Journal of Harbin Engineering University
ISSN: 1006-7043
Vol 44 No. 9
September 2023

Remote Expermentations of Artificial Intelligence in 3d Virtual World

Vijay Kumar Gumasa[1]
Research Scholar, Department Of Computer Science & Engineering, Faculty of Engineering& Technology, Dr. A. P. J. Abdul Kalam University Indore, MP, India
Dr. Manoj Eknath Patil[2]
Research Guide, Department Of Computer Science & Engineering, Faculty of Engineering& Technology, Dr.A. P. J. Abdul Kalam University, Indore, MP, India

Abstract: Visual Artificial Intelligence (AI) research has experienced significant advancements in recent years. To benefit from gathering a vast quantity of data across multiple conditions and environments. However, the task of gathering such data requires significant time and labor. In addition, the development and testing of visual AI algorithms aimed at multi-sensory models can be both costly and, in certain instances, pose potential risks in the real world. A 3D environment simulator that employs a view synthesis module to generate photo-realistic simulations and enables the adaptable configuration of multimodal sensors is designed to tackle these specific challenges. To enhance the capabilities of the view synthesis module, we integrate innovative techniques such as adaptive view selection, depth refinement, and layered rendering, with the objective of producing highly realistic imagery. This suggests that PreSim encompasses multiple advantages: (i) by showcasing a photo-realistic 3D environment, it facilitates the seamless integration of multisensory models within the virtual realm and enables these models to perceive and navigate scenes, (ii) Through the incorporation of an internal view synthesis module, PreSim facilitates the transfer of algorithms developed and tested in simulation onto physical platforms, eliminating the need for domain adaptation. (iii) Moreover, PreSim has the capacity to generate ample data for vision-based applications, encompassing object pose estimation and depth estimation.

Keywords:- Simulation and Animation, RGB-D Perception, Sensor Fusion,Remote Experimentation, 3D virtual worlds

I. Introduction

In recent years, data-driven approaches utilizing deep networks have resulted in remarkable progress in computer vision tasks, including depth estimation [3] and 6D object pose estimation. A substantial volume of data is required to train and evaluate the models of these data-driven methods. However, the task of data labeling and collection can be both tedious and time-consuming. Over time, simulated environments have emerged as an effective solution to address these challenges, as they can offer a substantial amount of annotated data for a wide range of AI tasks. One of the primary focuses of environment simulators is to achieve exceptional rendering of real scenes from various viewpoints. Several open-source simulators [1] exist with the goal of achieving this objective through the manipulation of various scene parameters, such as geometry, lighting, texture, and the 3D representation of stationary objects.On the other hand, configuring the parameters can be a labor-intensive and time-consuming endeavor. Despite possessing precise modeling and appropriate parameter settings, the virtual environment continues to fall short of capturing the diversity and richness observed in the real world. The smooth transfer of algorithms, which have been developed and tested through simulation, to physical platforms in order to achieve various vision-based tasks including obstacle avoidance, object recognition, and visual navigation, can be potentially hindered by the limitations of the virtual environment. This challenge is commonly referred to as the reality gap, which denotes the disparity between real data and synthetic data.

Game engines have been utilized to construct virtual environments, capitalizing on their capability for photorealistic rendering in order to address this issue. Regardless, users are unable to create their own environments with their own datasets due to the heavy reliance of the

simulated environment on the detailed datasets provided by the game engine's.However, to offer real-time rendering, game engines commonly employ 3D graphics pipelines. As a result, there exists a linear correlation between the rendering time and the number of polygons that require rendering (scene complexity). Dedicated hardware and architectural designs for 3D graphics are required to achieve real-time performance. In contrast, image-based rendering, enabling real-time realistic imagery, overcomes such limitations. By utilizing a restricted set of captured images, this method enables the realistic visualization of a 3D scene without the need for complete 3D reconstruction. In various environments, this approach showcased outstanding outcomes [4]. Moreover, when it comes to image-based rendering, the runtime is primarily dictated by the rendered image's display resolution, as opposed to being influenced by the complexity of the scene.

