**Activity Recognition System**

Project report submitted for

**4h Semester Minor Project**

**in**

**Department of CSE**

By

**Aditi Agrahari (17100002)**

**Charchit Dhawan (17100013)**

**Harsh Singh (17101021)**



**Department of Computer Science and Engineering**

**Dr. Shyama Prasad Mukherjee**

**International Institute of Information Technology, Naya Raipur**

**(A Joint Initiative of Govt. of Chhattisgarh and NTPC)**

**Email:** [**iiitnr@iiitnr.ac.in**](mailto:iiitnr@iiitnr.ac.in)**, Tel: (0771) 2474040, Web:** [**www.iiitnr.ac.in**](http://www.iiitnr.ac.in/)

**CERTIFICATE**

This is to certify that the project titled “Activity Recognition System” by Aditi Agrahari, Charchit Dhawan and Harsh Singh has been carried out under my/our supervision and that this work has not been submitted elsewhere for a degree/diploma.

(Signature of Guide)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Dr.Muneendra Ojha**

**Asst. Professor**

**Department of CSE**

**Dr. SPM IIIT-NR**

**May, 2019**

**Declaration**

I declare that this written submission represents my ideas in my own words and where others' ideas or words have been included, I have adequately cited and referenced the original sources. I also declare that I have adhered to all principles of academic honesty and integrity and have not misrepresented or fabricated or falsified any idea/data/fact/source in my submission. I understand that any violation of the above will be cause for disciplinary action by the Institute and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been taken when needed.

**Date: 13 May 2019**

**Aditi Agrahari**

**17100002**

**Charchit Dhawan**

**17100013**

**Harsh Singh**

**17101021**

**Approval Sheet**

This project report entitled “Activity Recognition System” by Aditi Agrahari, Charchit Dhawan and Harsh Singh is approved for 4th Semester Minor Project.

(Signature of Examiner - I)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of Examiner -I

(Signature of Examiner - II)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of Examiner -II

(Signature of Chair)

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

Name of Chair

Date: \_\_\_\_\_\_\_\_\_\_\_\_ Place: \_\_\_\_\_\_\_\_\_\_\_\_

**Table of Contents**

**Title Page No.**

**ABSTRACT…………………………………………………………. i**

**LIST OF TABLES………………………………………………….. ii**

**LIST OF FIGURES……………………………………………….... iii**

**CHAPTER 1. INTRODUCTION 1**

**CHAPTER 2. LITERATURE REVIEW 3**

* 1. Section 2.1………………………………………………………………. 3
     1. Sub Section 2.1.1………………………………………………… 6
     2. Sub Section 2.1.2………………………………………………… 7
  2. Section 2.2………………………………………………………………. 8

**CHAPTER 3. PROPOSED SOLUTION 10**

**CHAPTER 4. RESULTS 11**

* 1. Section 4.1………………………………………………………………. 12
  2. Section 4.2………………………………………………………………. 13

**CHAPTER 5 CONCLUSIONS 14**

* 1. Section 5.1………………………………………………………………. 14
  2. Section 5.2………………………………………………………………. 17

**REFERENCES 18**

**List of All Tables**

|  |  |  |
| --- | --- | --- |
| **Table No.** | **Table Title** | **Page Number** |
| 1.1 | Table Title…………………………………………………………. | 1 |
| 2.1 | Table Title…………………………………………………………. | 6 |

**List of All Figures**

|  |  |  |
| --- | --- | --- |
| **Figure No.** | **Figure Title** | **Page Number** |
| 1.1 | Figure Title………………………………………………………… | 1 |
| 2.1 | Figure Title………………………………………………………… | 6 |

