PoB: Toward Reasoning Patterns of Beauty in Image Data

Diep Thi Ngoc Nguyen*
University of Engineering and Technology
Vietnam National University
Hanoi, Vietnam
ngocdiep@vnu.edu.vn

Hideki Nakayama
National Institute of Advanced
Industrial Science and Technology
The University of Tokyo
Tokyo, Japan
nakayama@ci.i.u-tokyo.ac.jp

Naoaki Okazaki
National Institute of Advanced
Industrial Science and Technology
Tokyo Institute of Technology
Tokyo, Japan
okazaki@c.titech.ac.jp

Tatsuya Sakaeda
National Institute of Advanced
Industrial Science and Technology
Tokyo, Japan
daesaka@gmail.com

ABSTRACT

Aiming to develop of computational grammar system for visual information, we design a 4-tier framework that consists of four levels of 'visual grammar of images.' As a first step of realization, we propose a new dataset, named the PoB dataset, in which each image is annotated with multiple labels of armature patterns that compose the pictorial scene. The PoB dataset includes of a 10,000-painting dataset for art and a 4,959-image dataset for photography. In this paper, we discuss the consistency analysis of our dataset and its applicability. We also demonstrate how the armature patterns in the PoB dataset are useful in assessing aesthetic quality of images, and how well a deep learning algorithm can recognize these patterns. This paper seeks to set a new direction in image understanding with a more holistic approach beyond discrete objects and in aesthetic reasoning with a more interpretative way.

CCS CONCEPTS

• **Computing methodologies** → **Scene understanding**; **Computer vision problems**; • **Applied computing** → **Fine arts**;

KEYWORDS

Image dataset; Armature pattern; Composition; Visual grammar; Aesthetic assessment

ACM Reference Format:
Diep Thi Ngoc Nguyen, Hideki Nakayama, Naoaki Okazaki, and Tatsuya Sakaeda. 2018. PoB: Toward Reasoning Patterns of Beauty in Image Data. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3240508.3240711

*Corresponding author.

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10.
https://doi.org/10.1145/3240508.3240711

1 RECALL THE DREAM

Once upon a time, David Marr [21] wondered how our eyes see and developed a general framework for understanding visual perception in which the construction of the visual world has three stages: 'primal sketch' (e.g., with edges, regions), $2\frac{1}{2}D$ sketch (e.g., with depth), and $3D$ scene. Marr with this elegant inversion to imaging process of the functionality of our visual system has pioneered the study of computer vision ever since[14, 33]. We have seen the development of computer vision from 'low-level image processing' algorithms (e.g., edge detection, segmentation, feature extraction) to 'object' detection, tracking and recognition, to 'spatial analysis' and 'scene understanding', to currently 'visual question-answering' and 'image captioning'. Besides computational modeling of visual information, many researchers have tried to assess the beauty based on Marr's framework (i.e., using describing features of images).

Once in a while, Leland Wilkinson [36] figured that a pie chart and a divided bar are only different in the coordinating systems (i.e., polar versus Cartesian coordinates). By realizing the importance of graphical rules over the appearance of graphics, he emphasized on a development of a system of grammar for graphics. "The grammar of graphics takes us beyond a limited set of charts to an almost unlimited world of graphical forms [36]."

Now and then, while we follow Marr's framework, we reflect on Wilkinson's grammar. Visual objects are like words in natural languages, which needs a system of grammar to construct a rich and seemingly unlimited set of meaningful sentences. We come to believe that the visual world also has such a system of grammar to organize and make 'stories' for objects in pictures.

We propose a 4-tier framework to realize a computational system of visual grammar, which answer the following questions:

- Geometrical grammar: Given a 2D blank rectangular frame, is there any location in this frame which is more importantly perceived than others?
- Armature-based grammar: Are there typical patterns of composing objects in pictures?
- Perspectival grammar: What makes a picture look realistic (i.e., logical representation of the physical world)?

- Contextual/Semantic grammar: What makes aesthetic values of images? Is there contradictory meaning of objects given a context? and so on.

"Human's earliest ambition is to create," but there is nothing that gets out of nothing [15]. Like grammaticalization happens all the times in languages [3], grammar can be found as useful practices. We approach to visual grammar for images with the same manner, by analyzing frequent practices in fine arts, photograph, and design.

In this paper, we particularly focus on the 'armature-based grammar,' which are common rules to compose a picture. The main contributions of this paper are:

- A new image dataset, the PoB dataset[1], which emphasizes on a new aspect of scene understanding: a holistic construction of pictures;
- A demonstration of the usefulness of armature patterns in aesthetic assessment, in which we could achieve higher prediction accuracy using additional armature information;
- A promising classification learning model for armature patterns using a fine-tuned convolutional neural network.

In Section 2 and Section 3, we discuss how different definitions of 'composition' have been used in current literature and how we collect and annotate the PoB dataset as well as analyze the labeling consistency. In Section 4 and Section 5, we demonstrate the feasibility of the armature information in classification problems. We conclude our findings after some discussions about limitations and future works of our current proposal.

