A Deep Learning Approach for Painting Retrieval based on Genre Similarity

Tess Masclef, Mihaela Scuturici, Benjamin Bertin, Vincent Barrellon, Vasile-Marian Scuturici, Serge Miguet

A Deep Learning Approach for Painting Retrieval based on Genre Similarity

Tess Masclef, Mihaela Scuturici, Benjamin Bertin, Vincent Barrellon, Vasile-Marian Scuturici, and Serge Miguet

Univ Lyon, Univ Lyon 2, CNRS, INSA Lyon, UCBL, LIRIS, UMR5205

Abstract. As digitized paintings continue to grow in popularity and become more prevalent on online collection platforms, it becomes necessary to develop new image processing algorithms to effectively manage the paintings stored in databases. Image retrieval has historically been a challenging field within digital image processing, as it requires scanning large databases for images that are similar to a given query image. The notion of similarity itself, varies according to user's perception. The performance of image retrieval is heavily influenced by the feature representations and similarity measures used. Recently, Deep Learning has made significant strides, and deep features derived from this technology have become widely used due to their demonstrated ability to generalize well. In this paper, a fine-tune Convolutional Neural Network for the artistic genres recognition is employed to extract deep and high-level features from paintings. These features are then used to measure the similarity between a given query image and the images stored in the database, using an Approximate Nearest Neighbours algorithm to get a real time result. Our experimental results indicate this approach leads to a significant improvement in the performance of content-based image retrieval for the task of genre retrieval in paintings.

Keywords: Art classification · Deep Learning · Content-Based Image Retrieval · Approximate Nearest Neighbours Algorithm.

1 Introduction

Our work is part of the Augmented Artwork Analysis (AAA) project which aims to produce a tool for assisted interpretation of artistic images. This tool has to adapt to a specific museum, by allowing the study of different aspects and levels of organization of an artwork being observed in situ and by cross-linking it with an open access corpus of images to highlight genealogies and to stimulate intertextual dialogues. The goal is to help guides visits, pedagogical activities and scientific research. Our focus is on searching by image similarity and identifying the closest neighbours of the artwork in real-time. In recent years, the increasing digitisation of artworks has led to the creation of datasets from museums and private collections that are available online, like the WikiArt collection[1], which

[1] http://www.wikiart.org

provides expert metadata such as artist, date, artistic genre, style, etc. Louvre[2], Tate[3], Metropolitan[4] and many other museums have on-line galleries. Thus, projects such as Henri Focillon's [6], which tends to bring out the genealogy of forms, unaccomplished due to lack of data, are becoming technically feasible.

The genealogy of forms in painting is the study of the evolution of artistic elements, techniques and styles and their transmission from one artist or period to another. It involves tracing the development and transformation of visual elements, such as composition and the use of colour, in painting. By analysing these links, art historians can understand how artists have built on and responded to the work of their predecessors, contributing to the diversity and constant evolution of painting throughout history.

The literature offers many methods to enable such a study, notably in the exploration of painting based on images, based on textual descriptions or by combining images and textual descriptions. In the context of this project, we are mainly interested in the exploration of paintings based on images and the search for similarities between them. Content-based image retrieval is one of the methods used to obtain artworks similar to the one being explored.

Content-based image retrieval (CBIR) is a technique for searching images on the basis of their visual characteristics. Traditionally, these features are obtained using conventional extraction techniques (HSV histogram, local binary pattern, etc.)[13]. However, we propose a deep learning approach to identify visually similar images. Deep learning has successfully addressed the challenges posed by the semantic gap, which refers to the disparity between the low-level features extracted from images and the high-level semantic meanings perceived by humans. Using deep neural networks, we can extract complex features encompassing texture, colour and composition, exceeding the capabilities of conventional extraction methods [11][19]. These feature vectors can be fed into a query system, represented as a nearest neighbour graph, which suggests similar images. To guide our unsupervised search, we integrate a classification system that matches artworks by artist, genre and style.

