Viewpoint Quality Evaluation for Augmented Virtual Environment

Ming Meng, Yi Zhou, Chong Tan, and Zhong Zhou[✉]

State Key Laboratory of Virtual Reality Technology and Systems,
Beihang University, Beijing 100191, China
zz@buaa.edu.cn

Abstract. Augmented Virtual Environment (AVE) fuses real-time video streaming with virtual scenes to provide a new capability of the real-world run-time perception. Although this technique has been developed for many years, it still suffers from the fusion correctness, complexity and the image distortion during flying. The image distortion could be commonly found in an AVE system, which is decided by the viewpoint of the environment. Existing work lacks of the evaluation of the viewpoint quality, and then failed to optimize the fly path for AVE. In this paper, we propose a novel method of viewpoint quality evaluation (VQE), taking texture distortion as evaluation metric. The texture stretch and object fragment are taken as the main factors of distortion. We visually compare our method with viewpoint entropy on campus scene, demonstrating that our method is superior in reflecting distortion degree. Furthermore, we conduct a user study, revealing that our method is suitable for the good quality demonstration with viewpoint control for AVE.

Keywords: Augmented Virtual Environment · Viewpoint quality evaluation
Texture distortion · Depth estimation · Semantic image segmentation

1 Introduction

Augmented Virtual Environment (AVE), known as one part of mixed reality (MR), defined as a dynamic fusion of the real imagery with the 3D models [1]. Broadly speaking, AVE is a virtual-reality environment augmented by fusing real-time, dynamic, multiple information with virtual scenes. The technology was first introduced in 1996 [2], and had made great progress over the last several years. Many kinds of AVE systems have been created, such as Photo Tourism [3] and HouseFly [4], and applied in 3D video surveillance, public security management, city planning and construction [5].

The fusion results directly rely on the texture projection techniques, projecting real-time video onto a 3D model. The 3D model is represented as sample boxes and can't display objects that not belong to this model, resulting unavailable texture distortion, such as the stretch distortion of pedestrians, road facilities, cars and trees. Due to limitation of passive modeling, the modeled depth of each image pixel is not completely accurate. Although the texture distortion, such as texture stretch and object fragment, looks seamless when the user's viewpoint is consistent with the camera's

viewpoint, it will become obvious as user viewpoint increasingly deviating from camera viewpoint. The illustrations of distortion are shown in Fig. 1.
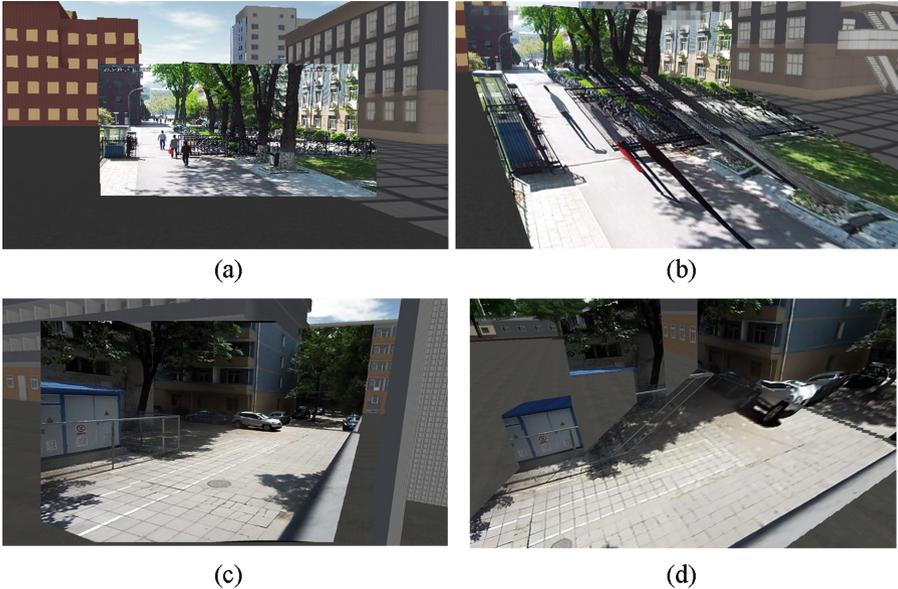

(a)   (b)

(c)   (d)

Fig. 1. Distortions of images/videos in AVE. (a) (c) Image model from camera's viewpoint. (b) (d) are respectively stretch distortion and object fragment, where the viewpoint deviate from camera's viewpoint.

