IOP Conference Series: Materials Science and Engineering

PAPER • **OPEN ACCESS**

Evaluation of visual descriptors for painting categorisation

To cite this article: Francesco Bianconi and Raquel Bello-Cerezo 2018 *IOP Conf. Ser.: Mater. Sci. Eng.* **364** 012037

IOP Conf. Series: Materials Science and Engineering **364** (2018) 012037 doi:10.1088/1757-899X/364/1/012037

Evaluation of visual descriptors for painting categorisation

Francesco Bianconi[1] **and Raquel Bello-Cerezo**[1]

[1]Department of Engineering, Università degli Studi di Perugia, Via Goffredo Duranti, 93 – 06125 Perugia, Italy

E-mail: bianco@ieee.org, bellocerezo@gmail.com

Abstract. The ever increasing availability of digital data from the Arts and cultural heritage calls for efficient methods to organise, categorise, and retrieve such information in an effective and reliable way. In this context, painting classification has attracted significant research interest in recent years. In this work we address the problem of style classification, which involves determining the school, period and art movement to which a painting belongs. Notably, this job is peculiarly different from other machine vision applications – such as material, object and scene recognition – since the concept of 'similarity' is much more difficult to define in this case. For this specific task we evaluate, in this study, the effectiveness of an array of hand-designed visual descriptors against a set of feature extractors based on last-generation convolutional neural networks. We also investigate the effect of pre-processing methods such as image split and pyramidal decomposition. The experiments are based on the open-access Pandora dataset. The results show that pre-trained models can significantly outperform hand-designed descriptors with overall accuracy surpassing 67%. This represents an improvement on the state-of-the-art by ≈12 percentage points.

1. Introduction

The Internet has long established itself as a common, open platform for cultural activities, and its role in shaping and factoring cultural interests can hardly be objected today [1]. Museums have not been oblivious to the potential of the Internet for reaching new users and, as a consequence, the availability of visual artistic data in digital format has increased dramatically in recent years. If, on the one hand, this calls for efficient methods to organise, categorise, and retrieve such information in an effective and reliable way, it makes, on the other, artistic data amenable of being processed by last-generation machine learning and data mining methods [2]. Within this context the present work is concerned with painting style classification. Notably, this job is peculiarly different from other machine vision tasks, as are for instance material classification, object and face recognition (see Refs. [3, 4] for a review), for the idea of *similarity* is more challenging to define in this context [5]. The very same concept of *style* is rather vague and imprecise: the Oxford English Dictionary defines this as the 'way of painting, writing, composing, building, etc., characteristic of a particular period, place, person or movement'. From a practical standpoint, however, few would object that among the visual elements that play a major role in making a style different from another are: the use of colour and its distribution across the painting, the shape and spatial layout of the geometric elements and the texture of the stroke.

Moving from these premises, various authors have tackled the problem of style categorisation

using different techniques. Zujovic *et al.* [6] for instance proposed a combination of grey-scale features from steerable filters and colour statistics in the HSV space. Khan *et al.* [7] comparatively evaluated the performance of a set of global and local hand-crafted visual descriptors, and obtained the best results with a combination of Scale-Invariant Image Transform (SIFT) and colour names. Florea *et al.* [8] assessed a number of hand-designed descriptors for painting style classification and achieved the best performance with a combination of Local Binary Patterns (LBP) and Colour Structure Descriptor (CSD). Agarwal *et al.* [9] benchmarked five grey-scale local image descriptors and colour features obtaining the best performance with a combination of SIFT and colour histogram in the CIELab space. In recent years convolutional neural networks (CNNs) have dramatically changed the outlook in many computer vision tasks [10, 4] and therefore have been attracting much interest in computer analysis of visual arts too. Saleh and Elgammal [11] recently investigated the use of a pre-trained net along with metric learning schemes for large-scale classification of fine-art paintings. For the same task Tan *et al.* [5] proposed an end-to-end CNN trained from scratch. Of late, a deep multi-branch convolutional network for artist recognition and style classification has been proposed by Bianco *et al.* [12].

