Declaration

I hereby declare that the work presented in this thesis titled *Fine Grained Feature Representation using Computer Vision Techniques for Understanding Indoor Space* submitted to the Indian Institute of Technology Jodhpur in partial fulfillment of the requirements for the award of the degree of Doctor of Philosophy, is a bonafide record of the research work carried out under the supervision of Dr. Chiranjoy Chattopadhyay and Dr. Gaurav Bhatnagar. The contents of this thesis in full or in parts, have not been submitted to, and will not be submitted by me to, any other Institute or University in India or abroad for the award of any degree or diploma.

Shreya.

Shreya Goyal
P16CS003

Certificate

This is to certify that the thesis titled *Fine Grained Feature Representation using Computer Vision Techniques for Understanding Indoor Space*, submitted by *Shreya Goyal (P16CS003)* to the Indian Institute of Technology Jodhpur for the award of the degree of *Doctor of Philosophy*, is a bonafide record of the research work done by her under our supervision. To the best of our knowledge, the contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Chiranjoy Chattopadhyay
Ph.D. Thesis Supervisor

Gaurav Bhatnagar
Ph.D. Thesis Supervisor

Acknowledgments

I wholeheartedly thank my Ph.D. Thesis Supervisors, *Dr. Chiranjoy Chattopadhyay* and *Dr. Gaurav Bhatnagar*, for being mentor, guide and friend in the truest form during the entire course of my PhD programme. Qualities like patience, optimism and seeking perfection in one's work are just some of the wonderful things that I have learnt from them and would carry with me throughout my life. This dissertation would not have been possible without their constant support and guidance.

I thank Head of the Computer Science and Engineering Department, *Professor Richa Singh* for her guidance. I would also like to thank the members of the Doctoral Committee, *Dr. Anil Kumar Tiwari*, *Dr. Mayurakshi Chaudhury*, *Dr. Suman Kundu*, for their enthusiastic and continued guidance during the research work. I am also thankful to all the CSE department, IIT Jodhpur staff for their continuous and valuable support.

A big thanks to *Dr. Naimul Khan* from multimedia laboratory, Ryerson University, Toronto for providing me a welcoming and conducive working environment and all the brainstorming ideas and discussions during my visit to his laboratory.

My stay at the Institute was a wonderful experience because of my friends, *Abhra Paul*, *Devika Laishram*, *Rohit Shandley*, *Vishwa Deepak*, *Gaurav Bahuguna*, *Ajay Urgunde.* I thank them all for all the liveliness they infused into the non-academic part of the days at IIT Jodhpur.

I would like to thank my PhD colleagues and lab-mates *Divya Shrivastava*, *Arka Ujjal Dey*, *Dipti Trivedi* and *Jaspreet Kaur* for being the major source of support. A special mention to my friends *Shubhra Sharma* and *Anirudhha Singhal* for always being there for me with the words of encouragement.

Lastly, I deeply acknowledge and thank my parents and my brother *Aditya*, for their unconditional trust, endless patience, and giving me all the opportunities that have helped me shape into the individual that I am today.

List of Figures

Figure *Title* *Page*

1.1 Illustration of various subareas of document image analysis (DIA) on a sample scanned document [Rusiñol *et al.*, 2010]. 2
1.2 Motivation of the problem statement. 5
1.3 Brief overview of the work done shown over a tree visualization. 7

2.1 Process of text graphics separation in method proposed in L.A. Fletcher and R. Kasturi [1988]. 12
2.2 Process of patch based wall segmentation proposed in de las Heras *et al.* [2014]. 14
2.3 Symbol spotting framework proposed in Dutta *et al.* [2013] using sub-graph matching. 16
2.4 Object detection by (a) Faster-RCNN proposed by Ren *et al.* [2015] and (b) YOLO proposed by Redmon *et al.* [2016]. 17
2.5 Image description generation by (a) Densecap proposed by Johnson *et al.* [2016] and (b) Hierarchical recurrent network based approach as proposed by Krause *et al.* [2017]. 18
2.6 3D room layout synthesis proposed by LayoutNet in the document Zou *et al.* [2018]. 20

