2
Background and Related Work

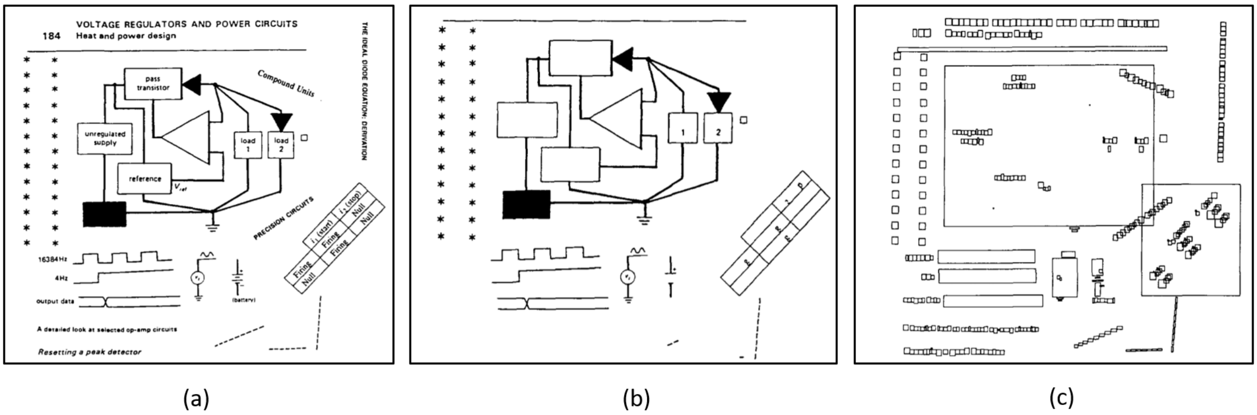
2.1 INTRODUCTION

In this chapter, the related work in the research area of document image processing in the context of graphic recognition is discussed. With the growing number of digital documents these days, graphic recognition has become an important task and gained popularity. All scanned documents containing drawings, diagrams, characters, words, images, and objects are considered graphical documents in document images. These documents include various circuit diagrams, geographical maps, engineering diagrams, architectural diagrams, musical notations, flowcharts, facility diagrams, mathematical operators, etc. They can be printed, scanned documents, or handwritten. These digital documents are ready for analysis. Their extraction, classification, indexing, and conversion into other modalities can be useful for engineers, architects, and other users to generate new documents. The extracted information from them can help an everyday user understand the technicalities in these documents. The focus of this thesis is to analyze and understand architectural drawings. Such drawings contain various graphical components such as thick and thin lines representing walls, the text used for annotations, and graphical symbols representing decor components. The works related to understanding floor plans, images and extracting information from them and the generation of text from images in sentences and paragraphs are also presented here. Additionally, work related to general graphic recognition in engineering drawings, architectural drawings, handwritten documents is discussed. The work related to generating plans/ layouts from indoor scene information is also discussed.

The rest of the chapter is divided in Sec. 2.2, which describe state-of-the-art methods in general graphical recognition, Sec. 2.3 discuss the existing work proposed in the floor plan analysis, Sec. 2.4 discusses various publicly available floor plan datasets available in the literature, Sec. 2.5 describes the existing systems proposed for symbol recognition and classification in document images, Sec. 2.6 describes various methods proposed for object detection and classification, where symbol spotting in documents is presented as an object detection problem, Sec. 2.7 discuss methods proposed for image description generation for natural images, Sec. 2.8 presents the literature about language modelling and evaluation in the line of textual description generation from images, Sec. 2.9 discuss the existing techniques in the literature for 2D/3D layout generation from RGB-D, panorama or 360° indoor scene images.

2.2 STATE OF THE ART IN GRAPHIC RECOGNITION

In the field of document image analysis, graphics recognition is a widely explored area. In the literature, the authors have explored various learning and non-learning-based methods to recognize and classify these graphical symbols. In this context, the authors in the work Chhabra [1997], have given the survey of graphical symbol spotting in various types of documents like engineering drawings and mentioned the general issues addressed in this area. In the work proposed in Rusiñol


VOLTAGE REGULATORS AND POWER CIRCUITS
184 Heat and power design
THE IDEAL DIODE EQUATION: DERIVATION
pass transistor
Compound Units
unregulated supply
load 1
load 2
reference
V_{ref}
PRECISION CIRCUITS
i_1 (start)
i_2 (stop)
Firing
Null
16384Hz
4Hz
output data
(battery)
A detailed look at selected op-amp circuits
Resetting a peak detector
1
2

(a)

(b)

(c)

Figure 2.1 : Process of text graphics separation in method proposed in L.A. Fletcher and R. Kasturi [1988].