In this work, we propose a novel viewpoint quality evaluation approach, using texture distortion as the metric of viewpoint quality for AVE. This approach includes stretch distortion and object fragment. We formulate the stretch distortion as accumulated relative error between model depth and "real depth" from depth estimation method, and object fragment as cumulative distance between semantic objects edge to the fragment model boundary. We combine these distortions in a weighted form for VQE. The main contributions of this work include: (1) we propose a new VQE method based on texture distortion. (2) We make a theoretical analysis of distortion phenomenon and the problem is mathematized. (3) We consider the effect of object semantic information on the metric of object fragment.

2   Related Work

Augmented Virtual Environment. AVE system displays still images onto scene models, and observers view them from arbitrary viewpoint. Neumann et al. [1] firstly introduced AVE concept and integrated it into a prototype system, supporting dynamic information extraction and complex scenes analysis through scene models reconstruction, real-time imagery collection and dynamic texture fusion. Sebe et al. [6] made

Neumann's technical extension to AVE, by proposing a novel virtualization system to make observers have an accurate comprehension and perception of dynamic events from arbitrary viewpoints. The Photo Tourism [3] was an end-to-end photo explorer used to interactively browsing 3D photos of popular scenic and historic sites. The HouseFly [4] was developed to project high-resolution videos onto 3D model, and generated the multi-modal virtualization of domestic and retail scene. However, the principle of this system, "directly project", brought insurmountable problems, such as hard to align real with virtual, unexpected video frames distortion. Zhou et al. [7] presented a new AVE video fusion technology based on active image-model technology, extracting video structure to generate image background model of the virtual scene, and projecting the real-time imagery onto model to enable users browse 3D videos from different viewpoint.

Viewpoint quality evaluation. Viewpoint quality is used to describe visual effect from viewpoint, and the higher the score, the better the viewpoint obtaining more detailed visual information in AVE. Generally, viewpoint quality is quantified through the information of 3D scene, such as geometry and texture. Previous methods [8–11] were mostly based on scene geometric information, which are difficult to evaluate high-quality viewpoints in complex scenes with multiple models. Relevant researchers performed the method of VQE based on user's visual perception [12–16], the typical methods include curvature entropy [13] and mesh saliency [14]. The results of these methods are not satisfied for the lack of model geometric information. In order to heighten the user's visual experience to some extent, Christie and Normand [17] investigated VQE method based on semantic information through the basic analysis of geometric and visual information. However, these methods, which are restricted by semantics understanding level of current scene are not suitable for multi-model scenarios.

Single image depth estimation. We compute the degree of stretch distortion through the accumulated relative error between model depth and "real depth", obtained from image depth estimation. Traditional methods of depth estimation were mostly based on geometric priors [18, 19]. Under the rapid development of machine learning, Liu et al. [20, 21] utilized the conditional random fields (CRF) to improve the accuracy of depth estimation for single image. Then Roy and Todorovic [22] adopted neural-random forest for depth estimation of single image, acquiring the same excellent depth estimation result as the above methods. Godard et al. [23] proposed a novel unsupervised depth estimation method, utilizing the unsupervised deep neural network to achieve more accurate results of depth estimation.

