Building 3D Scences from 2D Images using Machine Learning Algorithms

Cameron Cipriano
Department of Electrical and Computer Engineering
Boston University
Boston, MA
cmc322@bu.edu

Keywords—Image-based rendering, multiview stereo, neural rendering, multi-sphere images, superpixels, object reconstruction, shape induction, object detection, pose estimation, free-viewpoint navigation

I. Problem Statement

Experiments involving computer vision have been conducted as early as the 1950s utilizing some of the earliest neural networks to "detect the edges of an object and to sort simple objects into categories like circles and squares" [1]. Graduating from this very simplistic task, computer vision was first used commercially to recognize both typed and handwritten characters, known as Optical Character Recognition (OCR), providing the blind with the ability interpret text, as well as becoming the backbone of the modern mail system in many countries to decipher addresses and letter text [2]. Finally, with the advent and growing maturity of the internet, computer vision has seen drastic improvement in recognition tasks due to the large influx in available image databases, smartphone camera prevalence, computing cost reduction, novel and designated hardware designs, as well as inventive concepts and algorithms such as deep learning and convolutional neural networks (CNN) [2]. Regarding these milestones, the field of computer vision has seen abundant success in the purely two-dimensional (2D) world, revolutionizing the way we communicate, automate certain tasks, manufacture goods, prevent fraud, and improve security, however, we cannot ignore the three-dimensional (3D) world that every physical entity, including ourselves, resides in. Humans, and most animals, are only able to see the world via its two-dimensional (2D) projections given our stereo visual systems. This central fact is the basis and main challenge in the field of computer vision: to understand and reconstruct our lower-dimensional projections into the true space we perceive. We perform this seemingly menial task successfully each day, allowing us to navigate our environments, reason about the tangible items and people around us, and most importantly, interact with the world. In this way, constructing 3D scenes from 2D images is one of the most fundamental problems to solve, and will drastically expand the capabilities of technology in numerous fields such as autonomous driving, manufacturing, entertainment, retail/commerce, security, and many more.

II. Applications

Computer vision is widely applicable in the modern world we live in considering its goal is to mimic and potentially out-perform our own visual capabilities of recognizing, understanding, and even acting on visual cues. In fact, some systems are already far more capable at detecting and reacting to visual inputs [1], however, these systems' scopes are severely limited to the one task they were trained for. Looking more closely at building 3D scenes from 2D images, the applications are quite limitless.

A. Autonomous/Self-Driving Vehicles

The National Highway Traffic Safety Administration (NHTSA) claims that "fully automated cars and trucks that drive us, instead of us driving them, will become a reality" [3]. This claim is well supported by many in the industry, resulting in the development of the six levels of autonomy: "no automation", "driver assistance", "partial automation", "conditional automation", "high automation", and "full automation" [3]. Computer vision plays a crucial role in this growing area as camera technology is being incorporated more readily than other, more expensive sensors, such as LiDAR, radar, and ultrasonic sensors. As an example, one of the most popular electric cars capable of level 2 autonomy ("partial autonomy") has opted for a "vision-only" approach, employing eight separate cameras providing "360 degrees of visibility around the car at up to 250 meters of range" [4]. This fact demonstrates the extreme importance of being able to construct 3D scenes from these 2D video streams, providing the car with the ability to drive in our 3D world without colliding with other vehicles, obstacles, and pedestrians, getting us to our desired locations safely.

B. Manufacturing

Industry 4.0, otherwise described by near complete automation of factory facilities through employing advanced sensors, embedded software and robotics, internet of things (IoT), cloud computing, analytics, artificial intelligence (AI) and machine learning (ML) [5], is laden with computer vision. It is nearly impossible to find a modern factory that does not have process automation via robotics and computer vision systems as they provide unprecedented repeatability, tireless and continuous operation, as well as the ability to gather data and improve themselves. In the 2D plane, computer vision is used to identify manufacturing defects among work-in-progress (WIP) and finished products, guide robotic arms through complex motions defined by CAD models, provide a basis for predictive maintenance, inventory management, and even safety protocols (like determining if workers are wearing hardhats) [6]. Translating these features in 3D, the most promising applications would be for defect detection. A multi-camera rig capable of producing a 3D object would be invaluable for a growing internal database of defect-free and defective products to continuously improve the detection and mitigation of errors in the manufacturing process. In addition, robots with multiple cameras capable of producing 3D renderings of their environment could increase operational safety, as well as more accurate positioning to match against an inherently 3D CAD model, improving product reliability, quality, and closeness to the original design.