and Lladós [2010], authors have given an exhaustive survey for spotting symbols and words in document images and various geometric and photometric descriptors. Authors of the paper, Z. Lu [1998], presented an algorithm to separate text and graphical regions from engineering drawings by erasing graphical components and leaving only text by performing geometric operations on graphical elements. In the document proposed in Lai and Kasturi [1991], authors have given an algorithm to detect dashed line segments by comparing the slopes of the head and tail of each component of a dashed line. In another work, Trier *et al.* [1995], authors have presented a method to extract characters from grayscale images instead of a binary image to save information loss. In L.A. Fletcher and R. Kasturi [1988], authors have proposed an algorithm for text string separation from graphics in document images by extracting connected components and using Hough transform for grouping them. In the context of text spotting, the authors in Almazán *et al.* [2014] have spotted and recognized words in images by representing both words and text strings in a common vectorial space and casting recognition and retrieval as the nearest neighbor problem.

In the proposed work K. Tombre and S. Tabbone and L. Pélissier, B. Lamiroy and P. Dosch [2002], the authors have extended previous work by discussing the right choice of threshold and postprocessing steps for extracting text components. The proposed work in Freeman [1974] discusses various encoding techniques and geometric features for line drawing representations. In another work V. Yadav and N. Ragot [2016], text extraction in document images is done by grouping the densest group of points in the image and finding the connectivity between them. In another work Hilaire and Tombre [2006], authors present a method to vectorize graphics in line drawings. They separated each layer of the input binary image, followed by skeletonization and segmentation of each layer using a random sampling-based method. Figure. 2.1 depicts the process of text graphic separation in graphical documents as proposed in L.A. Fletcher and R. Kasturi [1988].

2.3 STATE OF THE ART IN FLOOR PLAN ANALYSIS

Understanding floor plans is one of the most popular applications due to its wide applications in various real-world areas. Hence, in recent years, numerous systems have been proposed which deal with floor plan understanding. Some examples of such tasks are retrieval, 2D to 3D reconstruction, graphical element recognition, and robot navigation. With the latest artificial intelligent models, these existing approaches can be made more accurate and efficient. However, dealing with these architectural documents has always been challenging due to the inherent heterogeneity in these plans and lack of domain knowledge. In the following paragraph, the most recent techniques to understand floor plans and their interpretations using conventional machine learning methods and current intelligent models are discussed.

In the work Aoki *et al.* [1996], authors proposed a system to convert hand-drawn floor plan drawings to CAD format by offering solutions to positional ambiguity and structural distortions in hand-drawn floor plans. In the same line, the authors in Dosch *et al.* [2000] present a complete system for analyzing architectural drawings by low level preprocessing and recognizing symbols and later performing a 3D reconstruction of the plan. Moreover, in the work Ah-Soon and Tombre [1997] the authors proposed an end to end system to analyze architectural drawings by their segmentation, vectorization, detection of arcs for specific symbols, and structural and symbol analysis. In S. Ahmed, M. Liwicki, M. Weber and A. Dengel [2011], authors have performed text recognition in floor plan images after removing the walls and detecting connected components. In the same line, the authors in Ahmed *et al.* [2011] have introduced a method for preprocessing the floor plan image by differentiating the walls with their thickness and extracting other symbols. As a continuation, the authors in Ahmed *et al.* [2012] have performed room detection and labeling in floor plan images by taking polygon approximation over the detected walls and closing gaps and assigned labels using OCR. In another work proposed in, Mello *et al.* [2012], authors performed image segmentation on ancient topographic maps and floor plan by removing the nontextual part and using OCR. In a similar line, the documents presented in de las Heras *et al.* [2014] propose a statistical patch-based segmentation method used to detect symbols and walls, and a graph-based approach is explored for identifying rooms in the floor plan. In Macé *et al.* [2010] this work, the authors detected rooms by recursively decomposing images into nearly convex regions after detecting walls and doors in the floor plans. In the work proposed in Dodge *et al.* [2017], the authors have introduced a new floor plan dataset and used a fully convolutional network for wall segmentation. In another work, de las Heras *et al.* [2013], authors present an unsupervised method to detect walls in floor plans with different textures without the use of any annotated data. In Sharma *et al.* [2017] authors have proposed a new floor plan dataset and CNN-based algorithm to retrieve similar floor plan images from the dataset. In the work presented in Zeng *et al.* [2019], authors have proposed a deep learning framework with room boundary guided attention-based mechanism for detecting elements in floor plans. Figure 2.2 shows the process of patch-based text and wall segmentation from floor plan images as proposed in de las Heras *et al.* [2014].