Taking advantage of image-based rendering, we introduce PreSim, a 3D photo-realistic environment simulator for training and testing AI algorithms. The objective is to minimize the gap between reality and simulation by offering a significant array of virtual Red Green Blue (RGB)-D views, generated with remarkable photo-realism from arbitrary viewpoints, thereby improving the suitability of vision-based applications. A primary contribution of our simulator is its provision of a virtual environment that is both 3D and exhibits remarkable photo-realism. In regions where the overall 3D representation of the environment may be inaccurate or have missing data, our simulator ensures that users still have access to precise poses of the multisensory model and color-and-depth image pairs from free viewpoints. This system incorporates a comprehensive visualizer that offers real-time positions and complete paths of moving sensors, as well as a global 3D map. The system integrates recorder components and sequence controller, responsible for regulating sensor motion and capturing essential information required for the development of AI algorithms. Our innovative view synthesis module, constructed using image-based rendering, incorporates adaptive view selection, depth refinement, and layered 3D warping techniques. The primary aim of this combination is to enhance the overall excellence of the synthesized images, with the added benefit of minimizing rendering complexity.

Artificial intelligence techniques extensively find application within the domain of entertainment, either through the utilization of embodied agents or the automated creation of artistic content. An alternative method involves utilizingAI to elevate the user experience by implementing innovative interactivity techniques based on AI. This holds particular significance in the development of artistic installations based on interactive 3D worlds [12]. An important challenge in the development of such installations lies in accurately translating the artistic intention into actual interactive elements, as these elements ultimately shape the experience for the users. The primary aim of this investigation revolved around enabling the representation of high-level behaviors within virtual worlds, with the purpose of integrating them as fundamental elements in art installations utilizing Virtual Reality (VR). The fundamental hypothesis posits that AI representations, inspired by planning formalisms, alongside AI-based simulations derived from them, can serve as the foundation for defining the behavior of virtual worlds within such installations.

As part of our approach to interactivity, dynamic computation of the consequences of user interaction is employed to generate cascading effects that evoke a specific user experience. Based on fundamental principles, a sequence of events is calculated by integrating components from an artistic brief.This involves the artist's foundational conceptualization for an interactive installation, as well as the envisioned user experience. Put simply, AI techniques are employed because they possess the capability to represent actions and carry out analogical transformations, ultimately resulting in the creation of a user experience [13].

Furthermore, the utilization of simulating behavior in virtual worlds through AI techniques is made feasible by capitalizing on a specific characteristic of game engines: their reliance on event-based systems for representing various forms of interaction. Event-based systems emerged as a means to distinctive physical interactions, simplifying the computational demands of simulating real-world dynamics. In contrast,

numerical simulation is used to control the movement of objects, while the outcomes of specific physical interactions, such as the shattering of glass following the collision with a hard object, can be calculated within a discretized system. This approach eliminates the requirement for complex real-time mechanical simulations. The installation encompasses a virtual world where individuals possess the capability to modify the conventional laws of physical causality. This is achieved by replacing the default effects of physical actions with new sequences of events.

VR art continues to delve into intricate user interactions within virtual environments, thus continuing the longstanding tradition of constructing alternative worlds in the realm of VR art. Take, for instance, Char Davies' Osmose VR installation or Maurice Benayoun'sQuarxs (part of the eponymous animation series), where unseen entities challenge the laws of physics. Contrarily, progress in VR art installations primarily revolved around the creation of virtual worlds and the adoption of ad hoc interaction techniques. As VR art continues to evolve, it presents a challenge in developing technologies that enable the establishment of world behavior from first principles. The creation of such technologies would make it possible to explore new possibilities and undertake more determined experiments in VR installations.

Ii. Literature Survey

C. Hong and colleagues [9] described the modern service industries as relying heavily on advanced technologies and the proficiency of service professionals to effectively conduct business operations in a globalized world. They also emphasize that service skills differ significantly from basic concepts or familiarity with science and technology.Without practical experience, individuals cannot receive effective training within a classroom. Service skills primarily involve soft skills, particularly those that require service staff to engage with applicable stakeholders, demonstrate an understanding of the service domain, address requests from multiple groups, evaluate potential solutions, prioritize tasks, and all agreed-upon final decisions.