3.1 Various tasks performed on floor plan images. 23
3.2 Sample images of floor plans in some publicly available datasets. 25
3.3 Sample images of floor plans in annotated ROBIN dataset. 26
3.4 Example of a sample floor plan image from ROBIN dataset and the annotation collected for the same to synthesize A-ROBIN. 26
3.5 Graphical User Interface of LabelImg tool for annotation proposed by Tzutalin [2015]. 27
3.6 Various classes of decor items available in the dataset. 28
3.7 Illustrations of the several steps of annotation process. 29
3.8 An example of paragraph annotation for a floor plan image. 30
3.9 Frequency distribution of the most frequent words occurring in the paragraph corpus. 31
3.10 Distribution of various objects in the training dataset. 32
3.11 Quantitative analysis of the symbol spotting task. 33
3.12 Qualitative results of symbol spotting using the YOLO model. 33
3.13 Qualitative results of symbol spotting using the faster-RCNN model. 34
3.14 Qualitative results of dense captioning with the proposed BRIDGE dataset. 35
3.15 Paragraphs generated using Densecap-concat, Template based method and evaluation. 36

4.1 An illustration of the problem of egocentric vision based narration synthesis from a given input floor plan image. 40
4.2 Block diagram of our proposed framework. 40
4.3 Illustration of the text segmentation and recognition. 42
4.4 Intermediate results generated while textured wall segmentation. In part (k), 'red' denotes Brick, 'green' denotes concrete, and 'blue' denotes wood. The corresponding textures are shown in part (l), which are adopted from Kilmer and Kilmer [2009]. 43
4.5 Flow chart for entry door detection in a floor plan image. 45
4.6 A schematic representation of room decor characterization and final schema generation. 46
4.7 A flow chart for egocentric narration synthesis and set of questions to extract salient features of a room. 48
4.8 An illustration of the results of egocentric narration synthesis based on the proposed model. 49

5.1 An illustration of the problem of narration synthesis from a given input floor plan image with navigation path. 52
5.2 Block diagram depicting various modules and work-flow within SUGAMAN. 52
5.3 An illustration of the room segmentation and room partitioning process. 54
5.4 Pictorial depiction of the twelve classes of decor models used in the experiments. 54
5.5 Qualitative comparison between ours and D. Sharma, C. Chattopadhyay and G. Harit [2016] for the task of Decor identification. 56
5.6 An example of a Bag of Decor (BoD) feature descriptor computed for a sample room image. 57
5.7 AN schematic illustration of computing the LOFD descriptor for four sample room images. 57
5.8 Schematic representation of automatic room label classification. 58
5.9 Structure of the XML file generated for description synthesis. 59
5.10 An example of the boundary tracing process for different values of the shrinkage factor t. 59
5.11 An illustration of the global and local coordinate systems with respect to a floor plan image. 60
5.12 A schematic representation of the non-uniform and uniform binning. 61
5.13 A visual illustration for explaining the rationale of non-uniform binning. 62
5.14 Illustration of path detection process. 65
5.15 Examples illustrating path detection by avoiding obstacles. 65
5.16 Training, testing and validation ROC curves for BoD and LOFD features. 68
5.17 Performance analysis of multi-layered perceptron by BoD and LOFD. 69
5.18 Generated descriptions for three floor plan images from A-ROBIN dataset. 72
5.19 Generated descriptions for three floor plan images from A-ROBIN dataset. 73
5.20 An illustration of incorrect creation of the LOFD feature. 73

6.1 An illustration of the proposed problem domain with the desired output. 78
6.2 An illustration of the proposed work of generating textual description from floor plan images. 79
6.3 Hierarchical RNN to yield paragraph from floor plans. 80
6.4 Framework of the proposed TBDG model of generating paragraph description from input floor plan image. 81
6.5 An illustration of a floor plan image and its corresponding annotations in the BRIDGE dataset Goyal *et al.* [2019b]. 84
6.6 The object classes available in BRIDGE dataset proposed in Goyal *et al.* [2019c]. 86
6.7 Qualitative result of the spotted decor symbols in two floor plan images. 87
6.8 A visualization of the image classification network with top-6 activation maps from each layers. 88
6.9 Qualitative result of room classification for 5 classes on two different floor plan images. 89
6.10 An illustration of semi-structured description generation from floor plan images. 90
6.11 Descriptions generated with proposed models and various baseline models. 90
6.12 A comparison of YOLO and Faster-RCNN models for decor identification. 91
6.13 Performance evaluation of room classification. 92
6.14 Performance evaluation for TBDG & DSIC models. 93
6.15 A comparison of the loss curves of language models. 93
6.17 An illustration of a failure case with the TBDG model. 93
6.16 Descriptions generated with proposed models and various baseline models. 94
6.18 An illustration depicting the robustness of the TBDG model over DSIC for an unknown sample. 94

