ISSN 2319 – 7595
Volume 8, No.2, March - April 2019
International Journal of Information Systems and Computer Sciences
Available Online at http://warse.org/IJISCS/static/pdf/file/ijiscs25822019.pdf
https://doi.org/10.30534/ijiscs/2019/25822019

2D-to-3D Scene Generation and Rendering using CNN and Spatial Knowledge Representation

Gokula Nath G[1], Praveen Sankar[2], Vishnu Vijayakumar[3]
Thejus K Nair[4], Paul Joseph[5]
[1]MLMCE, India, ggokulanath@gmail.com
[2]MLMCE, India, sankar_praveen@hotmail.com
[3]MLMCE, India, ktpna.vishnu@gmail.com
[4]MLMCE, India, thejusknair123@gmail.com
[5]MLMCE, India, pauljoseph24y@gmail.com

ABSTRACT

Picture to 3D scene generation framework creates a conceivable 3D graphical scene from a given 2D advanced picture input. Because of overpowering utilization of 3D models in computer games, robot routes, virtual situations and even in field of interior designing there is a developing enthusiasm for 3D scene generation, scene comprehension and 3D demonstrate recovery. Here proposed another framework which is the mix of profound learning with spatial information portrayal to construct a rearranged 3D demonstrate from a solitary picture. The info picture is changed over into characteristic language depiction by utilizing a neural system. At last an improved 3D scene created from this depiction utilizing semantic parsing and spatial learning. The framework is extremely compelling at delivering important 3D scene straight forwardly from a 2D picture. The new framework additionally presented a system for rendering and controlling the scene through iterative info directions.

Key words: Deep learning**,** Natural language description, Neural framework, Rendering, Semantic parsing, Spatial knowledge

1. INTRODUCTION

We are living in a 3D world since 3D graphics are used in many applications, such as cartoons, animations and video games. Creating 3D graphics on computers is complex and time consuming. User should learn to use a complex software package before he can actually create it. Basic model is shown in figure 1.Most previous works have proposed to describe 3D scenes directly from natural language [3], [6], [7], [11]. A new system which makes the creation of 3D graphics directly from 2D scene effortless and convenient is needed.

In this paper, proposed a framework which combines both Convolution Neural Network (CNN) model [1] and spatial knowledge representation (SKR) with semantic parsing [2]. Spatial knowledge is an important aspect of the world and is often not expressed explicitly in natural language. The new system incorporates user interaction by giving a query in the form of 2-dimensional image. The model working in two phases. At first converts it into natural language description and then generate a relevant 3D scene from this with good

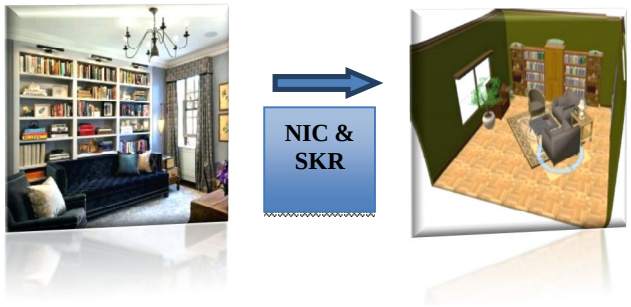

Figure 1: Basic Model, Input 2D scene to output 3D scene

clarity in the final stage. In I2T [4] framework, use semiautomatic method to parse images from the Internet in order to build an and–or graph (AoG) for visual knowledge representation. In baby talk [5], it automatically generates natural language descriptions from images that exploits both statistics gleaned from parsing large quantities of text data and recognition algorithms from computer vision. Generating image descriptions using deep visual-semantic alignments [8], sentence model [9] and grounded learning of a CCG semantic parser [10] are the other introduced approaches.

There are some existing work for generating 3D scene from textual descriptions. The WordsEye system [3], Learning Spatial Knowledge [2] and Rich Lexical Grounding [11] had addressed the text to 3D scene generation. WordsEye is a system for automatically converting text into representative 3D scenes. It relies on a large database of 3D models and poses to depict entities and actions. In Rich lexical grounding, introduced a dataset of 3D scenes annotated with natural language descriptions and learn from this data how to ground textual descriptions to physical objects. Also

proposed another system for text to 3D scene generation that incorporates user interaction [7].