2.4 PUBLICLY AVAILABLE FLOOR PLAN DATASETS

In the literature, the publicly available datasets are: ROBIN proposed in Sharma *et al.* [2017], CVC-FP proposed in de las Heras *et al.* [2015], SESYD proposed in Delalandre *et al.* [2010], BRIDGE, proposed by authors in Goyal *et al.* [2019c], CubiCasa5K, proposed in, Kalervo *et al.* [2019], and FPLAN-POLY, proposed in the document, Barducci and Marinai [2012]. Table 2.1 describes the various datasets available in literature with the sample of floor plan images in them and purpose. The ROBIN dataset contains 510 samples of floor plan images with 3 broad categories based on several rooms. This dataset is handcrafted and created primarily for retrieval purposes. Another popular dataset of floor plan images is SESYD, which has 1000 samples of floor plans divided into 10 classes. This dataset contains synthetically generated floor plans, and its primary purpose was symbol spotting and retrieval. This dataset consists of low inter-class similarity in plans while it has high intra-class similarity. Moreover, CVC-FP contains 122 samples of floor plans divided into 4 different layout classes, which are scanned documents. This dataset was aimed at learning and analyzing the structural arrangements in a floor plan. Another dataset, FPLAN-POLY, contains 42 floor plan vectorized images, and its primary purpose was to evaluate

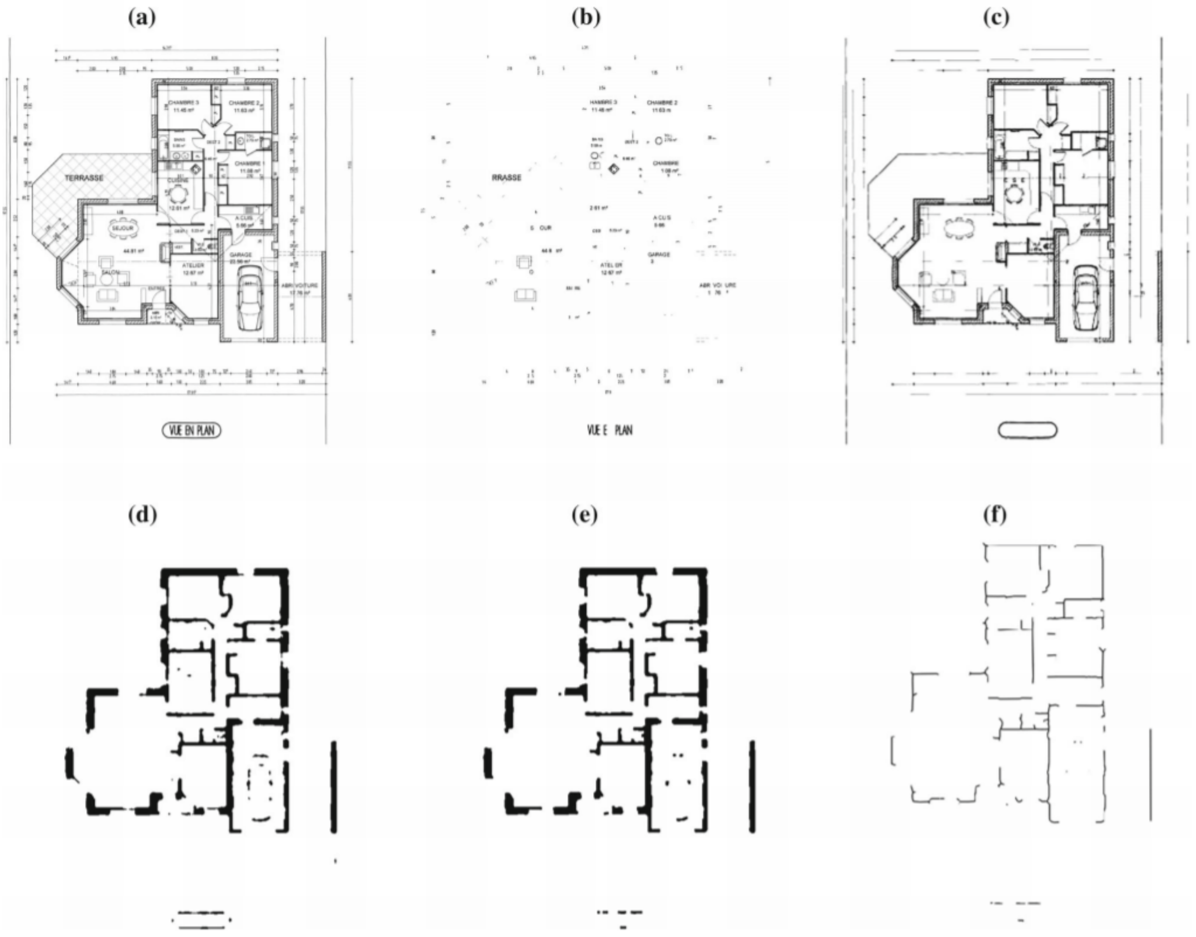


Figure 2.2 : Process of patch based wall segmentation proposed in de las Heras *et al.* [2014].

different symbol spotting methods. Another dataset, CubiCasa5K, contains 5000 floor plan images annotated into 80 floor plan object categories. This dataset is constructed primarily for detecting floor plan components such as rooms, walls, and decors.

