

SUGAMAN: Describing Floor Plans for Visually Impaired by Annotation Learning and Proximity based Grammar

Shreya Goyal, Satya Bhavsar, Shreya Patel, Chiranjoy Chattopadhyay, Gaurav Bhatnagar

Abstract

In this paper, we propose SUGAMAN (Supervised and Unified framework using Grammar and Annotation Model for Access and Navigation). SUGAMAN is a Hindi word meaning “easy passage from one place to another”. SUGAMAN synthesizes textual description from a given floor plan image for the visually impaired. A visually impaired person can navigate in an indoor environment using the textual description generated by SUGAMAN. With the help of a text reader software the target user can understand the rooms within the building and arrangement of furniture to navigate. SUGAMAN is the first framework for describing a floor plan and giving direction for obstacle-free movement within a building. We learn 5 classes of room categories from 1355 room image samples under a supervised learning paradigm. These learned annotations are fed into a description synthesis framework to yield a holistic description of a floor plan image. We demonstrate the performance of various supervised classifiers on room learning. We also provide a comparative analysis of system generated and human written descriptions. SUGAMAN gives state of the art performance on challenging, real-world floor plan images. This work can be applied to areas like understanding floor plans of historical monuments, stability analysis of buildings, and retrieval.

I. INTRODUCTION

One of the primary goals of a combined framework involving computer vision (CV) and natural language processing (NLP) is to understand an image and describe it. The techniques available in CV and digital image processing (DIP) helps to localize the objects in the image, identify key attributes and provide a relationship among them. On the other hand, NLP provides us with an end to end description of that image, and thereby connecting the output of CV with the text. For the purpose of having a better understanding of a floor plan image, we propose SUGAMAN (Supervised and Unified framework using Grammar and Annotation Model for Access and Navigation), which is an attempt to connect the these two modalities, i.e. CV and NLP, in the context of document images and text data. SUGAMAN is a Hindi word which translates to easy passage from one place to another. Apart from describing the general information about the floor plan images, it also generates room to room navigation information, while avoiding the obstacles. This navigation information can be very useful for visually impaired people as it becomes difficult for them to move in an indoor environment. It will be really helpful for them if there is a system that tells about the surroundings environment in natural language. SUGAMAN generates such natural language description of an indoor environment from building floor plan images, which gives a detailed idea of the indoor environment. Here the input is a building floor plan image and the output is a textual description of the same. The description includes detail about the (i) rooms, (ii) connectivity among the rooms, (iii) type of decor within every room, and (iv) their relative position, and (v) navigational information, while avoiding obstacles.

In the past few years, unconventional documents like floor plans, engineering drawing gained a lot of attention from the document analysis and recognition (DAR) community. Engineering drawings contains many symbols, texts, line drawings, which need to be recognized and understood. Understanding of circuit diagrams, floor plan images, machine designs, building diagrams, maps etc. has its own importance. While researchers have looked into the problems of segmentation, symbol spotting, text recognition, symbol classification etc., the problem of narration synthesis from those documents was overlooked. Generating a textual description (narration) from an engineering drawing document image can be very useful for a layman to understand the document. As a particular example, building floor plan images involve a lot of technicalities such as symbols, dimension of the building, its design, orientation, as well as the interior decors, which are not known by common users. Various techniques have been proposed for its wall segmentation, text, room and decor segmentation in order to understand them. Even though text synthesis has been an area covered under natural language processing for years and also explored with real world images, document images were always ignored.

Figure 1 exemplifies the problem and the potential solution for a real-world floor plan images. The key characteristics that makes this work unique are: (i) proposing a unified framework for narration synthesis from floor plan images, (ii) improvement in the previously available techniques for decor characterization, (iii) proposal of a novel feature to represent a room within a floor plan, (iv) learning the room annotations for room classification, and (v) augmentation of an publicly available dataset by annotating floor plan images with textual descriptions.

The paper is organized in the following manner: in Sec. II we discuss the known results that are published in the literature and are related to our framework. Overview of the proposed system is given in Sec. III. Description about the dataset used is given

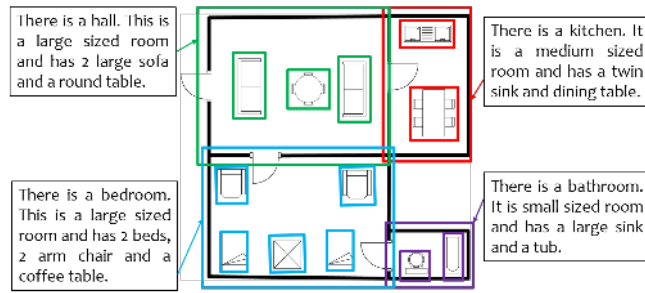


Fig. 1. Introducing the problem of narration synthesis from a given input floor plan image.

in Sec. IV. Room annotation learning and classification framework is presented in Sec. V. The method of description synthesis is discussed in Sec. VI. Results of the intermediate processing stage is discussed in Sec. VII. Results of description synthesis is given in Sec. VIII. Finally, the paper is concluded in Sec. IX.

II. RELATED WORK

In this section, first we describe works related to the document image processing and recognition tasks, followed by the ones proposed for description synthesis.

A. Document image processing and recognition

Symbol spotting, character recognition and feature extraction has been an evergreen research topic for classification of document images. An overview of graphics symbol recognition can be found in [51]. In this matter [65] provides a detailed state of the art survey for symbol spotting techniques and there performance analysis. In [49], a class of drawings are analyzed which includes flow charts by pre-processing scanned binary images, symbols and connection lines followed by extraction of features like shapes, orientation of lines and connections. Trier et al. [45] have also imparted knowledge based systems for interpreting symbols in engineering drawings. In [43], with respect to engineering drawings, an approach for detecting and classifying dashed line segments was proposed. In [8], a new rule based algorithm was proposed to differentiate between text and graphics part from engineering drawings.

In a floor plan, the decors are an inseparable component and visually similar to characters. Properties of characters and their recognition methods could be adopted for recognizing decors. The authors of [47] used topological analysis method to recognize characters and lines from gray scale scanned images of hand printed documents, whose information has been lost after binarization. In [5], CCA and histogram based thresholding is applied to separate text and graphics and Hough transform is used to group together components into logical strings, which was later improved by Tombre et al. [6] through right choice of threshold and its stability.

Language invariant script recognition was proposed in [72] using spatial relationships of features. Two classes of languages are determined using optical density distribution and most frequently occurring word shapes characteristics. Decors symbols can also be considered as line drawing images. In that context Freeman [71] discusses various forms of line drawing representations, processing of line drawing structures derived from images by extracting their features like chain coding scheme, polygon approximation. Qureshi et al. [67] have proposed a solution for symbol spotting using a graph representation of graphical documents. In the same line Dutta et al. [27] has proposed a symbol spotting technique in graphical documents, in which graph represents the document and a sub-graph matching is used to spot symbols. Viola and Jones [69] proposed a learning algorithm by selecting salient visual features and a novel representation called “integral image”, which allows the features used by detector to be computed very quickly. They combined increasingly more complex classifiers in a “cascade”, which allows background regions of the image to be quickly discarded speeding up object detection. [68], proposed a new rotated Haar like features was proposed, which improves the accuracy of object detection by great extent. In [62], Joachims discussed about bag of words approach for objects categorization. However in another work [11], an image is divided into blocks, and blocks with higher density are considered as text by grouping them, where key points are extracted by using FAST [12] method.

In the context of floor plans text/graphic segmentation were performed in [1], [2], [7]. In [32], a new algorithm to segment the ancient maps and floor plans was proposed by removing non textual elements and recognizing characters to identify the plans. In [7], various rooms are detected and labelled by Optical Character Recognition (OCR) on localized text regions. Heras et al. [3] have proposed a Statistical patch based Bag of visual words (BoVW) model to segment floor plan image. Attributed graph of line segments is generated with nodes labelled for wall segmentation, followed by DFS search to obtain walls in [3]. In [3], doors and windows are detected using symbol spotting techniques using SURF feature by detecting the key points in the image. In [9], walls in a floor plan image are detected by exploiting the properties of Hough transform on vectorized image.

Also they have used bag of visual words for the same task and later A* search is performed to find a path between doors. In [3], rooms are detected in the floor plan images by finding the closed regions in WDWC graph, preceded by removal of all the terminals of the graph. Wall contained image is decomposed into convex regions to detect the rooms and holes in the polygons are resolved. Later convex regions are checked for over segmentation by identifying the fitting rectangle of that region in [9]. In [24], a method for recognizing “uniform” binary patterns was proposed.

B. Image description generation

Description generation from images has been an interesting area for aligning images to their corresponding text. Although literature is available on natural images, description of document images is still an untouched domain. In one of our earlier work [18] we have introduced the problem of description generation from floor plan. An extensive survey of the existing techniques is given in [55]. Vinyals et al. [60] presented a generative model, based on deep recurrent architecture, which takes an input image I , and trained to maximize the likelihood $p(S|I)$ of producing a target sequence of words $S = S_1, S_2, \dots$. In the same line [34] discusses a model to generate descriptions from images, which automatically learns to describe the content of images. The work in [59] have presented a holistic data-driven approach to image description generation, which uses vast amount of image data and associated description available over internet. They recognize and predict the contents of an image and then use existing human composed description to generate natural captions for images. In [58], [37], authors have generated dense description of images, by developing a deep neural network model. They introduced a multimodal Recurrent Neural Network architecture that takes an input image and generates its description in text. In [57], authors have used image meta-data for automatically generating textual image collection descriptions that include both image content and context information. Moreover, they convert and expand the meta-data, by using publicly available information and services over internet. Kulkarni et al. [56] generated description of images by detecting objects, modifiers (adjectives) and spatial relationships (prepositions) in an image, comparing and smoothing these detection using pre-available description text.

