5-2024

Learning Scene Semantics for 3D Scene Retrieval

Natalie Gleason

Learning Scene Semantics for 3D Scene Retrieval

by

Natalie Jean Gleason

A Thesis
Submitted to the Honors College of
The University of Southern Mississippi
in Partial Fulfillment
of Honors Requirements

May 2024

Approved by:

Dr. Bo Li, Ph.D., Thesis Advisor,
School of Computing Sciences and Computer Engineering

Dr. Sarah Lee, Ph.D., Director,
School of Computing Sciences and Computer Engineering

Joyce Inman Ph.D., Dean
Honors College

ABSTRACT

This project presents a comprehensive exploration into semantics-driven 3D scene retrieval, aiming to bridge the gap between 2D sketches/images and 3D models. Through four distinct research objectives, this project endeavors to construct a foundational infrastructure, develop methodologies for quantifying semantic similarity, and advance a semantics-based retrieval framework for 2D scene sketch-based and image-based 3D scene retrieval. Leveraging WordNet as a foundational semantic ontology library, the research proposes the construction of an extensive hierarchical scene semantic tree, enriching 2D/3D scenes with encoded semantic information. The methodologies for semantic similarity computation utilize this semantic tree to bridge the semantic disparity between 2D sketches/images and 3D models, enhancing retrieval performance. Furthermore, the project proposes a semantics-driven framework for 2D scene sketch-based and image-based 3D scene retrieval, aiming to unlock new opportunities for applications spanning virtual reality, 3D entertainment, and autonomous systems. Overall, this thesis contributes valuable insights and methodologies to the field of semantics-driven 3D scene retrieval, laying a solid foundation for future advancements and interdisciplinary collaborations whilst promoting research development in the developing realm of visual computing.

Keywords: 3D Scene Retrieval, Semantics, Semantic Information, 3D Object Retrieval, Deep Learning, Machine Learning, Visual Computing

DEDICATION

I dedicate my work to my thesis advisor and incredible mentor Dr. Bo Li, my caring family, and my loving partner Mason.

ACKNOWLEDGMENTS

I would like to thank Dr. Bo Li, my thesis supervisor, for his patience and invaluable guidance throughout the thesis development process. His provision of learning materials, research guidance, and encouragement to develop my skills as a researcher has proven beneficial and for that I am incredibly thankful. To be able to provide me with the skills and knowledge necessary to finish this project having started from a minimal understanding is incredibly encouraging and inspiring.

I would also like to give special thanks to my family for their encouragement throughout all stages of my academic career. To have evolved from a young child struggling with their times tables into a graduating senior in a technical field is an incredible advancement only with the encouragement of you all. However, I am still not great at my times tables.

Lastly, I would like to thank my partner Mason for the incredible patience that he has granted me throughout this process. I would be nowhere without your undying support.

TABLE OF CONTENTS

LIST OF ABBREVIATIONS

2D	Two-dimensional
3D	Three-dimensional
CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
VR	Virtual Reality

CHAPTER I: INTRODUCTION

Background

In recent decades, the advancement of computer technology has revolutionized numerous fields, including the realm of three-dimensional (3D) computer models. These models are virtual representations of physical objects or environments and serve as indispensable tools across a multitude of disciplines. Using digital computation, 3D computer models enable the visualization, analysis, and manipulation of complex spatial data with precision and flexibility. As the demand for immersive and interactive digital experiences continues to grow, such as with the popularization of virtual and augmented reality (VR/AR), understanding the principles and techniques underlying 3D modeling becomes increasingly important.

3D object retrieval refers to the process of searching for and retrieving three-dimensional (3D) objects from large databases based on their geometric properties, spatial relationships, or semantic attributes. Unlike traditional text-based or 2D image retrieval systems, 3D object retrieval involves querying for objects in a three-dimensional space, considering factors such as shape, size, orientation, and appearance. This field plays a crucial role in various applications, including computer-aided design (CAD), VR, and AR. The primary goal of 3D object retrieval is to develop efficient algorithms and techniques that can accurately and quickly locate relevant 3D objects from vast and diverse repositories, facilitating tasks such as content-based modeling, shape analysis, object recognition, semantic search, and scene understanding.

Increasingly complex as compared to single 3D object retrieval, **3D scene retrieval** involves the retrieval of entire three-dimensional (3D) scenes from large

databases based on their spatial configurations, semantic content, or contextual relationships. Unlike 3D object retrieval, which focuses on individual objects within a scene or as a singular entity, 3D scene retrieval aims to identify and retrieve entire scenes, including their overall layout, arrangement of objects, and environmental context. The challenge in 3D scene retrieval lies in effectively representing and matching complex 3D scenes, considering factors such as scene topology, object interactions, lighting conditions, and semantic annotations. Research in this area aims to develop robust algorithms and methodologies for indexing, searching, and retrieving 3D scenes from large-scale repositories, enabling tasks such as scene understanding, content-based scene retrieval, and scene-based navigation in virtual and augmented reality applications.

The integration of 2D scene images presents a significant dimension to the process, particularly concerning the retrieval of 3D scene shapes. While traditional 3D scene retrieval primarily deals with the identification and retrieval of complete 3D scenes based on their spatial configurations and semantic content, the incorporation of 2D scene images expands the scope to include the retrieval of 3D scene shapes derived from 2D representations. This approach entails extracting relevant shape features from 2D scene images and matching them with corresponding 3D scene shapes in the database. By leveraging 2D scene images, researchers aim to enhance the accuracy and efficiency of 3D scene shape retrieval algorithms, enabling tasks such as shape-based scene categorization, scene reconstruction from 2D images, and content-based scene retrieval in multimedia applications. However, challenges persist in effectively bridging the semantic gap between 2D scene images and their corresponding 3D shapes, as well as in handling variations in viewpoint, lighting conditions, and image quality. As such, ongoing research

endeavors focus on developing robust methodologies and algorithms to address these challenges, advancing the capabilities of 3D scene retrieval systems in analyzing and interpreting complex spatial scenes.