The BRIDGE dataset contains 13000+ images and annotations. These annotations include decor symbol annotations, region-wise captions, and paragraph-based annotations. This dataset was primarily created for extracting information from floor plan images and generating textual descriptions from them. Table 2.1 contains the details of some publicly available floor plan datasets.

Table 2.1 : Details of publicly available existing floor plan datasets

Floor plan Datasets		
Dataset	Count	Remarks
CVC-FP [de las Heras *et al.*, 2015]	122	4 Sub-Categories, varying in wall textures, to study graphical notations in floor plans
FPLAN-POLY [Rusiñol *et al.*, 2010]	42	Used for floor plan analysis and room analysis
SESYD [Delalandre *et al.*, 2007]	1000	100 layouts/ class, differ in arrangement of symbols, used for symbol spotting tasks
ROBIN [Sharma *et al.*, 2017]	510	Used for retrieval and symbol spotting tasks
CubiCasa5K [Kalervo *et al.*, 2019]	5000	Used for floor plan object detection tasks
BRIDGE [Goyal *et al.*, 2019c]	$\sim$ 13000	Used for information extraction and description generation

2.5 SYMBOL SPOTTING IN DOCUMENT IMAGES

Symbol spotting in floor plan images deals with extracting information from floor plans for understanding the contents. Symbol spotting in document images has been widely explored using hand-crafted features and conventional machine learning classifiers. In the proposed work in Dutta *et al.* [2013], the authors presented a method for symbol spotting by sub-graph matching and reduced the computational complexity by using graph serialization. In another work Viola and Jones [2001] , authors have introduced a new image representation as "integral image" and a learning algorithm which selects a small number of critical visual features for efficient classification. In a similar line Le Bodic *et al.* [2012] authors have posed symbol spotting as a substitution-tolerant subgraph isomorphism problem, formulating it as integer linear programming. In the work proposed by Barducci and Marinai [2012], authors have done symbol spotting in floor plans by converting them into Region Adjacency Graphs where each symbol is a node. In another work Nayef and Breuel [2011], authors have performed symbol spotting in architectural line drawings by partitioning them into shapes and finding salient convex groups of geometric primitives. In the document presented in Fornés *et al.* [2010], a dynamic time wrapping based invariant rotation algorithm is proposed for handwritten symbol recognition. In Ah-Soon [1997] this work, the authors proposed an exact and inexact graph matching network based on Messmer's network, where symbols are represented as a constraint on segments and arcs of symbols. In another work, Valveny and Martí [1999], authors proposed a method to recognize handwritten symbols in architectural drawings by finding the best fit between the input image of a symbol and deformable templates of symbols. In the context of symbol recognition as presented in Baró *et al.* [2019], authors have proposed a recurrent convolutional neural network-based baseline for recognizing handwritten music scores. In Qureshi