C. *Entertainment*

Computer vision and constructing 3D scenes from 2D images is not just limited to “practical” implementations such as autonomous driving and manufacturing but can be used for entertainment as well. Analyzing the hugely profitable videogame industry, generating $155 billion in 2020 with projections to hit $260 billion by 2025 [7], this technology of constructing 3D scenes from 2D images can be used to supplement game designer’s job of modeling real-world locations amongst development companies with lower budgets. Having the ability to photograph physical spaces and translate them to 3D scenes could assist the design of game worlds based on real locations, the incorporation of real objects that may be obscure to model, as well as create more realistic virtual reality (VR) games.

D. *Retail/Services*

1) Retail: In the world of retail, computer vision has made a significant appearance through Amazon Go stores. Amazon Go is a new branch and type of grocery store where there are “no lines, no checkout… Just Walk Out Technology” [8], which makes extensive use of computer vision and additional sensor data to track items that are removed from shelves and taken out of the store. With hundreds of cameras looming high above customers’ heads, it would be a great application of creating 3D scenes from these 2D video streams. Amazon would be able to generate a more complete view of their stores, enabling more accurate item tracking, and potentially reduce the number of employees required to operate the store in case of technical errors or ID verification for alcohol purchases.

2) Technology Services: With the ubiquitousness of smartphones and society’s use of GPS for navigation, computer vision will be an invaluable asset to the core mapping technologies employed by companies like Google, Apple, Microsoft, and other mapping solutions. Through capturing consecutive 360-degree images and potentially fusing other sensory data like range/depth from LiDAR, constructing 3D scenes from these inherently 2D images can help to improve features like Google Street View. Being able to generate novel views will provide more interactivity for users, ultimately allowing for greater spacial understanding, confidence during travel, and even provide the basis for more comprehensive 3D tours and tourist attractions for those who cannot physically visit. When linked with VR, this technolgoy can create truly immersive and life-like experiences.

E. *Security/Law Enforcement*

Taking computer vision to yet another field, we can show that this technology could prove quite effective in assisting law enforcement in preventing and solving crimes. A large majority of public retail locations are equipped with multiple cameras to monitor the behavior of customers and provide a record of events in case crimes are committed. Outfitting existing systems with the ability to generate 3D scenes from these 2D video feeds could provide instrumental context and additional evidence that might have been obstructed/overlooked. Knowing that store owners can virtually reconstruct the entire environment might also deter criminals from carrying their plans for fear of being watched from essentially “all angles”.

III. Literature Review

Performing this retrieval of a 3D scene from 2D images is no small feat, and thus has numerous and varied methods for attempting to do so. Previous methods include:

1) Image Based Rendering: This class of algorithms proved to be an integral starting point to where reconstructing 3D scenes from 2D images spawned. As mentioned in [9]–[10], some of the earliest approaches date back to 1995 through work on plenoptic modeling, light fields in 1996, and unstructured lumigraphs in 2001. Despite not developing into the traditional reconstruction capabilties we can achieve today, these early methods translated into technologies we still utilize heavily today such as camera stabilization techniques, video enhancement, and strated commercial products like Google Street View [9]. These methods “demonstrated the power of *blending* a set of images to allow various effects such as viewpoint changes or depth-of-field, without the need for manually building and rendering an entire new scene” [10].

2) Image Interpolation and Geoemetric Proxies: New methods of performing novel-view construction from a set of input images came about called Structure from Motion (SfM) and Multi-View Stereo (MVS) which povided for the automatic reconstruction of approximate 3D geometry proxy [9], [10]. However, as pointed out in [9], regions that have poor reconstruction end up having poor geometric proxies, resulting in significant visual artifacts, or incorrect translations, during rendering. Methods of correcting for this include Ambient Point Clouds which employ non-photorealistic rendering (NPR) to effectively “patch” these areas where poor reconstruction occurs, restricting the output image to interpolations instead of extrapolations.

3) Variational waps guided by sparse multiview stereo (MVS) point clouds: This method of generating 3D scenes takes advantage of a multi-camera setup to generate sparse point clouds that provide anchor points to warp images to novel views [9]. Work has been done with these sparse point clouds to warp video frames to new camera positions, effectively becoming a video stabilizer. Furthermore, when translating images using this variational warp, limitations have come up such as requiring manual labeling of object/feature silhouettes near the variational warp, as well as highly specialized camera rigs and/or setups that average individuals, companies may not have access to given their complexity.

4) 3D Reconstruction and Depth Propagation: This method relies on modern multiview stereo, however, only reconstructs about 100-200K pixesl from a 5-6MP image, resulting in a distribution of depth samples being highly irregular, sparse, and or erroneous near silhouettes [9]. These depth maps can be merged with 3D point clouds via surface reconstruction, however, usually optimize for photo-consistency which can become very "challenging for texture-poor surfaces, complex (dis)occlusions (e.g. leaves), [and] non-lambertian surfaces" [9].

