Combining Holistic and Part-based Deep Representations for Computational Painting Categorization

Rao Muhammad Anwer
Department of Computer Science
Aalto University, Finland
rao.anwer@aalto.fi

Fahad Shahbaz Khan
Computer Vision Laboratory
Linköping University, Sweden
fahad.khan@liu.se

Joost van de Weijer
Computer Vision Center
CS Dept. Universitat Autonoma de Barcelona, Spain
joost@cvc.uab.es

Jorma Laaksonen
Department of Computer Science
Aalto University, Finland
jorma.laaksonen@aalto.fi

ABSTRACT

Automatic analysis of visual art, such as paintings, is a challenging inter-disciplinary research problem. Conventional approaches only rely on global scene characteristics by encoding holistic information for computational painting categorization. We argue that such approaches are sub-optimal and that discriminative common visual structures provide complementary information for painting classification.

We present an approach that encodes both the global scene layout and discriminative latent common structures for computational painting categorization. The region of interests are automatically extracted, without any manual part labeling, by training class-specific deformable part-based models. Both holistic and region-of-interests are then described using multi-scale dense convolutional features. These features are pooled separately using Fisher vector encoding and concatenated afterwards in a single image representation. Experiments are performed on a challenging dataset with 91 different painters and 13 diverse painting styles. Our approach outperforms the standard method, which only employs the global scene characteristics. Furthermore, our method achieves state-of-the-art results outperforming a recent multi-scale deep features based approach [11] by 6.4% and 3.8% respectively on artist and style classification.

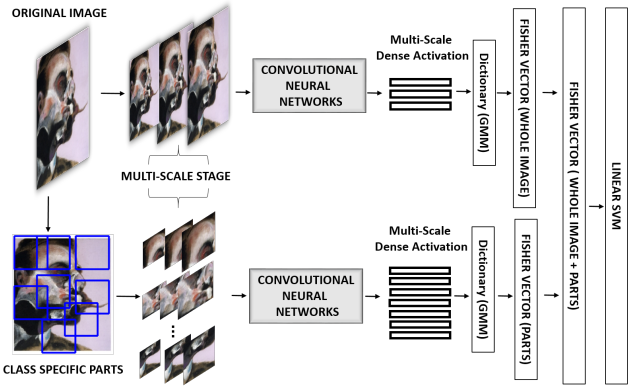

Figure 1: An overview of our approach for painting categorization problem. Both holistic and part-based information are encoded in our method. The discriminative parts are automatically captured, without any manual part labeling, by training a class-specific DPM detector. The holistic and class-specific parts are then described using multi-scale dense convolutional features from a pre-trained CNN. The two sets of features are later pooled separately using the Fisher vector encoding and concatenated afterwards in a single image representation.

1. INTRODUCTION

Digital analysis of art, such as paintings, is a challenging cross-disciplinary research problem. It has gained much attention recently [8, 2] due to the emergence of significant amount of visual artistic data on the web. Computational techniques to manage online large digital art data have several applications, such as, e.g. art recommendation systems in the tourism industry, analysis and labeling tools for experts in museums and detection systems to identify art forgery. In this paper, we investigate the task of automatically categorizing a painting to its artist and style.

The artist classification task involves associating a painting image to its respective painter. The style categorization problem tackles inferring artist paintings with respect to a school of art or art movement, e.g. renaissance, impressionism and modernism. Both these tasks are challenging due to large amount of inter-class and intra-class variations. For instance, Picasso's paintings span multiple styles such as cubism and surrealism. Other factors, such as stroke patterns, choice of color palette, scene composition and line styles, influence the art work making the problem of computational painting categorization further complicated.

ICMR'16, June 06-09, 2016, New York, NY, USA

DOI: http://dx.doi.org/10.1145/2911996.2912063

Most existing approaches [6, 2] tackle the problem of painting categorization using different low-level features such as appearance, texture and color. These low-level features are further used within the bag-of-words framework to obtain histogram-based image representations. Within the bag-of-words framework, several encoding schemes, e.g. hard assignment [6] and Fisher vector [8] have been employed for painting classification. Recently, deep features based representations, extracted from the Convolutional Neural Network (CNNs), have been applied for painting categorization [11, 10]. The CNNs are trained using raw image pixels and consist of a series of convolution and pooling operations followed by one or more fully connected (FC) layers. The deep features, trained using large amount of labeled data (e.g. ImageNet), are known to be generic and applied for many vision applications. Generally, the deep features are extracted using the activations from some FC layers of the network for the classification task.

In the context of object and texture recognition, it has been shown that activations from the last convolutional layers are more discriminative and provide superior performance compared to standard features from the FC layers of the network [3]. The convolutional layers mitigate the requirement of fine-tuning for a specific dataset. The activations from the convolutional layers can be further used as dense local features encoded using Fisher vectors. In this work, we also employ activations from the convolutional layer as dense local features and pool them using Fisher vectors for computational painting categorization.