The rest of the paper is organized as follows. The following section discusses about related work. The section 3 shows the framework description which is the new framework. The section 4 discusses about result and comparison with other approaches. Finally, conclude the work at section 5 and also shows the future scope.

2. RELATED WORKS

Natural language descriptions from visual data has long been studied in previous works. Baby talk [5] is one of the system where it automatically generate natural language descriptions from images that exploits both statistics gleaned from parsing large quantities of text data and recognition algorithms from Andrej Karpathy suggested a deep visual-semantic system model [8]that generates natural language descriptions of image regions based on weak labels in form of a dataset of images and sentences, and with very few hard coded assumptions. Benjamin Z. Yao discusses about I2T [4] which follows three steps: 1) the input images are decomposed into their constituent visual patterns by an image parsing engine 2) the image parsing results are converted into semantic representation in the form of Web ontology language (OWL) and 2) the image parsing results are converted into semantic representation in the form of Web Ontology Language(OWL) and 3) a text generation engine converts the result from previous steps into

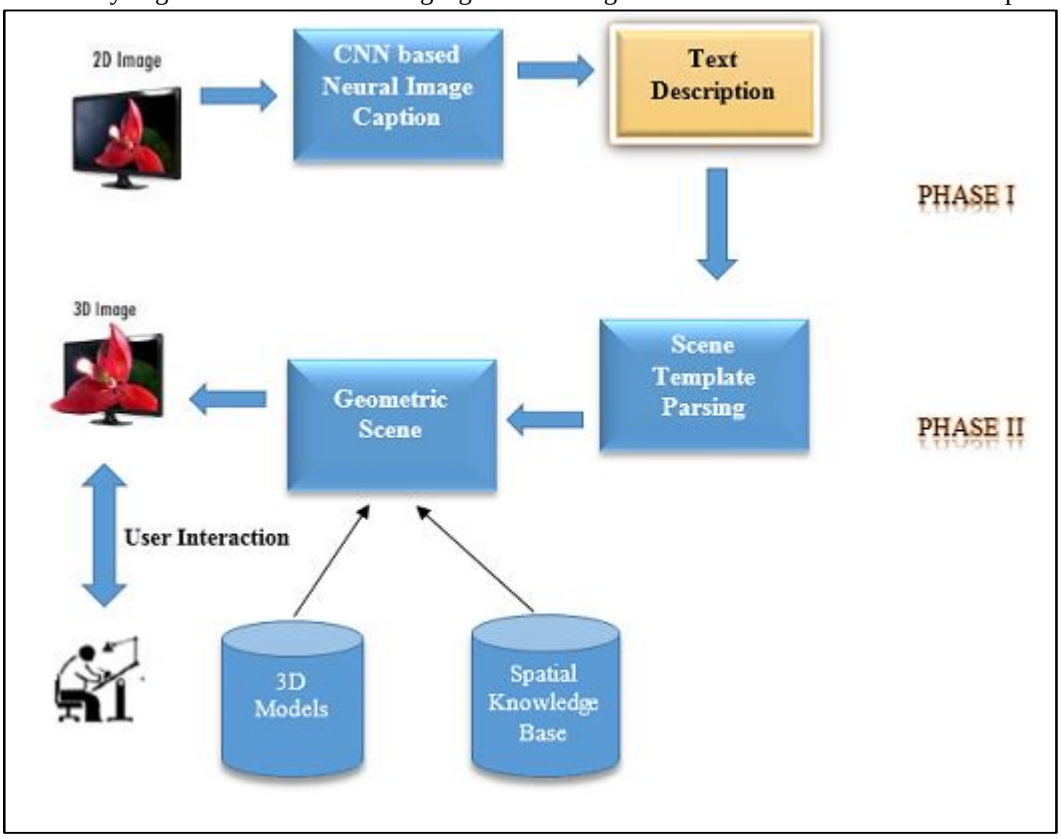

Figure 2: System Architecture

semantically meaningful, human readable, and query-able text reports.

