Journal of Imaging Science and Technology® 61(4): 040404-1–040404-10, 2017.

Analysis of Material Representation of Manga Line Drawings using Convolutional Neural Networks

Takahiko Horiuchi▲, Yuma Saito, and Keita Hirai▲
Department of Imaging Sciences, Chiba University, Japan
E-mail: horiuchi@faculty.chiba-u.jp

Abstract. *Visual perception of materials that make up objects has been gaining increasing interest. Most previous studies on visual material-category perception have used stimuli with rich information, e.g., color, shape, and texture. This article analyzes the image features of the material representations in Japanese "manga" comics, which are composed of line drawings and are typically printed in black and white. In this study, the authors first constructed a manga-material database by collecting 799 material images that gave consistent material impressions to observers. The manga-material data from the database were used to fully train "CaffeNet," a convolutional neural network (CNN). Then, the authors visualized training-image patches corresponding to the top-n activations for filters in each convolution layer. From the filter visualization, they found that the filters reacted gradually to complicated features, moving from the input layer to the output layer. Some filters were constructed to represent specific features unique to manga comics. Furthermore, materials in natural photographic images were classified using the constructed CNN, and a modest classification accuracy of 63% was obtained. This result suggests that material-perception features for natural images remain in the manga line-drawing representations.* © *2017 Society for Imaging Science and Technology.*
[DOI: 10.2352/J.ImagingSci.Technol.2017.61.4.040404]

INTRODUCTION

In everyday life, we can easily distinguish object categories by recognizing the visual information that makes up objects' different shapes and functions. Building on this and rich scientific information, researchers have recently undertaken studies on how the perception of materials contributes to the understanding of object perception.[1,2] Recently, a few approaches have been proposed to directly study the relationships between image features and several perceptual attributes, and to estimate the attribute values of a given image.[3–6]

Material analysis using natural and texture images is often considered as a pattern-recognition problem. As reviewed by Zhang et al.,[7] techniques such as the scale invariant feature transform (SIFT)[8,9] and the histogram of oriented gradients (HoG)[10] have been manually developed. However, over the last few years, the machine-learning method called deep learning has achieved state-of-the-art performance in the ImageNet Large-Scale Visual-Recognition Challenge (ILSVRC)[11] for general object recognition. Since 2012, the top position in the competition has shifted to machine-learning methods, and the error rate has decreased year by year, achieving surprising learning accuracy. Krizhevsky et al.[12] used a convolutional neural network (CNN) to win the ILSVRC. The CNN has been established as a state-of-the-art technology in object recognition, and many studies have shifted from extracting features by hand to extracting features automatically using a CNN. These studies have been performed using large image datasets, e.g., the Flickr Materials Database.[13]

These datasets contain a great deal of information. As has been shown in many previous studies, both surface properties (e.g., color, texture, surface reflectance, etc.) and shapes are influential in distinguishing materials, as they provide relevant visual information. However, we can also perceive material properties from more primitive information. Tanaka and Horiuchi[14] investigated the perceptual qualities of static surfaces using real materials and four types of image reproduction by varying the chromaticity (color versus gray) and resolution (high versus low) of the material images. To study material perception that was not influenced by the shape or saturated color, they used square images of materials in gray tones.

Analysis of such an image with a small amount of information is expected to elucidate the mechanism that humans use for perceiving the material. In this study, the authors focus on binary line drawings, which have the simplest and smallest amount of information. For example, we can perceive the wood grain in the line-drawing images shown in Figure 1. Studies dealing with such line-drawing image features have included identifying the names of objects from sketched images[15,16] and searching for a manga scene similar to an input sketch.[17] However, most studies on line-drawing images focus on object identification. To the best of our knowledge, no studies have analyzed the material representation of line drawings.

Some popular images can convey scenes to people using only black-and-white line-drawing images. These are known in Japan as "manga" line drawings. Modern-day manga, or comic books, correspond to a Japanese style that originated during the mid-1900s. Japanese manga are distinct from traditional Western comic books in that they present fine details. Color manga can express even more semantics and artistic styles; however, manga are seldom colored. Many

▲ IS&T Members.

Received Feb. 28, 2017; accepted for publication June 19, 2017; published online July 20, 2017. Associate Editor: Henry Y. T. Ngan.
1062-3701/2017/61(4)/040404/10/$25.00

Figure 1. Examples of line drawings representing wood material.

Table I. The collected images.

	Fabric	Rock	Wood	Foliage	Water	Grass	Metal	Sand	Total
Ref. 18	75	63	77	4	7	9	16	11	262
Comics	11	6	6	9	11	10	6	6	65
Manga 109	77	62	56	127	72	28	30	20	472
Total	163	131	139	140	90	47	52	37	799

manga comics have been digitized as electronic books, and more people are reading them than ever before.

In our previous study,[18] we constructed a manga materials database, consisting of object and patch images taken from black-and-white line drawings. By analyzing the database drawings, we found that the material appearance was represented by certain low-dimensional image features, e.g., the ratio of black pixels to total pixels and the average run length of horizontal black pixels. However, the derived features were not optimal, and their material-classification accuracy was insufficient.

In this article, we analyze the image features contributing to the material representation of manga images by constructing a CNN. By visualizing the layers of the trained CNN, we analyze the image features involved in the material identification of manga images.