In the realm of 3D scene retrieval algorithms, contextual relationships play a pivotal role in enhancing the accuracy and efficiency of retrieval algorithms. Contextual relationships refer to the spatial and semantic dependencies between objects, as well as the overall scene layout and environmental factors. By considering contextual relationships, retrieval algorithms can better understand the spatial organization of objects within a scene, their functional or semantic associations, and the overall scene semantics. This information enables more meaningful scene matching and retrieval, allowing algorithms to identify scenes that not only contain similar objects but also share similar spatial arrangements or contextual semantics. Leveraging contextual relationships also facilitates tasks such as semantic scene categorization, scene completion, and scene reconstruction by providing additional cues and constraints to guide the retrieval process. As research in 3D scene retrieval progresses, further exploration and exploitation of contextual relationships are likely to lead to more advanced and effective retrieval algorithms.

One such research direction involves SceneNet, a large-scale dataset designed for training and evaluating artificial intelligence algorithms in the field of scene understanding and computer vision. Unlike many existing datasets that focus on individual images or objects, SceneNet provides rich, detailed 3D scenes captured from various indoor environments. These scenes include not only the geometry of the surroundings but also information about object semantics, textures, lighting conditions,

and spatial relationships. SceneNet is particularly valuable for tasks such as scene recognition, object detection and localization, semantic segmentation, and depth estimation in complex and realistic indoor environments. The dataset is widely used in research to develop and benchmark algorithms for tasks related to scene understanding, providing a valuable resource for advancing the capabilities of AI systems in real-world environments.

Overview

This thesis research project expands upon the proposed directions by Dr. Juefei Yuan's research in 3D scene shape retrieval with the development of SceneNet for large scale 3D scene retrieval (Yuan, J., et al., 2020). Because there is little research in the niche of 3D scene retrieval, we encounter difficulties in the conception of further development due to the lack of related retrieval benchmarks. However, there is enormous potential in addressing this semantic gap between 2D scene sketches and 3D scene models and provide innovative and meaningful research to contribute to the fields of computer vision and deep learning.

In this project, we propose a semantic sketch/image based 3D scene retrieval approach that will be able to accurately retrieve 3D scenes given users' sketches and images with a low computational cost that can be scalable to a large-scale retrieval. This strategy is based on a semantic tree for 2D/3D scenes which bridges the gap between 2D sketches/images and 3D scenes by introducing the common factor of semantic information. This is accomplished through a three-step process. Firstly, a scene semantic tree is built based on the semantic ontology from WordNet, the large lexical database of English (Yuan, J., et al., 2020). Next, the semantic attributes will be identified in the 2D query sketch/image via a deep learning based classification approach. Finally, the semantic similarity will be measured between the 2D query sketch or image's semantic attributes and compared to the nodes in the semantic tree, we will be able to compute the similarities between the 2D query sketch/image and 3D scene to determine the most relevant 3D scenes.

This thesis delves into the realm of 3D scene retrieval, exploring its complexities and potential applications in the context of evolving computer technologies. Beginning with an overview of the significance of 3D modeling in various disciplines and the growing demand for immersive digital experiences, the thesis navigates through the intricacies of 3D object and scene retrieval, elucidating the challenges and advancements in the field. The integration of 2D scene images and the exploration of contextual relationships further enrich the discussion, paving the way for innovative approaches to scene understanding and retrieval. Building upon existing research and proposing a semantic sketch/image-based 3D scene retrieval approach, this thesis seeks to bridge the semantic gap between 2D representations and 3D scenes, offering a scalable and efficient solution for retrieving relevant 3D scenes. By following a structured organization, this thesis aims to contribute to the advancement of knowledge in computer vision and deep learning, providing insights, methodologies, and directions for future research endeavors in the dynamic domain of 3D scene retrieval.

Thesis Organization

The organization of this thesis follows a structured approach designed to present a comprehensive understanding of the topic. The introductory section provides background information, outlining the research problem, and stating the objectives of the study. Following this, the literature review critically analyzes existing scholarship to contextualize the research within the broader academic discourse. The methods section details the proposed research methodology, alongside a thorough analysis of proposed algorithms addressing each research direction. Finally, the conclusions and future work section synthesizes the findings, discusses their implications, and proposes avenues for future research, thereby contributing to the advancement of knowledge in the field.

CHAPTER II: LITERATURE REVIEW

Machine Learning

Machine learning is a subfield of artificial intelligence (AI) that focuses on developing algorithms and statistical models that enable computers to learn from and make predictions or decisions based on data, without being explicitly programmed to do so. At its core, machine learning involves the construction of mathematical models that can recognize patterns and relationships within datasets, and then use this knowledge to generalize and make predictions about new, unseen data. This process typically involves training a model on a labeled dataset, where the model learns from examples provided, adjusting its parameters iteratively to minimize errors or discrepancies between its predictions and the actual outcomes. Machine learning encompasses various techniques, including supervised learning, where the model is trained on labeled data; unsupervised learning, where the model identifies patterns in unlabeled data; semi-supervised learning, where the model using both labeled and unlabeled data; and reinforcement learning, where the model learns by interacting with an environment and receiving feedback on its actions. Through its ability to extract insights and patterns from vast amounts of data, machine learning has become a cornerstone of numerous applications across industries, from image and speech recognition to recommendation systems, medical diagnosis, financial forecasting, and beyond.

Batta Manesh, 2019, delves into the vast realm of machine learning (ML), which is the scientific study of algorithms and statistical models enabling computer systems to execute tasks without explicit programming. ML algorithms play a pivotal role in numerous daily applications, shaping our digital experiences. For instance, the

effectiveness of web search engines like Google relies significantly on learning algorithms proficient in ranking web pages. Beyond search engines, ML algorithms find utility across diverse domains including data mining, image processing, and predictive analytics. The inherent advantage of employing machine learning lies in its ability to automate tasks once the algorithm acquires the requisite understanding of the data. Through this paper, a concise exploration and a glimpse into the future prospects of the expansive applications of machine learning algorithms are presented.

