EasyChair Preprint

№ 557

CNN-based Classification of Illustrator Style in Graphic Novels: Which Features Contribute Most?

Jochen Laubrock and David Dubray

October 4, 2018

CNN-based Classification of Illustrator Style in Graphic Novels: Which Features Contribute Most?*

Jochen Laubrock[0000−0002−0798−8977] and David Dubray

University of Potsdam, Germany
{laubrock,ddubray}@uni-potsdam.de

Abstract. Can classification of graphic novel illustrators be achieved by convolutional neural network features (CNN) evolved for classifying concepts on photographs? Assuming that basic features at lower network levels generically represent invariants of our environment, they should be reusable. However, features at what level of abstraction are characteristic of illustrator style? We tested transfer learning by classfiying roughly 50,000 digitized pages from about 200 comic books of the Graphic Narrative Corpus (GNC, [5]) by illustrator. For comparison, we also classified Manga109 [16] by book. We tested the predictability of visual features by experimentally varying which of the mixed layers of Inception V3 [26] was used to train classifiers. Overall, the top-1 test-set classification accuracy in the artist attribution analysis increased from 92% for mixed-layer 0 to over 97% when adding mixed-layers higher in the hierarchy. Above mixed-layer 5, there were signs of overfitting, suggesting that texture-like mid-level vision features were sufficient. Experiments varying input material show that page layout and coloring scheme are important contributors. Thus, stylistic classification of comics artists is possible re-using pretrained CNN features, given only a limited amount of additional training material. We propose that CNN features are general enough to provide the foundation of a visual stylometry, potentially useful for comparative art history.

Keywords: Convolutional Neural Network · Classification · Graphic Novels · Stylometry.

1 Introduction

What constitutes pictorial style? Pictorial style has been one of the major interests of art history since its beginning, and some of the founding fathers of modern art history were greatly interested in developing a formal theory of style [28, 17]. A system enabling a formal and dimensional description of style is extremely useful, because it enables comparative work and allows a scientific judgment and classification of art in terms of agreed-upon categories. Placing

* Supported by a BMBF eHumanities grant to JL

artists, artworks and epochs in such a dimensional system enables comparative art-historical and art-critical enquiry into style. Ideally, comparative analysis of the formal organization of pictures will benefit from a set of objective parameters.

Perception-based approaches to style Artists often play with human perception, and explore its boundaries. Similarly, art historians have been interested in relating art to fundamentals of human sensation and perception. Ernst Gombrich believed that a proper theory of representation had to consider results of scientific research. Perceptual psychology deals with how we perceive and build rich internal representations from a flat retinal image, and how different images differ in memorability and emotional evaluation (e.g. how pleasing is this image?). Art history wants to understand how a flat pictorial image can represent, how representation evolved, and what are the patterns of beauty. Both disciplines are thus naturally interested in the nature of representations, albeit at different levels.

A fundamental insight from psychologist studying perception is that our visual system is hierarchically organized and works in a highly constructive manner: higher levels of visual processing re-construct semantics and depth of real-world scenes from an initial pixel-like coding available at the retinal image. Visual and cognitive processing along the ventral pathway of the brain eventually leads to object representations, but the representations at the initial (lower) stages of cortical processing are rather simple, coding for information such as boundaries, contrast, and color. These simple features are likely to code invariants of our evolutionary environment. The receptive fields of cells at the lower stages are small, so that each neuron only "sees" a small part of the visual scene. Mid-level vision takes their computations as input and combines them to code texture, shape, object parts and proto-objects as well as to separate figure from ground, resolve shadows, etc. Receptive fields get larger the higher up the processing hierarchy we go.