2 LITERATURE REVIEW

Many studies have realized the importance of compositional information in image understanding in additional to low-level features such as [1, 7, 9, 12, 18, 20, 26]. Their description of 'composition' from the use belongs to either photographic rules or other 'low-level' features. In details, they are: simplicity of the scene, salient object, size of object along with its relative brightness [9, 18, 26]; rule of thirds [1, 7, 12, 26]; rule of fifths [12]; golden ratio [1]; golden mean or golden triangle [26]; visual weight balance [1, 26]; viewpoint [12]; wavelet-based texture [7]; photographic rules like low depth of field, color contrast, lighting contrast [9, 18]; and CNN-based features (implicit composition) [4, 20].

We argue that those 'composition' information is not sufficient in giving a holistic way of how a picture is composed. A focal object can be placed at the 'third-points' of a picture by several ways: framing using edges of pictures, using a leading line that guides human eyes, or simply by making it really large. Additionally, those photograph-inspired rules may meet difficulty in assessing aesthetic value of other non-representative arts (e.g., decorative art, conceptual art or abstract art).

To avoid confusion with the use of 'composition' in current computer vision literature, we therefore use 'armature' to indicate how lines and regions and objects are organized in a picture. Artist Robert [31] also suggests using this term to denote the "backbone" of a picture.

Regarding datasets in computer vision, we have seen many which for general objects or domain-specific objects or attributes datasets (e.g., medical, traffic, surveillance, fingerprints, textures, faces, etc.)[2] or multi-modal datasets of text and images. However, those datasets emphasize on objects and attributes in the images rather than the holistic construction of the image. Our PoB dataset approaches from a more structural way to understand the image in order to realize a system of grammar for image data.

3 MAKING OF POB DATASET

3.1 Armature description

We refer to famous practices of composing pictures in fine arts [10, 25, 27, 31] and carefully compare and combine them into the most 15 frequently used armature patterns. Our selected 15 patterns are: 'O', '/', 'L', 'S', 'R', '+', 'Z', 'II', 'Y', 'X', '_', 'C', 'M', '∧', and 'V', in which they can be grouped into four groups based on their functionality in emphasizing a focal point in a picture. Four groups are: (1) *framing* ('O', '/', 'L'); (2) *leading line* ('S', 'R', '+'); (3) *repeating* ('Z', 'II', 'Y', 'X'); and (4) *space & mass* ('_', 'C', 'M', '∧', 'V').

The description of these 15 armature patterns is as follows:

Circular framing (O) : Using three or four edges of a picture to frame the objects of interest at the center.

Diagonal (/) : An edge of a object or a series of lines running along two main diagonals of a picture to create a dynamic atmosphere.

Ell (L) : Using two perpendicular edges of a picture plane to create a L-shape frame that either surrounds or holds the objects of interest.

S or compound curve (S) : Using a S-shape curve to lead the eyes to the objects of interest or to create dynamic in a picture.

Radiating (R) : Using lines that converge to a point that emphasizes the objects of interest. Frequently used in one-point perspective drawing.

Cross (+) : Vertical lines cross horizontal lines.

Horizontal overlapping (Z) : Spreading of horizontal regions from bottom of a picture.

Repeated vertical (II) : Many vertical objects that spread from left to right or by depth of a picture.

Pattern repeating (Y) : Repeating of a same pattern to emphasize the pattern.

Symmetry/Reflection (X) : Symmetry over a vertical or horizontal line.

Extreme horizon (_) : Using an extreme low/high horizon line.

Covering (C) : An elevated view from above in which there is a curved shape (e.g., full or half circle, eclipse shape) that surrounds a picture.

Group mass (M) : A macro view of the object of interest in a picture.

Pyramid (∧) : Composed of a flat plane at the front side (bottom of a picture) and a triangular object perpendicular to the plane at the back side (top of a picture).

Three-spot or triangle (V) : Using relative relations between three objects to create a balancing sense in a picture.

[1]The full dataset and related data are published at https://github.com/chupibk/PoBDB_Patterns_of_Beauty.

[2]See http://homepages.inf.ed.ac.uk/rbf/CVonline/Imagedbase.htm for an extended list of image databasets.

Table 1: Armature patterns in practices and our proposal

Author	# of armatures	Proposed armatures
Payne [27]	15	O, /, L, S, R, +, II, Y, X, _, M, ∧, V, steelyard, silhouette
Roberts [31]	9	S, L, /, R, triangle, fulcum, O, portrait, +
Nus [25]	14	Z, S, +, L, triangle, M, steelyard, V, Y, R, U-shape, O, /, tic-tac-toe
Dow [10]	5	opposition, transition, subordination, repetition, symmetry
PoB (ours)	15	O, /, L, S, R, +, Z, II, Y, X, _, C, M, ∧, V

Table 1 shows a comparison of our proposed 15 patterns and armature patterns in conventional practices. We argue that the patterns which are in other practices but not in ours can be transformed into ours. For example, 'steelyard' and 'U' are similar to 'O' as 'triangle' to 'V' or 'R' and 'fulcrum' to '+'. 'Portrait' and 'tic-tac-toe' should not be armature patterns. In a particular practice [10], the five ways of composing a picture can be considered as a highly abstract description of our proposed patterns.