CONSTRUCTION OF THE MANGA MATERIALS DATABASE

In our previous study,[18] we collected 276 material objects taken from manga comics, and constructed a manga materials database. From the database, as shown in Table I, we identified eight material categories: 75 Fabric, 63 Rock, 77 Wood, 4 Foliage, 7 Water, 9 Grass, 16 Metal, and 11 Sand. However, the amount of data was biased according to the material categories in the constructed database. In addition, since commercial comics were used, there were biases in the ages and genres of the collected manga, and various image types could not be collected. In this study, we extended the database by collecting a new image dataset from manga comics and the Manga 109 database.[19]

Collection from Comics

We newly collected 472 material images from commercially available Japanese comics. As shown in Table I, the images included eight material categories: 77 Fabric, 62 Rock, 56 Wood, 127 Foliage, 72 Water, 28 Grass, 30 Metal, and 20 Sand. We digitized the material images with an Epson GT-X820 scanner as close-up images of the objects, excluding the shape information. The scanner used 600 dpi resolution and saved images as 8 bit grayscale bitmaps. Since manga consist of frames of various sizes, the shapes of the rectangular object images differed from one another. The locations and sizes of the material images were determined by the authors' subjective judgment. However, through the same subjective evaluation process as was used in Ref. 18, all of the images were assigned the same material category by more than half of the subjects. Finally, these captured grayscale images were converted into binary images using Otsu's algorithm.[20]

Collection from the Manga 109 Database

The Manga 109 dataset is composed of 109 manga books published between the 1970s and 2010s.[19] The database covers a wide range of ages, genders, and genres. The database contains more than 20,000 pages digitized at 96 dpi in JPEG format. We selected 165 material images from the database and converted them into binary images using Otsu's algorithm. Through the subjective evaluation process described in Ref. 18, as shown in Table I, we extracted 65 material images that were assigned the same material category by more than half of the subjects, including 11 Fabric, 6 Rock, 6 Wood, 9 Foliage, 11 Water, 10 Grass, 6 Metal, and 6 Sand.

Validity of the Extended Database

We summarize the extended database in Table I. As shown in the table, the database consists of 799 material images, including eight material categories.

The images collected from comics have a 600 dpi resolution, whereas the resolution of images in the Manga 109 dataset is only 96 dpi. We investigated whether this difference in resolution affected the material perception. We downsampled the 600 dpi images collected from comics to 96 dpi. Then, we conducted material-classification experiments for both the original and low-resolution images,

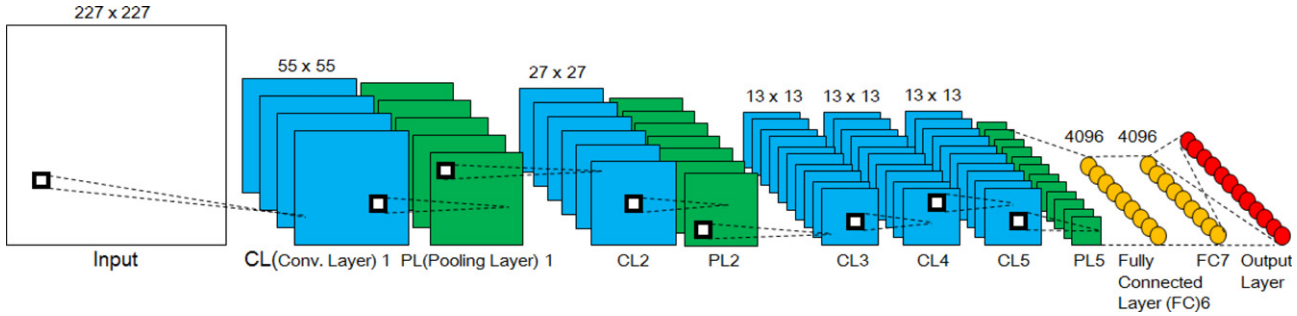


Figure 2. A CNN architecture used in this study.

Table II. Manga type in the database.

(a) Publication year		(b) Genre	
Date	Percentage	Genre	Percentage
1970's	9.0%	Science fiction	23.1%
1980's	33.3%	Gag	5.1%
1990's	21.8%	Historic	23.1%
2000's	32.1%	Animal	1.3%
2010's	3.8%	Battle	21.8%
		Fantasy	1.3%
		Romantic comedy	9.0%
		Romance	15.4%

using low-dimensional image features as explained in Ref. 18. The resulting classification accuracies were 60% and 66% for the 600 and 96 dpi images, respectively. Thus, the classification accuracy was not significantly influenced, even though the image resolution was different.

Since the drawing style of manga changes according to publication year and manga genre, we checked the variance. Tables II(a) and (b) show the percentages of images in the database for the publication year and manga genre, respectively. We can thus confirm that a wide range of data was included in the constructed database.

TRAINING OF THE MANGA MATERIALS DATABASE

Convolutional neural networks were first proposed by Fukushima,[21] developed in Ref. 22 by LeCun et al., and improved by others. Typical CNNs are composed of a convolutional layer, a pooling layer, and a fully connected layer. A CNN architecture used in this study is shown in Figure 2. In this study, we extract CNN features using the Caffe open-source package[23] with its ImageNet pre-trained CaffeNet. CaffeNet is a variant of AlexNet.[12]

The two-dimensional (2D) raw image pixels can be directly accepted as the CNN's input. The image is then convolved with multiple learned kernels using shared weights by CL1 in Fig. 2. An example of visualized weights in CL 1 will be represented in the section 'Material Feature Analysis' using Figure 5. Next, the pooling layer (PL1) reduces the image size while trying to maintain its information. The pooling layers achieve spatial invariance by reducing the resolution of the feature maps. The convolutional layer and pooling layer compose the feature-extraction part. As shown in Fig. 2, we have three sets of convolution and pooling layers (CL1–PL1, CL2–PL2, and CL5–PL5) and the second convolution layer performs convolution on the output of the first pooling layer. Between the second and fifth layers, there are two convolution layers. Afterwards, the extracted features are weighted and combined into one or more fully connected layers. This represents the classification part of the CNN. Finally, the output layer has one neuron per class (object category) in the classification task.

In this section, we describe the CNN that classifies the manga images into material categories. It consists of five convolutional layers, two fully connected layers, and a final 1000-way softmax. Other layers, e.g., max-pooling layers and normalization layers are also included.

Training Method

We used part of the dataset shown in Table I as training images for the CNN. The input images were resized to 227×227 pixels to match the CaffeNet input size.[12] In order to confirm the monotonicity of the classification accuracy for the different size, we prepared the following three datasets.

