Contour and Object-Oriented Learning for Indoor Cleanliness Classification

Sucheng Qian
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, China
+86 13328179990
qiansucheng@sjtu.edu.cn

Zhaoyu Li
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, China
Telephone number, incl. country code
Apollo19990425@sjtu.edu.cn

Weibang Jiang
Shanghai Jiao Tong University
800 Dongchuan Road
Shanghai, China
Telephone number, incl. country code
935963004@sjtu.edu.cn

ABSTRACT
The evaluation of indoor cleanliness is a meaningful task for vision-based household service systems. However, the perception of cleanliness is determined by diverse visual features and multiple criteria, which is subjective to the observer. We find the existing dataset and method fail to truthfully capture the concept of cleanliness because the feature used is not representative to human subjective judgement. Therefore, we create a dataset for indoor cleanliness classification from a group of annotators based on SUN-RGBD, a richly annotated scene understanding benchmark. Based on such analysis, we propose Contour and Object-orient Learning (COOL) model that integrates pretrained convolutional feature, low-level contour feature, and object arrangement in order to truthfully model the notion of cleanliness. Our design choices are justified in ablation studies, and our model outperforms the previous method in our dataset for cleanliness classification.

CCS Concepts
• **Computing methodologies → Artificial intelligence → Computer vision → Computer vision tasks → Scene understanding**

Keywords
Cleanliness classification; Feature analysis, Convolutional neural networks, Edge detection, Graph convolutional neural networks.

1. INTRODUCTION
Smooth bed sheets, ordered classroom desks, neat and clean restaurants, scene cleanliness is closely related to people's quality of life. Perceiving and evaluating indoor cleanliness computationally could be a helpful basic task for applications in future smart service industry, which is expected to create a more convenient and comfortable living environment.

To the best of our knowledge, few previous works has been done on scene cleanliness. [1] focuses on a specific application in measuring restroom cleanliness and performs detailed convolutional neural networks feature analysis. Visual complexity also partially relates to scene cleanliness, which measures the level of intricacy and details [2] or the amount of information conveyed in the image [3]. These works provide potential insights for cleanliness classification, but are not directly applicable. We compare our work with [17] which considers indoor cleanliness classification, and we will show that their oversimplified dataset and feature fail to capture the notion of cleanliness.

Figure 1. Cleanliness is reflected by two major criteria. The pair of images at the top are distinguished by color, contour, and texture. While the second image pair's cleanliness differs in the arrangement of chairs.

Scene cleanliness classification is an inherently complex task because humans rely on various standards to evaluate scene cleanliness, and their judgement can be subjective. As illustrated in Figure 1, cleanliness can be characterized not only by visual features including color, contour, texture, but the properly ordered arrangement of objects is also a crucial factor. Previous works exceedingly rely on handcrafted or biased feature, which is limited and cannot truthfully reflect human's evaluation of scene cleanliness.

Our work analyzes the limitations of the previously used image feature for scene cleanliness. Based on such analysis, we propose a comprehensive Contour and Object-orient Learning (COOL) model that integrates pretrained convolutional feature, low-level contour feature, and object arrangement in order to truthfully reflect human's subjective understanding of cleanliness. On the one hand, we make use of convolutional neural networks (CNN) pretrained on ImageNet to extract abstract visual feature. And do contour feature extraction to compensate for CNN feature. On the other hand, in order to model object-level placement, we use the positions and orientations of the main objects in the scene

obtained by 3D object detection. Then construct a scene graph by regarding objects as attributed nodes, and their relative displacement as attributed edges. Our model summarizes object arrangement as graph embedding using graph convolutional networks, which serves as another feature for scene cleanliness classification.

Our method achieves state-of-the-art performance on the dataset created by ourselves compared with previously proposed method. We further justify the effectiveness of our design in ablation studies in section 4.3. The code of this project is publicly available at GitHub repository https://github.com/ApolloLiZhaoyu/Cleanliness_Classification.

2. RELATED WORK

In this section, we introduce existing methods for cleanliness classification as well as tasks and techniques related to our work.

