

# Similar Floor Plan Retrieval Featuring Multi-Task Learning of Layout Type Classification and Room Presence Prediction

Yuki Takada, Naoto Inoue, Toshihiko Yamasaki, Kiyoharu Aizawa  
The University of Tokyo  
Email: {takada,inoue,yamasaki,aizawa}@hal.t.u-tokyo.ac.jp

**Abstract**—In this paper, a new framework for real estate property searches is presented in which a floor plan image is used as a query. In similar property searches, appearance-based similar image retrieval does not work well because similar properties have totally different floor plan images. Therefore, a multi-task learning method using deep neural networks to solve this problem is presented. Convolutional Neural Networks (CNNs) are trained to solve the two tasks: layout type classification and room presence classification. The feature vectors obtained from the CNNs are then applied to the retrieval task. Experiments using 22,140 floor plan images in Tokyo, Japan were conducted, and the proposed method achieved the best performance (15.7% with precision@5) compared to other possible approaches.

## I. INTRODUCTION

In recent years, information technology has been applied to real estate businesses, which is called real-estate tech (ReTech). For example, ReTech was the theme for the second consecutive year at Marche International des Professionnels de l’Immobilier (MIPIM), the world’s largest real estate trade fair, held in March 2017 in France [1]. However, in similar property retrieval, searches using basic metadata as shown in Fig. 1 have been the only choices so far. The amount and quality of detailed metadata vary widely depending on the property, and it is costly to manually annotate metadata. Therefore, retrieval by another method independent of metadata is required.

The purpose of this research is to retrieve similar real estate properties using only floor plan images. [2] proposed a graph-based similar floor plan search system. Namely, all the floor plan images were represented as graph structures in advance and the similar property search was converted to a similar graph search problem. This approach also showed that labeling is possible with an inexpensive price level of 20 yen (US\$0.20) per floor plan image by using crowd working services. However, a large number of new properties are registered every day, and it is difficult to create a system to label all new floor plan images by crowd working services. In this paper, we present a method that can directly use the floor plan image as a query. Conventional CNNs do not work well because even similar floor plans have a wide variety of drawing styles, and therefore floor plan images can appear to be totally different. To solve this problem, we propose a multi-task learning framework where both a layout type classifier and a room presence classifier are jointly optimized. Then,

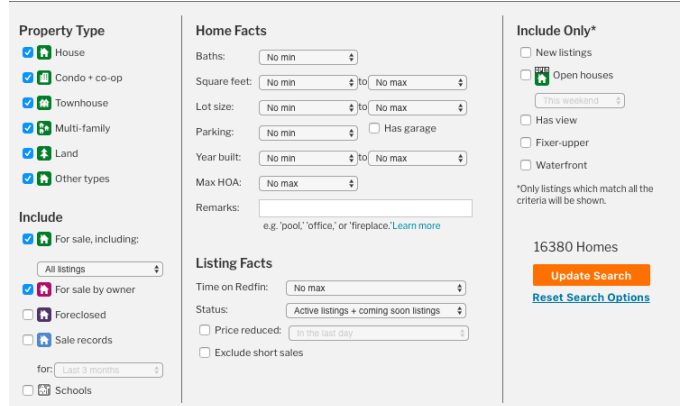
The image shows a screenshot of a real estate search interface. It is divided into several sections: 'Property Type' with checkboxes for House, Condo + co-op, Townhouse, Multi-family, Land, and Other types; 'Home Facts' with input fields for Baths, Square feet, Lot size, Parking, Year built, Max HOA, and Remarks; 'Include' with checkboxes for 'For sale, including' (All listings, For sale by owner, Foreclosed, Sale records, Last 3 months, Schools); 'Listing Facts' with input fields for Time on Redfin, Status, Price reduced, and Exclude short sales; and 'Include Only\*' with checkboxes for New listings, Open houses, This weekend, Has view, Fixer-upper, and Waterfront. At the bottom right, it says '16380 Homes' and has buttons for 'Update Search' and 'Reset Search Options'.

