Similar Interior Coordination Image Retrieval with Multi-view Features

Ren Togo[1*], Yuki Honma[2,3], Maiku Abe[3], Takahiro Ogawa[1] and Miki Haseyama[1]

[1]Faculty of Information Science and Technology, Hokkaido University, Sapporo, Japan.
[2]Information System Innovation Office, Nitori Holdings Co., Ltd, Sapporo, Japan.
[3]Education and Research Center for Mathematical and Data Science, Hokkaido University, Sapporo, Japan.

*Corresponding author(s). E-mail(s): togo@lmd.ist.hokudai.ac.jp;

Abstract

This paper presents a novel similar image retrieval method for interior coordination. Interior coordination is very familiar; however, it is still an abstract and difficult concept. Even if we are involved in coordination every day, it does not mean we can become professional coordinators. By realizing the retrieval that can provide similar interior coordination images from a query room image, inspiring users' ideas for interior coordination becomes feasible. In the proposed method, we extract image features specialized for interior coordination and realize similar interior coordination image retrieval. We employ multi-view features: object-based, color-based, and semantic-based features, in the feature extraction phase. The extracted features are used to calculate similarity between the query image and the database images for the retrieval. We conducted experiments using a sophisticated real-world interior coordination image dataset. Furthermore, we qualitatively and quantitatively evaluated the effectiveness of the proposed method.

Keywords: interior coordination, deep learning, similar image retrieval, multi-view features

1 Introduction

Modern society has entered a new development phase with the rise in advanced technologies, such as the Internet of things, artificial intelligence, and big data analysis [1, 2]. Such technological innovation has made it possible to easily access much information and has created an increasingly complex and diverse set of individuals' needs [3, 4]. Currently, we always use smartphones and wearable devices to obtain all kinds of information in real-time. Many digital devices allow us to easily access information about the weather, news, and information about our work according to their preferences. However, information that satisfies our needs is only part of the explosive growth of information, and it is becoming increasingly difficult for them to search for the desired information [5, 6]. Therefore, it is essential to establish efficient information retrieval and recommendation technology that can provide information according to our preferences.

In the industrial sector, understanding the users' needs is one of the most critical business issues [7, 8]. Since user needs are greatly reflected in user behavior, there is an urgent problem to produce products that can flexibly meet these changes. A proper understanding of users' interests can increase product sales. Particularly, in the manufacturing and retailing sectors, which have a close relationship with consumers, it is essential to develop products that understand users' needs [9, 10]. For example, by understanding or anticipating users' needs, retailers can develop products with a high level of consumer satisfaction while controlling production costs. Efficient production and consumption are desirable for industries and consumers and from the sustainable development goals (SDGs)' perspective [11]. Introducing the SDG perspectives into business is necessary for the sustainable development of the global environment and society. It can lead to the growth of the business and an increase in the company's value.

The living environment is a fundamental part of our lives [12]. It reflects our preferences and has a significant impact on our daily lives. Interior coordination, which is the design of furniture, lighting, etc., to suit an individual's lifestyle, is one of the fields where personal preferences are greatly reflected. Interior coordination is one of the most important factors when designing an individual's living environment; however, this is an abstract concept. Additionally, it can be challenging to coordinate with the current room conditions or design an ideal space from scratch. Interior coordination is also a concept that is not easy to evaluate quantitatively, as it is evaluated from various perspectives. Although users can access information about interior design in different ways using their smartphones, it is becoming increasingly difficult to find the information they want. Therefore, the realization of a manufacturing method that can reflect the users' preferences in interior coordination will significantly improve users' satisfaction. For retailers proposing interior coordination, the realization of coordination proposals based on the users' preferences can improve productivity and cost reduction. In this way, interior coordination has become one of the most important themes for consumers and retailers.

In the past, the main way of proposing interior coordination was the consultation by specialists at stores. The problem with this approach is that the number of users who can receive the service is limited, and the efficiency of the service is low. Recently, interior coordination proposals have been made using information technology (IT). Research on interior coordination has included three-dimensional (3D) interior simulation and augmented reality (AR) technology for furniture arrangement, which have been applied to the real-world [13, 14]. The IT allows users to express their coordination; however, it is still difficult to represent complex and diverse personal preferences using these conventional technologies because it is necessary to create individual coordination. Additionally, it is necessary to use AR furniture samples with varying textures from those used in the original products.

As a potential platform that can reflect user preferences, Web applications, such as social networking services, can be mentioned [15]. These platforms are used as a new infrastructure worldwide, and a wide variety of data is collected [16, 17]. For example, users can collect images of coordinated cases that reflect their preferences from the platform as a reference for their coordination. In this way, users can easily obtain coordination examples that reflect their preferences through Web applications. One of the most effective ways to support interior coordination is to use a retrieval system. For retailers, the retrieval system has already become the foundation for user acquisition [18, 19]. When we want to buy something, we input it as a query to the retrieval system, and we purchase it based on the retrieved results. Currently, we can easily search for the desired furniture or lighting by specifying the query as either a text or an image. In design, advertising, and travel, images on the Web are used to stimulate user inspiration. For example, Pinterest, one of the Web platforms, supports users' creative activities by presenting them with relevant images. Similar image retrieval can meet the users' needs due to its simplicity [20].

In this paper, we propose a new image retrieval method to support interior coordination by reflecting personal preferences. The proposed method consists of furniture recognition based on multi-view feature extraction and retrieval technology that considers the features for coordination. When coordinating interiors, the information in multiple objects and their relationships are essential in addition to the overall atmosphere. The proposed method extracts features that consider the relationships between objects and uses them to compare the similarity of coordinated images. Employing this object detection-based bottom-up attention can also reduce annotation costs. Furthermore, updating the retrieval candidate database becomes easier by automatically detecting objects and their positions. In this way, the similarity evaluation specialized for interior coordination becomes possible. Note that this paper is the extension of our previous study in [21].