Image semantic segmentation. We take the semantic information of object into consideration when measure the degree of object fragment. Previously, the methods of semantic segmentation were mostly classified pixel-wise based on geometric information [24] and statistical methods [25]. The DeepLab [26–28] combined deep-convolutional neural networks (DCNNs) with probability map models without increasing network parameters. The RefineNet [29] aggregated low-level semantic features and high-level semantic classification, to further refine segmentation results with long-range residual links. Zhao et al. [30] proposed PSPNet, extracting multi-scale

information through the introduction of pyramid pooling module and achieving more accurate results of semantic segmentation.

Through the analysis of the above work, we extract two main factors that are related to the measurement of VQE, including stretch distortion and object fragment. Taking these distortions into VQE is necessary for improving the roaming experience in AVE. When measuring the stretch distortion, the "real depth" of single image is obtained by Godard's method. And the metric of object fragment is based on the results of semantic image segmentation by Zhao's network structure.

3   Proposed Approach

3.1   Problem Formulation

The key for getting better visual effects lies in how to reduce the visual distortions of AVE. In our scenario, we analyze the following two distortions, stretch distortion and object fragment, to evaluate the viewpoint quality.

Stretch Distortion. The generation schematic diagram of stretch distortion is shown in Fig. 2(a). Suppose we have a source image I for texture projection, captured from a camera viewpoint v_{cam}. When user observes the built image model (or projected image) from a virtual viewpoint v_{usr}, the texture distortion will occur, including stretch distortion $D_{stretch}$ and object fragment $D_{fragment}$. Assuming that the established image model R has a corresponding 3D model C based on true depth, and the spatial point set of C and R is separately denoted as S and S'. The process of projection transformation is defined as

$$\begin{cases} t = M \times S(p_i) \\ t' = M \times S'(p_i) \end{cases},\qquad(1)$$

where t and t' respectively denote the screen position of $S(p_i)$ and $S'(p_i)$ for pixel i. M is perspective transformation matrix, defined as $M = M_w \cdot M_p \cdot M_v \cdot M_m$. The four matrixes respectively indicate viewport matrix, projection matrix, view matrix, and model matrix.

The projection offsets of P pixels cause distortion phenomenon, such as pedestrians and vehicles are stretched. Denoting $L\,(p_i, v)$ as the distance error of each pixel projected onto screen, and the overall stretch distortion of scene is formulated as

$$D_{stretch}(v, R) = \sum_{i=1}^{P} L(p_i, v),\qquad(2)$$

where $v \in v_{usr}$ and $v_{cam} \in v_{usr}$. And $L\,(p_i, v) = |t - t'|$, if $v \neq v_{cam}$, then $t \neq t'$, indicating that the space coordinates projected on the screen are inconsistent, defined as pixel offset, resulting in stretch distortion. Otherwise, $L\,(p_i, v) = 0$ represents they are projected to the same screen position, revealing that stretch distortion does not exist.

Object fragment. The generation schematic diagram of object fragment is shown in Fig. 2(b). Each R consists of a group of triangle patches represented as triangle-patch set T_R. The boundary set of R is defined as E_R. The left and right sides of each $e_i \in E_R$ are uniformly sampled, generating pair-wise space coordinates (V_l, V_r), where $V_l \in T_R$ and $V_l \notin E_R$, $V_r \in T_R$ and $V_r \notin E_R$. If $V_l \neq V_r$ and there is no common boundary between them, e_i is defined as fragment boundary, dividing the object into two parts. The projection transformation of each fragment boundary is calculated as

$$\begin{cases} w_1 = M \cdot V_1 \\ w_2 = M \cdot V_2 \end{cases}, \tag{3}$$

where w_1, w_2 respectively represent the screen position of space coordinate $V_1 \in E_R$ and $V_2 \in E_R$.