Furthermore, the paper expounds on the profound impact of ML algorithms on various facets of modern society, elucidating their pervasive presence in critical domains such as healthcare, finance, and autonomous systems. For instance, in healthcare, ML algorithms are instrumental in diagnosing diseases, analyzing medical images, and personalizing treatment plans. Similarly, in the financial sector, these algorithms aid in fraud detection, risk assessment, and algorithmic trading. Moreover, ML algorithms underpin the development of autonomous vehicles, drones, and robotic systems, driving innovation and reshaping industries. By shedding light on the multifaceted applications of ML, the paper underscores the transformative potential of these algorithms in revolutionizing how we interact with technology and address complex challenges across various domains (**Mahesh, B., 2019).**

Deep Learning

Deep learning is a subset of machine learning that focuses on utilizing artificial neural networks with multiple layers to model and interpret complex data representations. Unlike traditional machine learning algorithms, which may involve handcrafted feature extraction and selection, deep learning algorithms automatically learn hierarchical representations of data through successive layers of abstraction. Each layer of a deep neural network extracts increasingly abstract features from the input data, allowing the model to discern intricate patterns and relationships that may be difficult to capture with simpler models. Deep learning architectures, such as convolutional neural networks (CNNs) for image recognition and recurrent neural networks (RNNs) for sequential data, have demonstrated remarkable success in a wide range of tasks, including image and speech recognition, natural language processing, and autonomous driving. The scalability and adaptability of deep learning algorithms, coupled with advancements in computational power and data availability, have propelled deep learning to the forefront of AI research and applications, revolutionizing fields such as computer vision, healthcare, robotics, and more.

Voulodimos et. al. document the remarkable advancements of deep learning methodologies, which have surpassed previous state-of-the-art machine learning techniques across various domains, with computer vision emerging as a standout example. Offering a succinct overview, the paper delineates the key deep learning paradigms pivotal in tackling computer vision challenges, including Convolutional Neural Networks (CNNs), Deep Boltzmann Machines (DBMs), Deep Belief Networks (DBNs), and Stacked Denoising Autoencoders (SDAs). It provides a historical backdrop,

structural insights, and an assessment of their advantages and limitations, laying a foundation for understanding their effectiveness in dealing with diverse computer vision tasks such as object detection, face recognition, action and activity recognition, and human pose estimation.

Furthermore, the article offers a glimpse into the future trajectory of deep learning schemes in addressing complex computer vision problems, hinting at forthcoming directions and the attendant challenges. By encapsulating the current landscape and potential avenues for advancement, the article not only enriches the discourse on deep learning in computer vision but also serves as a guide for researchers navigating this dynamic field.

Visual Computing

Visual computing is a multidisciplinary field that integrates computer science, mathematics, and psychology to develop algorithms and techniques for processing, analyzing, and understanding visual data. It encompasses a broad range of applications, including image and video processing, computer vision, computer graphics, and visualization. Visual computing aims to bridge the gap between raw visual data and meaningful information, enabling computers to perceive, interpret, and interact with visual content in a manner akin to human perception. Techniques in visual computing include image filtering, feature extraction, object detection and recognition, 3D reconstruction, rendering, and virtual reality. From enhancing image quality to simulating realistic environments, visual computing plays a crucial role in diverse domains such as entertainment, healthcare, automotive, education, and scientific research. By harnessing the power of visual data, visual computing enables innovative solutions to complex problems, ranging from medical image analysis and autonomous navigation to augmented reality experiences and digital content creation.

Xie, Y., et al. report on the recent work in utilization of neural fields in visual computing. Neural fields, a subset of coordinate-based neural networks, are adept at parameterizing physical properties across space and time within scenes or objects, demonstrating remarkable success in addressing diverse visual computing challenges. Noteworthy applications include 3D shape and image synthesis, human body animation, 3D reconstruction, and pose estimation. The report meticulously reviews over 250 papers in the domain, offering a meticulous examination of neural field techniques. Part I of the report focuses on dissecting neural field methodologies, elucidating common components

such as conditioning, representation, forward mapping, architecture, and manipulation techniques. It provides insights into various approaches like prior learning, hybrid representations, and differentiable forward maps, laying a robust foundation for understanding the intricacies of neural field methods. ***(Xie, Y., et al., 2022).***

Furthermore, Part II of the report delves into the expansive array of applications where neural fields have made significant contributions to visual computing and beyond. Encompassing realms like 2D image processing, 3D scene reconstruction, generative modeling, digital humans, and robotics. It underscores the versatility of neural fields across diverse domains. Additionally, the report sheds light on their applications in adjacent communities such as medical imaging, audio processing, and physics-informed problems, further accentuating their transformative potential. Through its comprehensive coverage, the report not only delineates the breadth of topics addressed by neural fields but also underscores their enhanced quality and adaptability. Moreover, it introduces a dynamic companion website, facilitating community engagement and continual updates to track new developments in the field of neural fields applications in visual computing. In essence, the report stands as a cornerstone in understanding the current landscape, future prospects, and evolving applications of neural fields in visual computing. (***Voulodimos, A., et al., 2018).***

3D Object Retrieval

3D object retrieval is a specialized field within visual computing that focuses on developing algorithms and methodologies to retrieve three-dimensional objects from large databases based on their geometric properties, shape, texture, and other relevant features. Unlike traditional 2D image retrieval, where the focus is on matching visual content in 2D space, 3D object retrieval involves analyzing the spatial structure and surface characteristics of 3D models to identify similar objects or shapes. Techniques in 3D object retrieval include shape descriptors, which encode geometric information in a compact and discriminative manner, as well as indexing and similarity search algorithms tailored for handling 3D data efficiently. Applications of 3D object retrieval span various domains, including computer-aided design (CAD), manufacturing, cultural heritage preservation, virtual reality, and robotics. By enabling efficient search and retrieval of 3D objects, this field facilitates tasks such as shape recognition, object categorization, content-based modeling, and scene reconstruction, contributing to advancements in areas like product design, medical imaging, gaming, and architectural visualization.

View-Based 3D Object Retrieval

View-based 3D object retrieval is a specialized approach within 3D object retrieval that emphasizes the use of multiple views or perspectives of three-dimensional objects to facilitate their retrieval from large databases. Instead of relying solely on the geometric properties or shape descriptors of 3D models, view-based methods consider the appearance of objects from different viewpoints as crucial information for retrieval. This approach typically involves capturing or generating a set of 2D views or images of a 3D object from various angles and then extracting features or descriptors from these views to represent the object.