Sentence model [9] predict good sentences for images that people like. The intermediate meaning representation is one key component in this model as it allows benefiting from distributional semantics. YoavArtzi introduced a system which uses grounded learning of a CCG semantic parser [11] that includes a joint model of meaning and context for executing natural language instructions.

Many attempts are also studied in 3D scene generation from natural language description. Bob Coyne discusses one of such system, WordsEye [3] where it automatically converting text into representative 3D scenes. WordsEye relies on a large database of 3D models and poses to depict entities and actions. Since semantic intent is inherently ambiguous, the resulting 3D scene might only loosely match what the user expected. Angel Chang discusses Rich lexical grounding [11] in introduced a dataset of 3D scenes annotated with natural language descriptions and learn from this data how to ground textual descriptions to physical objects.

Sneha N. Dessai introduced a Text to 3D Scene Generation [7] 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.

3. PROPOSED MODEL

In the proposed framework, user will able to generate 3D scene from a single input 2D image. The system working in two phases as shown in figure 2. At first, input is converted into relevant textual description which is then transformed to corresponding 3D scene in the final phase.

a) Neural Image Caption (NIC)

Image to text conversion can be done using a deep convolution neural network (CNN) [1]. CNNs can produce a rich representation of the input image by embedding it to a fixed-length vector. The last hidden layer in CNN used as an input to the Recurrent Neural Network(RNN) decoder that generates sentences. This model is called the Neural Image Caption, or NIC. RNN encodes the variable length input into a fixed dimensional vector, and uses this representation to decode it to the desired output sentence.

b) Scene Template Parsing

Textual description of the scene is parsed into a set of constraints on the objects present and spatial relations between them. To capture the objects present and their arrangement, represent scenes as graphs where nodes are objects in the scene, and edges are semantic relationships between the objects. Figure 3 shows an example.

One important property which is captured by the scene graph representation is that of a static support hierarchy, i.e., the order in which bigger objects physically support smaller ones: the floor supports tables, which support plates, which can support items. Static support and other constraints on relationships between objects are represented as edges in the scene graph [2].

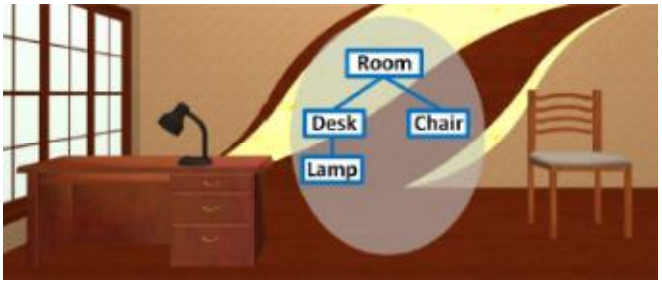

Figure 3: Generated scene for "There is a room with a desk and a lamp. There is a chair to the right of the desk." The inferred scene hierarchy is shown in the center.

c) Geometric Scene

Concrete geometric representation of a scene called as a geometric scene. It consists of a set of 3D model instances – one for each object– that capture the appearance of the object. Generate a geometric scene from a scene template by selecting appropriate models from a 3D model database and determining transformations that optimize their layout to satisfy spatial constraints.

d) Spatial Knowledge Base

It is on the idea of abstract scene types describing the occurrence and arrangement of different categories of objects within scenes of that type. For example, kitchens typically contain kitchen counters on which plates and cups are likely to be found. The type of scene and category of objects condition the spatial relationships that can exist in a scene.

e) User Interaction

The system provides a user to interactively adjust and rendering the scene by repositioning the objects, changing object's geometry through direct manipulation and textual commands.

4. RESULTS AND DISCUSSION

Proposed framework generates 3D scene which requires information that must be extracted from query image. Input query image has to be processed for extracting a description from it and then convert into corresponding 3D scene.

