A Survey on Challenges in Text to Scene Generation

Yashaswini.S[1], Dr Shylaja S S[2]

[1]Assistant Professor, Department of CSE, Cambridge Institute of Technology

[2]Chairperson &Head, Department of CSE, PES University

Abstract: ***A long standing goal of computer vision is to build a system that can automatically understand a scene from an image by extracting semantic concepts and depicting information. Visual representations are always better than narrations, images can be visually expressed as 3D scenes by creating a virtual environment. NLP techniques can be used in text processing and intelligence can be built artificially to the system using deep learning techniques.***

Keywords: visual representation, virtual environment, 3D scenes

1. Introduction

Designing scenes is currently a creative task that requires significant expertise and effort in using complex design interfaces has people use natural language to describe real and imaginary environments. The Scene construction is the process of building realistic, three-dimensional representations or models, of real world environments, such as rooms, landscapes or buildings. The generated scenes can be evaluated by human experts. However, the benefits of having such models are great. Scene Construction also known as Scene Rendering or Scene Modeling which involves Computer Vision and Pattern Recognition.

Computer vision is a field that includes methods for acquiring, processing, analyzing and understanding images in general, high-dimensional data from the real world in order to produce numerical or symbolic information. Computer Vision teaches computers to see, to tell natural language sentence understood by image or scene or vice versa. Computer Vision changes mental behavior of system, it is important for better future. Artificial intelligence is a field of computer vision having numerous applications like Scene Generation, Entertainment, HealthCare, Education, Medicine, Surveillance, Security and Robotics. The challenges in computer Vision domain is to map 2D image projection of an object caused in retina to 3D scale because objects occlude themselves due to limited resource of 2D images.

This is an approach to propose a hybrid framework involving Graphics and Artificial Intelligence for constructing scene from text description, in which knowledge-based technology is adopted to represent the common sense of the real world. The framework extracts spatial relationships of objects described by texts and applies Part of Speech (POS) tagging, which is a machine learning technique used to assigns a part of speech to each word in an input sentence and parsing, assigns a parse tree to an input sentence, describing the syntactic structure of the sentence and helps in object identification from textual input.

A spatial reasoning algorithm is engaged to decide the object layout in the scene. Several critical issues like occlusion, constraints (implicit, contact, Co-reference), degrees of freedom, shape, size , Placement strategies and resizing of objects are taken into consideration in the spatial reasoning. This framework provides a scene which is realistic and interactive, the challenge is to devise a better algorithm to increase accuracy and speed of automatically generated Scene.

2. Survey on Challenges

During the investigation, the work carried out so far on text to Scene generation includes parsing the given natural language text, finding out the noun, preposition in the sentences. The noun in the sentences helps to identify the objects that are required to build the scene and preposition helps to find the semantic and spatial relation among the objects using NLP and Machine Learning Techniques. Once the objects are identified the placement criteria of objects have to be decided based on constraints like implicit constraints and relation between them.

The Highlights on challenges in Existing Works are as follows:-

Angel Chang, Will Monroe, Manolis Savva, Christopher Potts and Christopher D. Manning [1] have discussed the challenges involved in fidelity and plausibility of generated 3D scenes using human judgment as metric for evaluation of generated scenes

A new approach for Text to 3D Scene generation system has been proposed by Sneha N. Dessai and Prof. Rachel Dhanaraj [2] 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 implicit constraints of the objects are identified along with scene type and 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.

