1
Introduction

In the current digital era, the use of paper-based documents continues to grow even with the existence of electronic documents because of preferences for reading paper-based documents as proposed by Jain and Yu [1998]. The amount of digital data, specifically in the form of images, has increased exponentially over the years. Accordingly, the complexity in their understanding and interpretation has been increased with the wide variety involved. Document images are scanned/camera images of the paper-based documents, taken to make their content computer-readable and reduce the amount of paper used. The electronic form of paper-based documents is easy to store and transmit, and accordingly, their content can be stored, retrieved, and transmit with ease after recognition and analysis. Documents images are of two broad categories, the born-digital documents and digitized documents as proposed in Doermann *et al.* [2014], Tigora *et al.* [2013]. The born-digital documents include plain text documents such as emails, PDF documents, and web documents. In contrast, digitized documents are the scanned pages of existing physical paper-based documents which are digitally encoded and given pixel structure for their wide availability, extraction, and processing of the information contained as proposed in Baird [2003] and Tigora *et al.* [2013].

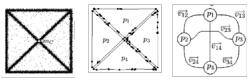
Given the popularity of paper-based documents over electronic documents, the inconvenience involved in their storage and retrieval, and information loss while their digitization, Document Image Analysis (DIA) systems have become an important research area. These DIA systems possess applications in storage, extraction, retrieval, processing, and document image interpretation. DIA systems research has essential and challenging machine learning applications, computer vision, pattern recognition, image processing, and natural language processing. DIA comes with an objective of human-like recognition of text and graphics present in the document and extracting the intended information. These documents can be hand-written or printed paper-based documents, and their conversion to electronic form opens up avenues for significant research areas.

Document images can contain components such as text and graphics. Analyzing the textual part and reading it with Optical Character Recognition (OCR) is one aspect of DIA research. At the same time, graphic recognition and interpretation is another critical research area in the same. Some document images can contain text with graphics, while others can be graphic dominating. The graphical components are binary-valued, and they can also exist along with text and pictures as proposed in Kasturi *et al.* [2002]. The research objective is to extract information depicted by the graphical components and use it for further interpretation by analyzing the visual elements. Interpreting the data extracted by graphics can help understand the document image in textual description, textual rules, or any other human-readable format.

Figure. 1.1 (a) and (b) depicts various tasks that are performed in the DIA research area. The documents shown can have graphical symbols, lines, text within graphics, mathematical equations, tables for detection, and recognition. Figure 1.1 (a) shows DIA includes analysis of a page's layout. Also, it provides text analysis, which detects and recognizes the text within the document image. Graphic recognition is another task covered in DIA, which detects and identifies the graphical symbols and line drawings within the document image. It also includes the detection and recognition of text within the graphics. Figure. 1.1 (b) shows another example of a