- Set 1: 433 images consisting of three categories, Fabric, Rock, and Wood, with 30 training images randomly extracted from each test category, and 343 test images.
- Set 2: 663 images consisting of five categories, Fabric, Rock, Wood, Foliage, and Water, with 15 training images randomly extracted from each test category, and 588 test images.
- Set 3: 799 images consisting of all eight categories, with 15 training images randomly extracted from each test category, and 679 test images.

Figures 3(a), (b), and (c) show the training images for Sets 1, 2, and 3, respectively.

When a CNN has only a small number of training data, a fine-tuning strategy, based on the concept of transfer learning,[24,25] is often used. The fine-tuning replaces the classification function and optimizes the network again to minimize error in another purpose. The fine-tuning

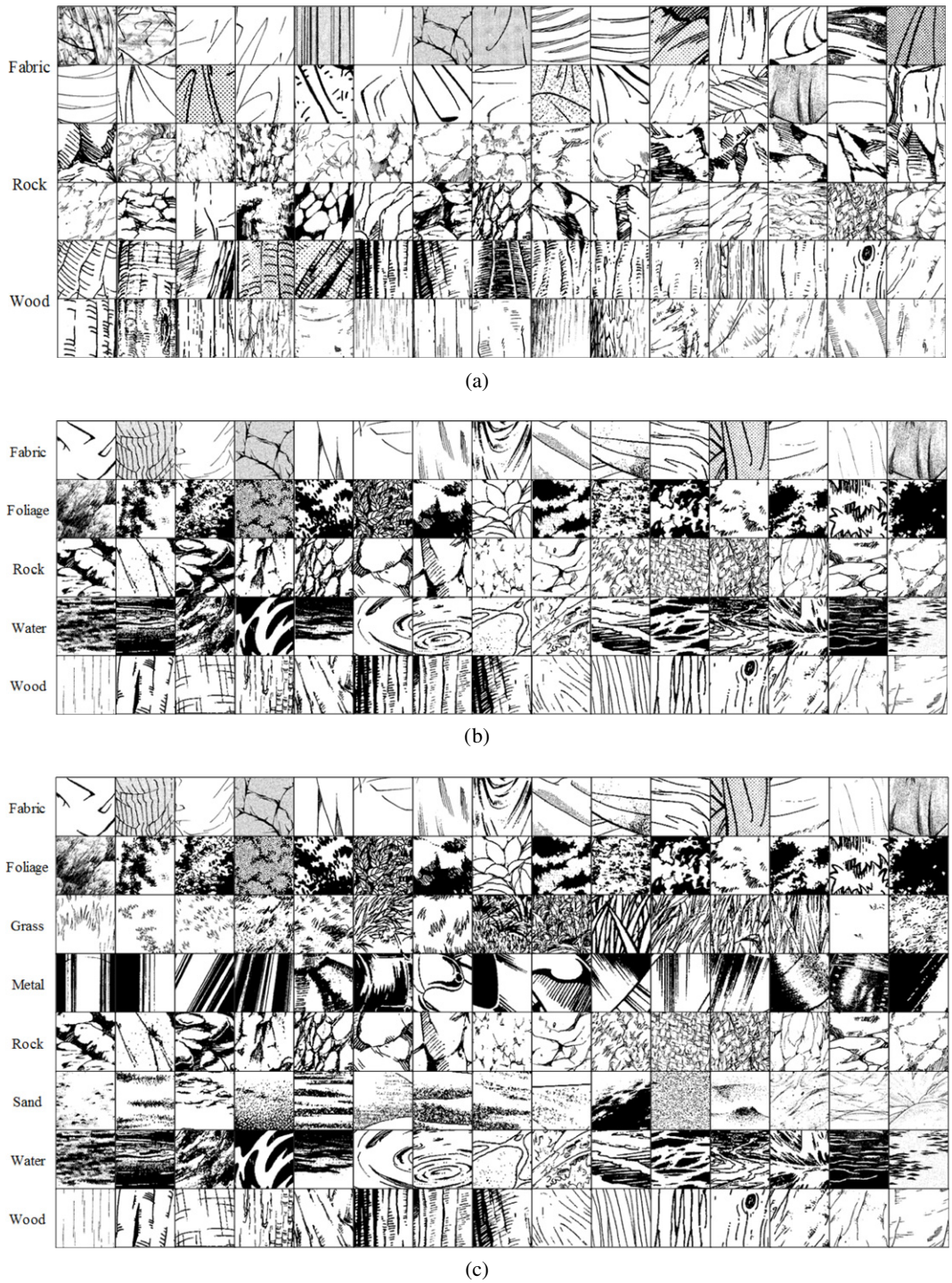


Figure 3. The training-image datasets. (a) Set 1, (b) Set 2, (c) Set 3

strategy usually reuses a CNN initially trained on a large natural image-recognition dataset. However, conventional huge datasets are typically represented as color images; none of them are represented as binary images. Although our dataset was small, CaffeNet was fully trained using the manga-material data. The full training requires learning from scratch with all the network layers initialized randomly.

The CNN was trained using the stochastic gradient descent method, with a momentum of 0.9 and a weight decay of 0.0005. When dealing with Set 1, the initial learning rate and the number of iterations were 0.001 and 50000, respectively. For Sets 2 and 3, the initial learning rate and the number of iterations were 0.0001 and 100000, respectively. These parameters were default settings suggested by Ref. 12.

(a) (b) (c) (d)

Figure 4. Example images misclassified by CaffeNet. (a) Classified into Rock, but actually Water. (b) Classified into Wood, but actually Fabric. (c) Classified into Wood, but actually Rock. (d) Classified into Foliage, but actually Water.

Table III. Top-one classification accuracy for each dataset.

Dataset	Accuracy % (top one)	
	CaffeNet	Previous[18]
Set 1	77.8	64.4
Set 2	69.3	66.7
Set 3	55.8	50.8

Training Results

We report the experimental results for datasets 1, 2, and 3. The quantitative results based on the maximum likelihood are shown in Table III. For comparison, we conducted classification experiments using our previous method, described in Ref. 18. The previous method used linear discriminant classification based on 14 manually designed low-dimensional features, e.g., the frequency of the horizontal and vertical black-and-white longer run length.

