Review Article
pISSN 2383-5389 / eISSN 2383-8116
Journal of International Society for Simulation Surgery 2016;3(2):39-45
https://doi.org/10.18204/JISSiS.2016.3.2.039

ORCID
Suwoong Heo: orcid.org/0000-0001-5516-8028
Hyewon Song: orcid.org/0000-0003-0681-0904
Jinwoo Kim: orcid.org/0000-0002-1437-2206
Anh-Duc Nguyen: orcid.org/0000-0001-7759-1134
Sanghoon Lee: orcid.org/0000-0001-9895-5347

Essential Computer Vision Methods for Maximal Visual Quality of Experience on Augmented Reality

Suwoong Heo, Hyewon Song, Jinwoo Kim, Anh-Duc Nguyen and Sanghoon Lee

The Department of Electrical and Electronic Engineering, Yonsei University Yonsei University, Seoul, Korea

The augmented reality is the environment which consists of real-world view and information drawn by computer. Since the image which user can see through augmented reality device is a synthetic image composed by real-view and virtual image, it is important to make the virtual image generated by computer well harmonized with real-view image. In this paper, we present reviews of several works about computer vision and graphics methods which give user realistic augmented reality experience. To generate visually harmonized synthetic image which consists of a real and a virtual image, 3D geometry and environmental information such as lighting or material surface reflectivity should be known by the computer. There are lots of computer vision methods which aim to estimate those. We introduce some of the approaches related to acquiring geometric information, lighting environment and material surface properties using monocular or multi-view images. We expect that this paper gives reader's intuition of the computer vision methods for providing a realistic augmented reality experience.

Key Words Augmented reality · Segmentation · Rendering · Lighting · 3D reconstruction.

Received: November 11, 2016 / **Revised:** November 14, 2016 / **Accepted:** November 19, 2016
Address for correspondence: Sanghoon Lee
The Department of Electrical and Electronic Engineering, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea
Tel: 82-2-2123-2767, **Fax:** 82-2-313-2879, **E-mail:** slee@yonsei.ac.kr

Introduction

The development of the smart device which contains the imaging module in addition to computing module enables us to build augmented reality (AR) system which visualizes the real-image and valuable information processed inside of the device simultaneously. There are several researches about the AR system for supporting medical surgery (1-5). Those type of research can be applied in the various area of medicine (e.g. dentistry, hepatic surgery). For example, Liao (1) proposes the 3D AR navigation system with autostereoscopic images. Their system aims to provide integrated videography (IV) image-guided therapy. AR can also support the medical education. Kamphuis (6) gives a comprehensive review of the potential to offer a highly realistic situated learning experience supportive of complex medical learning and transfer.

In the AR system, the virtual image synthesized with the real image is drawn by the rendering of virtual objects. Rendering is the computer graphic process which visualizes the virtual object in an image. There is sort of rendering techniques which draw the 3D virtual objects into an image – rasterization, ray-tracing.

For the reason of computational complexity, rasterization method which belongs to the local illumination category is widely used for rendering in AR system. The rasterization is pipelined processed. At first, the virtual objects described as

polygons in the 3D scene are transformed into the camera coordinate. Then those in camera space are projected onto the 2D plane via orthographic projection. Once polygons of each virtual object projected, those polygons outside of viewing window are clipped out. The scan conversion step is followed. It converts the vector image produced from orthographic projection into a raster image which is displayed on the screen. Since those steps can be carried out by fixed function hardware within the graphics pipeline, it is considered as suitable choice for real-time process of AR system.