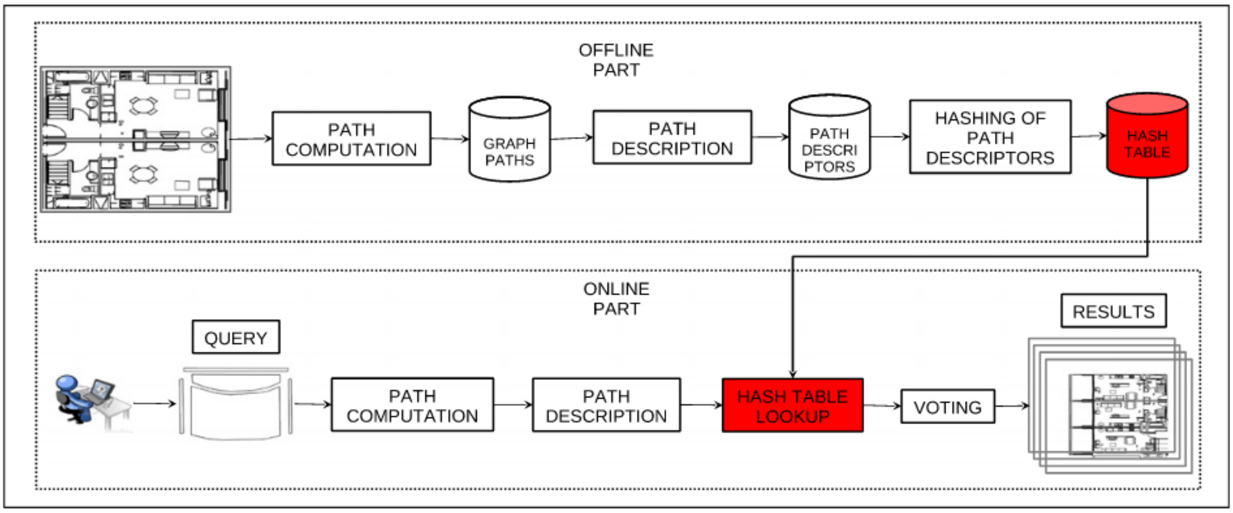
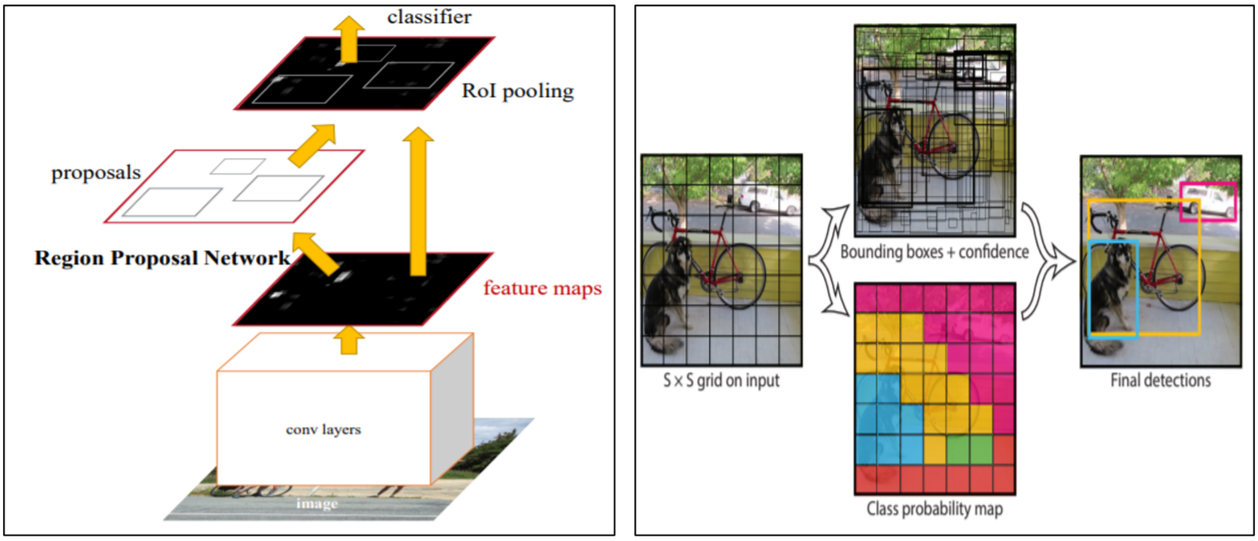


Figure 2.3 : Symbol spotting framework proposed in Dutta *et al.* [2013] using sub-graph matching.

et al. [2007] work, the authors proposed a graph-based representation of document images and spotting, identifying symbols as sub-graphs of that document. Figure 2.3 shows the symbol spotting framework as proposed in Dutta *et al.* [2013], in which the offline part runs to evaluate the paths of the acyclic graphs and their descriptors, and the online part does the sub-graph matching for the query documents for the symbol spotting task.

2.6 OBJECT DETECTION AND CLASSIFICATION

Symbol spotting tasks in document images can be seen as an object detection problem in images. However, object detection is a widely explored area in natural images by using conventional machine learning methods. With the advent of deep neural networks, similar tasks such as object detection, localization, and classification are being performed with much better accuracy. In the similar context, several schemes have been proposed. For example, the schemes presented in YOLO Redmon *et al.* [2016], Fast-RCNN Girshick [2015], Faster-RCNN Ren *et al.* [2015] for natural images. There are majorly two types of models available, region proposal based and region classification based. The YOLO family-based algorithms, presented by Redmon *et al.* [2016], Redmon and Farhadi [2017], Redmon and Farhadi [2018] are region classification-based methods, which is a single neural network trained end-to-end and predicts bounding boxes and class labels directly. However, all the R-CNN family-based proposed methods, for example, Girshick [2015], Ren *et al.* [2015], He *et al.* [2017], He *et al.* [2015], are region proposal based methods, which extract several regions from the input image and extract features from those regions using CNN and classify them using a classifier. In one of the works presented in, Ziran and Marinai [2018], symbol spotting in floor plans using YOLO and Fast-RCNN has also been explored. Object detection in document images in the form of detecting objects such as tables, equations, figures has been done using convolutional neural networks and region proposal networks has been presented in Yi *et al.* [2017], Saha *et al.* [2019], Schreiber *et al.* [2017]. In a similar direction, the authors detect signatures and logos using deep learning models in Sharma *et al.* [2018]. Figure 2.4 shows the object detection schemes proposed by, Faster-RCNN Ren *et al.* [2015] and YOLO Redmon *et al.* [2016], showing the difference between region proposal based and region classification based methods.