We conducted experiments to verify the effectiveness of the proposed method using actual interior coordination examples. We construct our retrieval method using images of coordination samples provided by Nitori Holdings Co., Ltd., a Japanese retail company (hereinafter, Nitori). Figure 1 shows examples of coordination image data and their coordinate styles used in this study. Each coordinate style (simple, natural, vintage, feminine, and Japanese modern) is defined by Nitori. We can see that each coordinate style has a different appearance and atmosphere; however, it is difficult to describe the abstract difference. Similar coordinated images can be retrieved by using an image as a query. This will help users have a more concrete imagination for purchasing. Besides, it will be possible to grasp the manufacturer/retailer's side trend by analyzing the uploaded query images. Although various methods for image retrieval have been proposed in the past, to the best of our knowledge, there have not been any feature extraction techniques specialized for interior coordination image retrieval.

The contributions of this study are summarized as follows.

- We propose a similar image retrieval method based on multi-view features to support interior coordination without any detailed annotations.
- The effectiveness of the proposed method was evaluated using the sophisticated interior coordination images provided by a retail company from Japan.

2 Related Works

In this section, we present some related works and discuss the differences between conventional methods and our approach.

2.1 Interior Coordination

In interior coordination, it is essential to consider the overall atmosphere of the room [22]. Interior decoration is a creative and fascinating topic that involves creating a new atmosphere from nothing using color palettes and trends. Various interiors of a room have unique atmospheres and are the focus of interior design studies [23]. How the interior's atmosphere is created and how space is perceived are important in creating the atmosphere. In interior atmosphere creation, decorative backgrounds, indoor props, light, special effects, shadows, and colors are often used to express a particular atmospheric state. The combination of the physical and psychological suggests a certain "reading" of the space, giving the interior an emotional resonance in the process. It provides a framework to understand and create complex interior atmospheres.

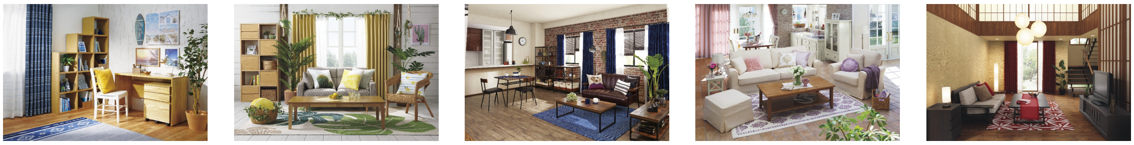
Interior design is understood as the design or renovation of the interior of a building [24]. Although the enclosure of buildings defines interior design, its actual usage is broader and encourages a more complementary understanding [25]. In existing interior designs, the structure of the building and physical conditions (e.g., furniture) and space and virtual conditions (e.g., the flow of people) are considered.

When we discuss the ideal form of interior coordination in the future, consideration for the environment, incorporating the SDG perspective, is an essential issue [26, 27]. However, environmentally sustainable interior design criteria are unclear, and few studies have been conducted. The use of sustainable interior decoration materials will become one of the new criteria for consumers to choose products. Manufacturers and retailers are expected to produce products with this concept in mind [28]. The first step is to recognize and share the importance of this concept.

In this way, interior coordination can be discussed from various perspectives. However, this concept is difficult for general consumers to understand, and it is difficult for them to incorporate it into their actual lives [29]. It is required to be able to understand and express interior coordination intuitively.

2.2 Image Retrieval Technology

Image retrieval can be applied to various tasks and retrieval targets, and many methods have been proposed [30]. Among them, tag-based image retrieval is one of the most representative image retrieval methods. The user inputs a query text associated with a tag representing the desired



(a) Simple (b) Natural (c) Vintage (d) Feminine (e) Japanese Modern

Fig. 1: Coordination image samples used in this study

image [31]. Images in the database are pre-assigned text tags, and images with relevant tags are presented at the top of the retrieval results by comparing the query with the tags. Tag-based image retrieval is the most accurate way to find the desired images if the database is given the correct tags and the user inputs the best query. However, tagging images is a time-consuming task, and there is still the problem of ambiguity in the user's query.

Cross-modal retrieval is an image retrieval method that can be used even when candidate images have not been tagged with metadata; it can retrieve images of different modalities with text as the query [32–35]. In this approach, text and images are projected into the same feature space, allowing for the comparison of different modalities. Cross-modal retrieval does not require tagging of images to be retrieved, significantly reducing the burden of database construction. However, user query ambiguity remains an issue, and various studies are being conducted to solve this problem.

In content-based image retrieval, the input query is an image [36–38]. Similar image retrieval can be implemented by comparing the input query image features with the features of the database images. Since the amount of information in the query is more significant than that of the text, it is easier to retrieve the desired image. It is worth noting that this retrieval approach assumes that the user already owns the query's contents, limiting its availability.

In this study, we focus on content-based image retrieval for interior coordination. We assume that the user is a general consumer who does not have specialized knowledge about interior coordination. Here it is not easy to express the ideal interior coordination linguistically. Images of interior coordination rooms that reflect personal preferences can easily be obtained from the Web. Therefore, we consider content-based image retrieval as the most appropriate method. The task of image retrieval in interior coordination is not similar to the task of identical object detection. Even if an image contains the same object, the coordination depends on the relationship with other furniture, the overall color of the room, and the angle from which the room is captured.