Techniques in view-based 3D object retrieval often include methods for view selection, feature extraction, and similarity measurement. View selection strategies aim to determine the most informative or discriminative views of objects to represent them effectively. Feature extraction involves extracting relevant visual features from each view, such as local descriptors, color histograms, or texture features. These features are then used to represent the object's appearance from different perspectives. Finally, similarity measurement techniques quantify the similarity between the views of query objects and those in the database, enabling the retrieval of objects with similar appearances from the dataset.

View-based 3D object retrieval has applications in various fields, including cultural heritage preservation, virtual reality, and augmented reality. By leveraging multiple views of 3D objects, this approach enhances the robustness and accuracy of retrieval systems, enabling more effective search and retrieval of objects in applications such as digital libraries, e-commerce, and content-based modeling.

Li (2012) proposes several 3D model retrieval techniques to deal with several major challenges in 3D model retrieval: developing better shape descriptors, supporting multi-modal queries, and improving 3D model alignment methods. The suggested approach to this is a view-based 3D model retrieval algorithm. Multiple chapters are included that present different algorithms for query-by-model retrieval using a so-called view context descriptor, query-by-sketch retrieval incorporating 2D-3D alignment, and query utilizing class information and hybrid features. The results of this research demonstrate the effectiveness of the proposed algorithms.

The article delves into the pressing need for efficient and accurate methods of searching for 3D models across various applications like Computer-Aided Design (CAD), online 3D model shopping, and entertainment production. Recognizing the significance of 3D model retrieval in these domains, the research explores different algorithms for extracting features from 3D models, focusing particularly on addressing gaps in supporting multi-modal queries and improving alignment accuracy by proposing a new 3D normalization technique named Minimum Projection Area (MPA)-based 3D model alignment algorithm (***Li, B., 2012).***

Gao et al. (2020) address the gap in deep learning success for computer vision, considering its application to 3D model retrieval is relatively limited and lacks a standardized benchmark for evaluating different deep learning-based features in this context. To address this, the performance of conventional neural network (CNN)-based deep learning features is evaluated across four widely used shape retrieval benchmarks (ETH, NTU60, PSB, and MVRED), based on various similarity measurement methods. Their research finds that Multi-View Convolutional Neural Network (MVCNN), which

explores the feature relationships among multiple views, outperforms single view based CNNs. Deep learning features also consistently remain robust to noise, highlighting one advantage of multi-view deep learning in enhancing retrieval accuracy (**Gao, Z., Li, Y., & Wan, S., 2020).**

3D Scene Retrieval

3D scene retrieval is a specialized field within content-based retrieval that focuses on retrieving entire three-dimensional scenes from large databases based on their visual content and structural properties. Unlike 3D object retrieval, which targets individual objects within scenes, 3D scene retrieval aims to identify complete environments or scenes that are similar or relevant to a given query scene. This task is particularly challenging due to the complexity and variability of 3D scenes, which may contain multiple objects, varying lighting conditions, occlusions, and spatial arrangements.

Techniques in 3D scene retrieval typically involve representing scenes using descriptors that capture their spatial layout, object distribution, and other relevant characteristics. These descriptors may include global features such as scene histograms, which summarize the distribution of visual properties within the scene, as well as local features describing specific regions or objects within the scene. Additionally, methods for scene segmentation and object recognition may be employed to identify and characterize individual components within the scene.

Similarity measurement plays a crucial role in 3D scene retrieval, as it quantifies the similarity between the query scene and scenes in the database based on their feature representations. This often involves comparing the descriptors of scenes using distance metrics or similarity measures such as cosine similarity or Euclidean distance. Advanced techniques such as deep learning-based approaches may also be utilized to learn representations directly from the data, capturing complex relationships and semantics within 3D scenes.

Applications of 3D scene retrieval span various domains, including virtual reality, gaming, urban planning, and augmented reality. By enabling efficient search and retrieval of 3D scenes, this technology facilitates tasks such as scene recognition, content-based scene modeling, and immersive virtual experiences, contributing to advancements in fields such as digital entertainment, architectural design, and urban simulation.

Semantic scene completion (SSC) aims to infer complete 3D geometry and semantics of a scene from limited observations. This area of research within computer vision is especially challenging due to occlusions and a limited field of view. Existing methods commonly use light detection and ranging (LiDAR), a remote sensing method that uses light in the form of a pulsed laser to measure variable distances. However, Kim et al. (2013) devises a camera-based SSC approach named Voxel-CRF (Please confirm) by utilizing MonoScene, a dense feature projection-based SSC technique, as the baseline.

VoxFormer, a two-stage framework, is proposed for SSC from images consisting of class-agnostic query proposals followed by class-specific semantic segmentation. It uses depth estimates to generate sparse voxel queries and employs deformable cross-attention and self-attention to complete the scene representations.

Voxel-CRF outperforms MonoScene on the widely used SemanticKITTI benchmark dataset, comprising Dense Semantic Annotations, Scene Dimensions, Voxel Grid Representation, and Semantic Labels (***Kim, B.-S., Kohli, P., & Savarese, S., 2013).***

Ansary et al. (2007) presents a 3D model search engine called FOX-MIIRE built upon the Adaptive Views Clustering (AVC) algorithm which determines the optimal number of views required to effectively describe a given 3D model while minimizing

redundancy and computational complexity. Additionally, FOX-MIIRE incorporates a probabilistic Bayesian method for visually retrieving 3D models similar to a query model, whether provided as a 3D model itself, a photo, or a sketch. This algorithm stands out as the first search engine capable of retrieving 3D models from photos, marking a significant advancement in the field of early 3D scene retrieval (**Ansary, T. F., Vandeborre, J.-P., & Daoudi, M., 2007).**