In this work we investigated the effectiveness of pre-trained convolutional networks as opposed to traditional, hand-crafted descriptors for painting style classification. We considered nine hand-designed visual descriptors combined with three pre-processing schemes on one side, and five pre-trained convolutional networks on the other. The pre-trained models surpassed the hand-designed descriptors by a large margin, allowing to reach overall accuracy in excess of 67%. This figure marks an improvement of ≈12 percentage points compared with the best result obtained in [8].

In the remainder of the paper we first describe the materials and methods used in this study (Secs. 2–3), then detail the experimental set-up in Sec. 4. We present and discuss the results in Sec. 5 and conclude the paper with some final considerations in Sec. 6.

2. Materials

We based our experiments on the recently released 'Pandora' dataset [8, 13]. This open-access, digital collection contains a total of 7724 images of paintings organised into 12 classes representing as many art movements: *Abstract expressionism*, *Baroque*, *Cubism*, *Fauvism*, *High Renaissance*, *Iconoclasm*, *Impressionism*, *old Greek pottery*, *Realism*, *Rococo*, *Romanticism* and *Surrealism*. The number of samples for each class is detailed in Fig. 1. The size, spatial resolution and shape factor (width/height) vary from one image to another; the imaging conditions such as illumination, camera/scanner type and acquisition settings are unknown.

3. Methods

We used a standard image classification pipeline based on three steps: image pre-processing, feature extraction and classification.

3.1. Image pre-processing

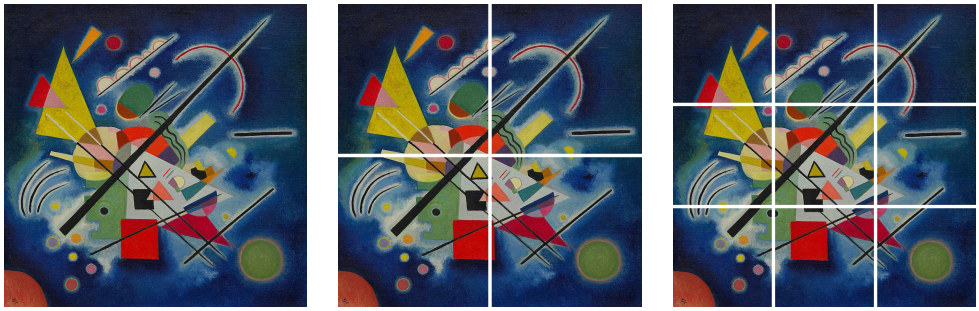
Image preprocessing consisted of either image split, pyramidal decomposition or no pre-processing whatsoever. No further preliminary operations such as colour normalisation, histogram equalisation, filtering or deblurring were applied.

3.1.1. Image split The original image is partitioned into s sub-images of equal size (Fig. 2). The visual descriptors are extracted from each sub-image separately and then concatenated into one single feature vector. As a result, the dimensionality of the original descriptor increases s-fold. In the experiments we used $s = 4$ and $s = 9$, which we indicate with subscripts '4×' and '9×' in the remainder.

Figure 1: Sample images from the 12 classes of the Pandora dataset. The figures in parentheses indicate the number of images for each class.

3.1.2. Pyramid decomposition A set of copies of the input image is obtained via recursive filtering and sub-sampling (Fig. 3). This enables a multi-scale analysis of the input image [14, 15]. Herein we used a Gaussian decomposition into four levels (including the original image – level 'zero'). The visual descriptors are extracted from each downsampled image separately and concatenated into one feature vector. The dimensionality of the original visual descriptor increases four-fold in this case. We shall indicate this option with with subscript '$p4$' in the remainder.