Nabeel Sabir Khan , Adnan Abid,Muhammad Shoaib Farooq, Yaser Daanial Khan, Bilal Hassan,Awais Kamran, Aneesa Abbasi[3] have proposed an effective framework to process input in textual form using NLP engine. The Constraint Based Grammar (CBG) is used to eliminate maximum occurrence of ambiguity in noun fragmentation process. The sentences are tokenized, normalized and tagged using penn Tree Bank and objects identified and spatial relations are represented in 3DScene

The Natural Language can be grounded into concrete spatial and implicit constraints. The explicit constraints are extracted with learned spatial knowledge to infer missing objects and layouts in scene. The text is converted into scene type and objects by checking its noun phrases, the adjectives define the properties of objects .The spatial relation is obtained by dependency pattern. The text is converted into graph of nodes and links representing objects and semantic relationships. The scene graphs should support static hierarchy, object properties and constraints then each object is transformed into matrix of position, orientation and scaling. The priors like object occurrence, support hierarchy, segmenting parent surface, and relative positions are determined. Cosine similarity is applied to obtain scores to match keywords to spatial relations as proposed by Angel Chang,Will Monroe,Manolis Savva,Christopher Potts and Christopher D. Manning [4]

Angel X. Chang, Manolis Savva and Christopher D. Manning[5] have used interactive text to 3Dscene generation that learns explicit constraints, prior observation of spatial arrangements using natural language. interactive system allows to manipulate and refine spatial knowledge using Amazon Mechanical Turk , a crowd sourcing platform to infer constraints, co -references and spatial relations using Gaussian on pair wise object position and orientation. Automatic resizing, addition of new objects and deletion of existing objects are supported with respect to view centric spatial relation.

The natural sentential descriptions of RGB-D scenes are exploited in order to improve 3D semantic parsing. Nouns, pronouns and verbs are identified for each object and also solves co-reference and resolution problem. This prediction model exploits potentials computed from text and RGB-D imagery to reason out class of 3D objects, scene type ,aligns noun pronoun with referred visual objects. The Parts Of Speech(POS) tags of all sentences are extracted and problem of co-reference resolution problem is solved by using Stanford co-reference system as suggested by Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, and Sanja Fidler [6] .

Daniel Bauer [7] has proposed a technique for semantic parsing into directed graphs. Graphs conveniently capture co-reference and hierarchical meaning. Input text describes functional aspects. The low level graphs depicts spatial relations between objects.HRG are used to generate context free strings. A sentence is decomposed into low level graphs to represent spatial complexities corresponding to 3D-Scene.

The framework which automatically converts arbitrary descriptive text into 3D-Scene is proposed by Christian Spika, Katharina Schwarz, Holger Dammertz and Hendrik P A Lensch[8].The system parses a user-written input text, extracts information using NLP and identifies relevant units. Every object is associated with appropriate 3D model based on object-to-object relations which evaluates spatial dependencies of entities. The resulting location is obtained based on heuristic like distance and position to create realistic and interactive 3D-Scene in natural looking virtual environment.

Bob Coyne , Owen Rambow , Julia Hirschberg, and Richard Sprout[9] have used series of complex menu, dialog boxes to easily create a wide variety of 3D-scene . Natural language offers an interface that is intuitively used by anyone without special training. Scenario Based Lexical Knowledge Resource(SBLR) is used to semantically categorize words, relation between predicates its linguistic information based on valance patterns. Frame Semantic supports verbs, nouns and relations robustly. Text is converted into semantic nodes and roles. The semantic resources used are WordNet, FrameNet, and PropBank. The input is parsed and semantically analyzed and mined using FrameNet, WordNet and textual corpora. Depiction strategies are decided based on spatial tags and metadata to create spatial reasoning and scene composing virtual 3D scene.

The importance of spatial relation in understanding language, converting them to 3D scenes intuitively and immediately by anyone without special skill or training. This system incorporates geometric and semantic knowledge about objects and their parts, thus improving system ability to infer relations automatically. Bob Coyne , Richard Sprout and Julia Hirschberg[10] have proposed the system that resolves verbs to semantic frames using SBLR and maps them to corresponding poses and spatial relations. It gives contextual information about action and location using human annotation extracted via Amazon Mechanical Turk.