As shown in the table, CaffeNet's accuracy was higher than that of the previous method. This discrimination rate was not sufficiently accurate. However, despite having only a few hundred training images, an acceptable classification rate was obtained. This result suggests that appropriate features were built inside the network. Figure 4 shows examples of images misclassified by CaffeNet.

MATERIAL FEATURE ANALYSIS

Neural networks have long been known as "black boxes" because it is difficult to understand exactly how any particular trained neural network functions due to the large number of interacting, non-linear parts. Large, modern neural networks are even harder to study because of their size. For example, an understanding of a typical CNN involves making sense of the values taken by tens of millions of trained network parameters. Therefore, our understanding of how these models work, especially what computations they perform at intermediate layers, has lagged behind; few studies have been published that analyze the intermediate layers.[26,27]

In this section, we consider the visual features used for manga-material classification by visualizing and analyzing the image features detected in each layer in the CaffeNet described in the previous section.

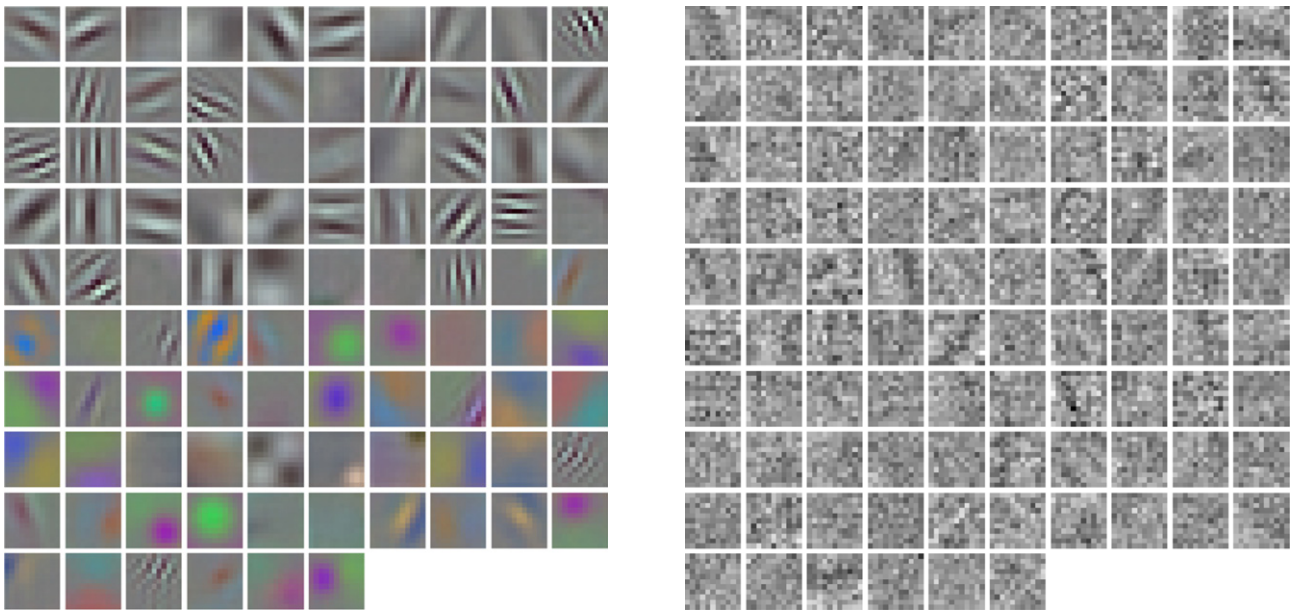
Visualization of Convolutional Layers

As the simplest method of visualizing feature quantities, we visualized the first convoluted layer connected to the input layer.[12] Since the first layer acts directly on the input image, it is possible to represent the features as images. By normalizing the parameter set for each filter in the convolution layer and replacing the values with RGB values, the convolution layer can be visualized. Fig. 5(a) shows the first layer of learned convolutional filters in CaffeNet using more than 1.2 million images from the ImageNet dataset. The filter outputs expand the dimensionality of the visual representation from the image's three-color channels to these 96 primitives. We can confirm that these filters are tuned to the edges of different orientations, frequencies, phases, and colors.

In contrast, Fig. 5(b) shows the first layer of learned convolutional filters in CaffeNet using our small manga dataset 3. As is visible, it is difficult to identify visual characteristics. This is because the features are not properly constructed in the CNN due to the small number of training images. However, even with our CNN, the classification results in Table III indicate that useful features were constructed. In the next subsection, we will further proceed to visualize the activations produced on each layer of our CNN.

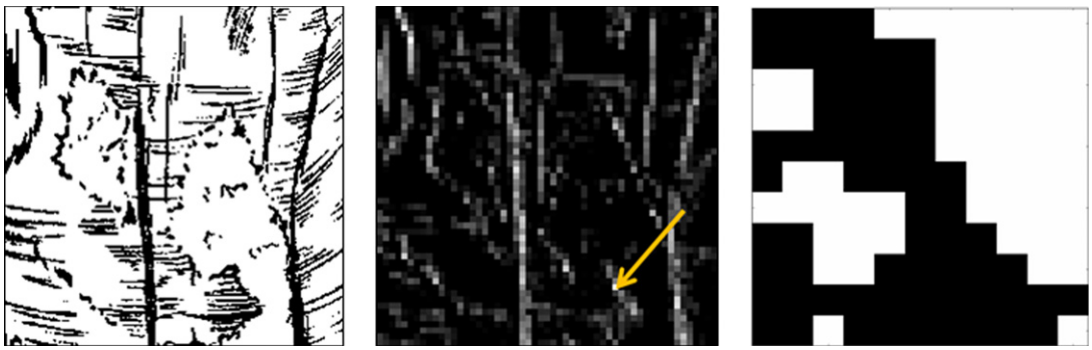
Visualization of Layer Activations

Since it was not possible to find effective features even by directly observing the weights of the convolution layer, we attempted to visualize corresponding patches in the input image from the highest activation in a convolution layer. An activation map is the visualization approach. This approach plots the activation values for the neurons in each layer in response to an image. In fully connected neural networks, the order of the units is irrelevant, so plots of these vectors are not spatially informative. However, in CNNs, filters are applied in a manner that respects the underlying geometry of the input; for 2D images, filters are applied in a 2D convolution over the two spatial dimensions of the image. This convolution produces activations on subsequent layers that are also arranged spatially for each channel.