In order to allow a supervised search, neural networks have indeed shown their efficacy for artwork processing, mainly for the recognition of a school, an era or a style (art movements) [16][3] [23]. In most cases, these tasks are done in parallel. For the classification of artistic styles with high complexity due to high intra-class variation and low inter-class variation, self-supervised methods are also used [10]. Algorithms for the recognition of different painting techniques such as the clustered multiple kernel learning algorithm that extracts features from three aspects (colour, texture and spatial arrangement) for the recognition of oil paintings can be used for this task [15]. Comparative studies of different pretrained neural networks on ImageNet, ResNet50, ResNet101 [7], and DenseNet [8] showed better performance [12] for artist classification than VGG16 [22].

---

[2] https://collections.louvre.fr/

[3] https://www.tate.org.uk/search?type=artwork

[4] https://www.metmuseum.org/art/the-collection

VGG Face have also shown their efficiency in the recognition of characters in a painting using domain transfer for data augmentation [17]. Siamese networks are used for the recognition of visual links with the help of experts in the annotation of clusters of similar paintings. These networks are more efficient when they are pre-trained [20][1]. A CNN approach to estimate the pose of characters in a painting has shown that the pose criterion can be effective in finding visual links between artworks [9].

Several approaches exist to perform a nearest neighbour search. The most classical one is to use an exact algorithm that from a chosen distance calculates each of the distances between the features of the query image and those of the dataset. Another approach is to use an Approximate Nearest Neighbour algorithm which has the advantage of being faster with relatively low error using tree forests, Voronoi diagrams [2], or other partitioning methods[14].

In this paper, we present an architecture composed of a neural network that has the task of classifying paintings by genre and that is used as a feature extractor combined with an Approximate Nearest Neighbour algorithm allowing a real-time search of similar images, with a quality close to an exact algorithm.

We obtain a more refined search after fine-tuning the model. Indeed, the search for similar neighbours focus more on the type of requested painting. Furthermore, optimising parameters for the Approximate Nearest Neighbour algorithm shows that we can obtain a search close to the exact one with a considerably shorter time than with an exhaustive approach with statistical guarantees for the quality of the results.

2 Methodology and experiments

In this section, we provide the architecture of the Convolutional Neural Network used. We briefly present the dataset and the functionality of the approximate nearest neighbour algorithm.

2.1 Convolutional Neural Network

We use the Resnet50 network as our feature extractor. After testing several pre-trained networks, ResNet50 is the one we found to be the most visually relevant. ResNet50 is a 50-layer Convolutional Neural Network (48 convolutional layers, one MaxPool layer, and one average pool layer).

To fine-tune ResNet50, we keep the convolutional layers (the encoder gives a vector of 2048 components as output) and use an average pooling followed by a fully connected layer with an output space size of 512 and a second fully connected layer using the activation Softmax for the classification. The idea behind the addition of the first fully connected layer permits to reduce the dimension of the feature vector.

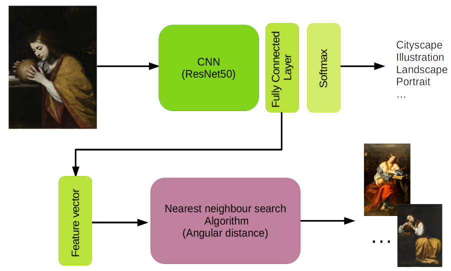


Fig. 1. Schematic representation of our similarity-based image search tool

2.2 Dataset

We use two datasets: the base ResNet50 is pre-trained on ImageNet [5] and the network is fine-tuned on WikiArt for genre recognition.

Among various subsets of the ImageNet dataset, the most used is "ImageNet Large Scale Visual Recognition Challenge (ILSVRC) $2012-2017$ image classification and localization dataset". A database composed of about 1.5 million annotated images and grouping $1,000$ object categories.