However, in the aspect of visual quality of experience (QoE), the image produced by rasterization is not satisfactory. This is because of the fact that it is local illumination technique which not takes the phenomenon which can occur in the real world into account. Effects such as indirect lighting, transparency and mirror like reflection are ignored. Even existence of several tricks to produce those phenomena, it is hard to exactly describe those effects. In contrast, ray tracing categorized as global illumination technique accounts for those effects by adopting an intuitive illumination model. The original work of ray-tracing technique is Whitted ray-tracing model proposed by Whitted (7). This model utilizes the concept of the ray. The rays shot from virtual light source (forward method) or the screen (backward method) traverse entire 3D scene. Then each ray computes the surface irradiance of objects which they intersect. Then, rays are reflected or transmitted based on material properties assigned to the virtual object. After several reflections and transmission, the pixel values of an image are calculated by blending those rays which arrive at that pixel. As we can see in the Fig. 1, the image produced by ray-tracing method is more plausible.

In general, the ray tracing method considered as not suitable method for real-time applications (e.g. AR) since this method needs to compute the path of all rays which incorporates drastic computational time. Recently there are several works about hardware implementation of the ray tracing. For example, Nah (8) demonstrates about real-time applicability of his ray-tracing hardware on mobile devices. Based on those, if information of real-scene such as light source, material parameters and transparency of objects are given, we can produce visually plausible virtual image which would be synthesized into an image in AR system.

However, most case those information for the ray-tracing are not available in real-scene. Therefore, it is essential to estimate those including light source, properties of each material in the scene. In this paper, we review several methods related to estimating those. The remainder of this paper is organized as follows. Section II is about the description of ray tracing method. Section III introduces the methods related to geometric inference including semantic segmentation and 3D reconstruction. Section IV is about estimating light source from a real scene. Then we will present some of works about estimating material properties which contain diffuse, specular reflectivity and transparency in Section V. In the last section, we present conclusion

Ray-Tracing

Tuner Whitted proposed the Whitted ray-tracing model to

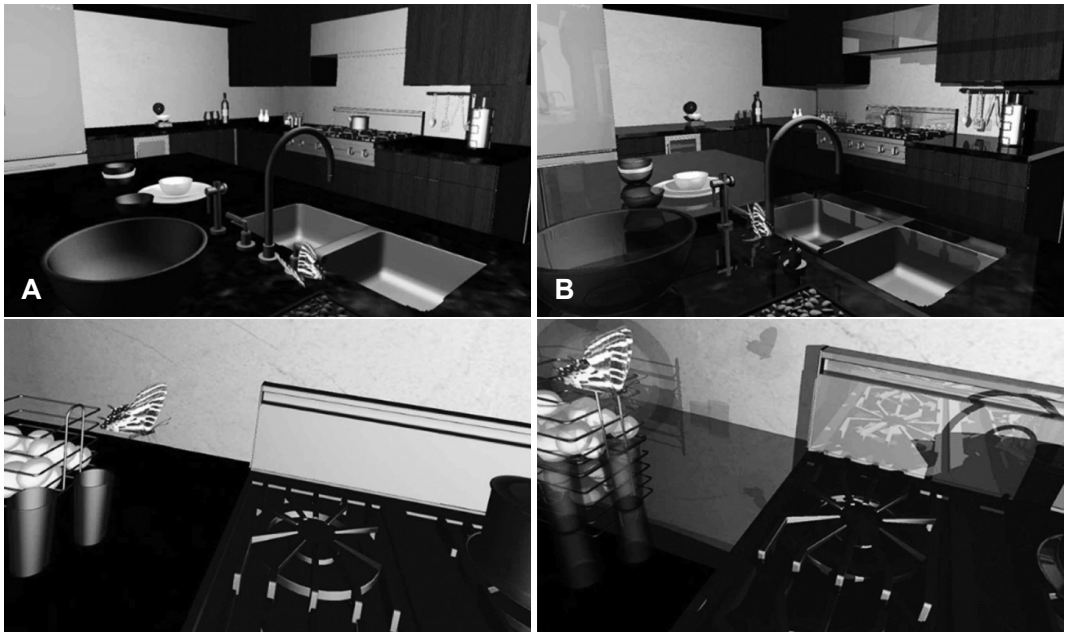

Fig. 1. Comparison between (A) Rasterization based and (B) Ray-tracing based Rendering Scenes.