S. Bronack and colleagues [11] employed a social constructivist framework to analyze the learning environment within the realm of 3D virtual worlds, emphasizing the significant disparities it possesses compared to web-based learning or traditional classroom environments. The authors argue that within the 3-dimensional (3D) world, students should be granted increased choices and support to create individual paths within this environment.

A. De Lucia and colleagues [7] present a virtual campus constructed as Second Life, incorporating four unique virtual spaces: a shared student campus, recreational areas, lecture rooms, and collaborative zones.Within a 3D multi-user virtual environment, the authors emphasize the importance of a user's connection to a learning community, along with their presence, awareness, and communication skills. A study was conducted with university students to assess the effectiveness of synchronous distance lectures using Second Life within the described learning environment. The findings of the study were highly positive.

D.C. Cliburn, J.L. Gross, et al. [8] employed a quasi-experimental design with pretest-posttest comparison groups approach to assess the impact of delivering a lecture in Second Life in contrast to conducting a lecture in the real world. The study demonstrated that individuals who attended the lecture in a physical setting outperformed participants who experienced a similar lecture within the virtual environment of Second Life during a terminal test quiz. In their comments, the researchers also highlighted the challenges faced by students, including difficulties in accessing the lecture material and the lack of restrictions on avatar behavior within the academic context.

P. Dev, et al [10] documented an extensive project that involved the development and evaluation of the computer-based simulator known as the Virtual Emergency Department. The objective of this project was to enable distance training for emergency medicine residency programs, emphasizing leadership and teamwork during trauma management. This aimed to successfully handle trauma without the need for practice with real patients.

L. Jarmon, et al [6] suggests that 3D virtual worlds have significant potential as suitable environments for experiential learning.They utilize a combination

of research methods, including focus groups, surveys, journal content analysis, as well as virtual world snapshots and video, to systematically evaluate the instructional influence that Second Life has for facilitating experiential learning in interdisciplinary communication.

C. M. Itin, et.al [15] contend that experiential learning encompasses deriving meaning from direct experience and emphasizes the significance of the individual's learning process.This approach establishes a deeper connection with the learner by addressing their specific needs and desires on a personal level.Based on this definition, a narrative script is developed for the educational service offered within Second Life.

N. Koenig and A. Howard, et.al [14] are widely acknowledged for their utilization of advanced physics engines to render both indoor and outdoor environments.While Gazebo possesses a range of features, it has limitations when it comes to creating visually rich environments on a large scale and providing realistic imagery. It has not kept pace with the rapid progress made in the latest rendering techniques that enable photo-realistic rendering. Another category of methodologies utilizes game engines with the capability to render camera streams with photo-realistic quality.

M. Savva et al., [2] describe the utilization of the Magnum engine for the creation of photo-realistic virtual environments. They also introduce a modular library that facilitates the development of AI tasks, such as visual navigation, within this framework. Nevertheless, the richness of simulated environments is constrained by their strong dependence on the capabilities of the engines. On the contrary, environment simulator empowers individuals to construct customized environments using the datasets they possess.

R. Ortiz-Cayon, G. Drettakis, and A. Djelouah, et al. [5] employ a strategy of dividing the image into super pixels to maintain the boundaries of objects. They then individually project each super pixel onto the virtual perspective using a local shape-preserving warping technique, with the aim of improving the blending quality. However, the specified approach overlooks photo-consistency and continues to encounter challenges such as inaccurate occlusion edges and the flattening of silhouettes. There have been several works that have improved the quality of synthesized images by filling holes.However, these methods have a fixed number of input views, which can result in failure to fill holes when the selected views are irrelevant or redundant.To avoid such a scenario, an adaptive approach is utilized for selecting views.