Graphic Recognition

- Symbol spotting

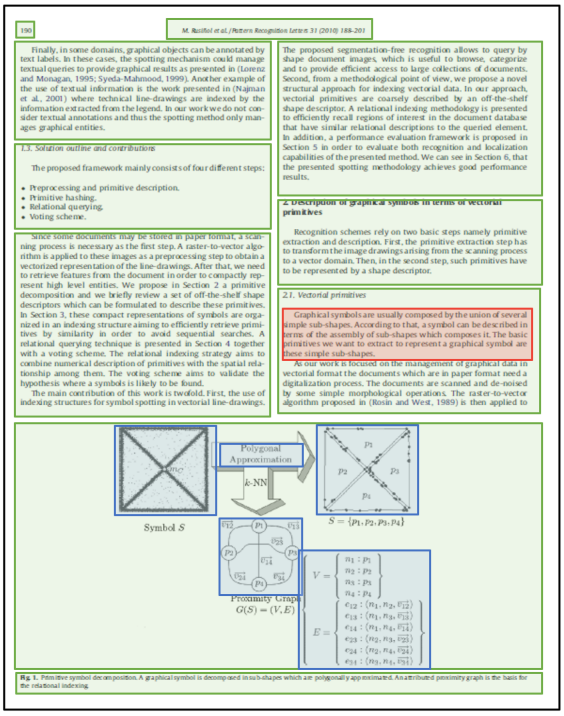


- Text in graphics

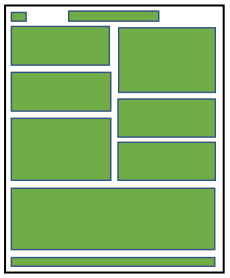
Polygonal Approximation

p_1, p_2, p_3, p_4

v_{12}, v_{13}, v_{23}, v_{14}, v_{34}, v_{24}



Page Layout Analysis

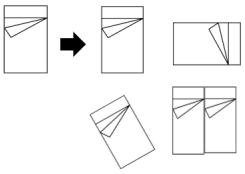


Text Analysis

Graphic Symbols are usually composed by the union of several simple sub-shapes. According to that, a symbol can be described in terms of the assembly of the sub-shapes which composes it. The basic primitives we want to extract to represent a graphical symbol are these simple sub-shapes.

(a)

Symbol spotting and retrieval



Mathematical equation detection

$$circ = \frac{4\pi A}{P^2}$$

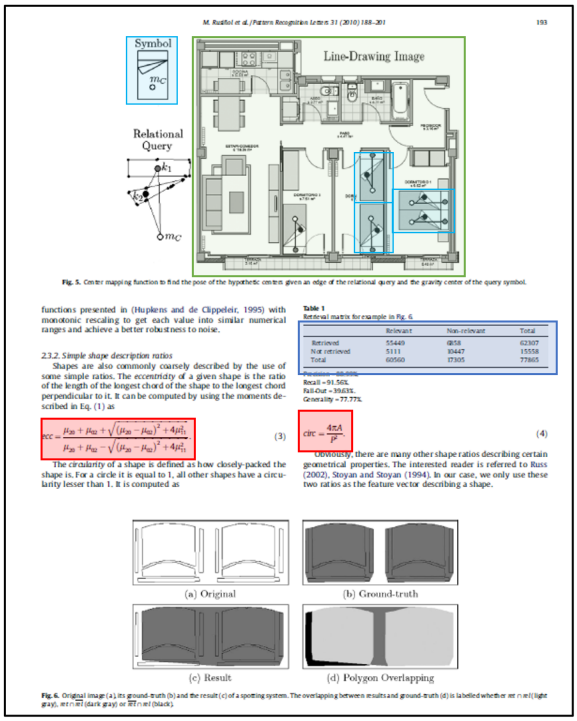


Image interpretation

This floor plan has a bedroom which has a double bed. There is a hall which has large sofa and tables. There is a kitchen which has a dining table, cooking range and sink. Bathrooms has bathtub and sinks.

Table layout detection

(b)

Figure 1.1 : Illustration of various subareas of document image analysis (DIA) on a sample scanned document [Rusiñol *et al.*, 2010].

document image, which shows tasks such as symbol spotting and retrieval. It also shows the task of interpreting a line drawing in the human-readable format, generating human-readable textual description from a line drawing. Tasks such as table detection and segmentation and mathematical equation detection are also covered under the area of DIA.

DIA primarily deals with text and graphics processing. Graphics recognition is one of the sub-areas in DIA that deals with the graphical content present in a document. Analyzing text in an image deals with recognizing the text content with OCR techniques and analyzing other text blocks and paragraphs. While graphic analysis deals with the straight, curve, and different kinds of lines, filled regions, and symbols present in the document by O'Gorman and Kasturi [1995]. Graphics are the visual images or representations of physical objects or diagrammatic representations of phenomena or information. These graphical documents include engineering drawings, facility maps, flow charts, circuit diagrams, musical notations, optical characters, mathematical expressions, architectural floor plans, and other line drawings. Graphical documents have their own set of visual language and rules (grammar) to convey semantics and the relevance for a given context. The graphic language combines complex symbols and grammatical rules that define the context-based relations between those symbols and other graphical components. Hence, understanding documents with graphical symbols and text and interpreting them becomes a challenging task. More often or not, such an analysis requires human expertise. Interpretation of document images involves graphic recognition, where graphics in a document are various symbols and line drawings. The graphic recognition (symbol spotting) process proceeds in two main steps where the first step extracts the representation (features) and the second steps match these representations with known models present in a pre-defined dictionary as proposed by [O'Gorman and Kasturi, 1995].