The projection errors of H pixels of each fragment boundary cause object fragment. Denoting $B(p_j, v)$ as distance error of fragment boundary e_i from v projected onto screen. The overall fragment of image model is formulated as

$$D_{fragment}(v, e_i) = \sum_{j=1}^{H} B(p_j, v), \tag{4}$$

where $B(p_j, v) = |w_1 - w_2|$, if $v \neq v_{cam}$, then $w_1 \neq w_2$ and $B(p_j, v) \neq 0$, indicating that the fragment boundary is projected to different positions on the screen, resulting in object fragment. Otherwise, the fragment boundary is projected to the same screen coordinates, $w_1 = w_2$, and $B(p_j, v) = 0$, symbolizing no object fragment occurs.
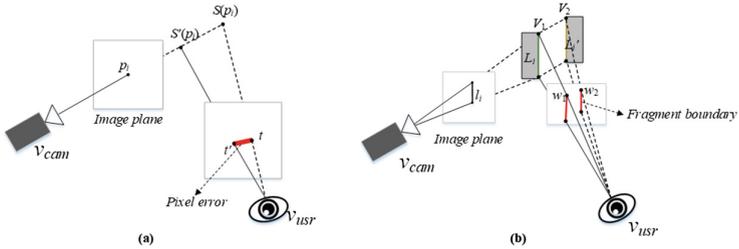

Fig. 2. Generation schematic diagram of distortion. (a) Stretch distortion. (b) Object fragment.

In summary, these two distortions are caused by the inconsistent depth, reflecting in screen when v_{usr} deviates from v_{cam}. The essential reason of stretch distortion is the offset of all pixels in the image, and it is inversely proportional to modeling accuracy. However, object fragment is caused by the offset of fragment boundary in the model, and it is proportional to modeling accuracy.

3.2   VQE Method for Stretch Distortion

Under the analysis of stretch distortion, we utilize the accumulated relative error of pixels projection as the metric of view quality evaluation. Using [23] to calculate the real depth of image, compared with image model depth to obtain projection error. Sampling the image model R to get the sampled pixel set $W(R)$, and the visible sampled pixels from v are denoted as $N(v, W(R))$. The degree of stretch distortion is computed as

$$L_{stretch}(v, R) = \frac{\sum_{p_i \in N(v, W(R))} |M \cdot S(p_i) - M \cdot S'(p_i)|}{|N(v, W(R))|}, \tag{5}$$

where $S(p_i) = l(R) + f(p_i, R) \cdot d$, $l(R)$ represents the location of v_{cam}, $f(p_i, R)$ is the unit vector indicating the orientation looking at p_i of image model, and d is the depth of mapped p_i.

The VQE method based on stretch distortion for single image model is denoted as

$$E_{stretch}(v, R) = \left(1 - \max\left\{\frac{L_{stretch}(v, R)}{L}, 1\right\}\right) \cdot \frac{Vis(v, R)}{r}, \tag{6}$$

where $Vis(v, R)$ denotes projection area of image model, r is screen resolution, and L is a fixed value, representing the acceptable maximum distance of pixel deviation, we take one-fifth of the screen diagonal as L.

3.3   VQE Method for Object Fragment

Analyzing the phenomena of object fragment above, we further propose a method of VQE based on object fragment. Using the cumulative error of fragment boundary projection to measure the degree of object fragment. The fragment degree of each fragment boundary is calculated using equation

$$L_{fragment}(v, e_i) = \frac{\sum_{p_i \in N(v, T(e_i))} |M \cdot V_1 - M \cdot V_2|}{|N(v, T(e_i))|}, \tag{7}$$

where $L_{fragment}(v, e_i)$ represents the fragment distance of ith fragment boundary. $T(e_i)$ is sampled pixel set of e_i and $N(v, T(e_i))$ is the visible sampled pixels of $T(e_i)$.