WikiArt from WikiArt.org is a visual art encyclopedia which contains $81,444$ artworks (paintings) covering the periods between the $15th$ and $20th$ centuries [18] (for our application, we are interested in artworks dating from before the beginning of the $20th$ century). They are labelled based on artist (129), genre (abstract painting, cityscape, genre painting, illustration, landscape, nude painting, portrait, religious painting, sketch and study and still life) and style (a total of 27 styles like romanticism, baroque, impressionism, cubism, realism, etc.). In this work, we are going to train the model to perform genre-based classification. In the WikiArt dataset, because the classes are unbalanced we chose to randomly sample of around 2500 images per class for fine-tuning, a total of $27,613$ images.

2.3 Nearest Neighbour Algorithm and similarity measure

To expect real time performances, we have chosen to use an Approximate Nearest Neighbour algorithm. In fact, the computation time of the distance between the features of the query image and those of the $81,444$ images of the dataset is long, knowing that these features are in a 512 dimensional space. The choice is ANNOY (Approximate Nearest Neighbour Oh Yeah)[5], developed in 2015 for the Spotify platform by Erik Bernhardsson, this algorithm has the advantage of being efficient in high dimensional spaces. ANNOY is selected due to its exceptional search efficiency and resilience in handling various datasets, along with its straightforward customization of hyperparameters. Another advantage of ANNOY is that it can achieve a superior balance between search performance and

[5] https://github.com/spotify/annoy

index size/construction time in comparison to proximity graph-based methods. This is because ANNOY allows for a reduction in the number of trees without compromising the search performance significantly [14]. ANNOY uses a tree forest to organise the vectors in the data space[6]. The forest is traversed in order to obtain a set of candidate points from which the closest to the query point is returned.

The next step is to calculate all distances and rank the points. Finally, the nodes are sorted by distance, so it can return the k-Nearest Neighbours (kNN). The distance used is the angular distance, obtained from the cosine similarity measure, between the feature vector of the query image and all the feature vectors of the dataset, to obtain the nearest neighbours of the query. This distance is defined as:

$$Cosine\ similarity = 1 - \frac{q.p}{||q||^2||p||^2}$$

$$Angular\ distance = \sqrt{2 * Cosine\ similarity}$$

where q is the feature vector of the request image and p, the feature vector of the image in the dataset.

The assembly of these two algorithms is shown in figure 1.

2.4 Experiments

In a first step, we fine-tune ResNet50 pre-trained on ImageNet, with the WikiArt dataset, using Adam as an optimizer, with a learning rate of 10^{-5}. For the loss function, we chose the categorical cross entropy [24] and train on 50 epochs. Since we have chosen to downsample our data to rebalance the classes, we increase the data by making small rotations (of the order of 10 degrees). For the repartition of the data, we consider 80% for the training and 20% for the test. Once the model is pre-trained for genre recognition in paintings, it is used as a feature extractor by removing the Softmax layer. For each image, we obtain a signature formed of a feature vector of size 512.

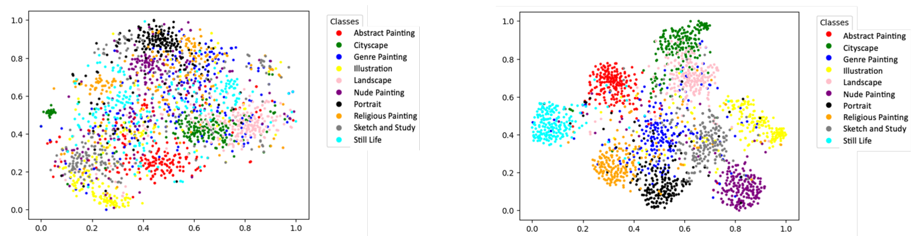
Each of these vectors is associated and indexed to an image. They are then given to the Approximate Nearest Neighbour algorithm, whose parameters are 25 for the number of trees in the forest, *angular distance* for the distance and 1,000 for the number of nearest neighbours searched, parameters we are seeking to optimise in section **3.3**.

In order to optimise these parameters, we search for a compromise between the execution time and the desired quality.