The understanding of a digitised document or document images requires analysis of its several components or parts, which involves:

- Detection and recognition of symbols in the document.
- Interpretation of the context and relationship between them.
- Converting them in a human readable form for common user.

To detect and recognize graphical symbols, techniques such as handcrafted feature matching Barducci and Marinai [2012]; Dutta *et al.* [2013] where symbols represented in graphic form are famous in the literature. In the work presented in [Goyal *et al.*, 2018a, 2019a], apart from detecting and recognizing symbols in architectural plans, a relationship among the nearby symbols and lines is identified. Also, authors in Ahmed *et al.* [2011] have proposed methods to detect lines based on their thickness and the relationship between the text present in the graphics. The work proposed in [Goyal *et al.*, 2018a, 2019a; Mondal and Jawahar, 2019] has interpreted the recognized graphics in the document in the human-readable form by generating a textual description out of them.

The most natural way of interpreting an image after its understanding is by describing it in natural language. Systems which describe image into the raw text are called image captioning systems. There are currently several artificially intelligent models available that interpret an image into a caption after its understanding. These captions could be of single or multi-sentence describing the salient parts of the image into natural language. Automatic image to text generation problems requires computer vision and natural language processing methods, both rooted in machine learning and artificial intelligence. This task requires an input image, and the system generated the image's interpretation into textual form. These systems have currently been developed for natural images where the generated description includes information about the visual content and their relationship among them. Hence, the right image captioning system requires an excellent understanding of the image's components and structured natural language.

Image captioning systems can be broadly divided into two categories

- Image visual feature based image to text generation
- Retrieval based image to text generation

A visual feature-based generation system extract information from images, as humans would have described them, by detecting objects, scenes, attributes, actions, and other visual content by using computer vision techniques. These systems use standard natural language generation methods, such as template-based, n-gram, or grammar rules, to generate a description from the extracted visual information. In this case, the accuracy of the generated description highly depends upon the accuracy of subsystems that detected the visual content in the image as proposed in Bernardi *et al.* [2016], Kulkarni *et al.* [2011], Elliott and Keller [2013].

The other category of methods poses image to text generation as a retrieval problem. They search for similar images in the database and generate a description for the new image by reusing the description for the most similar images or generate a new narrative based on descriptions of similar images as presented in the documents Bernardi *et al.* [2016], Ordonez *et al.* [2011], Kuznetsova *et al.* [2012]. In the same line of retrieval-based methods, a few authors have posed this problem in multi-modal space. They learn image and text features in a common multi-modal space, and use this joint representation to perform image-description retrieval given a query image as proposed in Bernardi *et al.* [2016], Karpathy and Fei Fei [2015], Donahue *et al.* [2015]. Some of the methods in this line, learn visual and textual features in an encoder-decoder framework. The encoder jointly encodes visual features with textual features, and the decoder uses this framework to generate the descriptions for new images as proposed in Kiros *et al.* [2014].

Another form of interpretation after understanding is synthesizing a similar and more understandable form of the input image. For example, indoor space can be understood and interpreted in natural language or its simplified layout re-created as a drawing. Hence, the re-creation of an intricate image in another simplified form is another dimension for understanding, interpreting, and expressing. Therefore building a system that could understand an image and generate its interpretation in another modality is necessary. Such a task requires knowledge from various disciplines such as computer vision, image processing, pattern recognition, machine learning, natural language processing, stereo geometry, etc.