2.1 Cleanliness Classification

There have been few works on the cleanliness classification. [1] focuses on cleanliness measure in restrooms, and [17] directly works on indoor cleanliness classification. They both apply principal component analysis (PCA) on feature extracted from convolutional neural networks to measure the cleanliness of a restroom. [1] also proposes a data augmentation algorithm and a PCA-based feature analysis schema to select the best suitable CNN architecture. As shown in Figure 2, [17] uses oversimplified dataset where the indoor images rarely appear in practice, thus only reflecting an oversimplified and biased notion of cleanliness.

2.2 Measuring Image Complexity

A number of studies have proposed various approaches to measure image complexity. [4] explores objective measures of complexity based on compression, and shows spatial information (SI) is closely related to compression-based complexity measures. [5] proposes a general method for measuring visual complexity by using perceptual similarity measurement and clustering. [6] also uses novel metrics based on the encoding size and compression error of JPEG and fractal compression to measure image complexity.

2.3 Edge Detection

Edge detection has been studied for many years and numerous methods have been proposed, such as Canny, Sobel, Laplacian operator. All these methods compute the measure of edge strength using first-order or second derivative. Recently, some edge detectors based on deep learning have been proposed to exploit more information, which achieve state of-the-art performance [7, 8]. Most of them use a deep CNN architecture, guided by deep supervision to performs image-to-image prediction.

2.4 3D Object Detection

Various 3D object detection methods based on RGB-D data have been proposed. [9] uses shape priors and occlusion patterns to infer the object 3D bounding boxes. [10] represents depth data in a 2D point map and applies 2D CNN to predict the bounding boxes in 2D images. [11] uses CNN on voxelized 3D grids in 3D point clouds. [12] converts point clouds into volumetric grids and uses 3D CNN to localize the object. Recently, [13] also proposes a 2D driven 3D object detection method.

2.5 Graph Convolutional Networks

[14] proposes graph convolutional networks (GCN) by generalizing convolution operation from grid to graph data. The main idea is to generate a node representation by aggregating its own features with features of neighboring nodes. There are a series of works following GCN, which propose various graph convolution operators for more expressive node embeddings, such as GraphSAG E [15], MPNN [16], etc.

Figure 2. Negative samples from dataset proposed in [17] rarely appear in practice.

3. METHOD

In this section, we first analyze the inadequacy of previous method and propose our own dataset to more truthfully reflect scene cleanliness problem. In our dataset, we observe people's subjective judgement on cleanliness is mainly based on visual complexity and object arrangement. For these two criteria respectively, we augment pretrained CNN feature with image contour feature, and apply 3D object detection and graph convolutional neural networks to model scene object arrangement. Our observation and module design are justified in ablation studies in section 4.3. We further integrate these modules and propose our COOL model for cleanliness classification.

3.1 Dataset Preparation

To the best of our knowledge, the only publicly available dataset for indoor cleanliness classification has a limited size of 200 images, and its negative samples are so messy that they could rarely appear in practice [17]. To reflect the notion of cleanliness in daily life, we collect 2264 indoor images from SUN-RGBD dataset by inviting 5 annotators to classify images as clean or messy, and use the mode of these results as labels to reduce variance in annotation.

3.2 Dataset Analysis

To compare the existing dataset and ours, we visualize their image distributions in Figure 3 by the first two principle components of abstract image feature obtained by Xception [18] pretrained on ImageNet. Clean and messy images appear closely mixed in our dataset, which suggests vanilla pretrained CNN feature may be inadequate to represent humans' subjective notion of cleanliness, and other criteria could get involved in people's subjective perception.

We empirically observe object arrangement may be another criterion for cleanliness classification in addition to abstract visual feature. We manually divide our dataset D into D_o and D_v. D_o relates to object arrangement, while D_v mainly reflects visual complexity. Overlapped images in both D_o and D_v are allowed. Then, we perform logistic regression on pretrained Xception CNN feature for the existing public dataset, our dataset D, D_o, and D_v. As a quick illustration, we down sample these datasets to obtain the same number of positive and negative samples. In Figure 4, we see CNN feature fails to capture object placement in the scene. And object arrangement plays an important role in our dataset for cleanliness classification, which justifies our empirical observation.