Iii. Methodology

Fig.1 displays the architecture of our simulator. The system encompasses a multisensory model, a global visualizer, scene datasets, controllers, and a view synthesis module. The simulator developed utilizes the Robot Operating System (ROS), which is well-known for its modular design, facilitating effortless customization, upgrades, and reusability. In the virtual environment, the process begins by importing the point cloud generated through 3D reconstruction of the real scene within the ROS. Then, the point cloud alongside the camera poses of the input images is showcased using Rviz, a graphical 3D visualization tool customized for utilization within the ROS framework. Following that, the movement of the virtual camera within the virtual world and real-time estimation of its 6Dimensional pose are accomplished by employing ROS. Utilizing the estimated pose as a reference, the system identifies the most closely matched pairs of color and depth images within a provided input dataset. Afterward, the chosen color and depth image pairs are employed to generate the synthesized view through the utilization of the view synthesis module. Simultaneously, the complete movement path of the mobile camera along with the generated color-and-depth image pairs is captured. Within the subsequent sections, detailed information is presented regarding the individual components of the simulator.

The aim is to construct a vision-based environment that is photo-realistic and allows free-viewpoint capabilities for tasks.Departing from previous techniques that build the entire virtual environment based on precisely reconstructed 3D geometry, the view synthesis module make use of utilizes a limited collection of RGB-D images for input. This module is capable of generating new color-and-depth image pairs with arbitrary viewpoints. The methodology incorporates innovative depth refinement and view selection

procedures, which are subsequently succeeded by a rapid rendering process.The collaborative efforts of these components aim to enhance the overall quality of the synthesized images, concurrently reducing rendering complexity.

Achieving high-quality rendering requires precise alignment of object boundaries across color-and-depth image pairs, as well as accurate depth values. This is due to the fact that inaccurate depth values and misalignment frequently result in noticeable artifacts.In the offline preprocessing stage, the aim is to accomplish this specific objective by utilizing an algorithm for pixel-level refinement of depth across multiple views.

In addition to rectifying misalignment between color-and-depth image pairs and filling in holes, achieving high-quality synthesized images also depends on the careful selection of input views.Selecting redundant or incorrect views by considering distances or angles among them commonly leads to the blurring of images. To prevent such scenarios, the selection of input images is conducted meticulously, taking into account the angles, distances, and overlaps between two views.

Our proposal revolves around employing a technique called layered depth image-based rendering to generate new sets of color-and-depth image pairs. A fundamental element of image-based rendering is 3D warping, encompassing the projection of pixels from the source image plane to the global coordinate system.Subsequently, these pixels are reprojectionto their new positions in a different image plane by utilizing camera intrinsic and extrinsic matrices.In cases where foreground and background entities occupy the same position during projection, it can result in the concealment or obstruction of the foreground objects by the background objects. This issue arises due to incorrect depth information or errors in the reprojection process. In order to address this issue, the depth map is partitioned into separate layers by utilizing the minimum and maximum depth values. For each individual layer, 3D warping is employed using matching color-and-depth image pairs to generate fresh images, followed by applying a median filter using a 3 × 3 window to complete any missing information within the generated image. Consequently, the newly generated images are combined in order to generate the final synthesized image. The method effectively addresses the issue of visibility by utilizing the capabilities of layered depths to represent concealed elements. A trade-off between speed and quality resulted in the determination that four layers are the optimal.

Upon the completion of the blending process, the synthesized image often contains gaps or holes resulting from the restricted quantity of input views utilized. In response to this challenge, we introduce an adaptive view selection approach, utilizing a flexible number of input images in order to efficiently address the gaps within the synthesized image. The process starts with the projection of a key image onto the virtual position, which is carefully chosen considering its distance, angle, and overlap with the virtual view. Subsequently, we identify and locate holes in the synthesized image. In the event that the size of the largest crater exceeds a predefined threshold (e.g., 0.04% of the entire image), an alternative input image is selected to fill the crater. This process is iterated as long as the size of the largest hole diminishes below the specified threshold or the maximum limit for the number of input views is reached.