render objects realistically as follows

$$I = I_a + k_d \sum_{j=1}^{j=l_s} (\bar{N} \cdot \bar{L}_j) + k_s S + k_t T \cdots (1).$$

where I is a reflected intensity, I_a is a reflection due to ambient light, k_d is a diffuse reflection constant, k_s is transmission coefficient, $\bar{N}$ is a unit surface normal, $(\bar{L}_i)$ is a vector in the direction of j^{th} light source, S is a intensity of light incident from the $\bar{s}$ direction and T is an intensity of light from the $\bar{t}$ direction.
According to the ray-tracing model, when an incident ray from the direction $\bar{i}$ is reflected to the direction $\bar{s}$ and refracted to the direction $\bar{t}$, the final intensity I of Eq. (1) is determined by the sum of the reflected light S and the refracted light T, where the coefficient k_s and the transmission coefficient k_t denote reflectance and transparency, respectively as shown in Fig. 2. The reflection direction $\bar{s}$ can be easily obtained, since the angle of incidence is equal to that of reflection. The refraction direction $\bar{t}$ is determined by the object characteristic, and the refractive index k_n is obtained by Snell's law.
By using the ray-tracing model, it is able to render objects realistically with shadow, reflection, and refraction effects of the object. However, we need to know the 3D geometric information of objects, properties of light sources (the number of light sources l_s, the direction of light sources $(\bar{L}_i)$, and the object properties (transmission or transparency coefficient k_$_t$, refractive index or the index of refraction k_n, diffuse coefficient k_d, reflection or specular coefficient k_s) in advance.

Geometric inference

Segmentation

Semantic segmentation is undoubtedly one of the most important topics in image processing and computer vision. There have been a huge number of methods proposed to solve the problem of salient object segmentation. Classically, there are three approaches: thresholding, region growing, and graph-based methods (9-13). Later, along with the development of deep neural networks (CNNs) in the problem of image classification, researchers also make use of these models to solve the problem of semantic segmentation due to the powerful ability to extract useful high-level features from a raw image as well as the capability of utilizing prior knowledge such as labels or ground-truth maps (14-18).

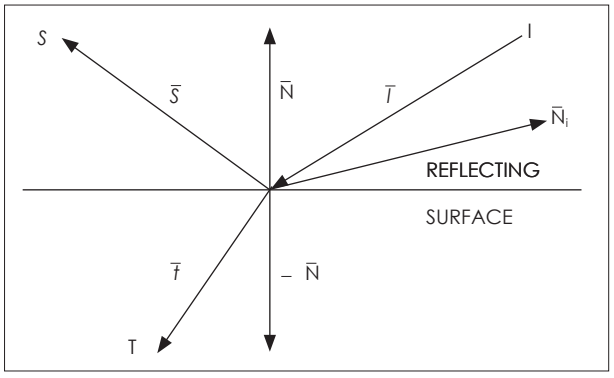

Fig. 2. Whitted ray-tracing model (7).

Conventional approach

One of the most basic methods in the conventional approach is thresholding. In this method, we usually compute the histogram of an image, and hopefully we can find a suitable threshold so that objects can be separated. The threshold can be found either manually or automatically. One way to automatically define the threshold is that we estimate some probability density functions for the histogram and then establish a threshold so that the total misclassification error is smaller. Two problems associated with this method are that it does not consider the spatial correlation at all and also, only gray values are used.

The second category in this approach is region growing methods. In general, the user initially selects some seed points and for each seed point, the algorithms will check whether the neighboring pixels belong to the region. The process is looped and in some way resembles to clustering algorithms. Famous methods in this category can be named such as split-and-merge and watershed algorithms. Several issues about region growing methods include the choice of initial seed points, computationally expensive, sensitivity to noise and no global view of the problem.

The last category in the conventional approach is graph-based models. In these methods, a graph is defined so that each node is associated with each pixel and there are edges connecting adjacent nodes. Each edge gives energy which is defined by some similarity measure so that edges in the same group have high energy while those in different groups have low energy. By finding the minimum cut to separate each group, we can optimally partition the graph and achieve good segmentation result. Some famous graph-based methods can be found in (13, 19).