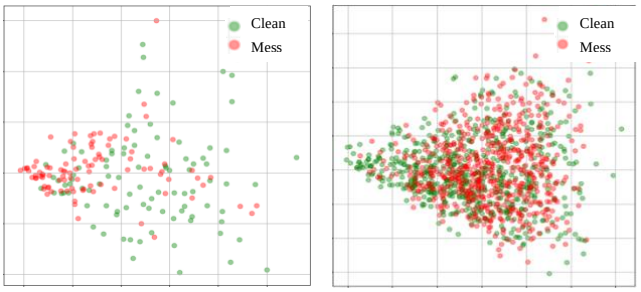

Figure 3. Data distribution visualization for previous work (Left) and our dataset (Right) using first two principle components of feature given by Xception pretrained on ImageNet.

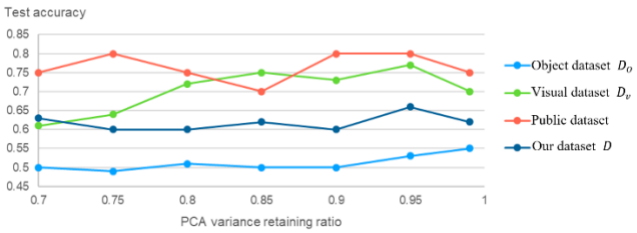

Figure 4. Performance of logistic regression on the principle components of Xception CNN feature pretrained on ImageNet. 10% images are used in testing.

The above analysis shows that

1. Object arrangement is a major criterion for human's notion of indoor cleanliness classification.
2. Abstract image feature extracted from pretrained CNN is inadequate to express scene cleanliness.
3. New model architecture is needed to capture object-level arrangement. Auxiliary image feature in addition to vanilla CNN is helpful to represent cleanliness.

3.3 Pretrained CNN Feature

As a baseline method, we use Xception [18] pretrained on ImageNet to obtain image feature, which is a complete representation generalizing from previous knowledge. To avoid overfitting, in practice we use principle component analysis (PCA) to further reduce feature dimension from 2048 to around 150 while retaining 95% variance.

We find current convolutional feature is biased in the following two aspects. On the one hand, CNN feature itself is incapable of object-level scene understanding. On the other hand, low-level feature vanishes under multiple convolution and pooling layers, but contour and texture may visually relate to cleanliness.

To compensate for convolutional feature, we additionally propose modules to model object arrangement and extract low-level contour feature and respectively.

3.4 Object Arrangement Feature

3.4.1 3D Object Detection

Ordered object arrangement is a main criterion for room cleanliness. Misaligned and cluttered furniture does not differ much from its organized counterpart in terms of color, contour or texture, but they have opposite cleanliness. The placement of main furniture constitutes the overall scene structure, and their arrangement representation is necessary for cleanliness classification.

Object arrangement is mainly contributed by main scene objects. According to the intuition, we only consider major objects categorized by table, chair, bed, and model it as a 3 dimensional one-hot vector. Then we can localize the main objects in indoor image and get the spatial information by using pretrained VoteNet [20] on SUN-RGBD. For each input image I, we obtain the bounding box B, centroid C, and orientation O of main scene objects in I:

$$B, C, O = VoteNet(I) \quad (1)$$

3.4.2 Object Arrangement Modeling

Scene objects are modeled as graph nodes with object class and size as attributes. Firstly, we use Delaunay triangulation to build the graph. We use volume of bounding box B to represent object size, and concatenate it with one-hot category vector as object feature h_v for node v. Graph edge expresses how a pair of objects is interacted and arranged. Edge feature e_{vw} between node v, w consists of the distance between two objects, the angle between source object orientation and their displacement vector, as well as the angle between two objects' orientations. We make above design choice to ensure the object graph is agnostic to translation and rotation, which increases model robustness.