Fig. 1: Example of existing real estate search system [3]

the feature vectors taken from fc7 layer of VGG16 are used as visual features for the floor plan images. We conducted experiments using images of 22,140 apartments in Tokyo, Japan. The results showed that the proposed method achieved the best performance (15.7% with precision@5) among the possible baselines.

## II. RELATED WORKS

### A. Graph Analysis of Floor Plan Images

There are some related works that analyze the floor plans of real estate properties graphically. [4] divided floor plans into four groups depending on the size of the dwelling unit and sampled approximately the same number for floor plans of each group. They attached graph structures to 486 floor plan images obtained by sampling, analyzed graph structures by using the real asymmetry value, indicating how far away a node is from the center in the graph value. They pointed out that the “individual room grouping tendency,” in which all rooms such as the living-dining (LD) and private rooms are connected to a corridor, is becoming a more popular trend in real estate.

[5] constructed a dataset of graph structures consisting of eight node types and five edge types using 996 floor plan images of apartments of 3K (three rooms, kitchen), 3DK (three rooms, dining room, and kitchen), and 3LDK (three rooms, living room, dining room, and kitchen) in Kyoto city. They extracted subgraph structures that involve the age of the building and rent information by using Emergence

Patterns. They also conducted a regression analysis of rent and showed that a regression model considering the extracted graph structure yielded the best prediction performance. In the studies introduced above, the graph structures were extracted manually and therefore it is not realistic to extend the methods to all the floor plan images in large databases.

### B. Image Analysis of Floor Plan Images

[6] conducted the pairing of floor plan images and photographs that were taken inside an apartment using the Siamese Network [7]. They used HOME’s dataset [8] provided by LIFULL Co., Ltd. [9], which is published by the National Institute of Informatics [10]. They estimated which floor plan image corresponds to the photographs inside the apartment (bathroom, kitchen, or living) and obtained higher accuracy than humans attempting the same task.

[11] estimated which part of the floor plan corresponds to each frame of video taken in the real estate property by performing matching and localization of the floor plan. [12] divided the information of the floor plan to text information, outer wall information, and inner wall information using the thickness of each line and processed these for each of the three pieces of information. This succeeded in increasing the accuracy of room detection by 4%.

Image analysis using floor plans is emerging as summarized above, but this paper is the first to propose similar property retrieval using floor plan images to the best of our knowledge.

### III. PROPOSED METHOD

The purpose of this study is to retrieve floor plan images that are structurally similar to the query floor plan image. We apply CNNs for learning the metadata of floor plan images, and extract features from each floor plan image and construct feature vectors by using the deep features of this network. We retrieve feature vectors similar to the feature vectors of the query floor plan image using a k-nearest neighbor algorithm to produce a set of structurally similar images.

We divided the dataset into training images, query images, and search images. The network used for this paper is shown in Fig. 2. We trained this network by using training images. This network contains fully connected layers of  $n$  dimensions and fully connected layers (FC) of  $2m$  dimensions behind the fc7 layer of VGG-16 [13] pre-trained on ImageNet. FC layers of  $n$  dimensions are connected to learn the floor type, and fully connected layers of  $2m$  dimensions are connected to learn the presence of a room, where  $n$  is the number of layout types and  $m$  is the number of types of room. The parameters in VGG-16 are updated by feeding back both layout type classification errors and room presence classification errors. We use softmax cross entropy loss to calculate errors. We use layers of  $2m$  dimensions, instead of  $m$  dimensions because of the imbalanced data to be learned. It is possible to obtain better deep features by learning two types of metadata jointly in this case.

Similar floor plan images are retrieved as follows. We construct feature vectors of 4,096 dimensions by inputting

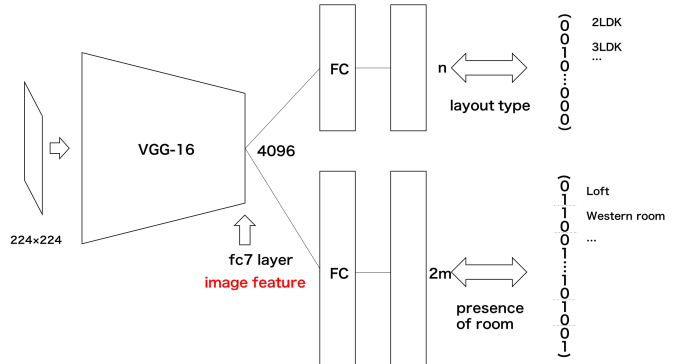


Fig. 2: Network architecture

Table I: Floor plan type

floor plan type	#	floor plan type	#
2LDK	5194	1LDK	758
2DK	5083	2SLDK	507
3LDK	2985	1DK	388
1K	2419	4LDK	214
2K	2065	3K	140
3DK	1112	3SLDK	121
one room	974	others	180