3 Our Retrieval Method

We employ three types of image features (object-based, color-based, and semantic features) in the proposed method to realize interior coordination retrieval. The object type and its positions are essential in interior coordination; therefore, we extract their information by applying a pre-trained object detection model, YOLOv5 [39]. The detected object names are converted into text feature representations using a pre-trained text classification model, FirstText [40]. Then, we divide the image into four rectangular regions and extract histogram-based color features from each region. Next, we extract semantic features from a pre-trained deep learning model DenseNet [41]. Finally, features, including the information of objects, positions, colors, and semantic information of the image, are used to calculate the similarity between the query and database. Figure 2 shows the overview of the proposed method.

First, we resize all images into the size of 768×512 pixels since various sizes of images are expected to be input as queries in practical use. This situation may affect retrieval performance. The aspect ratio of the images is 3:2, which is the standard aspect ratio of images taken by smartphones and other devices. We first extract object classes and their positional information from images $\mathbf{I}_n (n = 1, 2, \cdots, N;$ where N is the number of the target images) based on the recently proposed sophisticated object detection model YOLOv5 as follows:

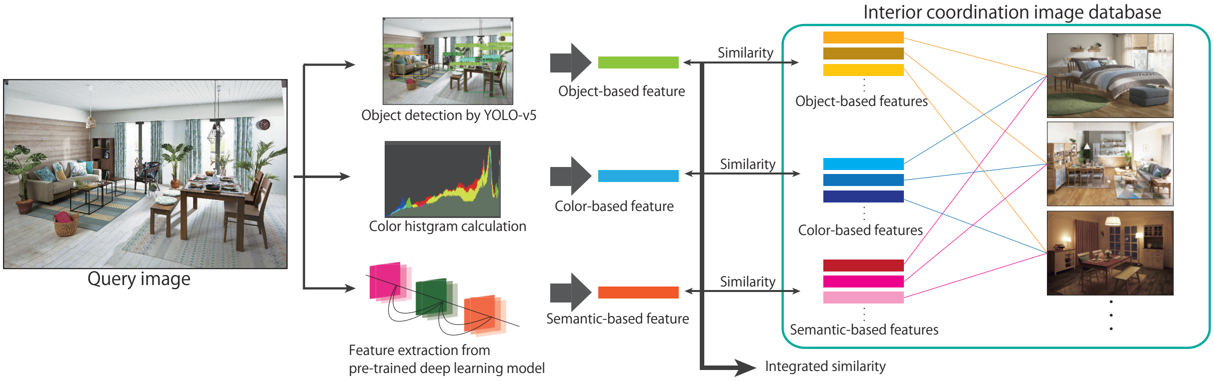


Fig. 2: Overview of the proposed method

$$o^n_{s,c} = \mathrm{YOLOv5}(\mathbf{I}_n),$$
$$(s = 1, 2, \cdots, S), (c = 1, 2, \cdots, C), \quad (1)$$

where $o^n_{s,c}$ is an object detected by YOLOv5; S is the number of detected objects from $\mathbf{I}_n$; C is the number of classes. Note that $o^n_{s,c}$ includes a detected object name and its positional information of the bounding box in the image $\mathbf{I}_n$ as follows:

$$o^n_{s,c} \in (o^{n,\mathrm{name}}_{s,c}, \boldsymbol{v}^{n,\mathrm{pos}}_{s,c}), \quad (2)$$

where $o^{n,\mathrm{name}}_{s,c}$ represents a text description of a detected object, and $\boldsymbol{v}^{n,\mathrm{pos}}_{s,c} \in \mathbb{R}^4$ represents coordinates in an image. YOLOv5 is trained on a large-scale dataset Microsoft common objects in context (MSCOCO) [42]. MSCOCO dataset consists of pairs of daily photos and their descriptions, each of which is associated with object labels from a set of 80 categories. MSCOCO dataset is widely used for evaluations in image classification, object detection, cross-modal vision, and language tasks. Therefore, in the proposed method, the number of classes that can be detected is 80, i.e., $C = 80$. In the practical situation, annotating coordinated images is not a realistic approach because it requires a lot of time and effort. This object detection-based bottom-up attention that detects the target objects automatically can solve the annotation cost problem.

Although object detection-based bottom-up attention can localize objects, the relationships between each object still cannot be considered. In a similar interior coordination image retrieval situation, it is considered that the same object is not always present in the query and database. Therefore, to enhance the robustness of our retrieval method, we transform a detected object name $o^{n,\mathrm{name}}_{s,c}$ into text features to consider the relationships between detected objects as follows:

$$\boldsymbol{v}^{n,\mathrm{txt}}_{s,c} = \mathrm{FirstText}(o^{n,\mathrm{name}}_{s,c}), \in \mathbb{R}^{300} \quad (3)$$

The function FirstText represents the text encoding model for feature extraction. By transferring object names into text features as $\boldsymbol{v}^{n,\mathrm{txt}}_{s,c}$, we can treat similar objects (e.g., sofa and bed) as similar features when calculating object similarities. This process is one of the advantages of the proposed method. Finally, object-based features from the images $\mathbf{I}_n$ can easily be extracted as follows:

$$\boldsymbol{v}^{n,\mathrm{obj}}_{s,c} = [\boldsymbol{v}^{n,\mathrm{txt}\top}_{s,c}, \boldsymbol{v}^{n,\mathrm{pos}\top}_{s,c}]^\top, \in \mathbb{R}^{304}. \quad (4)$$

In this way, we extract information of objects and their positions from interior coordination images. In retrieval situation, we do not have to provide object-wise annotations for a query and database since automatically detected objects and their positions are compared to those of the database.