3.2. Visual descriptors for feature extraction
Visual descriptors can be classified as 'hand-designed' (also referred to as 'engineered' or 'hand-crafted') or based on deep learning [16, 17]. In the first group most of the mathematics that makes up a descriptor (i.e.: functional form and parameters) is established a priori. As a result such descriptors usually require little or no training. By contrast, deep learning relies on large sets of parameters the values of which need to be determined by training. The advantage, in



a) Original image b) Partition into four sub-images (‘4×’) c) Partition into nine sub-images (‘9×’)

Figure 2: Image split.

Figure 3: Pyramid decomposition (‘*p*4’). From left to right: original image and downscaled images.

this case, is that a descriptor can be tailored to a specific application – and this can make it very accurate in that domain; the disadvantage that training requires large amounts of data and adequate computational resources. Fortunately, however, deep networks trained for specific tasks can be used effectively in completely different tasks as well (transfer learning [10]).

3.2.1. Hand-designed visual descriptors Hand-designed visual descriptors are generally classified into three main groups [3]: *spectral*, *spatial* and *hybrid* methods. Descriptors of the first group only consider the colour content of an image regardless of its spatial distribution. Conversely, those of the second disregard colour and only take into account the spatial variation of the image luminance (grey-scale intensity). Finally, hybrid methods combine spectral and spatial data in different ways. In this study we considered:

- Two spectral methods (joint colour histogram and marginal colour histograms)
- Three spatial methods (Histograms of Gradients – HOG, Local Binary Patterns and Improved Local Binary Patterns)
- Four hybrid methods (colour HOG, Local Colour Vector Binary Patterns, Improved Opponent Colour Local Binary Patterns and Opponent Colour Local Binary Patterns).

We briefly recall the basics of each descriptor here below and refer the interested reader to the references for in-depth explanations.

Colour histogram (ColHist) Joint, three-dimensional colour histograms in the RGB space [18]. Each colour channel was quantised into eight bins, which gives $8^3 = 512$ features.

Marginal colour histograms (MargHists) Concatenated marginal histograms of the intensity levels in each of the R, G and B channels [19]. In this case each colour was quantised into 256 bins, which results in a feature vector of dimension $256 \times 3 = 768$.

Histograms of oriented gradients (HOG) Histograms of the orientation of the local grey-scale gradient [20]. This was estimated using generalised Sobel filters at resolution 3px $\times$ 3px, 5px $\times$ 5px and 7px $\times$ 7px. The orientation was quantised into 64 bins. The number of features therefore is $3 \times 64 = 192$.

Colour HOG (ColourHOG) Histograms of oriented gradients extracted from each of the R, G and B channels. In this case the number of features is $3 \times 64 \times 3 = 576$.

Local Binary Patterns (LBP) Concatenation of directional, Local Binary Patterns histograms [21] computed over non-interpolated, eight-point digital circles of radius 1px, 2px and 3px (see [22] for details). This configuration generates $256 \times 3 = 768$ features.

Improved Local Binary Patterns (ILBP) Variant of LBP in which the local thresholding scheme is point-to-average instead of point-to-point [22]. The other settings were the same as LBP. The number of features in this case is $511 \times 3 = 1533$

Local Colour Vector Binary Patterns (LCVBP) Concatenation of Colour Norm Patterns (CNP) and Colour Angular Patterns (CAP) with the latter extracted from each of the R-G, R-B and G-B channel pairs [23]. The other settings were the same as LBP. Since CNP generates the same number of features as LBP and CAP three times as many, the final number of features is $256 + 256 \times 3 = 1024$.

Opponent Colour Local Binary Patterns (OCLBP) A colour variant of LBP obtained by applying LBP to each colour channel separately and in a pairwise manner between each of the R-G, R-B and G-B channel pairs [24]. This variant multiplies by six the dimensionality of LBP, therefore generates $256 \times 6 = 1536$ features.