floor plan images into the network and extracting the feature vector from the fc7 layer. We obtain similar feature vectors from the database by using the nearest neighbor method using Euclidean distance for distance calculation between vectors.

## IV. EXPERIMENTS

### A. Dataset

We used a dataset of floor plan images created by [2]. This dataset contains floor plan images from two sources: SUUMO [14] and HOME’s dataset [8]. [2] annotated the graph structure to these floor plan images. The graph structure is used to calculate the similarity between two floor plan images. We excluded the floor plan images with no graph structure and the duplicate real estate properties (with the same rent, address, and floor space) from the dataset, and obtained a total of 22,140 images.

This dataset contains floor plan images of 14 floor plan types. Table I lists the 14 floor plan types. For example, 2SLDK means that there are two bedrooms in addition to store room (S), living room (L), dining room (D), and kitchen (K). Floor plan types with fewer than 100 floor plan images are grouped into others. There are 24 room types in the original graph structure method discussed above. Table II lists the 24 room types. We use only 14 room types (**bold** / underlined) in our experiments, and do not use CL, E, or DR because of unbalanced data. We do not use room types containing L, D, or K because we already learned these room types in the floor plan type classification.

### B. Implementation details

We use 12,140 images for training and 10,000 images for testing. We perform near-duplicate removal of data on 10,000 test floor plan images before the experiment. The dataset used

Table II: Room types in graph structure

node name	explanation	node name	explanation	node name	explanation
<b>Loft</b>	<b>loft</b>	<b>BR</b>	<b>bedroom</b>	DR	dress room
<b>WR</b>	<b>western room</b>	<b>UB</b>	<b>modular bathroom</b>	L	living
<b>Bal</b>	<b>balcony</b>	<b>Ba</b>	<b>bathroom</b>	D	dining
<b>UPDN</b>	<b>stairs</b>	<b>WC</b>	<b>toilet</b>	K	kitchen
<b>JR</b>	<b>japanese room</b>	<b>Hall</b>	<b>corridor</b>	DK	dining kitchen
<b>WIC</b>	<b>walk-in closet</b>	<b>PR</b>	<b>powder room</b>	LD	living-dining
<b>Ver</b>	<b>verandah</b>	CL	closet	LDK	living-dining kitchen
<b>R</b>	<b>room</b>	E	entrance	Other	others

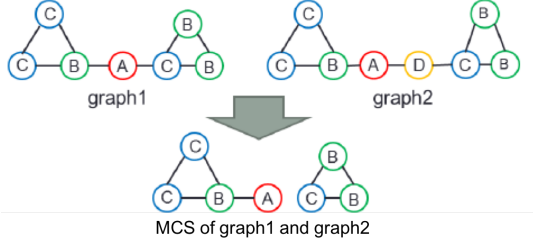


Fig. 3: Example of MCS [2]

in this study contains some images of the same floor plan, even though metadata such as address and rent are different. Therefore, we exclude any floor plan image that is identical to other images. As a result, 7,960 test images are obtained. These images are divided into 100 query images and 7,860 database images. We retrieve the top five images based on the proposed method for each query image to evaluate the retrieval.

The input to the network is an RGB image of resolution  $224 \times 224$ . We apply the Momentum SGD algorithm to train the model with batch size 16, 0.9 momentum, and 0.00001 weight decay. Learning rate is  $10^{-4}$  and total epochs are 90. The system is implemented in the open source deep learning framework called Chainer [15].

We define the ground truth similarity between two floor plan images as shown below. We use the graph structures of floor plan images and calculate the similarity between graph structures. The calculation of similarity between graph structures is based on the method of [2], outlined below. First, the maximum common subgraph (MCS) of the two graphs is calculated. Fig. 3 shows an example of an obtained MCS.