The field of 3D scene shape retrieval has garnered significant research attention based on the fact that (1) it is a natural and straightforward extension of the traditional meaning of content-based 3D shape retrieval which only involves retrieving a single 3D object; and (2) it requires an expansion of the knowledge, techniques, and algorithms used in single 3D object retrieval. Facing the challenge of having multiple objects within a singular scene, Yuan et al. (2020) present a comprehensive evaluation of 14 sketch-based and image-based 3D scene retrieval algorithms based on the two related benchmarks created by some of the authors. The benchmarks contain 2D sketch/image queries and 3D scene model targets across 10 categories for the basic benchmarks, and 30 categories for the extended benchmarks (**Yuan, J., et al. 2020).**

CHAPTER III: METHODS

Project Goals

This project represents a novel exploration into the matching of 2D scene sketches/images with their 3D counterparts at a semantic level, facilitated by a meticulously structured tree framework. To our knowledge, this endeavor marks the first attempt to develop a comprehensive framework specifically tailored for 3D scene retrieval through such an approach. The implications of this work are significant, potentially accelerating research in the field of sketch/image-based 3D scene retrieval while also shedding light on related areas of 2D/3D scene understanding. Our primary objectives in this project are delineated below, guiding our efforts toward the achievement of our overarching goals.

The first goal of this project proposal is to construct an **extensive hierarchical scene semantic tree**, leveraging WordNet as a foundation. This endeavor entails gathering a vast array of 2D/3D scenes to establish what will be the most comprehensive and singularly available scene semantic tree to date, encompassing over 10 million 2D/3D scene files across approximately 500 categories. Specifically, the repository will encompass around half a million 3D scene models (averaging 1,000 models per class), 5 million 2D scene images (averaging 10,000 images per class), and 5 million 2D scene sketches (averaging 10,000 sketches per class). Moreover, each scene category will be enriched with encoded semantic information, including distributions delineating scene object occurrence, co-concurrence, and spatial relations. This pioneering effort is poised to spearhead semantics-driven research in 3D scene retrieval, laying the groundwork for future advancements in the field.

The second goal of this project revolves around the proposal and implementation of a pioneering **semantic tree-based 3D scene retrieval framework**. This innovative approach aims to proficiently capture semantic information from both 2D sketches/images and 3D scene models, facilitating accurate measurement of similarities between their respective semantic representations. By effectively bridging the semantic disparity between 2D and 3D representations, this framework is poised to significantly enhance retrieval performance. The envisioned sketch/image-based 3D scene retrieval framework holds tremendous potential for diverse applications, spanning domains such as 3D printing, virtual reality, 3D cartoon animation, and mobile applications. Through this endeavor, we aim to unlock new frontiers in semantic-driven retrieval methodologies, fostering advancements with far-reaching implications across various industries and fields.

An extensive application research initiative is proposed for our semantics-driven 3D scene retrieval framework, with a particular focus on semantics-driven 2D scene sketch/image-based 3D scene retrieval. This endeavor aims to **explore and evaluate the efficacy of our framework** across a range of practical scenarios and use cases. By integrating semantic information extracted from 2D scene sketches/images into the retrieval process, we anticipate significant advancements in the accuracy and efficiency of 3D scene retrieval. This research thrust will entail rigorous testing and validation against diverse datasets and real-world applications, with the goal of demonstrating the robustness and versatility of our approach. Through this concerted effort, we aim to not only validate the effectiveness of our framework but also uncover new opportunities for

its application across various domains, including but not limited to virtual reality and 3D printing.

The fourth and final goal of this project is to establish a **robust scene semantic tree** that serves as a foundational infrastructure for a multitude of related applications. This scene semantic tree will not only bolster research efforts in sketch/image-based 3D scene retrieval but also provide invaluable guidance for advancements in related areas such as 2D and/or 3D scene understanding. By offering a comprehensive framework for encoding and organizing scene semantics, our work will facilitate endeavors ranging from object detection to scene classification, recognition, and retrieval. Through the provision of explicit guidance and direction, our scene semantic infrastructure promises to catalyze innovation and foster interdisciplinary collaboration across a spectrum of research domains. Ultimately, the establishment of this infrastructure will not only advance the field of 3D scene retrieval but also pave the way for transformative developments in the broader landscape of scene understanding applications.

Overview

Within this project, we introduce a semantic tree-based framework for 3D scene retrieval that comprises the following 5 steps:

1. Construction of 2D and 3D Scene Semantic Tree: The first step involves creating a semantic tree tailored for a large-scale warehouse environment, incorporating a diverse range of 2D/3D scene sketches, images, and models. This semantic tree functions as a structured network delineating semantic classes, attributes, and associated scene files.
2. Semantic Object Instance Segmentation of Queries: This step involves segmenting a scene query sketch or image into a cohesive set of semantic objects. For instance, an image depicting a kitchen would undergo segmentation to identify distinct semantic instances such as bottles, bowls, chairs, forks, tables, TV, and wine glasses. Each identified object instance is associated with its categorical name, along with information regarding its frequency of occurrence and spatial relations within the scene, collectively forming the semantic representation of the query.
3. Semantic Similarity Computation: Compute the semantic similarity between the semantics of the 2D query and that of each 3D target scene category, based on the extracted scene query's semantics information and the pre-learned scene semantics information for the 3D target scene category.
4. Hybrid Similarity Computation: This step involves computing a hybrid similarity measure between the query sketch/image and each 3D scene

model. This hybrid similarity is derived by integrating two distinct components: first, the high-level semantic similarity calculated in Step (3), which considers the semantic objects/stuffs contained within the scenes; second, the low-level pixel-based similarity, which is determined by training deep learning classification models directly on the query sketches/images and target 3D scenes or their view images. Deep learning methodologies are chosen due to their proven efficacy, as evidenced by their state-of-the-art performance across various applications (***Li, B., 2012).*** By combining both semantic and pixel-based similarities, this hybrid approach ensures a comprehensive evaluation of similarity, facilitating more accurate and robust scene retrievals.

5. 3D Scene Ranking: In this step, the query-target hybrid similarities are sorted, and the 3D scene models are ranked in descending order accordingly. This ranking process ensures that the most relevant 3D scene models, which exhibit the highest similarity to the query sketch/image, are positioned at the top of the list. By organizing the scene models based on their computed similarities, users can quickly identify and access the most suitable candidates for their specific needs or applications, streamlining the scene retrieval process and enhancing overall efficiency.