Existing image to text generation systems can understand details in natural images and describe them in human-like sentences. Describing the document images (graphics) in human-like sentences can help build assisting systems for students, navigation in building maps and historical monuments, online recommendation systems, self-learning tutorials, and many more. It will generate a standard user understandable format of that document by generating its description in natural language. For example, the work proposed in Goyal *et al.* [2019a] generated textual description from architectural floor plans of indoor house scene and built a door-to-door obstacle avoidance navigation system for user. The proposed system would help the visually impaired to walk through the indoor environment by using text-reading software. Also, authors in Mondal and Jawahar [2019] have proposed a method to generate textual description from mathematical equations, which will help students in self-understanding and assessment. The natural language reasoning from document images has also been explored for other graphical document domains such as scientific plots and charts. The authors in Methani *et al.* [2020]; Reddy *et al.* [2019]; Kafle *et al.* [2018]; Kahou *et al.* [2017] have proposed datasets and algorithms to reason over scientific plots, bar graphs, charts and other data visualization graphical documents.

Architectural floor plans are one of the documents rich in graphics and contain a lot of information. It has not been understood concerning its content and described in natural language.

It represents an indoor space of a house, office, hospital, or any other building with rich graphics. Hence, understanding an indoor space with its floor plan by understanding and its interpretation can be challenging given the graphics' complexity. How to develop a system that could understand the inherent information available in the indoor spaces by their floor plans and interpret them in human like sentences is a point that this thesis will reinforce. Hence, in this thesis, a study of indoor space is proposed, which takes architectural 2D floor plans as representation images and generates a textual interpretation from them after understanding. A generation of an image to image interpretation that is 2D floor plan generation for the given indoor space is also proposed in this thesis. The rest of the chapter is divided as Sec. 1.1 describes the motivation behind the proposed work in this thesis, Sec. 1.2 discusses the problem statement and brief details about the solutions proposed, Sec. 1.3 discusses the research issues in the proposed problem, and Sec. 1.4 presents the structure and organization of this thesis.

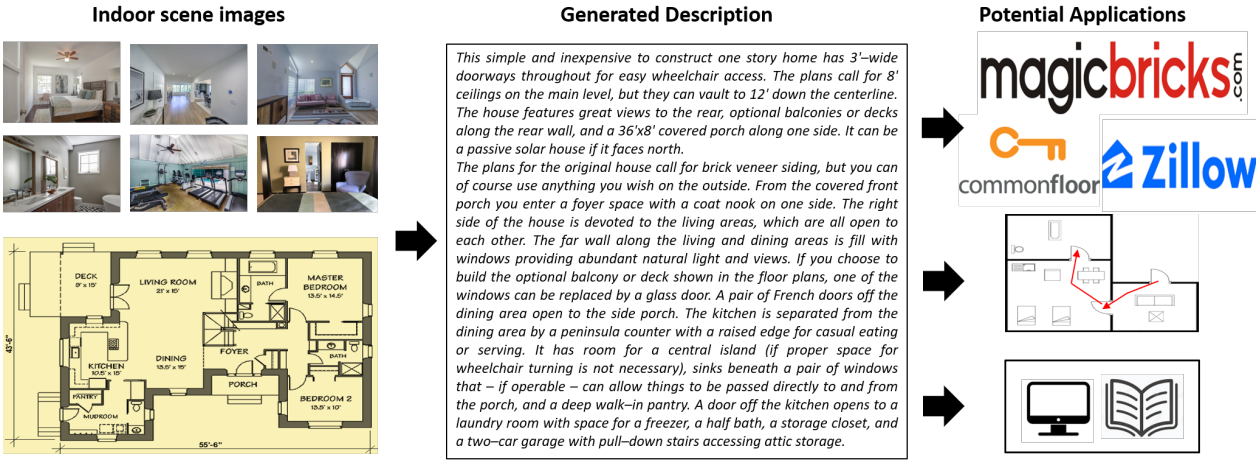


Figure 1.2 : Motivation of the problem statement.