3 Results

In this section, we will compare the performance of our approach with state-of-the-art methods. We show that fine-tuning our model allows us to guide our

[6] https://erikbern.com/2015/10/01/nearest-neighbors-and-vector-models-part-2-how-to-search-in-high-dimensional-spaces.html



(a) T-SNE visualisation before fine-tuning

(b) T-SNE visualisation after fine-tuning

Fig. 2. T-SNE visualisation before and after fine-tuning ResNet50 for artistic genre recognition

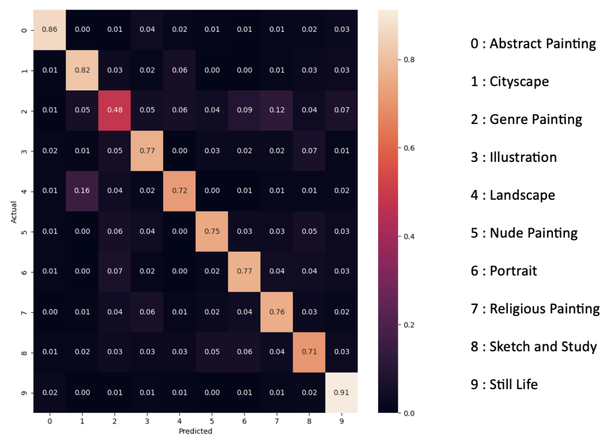


Fig. 3. Confusion matrix on WikiArt test data

search according to genre recognition. Indeed, even if we know the genre of most of the images, some experts do not agree. This could explain why some genres get confused, as shown on figure 2.

Genre recognition helps to list and classify paintings for quicker access to artworks of the same genre. Artistic genre recognition ensures that query results give priority to artworks of the same genre, whereas without it, results may not match the desired genre preference. The aim is not to classify but to guide the search, by maintaining visual similarity. This is why the use of neural networks is useful for extracting visual information from images in addition to genre recognition. We also show that by optimising the parameters of the Approximate Nearest Neighbour algorithm, we can obtain results close to the exact algorithm while keeping a reasonable execution time.

3.1 Classifier performance

Our first experiment evaluates the performance of our model in genre classification. Figure 3 represents the confusion matrix of the 10 genre classes of the WikiArt database.

The results show that the network performs relatively better on some classes than others because of their distinct visual appearance, such as still life (a 91% recall), abstract painting(a 86% recall) and cityscape (a 82% recall). In contrast, the network perfoms the worst for the genre painting class, mainly due to the fact that this class includes paintings featuring scene of anecdotal or familiar character (the illustration of scenes from everyday life, where the characters are anonymous human types) and can be confused with other genres such as religious painting (12%) or portraits (9%), as can be seen in the figure 3 representing the confusion matrix for the WikiArt dataset for each class.

The model obtains an accuracy of 75.12% with a standard deviation of 0.07 (with 97.27% for top 3 classes) on the validation data. In the table 1, each model is trained on a number of images from WikiArt and with a class number equal to 10. It can be seen that despite a smaller amount of data due to class rebalancing, we obtain an accuracy close to that obtained with a larger amount of data.

Table 1. Genre recognition: a State-of-the-art

	genre	
References	Samples	Acc (%)
Tan et al. [21]	66,993	74.14
Saleh et al.[18]	63,691	60.28
Cetinic et al.[4]	86,087	77.6
Zhong et al. [26]	28,760	76.27
Zhao et al. [25]	64,993	78.03
Ours	27,613	75.12

3.2 Comparison of CBIR performance before and after fine-tuning with specific domain knowledge

Our next experiment aims at comparing the performance of content-based image search before fine-tuning and after fine-tuning. Before fine-tuning, only the convolutional layers of ResNet50 are used for feature extraction, i.e. vectors of dimensions 2048. We make this comparison for all the test data on all genres and then on each genre treated separately. The results are presented in table 2.