We calculated the similarity between two graphs using MCS as (1), where  $|q|, |g|$  are the sum of edges and nodes in the graph, respectively, and  $MCS(q, g)$  is the MCS of two graphs. The similarity is 1 when the two graphs are perfectly matched, and 0 when there are no common parts at all.

$$sim(q, g) = \frac{|MCS(q, g)|}{|q| + |g| - |MCS(q, g)|} \quad (1)$$

Precision@5 is used as the evaluation method for this experiment, which indicates the ratio of the correct answer data included in the top five search results. In this experiment, if we can retrieve a floor plan image whose structural similarity

is more than  $p$ , we define it as a correct answer.  $p$  is set to 0.5, 0.6, 0.7, 0.8, 0.9.

We compare the following six methods.

- Random: picking floor plan images randomly.
- Metadata: picking floor plan images whose floor plan type is the same as a query image.
- ImageNet: using a network pre-trained with ImageNet.
- Layout type: using a network learning only layout type.
- Presence of room: using a network learning only presence of room.
- Proposed method: using a network that learned the layout type and the presence of the room at the same time.

We also conduct the experiment to calculate the accuracy of layout type classification and room presence classification by using the network learning only one task. In this experiment, 20,140 images are used for training, and 2,000 images are used for testing.

### C. Classification results

Table III shows a confusion matrix of layout type classification. This shows the correct labels and outputted labels of 2,000 floor plan images used for testing. The accuracy indicates how probable the layout type is recognized as the correct type. This table shows that the accuracy tends to be higher as the number of images of a layout type increases, and suggests that layout type classes with few occurrences are not sufficiently learned. However, 4LDK has relatively high accuracy even though there are few floor plan images of 4LDK in our dataset. There is the possibility that the semantical difference between 4LDK and other layout types is large, so the network can easily classify 4LDK floor plan images.

On the other hand, 2DK and 2K, and 1K and 1 room, are relatively difficult to distinguish. This is because these two types of layout type are semantically similar.

Table IV shows the accuracy of recognition based on the presence or absence of each room for each room type. The meaning of each line is defined below.

- TP: the room actually exists and it is recognized as an existing room.
- FP: the room actually does not exist, but it is mistaken as an existing room.

Table III: Confusion matrix

		output														accuracy
		2LDK	2DK	3LDK	1K	2K	3DK	1R	1LDK	2SLDK	1DK	4LDK	3K	3SLDK	others	
truth	2LDK	435	32	16	1	4	1	0	6	1	0	0	0	0	1	0.88
	2DK	23	379	3	2	13	3	0	5	0	3	0	0	0	1	0.88
	3LDK	18	4	271	0	0	6	0	1	5	0	0	0	3	0	0.88
	1K	2	0	0	192	5	0	16	5	0	0	0	0	0	0	0.87
	2K	2	28	0	6	126	0	2	3	0	1	0	0	0	0	0.75
	3DK	5	7	8	1	1	82	0	0	1	0	0	0	0	0	0.78
	1R	0	0	0	24	1	0	42	1	0	0	0	0	0	0	0.62
	1LDK	11	4	0	13	2	0	1	37	0	3	0	0	0	0	0.52
	2SLDK	3	0	5	0	0	0	0	0	29	0	0	0	0	0	0.78
	1DK	2	1	0	11	4	0	1	7	0	11	0	0	0	0	0.30
	4LDK	0	0	3	0	0	0	0	0	0	0	9	0	0	3	0.60
	3K	0	3	0	0	2	2	0	0	0	0	0	5	0	1	0.38
	3SLDK	1	0	8	0	0	0	0	0	1	0	0	0	5	0	0.33
	others	1	3	3	0	1	0	0	2	0	0	1	0	0	3	0.21

- TN: the room actually does not exist and it is recognized as an absent room.
- FN: the room actually exists, but it is mistaken as an absent room.

Accuracy and f-measure are expressed as (2)(3)(4).