Each of 15 armature patterns in our PoB dataset has a representative shape or 'template' as shown in Table 2 but images that have the same armature does not necessarily have the same shape. The shape can vary by some geometric transformation such as ±90° rotation, reflection, translation, dilation depending on the type of the pattern.

3.2 Data Collection and Annotation

3.2.1 Datasets. In order to demonstrate that armature patterns can be seen not only in photographs but also widely used in paintings, we collect two sets of images for creating PoB: photographs as a subset of the AVA dataset [23], and paintings from Wikiart [30]. Pictures are supposed to depict a clear story by highlighting clear focal points. Armature patterns are about the "wholeness" in the story. If a picture has a too strong figure (e.g., a person, familiar objects), the figure is always recognized firstly and becomes the focal point. That affects greatly how the picture will be composed. To amplify the recognition of the armature patterns themselves, we selected the 'landscape' image category because it is "natural", "figure-less", and can be collected by a large scale.

The photographs from the AVA dataset are selected from 'landscape' image list (4,959 images) and will be referred in our PoB dataset as the 'Photograph' dataset. The paintings from Wikiart are also from 'landscape' genre[3]. We implemented a crawling tool to download about 15,000 images. After that, we manually removed images which have low resolution, long height, circular frame, multiple images in one image, text border or which are too small or too dark. Finally, the 'Painting' dataset in our PoB dataset has 10,000 images. We will refer to our PoB dataset as 'PoB dataset' or 'PoB datasets' interchangeably.

3.2.2 Annotators. It's known that art-trained viewers and untrained viewers see art differently such as in [24, 29, 34]. Those studies show art-trained viewers are more attracted to the overall (structural and highly abstract) composition of a picture rather than features or objects which are in the picture. In making of the PoB dataset, we hire only annotators who are potentially sensitive to compositional

[3]https://www.wikiart.org/en/paintings-by-genre/landscape.

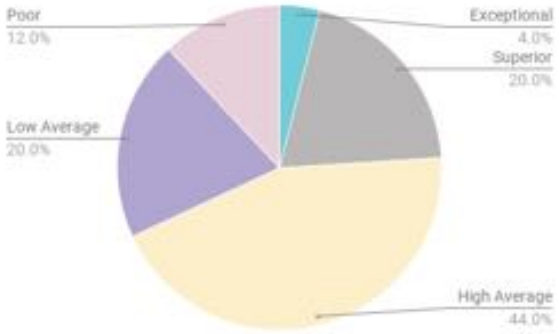


Figure 1: Distribution of Meier-Seashore art judgment test scores of 25 candidates for the PoB annotators.

designs. We use the Meier-Seashore art judgment test [22] to select annotators from candidates.

We asked 25 candidates with varying art background and professions (non-art students, art students in design, fine art, architecture, software engineers, accountants, managers, researchers) to complete the Meier test via a Web interface. The art training measure is collected by responses of each annotator for seven questions that follow the practice in [17].

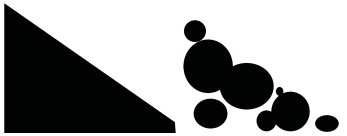
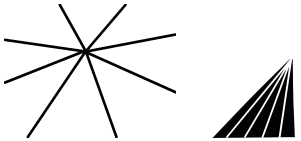
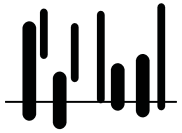
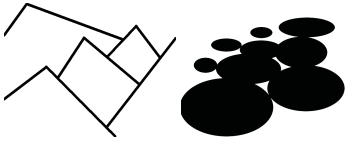
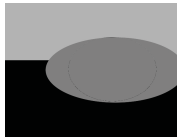
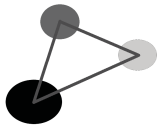
Each candidate answered 125 questions by select an image which has a 'better' composition out of each pair of non-representative images. The answers are compared to a golden list of results. The correct responses are counted, then divided by 125 to return a percentage score. The Meier test also provides a reference table to transform a percentage score into a reference rank. There are 6 ranks: 'exceptional' (85-100%), 'superior' (77-84%), 'high average' (70-76%), 'low average' (63-69%), 'poor' (53-62%), and 'zero score' (below 52%).

Figure 1 shows the distribution of scores of 25 candidates. It suggests a rarity degree to find people who are highly sensitive to good composition. Though it requires further investigation to make a stronger conclusion like [24, 29, 34], the data from answers of 25 candidates can be useful to many interested researchers. We publish them together with our PoB dataset.