Computational approaches to perception Recent advances in neural modelling using convolutional neural networks (CNNs; [14]) have led to computational models of low- and mid-level vision up to the conceptual level that are a rather good approximation of human vision. Research comparing visual representations in CNN-based models with actual brain recordings has found a striking correspondence [2, 1]. CNNs are a class of neural networks specialized in analyzing data with an implicit spatial layout, such as RGB images. CNNs are characterized by local connections, shared weights, pooling, and the use of many layers. Within each convolutional layer, a stack of different filters (feature maps) is trained. Each unit is connected to local patches in the feature maps of the previous layer through a set of learned weights, which describe a filter kernel, and learned by backpropagation [20]. A local weighted sum computed by applying the filter kernel to the image is passed through a non-linearity, often a rectified linear unit (ReLU). All units in a feature map share the same filter kernel; feature maps in a layer differ by using different kernels. The receptive

field size of each filter (i.e. the region of the image it responds to) is small at the lower layers, and becomes progressively larger at higher layers. Conversely, the higher the layer, the more complex the features encoded by the filters. Pooling layers typically replacing a local patch by its maximum value are added to further reduce the number of parameters and to provide a more coarse-grained and robust description.

Thus, like the human visual system uses a complex processing hierarchy in learning to categorize objects, CNN-based models learn what combinations of simple and intermediate features are most discriminative for a given object. Lower-level filters often respond well to edges and boundaries and thus resemble simple cells in human visual cortex. Higher-level features, in contrast, can code for complex stimuli like textures or facial parts. Just like the visual system, CNNs compose objects out of simple features by using compositional feature hierarchies. Edges combine into motifs, motifs into parts, and parts into objects. One fundamental insight is that especially the simpler lower-level features represent environmental invariants, so that they are quite generic and can be useful for many different images. For example, a low-level neuron in a CNN as well as simple cells in human early vision (V1, [9]) implement a filter responding to edges of a specific orientation, which may be useful in many contexts. Because lower-level features are quite generic, CNNs pre-trained on large-scale image classification tasks like ImageNet ([4], 14 million images with over 1,000 classes) can often be adapted to specific material by re-training just a few layers. The goal of the present work was to test for transfer to comics drawings.

Computational approaches to pictorial style The fairly new discipline of the *digital humanities* (DH) has been mostly interested in computational approaches to literature studies, with hallmarks being the so-called "distant reading" of texts [17] and the development of stylometric measures for text, enabling, for example, quantitative style comparison and author identification [10]. Very recently a shift in interest towards visual material can be observed in the DH community. Because CNN models contain examinable representations at several hierarchical levels (objects, proto-objects, textures, colors, orientations, etc), they are likely to provide a highly useful formal description of visual stimuli including comics and visual art. We are convinced that distant reading of art, exploration of the internal representations of deep CNNs trained on artworks, and correlating and mapping them to concepts from art history, will bring new perspectives and possibly even transform the field of art history, in a similar way that distant reading and computational linguistics contributed to the comparative study of literature.

2 Related work

Machine learning approaches combining visual features with similarity metrics have recently been used with some success to classify artists and styles [22]. Saleh and Elgammal also compared engineered features with CNN features and

somewhat surprisingly found that CNN features were not necessarily performing better. However, closer reading of their paper seems to suggest that they used a very late, 'semantic' output feature of a network trained on an unrelated task, thus the contribution of lower-level CNN features is not known.

There have been some successful attempts at classifying the style of artistic paintings using CNN features, which generally performed better than classic, engineered features [11, 3]. Benoît Seguin and colleagues have built an image search engine for a large collection of paintings using pre-trained CNN features and paying particular attention to visual links, a category important to art historians [23].

Leon Gatys has described an approach to a general description style using CNN features, and pioneered style transfer, which has become quite popular [7, 8]. Style transfer works by transfering style of one image to content of another image. This may be one of the most advanced descriptions of image style to date.

In the domain of comics, there were some approaches using relatively simple summary statistics to describe style. Lev Manovich pioneered this approach in combination with large-scale visualization [15]. Dunst and Hartel describe an approach to stylometry using somewhat more sophisticated engineered features such as a shape descriptor [6], and found some differences between genres.