1.1 MOTIVATION OF THE PROBLEM

The motivation behind understanding an indoor space and its interpretation originated from the current requirement of the real-estate industry and online-rental platforms. When everybody needs a quick solution to their needs in today's time, online rental websites for real estate are excellent platforms for buying, selling, or renting properties. Given the bulk of potential sellers and buyers, there is an exponential increase of people posting their ads on these platforms. Also, the number of such platforms has increased in recent years by many folds, given its potential. Hence, there is a requirement of a system that could give a brief interpretation of the property to be posted as an advertisement by understanding it. Floor plans represent a house, or an indoor space, which depicts its number of rooms, dimensions, functionality, and other components. Interpreting an indoor scene is a challenging task directly by looking at its natural images since there is a lot of subjectivity in the raw images. There is no direct method to identify the number of rooms, their global positioning, and other components unless the user himself makes them explicit. Hence, understanding and synthesizing its floor plan is the next best method to generate its interpretation. However, a floor plan itself is a graphical document that has inherent complexity and its graphical language. It is challenging to understand each component of a floor plan and interpret it in a human-like description. This generated description may have other applications, such as indoor navigation for robots, the visually impaired, navigation in historical monuments, and online property recommendation systems. Generalizing this system for general document images

such as engineering drawings, facility plans, and mathematical expressions may be developed into a self-learning system for students.

Figure. 1.2 depicts the inspiration behind the work presented in this thesis. Given the indoor scene images of a house or any other building, and its floor plan, its interpretation in textual form can be generated, which may be useful in automatically describing it in rental websites, or used as navigation for robots or visually impaired or can be developed into a self-learning tool for students for general graphical images.

The house or any other building has a variety of indoor spaces, but they do not depict the quantitative information or the objectiveness involved in constructing it. However, the floor plan represents the 2D plan of the entire construction. The description of the completed building is generated after understanding the intricate details in the floor plan. We also raised the question, "What if we could also understand these indoor spaces and generate the floor plan also without the user having to give it as input?".

The solution to that problem has many existing solutions, requiring specialized depth-sensing hardware or mobile phones with depth perception. Understanding an indoor scene has many components, such as the perception of the depth, knowledge of pixels belonging to walls, edges, objects, the sequence of images taken. For generating the geometry of the entire house, there is a requirement for an image of every scene and the sequence in which the photos were captured. Also, to have a plan that could agree with the real dimensions, there is a requirement of accurate depth perception and scale of the scene. The state-of-the-art techniques require specialized hardware, manual interruption, and less occlusion in the background. Since most of the online rental platforms have customers who are common users. Moreover, the availability of such hardware or techniques might not be possible, there was a requirement of a simpler system which could generate similar results with a conventional mobile phone's camera and with the help of a few captured images in the indoor space.

Thus, upon analyzing the existing work in the area of indoor space understanding and architectural floor plans, it was observed that online rental platforms, buyers and sellers are ubiquitous and hence automating their requirements is the need of the hour. Hence, this thesis focuses towards building a bridge between the modern age requirements and part of their solutions.

1.2 PROBLEM STATEMENT

The goal of the work proposed in this thesis is to design an indoor space understanding and interpretation system to perform the following task:

Given the indoor space of a house or any other building, the system could understand it and interpret its architecture in natural language using its 2D representation or by generating one. (refer Fig. 1.2).

As proposed in this thesis, understanding of indoor spaces and their interpretations, the input to the system would be a floor plan of indoor spaces. The expected output would be its interpretation in natural language as written by a human. Also, in the later stage, a floor plan's requirement is tried to be released to understand indoor spaces by generating a floor plan from indoor space images itself. Appropriate features are proposed which represent an entire floor plan to capture the intricate details in the description. Also, methods are proposed to generate a description which is close enough to human written sentences capturing all the details and grammatically well

placed.