In the same context Farhadi et al. [38] proposed a system that can obtain a score by linking an image to a sentence. In [59], authors have proposed a data driven approach for description generation by giving a query image and retrieving existing human composed phrases which describe similar images. Combining those phrases they generate a description for the given image. Verma and Jawahar [35] have proposed a system that achieve task like given an image, generating a textual description and vice versa, where both approaches are retrieval based. In the same line Zhu et al. [33] provides an approach to align visual contents of a movie release to their corresponding book by providing a description of the visuals. Image description is generated by creating visual dependency representation of natural images in [25]. Natural language description generation is also done for video for their retrieval purpose in [26], in which they capture relations between keywords associated with videos. Evaluation of machine translation with human generated description is also necessary. For that purpose several metrics for example BLEU [54], ROUGE [53], METEOR [52] etc, have been proposed. In [39], authors provided a correlation between automatic metrics and human judgments, using previously mentioned metrics and their variants. Various approaches have been developed to connect two domains, natural language processing and document image analysis. For example in [28] authors have gained performance by integrating features from linguistic analysis, image text recognition and image layout analysis. Looking at the existing literature we see that even though description synthesis from natural images has become common, the same from document images is still not there, and thus we propose SUGAMAN to bridge the gap. SUGAMAN is an extension of our earlier work [18]. The key differences between SUGAMAN and [18] are: (1) A feature based automatic room annotation learning method is proposed, as compared to OCR based method, (2) An improved proximity based sentence model is proposed in SUGAMAN, in stead of a template based model, (3) introduction of room to room navigational information for obstacle free movement.

III. SYSTEM OVERVIEW

Figure 2 shows a block diagram depicting various modules and work-flow within SUGAMAN. The whole system is divided into two stages (i) room annotation learning and (ii) description synthesis. At first, room semantic information is extracted by room segmentation process, which gives all the required information about an input floor plan image. For example, individual room area, door information which gives room neighborhood information, room locations (room coordinates). With those room locations, the floor plan image is partitioned into individual room image and taken as sample for room annotations learning. Decor characterization is applied over these room images and decors present in a room are labeled. We have proposed a Local Orientation and Frequency Descriptor (LOFD) feature, which is extracted from these room samples for automatic room annotation. A classifier is trained using LOFD feature matrix of room samples by assigning class labels to them. After this a new input image is taken as a test sample, features are extracted and room annotations are identified for it using previously trained model. An XML file is generated using the semantic information extracted by room segmentation and room classification. By parsing this XML file, textual description is generated. SUGAMAN also gives navigation path within the entire floor plan, starting from the entry door to the building. All such information about floor plan and navigation are fed to the proposed grammar model. The first stage of the proposed description synthesis method deals with “*what to say*” about the floor plan and the second stage will deal with “*how to say it*”. For ease of understanding, in rest of the paper, we demonstrate all our analysis on the input image shown in Fig. 1. Later, in the experimental results, we also show the results on other floor plan images. Next we discuss about the dataset used for experimentation.

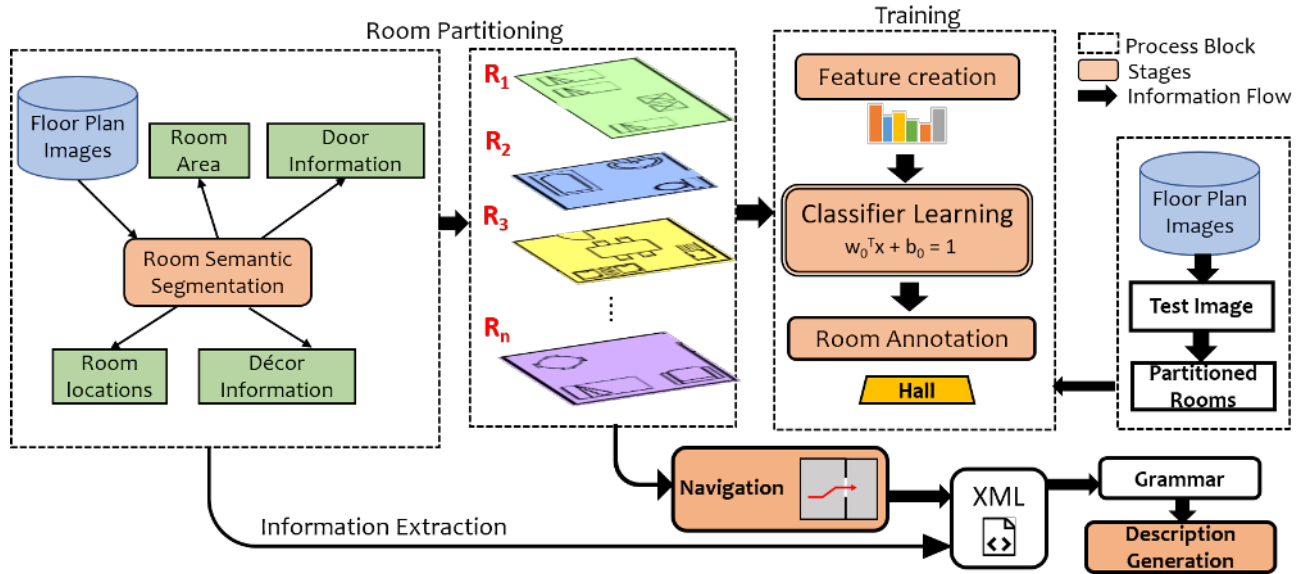


Fig. 2. Block diagram depicting various modules and work-flow within SUGAMAN.

IV. FLOOR PLAN DATASET

For room segmentation, symbol spotting, retrieval in floor plan images three public dataset were proposed. They are: (i) Systems Evaluation SYnthetic Documents (SESYD) [30] (ii) Computer Vision Center Floor Plan (CVC-FP) [31] and (iii) Repository Of BuildIng plaNs (ROBIN) [29]. SESYD has ten classes of floor plans, with 100 samples/class. On the other hand, CVC-FP has 122 scanned floor plan documents divided into four categories based on the origin and style. In ROBIN there are three broad categories, which are different from each other in terms of the number and type of rooms present in a floor plan. The three categories are (i) 3 room, (iv) 4 room, and 5 room floor plans. Each category is further classified into 10 sub-categories depending upon the global layout of the floor plan. ROBIN helps in better visualization of the floor plans and aids in efficient capturing of various high-level features while fine-grained retrieval. Since ROBIN has significant number of floor plans, as well as intra-class similarity and inter-class dissimilarity, it is suitable in our case. However, in ROBIN there is no textual description available for a given floor plan. For our purpose we further augmented ROBIN dataset by introducing textual description for each floor plan image.

A. A-ROBIN

In order to understand the floor plan images better, and produce automated textual description narrating them we propose a dataset *Augmented ROBIN* (A-ROBIN). In A-ROBIN there are four human written descriptions for each image in ROBIN dataset. In there literature, there are a few image datasets for example Flickr8k [22], COCO dataset [23], which has associated descriptions of the image samples. However, these datasets describes the natural images. In the context of document images, dataset consisting description of floor plan images was lacking. Hence we require a novel dataset which contains annotations and descriptions for document images. The descriptions for a floor plan images were collected from the volunteers in the following manner. Each volunteer was supplied with a set of 10 images in a Google form and asked to describe them in their own words based on a set of instructions. Each form was given to 4 volunteers, so that each image has 4 set of descriptions. These descriptions focus on the rooms and their decor content in the floor plan images. Also they focus on their relative positioning in respective rooms, and relative position of each room within the floor plan image. These descriptions vary in the sequence of the information given, the details of information provided, the sentence conjunctions and the vocabulary used for components in floor plan image. Figure 3 shows one of the floor plan and its corresponding descriptions. It can be observed that different descriptions of the same image vary in the amount of information provided, the sequence of describing each room, the names for each decor item could be different for different user. Also, a room may have variations in its name for different user. The length of the descriptions provided for each image is also varied. After the descriptions were collected, each set of 4 description were tagged to their respective images using image identifiers. Following were the instructions given to users:

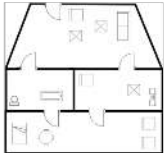
- Write information about the whole plan (number of rooms, name of the rooms).
- Write information of individual rooms- name, decor contained, relative size (small, medium, large, etc).

- Write information about the relative positioning of decors (north, east, west, south, aligned with wall, adjacent with something, etc).
- Write information about the relative position of each rooms (north, east, west, south, adjacent with bathroom, north of kitchen, etc).
- The description can have this information but they are not restricted to this. You can choose your own words and language.

Figure. 4 shows the user interface used to collect the descriptions. For 510 images in ROBIN datasets, there were 2040 descriptions collected in total. The dataset is tokenized and pre-processed for further processing. In our experiments we have compared machine generated description with these descriptions and an analysis regarding closeness of machine translations with human written descriptions is given. Next, we describe the steps in our framework in details.

V. SEMANTIC SEGMENTATION AND ROOM CLASSIFICATION

In all the previous approaches available in the literature, rooms have been classified by recognizing the textual label present in the floor plan image using Optical Character Recognition (OCR) techniques. Room classification in floor plans is not done by extracting salient feature from it. Room classification on the basis of their functionality is very useful in building information modelling (BIM). When a person enters a room in a house, he or she tells the functionality (class) of the room by looking at the decor items present inside the room. This inspired us to propose a unique feature for room classification. We have proposed a new



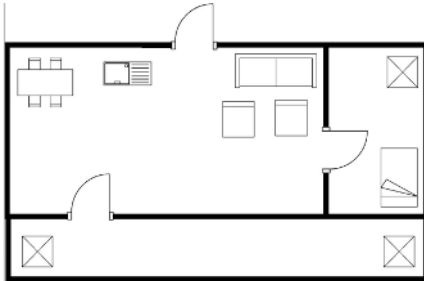
Description 1
The floor plan has 4 rooms. Rooms are entry, bathroom, bedroom and kitchen. Entry is in the north and has two tables in center, one small sofa in north and one large sofa in east corner. Bathroom is in east direction and has a tub and a sink in south west side. Kitchen is in west direction and has a sink, table in west side and sofa in east side. Bedroom is in south direction and has a bed, a table in west corner and two small sofa in the east corner. Entry is the largest room in floor plan, while bedroom being the second largest. Bathroom and kitchen are medium sized rooms.