Deep learning-based approach

The first practical CNN was trained to recognize hand-written digits through backpropagation and became worldwide famous after the ImageNet classification challenge (20-21). A myriad number of models have been proposed to deal with the recognition and classification problems. The power of CNNs lies in the high-level feature extraction after they are trained properly through efficient backpropagation. Acknowledging this potential of CNNs, researchers may either customize these mod-

els or come up with new network architectures to suit the problem of semantic segmentations. Inputs of the CNNs are a set of RGB images and sometimes include depth information. In contrast to the recognition or classification problem where outputs usually comprise a relatively small number of classes, the outputs of these methods are maps whose sizes are some multiplication of the input images because the semantic segmentation problem requires CNNs to make predictions for each pixel on in image. This is often referred as dense prediction. The dense maps produced by CNNs are very coarse and in order to enhance the results, researchers usually use some techniques such as superpixels, conditional random fields and recurrent neural networks to smooth the segment maps. Compared with the conventional techniques, CNN-based methods achieve significantly better segmented maps, especially on images having a large number of objects. However, a drawback of these models is that they need an enormous number of data to be properly trained. Nevertheless, many new models are increasingly introduced to increase the segmentation accuracy as well as decrease the computation time so that they can deal with real-time problems in our real-life situations.

3D reconstruction

In the Augmented Reality, it is necessary to obtain the 3D information of a scene, objects in the reality to synthesize real image and the virtual image naturally. Because we can calculate the position of virtual object using these 3D information.

Karsch suggested the method of rendering synthetic object in the image (22). In this method, at first, calculate the 3D structure of real image and then estimate the 3D position and illumination of light of real image (23-24). Also, calculate the 3D position of the objects which are needed for rendering in the real image and reconstruct the 3D space using pre-calculated 3D scene, light, and 3D objects. Finally, add a virtual object to obtain the final rendered image like below Fig. 3. However, there are several drawbacks to this method. First, they can obtain the 3D scene only in the restricted situation. And also, they cannot obtain the information of 3D objects automatically. The user has to decide the 3D objects manually. Finally, this method is inappropriate to use for Augmented Reality because it is not implemented in real-time.

For 3D reconstruction, using RGB-D camera is more common method than using only images. The representative method of 3D reconstruction in real-time is Kinect Fusion (25). Like Fig. 5, we can reconstruct human, object, scene in real-time in the process of depth map conversion, camera tracking, volumetric integration and ray-casting. This method is very useful for real-time reconstruction, but it takes long time to synthesize several scenes and the mesh is not accurate.

To compensate for weakness of Kinect Fusion, Whelan suggested "Elastic Fusion" which is based on SLAM (Simultaneous Localization And Mapping) (26). This system is able to capture, on line, comprehensive dense globally consistent surfel-based maps of a room scale environments with a RGB-D camera. The method uses a system which alternates between tracking and mapping phases. The map is separated into two parts, the active part recently observed, which is refined by new measurement and the inactive part no been observed during a period of time. Every frame, attempt to register the portion of the active model within the current estimated camera frame with the

Fig. 4. The result of Elastic Fusion.

Fig. 3. The rendered image (22).