Next, we extract color features from $\mathbf{I}_n$ to catch the atmosphere of the room. Color is one of the essential elements in interior coordination. Even if a room consists of the same furniture, changing the color can significantly change the image and mood of the room. For example, warm colors, such as red and yellow, can uplift the mood and create a sense of warmth. However, cold colors, such as light blue and blue, make one feel cooler and more focused. Additionally, neutral colors, such as green, have a relaxing and healing effect. Since it is not easy to capture such features using object detection techniques, the proposed method extracts new features focusing on colors.

We divide the image $\mathbf{I}_n$ into four parts and extract color histogram features $\boldsymbol{c}$ from each region as follows:

$$\boldsymbol{v}^{n,\mathrm{color}} = [\boldsymbol{c}_{n,1}{}^{\top}, \boldsymbol{c}_{n,2}{}^{\top}, \boldsymbol{c}_{n,3}{}^{\top}, \boldsymbol{c}_{n,4}{}^{\top}]^{\top}, \in \mathbb{R}^{64}. \tag{5}$$

The number of bins used in this study is set to 64 experimentally.

Next, we extract semantic-based features $\mathbf{I}_n$ to enhance the representation ability for image retrieval. It is necessary to extract more generic and robust features to represent various types of furniture. We use a pre-trained deep learning-based model DenseNet [41] to represent semantic robust features. Among various types of deep learning models, DenseNet is one of the sophisticated models with fewer parameters and high accuracy in the natural image classification task. The advantage of DenseNet is that the model has DenseBlocks to improve feature propagation in forward and backward fashions. Hence, we employ DenseNet as a feature extractor and extract semantic features $\boldsymbol{v}^{n,\mathrm{sem}} \in \mathbb{R}^{2,048}$ from $\mathbf{I}_n$.

Finally, we calculate the similarity of a query and target database images based on the above-derived features. Let $\mathbf{Q}$ denote a query image, and its object-based text feature $\boldsymbol{v}^{\mathrm{obj}}_{t,c}(t = 1, 2, \cdots, T)$ can be calculated using Eqs. 1 - 4. Note that T represents the number of detected objects from the query image $\mathbf{Q}$. Color features $\boldsymbol{v}^{\mathrm{color}}$ of $\mathbf{Q}$ can also be calculated using Eq. 5.

The similarity of object features between $\mathbf{Q}$ and $\mathbf{I}_n$ can be calculated as follows:

$$\mathcal{S}_{\mathrm{obj}} = \frac{1}{S+T} \sum_{s=1}^{S} \sum_{t=1}^{T} \frac{\boldsymbol{v}^{\mathrm{obj}}_{t,c} \cdot \boldsymbol{v}^{n,\mathrm{obj}}_{s,c}}{\|\boldsymbol{v}^{\mathrm{obj}}_{t,c}\| \|\boldsymbol{v}^{n,\mathrm{obj}}_{s,c}\|}. \tag{6}$$

Similarly, the similarity of color and semantic features $\mathbf{Q}$ and $\mathbf{I}_n$ can be calculated as follows:

$$\mathcal{S}_{\mathrm{color}} = \frac{\boldsymbol{v}^{\mathrm{color}} \cdot \boldsymbol{v}^{n,\mathrm{color}}}{\boldsymbol{v}^{\mathrm{color}} \boldsymbol{v}^{n,\mathrm{color}}}, \tag{7}$$

$$\mathcal{S}_{\mathrm{sem}} = \frac{\boldsymbol{v}^{\mathrm{sem}} \cdot \boldsymbol{v}^{n,\mathrm{sem}}}{\boldsymbol{v}^{\mathrm{sem}} \boldsymbol{v}^{n,\mathrm{sem}}}. \tag{8}$$

The integrated similarity is calculated as follows:

$$\mathcal{S} = \alpha \mathcal{S}_{\mathrm{obj}} \cdot \beta \mathcal{S}_{\mathrm{color}} \cdot \gamma \mathcal{S}_{\mathrm{sem}}, \quad \text{s.t. } \alpha + \beta + \gamma = 1, \tag{9}$$

where α, β, and γ are weighted parameters. From the obtained similarities $\mathcal{S}$, we can rank the candidate images in the database, and top k ranked images are provided to users as a retrieval result. In this paper, we experimentally set each weighted parameter as 0.33.

4 Results

4.1 Research data

In this study, we construct a similar image retrieval method using images of coordination samples provided by Nitori as the NITORI-dataset. As shown in Fig. 3, there are many types of coordination examples from the perspective of the types of the room (e.g., bedroom (Fig. 3 (a)), dining room (Fig. 3 (b)), and living room (Fig. 3 (c))). For example, the same bedroom can look different depending on the furniture, color, and atmosphere of the room. In this way, room coordination is an abstract concept with complex elements, and there have been few studies to express it quantitatively.