Different positions of fragment boundary in the object, causing various degree of object fragment. The greatest fragment occurs when the fragment boundary is in the middle of the object. We utilize the distance difference from fragment boundary to the two sides of object to measure the degree of object fragment. This paper obtains the results of semantic image segmentation by PSPNet [30], to get a more accurate distance difference, named as semantic weight λ. Therefore, the above-mentioned calculation function in Eq. (7) is extended into

$$L_{weight}(v, e_i) = \frac{\sum_{p_i \in N(v, T(e_i))} \lambda \cdot |M \cdot V_1 - M \cdot V_2|}{|N(v, T(e_i))|}, \tag{8}$$

where $\lambda = |1 - |d_1 - d_2||$, d_1 and d_2 respectively represent the distance of ith pixel from fragment edge to both segmentation edges of object, normalized to [0–1]. If the fragment boundary is in the middle of object, that is $\lambda \approx 1$, indicating that the fragment degree is most serious. Otherwise, the fragment edge is close to one of the object's segmentation edges, $d_1 \approx 1$ or $d_2 \approx 1$, revealing the fragment degree is not serious and can be ignored.

When determining how distortions affect the viewpoint quality, the score of VQE is in a weighted form, the computational formula is

$$E_{\text{distortion}}(v, R) = \alpha \cdot E_{stretch}(v, R) + \beta \cdot \sum_{i=1}^{N} L_{weight}(v, e_i) \tag{9}$$

where the hyper-parameters α and β are weight factors which control the contributions of the two terms, and we set $\alpha = \beta = 0.5$ empirically. N is the total number of fragment boundary of single image model R.

4 Experiments

4.1 Experimental Setups

We compare our method with viewpoint entropy with four campus scenes. We sample bounding sphere of each scene, getting a viewpoint set with 722 viewpoints. For a better visualization of our results, we utilize 7/8 view sphere with normalized heat map, same as [31], to illustrate viewpoints quality score. The view sphere's center is the source captured location of image, and its radius is the length of the vector from the sphere's center to the built image model's center. The sphere's north-east side is manually removed to make sure the visibility of inside section planes visible.

4.2 Experimental Results and Analysis

The results of VQE based on texture distortion are shown in Fig. 3. We select four image models and the blue cone is field of view (FOV), shown in Fig. 3(b). The red lines of Fig. 3(c) indicate the fragment boundaries.

Each view sphere demonstrates that the optimal viewpoint is the camera viewpoint, locating in the center of view sphere, where the distortion degree is 0. The back view from Fig. 3(d) shows that the more the viewpoint moves towards the rear of camera viewpoint, the larger the spatial range with higher viewpoint quality. The front view from Fig. 3(e) indicates that the viewpoint quality in front of camera viewpoint is deteriorating, that is the distance from viewpoint to image model is inversely proportional to the viewpoint quality.
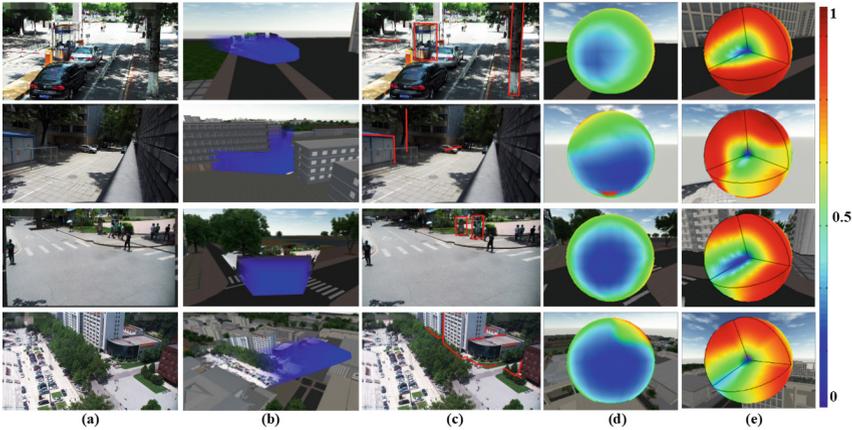
Fig. 3. Results of VQE based on texture distortion. Color values range of each view sphere from blue (good viewpoint) to red (bad viewpoint). (Color figure online)