Description 2
The entry door in the floor plan leads to the drawing room which is the largest room in the house. It contains a big sofa adjacent to the east wall and two square tables and a small sofa around it. This room leads to two rooms, out of which one is a bathroom containing a tub and a sink and the other is a kitchen containing a big sink to the right, a table and sofa to the left of the sink. The bathroom and kitchen give way to one more room that is the bedroom containing a bed and a round table adjacent to the west wall and two sofas near to the right wall. The shape of all the rooms is rectangular except the first room which resembles a parallelogram with two sides not parallel to each other.

Description 3
The house is has 4 rooms. Entrance is a big hall having 4 decors items. There is a big sofa to east wall, small sofa, and two tables in center. Neighboring rooms are bathroom and kitchen. Kitchen has a sink to east wall, a table in center and a small sofa adjacent to west wall. Bathroom has a tub and a large sink which are adjacent to south wall. Bedroom has a bed aligned to west wall and a round table in its side. While there are two small sofa aligned to east wall.

Description 4
In the given plan there is a big size hall located at west direction. It has main entry to the house. The hall contains 2 coffee table located in the middle, 1 small sofa beside the main entry door and one large sofa in the east corner wall. There are two other doors in the hall which are connected to the bathroom and one living room. Both bathroom and living room are besides of each other. Size of the bathroom is small and it is situated in the north direction. The bathroom contains one tub and one large sink situated in the opposite corners of the bathroom. The living room is situated on the east side beside the bathroom. The living room contains one small sofa and one coffee table and one twin sink. Both bathroom and living room have one more entry door to a big size bedroom. This bedroom is situated in the south direction. The bedroom contains one bed and one roundtable on the north corner wall and 2 small sofas on the east corner wall.

Fig. 3. Example of a sample floor plan image from ROBIN dataset and the annotation collected for the same to synthesize A-ROBIN



Write description for this image and mention image number.

Your answer

Fig. 4. User Interface used for data collection

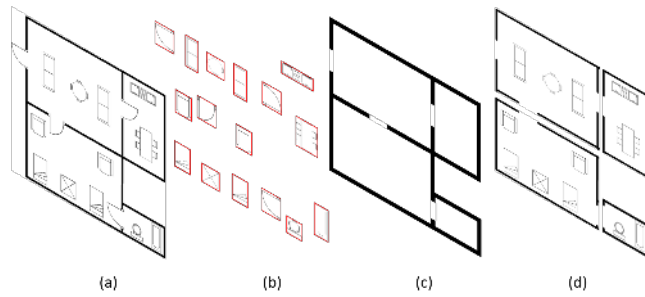


Fig. 5. Room segmentation and room partitioning process

feature called Local Orientation and Frequency Descriptor (LOFD), which represents the frequency of decors present in a room and their normalized distance from the center of the room. We proposed room classification approach as a 5 class classification problem, which annotates each room in a floor plan into one of the 5 classes namely, BEDROOM (label-1), BATHROOM (label-2), ENTRY (label-3), KITCHEN (label-4), HALL (label-5). The following subsections describes the details of room label learning and classification.

A. Room Segmentation

We have adopted the technique proposed in [14] for the identification of rooms. Walls are detected by performing morphological closing on the input floor plan image \mathcal{I} (see Fig. 5(c)). To delineate room boundaries, we detect doors using scale invariant features and close the gaps in wall image corresponding to the door locations. To obtain the rooms, we identified the connected components in the wall image by applying flood fill technique. The obtained connected components are the required rooms and their locations are obtained. Also, we calculate the areas of the respective rooms (polygon area), converted them into square feet (taking 100 pixels= 1 feet) and store all the information obtained, that is neighborhood, room area, room location coordinates, in a separate data structure.

B. Floor Plan Partitioning

A floor plan image is partitioned into rooms using the room coordinates extracted from the previous steps as shown in Fig. 5(d). These individual room images are the samples taken for training the room annotations. We have applied decor characterization in further stages on each of these individual room images to extract the features.

C. Decor Classification

In this section we describe the procedure employed for decor characterization and their classification. Figure 6 shows the 12 decor symbols used in the dataset [29]. We have improved the technique of decor characterization proposed in [14] by applying sequence of morphological operations. The technique in [14] uses a normalized area ratio of largest three components of a decor symbol for classification and characterization of decors. We have improved the technique by first collecting 10 different signatures for each symbol, taking an mean over them (symbols with different orientations) and stored them in a signature library. During classification, we first pre-process the symbol by applying a sequence of morphological operations (erosion and dilation), so that the symbol do not have broken lines. Then we applied blob detection over the image and cropped each decor symbol for signature comparison. Now we compare the test image’s signature with the signature stored in library and closest one is classified in its respective category. This modification in the technique greatly improved the classification accuracy for some symbols. Figure. 5 (b) depicts the detection of symbols in the floor plan input image Fig. 5 (a) with bounding boxes. These decors are classified in their respective categories shown in Fig. 6.

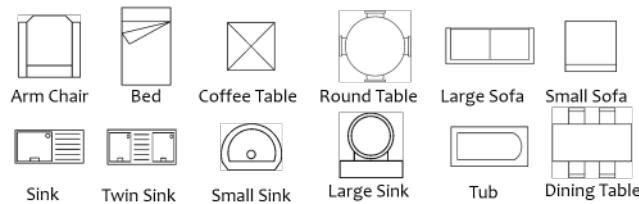


Fig. 6. Twelve classes of Decor models used in the experiments.

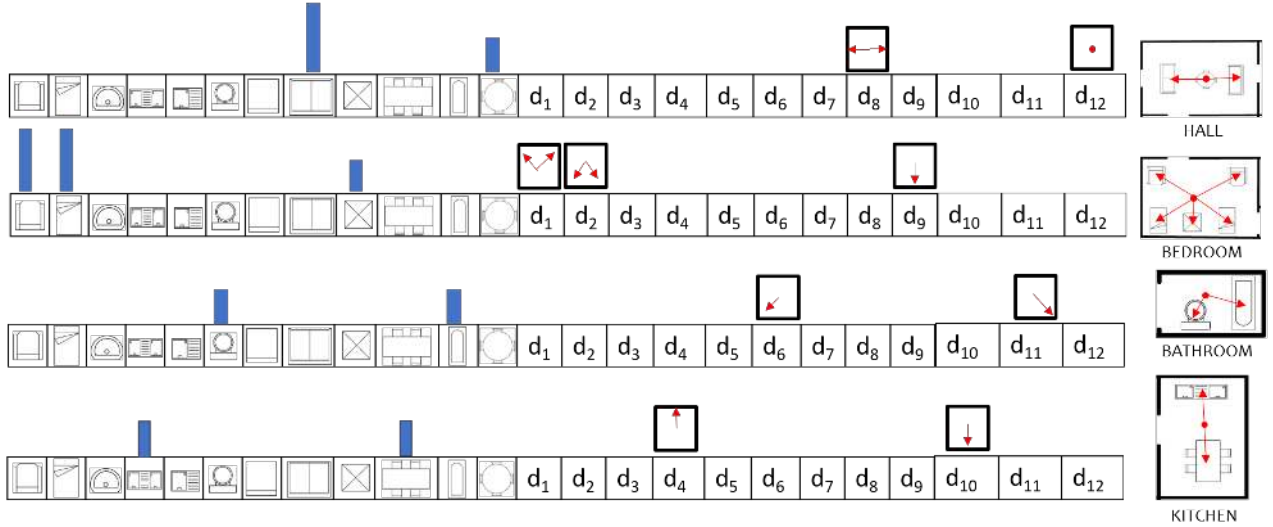


Fig. 7. Example of LOFD descriptor for four sample room images



Fig. 8. Output of room classification framework using LOFD

D. Local Orientation and Frequency descriptor (LOFD)

Once the decors inside a room are recognized, we compute the feature to classify a room. For room classification, we proposed a new feature named Local Orientation and Frequency descriptor (LOFD). The LOFD feature is required since no other feature descriptor for example SIFT [61], LBP [24] or SURF[21] could capture the room images clearly. The LOFD feature is a 1×24 vector containing the decor information of a room sample and their locally aggregated spatial information. Figure. 7 shows the LOFD feature matrix for the sample floor plan image. In LOFD, we have aggregated local information of room image in a vector form. LOFD is compact representation of frequency of the decor items and normalized distance of their centers from center of the room. The first 12 cells of the vectors are occupied by the 12 decor items from $D = \{D_1, D_2, \dots, D_{12}\}$ and next 12 cells are occupied by their normalized distances as $\mathcal{D} = \{d_1, d_2, \dots, d_{12}\}$ where d_n is the distance of each decor item from the center of the room. Here, d_n is calculated as:

$$d_n = \frac{\sum_{i=1}^k \text{dist}(\mathcal{R}_c, \mathcal{D}_c)}{\max(\mathcal{D}_n)} \quad (1)$$

Here, d_n is the normalized distance for each decor item, i is the count of each decor which may go up to k , which is the maximum number of a that decor item in the room, dist is the Manhattan distance between room center \mathcal{R}_c and decor center \mathcal{D}_c , $\max(\mathcal{D}_n)$ is the maximum of all the distances obtained for all the decors to normalize the distance value. Hence LOFD feature distinguishes each room uniquely by the frequency of each decor item and their spatial location in the room.

Since there are 12 decor models as shown in Fig. 6, first 12 elements of LOFD represents count of one decor item. However it is not necessary for any room to have all types of decor present, therefore LOFD is sparse in nature. In Fig. 7, depicts the room image followed by the corresponding LOFD feature vector. The colored bar over each cell represents the frequency count

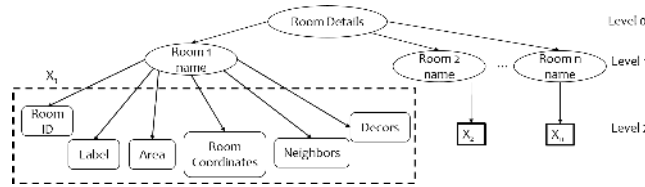


Fig. 9. Structure of the XML file generated for description synthesis

for each decor item, while arrows in red represents their relative spatial location in the room. In the next section training of the classifier using LOFD feature for room classification is explained.