Saito and Matsui used deep CNN features for computing similarity metrics between a large number of labeled illustrations, for which they were able to compute a semantic vector representation and create an impressive demonstration of 'semantic morphing' [21]. Interestingly, Matsui chose to use classic rather than CNN-based features in the equally impressive sketch-based image-retrieval for Manga109 [16], possibly for performance reasons. Current work from that group, albeit using different material (Kotenseki images), also seems to use deep CNN features [25].

Rigaud and Burie have a long history of studying comics from a computational perspective. In a recent conference presentation they showed that models based on deep CNN features perform significantly better than other approaches in detecting characters in comics [18]. Such object-level descriptions might well be characteristic of the style of an artist.

We have previously used deep CNN features from VGG-19 [24] in combination with Deep Gaze II [12] predictions of empirical saliency to show that gaze locations of human readers measured using eye-tracking can be well predicted by CNN features. [13].

3 CNNs applied to graphic novels

Here we propose a method for a visual stylometry of comics based on CNN features. We test transfer to comics by using a large corpus of graphic narratives. To illustrate the approach, we us CNN features to classify illustrator, genre, and publisher. We use an experimental approach to study the effect that some variables such as page layout may have on the classification. In closing we explore how the approach might be used in other domains such as art history.

3.1 Material: GNC

The material we used is the Graphic Narrative Corpus (GNC; [5]). The GNC is a representative collection of graphic novels, i.e., book-length comics that tell continuous stories and are aimed at an adult readership. At the time of analysis, the stratified monitor corpus included 209 graphic narratives amounting to nearly 50,000 digitized pages. A subset of the first chapter of these works is annotated by human annotators with respect to the location and identity of panels, main characters, character relations, captions, speech bubbles, onomatopoeia, and the respective text. Furthermore, eye movement data is collected for these pages to measure readers' attention. Metadata for the GNC includes information on author and illustrator ($N = 161$ at the time of analysis), publisher (78), and genre (24). For comparison, the classifiers were also fit on books of the Manga109 dataset [16].

3.2 CNN Model

In order to test generalization of the features and their transfer to graphic illustrations, we describe material from the GNC using a specific CNN, Inception V3 [26], using pre-trained weights from ImageNet. We chose Inception V3 for stylometry and artist attribution due to its state-of-the art performance, economic parameterization, and relative independence of input sizes. Scanned comic book pages were fed through the CNN to obtain feature maps. For classification, we trained a simple fully-connected neural network with one hidden layer of 1024 units, using average-pooled feature maps from mixed layers as input (see below) and illustrators, genre, or publisher (GNC) or book titles (Manga109) as outputs. The training set contained 90% of pages of each book, 10 % of randomly determined pages per comic were held out as a test set to evaluate performance.

4 Experiments

4.1 Experiments

We inspected the predictive value of pre-trained Inception V3 filters for classifiying comic illustratorship. Specifically, we used Inception V3 weights from ImageNet classification. Due to our interest in the convolutional part all of the fully connected layers were stripped off. Because we were particularly interested in what type of features are most useful in characterizing illustrator style, we lesioned the CNN at progressively lower layers and compared classification performance to the full model. In order to get a better understanding for what aspects of the material were used in classification, we artificially removed some cues from the material and re-fit the models. First, page layout should become less important if only a randomly cropped part of the page is used. Second, it is possible that a CNN is learning scanning artifacts and page frames for a given book rather than an artist's style. The chance for such artefact-based classification should be much reduced when the boundaries of a page are cropped. Third,

to study the contribution of color to style, we compared colored with grayscale versions, using several different methods of thresholding. Finally, we show how an image search engine for our corpus can be implemented using nearest neighbor search in feature space.

Representation As input to our classifiers we used feature maps obtained by processing the input image with Inception V3 until up to a set of mixed layers. The feature maps were then global average pooled, since we thought that a stylistic signature should be similar in several regions of the image. This results in a fairly compact representation. The mean size of the 49,009 input images was 814 kBytes. The size of the vector we used to represent feature maps ranged from 256 to 2048 entries for single-layer and 10,048 entries for cumulative representations. Table 1 shows the number of filters per layer used, as well as their cumulative sum and the average compression ratios for representations based on single and cumulative layers, assuming single precision and an average image size.