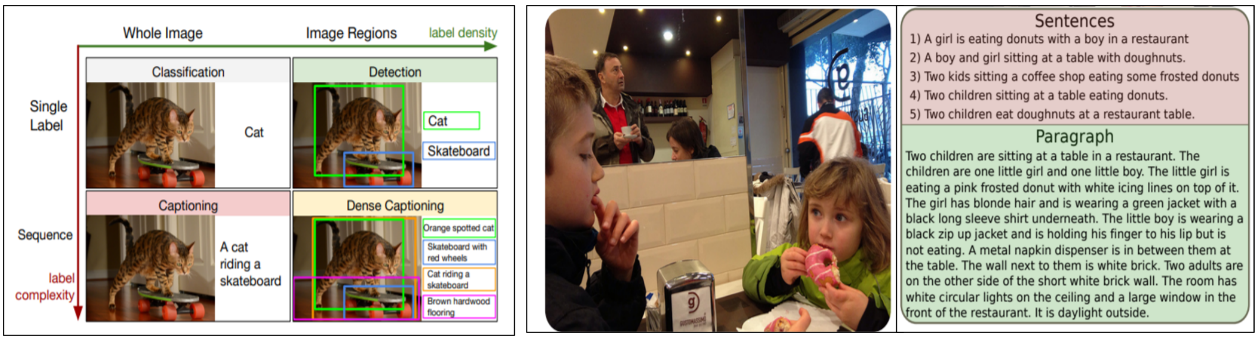
(a) (b)

Figure 2.4 : Object detection by (a) Faster-RCNN proposed by Ren *et al.* [2015] and (b) YOLO proposed by Redmon *et al.* [2016].

2.7 IMAGE DESCRIPTION GENERATION

Image description generation is a challenging task in computer vision and natural language processing. This area has been widely explored using traditional language generation techniques such as template based, retrieval, n-grams, grammar rules, and current deep neural networks such as RNN, LSTM, GRU, etc. These methods work with image modality features by extracting information related to image using conventional methods of object/scene recognition or CNN based architectures. Some of the initial works in this direction were proposed in the documents, Farhadi *et al.* [2010], Kulkarni *et al.* [2013], Li *et al.* [2011], Ordonez *et al.* [2011] which are using computer vision methods for extracting attributes from an image and generating sentences using retrieval and n-gram techniques. While using deep neural networks caption generation for natural images, Densecap Johnson *et al.* [2016] has proposed an algorithm that generates region-wise captions in images. Before Densecap, image captioning was dealt over the entire image instead of regions.

Since paragraphs are richer than the caption, image to paragraph generation was the next task to be considered. In this line, the hierarchical recurrent network Krause *et al.* [2017] based paragraph generation technique produces a paragraph-like description for the entire image by hierarchically using two RNNs. One RNN is used to regress over sentences, and the others will regress over words. In another work presented in Wang and Chan [2018], two CNN networks are used, where one of them is used as an encoder for image features and the other CNN is used as a decoder for language generation. In a similar line, the work proposed in Chatterjee and Schwing [2018] has used a sentence topic generator network from visual features of images, and RNN is used for sentence generation. In another work Wang *et al.* [2018], depth-aware attention model is used for generating a detailed paragraph from an image. In the work presented in Mao *et al.* [2018], latent Dirichlet allocation (LDA) is used to mine topics of interest from textual descriptions and generate multiple topic-oriented sentences to describe an image. In another work Yao *et al.* [2017], CNN and RNN framework is used to capture visual features and generate descriptions and enhanced descriptions by using attributes generated from images. The image description is also



Sentences

1) A girl is eating donuts with a boy in a restaurant
2) A boy and girl sitting at a table with doughnuts.
3) Two kids sitting a coffee shop eating some frosted donuts
4) Two children sitting at a table eating donuts.
5) Two children eat doughnuts at a restaurant table.

Paragraph

Two children are sitting at a table in a restaurant. The children are one little girl and one little boy. The little girl is eating a pink frosted donut with white icing lines on top of it. The girl has blonde hair and is wearing a green jacket with a black long sleeve shirt underneath. The little boy is wearing a black zip up jacket and is holding his finger to his lip but is not eating. A metal napkin dispenser is in between them at the table. The wall next to them is white brick. Two adults are on the other side of the short white brick wall. The room has white circular lights on the ceiling and a large window in the front of the restaurant. It is daylight outside.