We can observe a clear improvement in performance after fine-tuning. Indeed, among the 1,000 nearest neighbours found, the percentage of images belonging to the same class as the query image is higher after fine-tuning. The average improvement is 7%. Figures 4(a) and 4(b) show the query image belonging to religious painting genre with the seven most similar images. After the fine-tuning

Table 2. Percentage of images of the same genre as the query image in the nearest neighbours list before and after fine-tuning

	With 1,000 neighbours	
	Before fine-tuning	After fine-tuning
Genre	31.04%	38.09%
Abstract Painting	5.49%	19.24%
Cityscape	8.55 %	12.29%
Genre Painting	14.15%	14.62%
Illustration	2.03%	3.08%
Landscape	18.44 %	29.77%
Nude Painting	1.60%	2.66%
Portrait	23.50%	23.82%
Religious Painting	9.47%	14.50%
Sketch and Study	4.32%	4.66%
Still Life	2.47%	6.67%

(a) Before fine-tuning (religious painting)

(b) After fine-tuning (religious painting)

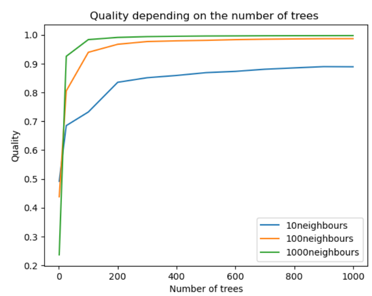
(c) Before fine-tuning (genre painting)

(d) After fine-tuning (genre painting)

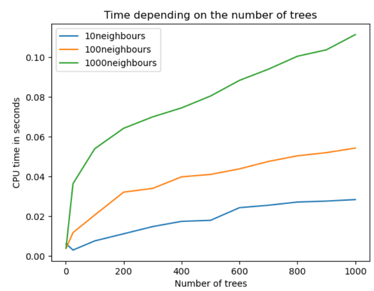
Fig. 4. CBIR performance on an example before and after fine-tuning on a religious painting and on a genre painting

of ResNet50, we can see that the model fine-tuned for genre recognition focuses more on genre attributes. Figures 4(c) and 4(d) show that this is also revealing in the case of a genre painting. Among the classes most dependent on fine-tuning (with the best recalls), there is a clear improvement, such as an increase of 14% for abstract painting, 4% for still life painting, or 10% for landscapes. Despite some incorrect class images, the CBIR tool can still find similar paintings in terms of contents by including some elements (such as colour, texture, etc.) and similar compositions.

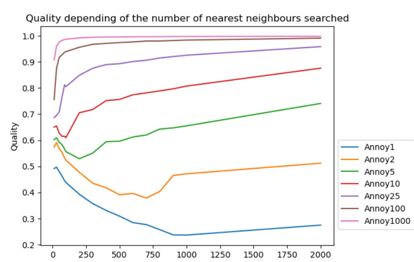
3.3 Parameters optimization of the Approximate Nearest Neighbour Algorithm



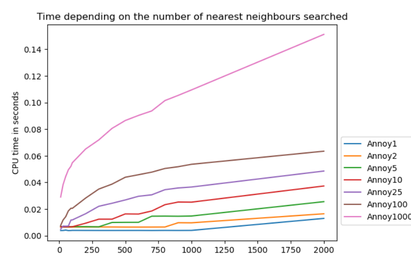
(a) Quality depending on the number of trees



(b) Time depending on the number of trees



(c) Quality depending on the number of nearest neighbours searched



(d) Time depending on the number of nearest neighbours searched

Fig. 5. Optimisation of the number of trees and of the number of nearest nieghbours

We use an Approximate Nearest Neighbour algorithm, mainly for its speed of execution. We are interested in optimising its parameters, with the objective of getting the best possible list of kNN while keeping low execution times.

The parameters to be taken into account are the number of trees, the number of nearest neighbours searched and the execution time.