# ABSTRACT

With the advent of technology, surveillance technology is becoming prominent in all fields. Surveillance cameras can be found everywhere these days, in shops, parking lots, offices etc. The footage collected through surveillance technology holds vital importance when it comes to data extraction based on purpose-specific applications. There has been an increasing demand and scope for manipulating the footage to serve for different applications. Surveillance footage can also be utilised to recognise the happening or possibility of nefarious activities. We are implementing a model which would recognise suspicious activities in trimmed video sequences, if there are any. We have employed the concept of image scene graph which is a graphical representation of a scene or an image. A scene graph constitutes nodes which represent objects and relationships between objects[21]. The model takes an image as input, produces a set of object proposals using a Region Proposal Network (RPN)[10], and then passes the extracted features of the object regions to a graph inference module, finally producing a scene graph[24]. RPN implementation uses Faster R-CNN, which first generates region proposals using Convolutional Neural Network (CNN) and then a CNN-based network is used to classify object classes and detect bounding boxes. The model then produces a scene graph using standard RNNsand learns to iteratively improve its predictions via message passing[3]. A video sequence would be broken down into a sequence of frames, for each frame, a scene graph would be generated and checked for relationships between objects that might individually/successively resemble any of the pre-defined relationships for suspicious activities.

**Keywords**: Image Scene Graph, Neural Network, Activity Recognition

**CHAPTER 1**

**INTRODUCTION**

Parking facilities are highly likely settings for crime – both violent and property, therefore, security is one of the most critical issues facing the owners and operators of parking facilities today. For enhancing the safety and security in parking lots we need a robust system to detectsuspicious activities. In the era of emerging advance technologies high quality cameras are being used in most of places where safety and security is prime concern. Footages we get from these high-quality cameras need to detect those activities which can cause threat to our security concern. Every technology available for solving this kind of critical problem deals with processing of image. There are various objects ,attributes of those selected objects and relationship among those objects through predicate in the image .Every part is essential to understand the image and use the information described in the image Current work in computer vision and Image processing has shown thata completely graph-basedsemanticrepresentation called a scene graph is very promising representation for entire image description . Relationships between objects also comprise effective information about the scene. Scene graphs, a visually-grounded graphical structure of an image[24]. It is a novel end-to-end model that generates such structured scene representation from an input image. A holy grail of computer vision is the complete understanding of visual scenes: a model that is able to name and detect objects, describe their attributes, and recognize their relationships[9]. Understanding scenes would create important applications such as image search, question answering, and human-robotic interactions. Good progress has been made in recent years towards this goal, including imageclassification. There is a pushing effort to put together the next generation advanced datasets to serve as training and making blueprint of these datasets for these deeper, social-scene understanding and question tasks, the most notable being visual genome.The Visual Genome dataset considers relationships and attributes as prime citizens of the labeled space, in addition to the conventional focus on objects. Identification of relationships and attributes is an important part for completely understanding the visual scene, and in many cases, these elements are key to the story of a scene (e.g., the difference between “a dog chasing a man” versus “a man chasing a dog”). The Visual Genome dataset is among the first to provide a detailed annotation of object relationships and attributes, grounding visual concepts to language.[18]. Now if we are provided such a big facility that comprises all possible objects in common life,their attributes and relationship among them then using these sets of information in our own dataset would be healthy choice. There are various steps required while creating a scene-graph. It starts from data preparation and goes up to prediction of relationship among objects having some unique attributes.

# CHAPTER 2

# LITERATURE REVIEW

Understanding a visual scene is not limited to recognizing individual objects in isolation. Relationships between objects also hold rich semantic information about the scene. Most of the present-day perceptual models tackle only objects in isolation. In order to capture the entire detailed description of an image, it is necessary to analyse the semantic relationships between objects. An important step towards a thorough understanding of visual scenes is building a structured representation that record objects and their semantic relationships. Such representation aids in both fundamental recognition tasks [8] and visual tasks [18]. The success of deep learning models has shifted focus on building such representations. One such representation proposed by Johnson et al. [1], is the image scene graph, which offers a means to clearly model objects and their relationships. A scene graph is a graphical representation of the object instances in an image, where the edges indicate their pairwise relationships. The worth of scene graph has been exhibited in a wide range of visual tasks such as automatic image caption evaluation [9], semantic image retrieval [6] and visual question answering [26]. Most methods [23] have paid attention to localizing a set of objects, that is, generating independent pairwise relationships between objects. Focusing on local predictions ignores the surrounding context which can be crucial in establishing unambiguous and detailed semantic information. The model used in this project implements RNN with iterative message passing [13]. It takes an image as input and outputs a scene graph that consists of object classes, their bounding boxes, and semantic relationships between pairs of objects. The model does not infer each component of a scene graph in isolation, instead it passes messages containing contextual information between a pair of bipartite sub-graphs of the scene graph, and iteratively refines its predictions using RNNs.