(a)

(b)

Figure 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].

generated from a stream of images in the proposed work by Park and Kim [2015] by using CNN and RNN for a visual feature and sequence encoding and retrieving sentences from an existing database. In a similar line, the method presented in Liu *et al.* [2017], the description has been generated as storytelling from images by using the Bi-directional attention-based RNN model. The reasoning over images in the form of captioning or question-answering is not limited to natural images. Authors in Mondal and Jawahar [2019] have proposed a method to generate textual description from mathematical equations, by using a CNN for visual feature extraction and LSTM for language generation. In another direction, the authors in Kahou *et al.* [2017] proposed a novel large-scale dataset, FigureQA, for visual reasoning over line plots, dot-line plots, vertical and horizontal bar graphs, and pie charts. The synthetically generated corpus FigureQA consists of around 100,000 common scientific plots and related questions and answers. Another synthetically generated dataset in the same line, DVQA is proposed by authors in Kafle *et al.* [2018] for reasoning over bar charts, by introducing variability in existing datasets and demonstrating the lack of generalisation in existing question-answering algorithms. Following the same line, the authors in Reddy *et al.* [2019] proposed a novel architecture FigureNet for reasoning and understanding of the scientific plots. FigureNet uses CNN for generating object representations by making color identification in scientific plots as key requirement, and LSTM for encoding questions, while demonstrating their experiments on FigureQA dataset. In this line, the authors in Methani *et al.* [2020] presented a novel dataset PlotQA for reasoning over scientific plots containing 224,377 plots on data from real-world sources and questions based on crowd-sourced question templates, for overcoming the gaps present in the previous datasets. They presented their experiments using multi-staged and end-to-end question-answering pipeline. Following the direction of reasoning over content different from natural images, authors in Kembhavi *et al.* [2017] presented dataset Textbook Question Answering (TQA) and algorithm for reasoning over textbook data that is a combination of textual content and visual content, which consist of both natural images and diagrams. The dataset consists of over 1,076 textbook lessons accompanied by 26,260 questions. Figure. 2.5 describes two types of image description methods, Densecap Johnson *et al.* [2016], generate a single sentence description for different regions of an image and Krause *et al.* [2017] generates multi-sentence paragraphs for describing intricate details in an image.

2.8 LANGUAGE MODELLING AND EVALUATION OF TEXT GENERATION

Since deep neural networks are very successful in natural language processing, learning text for generating descriptions using sequence-to-sequence models is a natural choice. In the work, Sutskever *et al.* [2014] has proposed seq2seq learning model for learning and modeling language by LSTM model. Moreover, Bahdanau *et al.* [2014] and Luong *et al.* [2015] are the initial models which model the language by aligning the input sequence to a target sequence using attention-based models. The neural machine translation model has also been used in text summarization tasks such as Rush *et al.* [2015] and Nallapati *et al.* [2016]. Evaluation of the generated description from images is an important step to identify the accuracy of the generation. For that purpose, several metrics, for example, BLEU (BiLingual Evaluation Understudy) as proposed in Papineni *et al.* [2002], ROUGE (Recall Oriented Understudy of Gisting Evaluation) as proposed in document Lin [2004], METEOR (Metric for Evaluation of Translation with Explicit ORdering) presented in document Denkowski and Lavie [2011], CIDEr (Consensus-based Image Description Evaluation) as proposed by Vedantam *et al.* [2015], SPICE (Semantic Propositional Image Caption Evaluation) as proposed by Anderson *et al.* [2016], BERTscore as presented by Zhang *et al.* [2019a], and their several variations have been proposed in the existing literature. In another work presented by Elliott and Keller [2014], the authors provided a correlation between automatic metrics and human judgments, using the previously mentioned metrics and their variants. These metrics compare the generated text with the human-written text on parameters such as uni-gram, bi-gram or $n-$gram and give the similarity score for each token between the candidate sentence and reference sentence in terms of precision, recall and F-score.

2.9 LAYOUT ESTIMATION

Layout estimation from RGB-D and panorama images has been a widely explored problem. Liu et al. Liu *et al.* [2015] and Zhang et al. Zhang *et al.* [2013] have reconstructed the indoor scene in 3D using monocular images and estimated the layout using vanishing points and depth features. In the work Zhang *et al.* [2019b], authors have calculated layouts on RGB images by using an encoder-decoder framework to jointly learn the edge maps and semantic labels of each image. In a similar line, the method presented in Bao *et al.* [2014], the room layout is generated from images taken from multiple views and reconstructed using SfM and region classification. In another work proposed by Dasgupta *et al.* [2016], the structure is estimated for the cluttered indoor scene by identifying label for a pixel from RGB images, using deep FCNN, and refined using geometrical techniques.

Since monocular images can not capture the entire scene, layout estimation from panorama images to increase the field of views has been explored. In the recent work proposed in, Zou *et al.* [2018] has proposed an encoder-decoder network which predicts boundary and corner maps and optimized the over-constrained geometrical properties of the room. Another work presented in Sun *et al.* [2019] estimated the room layout by regressing over boundaries and classifying the corners for each column representation of an image. Furthermore, Fernandez-Labrador *et al.* [2018] generated a 3D layout for 360-degree panorama image by reasoning between geometry and edge maps returned by the deep neural network. In a similar direction, the authors in Hsiao *et al.* [2019] proposed an approach to generate an indoor scene layout by learning the encoded layout representation in row vectors. In addition, another work proposed in Xu *et al.* [2017], Zhang *et al.* [2014], the layout is obtained from the panorama image by estimating object locations and pose in the room.