Let $(i_1, ..., i_n)$ be the image database sorted by increasing distance with a request image q. The list of k nearest neighbours of q is $(i_1, i_2 ... i_k)$, and the list of k-approximate nearest neighbours of q is $(i_{n_1}, i_{n_2}, ... i_{n_k})$, with n_i being an increasing integer function from $[1...k]$ to $[1...n]$. We define the quality $\mathcal{Q}$ of the request as:

$$\mathcal{Q} = \frac{k}{n_k}$$

Figure 5 shows that as soon as the number of trees is larger than 5, the quality increases with the number of trees and with the number of neighbours, and rapidly tends to 1. This means that the list of approximate nearest neighbours approaches the exact one. However, asking for a high number of neighbours or using a high number of trees requires a higher computation time. For our application, we suggest to fix the number of nearest neighbours and to make the number of trees vary to have the best possible quality, while keeping our time constraints under a acceptable threshold.

Suppose we want the 250 nearest neighbours with an execution time of less than 0.05 seconds. And taking into account the experiments done on figure 5. Figure 5(d) shows that we can choose a number of trees between 1 and 100. However, to obtain a result with a quality higher than 0.9 then the choice is made on a number of trees equal to 100, as shown on figure 5(c). Indeed, we obtain a quality of about 0.95 in about 0.03 seconds in CPU time on a MacBook Pro with an Apple $M1$ Max chip, 10 core and 64 Go RAM, which is much powerful than the final target device of our project, a tablet.

3.4 Introducing SimArt: a web application for efficiently searching similar artworks

To demonstrate the effectiveness of our approach, we developed SimArt[7], a web application for searching similar artworks. This application enables to explore wide and diverse art collections, spanning across different styles and eras, based on a nearest neighbour, image-based search engine. The request image can be a member of the dataset or can be uploaded by user.

We used two types of ANNOY indexes to enable efficient nearest neighbour search: one with features obtained from a pre-trained ResNet50 and the other with features obtained from a fine-tuned ResNet50 for genre recognition. The indexes were constructed using a 100 nodes tree structure to allow for efficient approximate nearest neighbour search.

We believe that SimArt can be a valuable resource for art enthusiasts and professionals alike, allowing for exploration and discovery of artworks across various styles and eras.

4 Discussion and conclusions

In this work, we presented a content-based image retrieval tool using Convolutional Neural Networks for feature extraction to measure similarity between images taking into account the genre of the paintings. We fine-tuned ResNet50 for genre recognition with an accuracy of 75%. We applied the features extracted from genre recognition model to increase the proximity of retrieved images belonging to the same genre as the request image. To obtain a real-time search, we used an Approximate Nearest Neighbour algorithm. Indeed, we were able to

[7] https://simart.datavalor.com

obtain the results with a factor of 1000 times faster (i.e. 0.04 seconds in CPU time instead of more than 30 seconds, for $81,444$ images). These results allow us to project how to guide a search for nearest neighbours, taking into account for example the style or the artist. The performance of classification by genre can still be improved, which would allow to better distinguish similar images by genre. This can be done by enriching the data from other collections or by using attention modules or using another pre-trained model. We plan to use style and artist recognition to offer several search dimensions, while having the possibility to do a multi-criteria search (e.g. by selecting style and genre), by doing multi-label image recognition with Graph Convolutional Networks. We also plan to refine the search using Siamese networks by proposing groups of images of the same and different classes, in order to observe whether this improves the search. We would also like to refine the search by including the detection and location of objects representing strong symbolism in the history of art.

5 Acknowledgement

This work was funded by french national research agency with grant ANR-20-CE38-0017.