Beside holistic image representations, intermediate image parts have often serve as region of interests that capture key object or scene regions. Typically, parts are extracted using object detectors such as deformable part-based models [4] or poselets [1] for object recognition. In a standard object detection settings, a deformable part-model (DPM) approach is trained using the positive examples annotated with the bounding box locations of the corresponding object class. The parts in the DPM framework are learned using a latent SVM (LSVM) formulation. Each object is then modeled as a deformable collection of parts with a root model at its center. The DPMs are also employed to capture latent common structures within a scene for scene recognition [9]. In such a case, no bounding box information is available for training and the root filter corresponds to the whole image with moveable parts representing the important scene regions. In the context of action recognition, class-specific action detectors [7] have shown to provide promising results. Here, we propose to employ DPMs to automatically capture latent common structures as region of interests within paintings for computational painting categorization tasks.

1.1 Contribution

We propose an approach that encodes both the global layout and the region of interests (ROI's) in digital paintings. Multi-scale dense convolutional features are computed over the entire image to capture the global scene characteristics. These multi-scale local convolutional features are pooled using the Fisher vector encoding [13] to obtain a single image representation. The ROI's are obtained automatically without any manual part labeling by training a class-specific DPM detector. The class-specific parts are then described similarly by using multi-scale dense convolutional features later pooled using the Fisher vector encoding. Both part-based and image-level deep Fisher vectors are then contacted into a single representation. Figure 1 shows an overview of our approach. Experiments are performed on the challenging painting-91 dataset [6] with two tasks: artist and style classification. On both artist and style classification tasks our approach improves the mean classification accuracy by 2.7% and 4.7% respectively, compared to using only the global scene characteristics. Further, our approach achieves state-of-the-art results on both tasks, outperforming a recent CNN-based approach [11].

1.2 Relation to Prior Works

Existing approaches [6, 2, 11] either employ low-level features or deep features for painting categorization. The ones [11, 10] based on deep features employ activations from the FC layers for the painting classification. Other than painting classification, it has been shown recently that activations from deeper convolutional layers provide superior results for object and texture recognition [3]. Typically, dense multi-scale convolutional features are extracted over an entire image. These features are then pooled using Fisher vector encoding to obtain an image representation.

In the context of fine-grained and scene recognition [12, 9, 15], part-based information is often exploited to capture the semantic content and composition of the integral regions in an image. Quattoni and Torralba [12] propose to incorporate part based information by using manually labeled data for scene recognition. The work of [9] uses a DPM framework to automatically capture part information for scene recognition. To the best of our knowledge, we are the first to incorporate part-based information in a multi-scale convolutional features based framework for painting categorization problem. Our approach goes beyond the conventional deep features based painting classification methods [11, 10] by capturing both the holistic and part-based information.

2. OUR APPROACH

Inspired by the recent success of CNNs, we base our approach on deep features and employ it for both components: holistic and part-based representations. We use the VGG-16 network [14][1] pre-trained on the ImageNet. In this network, the input images employ small 3×3 convolution filters per pixel. The stride is set to 1 pixel. The network comprises of several max-pooling layers that perform spatial pooling at a stride of 2 pixels over 2×2 pixel windows. At the end, the network contains 3 fully connected (FC) layers. The width of the network starts with 64 feature maps and goes to 512 feature maps at its widest. We refer to [14] for more details. Next, we describe our deep features based holistic and part-based representations.

2.1 Holistic Representation

As discussed earlier, most deep features based methods employ activations from the FC layer for computational painting classification [11, 10]. Instead, we employ activations from the last convolutional layer of the network since it was recently shown to provide superior results for object and texture recognition tasks [3]. The convolutional layer returns dense activations of 512 dimensions, computed over the entire image. We employ multi-scale strategy by rescal-

[1]The deep network models available at: http://www.robots.ox.ac.uk/~vgg/research/very_deep/

ing each image over a range of scales and pass them through the network to obtain dense convolutional activations. The use of convolutional layer also mitigates the need of resizing the original image to a fixed size. The number of local convolutional patches depend on the size of the input image.

As in [3], a visual vocabulary is constructed over the multi-scale dense local patches by using a Gaussian mixture model (GMM). The multi-scale features are then pooled in a single image representation via Fisher encoding scheme [13]. Our holistic representation can be seen as analogous to the recently introduced Fisher vector CNN (FVCNN) approach [3].

2.2 Part Representation

The DPM framework [4] is previously employed to automatically extract parts for fine-grained [15] and scene classification [9] tasks. Here, we employ the DPM framework [4] to obtain region of interests in digital paintings. The DPM framework consists of a root filter and a deformable collection of moveable parts. The deformation parameters are employed to penalize the movements of parts from their default locations relative to the root. In DPMs, both the root and parts are represented by a dense grid of non-overlapping cells and Latent SVM (LSVM) formulation is used for learning. A 31-dimensional HOG histogram is constructed for each cell. The detection score for each window is computed by concatenating the root filter, the part filters and the configuration deformation cost of all parts.