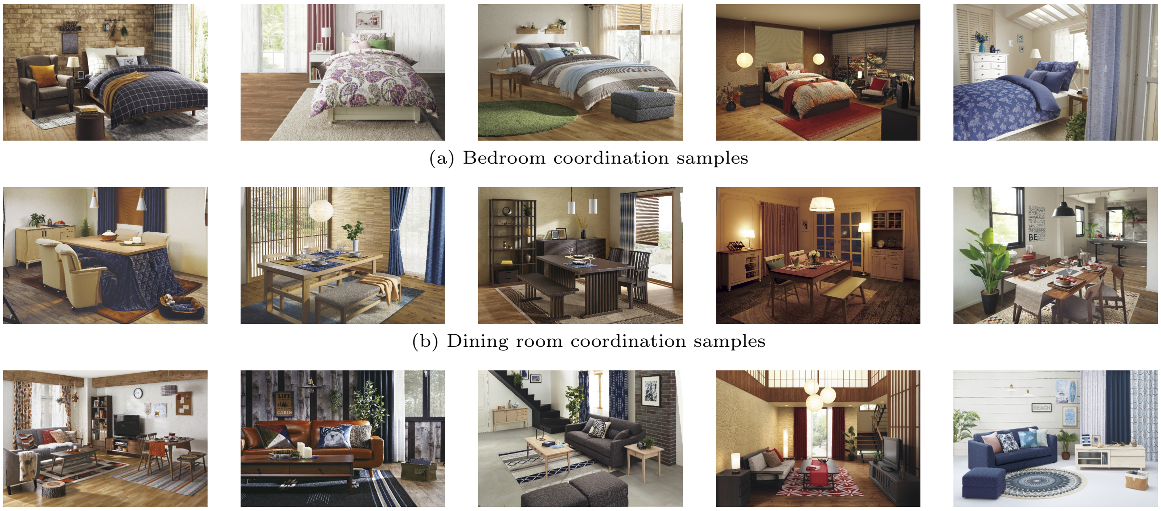
We also use another room image dataset for real-world applications, the CVPR-indoor-dataset, which was proposed in 2009 for room image recognition [43]. The dataset consists of images collected based on search engines on the Web, such as Google, and images collected from Flicker, an image-posting service. Note that this dataset excludes coordinated room images but various room images collected from the Web platforms. In this study, we use this dataset as a query image to qualitatively evaluate the interior coordination retrieval accuracy.

We construct a similar image retrieval method for interior coordination using the above dataset and verify the effectiveness of the proposed technique. As far as we know, no similar image retrieval technique focusing on interior coordination has been proposed. Additionally, the effectiveness of the method using images with well-organized coordination has not been verified so far.

4.2 Experimental Settings

We used the following two datasets in the experiment.

NITORI-dataset (Fig. 3) is images of interior

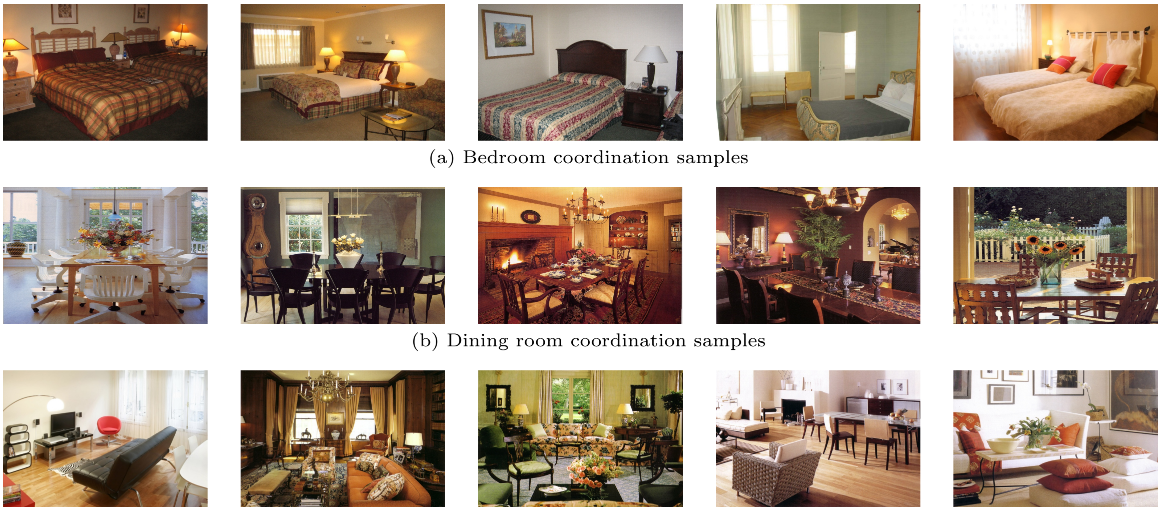


(a) Bedroom coordination samples

(b) Dining room coordination samples

(c) Living room coordination samples

Fig. 3: Examples of interior coordination images categorized by the type of the room from the NITORI-dataset



(a) Bedroom coordination samples

(b) Dining room coordination samples

(c) Living room coordination samples

Fig. 4: Examples of interior coordination images categorized by the type of the room from the CVPR-indoor-dataset

coordination samples provided by Nitori. It consists of 43 bedroom, 27 dining room, and 102 living room images (a total of 172 images). The annotation of the definition of the room was conducted by a Nitori employee.

CVPR-indoor-dataset (Fig. 4) consists of images for indoor scene recognition. It consists of 67 indoor categories and 15,620 images. The number of images is different for each category; however, each category contains at least 100 images. There is no consideration of room interior coordination. Therefore, we selected the classes and images according to the NITORI-dataset. In this experiment, we used 25 bedroom, 35 dining room,

Table 1: Explanation of the proposed method and comparison methods

Method	Overview
PM	Our proposed method using multi-view features
CM 1 (color + obj)	The method using color and object features
CM 2 (obj)	The method using object features
CM 3 (color)	The method using color features
CM 4 (DenseNet201) [41]	The method using the output feature of DenseNet201. DenseNet is the network using skip connection architectures and the improved version of ResNet.
CM 5 (InceptionResNetV2) [44]	The method using the output feature of InceptionResNetV2. InceptionResNetV2 is the network builds on the Inception architectures but incorporates residual connections.
CM 6 (InceptionV3) [45]	The method using the output feature of InceptionV3. InceptionV3 is the network builds on Label Smoothing, Factorized convolutions, and an auxiliary classifier.
CM 7 (ResNet50) [46]	The method using the output feature of ResNet50. ResNet is the network builds on residual blocks to alleviate the problem of very deep networks.
CM 8 (VGG19) [47]	The method using the output feature of VGG19. VGG is the network having improved the traditional convolutional neural networks.
CM 9 (Xception) [48]	The method using the output feature of Xception. Xception is the network builds on the depthwise separable convolutions.

and 23 living room images (a total of 83 images) as query images.