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad (3)$$

$$\text{f-measure} = \frac{2\text{Recall} \cdot \text{Precision}}{\text{Recall} + \text{Precision}}. \quad (4)$$

Accuracy is high as a whole, but the f-measure of “modular bathroom,” “walk-in closet,” and “room” is low. It seems that the modular bathroom is classified as a bathroom, and the walk-in closet is classified improperly as an ordinary closet. Moreover, “room” is considered to have less information than other room types, so it is difficult to recognize.

#### D. Retrieval results

Table V shows precision@5. The proposed method exhibits better performance than the other methods at all  $p$ , indicating that the proposed method is effective. The results for the presence of a room are better than those for the layout type, which indicates that the presence of rooms is considered to be more important for the structure of the floor plan when the proposed method is applied. The difference between the proposed method and other methods for the “presence of room” is largest by  $p = 0.9$ .

This can be explained for the following reasons. The network that learns only the presence of rooms cannot consider the connection between rooms and the layout of the room at all. Considering the definition of graph similarity, even if only room nodes are applied, we can still retrieve layout images that have a relatively high structural similarity to query images. In other words, if  $p$  becomes small, the possibility of retrieving the correct layout images increases without considering the layout type.

Fig. 4 and Fig. 5 show the retrieval results for several query images. Floor plan images that have a structural similarity higher than 0.8 to query images are surrounded by red flames.

Fig. 4a is an example where the proposed method performs well. A network that learns only the presence of a room finds that there is a Japanese-style room, a Western-style room, and a balcony in the query image, but does not determine that it is 2DK. On the other hand, the network also determines that the query image is 2DK using the proposed method. Fig. 4b is also an example where the proposed method performs well. A network that learns only the presence of a room does not determine that there are two Western-style rooms in the query image, and that it is 3LDK. On the other hand, our network does make these determinations. Fig. 5 is an example in which the proposed method (and all other methods) fails.

In the proposed method, 2DK images are retrieved even though the query image is 2LDK. This is because the difference between 2LDK and 2DK is small and difficult to detect. The DK node is in the center of the graph structure and the structural similarity becomes low if this node is selected by mistake. These examples show that it is possible to retrieve structurally similar floor plan images by learning both the layout type and the presence of rooms.

#### E. Subjective evaluation of retrieval results

Subjective evaluation was conducted under the following conditions. We show a query floor plan image and a retrieved floor plan image, obtained from the different methods, to subjects and ask if they think that the two floor plan images are similar. Ten subjects provide an answer for each pair presented. The Yahoo crowdsourcing tool [16] was used for this experiment. Table VI shows the subjective evaluation results. Values in the table indicate the rate of subjects who respond that two floor plan images are similar. From this table, we can observe the following findings. The proposed method is superior (from the subjective results) to the conventional method using metadata. This indicates the usefulness of the proposed method. However, the value of the method using graph similarity is also not very high. When converting a floor plan image into a graph structure, the information such as the positional relationship and direction of the rooms is lost. There is a possibility that such information is important to a user’s property similarity determination.

Table IV: Accuracy and f-measure of room type

room type	TP	FP	TN	FN	accuracy	f measure
loft	20	4	1964	12	0.99	0.71
western room	1538	56	328	78	0.93	0.96
balcony	1500	116	264	120	0.88	0.93
stairs	173	16	1798	13	0.99	0.92
japanese room	716	29	1217	38	0.97	0.96
walk-in closet	101	77	1764	58	0.93	0.6
verandah	131	67	1763	39	0.95	0.71
room	89	49	1793	69	0.94	0.6
bedroom	178	30	1754	38	0.97	0.84
modular bathroom	142	65	1561	232	0.85	0.49
bath room	1409	122	259	210	0.83	0.89
water closet	1768	123	84	25	0.93	0.96
corridor	1056	131	697	116	0.88	0.9
powder room	1194	79	527	200	0.86	0.9

Table V: Precision@5

$p$	random	metadata	ImageNet	layout type	presence of room	proposed method
0.5	0.084	0.240	0.294	0.400	0.470	<b>0.494</b>
0.6	0.038	0.140	0.224	0.294	0.368	<b>0.396</b>
0.7	0.022	0.065	0.151	0.202	0.259	<b>0.275</b>
0.8	0.013	0.032	0.112	0.151	0.174	<b>0.200</b>
0.9	0.008	0.019	0.107	0.112	0.120	<b>0.157</b>