[16] proposes Message Passing Neural Network (MPNN) framework based on GCN [14]. The forward pass contains a message passing phase and a readout phase. Message passing phase aggregates neighboring node features for multiple time steps $t = 1 \ldots T$. For each node v at time t, its feature at the next time step h_v^{t+1} is given by its current node feature h_v^t and the aggregated neighboring message m_v^{t+1}. Their interaction is modeled by node update functions U_t. In addition, m_v^{t+1} is determined by the sum over all neighboring messages, and each message is computed by a shared message functions M_t, which operates on current and neighbor node features h_v^t, h_w^t, together with their edge feature e_{vw}. The above updating rule is shown in equation 2, 3.

$$m_v^{t+1} = \sum_{w \in N(v)} M_t(h_v^t, h_w^t, e_{vw}) \quad (2)$$

$$h_v^{t+1} = U_t(h_v^t, m_v^{t+1}) \quad (3)$$

where $N(v)$ represents the neighbors of v in the graph.

The readout phase aggregates feature across all nodes at the last time step T for the global graph embedding H with some readout function according to

$$H = R(h_v^T \mid v \in G) \quad (4)$$

In our instantiation, the inputs of M_t and U_t are the simple concatenation of inputs, and the functions M_t, U_t themselves are both one fully connected layer with activation function.

In readout phase, we apply the same fully connected layer with activation, denoted by FC in equation 5 and 6, on each node feature h_v^T. Then apply max pooling on each dimension over all nodes to readout the global embedding H in the graph:

$$H = MaxPooling[FC(h_v^T \mid v \in G)] \quad (5)$$

The global graph embedding H learns effective feature representation of scene object arrangement. We may use a final fully connected layer to make a prediction y:

$$y = FC(H) \quad (6)$$

3.5 Low-Level Contour Feature

3.5.1 Edge Detection

We use low-level contour feature to compensate for pretrained convolutional feature to achieve better cleanliness classification.

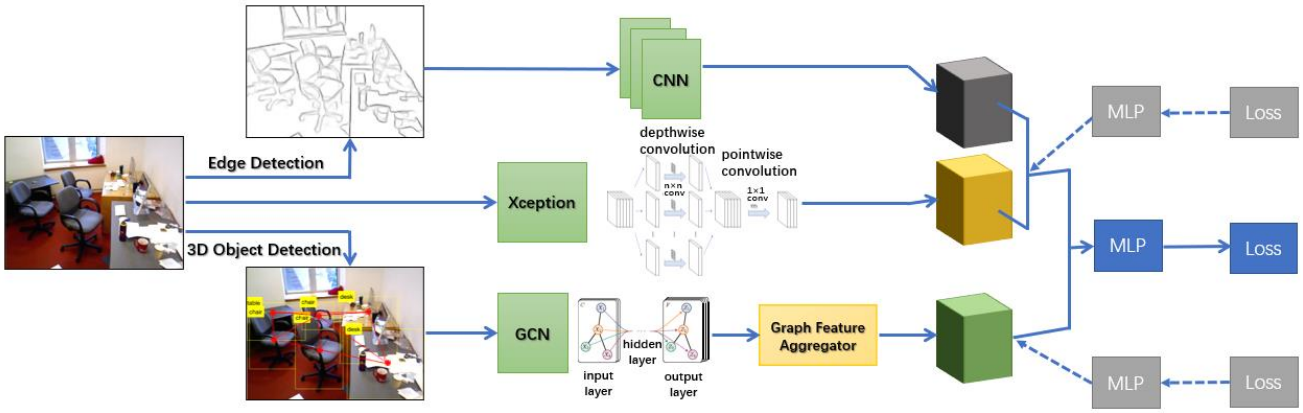

Figure 5. COOL model architecture for cleanliness classification.

Deep learning learns abstract, low frequency signal from data that exhibits overall statistical property, while filters out high frequency signal since it is statistically similar to data noise. However, cleanliness could directly relate to visual details, such as dirt, texture, and contour, which belong to high frequency signal. Therefore, we apply low-level contour detection with shallow convolution and pooling to preserve detailed low-level visual feature.

We apply contour detection for indoor images using method proposed in Edge Boxes [19]. Given the input image I, we use edgebox to generate the contour image I_e of I.