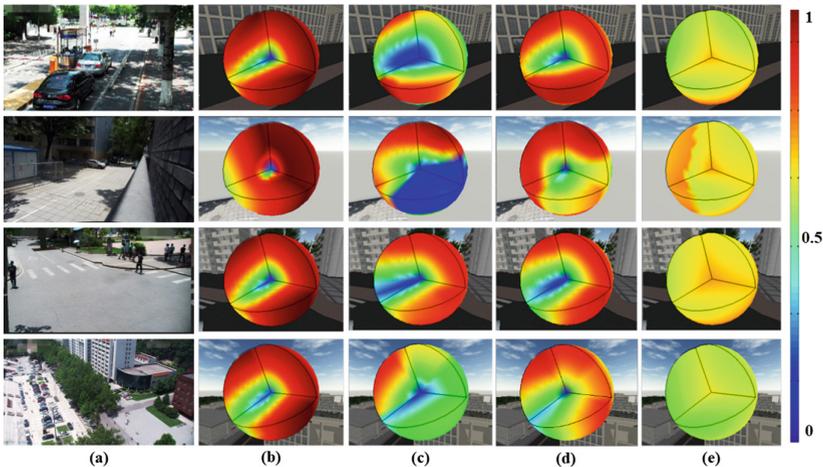The comparisons of four VQE methods are shown in Fig. 4.

Fig. 4. The results of four VQE methods. The last columns (b) (c) (d) (e) respectively denote our VQE method based on stretch distortion, our VQE method based on object, our VQE method based on texture distortion and the representative viewpoint entropy.

The Fig. 4(b) shows that the distribution of viewpoint quality varies slightly over the section of view sphere, where the quality of viewpoint is poor. This is because the stretch distortion exists in entire image. The Fig. 4(c) indicates that viewpoint quality from section drops rapidly when the viewpoint moves upward, due to the degree of object fragment is more severe from the top. While the viewpoint moves in the left and

right direction, the viewpoint quality deteriorates slowly, this is owing to the small fragment area and partial fragment being obscured by the foreground. We weight stretch distortion and object fragment equally, and the results are shown in Fig. 4(d). The last column (e) reveals that the viewpoint quality in the center of viewpoint sphere is lower than the outer edge viewpoint. The above results indicate that the VQE method based on texture distortion can better reflect viewpoint quality than geometric-based method.

In Fig. 5 we list the corresponding scenes from different viewpoints, intuitively displaying the good viewpoint and bad viewpoint by the distortion score of viewpoint.
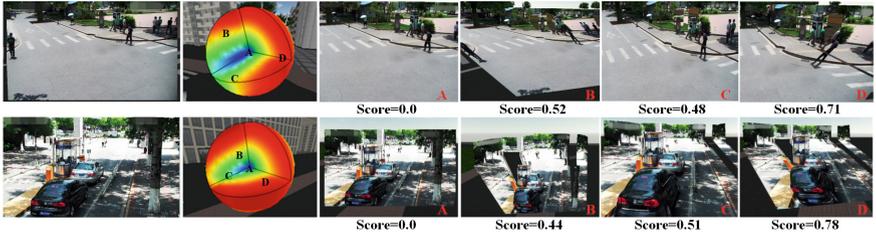

Fig. 5. The corresponding scenes and distortion degree from different viewpoints. 0 represents the best viewpoint, the closer the score to 1 the worse the viewpoint.