The standard DPM framework assumes bounding box information to be available at training time and the part locations are treated as latent information. Since no bounding box information is available in our case, the root filter surrounds the whole image. As in [9], we restrict the root filter to have at least 40% overlap with the image. A class-specific DPM is trained using instances from all other classes as negative samples. During the training, the root filter weights are initialized to cover the entire training images, whereas part filters are initialized same as in [4]. Each class-specific DPM is trained using a single mixture component and eight parts. Figure 2 shows a visualization of three style models together with five sample images for each respective style class. The root filter is shown in red, whereas the parts are visualized in blue.

To obtain part locations, the trained class-specific models are then applied on all images. The discriminative part regions are cropped from an image and rescaled over a range of scales before passing through the deep network. As described in Section 2.1, dense activations from the last convolutional layer of the network are extracted and a single GMM-based vocabulary is constructed for all parts. The multi-scale part based local features are then pooled together in a single representation via the Fisher encoding scheme. Finally, the holistic and part-based image representations are concatenated before input to a classifier.

3. EXPERIMENTS

We evaluate our approach on the challenging Painting-91 dataset [6] for both artist and style classification tasks. The dataset consists of 4266 paintings of 91 painters, such as Picasso, Ruben and Warhol. For style classification, the dataset contains 13 artistic styles such as Baroque, Cubbism and Realism. The train/test split is provided by the respective authors [6]. For multi-scale settings, we use 3 scales: $0.5, 1.0, 1.5$. For both holistic and part-based vocabularies,

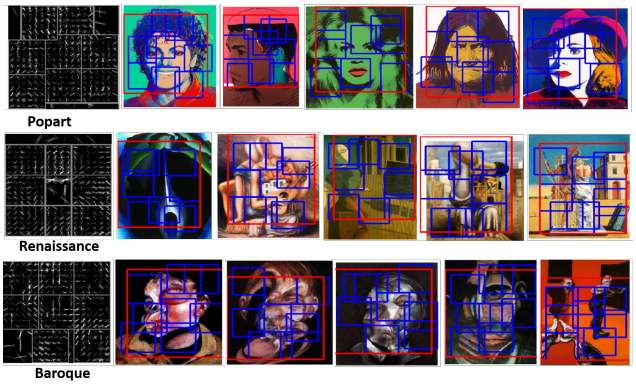

Figure 2: Style models for classes: popart, renaissance and baroque. Example images with root (red) and part locations (blue) are shown for each respective class. The visualizations show that the parts are able to capture latent common structures in a class.

Method	VGG-16 FC [14]	MOP [5]	Holistic [3]	Ours
Artist	51.7	59.7	61.8	**64.5**
Style	67.2	68.8	70.1	**74.8**

Table 1: Comparison (in average classification rates) of standard deep features (FC), the MOP approach, the baseline holistic alone, and our proposed approach combining the holistic and part-based deep representations. Our approach yields consistent improvements for both tasks.

we use a GMM with 16 components. For classification, we use one-vs-all linear SVMs using LibLinear package.For evaluation, we employ the same protocol as provided with the dataset [6]. Each test image is assigned the label of the classifier giving the highest confidence. The final performance is evaluated as the mean recognition rate.

Baseline experiment: We first compare our approach with different baseline methods on the Painting-91 dataset. Note that all approaches employ the same VGG-16 network. The VGG-16 FC method corresponds to using standard deep features with 4096 dimensions from FC layer of the network [14]. The multi-scale orderless pooling (MOP) approach [5] employs FC activations at three levels: 4096-D activations over the entire image, 128×128 of 4096-D pooled using VLAD encoding with 100 words and similar encoding with 64×64 image patches. The three levels are concatenated into a single image representation. Table 1 shows the comparisons on both artist and style classification tasks. On the artist classification task, the standard FC activation based method [14] achieves an average classification score of 51.7%. The MOP approach [5] obtains a mean classification accuracy of 59.7%. A mean classification score of 61.8% is obtained by using the holistic representation alone. Our approach that combines the holistic and part-based representations significantly improves the performance with a classification score of 64.5%. Similarly, on the style classification task, our combined approach improves the performance by 4.7% compared to the holistic alone method [3].

Figure 3 shows comparison on classes where our approach achieves maximum gain in performance, compared to the standard holistic alone approach. A significant gain in performance is achieved especially for Jacques David (+20%), Jean Ingres (+16%), and Arshille Gorky (+12.5%), compared to holistic only representation.

