a critical problem in hand pose estimation and could not be settled well. In [27], infrared images were explored and were reported nice performance with fast hand movement. To utilize previous techniques with depth image, domain transfer learning was used to transform unblurred infrared image to depth and then use neural network to process depth image and obtain a hand pose.

4 Illumination Estimation

Visual coherence is a key requirement in AR. As one of the most important factors, illumination consistency is typically achieved by two steps: real lighting estimation and virtual object rendering. Since the acquisition of real-world lighting directly affects the rendering result of virtual objects, much efforts have been made towards illumination estimation.

There are two main approaches for obtaining information about environment lighting in AR: (1) inserting active or passive light probes into a scene. This kind of methods use either an active camera with a fish-eye lens or a passive object with known shape and reflectivity such as chrome spheres to capture environmental illumination in real time. (2) estimating the illumination from the image of the main AR camera. This category of methods (probe-less methods) can estimate lighting information without additional equipment or arbitrary known objects in the scene. These methods typically use image features which are known to be directly affected by illumination, such as shadows, gradient of image brightness. Nowadays, data-driven approaches have been introduced which use a large dataset of panoramas to train an illumination predictor. Here, we mainly focus on the second category of methods since deep learning techniques could be applied.

Wang *et al.* [28] deployed a GAN-based method to achieve illumination consistency. A series of rendered images with different illumination and viewpoints were created first to serve as training data. Image transformation network is applied as a generator. Adversarial loss, feature matching loss and perceptual loss were considered and minimized to make the composite image more natural and realistic. However, the method is not that applicable since different virtual objects and scenes require different massive training process.

Kán and Kafumann [29] presented a novel method for dominant light estimation from RGB-D images. Authors used relative Euler angles calculated in camera coordinate space to serve as the direction of light source in training data since registration for different camera poses will lead to high complexity for network. Then, synthesized images and transformed light source positions are fed into a convolutional neural network equipped with residual blocks. A mean squared error was used as loss function for optimization. This method is relatively convincing since this network needs to be trained only once on a variety of scenes, and then can be applied in a new scene not seen before by network during training.

Mandl *et al.* [30] proposed a happy medium method that takes advantages of both light probe and probe-less light estimation, which was given the name "learned light probe". The system is split into a preparation and an online phase. A known object in the scene is selected arbitrarily as light probe at first. Then, authors reconstructed the geometry and albedo map of selected 3D object. For each uniformly distributed camera pose, they illuminated the retrieved 3D model with different spherical harmonics variations. The resulting images were then used to train a set of CNN instances for all camera poses which will be stored in the database at last. During the AR rendering, a live camera frame will be used to match the most similar camera pose in the database. Then the system would input that frame to the matched CNN model which outputs estimated spherical harmonics. This estimated lighting could be utilized to illuminate the augmented 3D object in the camera frame. "Learned light probe" works in an unmodified scene just like probe-less method but still retains the computational benefits of a normal light probe. However, it's computationally expensive to train CNN instances for all camera poses when the number of camera poses is considerably large.

5 Applications

In general, AR has diverse applications, consisting of education, medical, manufacturing and entertainment, etc. In this section, we demonstrate our survey on applications of AR, especially when equipped with deep learning techniques.

For navigation applications, Lin *et al.* [31] developed a campus navigation APP that uses object recognition and AR to make information of detected objects available and interactive on the smart phone screen. The APP first gets screenshots from phone camera and location data from GPS. These information is then treated as input to the search engine to figure out the rough camera pose and which model to be displayed. The output engine then uses a CNN to determine the actual place where the 3D model should be placed. Finally, computer graphics technologies were used to display 3D models and enabled them to interact with users. Moreover, Cruz *et al.* [32] proposed a system which was able to navigate users in big malls and display the interactive 3D models of some goods in a similar manner. With the help of argument reality and deep learning technologies, they make shopping and strolling easier and much more interesting. In [33], AR was applied in museum navigation. An android app pipeline for museum experience enhancing with deep learning was proposed.

AR could be introduced to enhance experiment in entertainment. In [34], combined traditional computer vision techniques and deep learning methods were used to solve Sudoku Puzzle in newspapers and could be displayed in an AR context.

In medical applications, [35] introduced Faster-RCNN in to visual inspection and installed it on an head mounted AR device. Besides, [36] developed a botulinum toxin injection application used for educational purpose, which relies on OpenCV and Dlib library to do injection point registration and show the results on patients' faces through AR techniques.