7.1 A schematic illustration of the entire input and output scenario while using GRIHA. 99
7.2 Panoramic image and multiple 2D images for the same scene. 100
7.3 Highlighted occluded portions of the scene and corresponding 2D images. 100
7.4 A schematic representation of the image acquisition framework developed for GRIHA. 100
7.5 Pipeline of the proposed framework. 101
7.6 An illustration of how depth is estimated by adapting the encoder-decoder model proposed in Alhashim and Wonka [2018]. 103
7.7 Results of depth generation on our dataset. 103

7.8 An illustration of how LayoutNet is adapted for the layout estimation in GRIHA. 104
7.9 Results of layout generation on our dataset. 104
7.10 An illustration of partial scene layout generation from 3-D point cloud. 105
7.11 Depiction of translation and global regularization. 106
7.12 An illustration of the intermediate stages of the regularization process. 107
7.13 Step by step generation of the floor plan of an indoor scene. 108
7.14 A schematic representation of the network architecture of YOLO for door detection. 109
7.15 An illustration of the performance of the door detection and placement algorithm in floor plan. 110
7.16 Sample images from the dataset. 111
7.17 Estimated layouts for Labs and office datasets. 112
7.18 Graphical User Interfaces (GUI) for different layout estimation applications. 115
7.19 Scenes captured in different environments. 116
7.21 Comparison of aspect ratio error across devices. 116
7.20 Comparative analysis of area error across devices. 117
7.22 Comparison of power consumption across devices. 117

List of Tables

Table *Title* *Page*

2.1 Details of publicly available existing floor plan datasets 15

3.1 Statistical details of the paragraph annotations 31
3.2 Evaluation of generated paragraphs with various metrices. 36

4.1 Room Quantity Info schema for the Example in Fig. 4.4. Room Names: Bed Room (BD), Bath Room (BT), Entry (EN), Kitchen (KN). 41
4.2 Wall info schema for the Example in Fig. 4.4. Material details: Concrete, Wood, and Brick. 45
4.3 Room Quality Info schema for the Example in Fig. 4.4. Decor details: Coffee table (CT), Bed (B), Small Sofa (s), Large Sofa (S),Dining Table (D), Sink (K), Tub (T). 47
4.4 Quantitative comparison of wall material characterization. 50
4.5 Quantititative results of quality of text synthesis. 50

5.1 Sentence model based on proximity 66
5.2 Classifiers results of room annotation learning by BoD feature. 66
5.3 Classifier-Results of room annotation learning by LOFD feature. 67
5.4 Performance analysis of text generation algorithm using ROUGE score. 68
5.5 Performance analysis of text generation algorithm using BLEU score. 69
5.6 Performance analysis of description synthesis using METEOR score. 70
5.7 Comparison of recognition accuracy (%) of ours vs. others. 74
5.8 Comparative analysis of various path finding approaches 74

6.1 Evaluation of generated paragraphs with different metrices (METEOR, BLEU, ROUGE). 92

7.1 Performance analysis of depth estimation on our dataset 112
7.2 Corner and edge estimation analysis on our dataset 112
7.3 Quantitative evaluation of the estimated layouts for different scenes(S) and methods(M) 113
7.4 Qualitative comparison of GRIHA and state-of-the-art. 114

List of Symbols

I	Input floor plan image
$CC()$	Connected Component Operator
ζ	Individual connected component
$\lvert.\rvert$	Cardinality operator
P_v	Pivot point
$\mathbb{W}$	Formed word
$\mathbb{V}$	Predefined word
$ED()$	Euclidian distance operator
e	Edit distance
W	Wall image
$\mathbb{E}$	Edge image of walls
W_H	Horizontal line image
W_V	Vertical line image
$\mathbb{C}$	Regions of intersection
$\mathbb{H}_B$	Normalized histogram for Brick wall material
$\mathbb{H}_C$	Normalized histogram for Concrete wall material
$\mathbb{H}_W$	Normalized histogram for Wood wall material
$\mathbb{H}_S$	Normalized histogram for given segments of wall materials
$\mathbb{S}$	Set of wall segments
$\mathbb{D}_\mathrm{k}$	Distance between a pair of histograms of material
(β, ξ)	Bins corresponding to histograms of current Wall segment
$\mathbb{F}_1$	Wall material characterized floor plan image
$(\mathrm{Cx}_s, \mathrm{Cy}_s)$	Centroid of each wall segment
D_p	Entry door pixel
C	Connected Components
$\mathscr{A}_k$	Area of each component
A	Sorted collected of areas for all components
F	UDI Signature
$\mathscr{R}_c$	Room center
$\mathscr{D}_c$	Decor Center
d_n	Normalized distances for decor items
D	Decor Items
N_r	Number of rooms
Dim_F	Dimension of the Feature Vector
t	Shrinking Factor
AM_D	Door based Adjacency Matrix
AM_N	Navigation based Adjacency Matrix
D_E	Door Entry
D_X	Door Exit
V_L	Vertex list
C_S	Strongest Corners
P^i	Navigation Route
S_i	Sentences
NN_r	neighboring Rooms