We finally selected two candidates from two highest score classes as our annotators of the PoB dataset. The candidate from 'exceptional' class will be referred as 'senior', and 'superior' as 'junior'. The senior annotator has a higher art training background.

3.2.3 Annotation interface. We implemented a Web interface as shown in Figure 2 for annotating the images. We configured the

Table 2: 15 armature patterns in our PoB dataset and their typical template images and sample images. Undecidable ('Unk') refers to unclear or confusing composition.

Armature	Armature name	Armature template(s)	Sample image*	Armature	Armature name	Armature template(s)	Sample image*
O	Circular framing			/	Diagonal		
L	Ell			S	Compound curve		
R	Radiating			+	Cross		
Z	Horizontal overlap-ping			II	Repeated vertical		
Y	Pattern repeating			X	Symmetry/ Reflection		
–	Extreme horizon			C	Covering		
M	Group mass			∧	Pyramid		
V	Three-spot or triangle			Unk	(undecidable)		

(*) Sample images are in public domain on Wikiart (https://www.wikiart.org/)

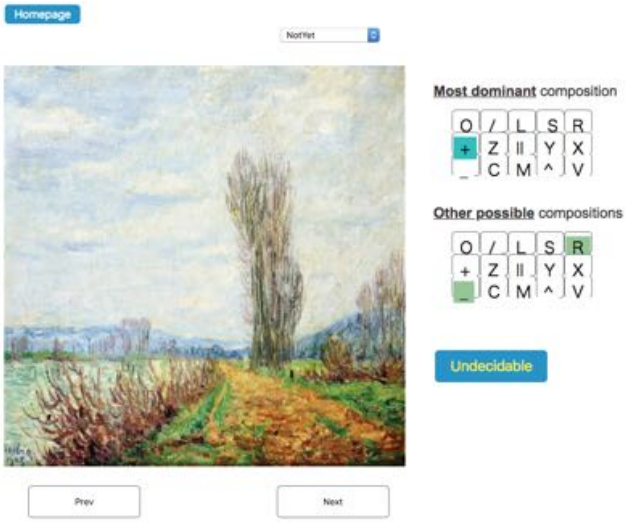


Figure 2: Web interface sample for annotating the 'most dominant' and 'other possible' armatures of an image. The image is in public domain on Wikiart (https://www.wikiart.org/).

annotation process as a multi-label annotation problem. However, instead of selecting several labels as equally important, we ask each annotator to select the 'most dominant' and 'other possible' compositions among our 15 pre-defined patterns for each image. If an annotator can not decide a pattern out of our 15 armature patterns, she selects 'Undecidable'. This case happens when an image is too low-quality, or its composition is unclear or confusing.

Display devices are personal computers (laptops). The annotators used the same web interface which controls the same display proportion of images. Other factors like color, resolution, distance to the eyes are not currently controlled.

3.2.4 Annotating process. We followed the annotation practices in PASCAL VOC2007 [11] to achieve consistency in annotation process by letting all annotation take place at a single place after following a set of guidelines, that includes: armature definitions, simple and difficult examples, how to decide the dominant composition. We trained annotators and let them annotate first 200 images. Then we collected the results, checked the agreement by a simple intersection operation, then we adjusted the guidelines until all annotators agree. This process was repeated about 4 times. After this, the annotators continued to annotate the whole dataset. Averagely, each annotator is required to work on an image from 30 seconds to 1 minute. However, one shouldn't continuously annotate images for a long time. Therefore, it took each annotator about 15 to 20 days (6 to 8 hours per day) to (only) annotate the PoB dataset (14,959 images).

Table 2 shows image examples of 15 armature patterns defined in Section 3.1 and one case where the image was selected as 'Undecidable' (*Unk*) by an annotator. We can see in this table, the 'Unk' image has an unclear composition since it shows only random distribution of colors while other images can be recognized by their corresponding compositional templates.

Figure 3 shows the distribution of selection frequency of two annotators. At a very first glance, we see (1) imbalance between patterns, and (2) different selection frequency between two annotators. The *leading line* and *repeating* patterns seem to be used widely in both Painting and Photograph datasets, though slightly more frequently in Painting. The 'O' framing and '_' space armatures also appear very often in both datasets. Interestingly, the '_' (*extreme horizon*) armature appear more frequently in Photograph than in Painting. This can be explained by its intrinsic simplicity in composition, which is very preferred in photography.

In Figure 3, we also see the higher imbalance in *junior* than in *senior*. Though there is no "truth" distribution to decide which is correct, we may prefer a more balancing distribution of *senior* or distribution that compensates for small classes since it gives more information of the images. We intend to investigate this in our future work.