References

1. Alaasam, R., Kurar, B., El-Sana, J.: Layout analysis on challenging historical arabic manuscripts using siamese network. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 738–742. IEEE (2019)
2. Aurenhammer, F.: Voronoi diagrams—a survey of a fundamental geometric data structure. ACM Computing Surveys (CSUR) **23**(3), 345–405 (1991)
3. Castellano, G., Vessio, G.: Deep learning approaches to pattern extraction and recognition in paintings and drawings: An overview. Neural Computing and Applications **33**(19), 12263–12282 (2021)
4. Cetinic, E., Lipic, T., Grgic, S.: Fine-tuning convolutional neural networks for fine art classification. Expert Systems with Applications **114**, 107–118 (2018)
5. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
6. Focillon, H.: Vie des formes. Librairie Ernest Leroux Paris (1934)
7. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
9. Jenicek, T., Chum, O.: Linking art through human poses. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 1338–1345. IEEE (2019)
10. Joshi, A., Agrawal, A., Nair, S.: Art style classification with self-trained ensemble of autoencoding transformations. arXiv preprint arXiv:2012.03377 (2020)

11. Kapoor, R., Sharma, D., Gulati, T.: State of the art content based image retrieval techniques using deep learning: a survey. Multimedia Tools and Applications **80**(19), 29561–29583 (2021)
12. Kelek, M.O., Calik, N., Yildirim, T.: Painter classification over the novel art painting data set via the latest deep neural networks. Procedia Computer Science **154**, 369–376 (2019)
13. Latif, A., Rasheed, A., Sajid, U., Ahmed, J., Ali, N., Ratyal, N.I., Zafar, B., Dar, S.H., Sajid, M., Khalil, T.: Content-based image retrieval and feature extraction: a comprehensive review. Mathematical problems in engineering **2019** (2019)
14. Li, W., Zhang, Y., Sun, Y., Wang, W., Li, M., Zhang, W., Lin, X.: Approximate nearest neighbor search on high dimensional data—experiments, analyses, and improvement. IEEE Transactions on Knowledge and Data Engineering **32**(8), 1475–1488 (2019)
15. Liao, Z., Gao, L., Zhou, T., Fan, X., Zhang, Y., Wu, J.: An oil painters recognition method based on cluster multiple kernel learning algorithm. IEEE Access **7**, 26842–26854 (2019)
16. Liu, S., Yang, J., Agaian, S.S., Yuan, C.: Novel features for art movement classification of portrait paintings. Image and Vision Computing **108**, 104121 (2021)
17. Madhu, P., Kosti, R., Mührenberg, L., Bell, P., Maier, A., Christlein, V.: Recognizing characters in art history using deep learning. In: Proceedings of the 1st Workshop on Structuring and Understanding of Multimedia heritAge Contents. pp. 15–22 (2019)
18. Saleh, B., Elgammal, A.: Large-scale classification of fine-art paintings: Learning the right metric on the right feature. arXiv preprint arXiv:1505.00855 (2015)
19. Saritha, R.R., Paul, V., Kumar, P.G.: Content based image retrieval using deep learning process. Cluster Computing **22**, 4187–4200 (2019)
20. Seguin, B., Striolo, C., diLenardo, I., Kaplan, F.: Visual link retrieval in a database of paintings. In: Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part I 14. pp. 753–767. Springer (2016)
21. Tan, W.R., Chan, C.S., Aguirre, H.E., Tanaka, K.: Ceci n'est pas une pipe: A deep convolutional network for fine-art paintings classification. In: 2016 IEEE international conference on image processing (ICIP). pp. 3703–3707. IEEE (2016)
22. Tan, W.S., Chin, W.Y., Lim, K.Y.: Content-based image retrieval for painting style with convolutional neural network. The Journal of The Institution of Engineers, Malaysia **82**(3) (2021)
23. Yang, Z.: Classification of picture art style based on vggnet. In: Journal of Physics: Conference Series. vol. 1774, p. 012043. IOP Publishing (2021)
24. Zhang, Z., Sabuncu, M.: Generalized cross entropy loss for training deep neural networks with noisy labels. Advances in neural information processing systems **31** (2018)
25. Zhao, W., Zhou, D., Qiu, X., Jiang, W.: Compare the performance of the models in art classification. Plos one **16**(3), e0248414 (2021)
26. Zhong, S.h., Huang, X., Xiao, Z.: Fine-art painting classification via two-channel dual path networks. International Journal of Machine Learning and Cybernetics **11**, 137–152 (2020)