The indoor layout has also been generated from monocular video sequences in the method

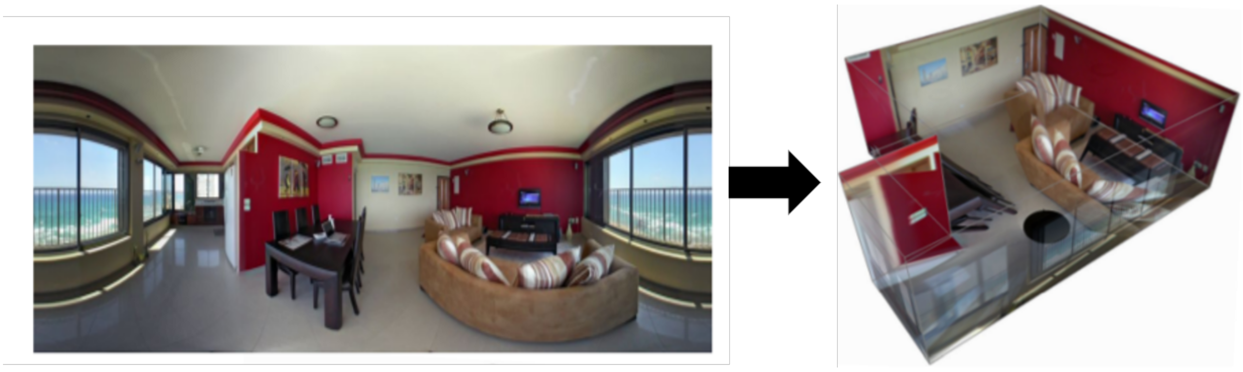


Figure 2.6 : 3D room layout synthesis proposed by LayoutNet in the document Zou *et al.* [2018].

proposed by Furlan *et al.* [2013] where SLAM/SfM techniques are used for 3D reconstruction, and the layout is generated by fitting the planes. In another work proposed by Angladon [2018], authors have developed a 2D structure of the room using depth data and 3D reconstruction using SLAM. In the context of floor plan generation, the schemes in Cabral and Furukawa [2014] have reconstructed the layout from the input panorama images using SfM and generated a 2D floor plan posing it as the shortest path problem. Another work in Lin *et al.* [2018] proposes methods to predict the global room layout and transformations using partial reconstructions of indoor scenes using RGB-D images without making use of feature matching between partial scans. In another work, Phalak *et al.* [2020] authors have presented a floor plan estimation method from 3D indoor scans by proposing a 2 stage method, where the first stage clusters the wall and room instances and the second stage predicts the perimeter of the rooms. In a similar line, work proposed by Murali *et al.* [2017], authors used a system to generate a Building Information Model of the house's interior from 3D scans by detecting walls and performing reasoning about their neighborhood relations.

In another work presented in Chen *et al.* [2015a] the authors proposed a system CrowdMap, which used sensor-rich video data for reconstructing floor plans from indoor scenes. Authors in Turner and Zakhor [2014] have generated floor plans by triangulating the space between wall samples and partitioning the space into the interior and exterior space. In a similar direction Chelani *et al.* [2018] the authors have generated a 2D floor plan by estimating camera motion trajectories from the depth and RGB sequences and by registering 3D data into a simplified rectilinear representation. In another work presented by Okorn *et al.* [2010] authors have proposed a system to create 2D floor plans for building interiors which takes 3D point clouds obtained from 3D scanners by utilizing height data for detecting wall and ceiling points and projecting the rest of the points in the 2D plane. Figure 2.6 shows the 3D layout generation system proposed by Zou *et al.* [2018] for an indoor scene using the panoramic image of the same.

2.10 SUMMARY

In this Chapter, various research areas such as graphic recognition, floor plan analysis, symbol spotting, object detection in natural images, and a textual description of them using multiple language models and layout synthesis from indoor scene images are discussed. The existing literature in these areas has proposed many methods to analyze indoor scenes and generate image descriptions from natural images. A lack of schemes could connect these two domains and interpret indoor settings in an everyday users' understandable format. Moreover, the existing publicly available datasets for floor plan images do not include text modality, and they are not suitable for textual interpretation. With the advancement of artificially intelligent models, there is a lack of adequate data and diversity in current datasets to train or fine-tune them. Hence, the next chapter describes the datasets used in the proposed work, the approaches taken for their creation, and experiments to validate their utility. Moreover, the consequent chapters in this thesis for indoor scene understanding and interpretation are proposing a few approaches to develop a system that will help typical users in various requirements related to buy or sell their property and other applications.