E. Room annotations Learning and Room classification

Room annotations for training samples (divided the 1355 room images into 70% and 30% for training and testing respectively) are learned by LOFD feature and classifier. For training purpose, we have manually annotated the room samples, and used those annotations during training. Extensive experiments were performed using various classifiers and the best classifier in term of highest training accuracy is taken for testing the model. For testing purpose, an image from test set is taken and class labels are evaluated accordingly for the room samples of that floor plan image. For each new test floor plan image, feature vector is evaluated for every room. Therefore dimension of feature matrix for a test floor plan image will be $N_r \times 24$ where N_r is the number of rooms in the floor plan. Trained classifier is used for this feature matrix and output class labels are evaluated. Figure 8 depicts the annotations obtained for each room in a floor plan image, where different colors signifies different rooms and their respective annotations. Thus for a given floor plan we obtain room names and the decors within the rooms.

VI. DESCRIPTION SYNTHESIS

Rooms are classified and their annotations are learned in Sec.V-E. Information extracted from room segmentation are combined and used for generating the description of the floor plan image. Information related to individual rooms are combined and stored in an XML file, which is parsed to generate description of the floor plan.

A. XML File generation

An XML file has many benefit in terms of cross platform portability, ease of understanding by novices, and extendability. We have created an XML file by combining the semantic information extracted from room segmentation and room annotations learned in previous steps. As shown in Fig. 9, the tree like structure of XML file contains “Room details” as root node at level 0, “Room names” as nodes at level 1, and information of rooms as nodes at level 2 (leaf nodes), which are Room ID, Room annotations, Room area, Room Coordinates, Room neighbors and Room Decors. Apart from room annotations, a room ID is given to each room since room annotations can be same for two rooms. to generate a description.

B. Coordinates systems

For defining the positions of rooms and decors present in the floor plan, we have defined two coordinate systems. The global coordinate system is to identify the global location of rooms with respect to the entire document. Local coordinate system is to define relative position of decors with respect to each room.

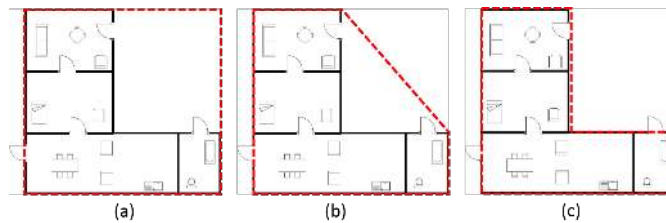


Fig. 10. Boundary tracing example for different values of t

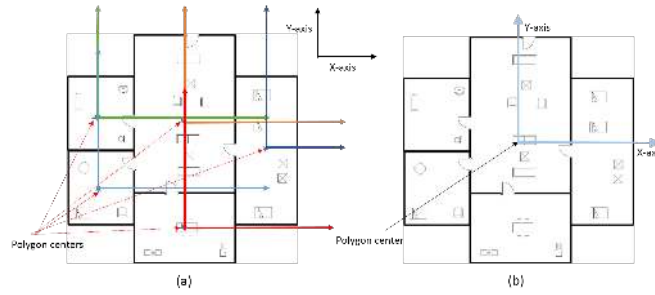


Fig. 11. Global and local Coordinate systems

1) *Boundary Tracing*: The origin of the global coordinate system is the center of the polygon, which makes the boundary of the floor plan. To identify the center of the floor plan, its boundary is traced. In order to trace the boundary, coordinates which compose individual rooms are plotted and an outer boundary is tracked which encloses all the outer points, since these points collectively makes the floor plan image. However, by tuning the value of shrinkage factor t between 0 and 1 we can switch between a convex hull of those points and a more close knit boundary. Shrinkage factor defines how closely the hull envelops the boundary points. For example, in Fig. 10 (a), the boundary traced is a convex hull for the floor plan image for the shrinking factor value $t = 0$, Fig. 10 (b) is the traced boundary for shrinking factor value $t = 0.5$ and Fig. 10 (c) is the close knit boundary for $t = 0.8$. Hence by tuning the shrinking factor value we can obtain a close knit boundary for the floor plan image.

2) *Global and local coordinate systems*: A global coordinate system defines the global position of all the rooms in a floor plan image (see Fig. 11 (b)). From the traced boundary obtained in the previous step, we calculate the origin of the global coordinate system. Equation 2 and 3 lists the governing equations.

$$a_i = x_i y_{i+1} - x_{i+1} y_i \quad (2)$$

$$A = \frac{1}{2} \sum_1^n a_i$$

Where, a_i in Eq. 2 is twice the signed area of the elementary triangle formed by (x_i, y_i) and (x_{i+1}, y_{i+1}) and the origin. A in Eq. 2 is the area of the polygon.

$$x_c = \frac{1}{6A} \sum_1^n a_i (x_i + x_{i+1}) \quad (3)$$

$$y_c = \frac{1}{6A} \sum_1^n a_i (y_i + y_{i+1})$$

In Eqn. 3, (x_c, y_c) is the center of the polygon. The local coordinate system (see Fig. 11 (a)) identifies the relative positions of all decors with respect to each room. Center of each room, for a local coordinate system is computed using Eq. 2 and 3.

C. Binning

We have performed global and local binning or radial partitioning of the floor plan (see Fig. 12). The non uniform binning angles were empirically determined. For identifying the direction of a decor, the center of the surrounding bounding box is taken as the reference point. While for the rooms, their respective centers, obtained in the previous steps is taken as reference point. As shown in the Fig. 12(a), the entire coordinate system is divided into 8 directions, north, north-east, east, south-east, south, south-west, west, north-west, in the clockwise direction. The binning depicted in Fig. 12(a) is a non uniform binning, while in Fig. 12(b) is a uniform binning.

The rationale behind non-uniform binning is to provide a more realistic direction information for rooms and decors. The idea of taking the direction from the center of the surrounding polygon may misguide the framework about the actual position of a room. E.g., if a room location in the west direction and stretched towards north, its center will lie in north west direction even if the room is in west. In order to avoid these kind of ambiguities, binning is done non uniformly and the angles are empirically taken. Figure 13 highlights examples for the above rationale. The highlighted room (Fig: 13 (a)) is more toward east direction, however it is also extended towards south. With non uniform binning we try to increase the span of east direction, shown purple line and arc where $(\theta_1 + \theta_2)$ is the angle of non uniform binning. While red line and arc shows the span of uniform binning

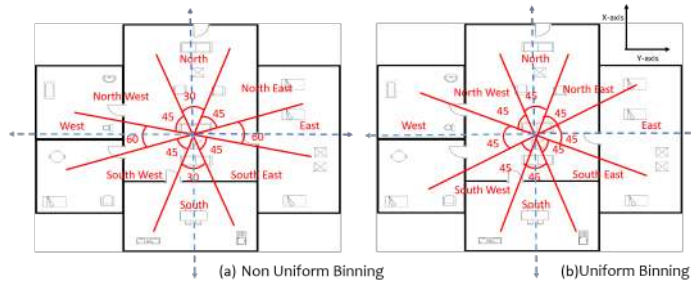


Fig. 12. Non uniform and uniform binning

which makes the room fall in south east direction and create ambiguity. Here θ_1 is the angle of uniform binning, C_1 and C_2 are the centers for floor plan and room respectively.

D. Navigation

Navigation in the indoor environment by avoiding the obstacles is an integral part of SUGAMAN. We have proposed a grammar based model that yields navigational directives to navigate through the house for a natural movement, from one door to the other door of each room. The algorithm is divided into two parts, first we create a data structure, which stores the room labels along with their respective doors and their corresponding index. The room information and the door coordinates are obtained from semantic segmentation in the earlier stages (see Sec. V-A). If a door is shared between two rooms then that door will be present in both room’s door structure and the index will represent the door’s identity. Next, we identify the entry room and the corresponding door, and generate a Depth First Search (DFS) ordering of the region adjacency graph of the floor plan taking the entry room as the start node. After that, a path to the next room is generated avoiding obstacles, by checking the visibility from first door to the other. We also create a door based adjacency matrix (AM_D), which stores the shared doors between rooms.

1) *Creating door structure*: The room coordinates, room labels and door coordinates are obtained in the semantic segmentation. After that an index i_d is assigned to each door. We have checked whether a given door i_d belongs to a particular room or not. We have performed an inside-outside test between the bounding polygons of the doors and the rooms to achieve the belongingness. The door structure contains each room with its corresponding doors having marked with their index i_d . As shown in Fig. 15(b), room and door information is stored in a door structure.

2) *Path Finding*: DFS search is performed over the region adjacency graph of the floor plan image taking the entry room as starting node. The door connected to the outer wall of the entry room is considered as the entry door and stored in the door structure. Here, entry door for the house is detected by the algorithm discussed in [18]. Algorithm. 1 describes the process of room to room navigation by obstacle avoidance. The route in each room is stored in the form of coordinates of movement and included in the description for narration of the path. Algorithm 1 traverse the rooms starting from the first room in the DFS graph, by checking if there is a door shared between them. This is checked by door based adjacency matrix (AM_D). If they do not share a door, the algorithm backtrack and explore other rooms. Also, it determines the route across the rooms for navigation. In Alg. 1, line 2 declares the flag, if the algorithm has to enter into backtracking. Line 1 describes the loop which traverse room to room finding the path. Line 5, algorithm checks if there is a shared door between the current room and the next room and continues traversal between rooms if there is a shared door. Line 6 directs the algorithm to further processing

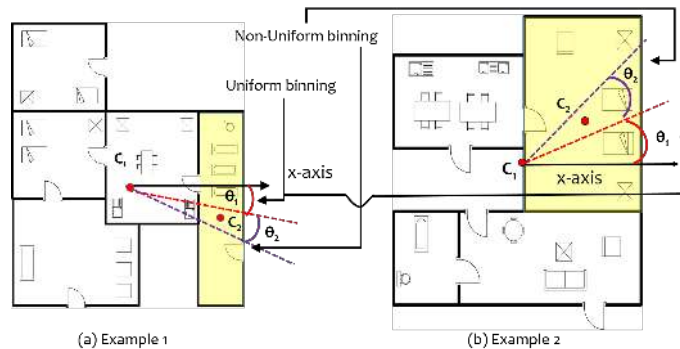


Fig. 13. Illustration for the rationale of non-uniform binning.