Table 1. Representations and compression ratio.

Layer	N Filters	Cumulative Sum	Compression Ratio	
			Single	Cumulative
mixed0	256	256	814	814
mixed1	288	544	724	383
mixed2	288	832	724	251
mixed3	768	1600	271	130
mixed4	768	2368	271	88
mixed5	768	3136	271	66
mixed6	768	3904	271	53
mixed7	768	4672	271	45
mixed8	1280	5952	163	35
mixed9	2048	8000	102	26
mixed10	2048	10048	102	21

Lesioning In the lesioning experiments, we used the output represenations of progressively deeper mixed layers of Inception V3 as inputs to the classifier. Additionally, we computed classifications based on single-layer outputs for mixed layers 0 to 10 vs. outputs up to mixed layer k, $k \in 0, 1, \ldots, 10$.

Binarization, Boundary Removal, Cropping *Binarization.* To investiage the use of color in classification, color and grayscale images were converted to black-and-white images using two different thresholding methods: Otsu's method [19], and adaptive local thresholding as implemented in scikit-image [27].

Boundary Removal. Boundaries were cropped so as to remove 36% of the image area (20% side, centered).

Cropping. Classification using the full page ($1024 \times 768\,px$) was compared to classification using various crop sizes: $200 \times 200\,px, 300 \times 300\,px, 400 \times 400\,px$

4.2 Results

Training and Test Training the neural network classifier on illustrators or book resulted in almost perfect training set performance after some 1000 iterations for both GNC and Manga109, respectively (Figure 1). The overall accuracy on

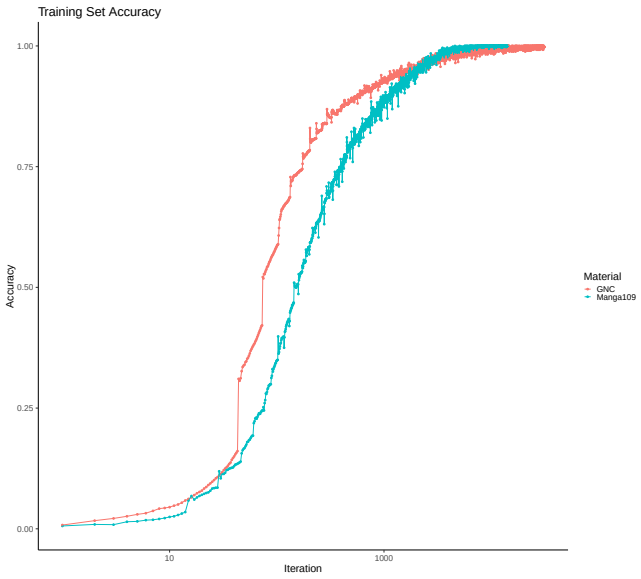

Fig. 1. Example of the evolution of training set accuracy during learning. Note that the x-scale is log10.

the test set was over 95% for the GNC and about 93% for the Manga109 data set. The somewhat lower performance on Manga109 might be due to its back-and-white only pages, meaning that features related to coloring scheme could not be used for classification. We further explore the relevance of color below.

Lesioning Results from the lesioning experiment are presented in Figure 2. Prediction using lower-level features from bottom layers are not as good as using higher-level features. Interestingly and importantly, very high-level features also don't perform quite as good. This probably indicates that they are too specifically representing concepts useful for ImageNet classification. They also seem

to be contributing to overfitting. Overall this result suggests that more generic mid-level features are better suited for classification of artistic style when using pre-trained features.