The model is evaluated on a new scene graph dataset Visual Genome [26], which contains human-annotated scene graphs on 108,077 images. Each image in this dataset is annotated with an average of 25 objects and 22 pairwise relationships.

There are two parsers: a rule-based parser and a classiﬁer-based parser[23]. Both of parsers workonalinguisticrepresentationwhichwe understand as a semantic graph. We obtain meaningful graphs by parsing the image descriptions to dependency trees followed by several tree transformations. These parsers translate the semantic graph to a scene graph.

Rule-based parser outputs objects, relations and attributes directly from the semantic graph[23]. We deﬁne in total nine dependency patterns using expressions. These patterns capture the following constructions and phenomena: • Adjectival modiﬁers • Subject-predicate-object formation and subject-predicate formation without an object • Copular constructions • Prepositional phrases • Possessive formation • Passive constructions • Clausal modiﬁers of nouns except of possessives for which we manually add a relation, all objects, relations and attributes are words from the meaningful graph[16].

Classiﬁer based parser includes two parts. First, we take out all candidate objects and attributes, and second we predict relations between objects and the attributes of all objects[23]. We use the semantic graph to extract object and attribute candidates. In a first step we extract all nouns, all adjectives and all in transitive verbs from the semantic graph. As this does not make sure that the extracted objects and attributes belong to well known object classes or attribute types and image retrieval model can only make use of known classes and types, we predict for each noun the most likely object class and for each adjective and in transitive verb the most likely attribute type. To tell about classes and types, we use an L2-regularized entropy classiﬁer. For each pair (x1,x2) where x1 is an object and x2 is an another object or an attribute we predict such relation which can be any relation observed in the training data, or one of the two prominent relations IS and NONE which denotes that x2 is an attribute of x1 or no relation exists . We noticed that pairs for which a relation exists, x1 and x2 are in the same constituent, i.e. their lowest ancestor is either one of the two objects or a word in between them[19]. We therefore consider only pairs which satisfy this limitation to improve precision and to limit the number of predictions.

# CHAPTER 3

# PROPOSED METHOD

We propose a model trained on image scene graph to recognise activities being performed in trimmed video sequences. An image scene graph is a visual scene representation of an image using nodes which depict objects and the relationships between these objects. Image scene graphs have increasingly been used in image captioning, image classification, 3D scene synthesis etc[17]. So far, most methods that visualise an image into a scene graph deploy reliance on ground-truth annotations or extracting information from text domain. In this project, we have implemented the method of Recurrent Neural Network (RNN) along with iterative message passing for image scene graph generation.

The first step to this approach is object detection in an arbitrary image which has been achieved through faster R-CNN. We use a model pretrained on the MS-COCO dataset, initially an image is passed through convolutional layers to extract feature maps which are then used to generate region proposals or region of interests, their coordinates are passed to the NMS and bbox libraries to produce non-overlapping bounding boxes on objects[16]. These are then fed as an input to the same R-CNN where object detection takes place.