As evaluation metrics, we used mean average precision (MAP)@k and normalized discounted cumulative gain (NDCG)@k. Both metrics represent the performance for retrieval. MAP@k reflects the accuracy of top-ranked items by a model and can be calculated as the mean of AP@k for each item in the test dataset. NDCG@k involves a discount function over the rank, while many other measures uniformly weight all positions. We used $k = 4$ in this experiment and set the ground truth to three classes: bedroom, living room, and dining room. As compararison methods, we employed deep learning-based semantic, object-based, and color-based features, as presented in Table 1.

4.3 Experimental Results

Tables 2 and 3 presents the retrieval performance of the proposed method and comparison methods. We can see that our method achieved higher performance than other methods on average. In case query images were from the NITORI-dataset, retrieval accuracy of each method was higher than that from the CVPR-indoor-dataset. It is confirmed that the performance of the proposed method is the highest when the query images are from the NITORI-dataset, which is a well-coordinated dataset. However, when query images are from the CVPR-indoor dataset, which is not intended for coordination, the retrieval performance of all methods decreases.

Figure 5 shows the retrieved samples of the proposed method query with the NITORI-dataset. As shown in Fig. 5, similar coordination images can be retrieved from the database using the proposed method. Even if the room types are different in the query image, our method based on multi-view features works well.

5 Discussion

The experimental results show that the proposed method can provide similar interior coordination images. Although there are various types of similar image retrieval methods, it is necessary to consider the coordination characteristics to realize interior coordination retrieval.

Interior coordination is essentially something that everyone in the world can enjoy individually; however, it is abstract and difficult to understand. Our retrieval approach will make interior coordination more accessible and concrete for users. One of the advantages of this study is that it

Table 2: MAP@4 on each dataset

Method	NITORI-dataset	CVPR-indoor-dataset	Mean
PM	0.850	0.223	**0.537**
CM1 (color+obj)	0.525	0.111	0.318
CM2 (obj)	0.650	0.205	0.428
CM3 (color)	0.483	0.008	0.246
CM4 (DenseNet201) [41]	0.817	0.226	0.522
CM5 (InceptionResNetV2)) [44]	0.483	0.108	0.300
CM6 (InceptionV3) [45]	0.817	0.201	0.509
CM7 (ResNet50) [46]	0.483	0.178	0.331
CM8 (VGG19) [47]	0.817	0.193	0.505
CM9 (Xception) [48]	0.750	0.229	0.490

Table 3: NDCG@4 on each dataset

Method	NITORI-dataset	CVPR-indoor-dataset	Mean
PM	0.858	0.219	**0.539**
CM1 (color+obj)	0.507	0.109	0.308
CM2 (obj)	0.643	0.193	0.418
CM3 (color)	0.500	0.008	0.254
CM4 (DenseNet201) [41]	0.843	0.222	0.533
CM5 (InceptionResNetV2)) [44]	0.508	0.104	0.306
CM6 (InceptionV3) [45]	0.824	0.213	0.519
CM7 (ResNet50) [46]	0.782	0.188	0.485
CM8 (VGG19) [47]	0.750	0.198	0.474
CM9 (Xception) [48]	0.792	0.234	0.513

does not require any annotation by the user. Even if the user is not familiar with interior coordination, they can retrieve similar coordinates by simply uploading an image. Additionally, from the manufacturer and retailer perspectives, tagging coordinates is a very labor-intensive task when creating new products one after another. In this study, we use furniture detection and feature extraction based on an object recognition technique, which enables retrieval without annotation. Using this method makes it possible to build a robust retrieval technology.

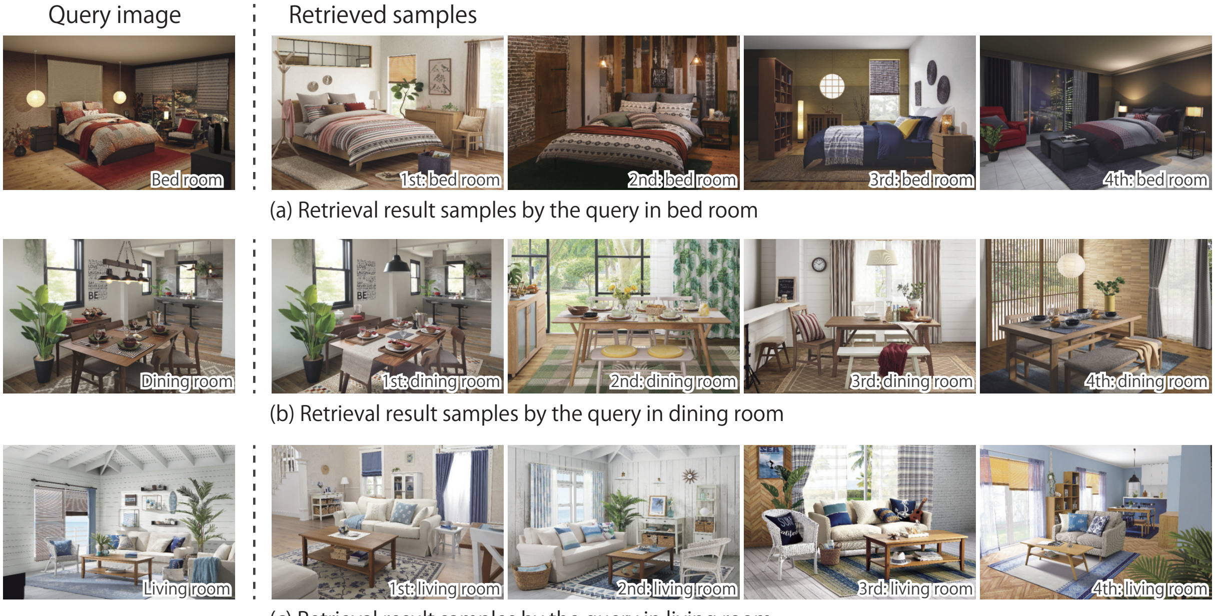
There are several limitations to this study. First, the proposed method was constructed with reference to the defined coordination. However, the definition of interior coordination is broad, and it is still difficult to evaluate quantitatively. We used room type as a criterion for evaluation; however, evaluation should be done using coordination tags from all perspectives. The evaluation was based on the accuracy of the retrieval, but it is thought that evaluation by professional coordinators will also be necessary. We will try to solve the above issues in the next step of this study.