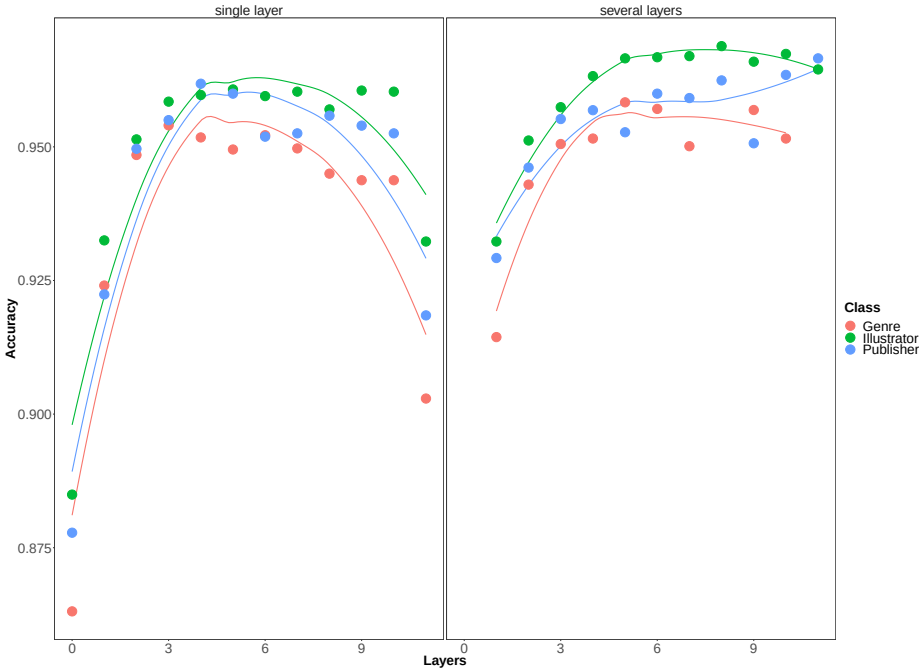

Fig. 2. Classification performance for several classes of the GNC corpus, using either single CNN layers (right panel) or several layers up to a certain layer (right panel). Lines are added by a loess smoother.

Binarization, Boundary Removal, Cropping Results from the experimental manipulations of the input images using image processing are shown in Figure 3. Removing color hat a much stronger effect on genre than on illustrator classification (left panel). Removing the boundaries didn't affect classification much, suggesting that scanning artefacts probably didn't play a large role for classification. Destroying the page layout by cropping random elements impaired performance quite strongly, suggesting that page layout is important for all three classifications.

Nearest Neighbors A nearest-neighbor search can be easily implemented using the computed representations, using a distant measure such as $\sum |X - X_i|$. A visual inspection of erroneous classifications shows interesting results. For example, given a Manga search image, many of the similar pages are also by the same