3.3 Annotation Consistency Analysis

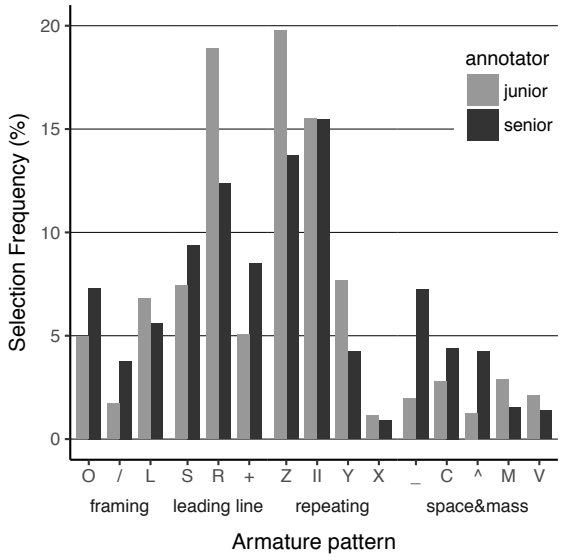
We used the Phi coefficients [13] to calculate the consistency degree of two annotated label variables. A Phi coefficient measure the degree of association between two binary variables. In this paper, Phi coefficients are calculated by two settings: (1) label-wise inter-annotator, and (2) image-wise inter-annotator. The first setting investigates the consistency of labeling while the second setting studies the *difficulty* of images to annotate. A Phi coefficient varies from -1 to $+1$. A value of -1 says very strong negative association and a value of $+1$ says very positive association between two variables. We expect positive association to conclude the consistency in our annotated datasets. We use Matthews correlation coefficient implement of Sklearn library [28] for calculating the Phi coefficients.

In the first setting, for each armature label A, we can calculate a Phi coefficient between two lists of annotated labels of two annotators, X_1 and X_2 where $X_i = (I_j, x_{ij})$ where $x_{ij} \in \{0, 1\}$ is the annotated label of the image I_j so that $x_{ij} = 1$ when the annotator i selected the corresponding label A.

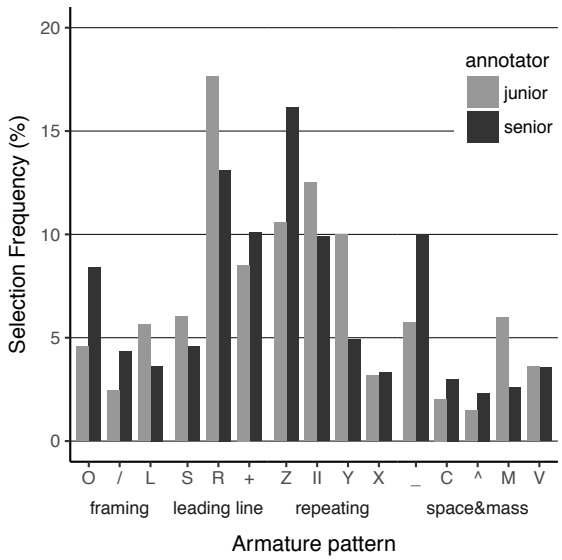
In the second setting, for each image I, we calculate a Phi coefficient between two lists of annotated labels of two annotators, Y_1 and Y_2 where $Y_i = (A_j, x_{ij})$ where $x_{ij} \in \{0, 1\}$ is the annotated label of the armature A_j so that $x_{ij} = 1$ when the annotator i selected the label A for the corresponding image I.

Figure 4 shows a value heatmap of Phi coefficients by armature labels between our annotators in the PoB datasets. In general, the values along the main diagonal line are the highest values. which suggests a high consistency in labeling of armature patterns. However, the Phi coefficients for infrequent patterns (refer to Figure 3) seem to be lower. This infrequent phenomena also appeared in a newspaper corpus [2]. Nevertheless, a better handling of these infrequent patterns may improve the consistency degree.

Figure 5 shows the histogram distributions of Phi coefficient values by images between two annotators of our PoB datasets. We see a higher consistency degree in labeling of images in Painting dataset than in Photograph. This can be due to two factors: (1) the

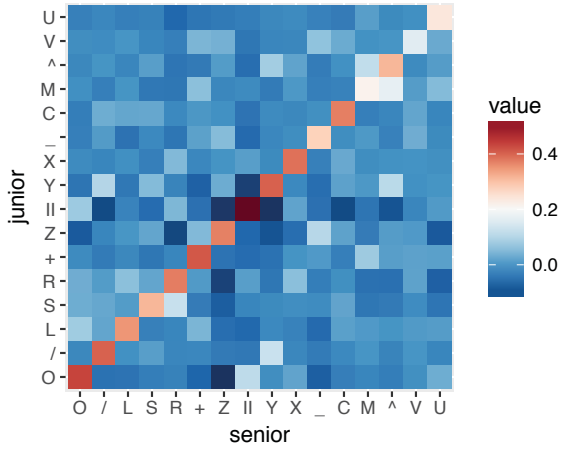


(a) 'Painting'