6 Conclusion

We have presented the novel similar image retrieval method for interior coordination. We employed object-based, color-based, and semantic-based features for interior coordination image retrieval. Our method integrating these multi-view features achieved higher retrieval performance than conventional methods in several datasets.

Acknowledgments. The authors thank Nitori Holdings Co., Ltd, for providing research data used in this study.

Funding. This research was partly supported by JSPS KAKENHI Grant Number JP20K19857



(a) Retrieval result samples by the query in bed room

(b) Retrieval result samples by the query in dining room

(c) Retrieval result samples by the query in living room

Fig. 5: Retrieved samples using the proposed method

and the NITORI Future Design Course at Education and Research Center for Mathematical and Data Science, Hokkaido Univ.

Data availability. The NITORI-dataset remains a private dataset, while the CVPR-indoor-dataset is a public accessible dataset.

Author contributions statement. Conceptualization, R. Togo; methodology, R. Togo, Y. Honma, and M. Abe; software, R. Togo; validation, R. Togo and Y. Honma; formal analysis, R. Togo; investigation, R. Togo, Y. Honma, and M. Abe; resources, R. Togo, T. Ogawa, and M. Haseyama; data curation, Y. Honma; writing—original draft preparation, R. Togo; writing—review and editing, R. Togo, Y. Honma, M. Abe, T. Ogawa; visualization, R. Togo; supervision, M. Abe, T. Ogawa, and M. Haseyama; project administration, R. Togo; funding acquisition, R. Togo. All authors have read and agreed to the published version of the manuscript.

Compliance with Ethical Standards. This paper contains no cases of studies with human participants performed by any of the authors.

Competing Interests. The authors declare that they have no conflict of interest.

References

[1] Lu, Y.: Artificial intelligence: a survey on evolution, models, applications and future trends. Journal of Management Analytics **6**(1), 1–29 (2019)

[2] ur Rehman, M.H., Yaqoob, I., Salah, K., Imran, M., Jayaraman, P.P., Perera, C.: The role of big data analytics in industrial internet of things. Future Generation Computer Systems **99**, 247–259 (2019)

[3] Marjani, M., Nasaruddin, F., Gani, A., Karim, A., Hashem, I.A.T., Siddiqa, A., Yaqoob, I.: Big iot data analytics: architecture, opportunities, and open research challenges. IEEE Access **5**, 5247–5261 (2017)

[4] Greco, L., Percannella, G., Ritrovato, P., Tortorella, F., Vento, M.: Trends in iot based solutions for health care: Moving ai to the edge. Pattern Recognition Letters **135**, 346–353 (2020)

[5] Li, S., Da Xu, L., Zhao, S.: The internet of things: a survey. Information Systems Frontiers **17**(2), 243–259 (2015)

[6] Al-Emran, M., Malik, S.I., Al-Kabi, M.N.: A survey of internet of things (iot) in education: opportunities and challenges. Toward social internet of things (SIoT): Enabling technologies, architectures and applications, 197–209 (2020)

[7] Savolainen, R.: Everyday life information seeking: Approaching information seeking in the context of "way of life". Library & Information Science Research **17**(3), 259–294 (1995)

[8] Gibbs, J.: Interior Design, (2005)

[9] Khanam, S., Jang, S.-W., Paik, W.: Shape retrieval combining interior and contour descriptors. In: International Conference on Future Generation Communication and Networking, pp. 120–128 (2011)

[10] Liu, M., Fang, Y., Choulos, A.G., Park, D.H., Hu, X.: Product review summarization through question retrieval and diversification. Information Retrieval Journal **20**(6), 575–605 (2017)

[11] Leng, J., Ruan, G., Jiang, P., Xu, K., Liu, Q., Zhou, X., Liu, C.: Blockchain-empowered sustainable manufacturing and product lifecycle management in industry 4.0: A survey. Renewable and Sustainable Energy Reviews **132**, 110112 (2020)

[12] Raanaas, R.K., Evensen, K.H., Rich, D., Sjøstrøm, G., Patil, G.: Benefits of indoor plants on attention capacity in an office setting. Journal of Environmental Psychology **31**(1), 99–105 (2011)

[13] Fu, Q., Chen, X., Wang, X., Wen, S., Zhou, B., Fu, H.: Adaptive synthesis of indoor scenes via activity-associated object relation graphs. ACM Transactions on Graphics (TOG) **36**(6), 1–13 (2017)

[14] Moares, R., Jadhav, V., Bagul, R., Jacbo, R., Rajguru, S., *et al.*: Inter ar: Interior decor app using augmented reality technology. In: Proceedings of the 5th International Conference on Cyber Security & Privacy in Communication Networks (ICCS), pp. 141–146 (2019)

[15] Dustdar, S., Schreiner, W.: A survey on web services composition. International Journal of Web and Grid Services **1**(1), 1–30 (2005)