Table VI: Subjective evaluation results

metadata	ImageNet	layout type	presence of room	proposed method	graph similarity
0.258	0.373	0.455	0.442	<b>0.469</b>	0.444

## V. CONCLUSION

In this study, we constructed a network that learns both the layout type and the presence of room types, and obtained feature vectors of floor plan images. We proposed a method to retrieve similar floor plan images using deep features and compared it with five methods such as a method using only metadata. As a result, the following results were obtained.

- In the classification of the floor plan, we demonstrated that learning of the layout type and the presence of the room were suitable to obtain feature vectors of floor plan images.
- We demonstrated that the proposed method, which learns the layout type and the presence of the room simultaneously, exhibits better classification performance than other methods. We were able to retrieve floor plan images that have a structural similarity higher than 0.9 at a precision of 15.7% with precision@5.

The limitations of this research are as follows. Only the node is used for the learning network, while the edge is not. Therefore, the network cannot directly learn the connections between rooms. Using graph similarity of the floor plan images, we can train Siamese Network [7] that learns to discriminate similar and dissimilar image pairs.













## ACKNOWLEDGMENTS

This work was partly supported by the Grants-in-Aid for Scientific Research (no. 26700008) from JSPS and The Asso-



























ciation of Real Estate Agents of Japan (Fudosan Ryutsu Keiei Kyokai, FRK).

## REFERENCES

- [1] <https://www.keieiken.co.jp/monthly/2016/0405/index.html>.
- [2] K. Ohara, T. Yamasaki, and K. Aizawa, "An intuitive real estate retrieval system that queries the layout and size," *The 78th national convention of IPSJ*, 2016(in Japanese).
- [3] <https://www.redfin.com/>.
- [4] T. Hanazato, Y. Hirano, and M. Sasaki, "Syantic analysis og large size condominium units supplied in the tokyo metroporitan area," *Journal of Structural and Construction Engineering*, no. 591, pp. 9–16, 2005.
- [5] A. Takizawa, kazuma Yoshida, and N. Kato, "Applying graph mining to rent analysis considering room layouts," *Journal of Environmental Engineerin*, vol. 73, no. 623, pp. 139–146, 2008.
- [6] C. Liu, J. Wu, P. Kohli, and Y. Furukawa, "Deep multi-modal image correspondence learning," *arXiv preprint arXiv:1612.01225*, 2016.
- [7] J. Bromley, J. W. Bentz, L. Bottou, I. Guyon, Y. LeCun, C. Moore, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," *IJPRAI*, vol. 7, no. 4, pp. 669–688, 1993.
- [8] <http://www.nii.ac.jp/dsc/ldr/next/homes.html>.
- [9] <http://lifull.com/>.
- [10] <http://www.nii.ac.jp/>.
- [11] H. Chu, D. Ki Kim, and T. Chen, "You are here: Mimicking the human thinking process in reading floor-plans," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 2210–2218.
- [12] S. Ahmed, M. Liwicki, M. Weber, and A. Dengel, "Automatic room detection and room labeling from architectural floor plans," in *DAS. IEEE*, 2012, pp. 339–343.
- [13] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [14] <http://suomo.jp/>.
- [15] S. Tokui, K. Oono, S. Hido, and J. Clayton, "Chainer: a next-generation open source framework for deep learning," in *Proceedings of workshop on LearningSys in the twenty-ninth annual conference on NIPS*, vol. 5, 2015.
- [16] <https://crowdsourcing.yahoo.co.jp/>.

query		1st	2nd	3rd	4th	5th
	metadata					
	ImageNet					
	layout type					
	presence of room					
	proposed method					

(a)

query		1st	2nd	3rd	4th	5th
	metadata					
	ImageNet					
	layout type					
	presence of room					
	proposed method					

(b)

Fig. 4: Successful cases



























query		1st	2nd	3rd	4th	5th
	metadata					
	ImageNet					
	layout type					
	presence of room					
	proposed method					

Fig. 5: Failure cases