To quantitatively evaluate the effectiveness of our methods, we conduct a user study in AVE. Comparing our method with viewpoint entropy, and each image model is evaluated by 20 participants. Each participant has normal vision and gives a score based on the perceived comfort level. For each method we select five viewpoints to compare the score of user's evaluation and VQE methods (see Fig. 6).
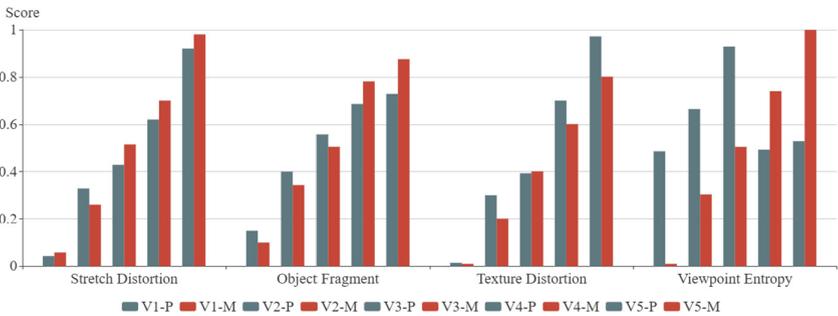

Fig. 6. Results of our user study. The charts report for each method the pair-wise scores of each viewpoint evaluated by user and VQE method.

As shown in Fig. 6, the differences of evaluation score between user and VQE method are not significant in the first three methods. However, the results of the fourth

method have significant differences. Our method is better than Viewpoint Entropy for the evaluation of distortion in AVE. We select Bhattacharyya Distance (BC) [32] to measure the similarity of participant evaluation and each method, the similarity value, abbreviated as BC, is shown in Table 1. The closer the user's score is to the score of our method, the greater the BC value is. Obviously, the similarity value of our method is higher than viewpoint entropy, which demonstrates that our method is more suitable for user's virtual perception.

Table 1. Bhattacharyya Distance of the user evaluation and results of VQE methods.

Method	Bhattacharyya Distance
Stretch distortion	0.9912
Object fragment	0.9670
Texture distortion	0.9942
Viewpoint Entropy	0.8952

5   Conclusion and Discussion

With the growth of the size and complexity of AVE, identifying good viewpoints automatically is an important requirement for good visual experience. Our method provides an elegant solution to achieve VQE. Comparing our method with other existing VQE methods, the main contribution of our method is the texture distortion metric for AVE. Experiments illustrated the effectiveness of the quality evaluation of the viewpoints.

References

1. Neumann, U., You, S., Hu, J.: Augmented virtual environments (AVE): dynamic fusion of imagery and 3D models. In: Proceedings of IEEE Virtual Reality, pp. 61–67. IEEE Computer Society (2003)
2. Moezzi, S., Katkere, A., Kuramura, D.Y., et al.: Immersive video. In: Proceedings of IEEE Virtual Reality (1996)
3. Snavely, N., Seitz, S.M., Szeliski, R.: Photo tourism: exploring photo collections in 3D. In: Proceedings of ACM Transactions on Graphics, pp. 835–846. ACM (2006)
4. Decamp, P., Shaw, G., Kubat, R.: An immersive system for browsing and visualizing surveillance video. In: Proceedings of the International Conference on Multimedia, pp. 371–380. ACM (2010)
5. Jian, H., Liao, J., Fan, X.: Augmented virtual environment: fusion of real-time video and 3D models in the digital earth system. Int. J. Digit. Earth **10**(9), 1–20 (2017)

