KaoKore: A Pre-modern Japanese Art Facial Expression Dataset

Yingtao Tian
Google Brain
Tokyo, Japan

Chikahiko Suzuki
ROIS-DS Center for Open Data
in the Humanities
NII

Tarin Clanuwat
ROIS-DS Center for Open Data
in the Humanities
NII

Mikel Bober-Irizar
University of Cambridge

Alex Lamb
MILA
Université de Montréal

Asanobu Kitamoto
ROIS-DS Center for Open Data
in the Humanities
NII

Abstract

From classifying handwritten digits to generating strings of text, the datasets which have received long-time focus from the machine learning community vary greatly in their subject matter. This has motivated a renewed interest in building datasets which are socially and culturally relevant, so that algorithmic research may have a more direct and immediate impact on society. One such area is in history and the humanities, where better and relevant machine learning models can accelerate research across various fields. To this end, newly released benchmarks (Clanuwat et al. 2018) and models (Clanuwat, Lamb, and Kitamoto 2019) have been proposed for transcribing historical Japanese cursive writing, yet for the field as a whole using machine learning for historical Japanese artworks still remains largely uncharted. To bridge this gap, in this work we propose a new dataset *KaoKore* [1] which consists of faces extracted from pre-modern Japanese artwork. We demonstrate its value as both a dataset for image classification as well as a creative and artistic dataset, which we explore using generative models.

Introduction

In pre-modern Japan, one well-established narrative form consisted of stories displayed entirely as one long continuous painting, usually in the form of a picture scroll (絵巻物, Emakimono) or a picture book (絵本, Ehon), accompanying cursive writing of the story itself. These stories include diverse arrays of characters (see Figure 1), and thus provide valuable materials for the study of Japanese art history.

In art history research, comparative style study, based on the visual comparison of characteristics in artworks, is a typical approach to answering research questions about works, such as the identification of creators, the period of production, and the skill of the painter. Among many attributes that appear in artworks, facial expressions offer especially rich information not only about the content but also about how the artworks were created. To accelerate comparative studies, *Collection of Facial Expressions* (Suzuki, Takagishi, and Kitamoto 2018) has been created as a collection of cropped faces from multiple artworks with basic metadata annotated manually by art history experts. It was made possible leveraging recent technological developments, such as mass digitization of historical Japanese artworks and image sharing using IIIF (Kitamoto, Homma, and Saier 2018), implemented using JSON-LD, a JSON serialization pattern.

Inspired by the recent success of developing new benchmarks (Clanuwat et al. 2018) as well as new models (Clanuwat, Lamb, and Kitamoto 2019) in the field of Japanese cursive writing, we believe that *Collection of Facial Expressions* provides a largely unexplored opportunity to

[1] Instruction for downloading and using *KaoKore* is available: `https://github.com/rois-codh/kaokore`

Figure 1: A cropped page of *Tale of the Hollow Tree* (宇津保物語, Utsuho Monogatari), a late-10th century Japanese story, presented in the form of a picture book (絵本, Ehon) in the 17th century (Unknown Late 10th Century). Picture scrolls and picture books usually have cursive texts telling the story (left) in addition to story-explaining paintings depicting many characters (right), often on the same page.

Figure 2: On the left is a face in the *Collection of Facial Expressions* showing one crop with annotated metadata “Gender: Male (性別：男)” and “Status: Nobleman (身分：貴族)”. On the right is an example of style comparison using the *Collection of Facial Expressions* , displaying similar faces from two different works (Suzuki, Takagishi, and Kitamoto 2018).

bridge research of historical Japanese artwork with machine learning research. It, however, is not designed for machine learning in terms of data format (JSON-LD), image size (of different and irregular sizes) and attribute values (of both related and unrelated sets). As a result, it presents an obstacle for easy adaption of machine learning techniques.

To bridge this gap, in this work we propose a novel dataset, *KaoKore*, derived from the *Collection of Facial Expressions*, Our contributions can be summarized as follows:

- We process the *Collection of Facial Expressions* to create the new *KaoKore* dataset of face images from Japanese artworks along with multiple labels for each face in a more simple, regular and easy-to-use format.
- We provide standard data loaders for both PyTorch and TensorFlow as well as official train/dev/test splits which make this dataset easy to use across frameworks and compare results with.
- We demonstrate the dataset’s utility for image classification by introducing a number of new baselines.
- We study how different types of generative models can be applied to this dataset and support different types of creative exploration.

