Literature Survey on speech to 3D scene generation.

Mr. Manthan Turakhia[#1], Mr. Umang Nandu[#2], Mr. Preyash Shah[#3], Mr. Siddharth Sharma[#4], Mr. Sagar Korde[*5]
Student, Dept. of Information Technology[#]
Professor, Dept. of Information Technology[*]
K. J. Somaiya College of Engineering, Mumbai, India.
manthan.turakhia@somaiya.edu[1], umang.nandu@somaiya.edu[2], prayesh.shah@somaiya.edu[3], siddharth.sharma@somaiya.edu[4] , sagar.korde@somaiya.edu[5]

Abstract— 3D scenes and graphics are widely used in the creative industry. However, the entire task of imagination and then depicting the same as 3D graphics is done manually today, which consumes a lot of time, not to mention the inability to depict the scene precisely as imagined. We aim to reduce human efforts for the same by generating 3D scenes described by the user with precision, and near real-time generation.

On the other hand, some industries currently lack the use of appropriate technology to make their tasks easier and more captivating, such as the education industry. We intend to replace the existing methods of teaching and learning by using speech to 3D scene generation to depict exactly what the professor is trying to explain.

Keywords— Speech to Scene, 3D Warehouse, Linguistic Analysis, Spatial Relationship, Natural language processing.

I. Introduction

We examine the task of speech to 3D scene generation. There is a myriad of applications for this technology, mainly for creative and educational industries. Designers can use this technology to interpret and display their thoughts and imaginations. Students can be taught with a near real-time graphical depiction of the topic. Commercial meetings and conference sessions can make the most of this technology.

While there has been some previous work researched and implemented on this technology, those projects have been implemented on static databases with a large, albeit fixed, set of images and database size. We aim to take it up a notch by dynamically generating scenes on-the-go; previous work is based on text to scene. consequently, previously researched or implemented projects involve use of a human language however, not natural, meaning that there have been restrictions to the kind of words used to describe the scene, whereas we target to generate a scene out of anything and everything spoken by the user.

In general, the algorithm uses the concept of semantic parsing for identifying objects and their nature, and spatial knowledge to map the identified objects and scenes according to the users' imaginations. The goal is to achieve a near real-time generation and display of consecutive scenes, one after another, so as to practically match the on-going speech of the user, such that the scene on the screen will keep changing and manipulating to match the speaker's requirements and descriptions. In future, we plan to introduce complex Artificial Intelligence algorithms to this project so as to make it learn the usual demands and imaginations for each user, which would make the scene generation more reliable, efficient and faster at the same time.

II. Problem definition

3D scenes and graphics are widely used in the creative industry. However, the entire task of imagination and then depicting the same as

3D graphics is done manually today, which consumes a lot of time, not to mention the inability to depict the scene precisely as imagined. We aim to reduce human efforts for the same by generating 3D scenes described by the user with precision, and near real-time generation.

On the other hand, some industries currently lack the use of appropriate technology to make their tasks easier and more captivating, such as the education industry. We intend to replace the existing methods of teaching and learning by using speech to 3D scene generation to depict exactly what the professor is trying to explain.

III. Literature Review

A couple of projects focus on text to 2D Graphics Generation, 3D Graphics Generation, Natural Language to scene Generation As well. These mechanisms are experienced in the projects like, "Learning Spatial Knowledge for Text to 3D Scene Generation", "Real-time Automatic 3D Scene Generation", etc. These systems act as the frameworks for depiction of natural language descriptions. [2]

WordsEye [1] is the project which focuses on converting the semantic intents of the user, as an input to the system, into a graphical representation. But as these semantic intents are inherently ambiguous, the generated scene might not accurately match the user's requirement.

A real-time system available in market was The Carsim system. [3] The system translated the traffic accident report to 3D scene. The limitations of this system was that it could not generate the scene outside of the traffic area.

These systems shows that the generation of scenes is a challenging task and time consuming as well, and accuracy is another major milestone for the application. The algorithms, methodologies used will be discussed in further part of the paper.

IV. Methodology

The process starts by taking an input from the user in the form of voice which will in turn be converted into text using the Google Cloud Speech-to-Text API.

A. Linguistic Analysis

The text generated from the speech to text API is than tagged and broken down and parsed using the parts-of-speech tagger. The resulting output is a parse tree that gives a structural representation of the sentence. This breakdown identifies subjects, objects, adjectives, articles, verbs and adverbs of the sentence.

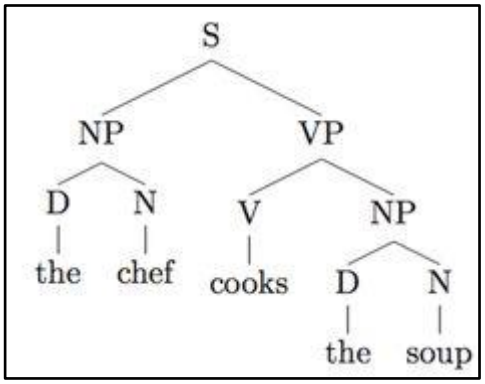

Fig 1: Parse tree

- S -> Sentence
- NP -> The Noun Phrase, in this case "chef" which is the subject of the sentence.
- VP -> The Verb Phrase, which is the predicate.
- V -> The Verb, in this case is "cooks".
- D -> Determiner.