1.2.1 Brief description of the Work Done

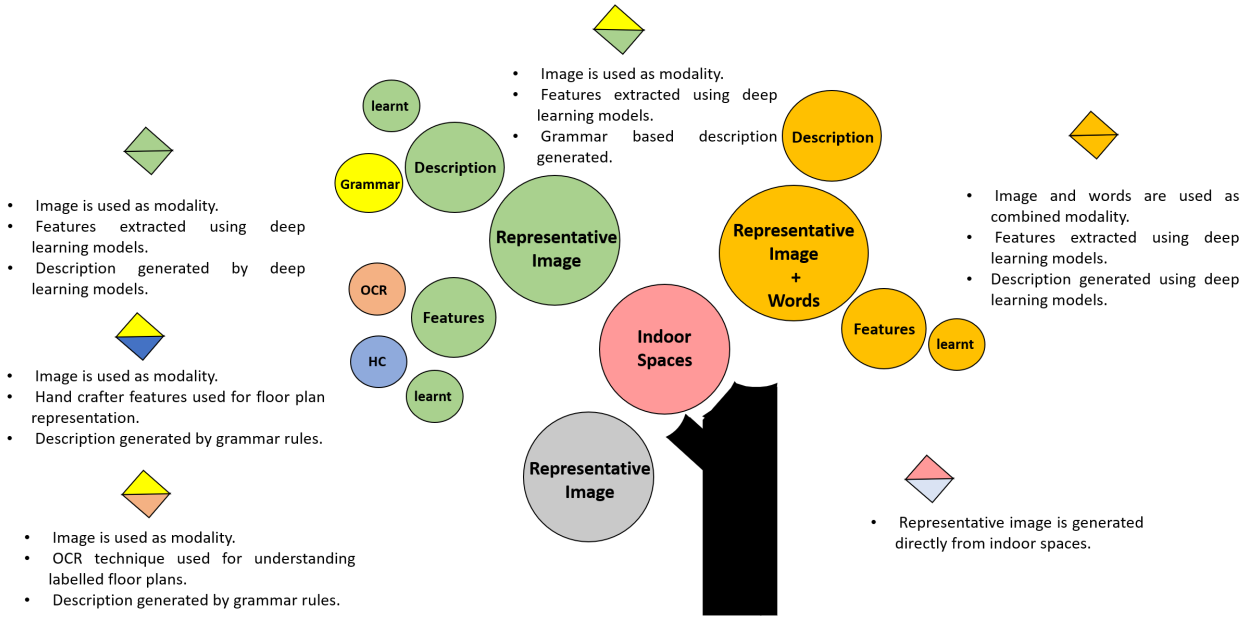


Figure 1.3 : Brief overview of the work done shown over a tree visualization.

Figure 1.3, indoor space understanding and interpretation in the textual description are represented with various possible approaches using a tree visualization. The two main branches of the tree show the two modalities used in this work: representative image (floor plan) cues and the same image with word cues. Modality is varied with image and image with words, where OCR can do image understanding, hand-crafted features (HC), and learned features using advanced deep learning methods. The description generation schemes are grammar-based and learning-based. The upper triangle in every text panel shows the approach used for generating a description, and the lower triangle indicates the procedure for feature extraction from images. Since indoor space understanding also requires a representative image, which may or may not be available to the user, the generation of floor plans as a 2D representation of an indoor scene is also proposed in this thesis.

The indoor environment floor plan can be understood and interpreted in the textual description by various methods, such as taking cues from the floor plan image directly and taking cues from the image along with some extra word-based knowledge. The techniques can be rule-based or learning-based to extract signals from the image. The rule-based methods can only be used for annotated floor plans and employ OCR techniques to read the annotations. For learning-based methods, feature extraction (image representation primitives) should be done using hand-crafted or learned features. The ways to generate textual descriptions could also be template-based, grammar rules, or learning-based. Floor plan generation is done by taking RGB images of the indoor environment taken from a conventional camera and generating a layout for each RGB image by 3D modeling. All partial 3D reconstructions are combined using information extracted from built-in SLAM technology in Google ARcore library and projected in 2D to generate a 2D floor plan of the entire indoor space.

1.3 RESEARCH CHALLENGES

Some of the key challenges encountered in the task of generating textual interpretation by understanding the indoor scene by their floor plans and later releasing the requirements for floor plans are highlighted below:

1. Floor plan images are complex graphical images and have several overlapping graphics, making the segmentation of different components a challenging task.