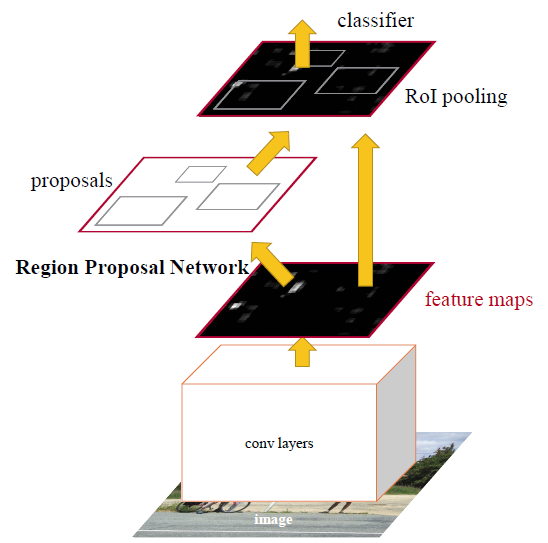


Figure 1. Faster R-CNN Architecture

Next, these object bounding boxes are fed into a graph inference model based on recurrent neural networks coupled with iterative message passing[22]. At each iteration, each hidden layer takes input from the previous layer, along with the output of the previous state and contextual information about the image. Therefore, at each iteration, predictions are revised and results with better probabilities are generate and relationship predicates are specified. Contextual information allows the model to view the image as a whole and provide a detailed description of the image, instead of a localised set of objects[17]. This message passing technique makes this approach superior to other methods employed in image scene graph generation. After a fixed number of iterations, a final scene graph is generated. At this stage, a detection network can be used to classify both objects and relationship predicates, and this step can be omitted at the previous stage that is the stage after generation of region proposals.

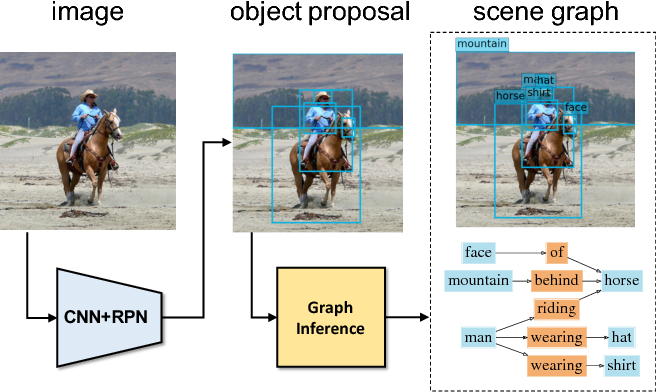


Figure 2. Image Scene Graph Generation

Finally, this model is employed for a trimmed video which is broken down into a sequence of frames. For each frame, scene graph is generated and analysed either individually (for instance, a man hitting car can be recognised in a single frame) or based on temporal features of successive frames (for instance a man forcing entry into a car would take multiple frames to identify)[10].

# CHAPTER 4

# PROJECT PLANNING

4.1. Dataset Preparation

All the datasets used in the framework belong to the hdf5 format. Hierarchical Data Format (HDF) is a set of file formats designed to store and organize large amounts of data. The data stored in hdf5 can be manipulated using NumPy.

Given a database of random images, we generate a generalised dataset named imdb in hdf5 format. In this dataset, each image is linked with its height and width.Next we generate ROI proposal database in hdf5 format using faster R-CNN and an ROI distribution for normalizing the bounding boxes.

4.2. Training the model

For training our model we used Visual Genome dataset. The dataset will include Scene graph database and Scene graph database metadata. All files in hdf5 format created during dataset preparation will be trained with Scene Graph database and Scene Graph metadata. Once the complete training is done, we will find checkpoint file and its metadata file. Checkpoint is a model format of TensorFlow which is a format dependent on the code that created the model. Here inference iteration 2 is deployed during this stage.

4.3.

We have generated a dataset of short video sequences that fall into the following categories of suspicious activities:

* A person forcing entry into a car
* A person vandalising a car
* Attempt to damage the car
* Group of people passing neary parking
* Moving around a car
* Sneaking in the car
* Person entering and exiting the car
* Touching the car
* Putting or withdrawing item in the car