Algorithm 1 Room to Room traversal within a floor plan

```

1: for  $i \leftarrow 1, N_r - 1$  do                                ▷  $N_r$ : No of rooms
2:    $Backtrack \leftarrow 0$ 
3:    $c_r \leftarrow i$                                         ▷  $c_r$ : Current room
4:    $n_r \leftarrow i + 1$                                   ▷  $n_r$ : Next room
5:   if  $AM_D(c_r, n_r) == 1$  then
6:     if  $Backtrack \neq 1$  then
7:        $DC \leftarrow \text{CLASSIFYDECOR}(R_i)$                 ▷  $R_i$ : Room Image
8:        $V_L \leftarrow \{DC\}$                                ▷  $V_L$ : Vertex list
9:        $\mathcal{D} \leftarrow \text{DETECTDOORCENTROID}(R_i)$ 
10:       $V_L \leftarrow V_L \cup \mathcal{D}$ 
11:       $R_i^{new} \leftarrow \text{REMOVEDOORS}(R_i)$ 
12:       $\mathcal{B} \leftarrow \text{Blobs}(R_i^{new})$ 
13:       $Corners \leftarrow \text{HARRISCORNER}(\mathcal{B})$ 
14:       $C_S \leftarrow \text{STRONGEST}(Corners)$ 
15:       $V_L \leftarrow V_L \cup C_S$ 
16:      for  $j \leftarrow 1, N_{V_L}$  do                          ▷  $N_{V_L}$ : No. of elements in  $V_L$ 
17:        for  $k \leftarrow 1, N_{V_L}$  do
18:           $visible \leftarrow \text{VISIBLE}(V_L(j), V_L(k), R_i^{new})$ 
19:          if  $visible$  then
20:             $AM_N^i(j, k) \leftarrow \text{ED}(V_L(j), V_L(k))$ 
21:          else
22:             $AM_N^i(j, k) \leftarrow 0$ 
23:          end if
24:        end for
25:      end for
26:       $D_E \leftarrow R_{c_r}(\text{Entry})$ 
27:       $D_X \leftarrow R_{n_r}(\text{Entry})$ 
28:    end if
29:     $Path(i) \leftarrow \text{DISJKSTRA}(AM_N^i, D_E, D_X)$ 
30:  else
31:     $Backtrack \leftarrow 1$ 
32:     $D_E \leftarrow R_{c_r}(\text{Exit})$ 
33:     $c_r \leftarrow c_r - 1$ 
34:    goto Step 5
35:  end if
36: end for
37: return  $Path$                                           ▷ Path to go to every room from the entry

```

Fig. 14. Algorithm for the navigation

if backtracking is not required. Line 7 to 10, detects the coordinates of bounding box of decor items and centroid of doors of current room and include in a vertex list. In 11, doors are removed from room image because they are not required for avoiding obstacles. Line 12 to 15 detect the corner points in the room image using Harris corner detector after detecting the blobs, and include maximum 1000 strongest corners in the vertex list. Line 16 to line 22 describes the construction of adjacency matrix for navigation (AM_N). It checks the visibility between every point in the vertex list and include the Euclidean distance between them in AM_N as the weight at $AM_N(V_L(j), V_L(k))$. Visibility between two points is checked by filling the line between equal intervals in those points and checking if there is a black pixel present. If there is a black pixel present, then there must be an obstacle between those two points and hence those points are not visible. $AM_N(V_L(j), V_L(k))$ will have a 0 in that case. Line 26 and 27 defined the entry (D_E) and exit (D_X) door for current traversal, where entry door is the entry of current room and exit door is the entry of the next room. Line 29 evaluates a route (P^i) for current traversal by applying Disjkstra's shortest path algorithm over AM_N taking D_E and D_X as start and end nodes. Line 31 to 34 defines the backtracking process if there is no shared door found between current room and next room. Algorithm will backtrack in the DFS path and find the navigation path between corresponding rooms. The route for i^{th} room (P^i) is a set of coordinates, which contains the start point, end point and intermediate turns which a person have to make for obstacle avoidance. Figure. 15 describes the entire process for the input image Fig. 15(a). The checker box (inset) depicts AM_D , where the dark box represents a 0 and a white box represents a 1.

Figure. 15(b) shows the door structure created in the previous step and the order of traversal with backtrack step, Fig. 15(c) shows the DFS search graph generated over the region adjacency matrix to obtain the order to traverse the each room and Fig. 15(d) shows the local coordinate system fitted over every point in a route while traversing through the floor plan, also showing the direction of movement by arrows. Figure. 16(a) represents the door to door path generated for navigation, avoiding obstacles in each room for the input image. Figure. 16 shows some other examples describing the path generated on various floor plan images.

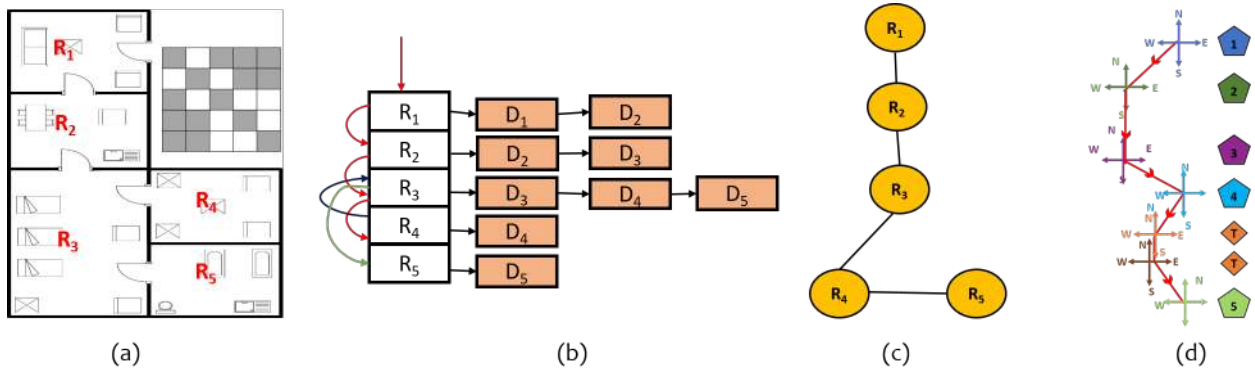


Fig. 15. Illustration of path detection process

E. Proximity based sentence model

Parsing of the XML file yields 5 types of information for each room, defined in separate sentences; Room name, area, neighbouring rooms, global position and contained decors with their relative position in room. For that purpose we defined, sentence model having 6 rules, which is based on proximity as shown in Tab. I. The first sentence (S_1) of description for every floor plan is a general sentence stating the number of rooms (N_r) in the floor plan. In S_2 , DT is a determiner which takes its value from the set {a, an}. Also, O_i is the object which takes its value from level 1 nodes (Room names), where value i varies from 1 to N_r . In S_3 , AREA takes its value from the RoomArea tag when XML file is parsed. In S_4 , s takes its value from the set {s, ϕ }, which is a proximity based value depends upon its previous word. Value s is chosen if the word in proximity (room) is a plural and ϕ otherwise. Also, AUX is an auxiliary verb, which takes its value from {is, are}, depending upon its proximity word and NR_j takes its value from Neighbors tag (neighboring rooms), when XML file is parsed. Here, value of j varies from 1 to NN_r which is number of neighboring rooms. In S_5 , LOC is the global position of room which takes its value from the set {North, North East, East, South East, South, South West, West, North West} described by binning. In S_6 , the value of k varies from 1 to DC i.e. decor count. Here, C is the count of individual decor item, D takes its value from the Decor tag in XML file, s takes its value from {s, ϕ } and $DLOC$ is the relative location of decor in the room which takes its value from {North, North East, East, South East, South, South West, West, North West} described by binning.

S_7 is the sentence narrating the navigation, where N_{step} is the number of steps to be taken. For calculating number of steps we took the euclidean distance between first coordinate and next coordinate in the route and calibrated the distance into steps (10 pixels= 1 step). Also, DIR is the direction in which the person has to move, for which local coordinate system is being fit on every coordinate of the route. It takes its values from the set {North, North East, East, South East, South, South West, West, North West}. The number of coordinates returned in the route of navigation inside a room, is the number of turns a person will