Research Objective 1: WordNet-Based Scene Semantic Tree Construction

WordNet offers a comprehensive taxonomy, comprising over 80,000 distinct synsets, rings of semantically equivalent words for the purpose of information retrieval, representing noun concepts, organized in a directed acyclic graph (DAG) network. This network delineates relationships like hyponyms, where, for instance, "table" is a hyponym of "furniture." A scene semantic tree mirrors this hierarchy, organizing 2D sketches, images, and 3D scene models according to WordNet's synset structure. Each class (or synset) in the tree possesses attributes such as "is-a", "has-part", or "is-made-of", derived from WordNet's gloss definitions. Leaf nodes in the semantic tree contain numerous 2D sketches, images, and 3D scene models corresponding to the leaf node's class, along with scene semantics information.

Traditionally, 3D model search approaches have focused on a Query-by-Model framework, utilizing existing 3D models as queries for retrieval. While this method offers simplicity, it deviates from the natural human inclination to search for 3D models. In scenarios such as product design, architectural planning, or cartoon animation creation, individuals typically begin by sketching their concepts on paper or digital devices like cellphones, tablets, or laptops to find similar 3D models. Similarly, in virtual reality environments and 3D games, users prefer interactive methods involving hand gestures or control consoles to craft 3D scenes based on existing models. Thus, the demand for retrieving 3D models based on human sketches has emerged as a crucial and desirable strategy, with implications spanning human-computer interaction, 3D animation, game design, virtual reality, and beyond. Despite recent advancements in sketch-based 3D model retrieval algorithms, many existing methods struggle with high computational

costs and low retrieval accuracy. This challenge stems from the significant semantic gap between 2D sketches and 3D models. Human sketches exhibit varied styles, iconic representations, high-level abstraction, and simplification, posing challenges in description and representation. In contrast, 3D models offer precise geometric representations. This semantic disparity complicates direct 2D-3D comparisons, rendering existing algorithms less effective, especially in large-scale sketch/image-based 3D scene retrieval scenarios.

To delineate the scene semantics for a specific category *S*, we construct three probability distributions based on its scene objects' occurrence, co-occurrence, and spatial relations.

1. **Object occurrence probability** $P(O_i|S)$: Describes the conditional probability that an object class appears within the scene.
2. **Object co-occurrence probability** $P(O_i, O_j|S)$: Describes the conditional probability that two differing object classes within a specified 3D scene appear simultaneously within the scene.
3. **Spatial relationship probability** $P(SR(O_i, O_j)|S)$: Describes the conditional probability that two differing object classes within a specified 3D scene have a certain spatial relationship that defines their occurrence. This relationship could describe an object that is near to another object, surrounds another object, or supports another object, for example.

Research Objective 2: 2D-3D Semantic Similarity Computation

By utilizing the scene semantic tree, our aim is to quantify the semantic similarity between 2D queries and 3D scenes. This involves assessing the semantic relatedness between the object semantics of the query—comprising names, object occurrences, co-occurrences, and their spatial relations distributions—and each scene category node in the 3D semantic tree. Consequently, this approach encompasses two key components: firstly, evaluating the semantic relatedness between their object and scene category names, and secondly, measuring the similarity between their respective semantics distributions. To gauge the similarities of these distributions, we will analyze and assess three distance metrics: the Jensen–Shannon distance D. M. Endres and J. E. Schindelin, 2003), Earth mover's distance (also known as Wasserstein-1 distance) (E. Levina and P. J. Bickel, 2001), and Frechet distance (alternatively referred to as Wasserstein-2 distance) (T. Eiter and H. Mannila, 1994). Our assessment of semantic relatedness is grounded in the utilization of WordNet, as elaborated below.

In measuring the semantic similarity between 2D queries and 3D scenes, our approach involves an intricate analysis rooted in the scene semantic tree developed in Research Objective 1. This analysis entails scrutinizing the object semantics of the query, encompassing factors such as names, object occurrences, co-occurrences, and spatial relations distributions, in relation to the scene category nodes within the 3D semantic tree. This process is structured around two pivotal aspects: firstly, establishing the semantic affinity between the names of objects and scene categories, and secondly, quantifying the resemblance between their respective semantics distributions. To effectively gauge these resemblances, we plan to employ three distinct distance metrics:

the Jensen–Shannon distance, Earth mover's distance, and Frechet distance. To highlight our assessment of semantic relatedness, we draw upon the comprehensive semantic framework provided by WordNet.

In determining the semantic relatedness value between a labeled semantic object of a 2D query and the name of a 3D scene category, the process of **word sense disambiguation** becomes crucial. This entails selecting the appropriate meaning among the various senses associated with the label name, as typically presented in resources like WordNet. Given that a word may carry different meanings in diverse textual contexts, determining the suitable sense is vital for accurate interpretation. Numerous methodologies have been suggested by researchers to address the challenge of word sense disambiguation.

Motivated by the Lesk algorithm's two hypotheses, our approach involves treating each semantic object's label within a query as its own contextual frame (Lesk, M., 1986). Through this lens, we conduct sense disambiguation by tallying the shared words between the gloss of the focal object's label and those of other objects' labels. Despite the logical interconnection between objects within a query, they often bear distinct semantic nuances, rendering the use of similarity metrics unsuitable. Thus, we opt for the Lesk relatedness metric to gauge their resemblances effectively. Nonetheless, several additional challenges persist, such as structuring the object labels within a segmented and labeled query into a coherent "sentence" for the Lesk algorithm's application, and addressing the issue of significantly duplicated words, where multiple instances of the same object may occur within a sentence.

To address these challenges, we delve into resolving specific queries surrounding the organization of object labels within a segmented and labeled query. Firstly, we aim to determine the optimal sequence to form a coherent "sentence" for the Lesk algorithm's application, ensuring that the contextual framework effectively captures semantic nuances. Secondly, we confront the issue of duplicated words, particularly prevalent when multiple instances of the same object occur within a sentence. Our task involves devising strategies to manage these repetitions effectively, maintaining the integrity of the semantic analysis while minimizing redundancy and maximizing computational efficiency.