$DLOC$	Relative location of rooms
DIR	Direction
N_m	number of turns
R_i	Region Proposals
PV	Pooled Vector
M	Projection Matrix
Th	Threshold
K	Paragraph Description
$Sent_{max}$	Maximum Sentences
$Word_{max}$	Maximum Words
$Prob$	Probability term
W_e	Word cues extracted from images
C_i	Captions
S_e	Target sentence embeddings
h_t	Target hidden state
h_s	Source hidden state
e_{ij}	Alighnment scores
α_{ij}	Attention weights
cv_i	Context vector
r_i	Characterized Rooms
o_i	Detected Objects
I_k	Image for input indoor scene
C^k	Camera pose
f	focal length
(C_x, C_y)	Camera center coordinates
$\mathbf{(u,v)}$	Coordinates in in RGB image
$D_{u,v}$	Depth value in the depth map
S	Scaling factor of the scene
$\mathbf{R}$	Total number of rooms in dataset
R_j	Individual rooms
P_i	Point clouds
K	Boundary coordinates
$\mathrm{c_k}$	Clusters
m_k	Mean for each cluster
RP_i	Regularized local point cloud
TP_i	Transformed point cloud
$\theta_x, \theta_y, \theta_z$	Rotation angles
(t_x, t_y)	Translation Coordinates
FP	Final Polygon
p	Number of sides
ϕ	Angle
$line_i$	Line joining two points
s_i	Sides of polygon
(X, Y, Z)	Coordinate system
$\mathbf{C}_{hull}$	Convex hull
$\mathbf{Ratio}_D$	Ratio used for door marking
C_{BB_I}	Centroid of bounding box for door in real world image
W_I	Wall in real world image
L_{W_I}	Width of wall in real world image
L_{W_F}	Width of wall in 2D mapping
C_{BB_F}	Centroid of a 2D door symbol

W_{I_F}	Wall in corresponding 2D mapping
(μ_x, μ_y)	Intensity terms
MAX_I	Maximum pixel value in image
(σ_x, σ_y)	Standard deviation terms

List of Abbreviations

DIA	Digital Image Analysis
OCR	Optical Character Recognition
DAR	Document Analysis and Research
NLTK	Natural Language Tookit
LDA	latent Dirichlet allocation
JSON	JavaScript Object Notation
LBP	Linear Binary Pattern
NLP	Natural Language Processing
DIA	Digital Image Analysis
SLIC	Simple Linear Iterative Clustering
XML	Extensible Markup Language
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
METEOR	Metric for Evaluation of Translation with Explicit ORdering
BLEU	Bilingual Evaluation Understudy
mAP	Mean Average Precision
IoU	Intersection over Union
BoD	Bag of Decors
LOFD	Local Orientation and Frequency Descriptor
BIM	Building Information Modelling
UDI	Unique Decor Identifier
BP	Brevity Penalty
SVM	Support Vector Machine
CNN	Convolutional Neural Network
TBDG	Transformer Based Description Generation
DSIC	Description Synthesis using Image Cues
RNN	Recurrent Neural Network
LSTM	Long Short-Term Memory
GRU	Gated Recurrent Units
RPN	Region Proposal Network
YOLO	You Only Look Once
F-RCNN	Faster- Region based Convolutional Neural Network
SSD	Single Shot Detector
FOV	Field Of View
HRNN	Hierarchical Recurrent Neural Network
SLAM	Simultaneous Localization And Mapping
AR	Augmented Reality
VR	Virtual Reality
ReLU	Rectified Linear Unit
CAD	Computer Aided Design
IMU	Inertial Measurement Unit
ROBIN	Repository Of BuildIng plaNs
SESYD	Systems Evaluation SYnthetic Documents
CVC-FP	Computer Vision Center- FLoor Plan

BRIDGE	Building plan Repository for Image Description Generation, Evaluation
SIFT	Scale Invariant Feature Transform
YOLO	You Only Look Once
IoU	Intersection Over Union
RCNN	Region-based Convolutional Neural Networks
SSD	Single Shot Detector
GUI	Graphical User Interface
PSNR	Peak Signal to Noise Ratio
SS	Structural Similarity
MSE	Mean Square Error
GT	Ground Truth
CE	Corner Error
PE	Pixel Error