Text to 3D Scene Generation

Sneha N. Dessai
Department of Computer Engineering
GEC, Farmagudi, Ponda, Goa, India

Prof. Rachel Dhanaraj
Department of Computer Engineering
GEC, Farmagudi, Ponda, Goa, India

Abstract- Creating 3D graphics is a difficult and time-consuming process. We see the need for a new paradigm in which the creation of 3D graphics is both effortless and immediate. Thus we propose a Text to 3d Scene generation system that incorporates user interaction. A user provides a natural language text as an input to this system and the system then identifies explicit constraints on the objects that should appear in the scene. From these explicit constraints system then uses various priors to identify implicit constraints on the objects. The system also identifies scene type from various constraints. Then candidate scene will be generated that will be continuously improved as per the user interaction and thus final scene will be rendered as an output.

Keywords – Natural language processing, implicit constraints, explicit constraints, scene template, render scene, scene graph.

I. INTRODUCTION

Nowadays 3D graphics are used in many applications, such as cartoons, animations and games. However, creating 3D graphics is a difficult and time-consuming task. The user must learn to use a complex software package before he/she can actually create the artwork. A new paradigm which makes the creation of 3D graphics effortless and convenient is needed. It should be possible to describe 3D scenes directly from natural language[3]. Language offers a convenient way for designers to express their creative goals. Systems that can interpret natural descriptions to build a visual representation allow non-experts to visually express their thoughts with language.

Unfortunately, several key technical challenges restrict our ability to create text to 3D scene systems. Natural language is difficult to map to formal representations of spatial knowledge and constraints. Furthermore, language rarely mentions common sense facts about the world, that contain critically important spatial knowledge. For example, people do not usually mention the presence of the ground or that most objects are supported by it. As a consequence, spatial knowledge is severely lacking in current computational systems.[1]

For a text to scene system to understand more natural text, it must be able to infer implicit information not explicitly stated in the text. For instance, given the sentence “there is an office with a red chair”, the system should be able to infer that the office also has a desk in front of the chair. This sort of inference requires a source of prior spatial knowledge. We propose learning this spatial knowledge from existing 3D scene data. However, since the number of available scenes is small, it is difficult to have broad coverage. Therefore, we also rely on user interaction to augment and grow the spatial knowledge. User interaction is also natural for scene design since it is an inherently interactive process where user input is needed for refinement.[1]

II. LITERATURE SURVEY

The conversion of natural language text into graphics has been investigated in a few projects. NALIG [5] is an early example of them that was aimed at recreating 2D scenes. One of its major goals was to study the relationship between space and prepositions. NALIG considered simple phrases in Italian of the type subject, preposition, and object that in spite of their simplicity can have ambiguous interpretations.
The Put system [2] which shared our goal of making graphics creation easier was limit to spatial arrangements of existing objects. And what’s more, its input was restricted to the form Put (X, P, Y)$^+$, where X and Y were objects, and P was a spatial preposition.

The Carsim system [6][7] which converted a traffic accident report to 3D scene cannot generate scenes outside the traffic field. We build on prior working data driven scene synthesis to automatically extract general spatial knowledge from data: knowledge of what objects occur in scenes, and their expected spatial relations.

Prior work has shown the task remains challenging and time intensive for non-experts, even with simplified interfaces. Prior work in text to 3D scene generation focused on collecting manual annotations of object properties and relations, which are used to drive rule based generation systems. Regrettably, the task of scene generation has not yet benefited from recent related work in Natural Language Processing.

A related line of work focuses on grounding referring expressions to referents in 3D worlds with simple colored geometric shapes. More recent work grounds text to object attributes such as color and shape in images[1] . [1]ground spatial relationship language in 3D scenes (e.g., to the left of, behind) by learning from pair wise object relations provided by crowd workers.

i.WordsEye[3]

WordsEye focuses on translating the semantic intent of the user, as expressed in language, into a graphic representation. Since semantic intent is inherently ambiguous, the resulting 3D scene might only loosely match what the user expected. Fig 1 and Fig 2 shows an example.