The computation of **semantic relatedness $\boldsymbol{S(q_i N_i)}$** between a semantic class N_i of a 3D scene category and an object label qi within a query q relies on leveraging WordNet gloss and the semantic hierarchy present in the 3D semantic tree. This process involves determining the object-wise relatedness $R(q_i N_i)$. Various semantic similarity and relatedness metrics have been proposed by researchers to measure the relatedness of semantic concepts in WordNet. Banerjee and Pedersen introduced several measures, including lch, wup, and path, based on path lengths between concepts, along with hso, lesk, and vector for semantic relatedness (Banerjee, S., and Pedersen, T., 2002) Additionally, alternative semantic relatedness and similarities have been suggested in other studies. As we aim to enhance retrieval performance, we propose novel definitions for 2D-3D semantic similarity and related metrics, with a focus on incorporating the spatial relationship between different object instances within a scene, which we anticipate will yield further improvements.

Research Objective 3: Semantics-Driven 2D Scene Sketch-Based Retrieval

To overcome the semantic gap in 2D scene sketch-based 3D scene retrieval, we suggest pursuing the following two different research directions to develop a semantics-driven system for this purpose.

1. Large-Scale and/or Multimodal 2D Scene Sketch-Based 3D Scene Retrieval Benchmark: The utilization of a large-scale and/or multimodal 2D scene sketch-based 3D retrieval benchmark will allow for excellent performance and extended scalability. To be able to extend the benchmark to incorporate more scene categories and modalities would implicate increased 2D/3D format diversity for related applications where datasets are incredibly diverse.
2. Semantics-Driven 2D Scene Sketch-Based 3D Scene Retrieval: To enhance the effectiveness or efficiency of a 2D scene sketch-based 3D scene retrieval algorithm, it is imperative to harness the semantic information present in both the 2D scene sketches used as queries and the 3D scene models being searched. We suggest that various practical domains, such as online 3D scene retrieval, development of 3D entertainment content, and the advancement of autonomous driving technologies, stand to gain from retrieval methodologies that integrate semantic insights derived from both query sketches and target scenes. Our strategic approach involves the adaptation or application of our previously proposed semantic tree-based 3D scene retrieval framework to effectively address this retrieval challenge.

Research Objective 4: Semantics-Driven 2D Scene Image-Based Retrieval

The augmentation of additional classes within a dataset poses a significant challenge, potentially leading to ambiguity during the retrieval process. In such scenarios, the model may struggle to discern between classes sharing similarities, resulting in a failure to accurately distinguish them. To address this issue, we propose leveraging semantic information extracted from detected objects in query images to establish correlations with scene semantics during the 3D scene retrieval process. By incorporating semantic object detection, we anticipate mitigating the inevitable ambiguities inherent in larger and more complex datasets, particularly in cases where similar classes coexist.

As an extension of our proposed work, we aim to enhance the accuracy and efficiency of 2D scene image-based 3D scene retrieval algorithms by integrating semantic information from both query images and target 3D scene models. Utilizing the scene semantic tree, we intend to incorporate semantic context into the retrieval process, a dimension not explored by any of the participating methods in the SHREC track. We believe that various applications, such as online 3D scene retrieval, 3D entertainment content development, and autonomous driving systems, stand to benefit significantly from leveraging extracted semantic information in both queries and targets.

To facilitate a semantics-driven 3D scene retrieval strategy, we propose the adoption of a Region-Based Convolutional Neural Network (R-CNN) for generating region proposals and extracting semantic objects from 2D image scenes. Augmenting existing R-CNN architectures, such as Mask R-CNN or Faster R-CNN, with a more robust segmented dataset holds the potential to enhance performance in both current and future benchmark evaluations. This approach underscores our commitment to advancing

the field by leveraging state-of-the-art techniques to address challenges associated with semantic understanding in 2D-to-3D scene retrieval.

CHAPTER IV: CONCLUSIONS AND FUTURE WORK

This project extends further than individual contributions and expanding research endeavors and directions at The University of Southern Mississippi and serves to make a lasting impact that reaches far beyond the academic community. This project has the potential to catalyze transformative shifts by pioneering novel pathways in research. Its impact extends beyond 3D scene retrieval, enriching broader research landscapes. The growing availability of 3D scene data is sparking excitement among researchers, opening new avenues for research endeavors. This research has applications in virtual reality, 3D printing, and beyond, promising major advancements.

In pursuit of constructing a foundational infrastructure, the research embarked on the ambitious goal of establishing an extensive hierarchical scene semantic tree. We propose constructing an extensive hierarchical scene semantic tree, leveraging WordNet as a foundational framework. The primary goal was to establish a comprehensive repository of 2D/3D scenes encompassing over 10 million files across approximately 500 categories. Each scene category was enriched with encoded semantic information, laying the groundwork for semantics-driven research in 3D scene retrieval and related areas. By pioneering this effort, the project aimed to spearhead advancements in semantic-driven retrieval methodologies, fostering innovation across various research domains.

In addressing the semantic gap between 2D sketches/images and 3D models, we aimed to develop methodologies for quantifying semantic similarity. Through the utilization of the scene semantic tree, this research sought to quantify the semantic

similarity between 2D queries and 3D scenes. By evaluating semantic relatedness metrics and assessing distributions, the goal was to bridge the semantic gap between 2D sketches/images and 3D models, enhancing retrieval performance. The proposed methodologies aimed to provide a comprehensive evaluation of similarity, facilitating more accurate and robust scene retrievals across diverse datasets and applications.

Focusing on leveraging semantic insights, this project aimed to develop a semantics-driven framework for 2D scene sketch-based 3D scene retrieval. The objective is to develop a semantics-driven framework for 2D scene sketch-based 3D scene retrieval. Through the establishment of large-scale benchmarks and the integration of semantic information into retrieval algorithms, the project aimed to enhance retrieval accuracy and efficiency. By leveraging semantic insights derived from both query sketches and target scenes, the goal was to unlock new opportunities for applications spanning virtual reality, 3D entertainment, and autonomous systems.