(a) (b)

Figure 5. The first layer of learned convolutional filters in CaffeNet. (a) ImageNet,[23] (b) Manga dataset.



(a) (b) (c)

Figure 6. Example of a reconstructed patch based on the activation map. (a) Input image, (b) Activation map, (c) Corresponding image patch in (a).

We set every feature to 0 except the highest activation, and passed it backwards through the network until it reached the input layer. Then, we could visualize the corresponding input patch. As the layer deepened from CL1 to CL5, the patch area of the input image corresponding to an activation pixel increased. Therefore, by observing the different convolution layers, we expected to sequentially extract meaningful lower-order features followed by higher-order features from the input image.

Visualization Method

We performed visualizations based on the methods in Refs. 26–28. This approach highlights the portions of a particular image that are responsible for firing each neural unit.

Figure 6 shows an example of the visualization. Figs. 6(a) and (b) show an input 227 × 227 raw manga image and its 55 × 55 activation feature map of the CaffeNet,[12] respectively. The yellow arrow in Fig. 6(b) shows the pixel with the highest activation. Fig. 6(c) shows the corresponding 11 × 11 image patches in Fig. 6(a). The reconstructions are not samples from the model; they are reconstructed patterns from the training set that cause high activations in a given activation map. Since the first convoluted layer of CaffeNet convolves with an 11 × 11 kernel, the affected area becomes 11 × 11 pixels.[12] The influence area can be acquired by tracing the neurons connected with this neuron inversely to the input layer.

Visualization Result

Figures 7–11 show the feature visualizations of image patches corresponding to strong activations in each convolution layer of our CaffeNet, once the training is complete using Set 3. These have greater variation than visualizations that solely focus on the discriminant structure within each patch. In this study, a small number of datasets were used to fully train the CNN without a fine-tuning process, but it was confirmed that appropriate features were built inside the network. Hereafter,

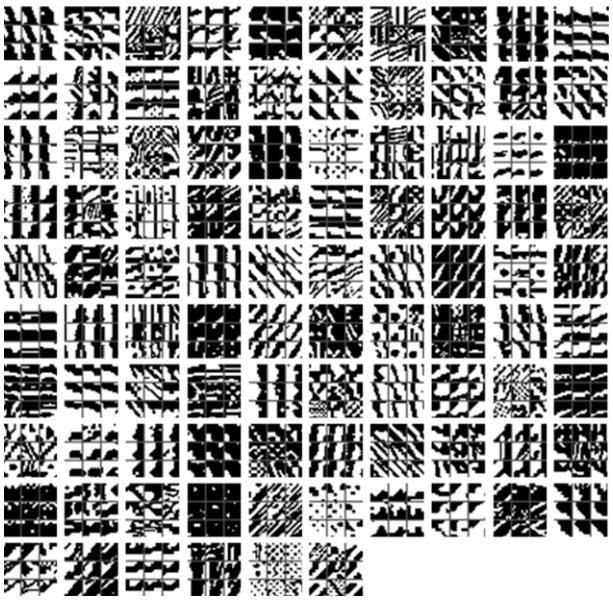


Figure 7. Visualization of image patches corresponding to the top-nine activations for 96 filters in Conv 1.

the visualization results of the features for each convolution layer will be described in detail.

Conv 1

Fig. 7 shows a visualization of image patches corresponding to the top-nine activations for 96 filters in the first convolution layer (Conv 1) of the CaffeNet.[12] In Conv 1, primitive image features, e.g., edges and gradients, are observed. This confirms that not only filters responsive to gradients and edges but also filters responsive to specific structures, e.g., diagonal lines and dots, are generated. Filters that respond to gradients and edges can also be confirmed in general object-recognition tasks,[26] but filters responding to dots and specific structures are simple image features unique to manga. With respect to the gradient direction, the number of filters that detect the gradient in the longitudinal direction (90°) is larger than the number of filters that detect the gradient in the lateral direction (0°). This indicates that manga include many images in the vertical direction, e.g., the Wood category.

Many filters for detecting diagonal lines were generated in the filter. This is presumed to be due to the technique of representing objects' shadows and color tints with diagonal lines in manga images. In the third filter from the left in the uppermost row in Fig. 7, filters for detecting multiple diagonal lines are generated. Furthermore, filters for detecting two oblique lines and filters for detecting diagonal white lines are also generated. This means that the diagonal-line feature is useful as a material representation of manga images.

Conv 2

Fig. 8 shows a visualization of image patches corresponding to the top-four activations for 35 filters in the second convolution layer (Conv 2) of our CaffeNet. Since there

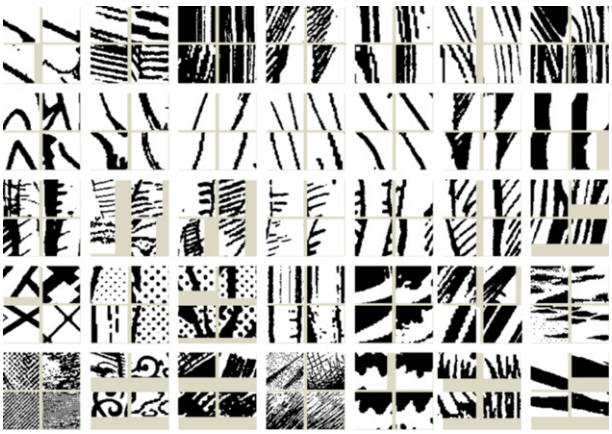


Figure 8. Visualization of image patches corresponding to the top-four activations for 35 selected filters in Conv 2.