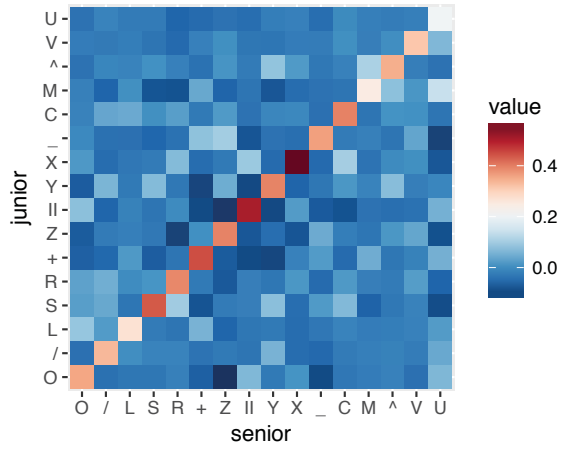


(b) 'Photograph'

Figure 3: Selection frequency of 15 armature patterns of two annotators in our PoB datasets.



(a) 'Painting'



(b) 'Photograph'

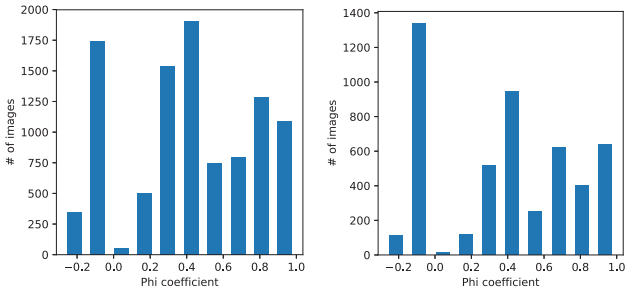
Figure 4: A value heatmap of Phi coefficients by armature labels between 'senior' and 'junior' annotators in our PoB datasets.

production of paintings often requires a clear mindset in composition, while photographs can be spontaneous; and (2) objects in paintings often are removed if needed in order to emphasize the compositions that result in more unambiguous compositions.

4 ARMATURES AND AESTHETIC VALUE

We investigate the aesthetic value of armature patterns by demonstrating its usefulness in aesthetic assessment. The Photograph dataset in our PoB inherits the aesthetic rating scores from the AVA dataset [23]. There have been many efforts in predicting aesthetic quality of images based on this dataset by a classification problem of 'low' or 'high' aesthetic value. We also selected a threshold of 5 to separate the images [16, 23].

We followed current practices of extracting deep-learning features for images then applying classification learning model such as in [16, 19, 35]. We used the pre-trained VGG19 model on ImageNet dataset [32] on Keras deep learning library [6] for extracting a 4096d vector for each image. We used the standard split of train-test data as in the original AVA dataset [23]. In addition to these features, each image has a 16d vector information that is a binary encoding vector of 15 armature patterns and the last element for 'Undecidable' case. Each 16d vector is created by an intersection operation



(a) 'Painting' (b) 'Photograph'

Figure 5: Histogram distributions of Phi coefficient values by images between the annotators in our PoB datasets.

Table 3: Accuracy of aesthetic classification on the AVA dataset and its subset 'landscape' as the 'Photograph' in PoB.

Method	Accuracy on test (%)	Cross validation accuracy (%)
AVA-baseline	70.94	-
AVA-CNN [8, 35]	78.08	-
AVA-CNN + semantic [16]	79.08	-
PoB-baseline	76.99	76.99
PoB-VGG19	75.45	72.17
PoB-armature	**77.19**	**77.19**
PoB-VGG19+PCA	78.27	78.86
PoB-VGG19+armature	**79.12**	**79.25**

between two binary vectors of two corresponding annotated results for the image.

Table 3 shows the accuracy of the aesthetic classification using only VGG features or when combining with the armature information from our PoB dataset. *Baseline* is calculated as the percentage of the most frequent label in the test set. It shows a competitive accuracy even when solely using armature information (*PoB-armature*), and a better accuracy comparing to image information alone. This may imply that the armature information already encodes information about images which are related to their aesthetic values. By applying a simple dimension reduction method (PCA) to the original VGG features, we achieved higher accuracy. By combining VGG features and armature information, we yielded the higher accuracy than the state-of-the-art accuracy which uses complicated semantic information [16] for the whole AVA dataset. This indicates that the armature information can considerably complement to highly contextual information from image features such as semantic textual features.

We conclude that the armature information are highly associated with aesthetic value of images. By recognizing armature patterns in images, we not only know how the pictures are composed but also promisingly have capacity to interpret the elements of beauty in the pictures.

Table 4: Classification accuracy (%) of armature patterns in PoB datasets of different methods using fine-tuned Xception CNN models.