Modern methods for performing this task of reconstructing 3D scenes from 2D images include:

5) *Neural Rendering:* This area is the culmination of years of research in deep learning to improve existing IBR tecniques. Neural rendering often builds upon MVS proxy geometry, utilizing a neural network to learn the *blend weights* for novel view synthesis of images [10]. As described in [10], MVS geometric proxies tend to have the limitation that the reprojected geometries are fixed or erroneous, sometimes hindering image quality and overall novel view synthesis. In their method, [10] describes the use of a differentiable pipeline so that they can backpropogate to fix errors generated by the MVS geometry proxies.

6) *Point-based Rendering:* Typically used in combination with neural rendering, points inherently come with RGBD sensors, as well as SfM and early steps of numerous MVS algorithms [10]. This is of great benefit because points have been well studied in graphics and can easily be reprojected from learned features.

7) *Oversegmentation (Superpixels):* While not inherently its own method for translating 2D images into 3D scenes, this technique provides the baseline for many of the aforementioed methods like point-pased and neural renderings. Superpixels consist of regions of images that contain similar information, usuaully preserving depth characteristics, allowing for more robust segmentation of images and observing higher-level patterns. There are a few methods to producing these superpixels, however, the most common is Simple Linear Iterative Clustering (SLIC) which [10] implements to identify its local, shape preserving warps to maintain the quality of poorly reconstructed areas.

8) *Applications and Smaller Pieces:* Utilizing these previously described methods, [11] combines stereo magnification, light field fusion, image interpolation, depth estimation, and novel view synthesis to 360-degree images, creating worlds where users are able to move around with six degrees of freedom (DoF) for truly immersive virtual reality experiences. Additionally, segmenting the larger issue of full 3D scene reconstruction, [13] researched a way of turning 2D projections of everyday items like chairs, motorcycles, planes, and more into their 3D objects. The paper was quite limited in its ability though as it's only able to produce 32x32x32 objects, however, further research can prove promising to translate simple pictures into full 3D objects. Furthermore, the work done in [12], [14] describe the use of creating individual objects from 2D images. [12] developed viewpoint, shape, and shape deformation models/algorithms to be able to take an input image and transform it into its 3D representation. [14] uses similar algorithms, not to reproduce a 3D model of the detected object, rather place a 3D bounding box around it instead of the traditional 2D one. This third dimension to the bounding box conveys much more information than that of a 2D one since it localizes it in space both locationally and orientationally.

IV. Open-Source Review

From an open-source perspective, most of the research studies mentioned have associated open-source repositories. [10] provides its source code under the Gitlab platform, however it's protected under an "All rights reserved" license. This indicates any individual attempting to use this code needs to explicitly contact and receive permission from the authors to be able to use it. Others such as [14] do not have an official codebase, rather multiple community implementations that are free to use and distribute. This demonstrates an interesting fact about open-source and research in that anyone can read and synthesize theoretical research papers and adapt them to become "real". [9] provides its source code under Gitlab as well, however, the researchers have described a custom license in which the code can only be used for research/non-profit reasons. The code also must be shared with the existing license to be distributed. Otherwise, it is free to distribute and use. Additional libraries exist which provide the basis for the other implementations, such as PyTorch, TensorFlow, Numpy, Scipy, CUDA, Pix3D, and many more. Overall, it's a flourishing research area with lots of open-source repositories. The unfortunate aspect, for myself, is I am mostly unable to run the projects due to the technology I have access to at home.

References

[1] https://www.sas.com/en_us/insights/analytics/computer-vision.html#history

[2] https://mailingsystemstechnology.com/article-2813-Optical-Character-Recognition-A-Backbone-for-Postal-and-Mail-Sorting-Applications.html

[3] https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

[4] https://www.tesla.com/autopilot

[5] https://www.ibm.com/topics/industry-4-0

[6] https://blog.vsoftconsulting.com/blog/top-usecases-of-computer-vision-in-manufacturing

[7] https://www.investopedia.com/articles/investing/053115/how-video-game-industry-changing.asp

[8] https://www.amazon.com/b?ie=UTF8&node=16008589011

[9] https://www-sop.inria.fr/reves/Basilic/2013/CDSD13/spixel_warp_preprint.pdf

[10] http://www-sop.inria.fr/reves/Basilic/2021/KPLD21/KopanasPointBasedNeuralRenderingPerViewOptimization.pdf#page14

[11] https://arxiv.org/pdf/2008.06534.pdf

[12] https://abhishekkar.info/categoryshapes.pdf

[13] https://arxiv.org/pdf/1612.05872.pdf

[14] https://arxiv.org/pdf/1612.00496.pdf