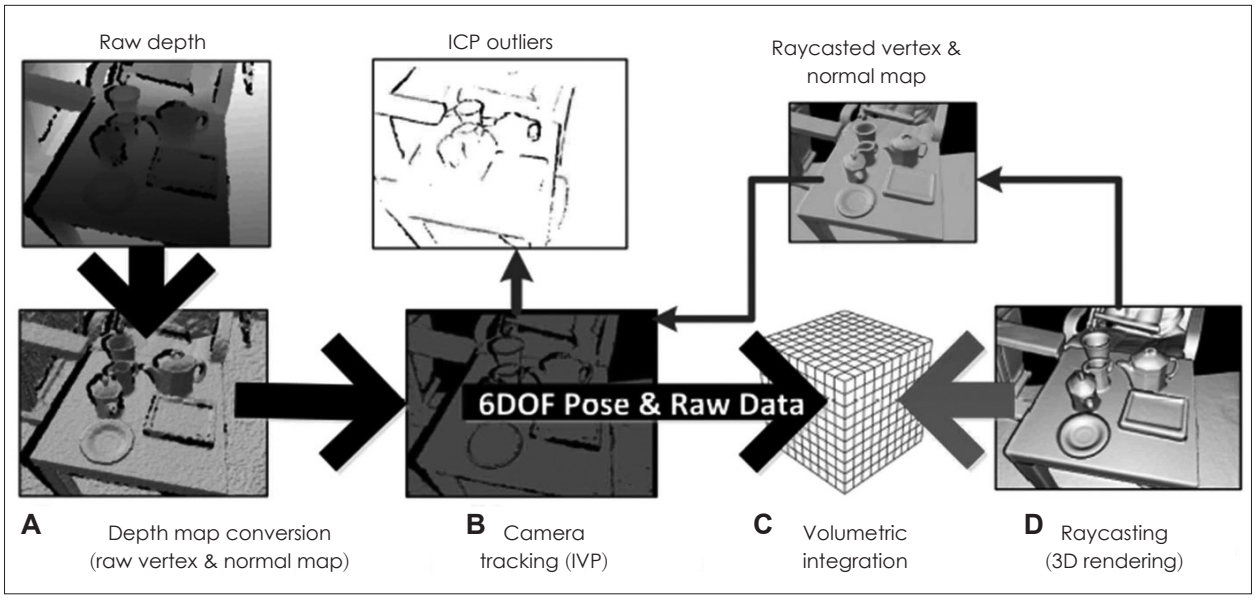

Fig. 5. The process of Kinect Fusion (25).

Table 1. Comparison of surface reconstruction accuracy (26)

System	kt0	kt1	kt2	kt3
DVO SLAM	0.032m	0.061m	0.119m	0.053m
RGB-D SLAM	0.044m	0.032m	0.031m	0.167m
MRSMap	0.061m	0.140m	0.098m	0.248m
Kintinuous	0.011m	0.008m	0.009m	0.150m
Frame-to-model	0.098m	**0.007m**	0.011m	0.107m
ElasticFusion	**0.007m**	**0.007m**	**0.008m**	**0.028m**

portion of the inactive model underlaid within the same frame. The result of Elastic Fusion is Fig. 4. Through the following Table. 1, Elastic Fusion is better than other SLAM algorithms in accuracy.

Light Source Estimation

Augmented Reality (AR) is one research area in the Computer Graphics (CG) that renders virtual objects into the real world which is shown by the camera. The main objective of this research is making the virtual object looks natural so that users think the object is really existed one.

To enhance a realistic AR experience, the virtual and the real world have to be rendered into the same light model. Ideally, the viewer should not be able to tell the difference between virtual and real. This is what is known as visual coherence. The main distinguishing aspects of AR are that visual coherent renderings must be generated in real time and with a limited amount of preparation. This turns out difficult problem and thus visual coherent rendering is still not a standard feature of commercial AR applications.

To estimate light source estimation, there are two ways, one is a light probe method and the other one is a probeless method. The visual effects in animation and games industries have successfully used a light probe to accurately capture the incident light field in a scene. The scene is photographed after inserting the light probe at one or multiple key location. The light probe is a simple calibration object of known size, shape and reflectance properties. This method is successfully used to acquire very detailed illumination environments, which can be used to render synthetic objects (27). However placing a light probe into a scene is not practical. If a scene was filmed or photographed without capturing the illumination by means of a light probe, then compositing and relighting tasks become much more difficult. The solution for estimating light sources then typically involve making significant and restrictive assumption about the nature of the scene. Therefore, current researchers do not insert a maker in the 3D real space. Instead, they utilize an arbitrary object in the real world as a light probe to increase visual coherent. In order to maximize the visual coherent, the assumption about the nature of the scene should be minimized.