AR can also be utilized in aviation areas. In [37], with the assistance of AR techniques, it will be easier to do mismatched pins inspection of complex aviation connector. Onsite connector image can be captured by the digital camera of AR glasses and transmitted to background deep learning-based detection architecture. Once mismatched pins are identified, the AR glasses will deliver warning information to the operators with an intuitionistic interface to highlight those mismatched pins.

AR also has application in information security. Li *et al.* [38] proposed an AR-based information hiding architecture. Through this architecture, users could first map secret messages into secret keys which can be objects, images and coordinates, and then transmit the secret keys and concealing models to their patterns. The secret keys can only be revealed when the secret key is detected by the AR system. In such a system argument reality and deep learning technologies are exploited to make the secret keys various and thus improve the secret embedding rate.

In summary, AR has a wide range of application aspects. Together with deep learning technology, they made our life more comfortable and convenient.

6 Discussion

In this section, we will discuss some future directions of the AR techniques mensioned above.

We notice that although deep learning method is developed quickly, non-learning methods are developed in parallel. In tracking techniques, optical flow algorithm were explored and obtained good implement performance. [39] We believe that non-learning methods would behave as a significant part in future AR development and provide theory fundamentals for deep learning methods of AR.

As we can see in hand tracking methods development, various image types were explored. As an example, depth image improved hand tracking techniques greatly. Furthermore, infrared image provide a new horizon for fast hand pose estimation. We think these new types of image would be encouraged to try and utilize.

End-to-end camera pose estimation methods have made great strides in the past few years. Systems in such a manner are known for their robustness. However, so far, they fail to beat the state-of-the-art SLAM/SFM based systems in terms of accuracy. Moreover, these systems do not self-adaptively expand to unknown space since they are essentially an encoding of the training data. Thus, there is still a significant amount of research to be done before deep learning–based end-to-end camera pose estimation approaches can be brought into real world systems.

In illumination estimation field, up to now, deep learning techniques are mostly associated with probe-less light estimation methods. And this category of methods mainly focus on point light source estimation rather than environmental light. Besides, since the training datasets are synthesized by computer and are relatively simple, when facing with real world scenes, the trained model's robustness is still a challenging problem for researchers.

Deep learning empowers AR in a variety of fundamental techniques, we think more cutting-edge deep learning methods, such as reinforcement learning, meta learning, could be explored and utilized in AR technology.

References

[1] Georgios Lampropoulos, Euclid Keramopoulos, and Konstantinos Diamantaras. Enhancing the functionality of augmented reality using deep learning, semantic web and knowledge graphs: A review. *Visual Informatics*, 4(1):32–42, 2020.

[2] Feng Zhou, Henry Been-Lirn Duh, and Mark Billinghurst. Trends in Augmented Reality Tracking, Interaction and Display: A Review of Ten Years of ISMAR. *2008 7th IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 193–202, 2008.

[3] Kangsoo Kim, Mark Billinghurst, Gerd Bruder, Henry Been-Lirn Duh, and Gregory F. Welch. Revisiting Trends in Augmented Reality Research: A Review of the 2nd Decade of ISMAR (2008–2017). *IEEE Transactions on Visualization and Computer Graphics*, 24(11):2947–2962, 2018.

[4] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[5] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011.

[6] Orner Akgul, H. Ibrahim Penekli, and Yakup Genc. Applying Deep Learning in Augmented Reality Tracking. *2016 12th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*, pages 47–54, 2016.

[7] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 224–236, 2018.

[8] Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Toward geometric deep slam. *arXiv preprint arXiv:1707.07410*, 2017.

[9] T. Sattler, Q. Zhou, M. Pollefeys, and L. Leal-Taixé. Understanding the limitations of cnn-based absolute camera pose regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3297–3307, 2019.

[10] A. Kendall, M. Grimes, and R. Cipolla. Posenet: A convolutional network for real-time 6-dof camera relocalization. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 2938–2946, 2015.

[11] Alex Kendall and Roberto Cipolla. Modelling uncertainty in deep learning for camera relocalization. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4762–4769, 05 2016.

[12] Alex Kendall and Roberto Cipolla. Geometric loss functions for camera pose regression with deep learning, 2017.

[13] I. Melekhov, J. Ylioinas, J. Kannala, and E. Rahtu. Image-based localization using hourglass networks. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, pages 870–877, 2017.

[14] F. Walch, C. Hazirbas, L. Leal-Taixé, T. Sattler, S. Hilsenbeck, and D. Cremers. Image-based localization using lstms for structured feature correlation. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 627–637, 2017.

[15] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen. Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2652–2660, 2017.

[16] S. Seifi and T. Tuytelaars. How to improve cnn-based 6-dof camera pose estimation. In *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3788–3795, 2019.