# CHAPTER 5

# IMPLEMENTATION

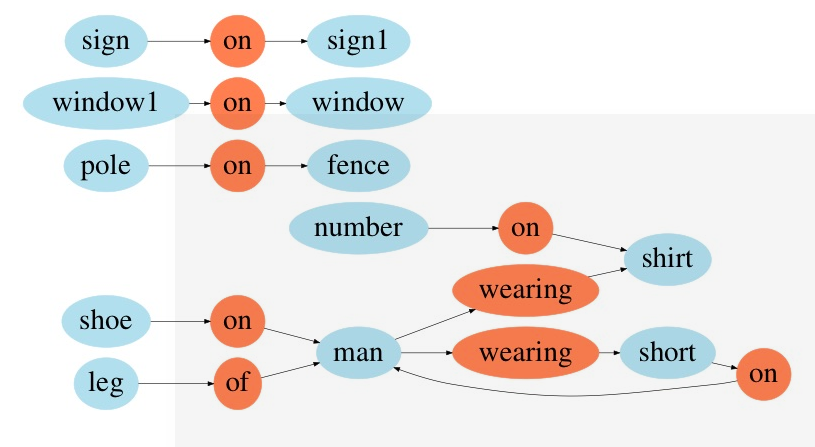
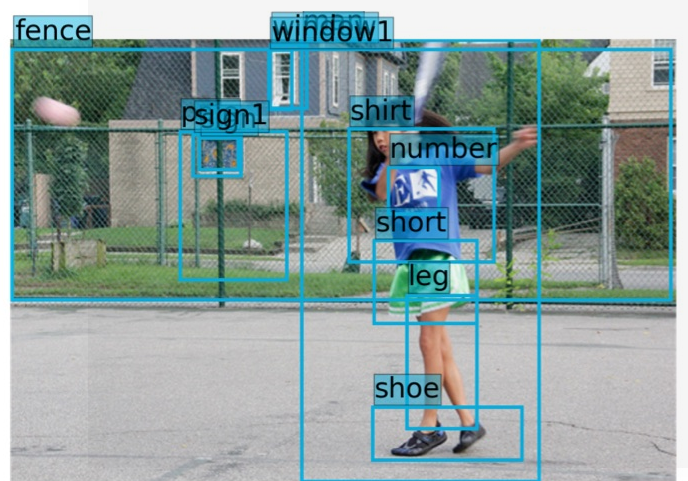
Establishing the scene graph generation model began with the environment setup. We installed all the required Python dependencies and the required tools and softwares. Next, the image database is passed to the faster R-CNN model as input. Each image goes through certain convolution layers and image’s feature maps are extracted. A sliding window is used for each location over the feature map and for each location, a specified number of boxes are used to generate region proposals. A neural network layer specifies whether there is an object for k boxes and another layer specifies coordinates, height and width for these boxes. RPN network passes locations and coordinates to the detection network which then detects objects and returns bounding box for the object[25]. The detection network uses the same CNN as the RPN network. Non-maximum suppression (NMS library) is used to reduce the number of bounding boxes for an object by eliminating overlapping bounding boxes. Further, ROI pooling layers are used to achieve fixed-size feature maps which are then exploited by Softmax and bounding box regressor to classify the bounding boxes into objects.

After obtaining bounding boxes with objects classified, generic Recurrent Neural Network (RNN) is used as graph inference module. In this module, at each iteration, each recurrent unit takes its previous hidden state and an incoming message as input, and produces a new hidden state as output. The incoming message constitutes contextual information about the image. This incoming message is utilised to improve the predictions/results for the next hidden layer. The final output layers of this neural network closely follow the faster R-CNN setup. A fully-connected layer is used to regress to the bounding box offsets for each object class separately[17]. After reaching a certain number of iterations or when there are no further improvements observed in the output, the processing is terminated and the final image scene graph is stored in a database. The final scene graph would be stored in the format of triplets where they will be inspected against a list of pre-defined relationships representing a suspicious activity.

# CHAPTER 6

# RESULTS

Scene graph generation while running the model.



# CHAPTER 7

# SUMMARY AND FUTURE SCOPE

We Creatied a suitable environment for generating Image Scene-graph,Trained the model with specified python libraries such as IMDB,ROIdb, nms,bbox etc..using VG-Dataset.We Implemented training to our model on VG-dataset and created our own dataset of arbitrary images in desired format(.hdf5). For recognition of suspicious training we trimmed video clips in different-different activities to create dataset .

Studying and establishing rules, regulations and constraints to classify activities as dubious and observing a common pattern for the same and training the model based on this dataset. Applying the built model to live surveillance footage and implement smart parking

# CHAPTER 8

# REFERNCES

[1] P. Anderson, B. Fernando, M. Johnson, and S. Gould. Spice:

Semantic propositional image caption evaluation. In ECCV,

2016.

[2] R. Baur, A. Efros, and M. Hebert. Statistics of 3d object

locations in images. 2008.