Liuzhou Wu, Zelin Chen[11] have proposed a system to convert text to 3D Scenes, the input is parsed to obtain dependency structure .The depiction rule converts semantic representation into set of low level depictors, poses, relation, color, heads, dependants. The spatial relation helps to find relative positions and distance among objects. This approach helps to handle implicit and conflicting constraints.

The automatic scene can be generated using text for storytelling ,scene descriptions are constructed in real time using spatial relationship among different objects by using the model Proposed by Lee M Seversky, Lijun Yin[12] . When natural language is used along with automatic scene generators it benefits non graphic areas to visualize in 3D ,reducing graphic specific knowledge .It uses placement algorithms where as a preprocessing step it converts polygon into voxel , voxelized image to object . The surface and regions of object are identified, partitioned to coherent regions, location is validated , and then object is added to scene. Then object name, relation types, attributes are identified. These phrases describe spatial relations among object, then valid placement is computed. The textual input undergoes part of speech tagging ,the tagger builds up a tree representing language components. The structure is traversed to locate components. Scene instance is constructed using geometric and voxel representation of referenced object.

The Framework that uses semantic database, boundary boxes ,occupancy space and contact constraints on each object, group ,regular and random interval placement generates natural 3Dscene considering six degrees of freedom is proposed by Yoshiaki Akazawa, Yoshihiro Okada and Koichi Niijima[13]. Here the input is mapped to high dimensional vectors whose surface is planar, the collision is avoided by occupancy space and occupancy distance considering parent child relationship. While placing different components together connection width, depth and face should be considered to find connectable faces based on size of floor. The placement of objects can be based on random layout or regular interval method by individually separating components.

Richard Johansson, Anders Berglund, Magnus Danielsson and Pierre Nugues[14] have proposed system that automatically converts narratives into 3D scenes. This system animates the generated scenes using temporal relations between events. CarSim system consists of narratives of car accidents consisting of space description ,movements and direction. The information is extracted using natural language and machine learning is used to solve co-references, order events temporally. So this approach enables people to imagine traffic situation.

WordsEye has used natural language as medium for describing visual ideas and images to acquire artistic skills on window based interface, it automatically convert text into 3D-scenes by depicting entities and objects involved, their poses, grips, shapes and spatial tags and relations, color, kinematics, attributes like twisting ,bending and tries to avoid conflicting constraints by specifying path, orientation and position as specified by Bob Coyne, Richard Sprout[15]

3. Conclusion

Prior Work in scene generation relies purely on rule based method, so the challenges like placement(explicit, implicit, contact) constraints, co-reference, occlusion, size and shape of the object ,metrics and human judgments can be used to evaluate generated scene by correctly mapping objects to images.

4. Future Work

A hybrid framework can be developed which involves more than one method so it can handle as many as challenges as possible, by creating a dynamic environment which supports resizing of objects and also provides a feedback mechanism to improve the scenes generated.

The system generates the static images of Scene, once the scene is generated the changes cannot be incorporated in same scene should repeat the processing steps again to get modifications. There is no feedback mechanism to express the view about generated Scene. The accuracy of object placed at correct place based on implicit constraints can be improved.

References

[1] Angel Chang , Will Monroe , Manolis Savva,Christopher Potts and Christopher D. Manning “Text to 3D Scene Generation with Rich Lexical Grounding” Stanford University, Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing(IJCNLP), Held at Beijing, China, 26-31 July 2015, Volume 1,P15-1006

[2] Sneha N. Dessai, Prof. Rachel Dhanaraj “Text to 3D Scene Generation” in International Journal of Latest Trends in Engineering And Technology(IJLTET),Volume 6 Issue 3 January 2016, pp 255-258

[3] Nabeel Sabir Khan , Adnan Abid,Muhammad Shoaib Farooq, Yaser Daanial Khan, Bilal Hassan,Awais Kamran, Aneesa Abbasi “ Constraint Based NLP Engine” Department of Computer Science/SST ,University of Management & Technology Lahore Pakistan. VAWKUM transaction on Computer Sciences volume 6, number 2, March - April 2015, pp 01-06