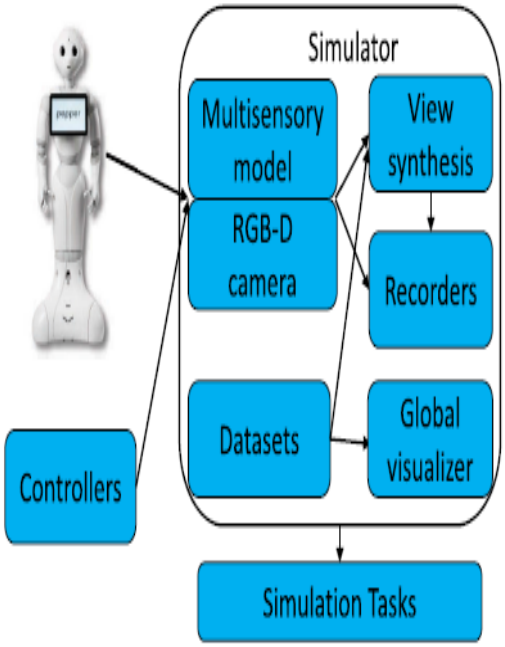

Fig.1: The architecture of simulator.

Multisensory Models and Controllers: PreSim is purpose-built to address the challenge of transferring domains between simulation and the real world.Therefore, the multisensory model must consistently adhere to constraints imposed by space and physics, including considerations for gravity and collision.

Universal Robotic Description Formats (URDF) are employed for the description of multisensory models, including humanoid robots. Thus, it is possible to customize the model and its characteristics (e.g., varieties of sensors). For demonstration purposes, this utilize the Pepper robot, a social humanoid robot developed by SoftBank.

To simplify the control complexity involved in the dynamic motions of the model, we offer a comprehensive range of practical controllers, such as navigation controllers and joint state. The joint state controller is utilized to regulate the movements of the model's joints, encompassing adjustments to the roll, pitch, and yaw angles. The navigation controller facilitates direct control of the model through the transmission of movement commands. Moreover, data recorders are provided, allowing for the storage of all the necessary data for learning-based approaches. Here's an example of a multisensory model and its trajectory.

Iv. Result Analysis

The evaluation of PreSim encompasses seven static datasets, which include three datasets we created ourselves (Table1, Table2, and Study room), four datasets (Playroom, Attic,Reading corner, Dorm) from and two dynamic datasets (Ballet and Break dancers). The seven static datasets consist of approximately 220 color-and-depth image pairs, encompassing objects such as black and texture-less items (e.g., writing boards and white walls), reflective objects (e.g., lights and bottles), and objects with small geometric features. In both dynamic datasets, there are sequences of 100 color-and-depth image pairs capturing individuals engaged in dancing activities. These sequences are captured using a set of eight static cameras arranged in an arc, each spaced 20 degrees apart.

In this approach, the ground truth image is selected at random from the initial captured dataset, representing a color image.Afterward, the remaining images are utilized to synthesize the selected image. It presents various instances of synthesized color images, displaying them alongside the corresponding ground truth images.The process of synthesizing a single image (1280 × 720) typically requires approximately 500–600 ms using a computer utilizing a 6-core Intel Core i7 8700 CPU operating at 3.19 GHz. While

achieves a faster processing speed that compared to our approach, it is important to highlight that is dependent on a GPU, whereas our method functions independently of a GPU. It is evident that our proposed method successfully generates high-quality synthesized images.

Table.1:Perforamnce Analysis

Methods	Average PSNR over 100 images (dB)	
	Ballet	Breakdancers
VSRS [24]	30.23	31.17
Liu [25]	32.52	33.33
Dai [26]	32.55	31.77
Loghman [27]	30.36	31.64
Ours	**33.41**	**33.61**

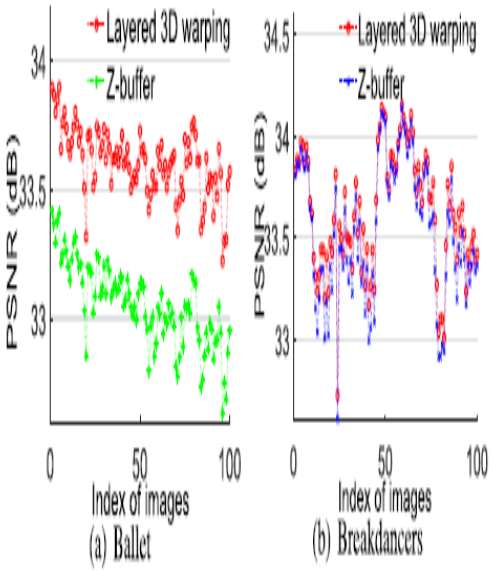

Fig.2: The PSNR comparison with layered 3Dwarping and Z-buffer on each frame.