[3] S. Bell, C. L. Zitnick, K. Bala, and R. Girshick. Inside-

outside net: Detecting objects in context with skip

pooling and recurrent neural networks. arXiv preprint

arXiv:1512.04143, 2015.

[4] A. X. Chang, M. Savva, and C. D. Manning. Learning spatial

knowledge for text to 3d scene generation. 2014.

[5] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico:

A benchmark for recognizing human-object interactions in

images. In ICCV, 2015.

[6] Y.-W. Chao, Z. Wang, Y. He, J. Wang, and J. Deng. Hico: A

benchmark for recognizing human-object interactions in im-

ages. In Proceedings of the IEEE International Conference

on Computer Vision, 2015.

[7] K. Cho, B. Van Merriënboer, D. Bahdanau, and Y. Bengio.

On the properties of neural machine translation: Encoder-

decoder approaches. arXiv preprint arXiv:1409.1259, 2014.

[8] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative mod-

els for static human-object interactions. In 2010 IEEE Com-

puter Society Conference on Computer Vision and Pattern

Recognition-Workshops. IEEE, 2010.

[9] C. Desai, D. Ramanan, and C. C. Fowlkes. Discriminative

models for multi-class object layout. International journal

of computer vision, 95(1), 2011.

[10] M. Fisher, M. Savva, and P. Hanrahan. Characterizing struc-

tural relationships in scenes using graph kernels. In ACM

SIGGRAPH 2011 papers, 2011.

[11] C. Galleguillos, A. Rabinovich, and S. Belongie. Object cat-

egorization using co-occurrence, location and appearance.

In Computer Vision and Pattern Recognition, 2008. CVPR

2008. IEEE Conference on. IEEE, 2008.

[12] R. Girshick. Fast r-cnn. In Proceedings of the IEEE Interna-

tional Conference on Computer Vision, 2015.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Region-

based convolutional networks for accurate object detection

and segmentation. IEEE transactions on pattern analysis

and machine intelligence, 38(1), 2016.

[14] A. Gupta and L. S. Davis. Beyond nouns: Exploiting prepo-

sitions and comparative adjectives for learning visual classi-

fiers. In European conference on computer vision. Springer,

2008.

[15] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning

for image recognition. CVPR, 2016.

[16] A. Jain, A. R. Zamir, S. Savarese, and A. Saxena. Structural-

rnn: Deep learning on spatio-temporal graphs. arXiv preprint

arXiv:1511.05298, 2015.

[17] Z. Jia, A. Gallagher, A. Saxena, and T. Chen. 3d-based rea-

soning with blocks, support, and stability. In Proceedings

of the IEEE Conference on Computer Vision and Pattern

Recognition, 2013.

[18] J. Johnson, R. Krishna, M. Stark, L. J. Li, D. A. Shamma,

M. S. Bernstein, and L. Fei-Fei. Image retrieval using scene

graphs. In IEEE Conference on Computer Vision and Pattern

Recognition (CVPR), 2015.

[19] P. Krähenbühl and V. Koltun. Efficient inference in fully

connected crfs with gaussian edge potentials. In Advances in

Neural Information Processing Systems 24, 2011.

[20] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz,

S. Chen, Y. Kalantidis, L.-J. Li, D. A. Shamma, M. Bern-

stein, and L. Fei-Fei. Visual genome: Connecting language

and vision using crowdsourced dense image annotations. In

arXiv, 2016.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet

classification with deep convolutional neural networks. In

NIPS, 2012.

[22] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Graph cut

based inference with co-occurrence statistics. In European

Conference on Computer Vision. Springer, 2010.

[23] X. Liang, X. Shen, J. Feng, L. Lin, and S. Yan. Semantic

object parsing with graph lstm. In European Conference on

Computer Vision, 2016.

[24] W. Liao, M. Y. Yang, H. Ackermann, and B. Rosenhahn. On

support relations and semantic scene graphs. arXiv preprint

arXiv:1609.05834, 2016.

[25] D. Lin, S. Fidler, and R. Urtasun. Holistic scene understand-

ing for 3d object detection with rgbd cameras. In Proceed-

ings of the IEEE International Conference on Computer Vi-

sion, pages 1417–1424, 2013.