While all these techniques estimate physically-based lighting from the scene, Khan (28) shows that wrapping an image to create the environment map can suffice for certain applications. i.e. these techniques make environment map from surrounding landscape, and regard the environment map as light source. In a previous study, the Light source estimation technique attempts to predict illumination with a data-driven matching approach. While data-driven approach using High Dynamic Range (HDR) panorama image is able to apply general image than light probe method, these method have a lot of computation. Thus this method should be fixed scene and difficult to

Fig. 6. Some input image to estimate light source using light probe method.

move virtual objects. Fig. 6 represents the result of relighting using environment map.

Material Properties Estimation

There are very few works about estimating material properties for rendering. Given geometric data and light information, we can estimate the diffuse k_d and specular k_s parameters as in Eq. (1). But it is hard to deal with transparency or refraction property of an object. Even detecting transparent or mirror like object in the scene also hard problem in computer vision. Some of works are concentrated on estimating diffuse and specular parameters only. Ko (29) proposes the method which recovers the parameters of Torrance-Sparrow model derived from BRDF (Bi-directional Reflectance Distribution Function). Their work is based on multi-view images. From those images, the diffuse albedo which arises with absence of illumination and the specular hemisphere which describes the illumination surround an object can be computed. Then rendering parameters except the transmission and refraction are estimated by comparing re-rendered scene and original images. In another approach proposed by Borom (30), they set a studio to reconstruct and capture the variance of the surface of a target object. They successfully recovers the diffuse and specular maps on the surface of an object. The surface normal of reflective part of the object is estimated to realize the mirror like reflection.

Conclusion

There are many researches to plug in the AR and medical treatment or educations. But there are few of those which interested in realistic synthesis between real and virtual image. We presented essential computer vision methods for maximal visual QoE on AR system. To produce realistic virtual image for the scene what user sees through the AR device, geometric inference and light source estimation are essential. Those informations can be applied to the rendering of virtual objects. For more visually natural synthetic image, virtual object can be rendered using ray-tracing method rather than rasterization. But estimating parameters related to transparency and mirror-like reflections are intractable. Those are very important issues in computer vision and graphics. Despite of the difficulty of estimating those, it is valuable to solve this problem and many researchers are engaged in solving this. We hope that this paper gives intuition about the methods related to generating realistic AR.

Acknowledgements

This work was supported by ETRI R&D Program (16ZC1400, The Development of a Realistic Surgery Rehearsal System based on Patient Specific Surgical Planning) funded by the Government of Korea.

References

1. Liao H, Inomata T, Sakuma I, Dohi T. 3-D augmented reality for MRI-guided surgery using integral videography autostereoscopic image overlay. IEEE Transactions on Biomedical Engineering 2010; 57(6):1476-1486
2. Wang J, Suenaga H, Hoshi K, Yang L, Kobayashi E, Sakuma I, et al. Augmented reality navigation with automatic marker-free image registration using 3-D image overlay for dental surgery. IEEE Transactions on Biomedical Engineering 2014;61(4):1295-1304
3. Shen F, Chen B, Guo Q, Qi Y, Shen Y. Augmented reality patient-specific reconstruction plate design for pelvic and acetabular fracture surgery. International Journal of Computer Assisted Radiology and Surgery 2013;8(2):169-179
4. Yamamoto T, Abolhassani N, Jung S, Okamura AM, Judkins TN. Augmented reality and haptic interfaces for robot-assisted surgery. The International Journal of Medical Robotics and Computer Assisted Surgery 2012;8(1):45-56
5. Ieiri S, Uemura M, Konishi K, Souzaki R, Nagao Y, Tsutsumi N, et al. Augmented reality navigation system for laparoscopic splenectomy in children based on preoperative CT image using optical tracking device. Pediatric Surgery International 2012;28(4):341-346
6. Kamphuis C, Barsom E, Schijven M, Christoph N. Augmented real-