V. Conclusion

Virtual environments and remote experimentation act as suitable tools that facilitate the collaborative process, offering an intriguing perspective on teaching collaboration and distributed learning across various applications. Such technologies have the capacity to enhance immersion, offering a sense of genuine presence and interaction. The objective of the exhibition is to showcase the seamless integration between3D remote experiments and virtual worlds, enhancing the appeal of fundamental concepts within science and technology careers. The generated data holds great potential for the training and testing data-driven approaches in a wide range of AI applications, including 6D object pose estimation

and depth estimation. The conducted experiments provide evidence of our simulator's capability to minimize the reality gap between the real scene and the virtual environment. As a result, vision-based algorithms formulated within the simulation have the capability to be directly applied to actual physical platforms without the need for domain adaptation.

Vi. References

[1] [C. Wang, "Densefusion: 6d object pose estimation by iterative dense fusion," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 3343–3352.

[2] M. Savva, "Habitat: A. platform for embodied ai research," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 9339–9347

[3] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab, "Deeper depth prediction with fully convolutional residual networks," in *Proc. 4th Int. Conf. 3D Vis.*, 2016, pp. 239–248.

[4] P. Hedman, T. Ritschel, G. Drettakis, and G. Brostow, "Scalable inside-out image-based rendering," *ACM Trans. Gr.*, vol. 35, no. 6, pp. 1–11, 2016.

[5] R. Ortiz-Cayon, A. Djelouah, and G. Drettakis, "A bayesian approach for selective image-based rendering using superpixels," in *Proc. Int. Conf. 3DVis.*, 2015, doi: 10.1109/3DV.2015.59.

[6] L. Jarmon, "Virtual world teaching, experiential learning, and assessment: An interdisciplinary communication course in Second Life," *Computers & Education*, vol. 53, no. 1, 2009, pp. 169-182.

[7] A. De Lucia, "Development and evaluation of a virtual campus on Second Life: The case of SecondDMI," *Computers &Education*, vol. 52, no. 1, 2009, pp. 220-233.

[8] D.C. Cliburn and J.L. Gross, "Second Life as a Medium for Lecturing in College Courses," *Proc. System Sciences, 2009. HICSS '09. 42nd Hawaii International Conference on*, 2009, pp.1-8.

[9] C. Hong, "Service Design for 3D Virtual World Learning Applications," *Book Service Design for 3D Virtual World Learning Applications*, Series Service Design for 3D Virtual World Learning Applications, ed., Editor ed.^eds., IEEE Computer Society, 2008, pp.795-796

[10] P. Dev, "Virtual Worlds and Team Training," *Anesthesiology Clinics*, vol. 25, no. 2, 2007, pp. 321-336.

[11] S. Bronack, "Learning in the Zone: A social constructivist framework for distance education in a 3D virtual world," *Proc.Society for Information Technology & Teacher EducationInternational Conference 2006*, AACE, 2006, pp. 268-275.

[12] M. Cavazza, "Causality and Virtual Reality Art," *Proc. 5th Conf. Creativity and Cognition*,ACM Press, 2005, pp. 4–12.

[13] J. Jacobson, "The CaveUT System: Immersive Entertainment Based on a Game Engine," *Proc. 2nd ACM SIGCHI Conf. Advances in Computer Entertainment Technology* (ACE 05), ACM Press, 2005, pp.184–187.

[14] N. Koenig and A. Howard, "Design and use paradigms for gazebo, an open-source multi-robot simulator," in *Proc. IEEE/RSJ Int. Conf. Intell.Robots Syst.*, vol. 3, 2004, pp. 2149–2154.

[15] C. M.Itin, "Reasserting the Philosophy of Experiential Education as a Vehicle for Change in the 21st Century," *The Journal ofExperiential Education*, vol. 22, no. 2, 1999, pp. 91-98.