3.5.2 Contour Feature Extraction

Having the contour information of each input image, we use convolutional neural networks with $tanh(\cdot)$ activation function to find the local patch clutter in the image:

$$U^0 = tanh[conv(I_e)] \quad (7)$$

$$U^i = tanh\left[conv\left(U^{i-1}\right)\right] \quad i = 1 \ldots \mathrm{L} \quad (8)$$

We also apply mean pooling operator to down sample the intermediate result, and flatten it to obtain final contour feature representation C.

4. EXPERIMENTS

4.1 Datasets

We construct our own dataset based on SUN-RGBD, which contains 10,335 densely annotated RGB-D images including 3D bounding boxes with object orientations. Our dataset consists of 2264 manually classified images for cleanliness classification. As described in section 3.2, our dataset is split into object arrangement subset and visual complexity subset D_o, $\mathrm{D_v}$ with 1193, 1332 respectively. Overlapped images are allowed.

4.2 Implementation Details

Our object arrangement module includes 3 graph convolutional layers followed by a feature aggregation across all nodes. Each graph convolutional layer contains a single fully connected layer with 16 hidden neurons and tanh activation. The bias in graph convolution is disabled. For global feature aggregation, we use one fully connected layer with 16 hidden neurons and ReLU activation to operate on each node, and do max pooling across all nodes. We use 3D object bounding box annotated in SUN-RGBD directly, because the 3D object detection isn't the core problem we are addressing. As stated in section 3.4.1, one can also do 3D object detection with existing methods.

For contour feature extraction module, after a 4×4 average pooling, we apply two convolutional layers with 4 channels, kernel size 3, stride 3, and without bias, followed by batch normalization and ReLU activation. In addition, one 2×2 max pooling goes between two convolutional layers and one 2×2 average pooling is applied in the end. The output contour feature is a 4-dimensional vector to be concatenated with the pretrained CNN feature.

4.3 Experiment and Analysis

We first present the experiments results for object arrangement feature and low-level contour feature on D_o and D_v to show their respective effectiveness. And then compare the cleanliness classification results of the integrated model on our dataset with the method used in previous work to show that our model design provides a more truthful modeling for the notion of cleanliness.

The models are trained using balanced sampler to handle data imbalance in classification problem, hence the presented accuracy can reasonably reflect model performance.

4.3.1 Object Arrangement

For object arrangement feature learning, we compare the graph convolutional networks proposed in section 3.4.2 with logistic regression used in previous method. The evaluation is performed on object dataset D_o to highlight the capability for object arrangement modeling. Our proposed graph convolutional networks module captures object arrangement feature that is unavailable for CNN feature.

Table 1. Performance of GCN module on D_o

Model	Proposed GCN	Logistic Regression
Accuracy	72.5%	52.1%

4.3.2 Low-Level Contour

For low-level contour experiment, we compare the performance between the presence and absence of contour feature extraction on top of pretrained CNN feature. The evaluation is performed on visual dataset $\mathrm{D_v}$ for comparison clarity. To enable the interaction between CNN feature and contour feature, we first use fully connected layer to embed the CNN feature after PCA into 16-dimensional vector, and then concatenate it with the extracted contour feature of 4 dimensions, followed by logistic regression. Because we introduce an additional hidden layer before output, we consider one more baseline model using vanilla pretrained CNN feature with a 16-dimensional hidden layer.

Table 2. Performance of contour feature augmentation on D_v

Model	Contour Feature	Logistic Regression	One Hidden Layer MLP
Accuracy	81.5%	77.4%	79.1%

4.3.3 Integrated Model

After justifying our feature analysis and design choice for both contour and object-oriented aspects. We integrate these two parts onto pretrained CNN feature to propose our model COOL for indoor cleanliness classification. COOL is compared with previous method on the dataset we create. Because general visual complexity and object arrangement are two complementary criteria for cleanliness, training COOL directly leads to unclear credit assignment for the two criteria. We introduce auxiliary guidance by using object arrangement feature alone to predict training images in D_o , and similarly using visual complexity feature to predict training images in D_v. These auxiliary losses L_o, L_v serve as regularization for COOL to stabilize training, and our total training loss is given in Equation 9.

$$L_{COOL} = L + \lambda_o L_o + \lambda_v L_v \quad (9)$$

Table 3. Performance of COOL on our proposed dataset *D*

Model	COOL	COOL without Auxiliary Loss	Logistic Regression
Accuracy	81.5%	68.6%	75.5%

As is illustrated in Table 3, COOL significantly outperforms the baseline method using auxiliary feature with $\lambda_o = \lambda_v = 0.1$.