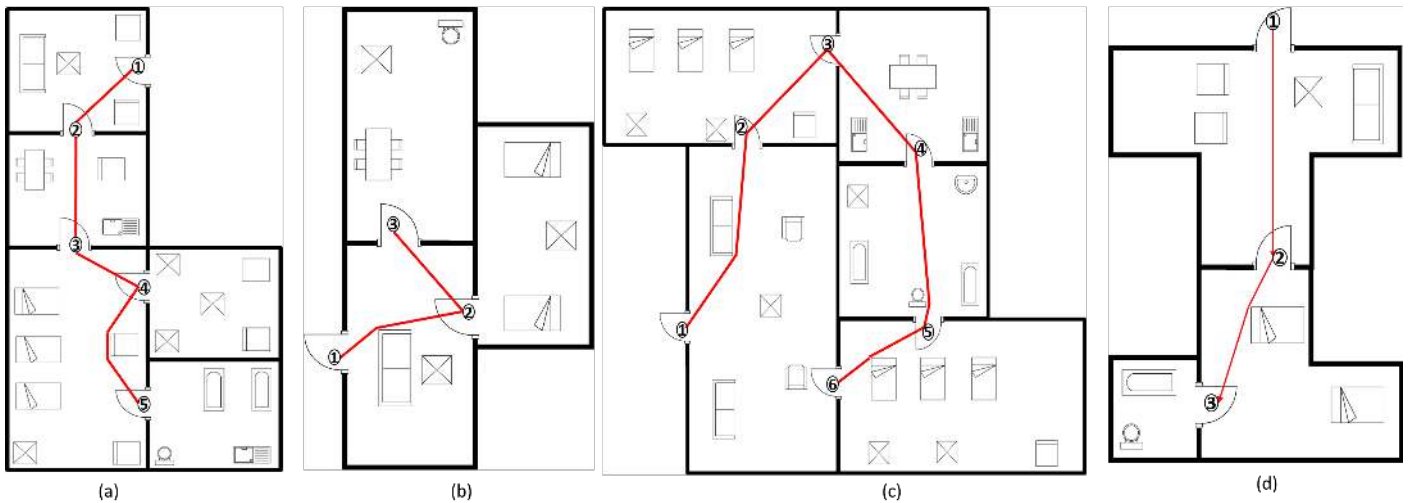


Fig. 16. Examples illustrating path detection by avoiding obstacles

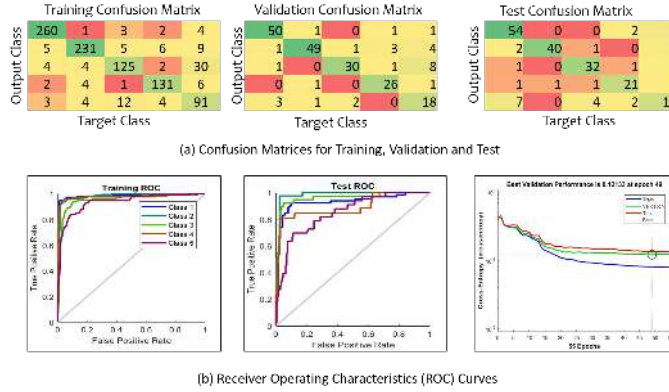


Fig. 17. Performance analysis of multi-layered perceptron.

have to make. Here N_m is the number of turns inside a room and N_r is the number of rooms. S_8 describes the door and room found after navigating through previous room. If the room has only one door and hence a dead end, the person will turn back and navigate further (S_9), else he will go straight and explore the other rooms by entering (S_7).

VII. ANALYSIS OF INTERMEDIATE STEPS

We have performed our experiments on a hardware platform with the following configurations. The system has an Intel core i7 (8^{th} generation), with a 1.87 GHz processor. It has a memory of 8 GB where, implementation has been done on Matlab 16a.

A. Room annotation learning and Classification

For this task a dataset of 1355 room images divided into 70% and 30% for training and testing respectively. LOFD features (see Sec. V-C and Sec. V-D) are used to train a multi-layered perceptron (1 hidden layer with 10 neurons). Performance of neural network is shown in Fig. 17 using confusion matrix and Receiver Operating Characteristics (ROC) curves. It is clear that ROC curve for class 5 moves maximum towards false positive axis, because of less number of training samples for class 5. Also, for class 1 and 4 it remains toward true positive axis due to more number of samples in training data. Figure. 17(b) (column 3) shows the performance curve for neural network in which best validation performance is achieved at epoch 49. The training, testing and validation accuracy obtained by neural network are 88.3%, 81.3% and 85.2%, respectively.

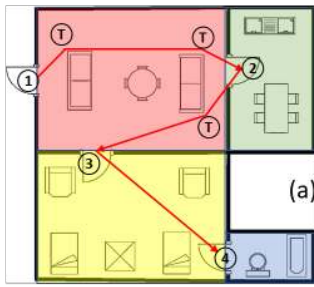
Experimental results on other supervised classifiers are shown in Tab. II. For training the classifiers, as shown in Tab. II, first we divided 1355 samples from ROBIN(R) dataset in 70% (training) and 30% (testing). Training and testing accuracy are shown

TABLE I. SENTENCE MODEL BASED ON PROXIMITY

Sentence	Rule
S_1	This floor plan has N_r rooms
S_2	There is DT O_i
S_3	It has an area of AREA
S_4	Its neighboring room{s} AUX NR_j
S_5	It is located in the LOC
S_6	This room has $\{C D\{s\} at the DLOC\}_k$
S_7	$\{\{Go N_{step} steps in DIR direction\}_{N_m}\}_{N_r}$
S_8	There is a door and a room. $\{S_9\}_{if\ dead\ end}$ $\{S_7\}_{else}$.
S_9	You have to turn back.

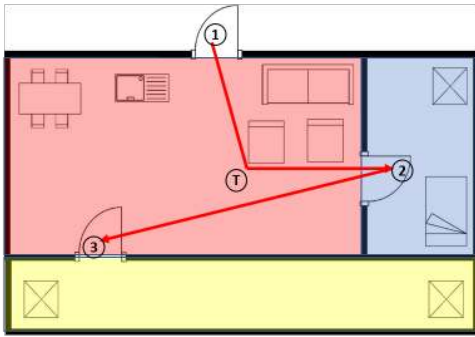
TABLE II. SVM CLASSIFIER-RESULTS OF ROOM ANNOTATION LEARNING BY SUPPORT VECTOR MACHINE

Variant	Training (R) (%)	Testing (R) (%)	Testing (S) (%)	Training (R+S) (%)	Testing (R+S) (%)
linear svm OVO	90.1	78	66.04	89.1	73.4
Quadratic SVM OVA	87.8	79	60	91.6	78.4
Cubic SVM OVA	87.8	77.11	63	91.2	76.6
Medium Gauss SVM OVO	88.3	76.67	58	90.5	75.3
Quadratic SVM OVO	87.1	77.11	60	90.9	74.2
Complex Tree	88.3	76.44	65.57	88.7	73



GD: In this architectural floor plan, there are 4 rooms. **There** is an **ENTRY**. It has an area of 39.5485 sq. ft. It has a door opening to the outside. Its neighboring rooms are **BEDROOM**, **KITCHEN**. It is located in the North side of the house. There is 1 large sofa at the West side of the room, 1 round table at the East side of the room, 1 large sofa at the East side of the room. **There** is a **BEDROOM**. It has an area of 34.64195 sq. ft. Its neighboring rooms are **ENTRY**, **BATHROOM**. It is located in the South West side of the house. There is 1 armchair at the West side of the room, 1 bed at the West side of the room, 1 coffee table at the South West side of the room, 1 bed at the South East side of the room, 1 armchair at the East side of the room. **There** is a **BATHROOM**. It has an area of 5.2914 sq. ft. Its neighboring room is **BEDROOM**. It is located in the South East side of the house. There is 1 large sink at the South West side of the room, 1 tub at the East side of the room. **There** is a **KITCHEN**. It has an area of 16.80245 sq. ft. Its neighboring room is **ENTRY**. It is located in the North East side of the house. There is 1 twin sink at the North side of the room, 1 dining table at the South side of the room.

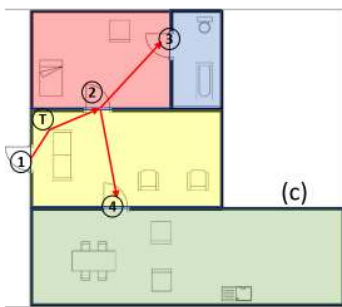
NV: Go 7 steps in North East Direction. Go 65 steps in East direction. Go 6 steps in South East Direction. There is a door and a room. You have to turn back. Go 9 steps in South West Direction. There is a door and a room. Go 70 steps in South East Direction. There is a door and a room.



GD: In this architecture floor plan, there are 3 rooms. **There** is an **ENTRY**. It has an area of 89.07 sq. units. It has a door opening to outside. Its neighbouring rooms are **BEDROOM** and **HALL**. It is located in the North side of the house. There is 1 dining table at the north east side of the room, 1 sink at the north side of the room, 1 large sofa at the north west side of the room, 2 small sofa at the west side of the room. **There** is a **BEDROOM**. It has an area of 42.76 sq. units. Its neighboring room is **ENTRY**. It is located in the west side of the house. **There** is 1 coffee table at the north side of the room, 1 bed at the south side of the room. **There** is a **HALL**. It has an area of 35.20 sq. units. Its neighboring room is **ENTRY**. It is located at the south side of the house. There is 1 coffee table at the west side of the room, 1 coffee table at east side of the room.

NV: Go 24 steps in South Direction. Go 27 steps in East Direction. There is a door and a room. You have to turn back. Go 51 steps in South West Direction. There is a door and a room.

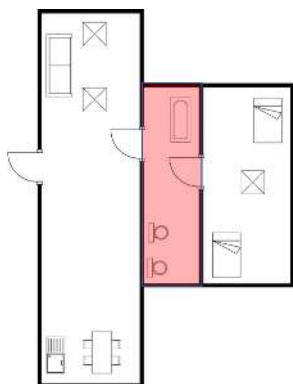
(b)



GD: In this architectural floor plan, there are 4 rooms. **There** is a **ENTRY**. It has an area of 49.66 sq. units. It has a door opening to outside of the house. Its neighbouring rooms are **BEDROOM** and **KITCHEN**. It is located in the west side of the house. There is 1 large sofa in the west side of the room, 1 armchair in the south of the room, 1 armchair in the south east of the room. **There** is a **BEDROOM**. It has an area of 37.45 sq. units. Its neighbouring rooms are **BATHROOM** and **ENTRY**. It is located in the north west side of the house. There is 1 bed in the south west side of the room, 1 small sofa in the north side of the room. **There** is a **BATHROOM**. It has an area of 20.34 sq. units. Its neighbouring room is **BEDROOM**. It is located in the north east side of the house. There is 1 tub in the south of the room, 1 large sink in the north of the room. **There** is a **KITCHEN**. It has an area of 81.22sq. units. Its neighbouring room is **ENTRY**. It is located in the south side of the house. There is 1 dining table in the west side of the room, 1 small sofa in the south side of the room, 1 small sofa in the north side of the room, 1 sink in the south west side of the room.