We hope that this dataset will help to strengthen the link between the machine learning community and research in Japanese art history.

KaoKore Dataset

We begin with describing *Collection of Facial Expressions* with is the foundation on which we build our work, the *KaoKore* dataset. *Collection of Facial Expressions* results from an effort by the ROIS-DS Center for Open Data in the Humanities (CODH) that has been publicly available since 2018. *Collection of Facial Expressions* provides a dataset of cropped face images extracted from Japanese artwork[2] from the Late Muromachi Period (16th century) to the Early Edo Period (17th century) to facilitate research into art history, especially the study of artistic style. It also provides corresponding metadata annotated by researchers with domain expertise. An example of our cropping process is shown on the left panel of Figure 2.

The *Collection of Facial Expressions* is built upon the International Image Interoperability Framework (IIIF) and IIIF Curation Platform (Kitamoto, Homma, and Saier 2018),

[2]Publicly available from National Institute of Japanese Literature, Kyoto University Rare Materials Digital Archive and Keio University Media Center.

Figure 3: Exemplary images in the *KaoKore* dataset. These examples demonstrate various faces dipicting different subjects in diverse yet coherent artistic styles.

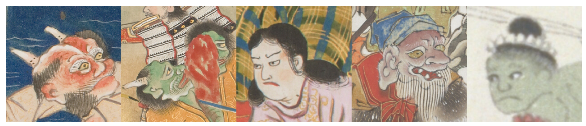
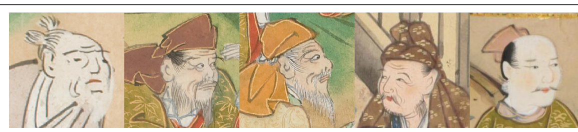
Class	*Labels*	*Examples*
gender (性別)	male (男)	
	female (女)	
status (身分)	noble (貴族)	
	warrior (武士)	
	incarnation (化身)	
	commoner (庶民)	

Table 1: Labels available in the dataset along with exemplary images belonging to each label.

which is a system designed to ease the burden of common tasks in humanities research such as viewing, searching, and annotating documents, by representing documents in a structured and machine-readable data format. *Collection of Facial Expressions* serves as the basis of several research projects in art history. One example, shown on the right panel of Figure 2, is the comparison of similar-looking faces from different artworks. This type of art history research provides insights into the history of trending art styles and is useful for determining authorship of works.

We derive *KaoKore* dataset (see Figure 3 for exemplary images) from *Collection of Facial Expressions* in a form that will be easily recognizable to the machine learning community, with the hope of facilitating dialogue and collaboration with humanities researchers. Concretely, we process the images and labels into industry-standard formats such that the resulting dataset is easy to use with off-the-shelf machine learning models and tools. Since the cropped faces from the *Collection of Facial Expressions* can have different sizes and aspect ratios, we pre-process them to ensure that all cropped images are normalised to the same size and aspect ratio.

The resulting *KaoKore* dataset contains 5552 [3] RGB image files of size 256 x 256. Figure 3 shows examples from the *KaoKore* dataset, a collection of various faces in diverse yet coherent artistic styles. The format follows that of ImageNet (Deng et al. 2009), making it a potential drop-in replacement dataset for existing machine learning setups.

To facilitate supervised learning, we provide two sets of labels for all faces: *gender* and *(social) status*, both from the frequently appearing subset of all expert-annotated labels from the *Collection of Facial Expressions*. Table 1 shows the classes and labels available in the *KaoKore* dataset, with exemplary images for each label. We setup tasks on the labels that appear most frequently. For example for class *(social) status* we choose *noble*, *warrior*, *incarnation* and *commoner* which each has at least 600 images, while discarding rare ones like *priesthood* and *animal*, each having merely a dozen. This is to avoid an overly unbalanced distribution over the labels. We also give official training, validation, and test sets splits to enable model comparisons in future studies.

Model	classifying *gender*	classifying *status*
VGG-11	92.03 %	78.74 %
AlexNet	91.27 %	78.93 %
ResNet-18	92.98 %	82.16 %
ResNet-34	93.55 %	84.82 %
MobileNet-v2	95.06 %	82.35 %
DenseNet-121	94.31 %	79.70 %
Inception-v3	96.58 %	84.25 %

Table 2: Test accuracy on classification tasks. See text for the discussion and citation for each model.

[3] We are experimentally extending the dataset by incorporating more images, which can be seen in our website.