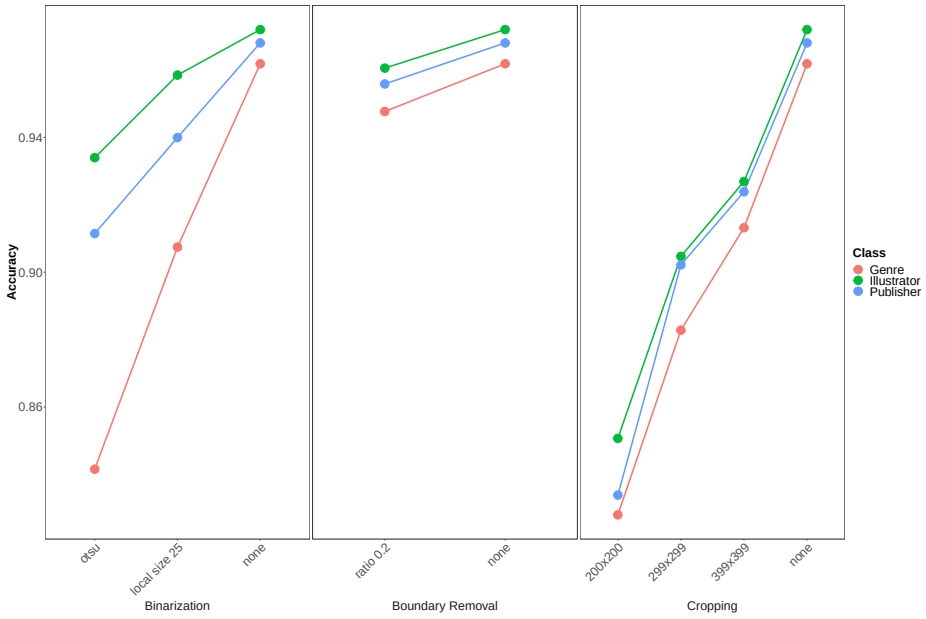

Fig. 3. Classification performance for several classes of the GNC corpus for experimental variation of the training material using image processing: binarization (left panel), boundary removal (middle panel), and cropping (right panel).

illustrator, and all of them are Mangas. So, implicitly, the illustrator classificator can distinguish between Manga and Western comics. Frank Miller's "Batman-The Dark Knight" closest cousins are also other pages from Miller—although there is some confusion with (also action-oriented) Manga. The features even allow to track historical developments within a book. For example, Vincent Mahé's 750 Years in Paris shows the evolution of a house in Paris over several centuries. Generally, using our embedding, pages from similar eras are grouped together.

5 Discussion

We presented a description of illustrator style using a compact representation of CNN features. Overall classification accuracy was very good, thus transfer to illustrations from features obtained by training on photographs generally works well. Mid-level features corresponding to texture and hatching seem to be more important for classification than higher-level features that are closer to object-level descriptions. Visualization of the most discriminative features (not shown) supports this interpretation. Page layout and color do play an important role, whereas scanning artefacts do not contribute much.

We think that the successful transfer suggest that the feature descriptions at the lower and middle levels of CNNs are rather generic, given that they were trained on enough real-world material. After all, an illustration is an abstraction of a visual reality and thus still related to a less abstract photographic depiction. We think that successful transfer implies that CNN features can be successfully employed in several new domains, for example, in describing and computing similarities of artworks in general. While we are pretty confident that they will work well with figurative art, it will be interesting to explore how pre-trained CNN features perform in classifying abstract art.

References

1. Cichy, R.M., Khosla, A., Pantazis, D., Oliva, A.: Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. Neuroimage **153**, 346–358 (Jun 2017). https://doi.org/10.1016/j.neuroimage.2016.03.063
2. Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A.: Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific reports **6**, 27755 (06 2016). https://doi.org/10.1038/srep27755
3. Crowley, E.J., Zisserman, A.: In search of art. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) Computer Vision - ECCV 2014 Workshops. pp. 54–70. Springer International Publishing, Cham (2015)
4. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR09 (2009)
5. Dunst, A., Hartel, R., Laubrock, J.: The Graphic Narrative Corpus (GNC): Design, annotation, and analysis for the Digital Humanities. In: 2017 14th

IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 03, pp. 15–20 (Nov 2017). https://doi.org/10.1109/ICDAR.2017.286, doi.ieeecomputersociety.org/10.1109/ICDAR.2017.286
6. Dunst, A., Hartel, R.: The quantitative analysis of comics: Towards a visual stylometry of graphic narrative. In: Dunst, A., Laubrock, J., Wildfeuer, J. (eds.) Empirical Comics Research: Digital, Multimodal, and Cognitive Methods, chap. 12, pp. 239–263. Routledge, New York (2018)
7. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2414–2423 (June 2016). https://doi.org/10.1109/CVPR.2016.265
8. Gatys, L.A., Ecker, A.S., Bethge, M.: Texture and art with deep neural networks. Curr Opin Neurobiol **46**, 178–186 (10 2017). https://doi.org/10.1016/j.conb.2017.08.019
9. Hubel, D.H., Wiesel, T.N.: Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology **148**, 574–91 (Oct 1959)
10. Juola, P.: Authorship attribution. Foundations and Trends® in Information Retrieval **1**(3), 233–334 (2008). https://doi.org/10.1561/1500000005, http://dx.doi.org/10.1561/1500000005
11. Karayev, S., Hertzmann, A., Winnemoeller, H., Agarwala, A., Darrell, T.: Recognizing image style. CoRR **abs/1311.3715** (2013), http://arxiv.org/abs/1311.3715
12. Kummerer, M., Wallis, T.S.A., Gatys, L.A., Bethge, M.: Understanding low- and high-level contributions to fixation prediction. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017)
13. Laubrock, J., Hohenstein, S., Kümmerer, M.: Attention to comics: Cognitive processing during the reading of graphic literature. In: Dunst, A., Laubrock, J., Wildfeuer, J. (eds.) Empirical Comics Research: Digital, Multimodal, and Cognitive Methods, chap. 12, pp. 239–263. Routledge, New York (2018)
14. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Computation **1**(4), 541–551 (Dec 1989). https://doi.org/10.1162/neco.1989.1.4.541
15. Manovich, L.: How to compare one million images? in david berry, ed., understanding digital humanities (palgrave, 2012). In: Berry, D.M. (ed.) Understanding digital humanities. Palgrave Macmillan (2012)
16. Matsui, Y., Ito, K., Aramaki, Y., Fujimoto, A., Ogawa, T., Yamasaki, T., Aizawa, K.: Sketch-based manga retrieval using manga109 dataset. Multimedia Tools and Applications **76**(20), 21811–21838 (Oct 2017). https://doi.org/10.1007/s11042-016-4020-z, https://doi.org/10.1007/s11042-016-4020-z
17. Moretti, F.: Distant Reading. Verso, London/New York (2013)
18. Nguyen, N., Rigaud, C., Burie, J.: Comic characters detection using deep learning. In: 2nd International Workshop on coMics Analysis, Processing, and Understanding, 14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017. pp. 41–46 (2017). https://doi.org/10.1109/ICDAR.2017.290, https://doi.org/10.1109/ICDAR.2017.290
19. Otsu, N.: A threshold selection method from gray-level histograms. IEEE Transactions on Systems, Man, and Cybernetics **9**(1), 62–66 (Jan 1979). https://doi.org/10.1109/TSMC.1979.4310076
20. Rumelhart, D.E., Hinton, G.E., Williams, R.J.: Learning representations by back-propagating errors. Nature **323**, 533—536 (10 1986), http://dx.doi.org/10.1038/323533a0

21. Saito, M., Matsui, Y.: Illustration2vec: A semantic vector representation of illustrations. In: SIGGRAPH Asia 2015 Technical Briefs. pp. 5:1–5:4. SA '15, ACM, New York, NY, USA (2015). https://doi.org/10.1145/2820903.2820907, http://doi.acm.org/10.1145/2820903.2820907
22. Saleh, B., Elgammal, A.M.: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. CoRR **abs/1505.00855** (2015), http://arxiv.org/abs/1505.00855
23. Seguin, B., Striolo, C., diLenardo, I., Kaplan, F.: Visual link retrieval in a database of paintings. In: Hua, G., Jégou, H. (eds.) Computer Vision – ECCV 2016 Workshops. pp. 753–767. Springer International Publishing, Cham (2016)
24. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. CoRR **abs/1409.1556** (2014), http://arxiv.org/abs/1409.1556
25. Sirirattanapol, C., Matsui, Y., Satoh, S., Matsuda, K., Yamamoto, K.: Deep image retrieval applied on kotenseki ancient japanese literature. In: 2017 IEEE International Symposium on Multimedia (ISM). pp. 495–499 (Dec 2017). https://doi.org/10.1109/ISM.2017.98
26. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. CoRR **abs/1512.00567** (2015), http://arxiv.org/abs/1512.00567
27. van der Walt, S., Schönberger, J., Nunez-Iglesias, J., Boulogne, F., Warner, J., Yager, N., Gouillart, E., Yu, T., the scikit-image contributors: Scikit-image: Image processing in python. PeerJ **2**(e453), 1–18 (2014)
28. Wölfflin, H.: Kunstgeschichtliche Grundbegriffe: Das Problem der Stilentwickelung in der neueren Kunst. Bruckmann, München (1915)