[17] Samarth Brahmbhatt, Jinwei Gu, Kihwan Kim, James Hays, and Jan Kautz. Geometry-aware learning of maps for camera localization, 2018.

[18] Luca Carlone, G. Calafiore, Carlo Tommolillo, and Frank Dellaert. Planar pose graph optimization: Duality, optimal solutions, and verification. *IEEE Transactions on Robotics*, pages 1–21, 05 2016.

[19] Tom Duckett, Stephen Marsland, and Jonathan Shapiro. Fast, on-line learning of globally consistent maps. *Auton. Robots*, 12:287–300, 05 2002.

[20] F. Lu and E. Milios. Globally consistent range scan alignment for environment mapping. *Autonomous Robots*, 4:333–349, 1997.

[21] Mai Bui, Christoph Baur, Nassir Navab, Slobodan Ilic, and Shadi Albarqouni. Adversarial networks for camera pose regression and refinement, 2019.

[22] N. Radwan, A. Valada, and W. Burgard. Vlocnet++: Deep multitask learning for semantic visual localization and odometry. *IEEE Robotics and Automation Letters*, 3(4):4407–4414, 2018.

[23] Ali Erol, George Bebis, Mircea Nicolescu, Richard D. Boyle, and Xander Twombly. Vision-based hand pose estimation: A review. *Computer Vision and Image Understanding*, 108(1):52–73, 2007. Special Issue on Vision for Human-Computer Interaction.

[24] Christian Zimmermann and Thomas Brox. Learning to Estimate 3D Hand Pose from Single RGB Images. *arXiv*, 2017.

[25] Liuhao Ge, Yujun Cai, Junwu Weng, and Junsong Yuan. Hand pointnet: 3d hand pose estimation using point sets. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[26] Gyeongsik Moon, Ju Yong Chang, and Kyoung Mu Lee. V2v-posenet: Voxel-to-voxel prediction network for accurate 3d hand and human pose estimation from a single depth map. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] Gabyong Park, Tae-Kyun Kim, and Woontack Woo. 3D Hand Pose Estimation with a Single Infrared Camera via Domain Transfer Learning. *2020 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, 00:588–599, 2020.

[28] Xiang Wang, Kai Wang, and Shiguo Lian. Deep consistent illumination in augmented reality. *2019 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Oct 2019.

[29] Peter Kán and Hannes Kafumann. Deeplight: light source estimation for augmented reality using deep learning. *The Visual Computer*, 35:873–883, 2019.

[30] David Mandl, Kwang Moo Yi, and *et al.* Learning lightprobes for mixed reality illumination. *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 82–89, 2017.

[31] C. Lin, Y. Chung, B. Chou, H. Chen, and C. Tsai. A novel campus navigation app with augmented reality and deep learning. In *2018 IEEE International Conference on Applied System Invention (ICASI)*, pages 1075–1077, 2018.

[32] Edmanuel Cruz, Sergio Orts, Francisco Gomez-Donoso, carlos rizo, Jose Rangel Ortiz, and Miguel Cazorla. An augmented reality application for improving shopping experience in large retail stores. *Virtual Reality*, Accepted, 09 2019.

[33] Mudassar Ali Khan, Sabahat Israr, Abeer S Almogren, Ikram Ud Din, Ahmad Almogren, and Joel J. P. C. Rodrigues. Using augmented reality and deep learning to enhance Taxila Museum experience. *Journal of Real-Time Image Processing*, pages 1–12, 2020.

[34] Augmented Reality on Sudoku Puzzle using Computer Vision and Deep Learning. *International Journal of Innovative Technology and Exploring Engineering*, 8(11S2):140–145, 2019.

[35] Shaohan Wang, Sakib Ashraf Zargar, and Fuh-Gwo Yuan. Augmented reality for enhanced visual inspection through knowledge-based deep learning. *Structural Health Monitoring*, 20(1):426–442, 2021.

[36] HyoJoon Kim, SangHui Jeong, and *et al.* Augmented reality for botulinum toxin injection. *Concurrency Computat Pract Exper*, 32, 2020.

[37] Shufei Li, Pai Zheng, and Lianyu Zheng. An ar-assisted deep learning-based approach for automatic inspection of aviation connectors. *IEEE Transactions on Industrial Informatics*, 2020.

[38] Chuanlong Li, Xingming Sun, and Yuqian Li. Information hiding based on augmented reality. *Mathematical Biosciences and Engineering*, 16:4777–4787, 05 2019.

[39] Peng-Xia Cao, Wen-Xin Li, and Wei-Ping Ma. Tracking Registration Algorithm for Augmented Reality Based on Template Tracking. *International Journal of Automation and Computing*, 17(2):257–266, 2020.