2. There is wide variability in decor symbols' representation. Coming up with a specific set of symbols is difficult.

3. Variability is also present in the global layout and shape of the floor plans, making it difficult for a network to learn shape-related information about a floor plan.

4. Floor plans are the 2D blueprints of a building or interior space, consisting of a lot of blank space (featureless space), unlike natural images. Hence, it is complicated for a deep learning model to learn the features directly from images.

5. While implementing image to text generation models, there is a lack of dataset that contains floor plans in both modalities, images, and corresponding text.

6. While generating the layouts from indoor scene images, it is difficult to capture the irregularity in the walls given the inaccuracy in the depth estimation systems.

7. Generating a layout for an indoor scene with occluded walls and corners is complicated since it gives inaccurate depth perception.

1.4 ORGANIZATION OF THE THESIS

In this chapter, the importance of the problem of indoor scene understanding is introduced. A brief overview of the motivation behind using the floor plans to understand the scene and the use of their textual interpretation is highlighted. This chapter also highlights the inherent research challenges and efficient solutions to solve the problem. As per the problem statement given in Sec. 1.2, Chapter 2 describes the datasets of floor plan analysis, Chapter 4-6 presents the description synthesis techniques, while Chapter 7 presents a technique of floor plan synthesis. A brief overview of the work done towards solving this problem is also presented.

In Chapter 2, a review of the current work is presented in Document Image Analysis, graphics analysis, and floor plan image analysis. It discusses various solved problems in floor plan image analysis and their components. In the context of image to text generation, it discusses different methods developed for image understanding and text generation along with the multimodal datasets available for natural images and floor plans. It also presents a review of the work done in floor plan generation from indoor images.

Chapter 3 presents the datasets used in the proposed work. Along with the augmentation of existing floor plan datasets, it also proposes a dataset "BRIDGE," which contains $\sim$ 13000 floor plan images along with their decor symbol annotations, region-wise captions, and multisentenced paragraphs for experiments and evaluation. This is a first-of-its kind dataset of floor plan images that connects document images with text modality. This dataset is used to train and test advanced

deep learning models for floor plan images for tasks such as decor symbol detection and classification, region-wise caption generation, and paragraph-based description generation.

In Chapter 4, a framework to generate textual descriptions from annotated floor plan images is presented. A novel end-to-end framework for understanding annotated floor plans and developing their ego-centric description is proposed by segmenting and understanding each component, such as walls and their materials, decors, rooms, and global positioning of each room and their neighborhood.

Chapter 5 extends the work proposed in chapter 4, for general un-annotated floor plan images using machine learning methods, is presented. Novel hand-crafted features, BoD and LOFD, are proposed, giving a holistic representation of floor plan images and descriptions generated using grammar-based methods. Moreover, the application of the generated description for obstacle avoidance door to door navigation is presented in this chapter.

Chapter 6 presents description generation models from floor plan images using learned features by advanced deep learning models. It describes a hierarchical RNN based model (DSIC) that learns features directly from the image and generates a paragraph-based description. An attention-based model (TBDG) is also proposed, which introduces a word layer between the image and descriptions, making it more robust than DSIC model. It presents the experiments using all language models and shows how introducing a word layer helps the model achieve higher accuracy.

Chapter 7 presents a floor plan generation model for indoor spaces, where we try to release the requirement of using floor plans as input to understand an indoor scene. This chapter presents a model that generates a floor plan using indoor space images that agree with the indoor space's real dimensions. The data is captured by using a conventional mobile phone's camera, relieving end-users from specialized hardware. All RGB images taken are reconstructed into 3D after depth estimation and compiled into a 2D floor plan using the camera motion data extracted by built-in SLAM technology in Google's ARcore library.

Chapter "Conclusions", concludes the thesis while providing a brief summary of the problem statement, research challenges, and solutions proposed throughout all chapters. It also draws an outline of the various future directions for the proposed work.