Figure 9. Visualization of image patches corresponding with to top-four activations for 40 filters in Conv 3.

are over 200 filters, the 35 filters were randomly selected. In Conv 2, we can observe the feature combinations that appeared on the Conv 1 filters. In particular, filters combining a vertical line and a dot, and filters combining multiple diagonal lines are confirmed as numerous. As in Conv 1, many filters that detect oblique lines were also observed in Conv 2, while filters reflecting features that could not be observed in Conv 1, e.g., crosses, triangle corners, and curves, were confirmed.

Conv 3

Fig. 9 shows a visualization of image patches corresponding to the top-four activations for 40 randomly selected filters

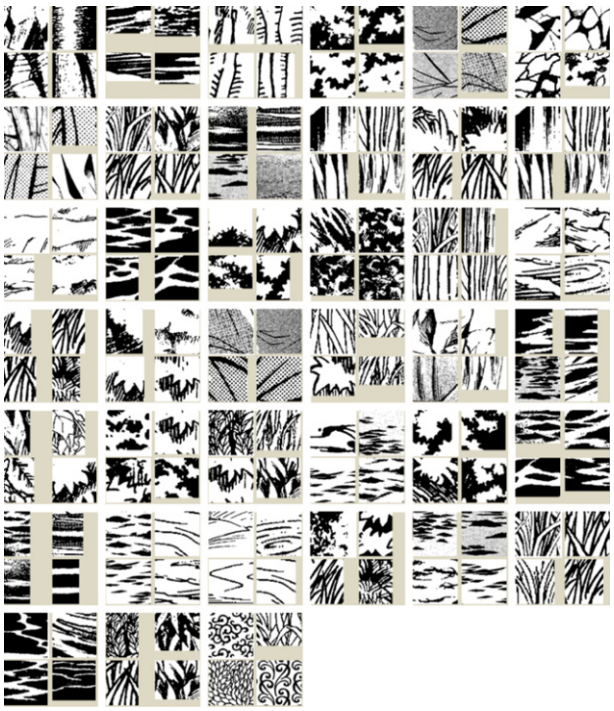


Figure 10. Visualization of image patches corresponding to the top-four activations for 39 filters in Conv 4.

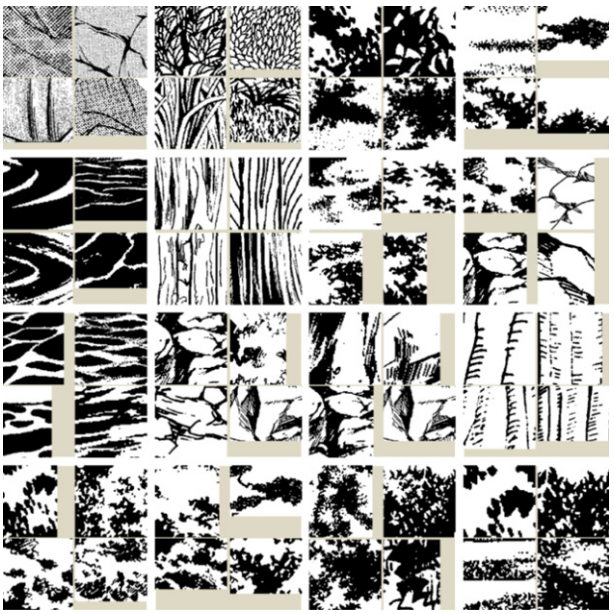


Figure 11. Visualization of image patches corresponding to the top-four activations for 16 filters in Conv 5.

in the third convolution layer (Conv 3) of our CaffeNet. In Conv 3, numerous filters were found to express the surface of a tree trunk, which combined the vertical direction, the horizontal direction, and the diagonal line. In addition, many images with a combination of diagonal lines and dots could be confirmed. This line-and-dot structure is a representation that is often seen in Fabric images using the screen tone; images corresponding to the Conv 3 filter were dominated by images from the Fabric and Wood categories. These categories can be confirmed by shallow CNN layers, i.e., materials that can be identified with relatively simple image features, e.g., hatched combinations.

Conv 4

Fig. 10 shows a visualization of image patches corresponding to the top-four activations for 39 randomly selected filters in the fourth convolution layer (Conv 4) of our CaffeNet. In Conv 4, numerous filters were detected that represent shapes such as leaves and the surface of water. Filters that detect such shapes could hardly be confirmed in Conv 3, so it is considered that complicated structures in the deep layers are necessary to represent the materials of Foliage and Water.

Conv 5

Fig. 11 shows a visualization of image patches corresponding to the top-four activations for 16 randomly selected filters in the fifth convolution layer (Conv 5) of our CaffeNet. In Conv 5, filters that can recognize each category are generated. For example, we can confirm many filters corresponding to a leaf structure and a technique unique to manga of drawing tree leaves with black spots. The structure of the boundary between rocks can also be checked.

The visualization results illustrate that Conv 1 and Conv 2 detect simple line features and combinations of multidirectional line segments, respectively. From Conv 3 onwards, filters representing materials are generated from combinations of the line segments. In particular, Fabric and Wood are represented early on, in Conv 3, and these two categories identify materials with relatively simple image features. However, in Conv 3, the Fabric and Water categories are mixed in the same filter as the Wood and Rock categories. From Conv 4 onwards, one filter matches each material category. It is considered that, starting from Conv 4, a filter capable of detecting the material is generated. Although the detected features were different, this hierarchical nature of the features in the network was similar to the result in Ref. 26 for color images.

Relationship with Material Features of Natural Images

In this study, a CNN was constructed and the features were visualized to analyze the material representations of manga binary images. In this subsection, to elucidate the relationship between the material representation of manga images and that of natural photographic images, we perform a classification experiment using grayscale natural images as test data.