Comparison to alternative approaches: One of the common methods used for 3D scene generation is semantic scene completion method [12]. We can compare our model with this approach. Due to the GPU memory constraints, our network output resolution is lower than that of input volume. This results in less detailed geometry and missing small objects. New model covers this issue and generating more accurate results with sample trained dataset. With this framework, user is free to interact by direct control or through commands. The system can then leverage user interaction to update its spatial knowledge and integrate newly learned constraints or relations.

5. CONCLUSION AND FUTURE SCOPE

In this paper, introduced a new user friendly model which combines both NIC and spatial knowledge representation. A 2D input image is converted to natural language description and then generate a 3D scene from this using spatial knowledge base. NIC is based on a convolution neural network that encodes an image into a compact representation, followed by a recurrent neural network that generates a corresponding sentence. Spatial inference is used for achieving plausible results in the text-to-3D scene generation task. The model is effective since it generates good quality image and scene rendering is also allowed by incorporating user interaction. An obvious improvement would be to include more advanced semantic parsing methods for the automatic learning of how to parse text describing scenes into formal representations.

ACKNOWLEDGEMENT

I would like to express my sincere gratitude and thanks to my colleagues for their consistent support in completing this work.

REFERENCES

1. OriolVinyals, Alexander Toshev, SamyBengio, and DumitruErhan.**Show and Tell: Lessons learned from the 2015 MSCOCO image captioning challenge**, *IEEE Trans. on Pattern analysis and Machine intelligence* , Vol .XX, No. XX, Month 2016. https://doi.org/10.1109/TPAMI.2016.2587640
2. Angel X. Chang, ManolisSavva and Christopher D. Manning**,Learning Spatial Knowledge for Text to 3D Scene Generation**, *International journal of latest trends in Engineering and Technology*.
3. Bob Coyne and Richard Sproatj. **WordsEye: an automatic text-to-scene conversion system**, In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques.*
4. Benjamin Z. Yao, Xiong Yang, Liang Lin, MunWai Lee, and Song-Chun Zhu. **I2T: Image parsing to text description**, *Vol. 98, No. 8, August 2010 | Proceedings of the IEEE.* https://doi.org/10.1109/JPROC.2010.2050411
5. G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg, **"Baby talk: Understanding and generating simple image descriptions,"** *in CVPR, 2011.* https://doi.org/10.1109/CVPR.2011.5995466
6. Sneha N. Dessai and Rachel Dhanaraj. **Creation of 3D scene from raw text,** *IEEE International Conference on Recent Trends in Electronics Information Communication Technology, May 20-21, 2016, India.* https://doi.org/10.1109/RTEICT.2016.7808075
7. Sneha N. Dessai and Rachel Dhanaraj. **Text to 3D scene generation**, *International journal of latest trends in Engineering and Technology*.
8. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma,Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, **ImageNet large scale visual recognition challenge,** *2014.* https://doi.org/10.1007/s11263-015-0816-y
9. Ali Farhadi, Mohsen Hejrati , Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian1, Julia Hockenmaier, David Forsyth, **Every picture tells a story:Generating sentences from images,** in ECCV, 2010. https://doi.org/10.1007/978-3-642-15561-1_2
10. YoavArtzi and Luke Zettlemoyer. 2013. **Weakly supervised learning of semantic parsers for mapping instructions to actions.***Transactions of the Association for Computational Linguistics.* https://doi.org/10.1162/tacl_a_00209
11. Angel Chang, Will Monroe, ManolisSavva, Christopher Potts and Christopher D. Manning. **Text to 3D scene generation with rich lexical grounding,***pages 53–62, Beijing, China, July 26-31, 2015. c 2015 Association for Computational Linguistics.* https://doi.org/10.3115/v1/P15-1006
12. Shuran Song, Fisher Yu, Andy Zeng, Angel X. Chang, ManolisSavva and Thomas Funkhouser. **Semantic Scene Completion from a Single Depth Image,***arXiv: 1611.08974v1 [cs.CV] 28 Nov 2016.*