NV: Go 5 steps in North East Direction. Go 7 steps in North East Direction. There is a door and a room. Go 41 steps in North East Direction. There is a door and a room. You have to turn back. Go 38 steps in the South Direction. There is a door and a room.

Fig. 18. Generated descriptions for three floor plan images from A-ROBIN dataset.



GD: In this architecture floor plan, there are 3 rooms. **There** is an **ENTRY**. It has an area of 91.73 sq. units. It has a door opening to outside. Its neighbouring room is **KITCHEN**. It is located in the west side of the house. There is 1 dining table at the south side of the room, 1 sink at the south west side of the room, 1 large sofa at the north west side of the room, 2 coffee table at the north side of the room. **There is a KITCHEN**. It has an area of 32.16 sq. units. Its neighboring rooms are **ENTRY** and **BEDROOM**. It is located in the north side of the house. **There** is 1 large sofa at the north west side of the room, 2 sink at the south east of the room. **There** is a **BEDROOM**. It has an area of 35.20 sq. units. Its neighboring room is **KITCHEN**. It is located at the east side of the house. There is 1 coffee table at the west side of the room, 1 bed at north east side of the room, 1 bed at the south west side of the room.

Fig. 19. Generated descriptions for three floor plan images from A-ROBIN dataset.

TABLE III. PERFORMANCE ANALYSIS OF TEXT GENERATION ALGORITHM USING ROUGE SCORE.

ROUGE	Average Recall	Average Precision	F score
ROUGE-1	0.5061	0.2715	0.3445
ROUGE-2	0.1545	0.5707	0.07616
ROUGE-3	0.0535	0.01093	0.01483

in first and second column respectively. We tried testing the sampled from SESYD (denoted as S) dataset from this trained model but testing accuracy statistics (column 3) are not up to the mark, hence we mixed the samples from both datasets. Taking 500 samples from SESYD and 1355 samples of ROBIN making it a collective dataset of 1855 images, another models were trained and training and testing are shown in column 4 and column 5, respectively. The best performing classifier is linear Support Vector Machine (SVM), one verses one, for ROBIN dataset and quadratic SVM (one verses all) for mixed samples making LOFD a highly accurate feature descriptor for room annotation learning in floor plan images.

B. Description synthesis

All the Reference Corpus available in the A-ROBIN dataset and the generated descriptions were tokenism using The Penn Treebank tokenizer [20] and kept for evaluation purpose. We have compared the machine generated description of the floor plan with human written descriptions in A-ROBIN. The generated description is evaluated by three metrics, Recall-Oriented Understudy for Gisting Evaluation (ROUGE) [53], Bilingual Evaluation Understudy (BLEU) [54] and Metric for Evaluation of Translation with Explicit Ordering (METEOR) [52]. The textual description generated by our framework is then compared with the descriptions in A-ROBIN to evaluate their agreement with human written descriptions . Table III depicts the average recall, average precision and F score for ROUGE-1, ROUGE-2, ROUGE-3. As the value of n in n -gram comparison increasing, the ROUGE precision score decreases, which is also clear from Tab. III. Table. IV depicts the BLEU score and Tab. V METEOR score for the description generated, which demonstrates high correlation with human judgments.

1) *ROUGE*: ROUGE is a set of metrics designed to evaluate the text summaries. The generated summary can be evaluated with a set of reference summaries. In our work, we have compared the generated descriptions with available human written descriptions using n -gram ROUGE by the following equation.

$$\frac{\sum_{S \in \{RS\}} \sum_{gram-n \in S} Count_m(gram-n)}{\sum_{S \in \{RS\}} \sum_{gram-n \in S} Count(gram-n)} \tag{4}$$

Where RS stands for reference summaries, n stands for length of the n -gram, $gram-n$, and $Count_m(gram-n)$ is the maximum number of n -grams co-occurring in the candidate summary and the set of reference summaries. In Tab. III, comparison with three type of ROUGE- n is shown, ROUGE-1, ROUGE-2 and ROUGE-3. It can be seen that average recall is decreasing with increasing n -gram in ROUGE. The reason behind this behaviour is natural as ROUGE-1 compares on uni-gram basis in the candidate to reference corpus, which is word is word matching. ROUGE-2 compares on bi-gram basis, which is taking a set of two words at a time. However ROUGE-3 compares on tri-gram basis which is by considering 3 words at a time. Since ROUGE-1, ROUGE-2, and ROUGE-3 use uni-gram, bi-gram and tri-gram comparisons respectively, the decreasing nature of average precision is natural. Machine generated descriptions has a fixed pattern for words to be used and the information to be displayed. However, human written descriptions can have any sequence and use of words and phrases.

2) *BLEU*: BLEU metric analyses the co-occurrences of n -grams between a machine translation and human written sentence. The more the matches, the better is the candidate translation is. The score ranges from 0 to 1, where 0 is the worst score and 1 is the perfect match. In Tab. IV, we have given 4 types of BLEU score, for 4 values of n -gram. They first compute n -gram modified precision score (p_n) by following equation

$$p_n = \frac{\sum_{C \in \{Cand\}} \sum_{gram-n \in C} Count_{clip}(gram-n)}{\sum_{C' \in \{Cand\}} \sum_{gram-n' \in C'} Count(gram-n')} \tag{5}$$

Where, $Count_{clip}$ limits the number of times a n -gram to be considered in a candidate ($Cand$) string. Then they computer the geometric mean of the modified precision (p_n) using n -gram upto length N and weights W_n which sums up to 1. A brevity

TABLE IV. PERFORMANCE ANALYSIS OF TEXT GENERATION ALGORITHM USING BLEU SCORE.

METRIC	Score
BLEU-1	0.6418
BLEU-2	0.4673
BLEU-3	0.3448
BLEU-4	0.2103

TABLE V. PERFORMANCE ANALYSIS OF DESCRIPTION SYNTHESIS USING METEOR SCORE.

Average Recall	Average Precision	F1	F mean	Final Score
0.555	0.218	0.313	0.450	0.184

penalty(BP) is used for longer candidate summaries and for spurious words in it, which is defined by the following equation:

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{\frac{1-r}{c}}, & c \leq r \end{cases} \quad (6)$$

Where c is the length of candidate summary and r is the length of reference summary. Then BLEU score for corpus level given equal weights to all n-grams is evaluated by the following equation:

$$BLEU = BP \cdot \exp \sum_{i=1}^N W_n \log(p_n) \quad (7)$$

Here W_n is the equally distributed weight in n-grams. E.g. in case of BLEU-4 the weights used are $\{(0.25), (0.25), (0.25), (0.25)\}$. The proposed dataset perform well on BLEU score as shown in Tab. IV.

3) *METEOR*: METEOR is a metric used for evaluating machine generated summaries with human written summaries by checking the goodness of order of words in both. METEOR score is a combination of precision, recall and fragmentation (alignment) in the sentences. It is a harmonic mean of the uni-gram precision and uni-gram recall given an alignment, and calculated as:

$$PN = \frac{1}{2} \left(\frac{\text{no of chunks}}{\text{matched uni-grams}} \right) \quad (8)$$

$$METEOR = \frac{10PR}{R + 9P} (1 - PN) \quad (9)$$

Where PN is the penalty imposed on the basis of larger number of chunks, P is the uni-gram precision, R is the uni-gram recall and METEOR is the final score obtained by multiplying the harmonic mean of uni-gram precision and uni-gram recall with penalty imposed. Table. V shows the METEOR score we have obtained while experimenting on A-ROBIN dataset. It can be said that generated descriptions are very close to the human written descriptions. Also, it is clear that the descriptions collected in the A-ROBIN dataset are grammatically correct and close to the descriptions generated by the proximity based grammar model.

VIII. QUALITATIVE RESULTS

In this section we describe the qualitative results. These result shows the generated descriptions for samples from A-ROBIN dataset, along with the navigation information in the narrative form.

A. Examples of description Synthesis

In SUGAMAN, rooms are labelled into one of the 5 classes using the trained model (see Fig. 8). Room annotations and semantic information are stored in an XML file, which is parsed for description synthesis. Figure 19 presents the resultant description for 3 floor plan images. To facilitate the reader of the manuscript in understanding the description, we have made the following adjustments, (i) the first word of the first sentence about any room is in **bold** face, (ii) in the floor plan image, every room is highlighted with a different color and the same color is used to highlight the room name in the first sentence about the room, (iii) in the floor plan image, the turning points are marked with 'T' and sequence of traversal of doors are marked with their respective numbers. There are two types of description synthesized for a given floor plan. The first kind of description is named as General description (GD), which contains information like name, area, global position in the floor plan, relative position of decors, and neighboring rooms in terms of its accessibility by a door is described for each room in the final output description, along with a room having a door opening to outside of the house is also described. The other one is Navigation description (NV), which contains navigation information from room to room avoiding obstacles. If a room has only one door, it is a dead end. Hence the navigating person will turn back.