Thirty images were selected as test images from the Flickr Material Database for three types of material: Fabric, Stone, and Wood. The test images were resized to square images with 227×227 pixels and then converted into grayscale images. The 30 test images are shown in Figure 12. We used a CaffeNet trained with the Set 1 dataset for the classification experiments. Here, the Stone category in Fig. 12

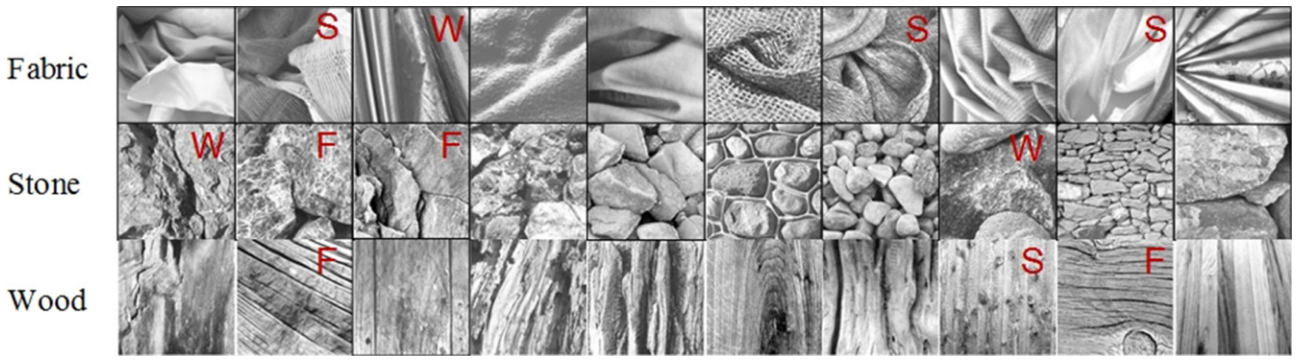


Figure 12. The test data for natural photographic images. Each red letter indicates the initial of the category into which the image was misclassified.

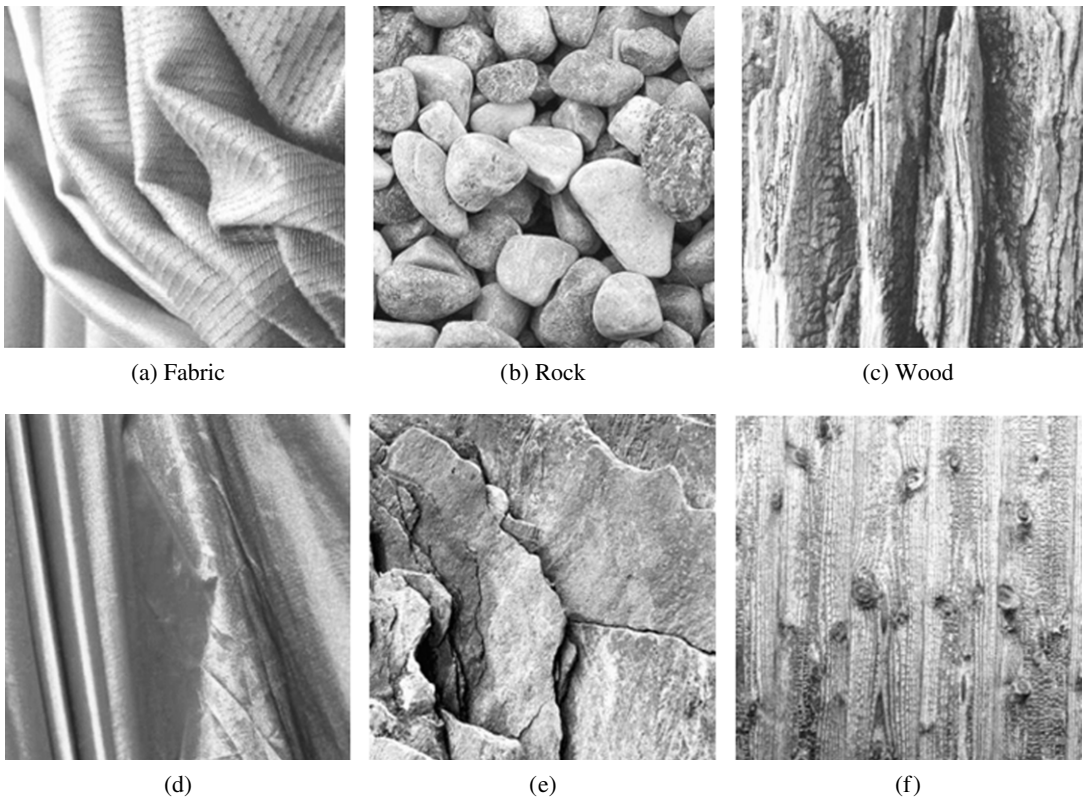


Figure 13. Example images misclassified by CaffeNet. (a–c) Correctly classified images. (d–f) Misclassified images. (d) Classified into Wood, but actually Fabric. (e) Classified into Fabric, but actually Rock. (f) Classified into Rock, but actually Wood.

was interpreted as the Rock category in the constructed CaffeNet.

The classification accuracies for Fabric, Stone, and Wood were 60%, 60%, and 70%, respectively. Despite the fact that the CNN was trained on binary manga images, modest classification accuracy was obtained for natural photographic images. This result suggests that the material-perception features of manga images and natural photographic images are similar. In other words, manga accurately represent the material features of natural images by compressing information. Figures 13(a–c) and (d–f) show examples of correctly classified images and misclassified images.

CONCLUSIONS

In this study, we extended the manga-material image database, and constructed a CNN that classified materials shown in manga images. By visualizing the intermediate layers of the trained CNN, we analyzed the image features involved in the material identification of manga images.

The manga-material images collected in this study were more generic, including elements such as a broader genre and publication year than our previous study. In addition, a subjective evaluation experiment confirmed that all of the images were assigned to the same material category by more than half of the subjects.