5. CONCLUSION AND FUTURE WORK

In this paper, we focused on indoor scene cleanliness classification. We went beyond from previous work on cleanliness classification by overcoming their limitations. We proposed our own cleanliness classification dataset based on SUN-RGBD, and analyzed two major criteria for human's subjective notion of cleanliness, namely object arrangement and visual complexity. We proposed Contour and Object-orient Learning (COOL) to capture scene object arrangement and augmented pretrained CNN feature with contour feature extraction. We justified our feature analysis and model design in experiments for each module, and achieved improved performance on our dataset. Our work shows human's subjective notion can involve diverse aspects of feature. Looking forward, we would design new model architecture to better model cleanliness.

6. REFERENCES

[1] L. Jayasinghe, N. Wijerathne, C. Yuen and M. Zhang, "Feature Learning and Analysis for Cleanliness Classification in Restrooms," in IEEE Access, vol. 7, pp. 14871-14882, 2019.

[2] M. Cardaci, V. Di Ges`u, M. Petrou, and M. E. Tabacchi. A Fuzzy Approach to the Evaluation of Image Complexity. Fuzzy Sets and Systems, 160(10):1474–1484, 2009.

[3] Ruth Rosenholtz, Yuanzhen Li, Lisa Nakano; Measuring visual clutter. *Journal of Vision* 2007;7(2):17.

[4] H. Yu and S. Winkler, "I mage complexity and spatial information," *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*, Klagenfurt am Worthersee, 2013, pp. 12-17.

[5] T. Guha and R. K. Ward, "Image Similarity Using Sparse Representation and Compression Distance," in *IEEE Transactions on Multimedia*, vol. 16, no. 4, pp. 980-987, June 2014.

[6] Juan Romero, Penousal Machado, Adrian Carballal & Antonino Santos (2012) Using complexity estimates in aesthetic image classification, *Journal of Mathematics and the Arts*, 6:2-3, 125-136.

[7] Xie, Saining, and Zhuowen Tu. "Holistically-nested edge detection." *Proceedings of the IEEE international conference on computer vision*. 2015.

[8] Yu, Zhiding, et al. "Casenet: Deep category-aware semantic edge detection." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[9] Mousavian, Arsalan, et al. "3d bounding box estimation using deep learning and geometry." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2017.

[10] Li, Bo, Tianlei Zhang, and Tian Xia. "Vehicle detection from 3d lidar using fully convolutional network." *arXiv preprint arXiv:1608.07916* (2016).

[11] Engelcke, Martin, et al. "Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks." *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.

[12] Song, Shuran, and Jianxiong Xiao. "Deep sliding shapes for amodal 3d object detection in rgb-d images." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016.

[13] Qi, Charles R., et al. "Frustum pointnets for 3d object detection from rgb-d data." *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2018.

[14] Kipf, Thomas N., and Max Welling. "Semi-supervised classification with graph convolutional networks." *arXiv preprint arXiv:1609.02907* (2016).

[15] Hamilton, Will, Zhitao Ying, and Jure Leskovec. "Inductive representation learning on large graphs." *Advances in Neural Information Processing Systems.* 2017.

[16] Gilmer, Justin, et al. "Neural message passing for quantum chemistry." *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017.

[17] Guanqiao Ding. (2019, March). Messy vs Clean Room. Retrieved November 20, 2019 from https://www.kaggle.com/cdawn1/messy-vs-clean-room

[18] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017.

[19] Zitnick, C. Lawrence, and Piotr Dollár. "Edge boxes: Locating object proposals from edges." *European conference on computer vision*. Springer, Cham, 2014.

[20] Qi, Charles R., et al. "Deep Hough Voting for 3D Object Detection in Point Clouds." *arXiv preprint arXiv:1904.09664* (2019).