[16] Liu, M., Zhang, K., Zhu, J., Wang, J., Guo, J., Guo, Y.: Data-driven indoor scene modeling from a single color image with iterative object segmentation and model retrieval. IEEE Transactions on Visualization and Computer Graphics **26**(4), 1702–1715 (2020)

[17] Yanagi, R., Togo, R., Ogawa, T., Haseyama, M.: Enhancing cross-modal retrieval based on modality-specific and embedding spaces. IEEE Access **8**, 96777–96786 (2020)

[18] Shih, J.-L., Chen, H.-Y.: A 3d model retrieval approach using the interior and exterior 3d shape information. Multimedia Tools and Applications **43**(1), 45–62 (2009)

[19] Kaothanthong, N., Chun, J., Tokuyama, T.: Distance interior ratio: A new shape signature for 2d shape retrieval. Pattern Recognition Letters **78**, 14–21 (2016)

[20] Zhang, J.: Visualization for Information Retrieval vol. 23, (2007)

[21] Togo, R., Ogawa, T., Haseyama, M.: Interior coordination image retrieval with object-detection-based and color features. In: International Workshop on Advanced Imaging Technology (IWAIT) 2021, vol. 11766, p. 1176616 (2021)

[22] Daniels, I.: Feeling at home in contemporary japan: Space, atmosphere and intimacy. Emotion, Space and Society **15**, 47–55 (2015)

[23] Lohr, V.I., Pearson-Mims, C.H., Goodwin, G.K.: Interior plants may improve worker productivity and reduce stress in a windowless environment. Journal of environmental horticulture **14**(2), 97–100 (1996)

[24] Brooker, G., Stone, S.: What Is Interior Design?, (2010)

[25] Cho, J.Y., Suh, J.: Understanding spatial ability in interior design education: 2d-to-3d visualization proficiency as a predictor

of design performance. Journal of Interior Design **44**(3), 141–159 (2019)

[26] Sun, P., Zhang, N., Zuo, J., Mao, R., Gao, X., Duan, H.: Characterizing the generation and flows of building interior decoration and renovation waste: A case study in shenzhen city. Journal of Cleaner Production **260**, 121077 (2020)

[27] Kishi, R., Araki, A.: Importance of indoor environmental quality on human health toward achievement of the sdgs. In: Indoor Environmental Quality and Health Risk Toward Healthier Environment for All, pp. 3–18 (2020)

[28] Nhamo, G., Nhemachena, C., Nhamo, S.: Using ict indicators to measure readiness of countries to implement industry 4.0 and the sdgs. Environmental Economics and Policy Studies **22**(2), 315–337 (2020)

[29] Ruff, C.L., Olson, M.A.: The attitudes of interior design students towards sustainability. International Journal of Technology and Design Education **19**(1), 67–77 (2009)

[30] Datta, R., Joshi, D., Li, J., Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys (Csur) **40**(2), 1–60 (2008)

[31] Yasmin, M., Mohsin, S., Sharif, M.: Intelligent image retrieval techniques: a survey. Journal of Applied Research and Technology **12**(1), 87–103 (2014)

[32] Zhen, L., Hu, P., Wang, X., Peng, D.: Deep supervised cross-modal retrieval. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 10394–10403 (2019)

[33] Wang, B., Yang, Y., Xu, X., Hanjalic, A., Shen, H.T.: Adversarial cross-modal retrieval. In: Proceedings of the 25th ACM International Conference on Multimedia (ACM MM), pp. 154–162 (2017)

[34] Wei, Y., Zhao, Y., Lu, C., Wei, S., Liu, L., Zhu, Z., Yan, S.: Cross-modal retrieval with cnn visual features: A new baseline. IEEE Transactions on Cybernetics **47**(2), 449–460 (2016)

[35] Yanagi, R., Togo, R., Ogawa, T., Haseyama, M.: Database-adaptive re-ranking for enhancing cross-modal image retrieval. In: Proceedings of the 29th ACM International Conference on Multimedia (ACM MM), pp. 3816–3825 (2021)

[36] Liu, Y., Zhang, D., Lu, G., Ma, W.-Y.: A survey of content-based image retrieval with high-level semantics. Pattern Recognition **40**(1), 262–282 (2007)

[37] Gandhani, S., Singhal, N.: Content based image retrieval: survey and comparison of cbir system based on combined features. International Journal of Signal Processing, Image Processing and Pattern Recognition **8**(10), 155–162 (2015)

[38] Li, Y., Li, W.: A survey of sketch-based image retrieval. Machine Vision and Applications **29**(7), 1083–1100 (2018)

[39] Jocher, G., Stoken, A., Borovec, J., NanoCode012, ChristopherSTAN, Changyu, L., Laughing, tkianai, Hogan, A., lorenzomammana, yxNONG, AlexWang1900, Diaconu, L., Marc, wanghaoyang0106, ml5ah, Doug, Ingham, F., Frederik, Guilhen, Hatovix, Poznanski, J., Fang, J., Yu, L., changyu98, Wang, M., Gupta, N., Akhtar, O., PetrDvoracek, Rai, P.: Ultralytics/yolov5: V3.1 - Bug Fixes and Performance Improvements

[40] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics **5**, 135–146 (2017)

[41] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4700–4708 (2017)

[42] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Proceedings of the IEEE European Conference on Computer Vision (ECCV), pp. 740–755 (2014)

[43] Quattoni, A., Torralba, A.: Recognizing indoor scenes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)

[44] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.A.: Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-first AAAI Conference on Artificial Intelligence (2017)

[45] Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2818–2826 (2016)

[46] He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 770–778 (2016)

[47] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)

[48] Chollet, F.: Xception: Deep learning with depthwise separable convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 1251–1258 (2017)