Improved Opponent Colour Local Binary Patterns (IOCLBP) An extesion of ILBP to the colour domain based on the same scheme as OCLBP's [22]. The number of features is $511 \times 3 + 512 \times 3 = 3069$.

3.2.2. Deep learning We considered five pre-trained deep network models (details below) and, for each model, used as image features the L_1-normalised output of the last fully-connected layer. This is usually referred to as the 'FC' configuration (see also Refs. [25, 17] on this point). Pre-processing only involved resizing the input image to 224px $\times$ 224px – the size of the receptive field of all the models considered. The implementation was based on the pre-trained models available on MatConvNet [26] with no further modification whatsoever.

GoogLeNet (GoogLeNet) A deep network with 22 learnable layers based on the 'Inception' architecture as described in [27]. The dimension of the feature vector is 1024.

ResNet 50 (ResNet-50) A very deep network with 50 learnable layers based on the 'residual' architecture [28]. The dimension of the feature vector is 2048.

ResNet 101 (ResNet-101) A very deep network with the same overall architecture as ResNet-50's but with 101 learnable layers instead of 50 [28]. The dimension of the feature vector is 2048.

VGG very deep 16 (VGG-VD-16) A deep network for large-scale image recognition based on 16 learnable layers and small (3×3) convolution filters [29]. The dimension of the feature vector is 4096.

VGG very deep 19 (VGG-VD-19) A network with a structure very similar to VGG-VD-16's but with 19 learnable layers instead of 16 [29]. The number of features is 4096.

3.3. Classification

The feature vectors generated by the visual descriptors presented in Secs. 3.2.1–3.2.2 were fed to nearest neighbour and support vector classifiers. We used L_1 (cityblock) and L_2 (Euclidean) distance for the nearest neighbour (1-NN) classifier and Gaussian radial basis kernel for the support vector machine (SVM). Multi-class extension of SVM was based on the one-vs-one architecture; optimal box constraint (C) and kernel scale (γ) were determined via trial and error (final values were $C = 1000$ and $\gamma = 1$).

4. Experiments

We carried out a set of supervised image classification experiments to evaluate the effectiveness of each combination pre-processing/visual descriptor/classifier. To avoid excessively long feature vectors ('curse of dimensionality') pre-processing via image split and/or pyramid decomposition was limited to those descriptors that generate less than 1000 features. Accuracy estimation was based on four-fold cross-validation using the same partitions that come in bundle with the Pandora dataset. In each experiment a single fold was retained as the test set, and the remaining three were used as training data. The estimated accuracy was the fraction of paintings of the test set classified correctly. The results (Tabs. 1–2) were averaged over the four folds.

5. Results and discussion

As can be seen, the visual descriptors based on pre-trained CNNs outperformed the hand-designed ones by a large margin. This is remarkable, especially if we consider that: a) we used CNN models trained for other tasks (i.e. object recognition) with no adjustment whatsoever, and b) the input images had to be resized heavily to fit the visual field of each pre-trained net. In absolute terms the overall was in excess of 67%, which marks an improvement of more than 12 percentage points with respect to previous benchmarks [8].

Among the hand-designed methods, the best performance (45.7%) was obtained through joint colour histogram and preliminary image split ('9×'). This outcome is also interesting in that it shows that colour alone plays an important role in discriminating between styles. The results of the other hand-crafted descriptors (i.e. LBP, HOG and colour HOG) were otherwise in line with those available in the literature [8].

Table 1: Overall classification accuracy (in %) of the hand-designed visual descriptors. Boldface and underlined figures respectively indicate the best overall accuracy and the best accuracy for each group.