Finally, with a focus on semantic understanding, we aim to advance methodologies for 2D scene image-based 3D scene retrieval. In addressing the challenges of semantic understanding in 2D-to-3D scene retrieval, this research proposed leveraging semantic information extracted from query images to establish correlations with scene semantics. By adopting region-based convolutional neural networks and augmenting existing architectures, the goal was to enhance retrieval performance and address challenges associated with semantic understanding. Through these efforts, the project aims to advance the state-of-the-art in 2D scene image-based 3D scene retrieval, with implications for diverse practical domains.

This project represents a comprehensive exploration into semantics-driven 3D scene retrieval, aiming to bridge the semantic gap between 2D sketches/images and 3D models. By pursuing distinct research objectives focused on semantic tree construction, similarity computation, and semantics-driven retrieval strategies, the project has contributed valuable insights and methodologies to the field. Through innovative approaches and rigorous methodologies, this research lays a solid foundation for future advancements, with implications spanning various industries and fields.

Applications

The realm of Virtual Reality (VR) sketching stands poised at the intersection of creativity and technology, offering a dynamic platform for artistic expression, design prototyping, and immersive storytelling. With the advent of advanced 3D scene retrieval techniques, VR sketching experiences are poised to undergo a profound transformation, empowering creators with unprecedented access to a vast repository of pre-existing models and scenes. This fusion of VR and semantic-level information search promises to streamline the creative process, enabling designers and artists to seamlessly integrate complex 3D elements into their virtual canvases. From architectural visualization to conceptual design ideation, the applications of VR sketching are multifaceted and far-reaching, unlocking new avenues for exploration and innovation across industries. Whether sculpting virtual landscapes or conceptualizing architectural blueprints, the integration of semantic-tree 3D scene retrieval frameworks holds the potential to revolutionize the way we conceptualize and interact with virtual environments, ushering in a new era of immersive digital creativity.

Recognizing the growing field of 3D printing and its growing popularity, this novel sketch-based 3D model/scene retrieval system shows great promise. Since those interacting with 3D printers often do not spend the time to build a 3D model or scene from scratch, our proposed retrieval framework can facilitate the building of 3D scenes based on hand-drawn sketches. This shows great potential for a wide range of applications including individuals, schools, or even companies.

Future Work

This work could be idealized by potentially expanding the scope of SHREC (Shape Retrieval Contest) tracks over the coming years. SHREC tracks are specific challenges within the SHREC competition where researchers and practitioners develop and evaluate algorithms for the retrieval of 3D shapes or scenes from large databases. Through the introduction of new SHREC tracks such as extended 2D scene sketch-based and image-based retrieval, as well as large-scale scene retrieval using semantic-tree-based approaches, we aim to provide a comprehensive platform for advancing 3D scene retrieval techniques.

By hosting and expanding the scope of these SHREC tracks, we can facilitate comparative evaluations and offer a common testing ground for researchers. This approach has the potential to significantly drive progress in the field and attract broader interest from both seasoned researchers and students. Furthermore, to address the challenges posed by increasing dataset complexity and potential ambiguities, future work could focus on leveraging semantic information to enhance the accuracy and efficiency of 2D scene image-based 3D scene retrieval algorithms.

By incorporating semantic object detection techniques, such as Region-Based Convolutional Neural Networks (R-CNN), we could extract meaningful semantic information from both query images and target models. This approach has the potential to mitigate ambiguities and improve retrieval performance, benefiting applications such as online 3D scene retrieval and autonomous driving. Moreover, augmenting existing R-CNN models with robust segmented datasets, such as Mask R-CNN or Faster R-CNN, could lead to further improvements in benchmark evaluations and future research endeavors. Expanding SHREC tracks to encompass these advancements could serve as a catalyst for innovation in the field of 3D scene retrieval.

REFERENCES

Ansary, T., Vandeborre, J., & Daoudi, M. (2007). 3D-Model search engine from photos. Proceedings of the 6th ACM International Conference on Image and Video Retrieval, 89–92.

Banerjee, S. and Pedersen, T. (2002). An adapted Lesk algorithm for word sense disambiguation using WordNet. In CICLing, pages 136–145.

Eiter, T. and Mannila, H. (1986). Computing discrete Fréchet distance. Technical report, 1994.

Endres, D. and Schindelin, J. (2003). A new metric for probability distributions. IEEE Transactions on Information Theory, 49(7):1858–1860.

Gao, Z., Li, Y., & Wan, S. (2020). Exploring Deep Learning for View-Based 3D Model Retrieval. ACM Transactions on Multimedia Computing, Communications, and Applications, 16(1), 1–21.

Kim, B., Kohli, P., & Savarese, S. (2013). 3D Scene Understanding by Voxel-CRF. 2013 IEEE International Conference on Computer Vision, 1425–1432.

Lesk, M. (1986). Automatic sense disambiguation using machine readable dictionaries: how to tell a pinecone from an ice cream cone. In Proceedings of the 5th Annual International Conference on Systems Documentation, SIGDOC 1986, Toronto, Ontario, Canada, 1986, pages 24–26.

Levina, E. and Bickel, P. (2001). The earth mover's distance is the mallows distance: Some insights from statistics. In Proceedings of the Eighth International Conference On Computer Vision (ICCV-01), Vancouver, British Columbia, Canada, July 7-14, 2001 - Volume 2, pages 251–256.

Li, B. (2012). View-based techniques for 3D model retrieval [Thesis]

Mahesh, Batta. (2019). Machine Learning Algorithms - A Review. 10.21275/ART20203995.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. Computational Intelligence and Neuroscience, 2018, 1–13.

Xie, Y., Takikawa, T., Saito, S., Litany, O., Yan, S., Khan, N., Tombari, F., Tompkin, J., Sitzmann, V., & Sridhar, S. (2022). Neural Fields in Visual Computing and Beyond (arXiv:2111.11426). arXiv.

Yuan, J., et al. (2020). A comparison of methods for 3D scene shape retrieval. Computer Vision and Image Understanding, 201, 103070.