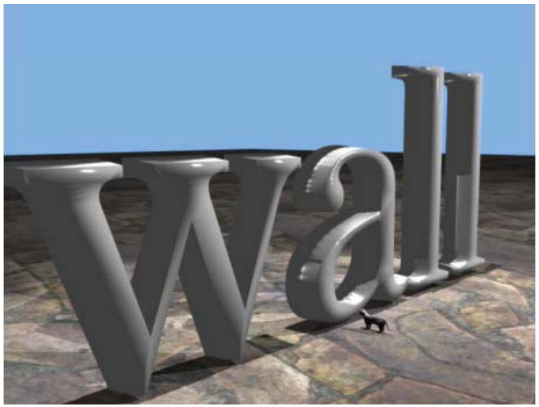


Fig1: The cat is facing the wall

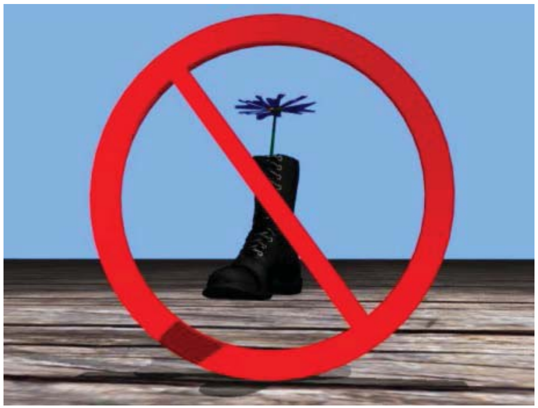


Fig2: The blue daisy is not in the army boot

ii. Real-time Automatic 3D Scene Generation [4]

This system is an initial framework which focuses on the depiction of natural language descriptions in real-time. They have assumed that the descriptions contain spatial relationships. The linguistic challenges of more general descriptions have not been addressed. Fig 3 and fig 4 shows an example.

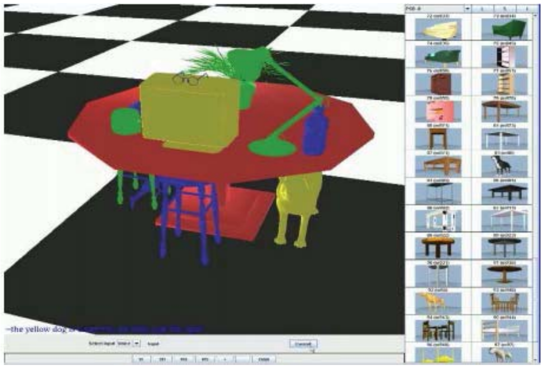


Fig 3: A scene composed of 12 objects generated from voice input in real-time. The following are excerpts from the input description : A green

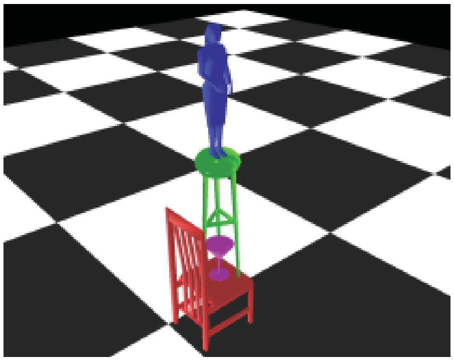
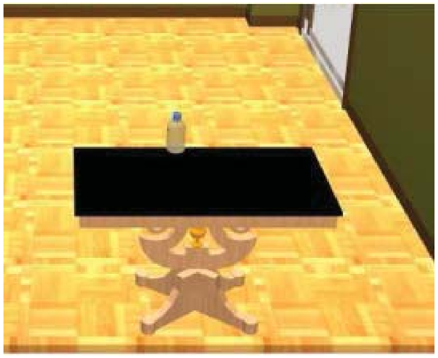


Fig.4: Placement of multiple objects from the following input descriptions: The green stool is on the red chair. A purple wine glass is under

plant is on the red table. The blue water bottle is next to the green lamp. The blue stool is under the red table to-wards the front. the green stool. The blue lady is on the green stool.

iii. Learning Spatial Knowledge for Text to 3D Scene Generation [1]

In this paper, they have presented a deterministic approach for mapping input text to the parsed scene template. But automatic learning of how to parse text describing scenes into formal representations by using more advanced semantic parsing methods is not included.



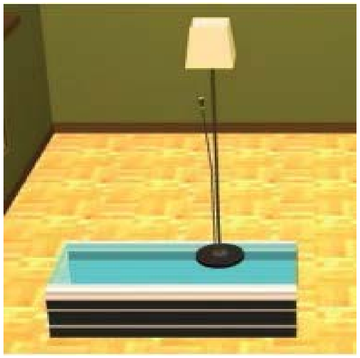


Fig.5 : A water bottle instead of wine bottle is selected for "There is a bottle of wine on the table in the kitchen". In addition, th selected table is inappropriate for a kitchen.

Fig. 6: A floor lamp is incorrectly selected for the input "There is a lamp on the table".

