Abstract

Understanding the indoor spaces from images, videos, or other visual information has become an essential task in the current scenario. With the rapid urbanization of cities and the common user's quick requirements, most real estate businesses have become web-based. It has become essential to understand an indoor space from its visuals and provide users with an interpretation. These users can be looking for a house with their desired features, architect, or an interior designer, trying to understand the given building. A person looking for a home or office space requires a quick solution for his/her property based needs with desired features. Hence, it is essential for showcasing the properties online, with the proper display of their indoor and outdoor environment and detailed interpretation. An interpretation provided in natural language is the most powerful way to communicate the information in the indoor space's visuals. Another form of interpretation could be a line drawing or map for much complex indoor space visuals.

While searching any property, every potential customer looks up for its detailed images and description available about the number of bedrooms, bathrooms, details of kitchen, halls, balconies, and global relations present in them. In case of lack of a proper description for any house listing, users tend to skip buying or renting it. A real estate website or any rental website may have thousands of listings. To reduce manual efforts for understanding and to give a correct interpretation for these listings, it is required to device an automation system that could simultaneously understand visual stimuli and generate an interpretation out of it.

In case of textual interpretation, just by looking at the stream of indoor scene images for different rooms, it is impossible to find the exact number of bedrooms, bathrooms, etc. available in the house since these images could be randomly ordered. It is possible to have extra information about one part of the property and to miss out on details about the other. It is also difficult to identify the global relationship between the rooms and their possible arrangement. Hence, instead of directly taking indoor scene images as input, a representation is the next best input that contains all the information required. This representation could be information extracted from these images by keywords or a more precise line drawing/ map that connects all the information. In indoor space images, its floor plan image is the best representation that captures the intricate details and the heart of all the construction drawings. In this thesis, an automation model is proposed which understand indoor space images and generate interpretation in the form of textual description from a floor plan image.

Analyzing the floor plan of a house or any other building is the research area covered under graphics analysis within the broad scope of document image analysis. Understanding every detail of the graphics involved in the floor plans and interpreting them in a form readable by the common user is not a trivial task. A floor plan image can compose of graphics such as symbols, various forms of lines (thin, medium, thick), text, and each of them may require different techniques for their understanding and analysis. Hence, the architectural floor plan image analysis aims to extract the structural details and understand the semantics involved. These extracted details should be converted into textual modalities to have a fair interpretation of the floor plan. There is a requirement of datasets for each specific task for the development of any machine learning algorithm. There are existing floor plan datasets in the public domain, which were constructed primarily for the task of symbol spotting, retrieval, and structural analysis in the floor plans. Examples include Systems Evaluation SYnthetic Documents (SESYD), Computer Vision Center-Floor Plan (CVC-FP), and Repository Of BuildIng plaNs (ROBIN). These datasets were necessary but not sufficient for the holistic understanding and interpretation of the floor plans in other

modalities. Another dataset, namely Building plan Repository for Image Description Generation and Evaluation (BRIDGE), is proposed to fill those gaps. It is a large-scale floor plan dataset containing textual annotations corresponding to each image. This dataset was targeted for multiple tasks such as symbol spotting, caption generation, paragraph generation, and sufficient for a complete understanding and interpretation of the floor plan image. To build machine learning-based models for understanding floor plans, features such as Bag of Decors (BoD) and Local Oriented Feature Descriptor (LOFD) were proposed, which captured the features of rooms in a floor plan in the form of a sparse histogram. These features and the decor symbol spotting method generated attributes for a floor plan that were later used to create a grammar-based textual description.

With the advent of advanced and more accurate artificial intelligent models for image understanding, it was the need of an hour to build such models for floor plan image understanding as well. Hence, in another attempt, experiments generate textual descriptions from floor plan images using just image cues by extracting features from the CNN-based model and using a hierarchical recurrent neural network-based model for generating paragraphs. To make these paragraphs more accurate and robust, another model was proposed which used a text layer in between image features and textual paragraphs, hence taking the image and word cues together to train a model.

While generating an interpretation of the indoor space, keeping their applications in mind for real estate business and architectural solutions, there might be a case where a user might not have the floor plan of the building or property due to old structures or any modifications made in the past or rented property. In such cases, it might be required to used indoor space images directly instead of using a representation for generating an interpretation. Hence, in another part of the work, we proposed a system that could create a floor plan as a pictorial interpretation of the house, taking a stream of indoor space as input. The system uses a conventional mobile phone's camera to capture images and data from IMU sensors of the phone to track the mobile device's motion. The system uses captured RGB images and this data and generates a floor plan of the indoor space. The proposed approach has other potential applications, such as robot navigation and AR/VR applications. The generated description can also be used for door-to-door indoor navigation for a visually impaired person, robot. Using such a model a self-teaching system for engineering students can be developed, which could automatically generate interpretations for engineering drawings. Hence the proposed work in this thesis opens new avenues of research in engineering and architectural drawing connecting the two modalities of images and text.