[4] Angel Chang , Will Monroe , Manolis Savva,Christopher Potts and Christopher D. Manning “Learning Spatial Knowledge for Text to 3D Scene Generation” .In Proceedings Of Empirical Methods in Natural Language Processing(EMNLP), held on 25-29 October 2014 in Doha, Qatar, pp 101,

[5] Angel X. Chang, Manolis Savva and Christopher D. Manning “Interactive Learning of Spatial Knowledge for Text to 3D Scene Generation” Proceedings of Workshop on Interactive Language Learning, Visualization and interfaces, Association for Computational linguistics , Baltimore ,Maryland, USA, 27 June 2014 ,pp 14-21

[6] Chen Kong, Dahua Lin, Mohit Bansal, Raquel Urtasun, Sanja Fidler “What are you talking about? Text-to-Image Coreference” Tsinghua University, University of Toronto TTI Chicago, Computer Vision and Pattern Recognition (CVPR), 2014, pp 3558-3565

[7] Daniel Bauer “Understanding Descriptions of Visual Scenes Using Graph Grammars” Columbia University New York. In proceedings of the twenty-seventh Association for the Advancement of Artificial Intelligence (AAAI) conference on Artificial Intelligence, 2013, pp 1656-1657

[8] Christian Spika, Katharina Schwarz, Holger Dammertz and Hendrik P A Lensch “AVDT- Automatic Visualization of Descriptive Texts” Institute of Media Informatics, Ulm University, Germany Proceedings of Vision, modeling and visualization, The Euro graphics Association Held at Berlin , Germany ,4-6 October 2011,pp 129-136

[9] Bob Coyne , Owen Rambow , Julia Hirschberg, and Richard Sproat “Frame Semantics in Text-to-Scene Generation ”Columbia University, New York NY, USA and Oregon Health & Science University, Beaverton, Oregon, USA,14th International Conference on Knowledge Based And Intelligent Information and Engineering Systems, ACM Digital Library ,KES 2010, Cardiff, UK, September 8-10, 2010, Proceedings, Part IV, pp 1-9

[10] Bob Coyne, Richard Sprout and Julia Hirschberg "Spatial Relations in text-to-Scene Conversion" Columbia University, New York NY, USA and Oregon Health & Science University, Beaverton, Oregon, USA, In Computational Models of Spatial Language Interpretation (COSLI), Workshop at spatial cognition, 8 August 2010, pp 9-16

[11] Liuzhou Wu, Zelin Chen " A constraint-based Text-to-Scene Conversion System " School of computer Science and Engineering South China University Of Technology, in the Proceedings of the 2009 International Conference on Computational Intelligence and Software Engineering (CISE 2009),Held at China on 11-13 December 2009,pp 1311-1313

[12] Lee M Seversky, Lijun Yin "Real-time Automatic 3D Scene Generation from Natural Language Voice and Text Description" In Proceedings of 14th annual ACM international conference on Multi-media, ACM ,New York ,USA, 2006 ,pp 61-64

[13] Yoshiaki Akazawa, Yoshihiro Okada and Koichi Niijima "Automatic 3D Scene generation based on contact constraints" International *Conference* on Computer Graphics and Artificial Intelligence, 3IA Held at Limoges, France, 23 May 2005

[14] Richard Johansson, Anders Berglund, Magnus Danielsson and Pierre Nugues "Automatic Text-to-Scene Conversion in the Traffic Accident Domain" LUCAS, Department of Computer Science, Lund University, International Joint *Conference* on Artificial Intelligence, 2005, pp 1073-1078

[15] Bob Coyne, Richard Sprout "WordsEye: An Automatic Text-to-Scene Conversion System" .In Proceedings of the 28th annual conference on Computer graphics and interactive techniques, SIGGRAPH, 2001, pp 487- 496