Descriptor	Classifier		
	NN_{L1}	NN_{L2}	SVM_{ovo}
Spectral methods			
ColHist	35.7	34.0	39.4
$ColHist_{\times 4}$	36.5	32.2	43.8
$ColHist_{\times 9}$	37.7	33.1	**45.7**
MargHists	30.9	30.9	27.7
$MargHists_{\times 4}$	32.8	30.8	35.0
$MargHists_{\times 9}$	33.8	31.6	38.7
Spatial methods			
HOG	27.6	28.0	15.9
$HOG_{\times 4}$	29.1	28.3	21.4
$HOG_{\times 9}$	26.8	26.5	28.0
HOG_{p4}	24.7	23.7	22.5
ILBP	39.5	35.8	20.8
$ILBP_{\times 4}$	39.3	36.5	31.7
$ILBP_{\times 9}$	36.5	35.3	39.0
$ILBP_{p4}$	<u>41.7</u>	38.9	37.0
LBP	38.2	35.0	21.3
$LBP_{\times 4}$	38.9	35.7	32.5
$LBP_{\times 9}$	36.3	34.6	38.2
LBP_{p4}	37.8	35.4	35.6
Hybrid methods			
ColourHOG	28.2	28.0	19.1
$ColourHOG_{\times 4}$	29.1	28.3	25.2
$ColourHOG_{\times 9}$	26.0	26.3	34.9
$ColourHOG_{p4}$	25.0	23.8	28.0
LCVBP	<u>40.8</u>	38.5	23.7
IOCLBP	36.5	26.7	25.7
OCLBP	36.4	29.1	26.6
Accuracy of random classifier (blind) = 8.3			
Accuracy of random classifier (with priors) = 9.9			

Table 2: Overall classification accuracy (in %) of the visual descriptors based on pre-trained convolutional neural networks (deep learning). The best accuracy is indicated in boldface.

Descriptor	Classifier		
	NN_{L1}	NN_{L2}	SVM_{ovo}
GoogLeNet	52.0	52.5	61.3
ResNet-50	59.7	60.7	**67.2**
ResNet-101	59.1	60.1	66.9
VGG-VD-19	51.2	51.9	60.9
VGG-VD-16	51.9	51.9	61.8
Accuracy of random classifier (blind) = 8.3			
Accuracy of random classifier (with priors) = 9.9			

6. Conclusions

In this work we have carried out an experimental evaluation of an array of visual descriptors for painting style categorisation. Compared with other machine vision tasks such as object, face, scene or material recognition this is a more challenging problem, for the concept of similarity is more difficult to define in this context. Besides, whereas humans can rely on previous background on art and art history for the job, the machine cannot [5]. In light of this, a success rate over 67% seems promising. Among the visual descriptors considered in this study, features from 'off-the-shelf' pre-trained CNNs emerged as the best strategy and largely outperformed hand-designed visual descriptors.

Acknowledgements

This work was partially supported by the Department of Engineering at the Università degli Studi di Perugia, Italy, under project *Machine learning algorithms for the control of autonomous mobile systems and the automatic classification of industrial products and biomedical images* (Fundamental research grants 2017), and by the Italian Ministry of Education, University and

Research (MIUR) within the framework *Finanziamento individuale annuale delle attività base di ricerca* (FFABR 2017).