Next, we fully trained a "CaffeNet" CNN, which classified the manga images into material categories. Despite there being a few hundred training images, an acceptable classification rate was obtained. This result suggested that appropriate features were built inside the network. Then, we visualized the image patches corresponding to training images with the top-n activations for filters in each convolution layer. The visualization result confirmed that the filters reacted gradually to complicated features, from the input layer to the output layer.

Furthermore, the materials in natural photographic images were classified using the CNN constructed using manga images, and a modest classification rate was obtained. This suggests that material-perception features for natural images remain in the manga image.

In this study, the number of images used for training the CNN was not sufficiently large, so direct visualization of the first convolution layer was not useful, as shown in Fig. 5(b). As one solution, fine-tuning may be an effective approach for producing highly accurate learning results with a small dataset. However, for example, if the fine-tuning is performed using the ImageNet model, the early convoluted layers are not related to material classification, but rather general object recognition. To capture more sophisticated features of manga-material images in the future, we should collect thousands of images or perform fine-tuning for an appropriate network.

ACKNOWLEDGMENTS

This work was supported by JSPS KAKENHI Grant Number JP15H05926 (Grant-in-Aid for Scientific Research on Innovative Areas "Innovative SHITSUKAN Science and Technology").

REFERENCES

[1] R. W. Fleming, S. Nishida, and K. R. Gegenfurtner (Ed.), "Perception of material properties (Part I)," Vis. Res. B **109**, 123–236 (2015).

[2] R. W. Fleming, S. Nishida, and K. R. Gegenfurtner (Ed.), "Perception of material properties (Part II)," Vis. Res. B **115**, 157–302 (2015).

[3] R. O. Dror, E. H. Adelson, and A. S. Willsky, "Recognition of surface reflectance properties from a single image under unknown real-world illumination," *Proc. Workshop on Identifying Objects Across Variations in Lighting at Computer Vision and Pattern Recognition* (IEEE, Piscataway, NJ, 2001).

[4] T. Abe, T. Okatani, and K. Deguchi, "Recognizing surface qualities from natural images based on learning to rank," *Proc. Int'l. Conf. on Pattern Recognition* (IEEE, Piscataway, NJ, 2012), pp. 3712–3715.

[5] M. Varma and A. Zisserman, "A statistical approach to material classification using image patch exemplars," IEEE Trans. Pattern Anal. Mach. Intell. **31**, 2032–2047 (2009).

[6] C. Liu, L. Sharan, E. H. Adelson, and R. Rosenholtz, "Exploring features in a Bayesian framework for material recognition," *Proc. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2010), pp. 239–246.

[7] J. Zhang and T. Tan, "Brief review of invariant texture analysis methods," Pattern Recognit. **35**, 735–747 (2002).

[8] J. Koenderink and A. van Doorn, "Representation of local geometry in the visual system," Biol. Cybern. **545**, 367–375 (1987).

[9] D. G. Lowe, "Distinctive image-features from scale-invariant keypoints," Int. J. Comput. Vis. **60**, 91–110 (2004).

[10] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc. Comput. Vis. Pattern Recognit.* (IEEE, Piscataway, NJ, 2005), vol. 2, pp. 886–893.

[11] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Image net large scale visual recognition challenge," Int. J. Comput. Vis. **115**, 211–252 (2015).

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," Proc. Adv. Neural Inf. Process. Syst. **25**, 1106–1114 (2012).

[13] L. Sharan, R. Rosenholtz, and E. Adelson, "Material perception: What can you see in a brief glance?," J. Vis. **9**, 784 (2009).

[14] M. Tanaka and T. Horiuchi, "Investigating perceptual qualities of static surface appearance using real materials and displayed images," Vis. Res. B **115**, 246–258 (2015).

[15] A. Yu, Y. Yang, Y. Z. Song, T. Xiang, and T. Hospedales, "Sketch-a-net that beats humans", (2015), arXiv preprint, arXiv:1501.07873, 2.

[16] X. Sun, C. Wang, C. Xu, and L. Zhang, "Indexing billions of images for sketch-based retrieval," *Proc. 21st ACM Int'l. Conf. on Multimedia* (ACM, New York, NY, 2013), pp. 233–242.

[17] Y. Matsui, K. Ito, Y. Aramaki, A. Fujimoto, T. Ogawa, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset," *Multimedia Tools Appl.* (2016), pp. 1–28.

[18] Y. Saito, K. Hirai, and T. Horiuchi, "Construction of manga materials database for analyzing perception of materials in line drawings," *Proc. IS&T CIC23: Twenty-third Color and Imaging Conf.* (IS&T, Springfield, VA, 2015), pp. 201–206.

[19] Y. Matsui, K. Ito, Y. Aramaki, T. Yamasaki, and K. Aizawa, "Sketch-based manga retrieval using manga109 dataset" (2015), arXiv:1510.04389 .

[20] N. Otsu, "A threshold selection method from gray-level histograms," IEEE Trans. Syst. Man Cybern. **9**, 62–66 (1979).

[21] K. Fukushima and S. Miyake, "Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position," Pattern Recognit. **15**, 455–469 (1982).

[22] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," Neural Comput. **1**, 541–551 (1989).

[23] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: convolutional architecture for fast feature embedding," *Proc. 22nd ACM Int'l. Conf. Multimedia* (ACM, New York, NY, 2014), pp. 675–678.

[24] Y. Bengio, "Deep learning of representations for unsupervised and transfer learning," *JMLR: Workshop and Conf. Proc.* (Bellevue, USA, 2012), Vol. 27, pp. 17–36.

[25] J. Donahue, Y. Jia, O. Vinyals, J. Homan, N. Zhang, E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition" (2013), arXiv:1310.1531 .

[26] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," *Lecture Note ECCV, LNCS* (Zurich, Switzerland and Springer, 2014), Vol. 8689, pp. 818–833.

[27] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," *Proc. IEEE CVPR* (IEEE, Piscataway, NJ, 2014), pp. 580–587.

[28] J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," *Proc. Int'l. Conf. on Machine Learning, Deep Learning Workshop* (Lille, France, 2015).