Figure 4: Uncurated images produced by StyleGAN (Karras, Laine, and Aila 2019) trained on our dataset. These samples demonstrate that the varieties in our datasets is well captured.

Experiments

We conduct two types of experiments. First, we provide *quantitative results* on the supervised machine learning tasks of gender and status classification of *KaoKore* images. Second, we provide *qualitative results* from generative models on the *KaoKore* images.

Classification Results for *KaoKore* Dataset

We present classification results for the *KaoKore* dataset in Table 2 using several neural network architectures, namely VGG (Simonyan and Zisserman 2014), AlexNet (Krizhevsky 2014), ResNet (He et al. 2016), MobileNet-V2 (Sandler et al. 2018), DenseNet (Huang et al. 2017) and Inception (Szegedy et al. 2016). We use PyTorch's reference implementation(Paszke et al. 2019) of common visual models, standard data augmentation, and Adam optimizer(Kingma and Ba 2014). We use early stopping on the validation set, and report the test set accuracy.

In providing accuracies across various models we demonstrate that standard classifiers are able to achieve decent but imperfect performance on these tasks. Additionally, we show that newer and larger architectures often achieve better performance, which suggests that even further improvement through better architectures will be possible.

Creativity Applications

As the *KaoKore* dataset is based on artwork, we also investigate its creative applications. While our hope is that people will find novel ways of engaging with this dataset artistically, we demonstrate that reasonable results can be achieved using today's best-performing generative models. We note that faces in the *KaoKore* dataset contain many correlated

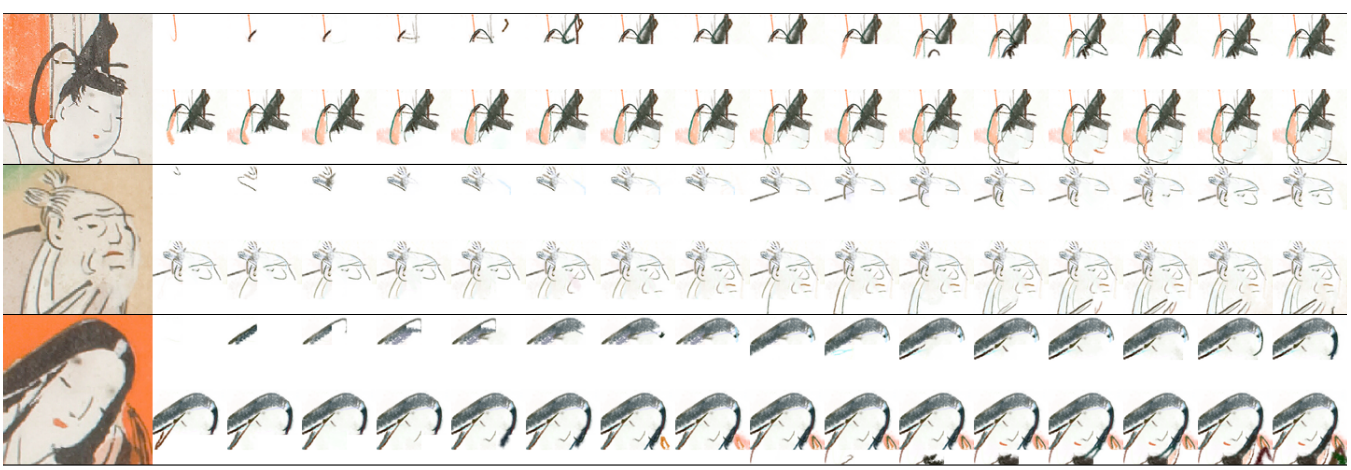


Figure 5: Painting sequences generated by *intrinsic style transfer* (Nakano 2019), a neural painting model, on a few exemplary images in the *KaoKore* dataset. On each row, the leftmost image is the reference image from *KaoKore* dataset, while the smaller images illustrate the generate painting sequence.

Figure 6: The second row shows the final produced canvas from *intrinsic style transfer* (Nakano 2019) after all painting steps; trained to approximate the reference image (first row).

attributes like face shape and hairstyle, giving a challenging tasks of correctly model correlations, yet are highly stylized and simpler than realistic images of faces, creating challenges in modeling data distribution. Since these two challenges are non-trivial for generative models, we hope the *KaoKore* dataset will be useful for generative modeling research.

Figure 7: *intrinsic style transfer* (Nakano 2019) produces more example of final canvas after all painting steps are generated.