References

[1] Karp C 2014 *Museum International* **66** 157–162
[2] Stork D, Coddington J and Bentkowska-Kafel A 2010 *Computer Vision and Image Analysis of Art* (*Proceedings of SPIE – The International Society for Optical Engineering* no 7531) (San José, USA) art. no. 753101
[3] González E, Bianconi F, Álvarez M and Saetta S 2013 *Advances in Optical Technologies* Art. no. 503541
[4] Nanni L, Ghidoni S and Brahnam S 2017 *Pattern Recognition* **71** 158–172
[5] Tan W, Chan C, Aguirre H and Tanaka K 2016 *Proceedings of the International Conference on Image Processing (ICIP)* (Phoenix, USA) pp 3703–3707
[6] Zujovic J, Gandy L, Friedman S, Pardo B and Pappas T 2009 *Proceedings of the IEEE International Workshop on Multimedia Signal Processing (MMSP)* (Rio De Janeiro; Brazil) art. no. 5293271
[7] Khan F, Beigpour S, Van De Weijer J and Felsberg M 2014 *Machine Vision and Applications* **25** 1385–1397
[8] Florea C, Condorovici R, Vertan C, Butnaru R, Florea L and Vrânceanu R 2016 *Proceedings of the 24th European Signal Processing Conference (EUSIPCO)* (Budapest, Hungary) pp 918–922 art. no. 7760382
[9] Agarwal S, Karnick H, Pant N and Patel U 2015 *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)* (Waikoloa Beach, USA) pp 588–594 art. no. 7045938
[10] Razavian A, Azizpour H, Sullivan J and Carlsson S 2014 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPRW 2014* (Columbus, United States) pp 512–519
[11] Saleh B and Elgammal A 2016 *International Journal for Digital Art History* **2**
[12] Bianco S, Mazzini D and Schettini R 2017 *Proceedings of the 19th International Conference on Image Analysis and Processing (ICIAP)* (*Lecture Notes in Computer Science* no 10484) (Catania, Italy) pp 414–423
[13] Image Processing and Analysis Laboratory 2016 Perceptual ANalysis and DescriptiOn of Romanian visual Art (PANDORA) available online at `http://imag.pub.ro/pandora/pandora_download.html`. Last accessed on Nov. 6, 2017
[14] Adelson E H, Burt P J, Anderson C H, Ogden J M and Bergen J R 1984 *RCA engineer* **29** 33–41
[15] González E, Bianconi F and Fernández A 2016 *Information Sciences* **361-362** 1–13
[16] Napoletano P 2017 *Proceedings of the 6th Computational Color Imaging Workshop (CCIW'17)* (*Lecture Notes in Computer Science* vol 10213) ed Bianco S, Schettini R, Tominaga S and Tremeau A (Milan, Italy: Springer) pp 259–271
[17] Bello-Cerezo R, Bianconi F, Cascianelli S, Fravolini M, Di Maria F and Smeraldi F 2017 *Intelligent Interactive Multimedia Systems and Services 2017. KES-IIMSS 2017* (*Smart Innovation, Systems and Technologies* vol 76) ed De Pietro G, Gallo L, Howlett R and Jain L (Vilamoura, Portugal)
[18] Swain M and Ballard D 1991 *International Journal of Computer Vision* **7** 11–32
[19] Pietikainen M, Nieminen S, Marszalec E and Ojala T 1996 *Proceedings of the International Conference on Pattern Recognition (ICPR)* vol 3 (Vienna, Austria) pp 833–838 art. no. 547285
[20] Dalal N end Triggs B 2005 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)* vol 1 (San Diego, USA) pp 886–893 art. no. 1467360
[21] Ojala T, Pietikäinen M and Mäenpää T 2002 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24** 971–987
[22] Bianconi F, Bello-Cerezo R and Napoletano P 2018 *Journal of Electronic Imaging* **27** art. No. 011002
[23] Lee S, Choi J, Ro Y and Plataniotis K 2012 *IEEE Transactions on Image Processing* **21** 2347–2353
[24] Mäenpää T and Pietikäinen M 2005 *Handbook of Pattern Recognition and Computer Vision (3rd Edition)* ed Chen C H and Wang P S P (World Scientific Publishing) pp 197–216
[25] Cimpoi M, Maji S, Kokkinos I and Vedaldi A 2016 *International Journal of Computer Vision* **118** 65–94
[26] Vedaldi A and Lenc K 2015 *Proceedings of the 23rd ACM International Conference on Multimedia (MM 2015)* (Brisbane, Australia) pp 689–692
[27] Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V and Rabinovich A 2015 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Boston, United States) pp 1–9
[28] He K, Zhang X, Ren S and Sun J 2016 *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (Las Vegas, United States) pp 770–778
[29] Simonyan K and Zisserman A 2015 *Proceedings of the 3rd International Conference on Learning Representations (ICLR)* (San Diego, USA)