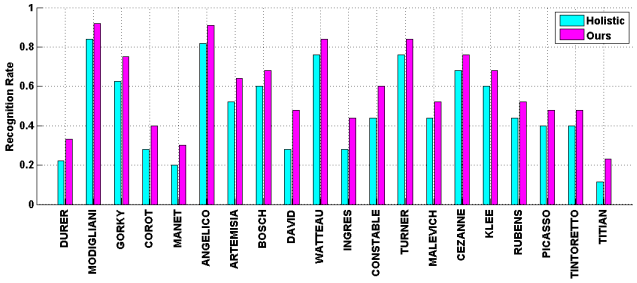

Figure 3: Average classification rates for 20 classes where maximum improvement is obtained with our approach, compared to the standard holistic method [3].

Num. of parts	0	2	4	6	8	10	12
Style	70.1	70.9	72.2	73.1	**74.8**	74.6	74.5

Table 2: Results when varying the number of parts for style classification. The first entry (0) corresponds to using holistic alone with no parts. The best results are obtained with 8 parts in combination with the holistic approach.

Number of part filters: As discussed in Section 2.2, we employ 8 parts for each DPM model. To further validate the impact of number of parts, we perform an experiment by varying the number of parts. Table 2 shows the results for the style classification task. The performance improves marginally with only 2 parts. The best results are obtained when using 8 parts and the performance starts to saturate when this number exceeds.

State-of-the-art comparison: Table 3 shows a comparison of our approach with state-of-the-art methods for both artist and style classification. For artist classification, the cross-layer CNN approach (CL-CNN) [10] based features from multiple CNN layers, achieves a mean accuracy of 56.4%. The multi-scale CNN approach (MS-CNN) [11] employs a multi-scale strategy for CNN features and obtains a mean accuracy of 58.1%. Our approach achieves a considerable gain of 6.4% in mean accuracy, compared to MS-CNN approach [11]. Similarly, our approach outperforms the state-of-the-art results for style classification, achieving a mean classification rate of 74.8%.

4. CONCLUSIONS

We have proposed an approach based on exploiting both holistic and part-based information for painting categorization. Our method automatically extracts discriminative part by training class-specific DPMs. Both holistic and part-based regions are described using multi-scale convolutional features. Experiments show that our approach outperforms the baseline methods, leading to the state-of-the-art results.

5. ACKNOWLEDGMENTS

This work has been funded by the grant 251170 of the Academy of Finland, VR (EMC^2), the projects TIN2013-41751 and of the Spanish Ministry of Science and the Catalan project 2014 SGR 221. The calculations were performed using computer resources within the Aalto University School of Science "Science-IT" project. We also acknowledge the support from Nvidia and NSC.

Method	MF [6]	CL-CNN [10]	MS-MCNN [11]	Ours
Artist	53.1	56.4	58.1	**64.5**
Style	62.2	69.2	71.0	**74.8**

Table 3: Comparison of our approach with state-of-the-art methods. Our approach achieves the best results compared to existing methods based on deep features.

6. REFERENCES

[1] L. Bourdev, S. Maji, and J. Malik. Describing people: A poselet-based approach to attribute classification. In *ICCV*, 2011.

[2] G. Carneiro, N. Silva, A. Bue, and J. Costeira. Artistic image classification: An analysis on the printart database. In *ECCV*, 2012.

[3] M. Cimpoi, S. Maji, and A. Vedaldi. Deep filter banks for texture recognition and segmentation. In *CVPR*, 2015.

[4] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32(9):1627–1645, 2010.

[5] Y. Gong, L. Wang, R. Guo, and S. Lazebnik. Multi-scale orderless pooling of deep convolutional activation features. In *ECCV*, 2014.

[6] F. S. Khan, S. Beigpour, J. van de Weijer, and M. Felsberg. Painting-91: a large scale database for computational painting categorization. *MVA*, 25(6):1385–1397, 2014.

[7] F. S. Khan, J. Xu, J. van de Weijer, A. Bagdanov, R. M. Anwer, and A. Lopez. Recognizing actions through action-specific person detection. *TIP*, 24(11):4422–4432, 2015.

[8] T. Mensink and J. Gemert. The rijksmuseum challenge: Museum-centered visual recognition. In *ICMR*, 2014.

[9] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[10] K.-C. Peng and T. Chen. Cross-layer features in convolutional neural networks for generic classification tasks. In *ICIP*, 2015.

[11] K.-C. Peng and T. Chen. A framework of extracting multi-scale features using multiple convolutional neural networks. In *ICME*, 2015.

[12] A. Quattoni and A. Torralba. Recognizing indoor scenes. In *CVPR*, 2009.

[13] J. Sanchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 105(3):222–245, 2013.

[14] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015.

[15] N. Zhang, R. Farrell, F. Iandola, and T. Darrell. Deformable part descriptors for fine-grained recognition and attribute prediction. In *ICCV*, 2013.