Generative Adversarial Networks We first explore Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) which have been successfully used as generative models for synthesizing high quality images(Karras et al. 2017; Karras, Laine, and Aila 2019; Zhang et al. 2018), and has also seen creative applications such as image-to-image translation (Zhu et al. 2017), controllable Anime character generation (Jin et al. 2017; Hamada et al. 2018; Jin and others 2020), photo-realistic face generation (Lu, Tai, and Tang 2018), apartment (Chaillou 2019), and fashion design (Chen et al. 2020).

Inspired by this, we leverage StyleGAN (Karras, Laine, and Aila 2019), a state-of-the-art GAN model for images. We implement and train it on our dataset and show the resulting generated images. In Figure 4, we show uncurated images produced by StyleGAN, showing that the varieties in our datasets are successfully captured by the generative model.

Neural Painting Models We find that GAN models are able to generate somewhat plausible-looking images of *KaoKore* faces (Figure 4). However, the GAN objective requires that the model directly generate *pixels*, where an artist would paint the image by applying strokes it-

Figure 8: Painting sequences produced by *learning to paint* (Huang, Heng, and Zhou 2019), a neural painting model, on a few exemplary images in the *KaoKore* dataset. On each row, the leftmost image is the reference image from the*KaoKore* dataset, while the smaller images illustrate the generated painting sequence. The reference images used are the same as in Figure 5 for easier comparison of style studies.

Figure 9: Final canvases produced by *learning to paint* (Huang, Heng, and Zhou 2019) (second row) after all painting steps have been completed in order to approximate the reference image (first row). Reference images match Figure 6 for easier comparison of style studies.

Figure 10: More examples of *learning to paint* (Huang, Heng, and Zhou 2019) on *KaoKore* images after all painting steps. Reference images match Figure 7 for easier comparison of style studies.

eratively on a canvas. Thus when a GAN makes mistakes, the types of errors it makes are generally quite unlike the variations that could be produced by a human painter. To give the synthesis process a more artwork-like inductive bias, we consider *Stroke-based rendering* (Hertzmann 2003) which produces a reference painting by sequentially drawing primitives, such as simple strokes, onto a canvas. Recent advances using neural networks have been proposed by integrating techniques such as differentiable image parameterizations (Mordvintsev et al. 2018; Nakano 2019) or reinforcement learning(Ganin et al. 2018; Huang, Heng, and Zhou 2019), which can greatly improve the quality of generated painting sequences. Given that our proposed *KaoKore* dataset is based on art collections, we are interested in applying these neural painting methods with image production mechanisms that better resemble human artists. In particular, we explore applying *intrinsic style transfer* (Nakano 2019) and *learning to paint* (Huang, Heng, and Zhou 2019) on the proposed *KaoKore* dataset.

***Intrinsic style transfer* (Nakano 2019)** is a set of methods combining differentiable approximation with non-differentiable painting primitives (e.g. strokes in the colored pencil drawing style) and an adversarially trained neural agent that learns to recreate the reference image on the canvas by producing these primitives sequentially. It is characterized by a lack of "style-loss" that are often used in style transfer methods to carry the reference image's style into the target one, which in turn exposes the intrinsic style derived from painting primitives mentioned above. In Figure 5, we show the produced painting sequences on a few exemplary images. It can be observed that the image has been decomposed into strokes that resemble how a human artist might create the pencil sketch, while the model has not been provided with any recording of sketching sequence. This is further epitomized in Figure 6 and Figure 7, which show the reference images and the final canvas after completing painting.

***Learning to paint* (Huang, Heng, and Zhou 2019)** is a neural painting model which differentiates itself from others in a few aspects, including using regions marked by quadratic Bézier curves as painting primitives as well as leveraging model-based deep reinforcement learning for training. As a result, its painting sequence is made of simple regions rather than brush-like strokes, and the sequence is as short as possible due to the optimization goal used in reinforcement learning training. As shown in Figure 8, 9 and 10, the method learns to assemble simple curve regions in recreating the image that emphasize the general arrangements of objects in the painting and resembles an abstract painting style.

The two neural painting models explored can, given a single image from the *KaoKore* dataset, produce painting sequences that could be applied on a (virtual) canvas and resemble human-interpretable styles. Yet due to each method's fundamentally different mechanism, the style, while being expressive, as a result also resort to different artistic style. By simultaneously presenting artistic familiarity in the style and surprising in how to decompose a single image, we hope this result can provide insights into the study of art styles.