ity in medical education?. Perspectives on Medical Education 2014; 3(4):300-311
7. Whitted T. An improved Illumination Model for Shaded Display. ACM SIGGRAPH Computer Graphics 1979:13.2
8. Nah JH, Kwon HJ, Kim DS, Jeong CH, Park JH, Han TD, et al, RayCore: A ray-tracing hardware architecture for mobile devices. ACM Transactions on Graphics 2014;33(5):162
9. Tobias OJ, Seara R, Image segmentation by histogram thresholding using fuzzy sets. IEEE Transactions on Image Processing 2002;11: 1457-1465
10. Haris K, Efstratiadis SN, Maglaveras N, Katsaggelos AK, Hybrid image segmentation using watersheds and fast region merging. IEEE Transactions on Image Processing 1998;7:1684-1699
11. Wu X, Adaptive split-and-merge segmentation based on piecewise least-square approximation. IEEE Transactions on Pattern Analysis and Machine Intelligence 1993;15:808-815
12. Sinha SN. Graph cut algorithms in vision. graphics and machine learning–an integrative paper, 2004
13. Shi J, Malik J, Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000; 22:888-905
14. Rouhi R, Jafari M, Kasaei S, Keshavarzian P, Benign and malignant breast tumors classification based on region growing and CNN segmentation. Expert Systems with Applications 2015;42:990-1002
15. Shelhamer E, Long J, Darrell T. Fully Convolutional Networks for Semantic Segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016
16. Noh H, Hong S, Han B, Learning deconvolution network for semantic segmentation. IEEE International Conference on Computer Vision 2015:1520-1528
17. Pinheiro PH, Collobert R, Recurrent Convolutional Neural Networks for Scene Labeling. ICML, 2014:82-90
18. Bruce NDB, Catton C, Janjic S. A Deeper Look at Saliency: Feature Contrast, Semantics, and Beyond. IEEE Conference on Computer Vision and Pattern Recognition 2016:516-524
19. Felzenszwalb PF, Huttenlocher DP, Efficient graph-based image segmentation. International Journal of Computer Vision 2004;59: 167-181
20. LeCun Y, Boser B, Denker JS, Henderson D, Howard RE, Hubbard W, et al. Backpropagation applied to handwritten zip code recognition. Neural Computation 1989;1:541-551
21. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. NIP, 2014:1097-1105
22. Karsch K, Hedau V, Forsyth D, Hoiem D. Rendering synthetic objects into legacy photographs. ACM Transactions on Graphics (TOG) 2011;30:6
23. Rother C. A new approach to vanishing point detection in architectural environments. Image and Vision Computing 2002;20(9):647-655
24. Coughlan JM, Yuille AL. Manhattan world: Compass direction from a single image by bayesian inference. Computer Vision, 1999:2
25. Newcombe RA, Izadi S, Hilliges O, Molyneaux D, Kim D, Davison AJ, et al. KinectFusion: Real-time dense surface mapping and tracking. Mixed and augmented reality (ISMAR), 2011
26. Whelan T, Leutenegger S, Salas-Moreno RF, Glocker B, Davison AJ. ElasticFusion: Dense SLAM without a pose graph. Proc. Robotics: Science and Systems, Rome, 2015
27. Perley RA. High dynamic range imaging. Synthesis Imaging in Radio Astronomy II. 1999:180
28. Khan EA, Reinhard E, Fleming RW, Bulthofff HH. Image-based material editing. ACM Transactions on Graphics (TOG) 2006;25(3): 654-663
29. Nishino K, Zhang Z, Ikeuchi K. Determining reflectance parameters and illumination distribution from a sparse set of images for view-dependent image synthesis. ICCV 2001;1:599-606
30. Tunwattanapong B, Fyffe G, Graham P, Busch J, Yu X, Ghosh A, et al. Acquiring reflectance and shape from continuous spherical harmonic illumination. ACM Transactions on Graphics (TOG) 2013; 32(4):109