 6. Sebe, I.O., Hu, J., You, S.: 3D video surveillance with augmented virtual environments. In: Proceedings of 1st ACM SIGMM International Workshop on Video Surveillance, pp. 107–112. ACM (2003)
 7. Zhou, Z., You, J., Yan, J., et al.: Method for 3D Scene Structure Modeling And Camera Registration From Single Image. US 20160249041 A1 (2016)
 8. Neumann, L., Sbert, M., Gooch, B.: Viewpoint quality: measures and applications. In: Proceedings of Eurographics Conference on Computational Aesthetics in Graphics, Visualization and Imaging, pp. 185–192. ACM (2005)
 9. Polonsky, O., Patané, G., Biasotti, S.: What's in an image? Vis. Comput. **21**(8–10), 840–847 (2005)
10. Zquez, P.P., Feixas, M., Sbert, M.: Viewpoint selection using viewpoint entropy. In: Proceedings of the 6th International Fall Workshop on Vision, Modeling, and Visualization, pp. 273–280 (2001)
11. Freitag, S., Weyers, B., Bönsch, A.: Comparison and evaluation of viewpoint quality estimation algorithms for immersive virtual environments. In: Proceedings of International Conference on Artificial Reality and Telexistence and Eurographics Symposium on Virtual Environments (2015)
12. Yamauchi, H., Saleem, W., Yoshizawa, S.: Towards stable and salient multi-view representation of 3D shapes. In: Proceedings of IEEE International Conference on Shape Modeling and Applications, pp. 40. IEEE Computer Society (2006)
13. Page, D.L., Koschan, A., Sukumar, S.R., Rouiabidi, B., Abidi, M.A.: Shape analysis algorithm based on information theory. In: Proceedings of International Conference on Image Processing, pp. 229–232 (2003)
14. Lee, C.H., Varshney, A., Jacobs, D.W.: Mesh saliency. ACM Trans. Graph. **24**(3), 659–666 (2005)
15. Vázquez, P.-P.: Automatic view selection through depth-based view stability analysis. Vis. Comput. **25**, 5–7 (2009)
16. Miao, Y., Wang, H., Hang, Z.: Best viewpoint selection driven by relief saliency entropy. J. Comput.-Aided Des. Comput. Graph. **23**(12), 2033–2039 (2011)
17. Christie, M., Normand, J.M.: A semantic space partitioning approach to virtual camera composition. Comput. Graph. Forum **24**(3), 247–256 (2005)
18. Tsai, G., Xu, C.H., Liu, J.E.: Real-time indoor scene understanding using Bayesian filtering with motion cues. In: IEEE International Conference on Computer Vision, pp. 121–128. IEEE Computer Society (2011)
19. Zeng, Y., Hu, Y., Liu, S.: GeoCueDepth: exploiting geometric structure cues to estimate depth from a single image. In: IEEE International Conference on Intelligent Robots and Systems (2017)
20. Liu, M.M., Salzmann, M., He, X.M.: Discrete-continuous depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 716–723 (2014)
21. Liu, F.Y., Shen, C.H., Lin, G.S.: Deep convolutional neural fields for depth estimation from a single image. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5162–5170 (2015)
22. Roy, A., Todorovic, S.: Monocular depth estimation using neural regression forest. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5506–5514 (2016)
23. Godard, C., Aodha, O.M., Brostow, G.J.: Unsupervised monocular depth estimation with left-right consistency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 270–279 (2017)

24. Lee, D.C., Hebert, M., Kanade, T.: Geometric reasoning for single image structure recovery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2136–2143 (2009)
25. Zhao, Y.B., Zhu, S.C.: Image parsing via stochastic scene grammar. In: Proceedings of the Conference and Workshop on Neural Information Processing System, pp. 73–81 (2011)
26. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K.: Semantic image segmentation with deep convolutional nets and fully connected crfs. Comput. Sci. **4**, 357–361 (2014)
27. Chen, L.C., Papandreou, G., Kokkinos, I.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
28. Chen, L.C., Yang, Y., Wang, J.: Attention to scale: scale-aware semantic image segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3640–3649 (2016)
29. Lin, G.S., Milan, A., Shen, C.H.: RefineNet: multi-path refinement networks with identity mappings for highresolution semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1925–1934 (2017)
30. Zhao, H.S., Shi, J.P., Qi, X.J.: Pyramid scene parsing network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2881–2890 (2017)
31. Zhou, Y., Xie, J.Q., Wu, W.: Path planning for virtual-reality integration surveillance system. J. Comput.-Aided Des. Comput. Graph. **30**(3), 514–523 (2018)
32. Guo, R.X., Pei, Q.C., Min, H.W.: Bhattacharyya distance feature selection. In: Proceedings of the 13th International Conference on Pattern Recognition, pp. 195–199 (1996)