Discussion and Future Work

We hope that our proposed *KaoKore* dataset provides a foundation on which future works can be built: including humanities research of historical Japanese artworks or machine learning research with creative models, given our dataset's dual value both as a classification dataset as well as a creative and artistic one. In the future, we plan to increase the number of facial expression images in our dataset by building a machine learning powered human-in-the-loop annotation mechanism that allows us to scale the labeling process. We would also like to construct new datasets in future work which help to expand machine learning research and its applications for more general Japanese art. Finally, we anticipate that further interdisciplinary collaboration between humanities research and the machine learning research community would contribute to better and more efficient cultural preservation.

References

Chaillou, S. 2019. Archigan: a generative stack for apartment building design. `https://devblogs.nvidia.com/archigan-generative-stack-apartment-building-design/`.

Chen, L.; Tian, J.; Li, G.; Wu, C.-H.; King, E.-K.; Chen, K.-T.; and Hsieh, S.-H. 2020. Tailorgan: Making user-defined fashion designs. *arXiv preprint arXiv:2001.06427*.

Clanuwat, T.; Bober-Irizar, M.; Kitamoto, A.; Lamb, A.; Yamamoto, K.; and Ha, D. 2018. Deep learning for classical Japanese literature. *arXiv preprint arXiv:1812.01718*.

Clanuwat, T.; Lamb, A.; and Kitamoto, A. 2019. KuroNet: Pre-modern Japanese kuzushiji character recognition with deep learning. *arXiv preprint arXiv:1910.09433*.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Ganin, Y.; Kulkarni, T.; Babuschkin, I.; Eslami, S. M. A.; and Vinyals, O. 2018. Synthesizing programs for images using reinforced adversarial learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1666–1675. PMLR.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

Hamada, K.; Tachibana, K.; Li, T.; Honda, H.; and Uchida, Y. 2018. Full-body high-resolution anime generation with progressive structure-conditional generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 0–0.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Hertzmann, A. 2003. A survey of stroke-based rendering. *IEEE Computer Graphics and Applications* 23(4):70–81.

Huang, G.; Liu, Z.; Van Der Maaten, L.; and Weinberger, K. Q. 2017. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4700–4708.

Huang, Z.; Heng, W.; and Zhou, S. 2019. Learning to paint with model-based deep reinforcement learning. *arXiv preprint arXiv:1903.04411*.

Jin, Y., et al. 2020. Crypko. `https://crypko.ai`.

Jin, Y.; Zhang, J.; Li, M.; Tian, Y.; and Zhu, H. 2017. Towards the high-quality anime characters generation with generative adversarial networks. In *Proceedings of the Machine Learning for Creativity and Design Workshop at NIPS*.

Karras, T.; Aila, T.; Laine, S.; and Lehtinen, J. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*.

Karras, T.; Laine, S.; and Aila, T. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4401–4410.

Kingma, D. P., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kitamoto, A.; Homma, J.; and Saier, T. 2018. IIIF Curation Platform: Next generation IIIF open platform supporting user-driven image sharing. In *Proceedings of IPSJ SIG Computers and the Humanities Symposium 2018*, 327–334. (in Japanese).

Krizhevsky, A. 2014. One weird trick for parallelizing convolutional neural networks. *arXiv preprint arXiv:1404.5997*.

Lu, Y.; Tai, Y.-W.; and Tang, C.-K. 2018. Attribute-guided face generation using conditional cyclegan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 282–297.

Mordvintsev, A.; Pezzotti, N.; Schubert, L.; and Olah, C. 2018. Differentiable image parameterizations. *Distill* 3(7):e12.

Nakano, R. 2019. Neural painters: A learned differentiable constraint for generating brushstroke paintings. *arXiv preprint arXiv:1904.08410*.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc. 8024–8035.

Sandler, M.; Howard, A.; Zhu, M.; Zhmoginov, A.; and Chen, L.-C. 2018. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 4510–4520.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Suzuki, C.; Takagishi, A.; and Kitamoto, A. 2018. 'Collection of facial expressions' with IIIF Curation Platform - Close Reading and Distant Reading for Style Comparative Studies. *Proceedings of IPSJ SIG Computers and the Humanities Symposium 2018* 249–256.

Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; and Wojna, Z. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2818–2826.

Unknown. Late 10th Century. Tale of the hollow tree (宇津保物語, utsuho monogatari). National Institute of

Japanese Literature http://codh.rois.ac.jp/pmjt/book/200017526/.

Zhang, H.; Goodfellow, I.; Metaxas, D.; and Odena, A. 2018. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.