Dataset	Method	Baseline	Top 1	Top 3
PoB photograph	whole	17.30	46.57	78.00
	combine_cls	39.43	57.71	**96.86**
	del_difficult_cls	20.92	51.67	82.67
	del_small_cls	21.62	**58.80**	91.20
	del_phi_cls	18.55	19.33	51.33
	del_phi_image	17.27	17.71	47.14
PoB painting	whole	23.44	47.18	78.35
	combine_cls	46.14	**57.88**	**96.37**
	del_difficult_cls	25.09	50.37	80.12
	del_small_cls	24.25	48.35	78.71
	del_phi_cls	24.39	24.71	62.00
	del_phi_image	23.39	23.53	59.53

5 AUTOMATIC CLASSIFICATION LEARNING OF PATTERNS

In this section, we introduce preliminary learning results in recognizing armature patterns in images. The problem is configured as a classification of an image into armature patterns.

We implemented a fine-tuned CNN model from a pre-trained Xception model [5] using Keras library [6]. We added two fully connected layer (1024 and 15) and an activation layer after pooling the output of the Xception model. We froze the layers of Xception model for some epochs (5) then unfroze the last 25 layers and trained the network for about 30-50 epochs.

We configured several settings for the data: (1) whole datasets, (2) combining 15 armature patterns by 4 functionalities (refer to Section 3.1), (3) removing (manually assessed) *difficult* armatures ('Y', 'M', 'V'), (3) removing *infrequent* armatures (< 5%), (4) removing *low-Phi coefficient* armatures (< 0.3), and (5) removing images with low Phi coefficients (< 0.3).

Table 4 shows the classification accuracy results. *Baseline* is calculated as the percentage of the most frequent label in the test set. Using the whole datasets, accuracy for Painting is higher than for Photograph. This may be because of the amount of data for training is almost double. Combining classes or deleting difficult or infrequent classes increase the accuracies. Surprisingly, deleting classes or images based on Phi coefficients decrease the accuracies significantly.

6 DISCUSSION

In Section 4 and Section 5, we have demonstrated the usefulness of armature information and the promising results for recognizing armature patterns in images. While those findings are encouraging, we acknowledge several problems that are still remaining in our current approach.

Firstly, the PoB dataset is constructed based on two annotators due to human and financial resources. This is the main limitation of our dataset that causes a difficulty in concluding the consistency analysis by any statistical significance. However, as a preliminary

work on realization of a visual grammar, we believe the PoB dataset can be useful in some ways. Also, recognizing armature patterns of images is not a simple task as recognizing objects since human already have knowledge of objects. We intend to hire more annotators in future.

Secondly, the next limitation is our approach in Section 5 for classification problem. We currently treat images as squared images, which can lead to an distorted composition. A further investigation of this phenomenon can be referred to [20].

We also work only with landscape-genre images. Other genres may use the same 15 armature patterns but it can be interesting to investigate other genres and discover different pattern distributions. With our knowledge, we believe the proposed 15 patterns should cover almost cases of fine arts.

Despite those limitations, we suggest potential applications using armature patterns that explore and understand deeply the visual world. For example, we can develop an armature-aware image cropping; we can use the armatures to actually compose an image (promisingly valuable to GAN image generation researches); or we can detect abnormals of focal point placement for each composition patterns.

7 CONCLUSION

By acknowledging the need for a new grammar-based approach to understanding visual data, we have proposed a 4-tier framework to realize a computational system of visual grammar that consists of geometrical grammar, armature-based grammar, perspectival grammar, and contextual/semantic grammar.

In this paper, we particularly focused on the 'armature-based grammar,' which are patterns to compose a picture. We introduced a new image dataset, the PoB dataset, which is a dataset of armature patterns in image data (paintings and photographs). By demonstrating the usefulness of armature patterns in aesthetic assessment, we argue the power of armature-based grammar in reasoning the patterns of beauty in image data. We also demonstrated promising armature classification learning model for recognizing armatures in images.

ACKNOWLEDGMENTS

This paper is based on results obtained from a project commissioned by the New Energy and Industrial Technology Development Organization (NEDO), Japan.

REFERENCES

[1] Subhabrata Bhattacharya, Rahul Sukthankar, and Mubarak Shah. 2010. A framework for photo-quality assessment and enhancement based on visual aesthetics. In *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 271–280.

[2] Thorsten Brants. 2000. Inter-annotator Agreement for a German Newspaper Corpus.. In *LREC*.

[3] Joan Bybee. 2012. Where does grammar come from? *The 5-Minute Linguist. Bite-sized essays on Language and Languages*. In E. M. Rickerson and B. Hilton (Eds) (2012), pp. 60–63.

[4] Yi-Ling Chen, Jan Klopp, Min Sun, Shao-Yi Chien, and Kwan-Liu Ma. 2017. Learning to compose with professional photographs on the web. In *Proceedings of the 2017 ACM on Multimedia Conference*. ACM, 37–45.

[5] François Chollet. 2016. Xception: Deep learning with depthwise separable convolutions. *arXiv preprint* (2016).