Figure. 19(a),(b),(c) are examples where the descriptions are successfully generated for the floor plan images. For example, in Fig. 19(a), the GD correctly describes the number of rooms, there connectivity, as well as count and the arrangements of the decors inside each room. On the other hand, the NV part of the description guides the user to navigate to each room, starting from the entry. It can be observed that starting from entry door (labelled as 1), the first obstacle to go to kitchen (as per the DFS navigation) is the sofa. Hence the user has to take a turn (marked as 'T') and then proceed. The directional information are obtained from the non-uniform binning technique discussed in Sec. VI-C. In Fig. 19(a), also the significance of "backtracking" can be understood. Once someone reaches the kitchen, then the next room to visit is the bedroom. Since there is no direct connection (as per Algo. 1, Line 5, $AM_D(c_r, n_r) \neq 1$). Thus the current room (c_r) is changed from kitchen to entry. However,

- [20] Marcus, Mitchell P and Marcinkiewicz, Mary Ann and Santorini, Beatrice *Building a large annotated corpus of English: The Penn Treebank*, Computational linguistics, 19(2), 313-330, (1993).
- [21] Bay, Herbert and Tuytelaars, Tinne and Van Gool, Luc *Surf: Speeded up robust features*, ECCV, 2006
- [22] Hodosh, Micah and Young, Peter and Hockenmaier, Julia *Framing image description as a ranking task: Data, models and evaluation metrics*, Journal of Artificial Intelligence Research, 47, 853-899, (2013).
- [23] Chen, Xinlei and Fang, Hao and Lin, Tsung-Yi and Vedantam, Ramakrishna and Gupta, Saurabh and Dollár, Piotr and Zitnick, C Lawrence, *Microsoft COCO captions: Data collection and evaluation server*, arXiv preprint arXiv:1504.00325, 2015.
- [24] Ojala, Timo and Pietikainen, Matti and Maenpaa, Topi *Multiresolution gray-scale and rotation invariant texture classification with local binary patterns*, IEEE T- PAMI, 24(7), 971-987, (2002).
- [25] Elliott, Desmond and Keller, Frank, *Image description using visual dependency representations*, EMNLP, 1292-1302, (2013).
- [26] Khan, Muhammad Usman Ghani and Gotoh, Yoshihiko, *Generating natural language tags for video information management*, MVA, 28(3), 243-265, (2017).
- [27] Dutta, Anjan and Lladós, Josep and Pal, Umapada, *A symbol spotting approach in graphical documents by hashing serialized graphs*, PR, 46(3), 752-768, (2013)
- [28] Chen, Jinying and Cao, Huaigu and Natarajan, Premkumar, *Integrating natural language processing with image document analysis: what we learned from two real-world applications*, IJDAR, 18(3), 235-247, (2015).
- [29] Sharma, D. and Gupta, N. and Chattopadhyay, C. and Mehta, S. *DANIEL: A Deep Architecture for Automatic Analysis and Retrieval of Building Floor Plans*, ICDAR, 2017.
- [30] Delalandre, Mathieu and Valveny, Ernest and Pridmore, Tony and Karatzas, Dimosthenis, *Generation of synthetic documents for performance evaluation of symbol recognition & spotting systems*, IJDAR, 13(3), 187-207, (2010).
- [31] de las Heras, Lluís-Pere and Terrades, Oriol Ramos and Robles, Sergi and Sánchez, Gemma, *CVC-FP and SGT: a new database for structural floor plan analysis and its groundtruthing tool*, IJDAR, 18(1), 15-30, 2015.
- [32] Mello, Carlos AB and Costa, Diogo C and dos Santos, TJ, *Automatic image segmentation of old topographic maps and floor plans*, SMC, 2012.
- [33] Zhu, Yukun and Kiros, Ryan and Zemel, Rich and Salakhutdinov, Ruslan and Urtasun, Raquel and Torralba, Antonio and Fidler, Sanja, *Aligning books and movies: Towards story-like visual explanations by watching movies and reading books*, ICCV, 2015.
- [34] Xu, Kelvin and Ba, Jimmy and Kiros, Ryan and Cho, Kyunghyun and Courville, Aaron and Salakhutdinov, Ruslan and Zemel, Rich and Bengio, Yoshua *Show, attend and tell: Neural image caption generation with visual attention*, ICML, 2015.
- [35] Verma, Yashaswi and Jawahar, CV, *Im2Text and Text2Im: Associating Images and Texts for Cross-Modal Retrieval.*, BMVC, 2014
- [36] Kuznetsova, Polina and Ordonez, Vicente and Berg, Alexander C and Berg, Tamara L and Choi, Yejin, *Collective generation of natural image descriptions*, ACL, 2012.
- [37] Karpathy, Andrej and Joulin, Armand and Fei Fei, Li, *Deep fragment embeddings for bidirectional image sentence mapping*, NIPS, 2014.
- [38] Farhadi, Ali and Hejrati, Mohsen and Sadeghi, Mohammad Amin and Young, Peter and Rashtchian, Cyrus and Hockenmaier, Julia and Forsyth, David, *Every picture tells a story: Generating sentences from image*, ECCV, 2010.
- [39] Elliott, Desmond and Keller, Frank, *Comparing automatic evaluation measures for image description*, ACL, 2014
- [40] Hu, Ming K., *Visual pattern recognition by moment invariants*, computer methods in image analysis, IRE trans. on Information Theory, 8, 1962.
- [41] Tombre, Karl, *Analysis of engineering drawings: State of the art and challenges*, GREC, 1998.
- [42] Dori, Dov and Velkovitch, Yelena *Segmentation and recognition of dimensioning text from engineering drawings*, CVIU 69(2), pp. 196-201, 1998
- [43] Lai, CP and Kasturi, R *Detection of dashed lines in engineering drawings and maps*, ICDAR, 1991
- [44] O’Gorman, Lawrence and Kasturi, Rangachar *Document image analysis*, IEEE Computer Society Press Los Alamitos, 1995.
- [45] Joseph, SH and Pridmore, Tony P. *Knowledge-directed interpretation of mechanical engineering drawings*, 14(9), pp. 928-940, 1992.
- [46] Kasturi, Rangachar and Bow, Sing T. and El-Masri, Wassim and Shah, Jayesh and Gattiker, James R. and Mokate, Umesh B., *A system for interpretation of line drawings*, IEEE T-PAMI, 12(10), pp. 978-992, 1990.
- [47] Trier, Oivind Due and Taxt, Torfinn and Jain, Anil K *Data capture from maps based on gray scale topographic analysis*, ICDAR, 1995
- [48] Deseilligny, Marc Pierrot and Le Men, Hervé and Stamon, Georges *Character string recognition on maps, a rotation-invariant recognition method*, PRL, 16(12), 1297-1310, 1995.
- [49] Yu, Yuhong and Samal, Ashok and Seth, Sharad *Isolating symbols from connection lines in a class of engineering drawings*, PR, 27(3), 391-404, 1994.
- [50] Mori, Shunji and Suen, Ching Y and Yamamoto, Kazuhiko *Historical review of OCR research and development*, Proceedings of IEEE, 80(7), 1029-1058, 1992
- [51] Chhabra, Atul K, *Graphic symbol recognition: An overview*, IWGR, pp. 68-79, 1997.
- [52] Denkowski, Michael and Lavie, Alon *Meteor 1.3: Automatic metric for reliable optimization and evaluation of machine translation systems*, WMT, 2011.
- [53] Lin, Chin-Yew *Rouge: A package for automatic evaluation of summaries*, ACL, 2004.
- [54] Papineni, Kishore and Roukos, Salim and Ward, Todd and Zhu, Wei-Jing *BLEU: a method for automatic evaluation of machine translation*, ACL, 2002
- [55] Bernardi, Raffaella and Cakici, Ruket and Elliott, Desmond and Erdem, Aykut and Erdem, Erkut and Ikizler-Cinbis, Nazli and Keller, Frank and Muscat, Adrian and Plank, Barbara *Automatic Description Generation from Images: A Survey of Models, Datasets, and Evaluation Measures*, JAIR, 55, 409-442, 2016.
- [56] Kulkarni, Girish and Premraj, Visruth and Dhar, Sagnik and Li, Siming and Choi, Yejin and Berg, Alexander C and Berg, Tamara L *Baby talk: Understanding and generating image descriptions*, CVPR, 2011.
- [57] Karlsen, Randi and Sundby, David and Nordbotten, Joan *Automatic generation of textual image collection descriptions*, ICWIMS, 26, 2013.
- [58] Karpathy, Andrej and Fei Fei, Li, *Deep visual-semantic alignments for generating image descriptions*, CVPR, 2015.

- [59] Kuznetsova, Polina and Ordonez, Vicente and Berg, Alexander C and Berg, Tamara L and Choi, Yejin *Collective generation of natural image descriptions*, ACL, 2012.
- [60] Vinyals, Oriol and Toshev, Alexander and Bengio, Samy and Erhan, Dumitru *Show and tell: A neural image caption generator*, CVPR, 2015.
- [61] Lowe, David G, *Distinctive image features from scale-invariant keypoints*, IJCV, 60(2), 91-110, 2004.
- [62] Joachims, Thorsten *Text categorization with support vector machines: Learning with many relevant features*, ECML, 1998.
- [63] Doermann, David Scott and Tombre, Karl and others, *Handbook of Document Image Processing and Recognition*, Springer, 2014.
- [64] Wang, Zhe and Xue, Xiangyang, *Multi-class support vector machine*, pp. 23-48, 2014.
- [65] Rusiñol, Marçal and Lladós, Josep, booktitle=Symbol Spotting in Digital Libraries, *State-of-the-Art in Symbol Spotting*, Springer, pp. 15-47, 2010
- [66] Liu, Lu and Lu, Xiaoqing and Li, Keqiang and Qu, Jingwei and Gao, Liangcai and Tang, Zhi, *Plane geometry figure retrieval with bag of shapes*, DAS, 2014.
- [67] Qureshi, Rashid Jalal and Ramel, Jean-Yves and Barret, Didier and Cardot, Hubert, *Spotting symbols in line drawing images using graph representations*, GREC, 2014.
- [68] Lienhart, Rainer and Maydt, Jochen, *An extended set of haar-like features for rapid object detection*, ICIP, 2002.
- [69] Viola, Paul and Jones, Michael *Rapid object detection using a boosted cascade of simple features*, CVPR, 2001.
- [70] Adam, Sébastien and Ogier, Jean-Marc and Cariou, Claude and Mullot, Rémy and Labiche, Jacques and Gardes, Joël, *Symbol and character recognition: application to engineering drawings*, IJDAR, 3(2), pp. 89-101, 2000.
- [71] Freeman, Herbert *Computer processing of line-drawing images*, ACM Computing Surveys, 6(1), 57-97, 1974.
- [72] Spitz, A Lawrence *Determination of the script and language content of document images*, IEEE T-PAMI, 19(3), pp. 235-245, 1997.