B. Semantic Representation

The next phase involves building a relationship between these fragments of words called the dependency structure and converting this structure into a semantic representation. Fig 2 shows the dependency structure of the sentence "John said that the cat was on the table". Fig 3 shows the semantic representation of that sentence.

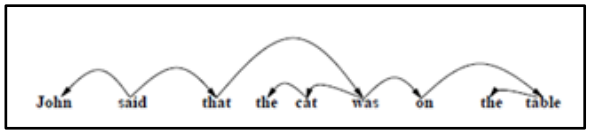

Fig 2: Dependency structure for John said that the cat.

```
(("node2" (:ENTITY :3D-OBJECTS ("mr_happy")
          :LEXICAL-SOURCE "John" :SOURCE SELF))
 ("node1" (:ACTION "say" :SUBJECT "node2"
          :DIRECT-OBJECT  ("node5" "node4" "node7")...))
 ("node5" (:ENTITY :3D-OBJECTS  ("cat-vp2842")))
 ("node4" (:STATIVE-RELATION "on" :FIGURE "node5"
          :GROUND "node7"))
 ("node7" (:ENTITY :3D-OBJECTS
          ("table-vp14364" "nightstand-vp21374"
           "table-vp4098" "pool_table-vp8359" ...))))
```

Fig 3: Semantic Representation was on the table.

C. 3-D Objects and Object Database

Ultimately, the semantic fragments will be used to define the low level 3-dimensional graphical figures also known as depictors. These depictors will control the object visibility, colour, position, size and orientation of the object across 3 dimensions.

D. Google 3D Warehouse

Google 3D Warehouse is an open source dataset library consisting of over 12000 3D models from various different categories. Fig 6 show the classification of each model into various sub categories. For example, couch and chair can be found under the same category of "couch". This can reduce retrieval time from the database thus reducing the application's latency time and improving its response time.

The 3D models can be downloaded and rendered using "Sketchup" which is a model viewer.

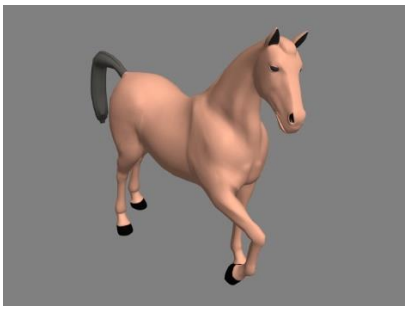

Fig 4: 3-D Model of a horse

Fig 5: 3-D Model of a bicycle

text	category	text	category
chair	Chair	round	RoundTable
lamp	Lamp	laptop	Laptop
couch	Couch	fruit	Bowl
vase	Vase	round table	RoundTable
sofa	Couch	laptop	Computer
bed	Bed	bookshelf	Bookcase

Fig 6: Categorization of lexical terms in the dataset.

E. Spatial Relationship

Spatial relations define the look and feel of the scene. These relations define how objects in a scene are spatially related to each other, their positions, distances and orientation. Fig 6 shows a similar example; where the dog's position is between the T.V and the couch, dog facing towards the T.V is its orientation, and the dog being 3 meter away from the T.V is the distance.

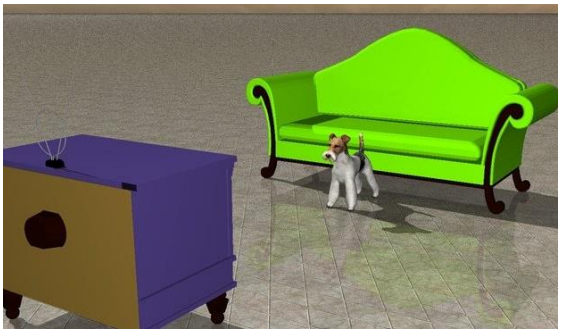

Fig 7: The dog is between a television and a couch. The dog faces towards the television which is 3 meters away.

VI. Conclusion

This paper proposes a Speech to 3D-Scene generation system. As an input, natural language text it provided, to construct 3D scenes. These scenes are based on spatial relationships. Various systems are available with their concepts and applications, but the accuracy is compromised and this is required to be addressed and propose a better solution. We aim to develop a system with speech input freedom, accuracy and dynamicity to

user's expectation. We also aim to provide the Geometric scene learning experience.

VII. References

[1] Bob Coyne and Richard Sproat.[2001] "WordsEye: an automatic text-to-scene conversion system". In Proceedings of the 28th annual conference on Computer graphics and interactive techniques.

[2] Lee M Seversky and Lijun Yin.[2006]. "Real-time automatic 3D scene generation from natural language voice and text descriptions". In Proceedings of the14th annual ACM international conference on Multimedia.

[3] R. Johansson, A. Berglund, M. Danielsson and P. Nugues,[2005] "Automatic Text-to-Scene Conversion in the Traf fic Accident Domain", The Nineteenth International Joint Conference on Artificial Intelligence, pages 1073–1078, 30 July-5 August 2005.