[6] François Chollet et al. 2015. Keras: Deep learning library for theano and tensorflow.(2015). *There is no corresponding record for this reference* (2015).

[7] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z Wang. 2006. Studying aesthetics in photographic images using a computational approach. In *European Conference on Computer Vision*. Springer, 288–301.

[8] Yubin Deng, Chen Change Loy, and Xiaoou Tang. 2017. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine* 34, 4 (2017), 80–106.

[9] Sagnik Dhar, Vicente Ordonez, and Tamara L Berg. 2011. High level describable attributes for predicting aesthetics and interestingness. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 1657–1664.

[10] Arthur Wesley Dow. 1997. *Composition: A series of exercises in art structure for the use of students and teachers*. Univ of California Press.

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. 2009. The PASCAL Visual Object Classes Challenge 2009 (VOC2009). In *Summary presentation at the 2009 PASCAL VOC workshop*, Vol. 10.

[12] Bruce Gooch, Erik Reinhard, Chris Moulding, and Peter Shirley. 2001. Artistic composition for image creation. In *Rendering Techniques 2001*. Springer, 83–88.

[13] Joy P Guilford. 1941. The phi coefficient and chi square as indices of item validity. *Psychometrika* 6, 1 (1941), 11–19.

[14] T Huang. 1996. Computer vision: Evolution and promise. (1996).

[15] Owen Jones. 1865. *The grammar of ornament*. Day and Son, Limited.

[16] Yueying Kao, Ran He, and Kaiqi Huang. 2017. Deep aesthetic quality assessment with semantic information. *IEEE Transactions on Image Processing* 26, 3 (2017), 1482–1495.

[17] Frances F Kaplan. 1991. Drawing assessment and artistic skill. *The Arts in psychotherapy* (1991).

[18] Yan Ke, Xiaoou Tang, and Feng Jing. 2006. The design of high-level features for photo quality assessment. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, Vol. 1. IEEE, 419–426.

[19] Xin Lu, Zhe Lin, Hailin Jin, Jianchao Yang, and James Z Wang. 2015. Rating image aesthetics using deep learning. *IEEE Transactions on Multimedia* 17, 11 (2015), 2021–2034.

[20] Long Mai, Hailin Jin, and Feng Liu. 2016. Composition-preserving deep photo aesthetics assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 497–506.

[21] David Marr. 1982. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY, USA.

[22] Norman Charles Meier and Carl Emil Seashore. 1929. The Meier-Seashore art judgment test. (1929).

[23] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2408–2415.

[24] Calvin F Nodine, Paul J Locher, and Elizabeth A Krupinski. 1993. The role of formal art training on perception and aesthetic judgment of art compositions. *Leonardo* (1993), 219–227.

[25] B. Nuss. 2004. *14 Formulas for Painting Fabulous Landscapes*. North Light Books. https://books.google.co.jp/books?id=3egHAAAACAAJ

[26] Pere Obrador, Ludwig Schmidt-Hackenberg, and Nuria Oliver. 2010. The role of image composition in image aesthetics. In *Image Processing (ICIP), 2010 17th IEEE International Conference on*. IEEE, 3185–3188.

[27] Edgar Alwin Payne. 1957. *Composition of outdoor painting*. EP Payne.

[28] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research* 12, Oct (2011), 2825–2830.

[29] Elina Pihko, Anne Virtanen, Veli-Matti Saarinen, Sebastian Pannasch, Lotta Hirvenkari, Timo Tossavainen, Arto Haapala, and Riitta Hari. 2011. Experiencing art: the influence of expertise and painting abstraction level. *Frontiers in human neuroscience* 5 (2011), 94.

[30] Roberto Pirrone, Vincenzo Cannella, Orazio Gambino, Arianna Pipitone, and Giuseppe Russo. 2009. Wikiart: An ontology-based information retrieval system for arts. In *Intelligent Systems Design and Applications, 2009. ISDA'09. Ninth International Conference on*. IEEE, 913–918.

[31] I. Roberts. 2007. *Mastering Composition: Techniques and Principles to Dramatically Improve Your Painting*. F+W Media. https://books.google.com.vn/books?id=hOHGV-e4o38C

[32] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[33] Milan Sonka, Vaclav Hlavac, and Roger Boyle. 2014. *Image processing, analysis, and machine vision*. Cengage Learning.

[34] Stine Vogt and Svein Magnussen. 2007. Expertise in pictorial perception: eye-movement patterns and visual memory in artists and laymen. *Perception* 36, 1 (2007), 91–100.

[35] Zhangyang Wang, Shiyu Chang, Florin Dolcos, Diane Beck, Ding Liu, and Thomas S Huang. 2016. Brain-inspired deep networks for image aesthetics assessment. *arXiv preprint arXiv:1601.04155* (2016).

[36] Leland Wilkinson. 2006. *The grammar of graphics*. Springer Science & Business Media.