III. PROPOSED FRAMEWORK

In this proposed system, the user will be able to provide a brief scene description in natural language as input. The system parses this text to a set of explicitly provided constraints on what objects should be present, and how they are arranged. This set of constraints should be automatically expanded by using prior knowledge so that "common sense" facts are reflected in the general scene – an example is the static support hierarchy for objects in the scene (i.e. plate goes on table, table goes on ground). The system generates a candidate scene and then the user is free to interact with it by direct control or through textual commands. The system can then leverage user interaction to update its spatial knowledge and integrate newly learned constraints or relations. The final output is a 3D scene that can be viewed from any position and rendered by a graphics engine.

Algorithm

i. Template Parsing [1]: Parse the textual description of a scene into a set of constraints on the objects present and spatial relations between them. In order to identify objects the system will use extraction algorithm like Rapid Automatic Keywords Extraction (RAKE) algorithm. Some tools like Named Entity Recognition(NER) or Parts-Of-Speech (POS) tagger can also be used.

ii. Inference [1] : Expand this set of constraints by accounting for implicit constraints not specified in the text using learned spatial priors.

iii. Grounding [1] : Given the constraints and priors on the spatial relations of objects, transform the scene template into a geometric 3D scene with a set of objects to be instantiated.

iv. Scene Layout [1] : Arrange the objects and optimize their placement based on priors on the relative positions of objects and explicitly provided spatial constraints.

v. Interaction and Learning : Provide means for a user to interactively adjust the scene through direct manipulation and textual commands. Use any such interaction to update the system's spatial knowledge so it better captures the user's intent.

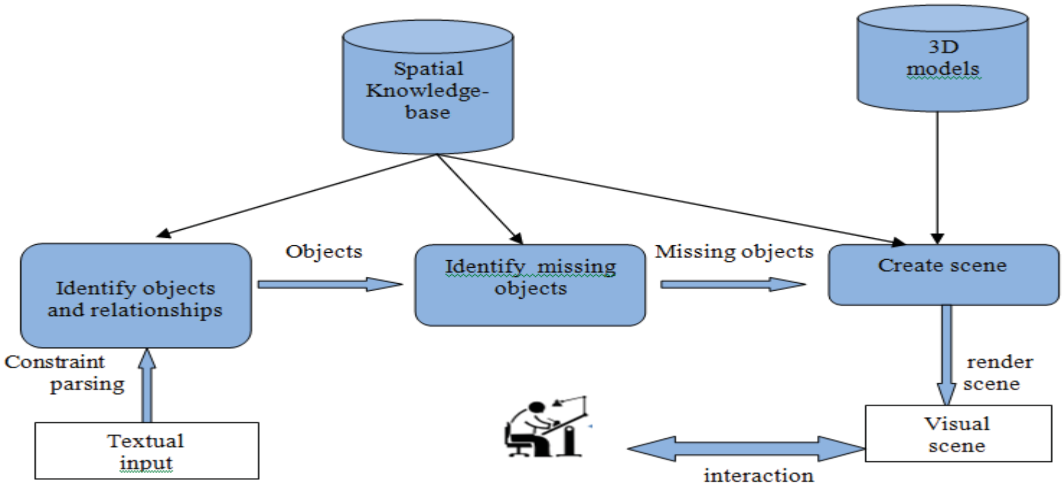


Fig 7. Architecture of Text to 3D scene

IV.CONCLUSION

This paper proposes a text to 3D scene generation system that takes natural language text as input to construct 3D scenes based on spatial relationships and provides user interaction. There are various existing papers that has done considerable amount of work in the same domain but since there were lot of flaws and limitations, there is a need for advanced system that will improve and refine such existing systems. This paper mainly focuses on user friendly system that would give more benefits to the user. Also choices made by the user will be learned and that knowledge would be used in future queries. Thus the author feels that such a system would provide blank slate to the user where he can actually paint with words.

REFERENCES

[1] Angel X. Chang, Manolis Savva, and Christopher D. Manning.[2014]. "Learning spatial knowledge for text to 3D scene generation". In Proceedings of Empirical Methods in Natural Language Processing(EMNLP) .

[2] S. R. Clay and J. Wilhelms[1996]. "Put: Language-Based Interactive Manipulation of Objects". IEEE Computer Graphics and Applications, pages31–39,March1996.

[3] 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.

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

[5] M. Di Manzo, G. Adorni, and F. Giunchiglia,[1986] "Reasoning about scene descriptions", IEEE Proceedings – Special Issue on Natural Language, 74(7):1013–1025

[6] K. Church.[1988] "A Stochastic Parts Program and Noun Phrase. Parser for Unrestricted Text." In Proceedings of the Second Conference on Applied Natural Language Processing, pages 136–143, Morristown, NJ.

[7] 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.