The Rijksmuseum Challenge: Museum-Centered Visual Recognition

Thomas Mensink and Jan van Gemert
ISLA Lab – Informatics Institute
University of Amsterdam
{thomas.mensink,j.c.vangemert}@uva.nl

ABSTRACT

This paper offers a challenge for visual classification and content-based retrieval of artistic content. The challenge is posed from a museum-centric point of view offering a wide range of object types including paintings, photographs, ceramics, furniture, *etc*. The freely available dataset consists of 112,039 photographic reproductions of the artworks exhibited in the Rijksmuseum in Amsterdam, the Netherlands. We offer four automatic visual recognition challenges consisting of predicting the artist, type, material and creation year. We include a set of baseline results, and make available state-of-the-art image features encoded with the Fisher vector. Progress on this challenge improves the tools of a museum curator while improving content-based exploration by online visitors of the museum collection.

Keywords

Cultural Heritage, Art dataset, image classification

1. INTRODUCTION

With art museums digitizing their collection for cultural heritage or for (online) visitors there is a need for automatic tools to organizing large quantities of visual artistic data. Such tools can help the museum conservator with labeling art objects by automatically suggesting the artist, type, year, or used material. Another application is in querying the art collection which allows visitors to link, browse and explore the set online, or with a smart phone pointed to a physical art object on display.

In this paper we introduce a large, diverse and open dataset of art objects to support and evaluate automatic art classification and retrieval techniques. The collection is a set of over 110,000 objects consisting of digital images and their metadata descriptions from the Rijksmuseum collection made public online[1]. The works of art date from ancient times, medieval ages and the late 19th century. They provide an excellent overview of the richness, diversity and beauty of Dutch and international cultural heritage, see figures 1 and 2 for examples. The set includes paintings and prints ranging from the great masters to anonymous biblical paintings, 19th-century photographs, ceramics, furniture, silverware, doll's houses, miniatures, *etc*. The diversity and size of the dataset shifts the scope of previous single-type approaches such as paintings [2] or vases [3] to a broad and realistic *museum-centered* view to art collection management.

Figure 1: Visualization of the dataset with t-SNE

To allow a structured study of automatic image analysis we introduce four open challenges, namely: (i) predict the **artist**, (ii) predict the **art-type**, (iii) predict the used **material** and (iv): predict the **creation year**. To facilitate researchers on image features we include a standard classification baseline [4] and to aid machine learning researchers we include a baseline of state-of-the-art image features based on the Fisher vector [12]. We will make the dataset, including API details, experimental settings and visual features available for download[2].

2. RELATED WORK

Multimedia tools have aided conservation, analysis and study of cultural, historical and artistic content [2]. For example, the digital Michelangelo project [8] created high quality 3D models of Michelangelo's sculptures and architecture. Paintings can automatically be classified as aesthetic or not [9] and such visual composition can aid category labeling by style pooling [14]. Furthermore, wavelet analysis of brush strokes in paintings can reveal artist identity [7, 10], and help in painter authentication [11]. Our work aids art

[1] www.rijksmuseum.nl/en/api

[2] http://www.science.uva.nl/~tmensink/rijks

ICMR'14, April 01 - 04 2014, Glasgow, United Kingdom Copyright 2014 ACM 978-1-4503-2782-4/14/04...$15.00.
http://dx.doi.org/10.1145/2578726.2578791




Title	The Milkmaid	Seated Cupid	Reclining Lion	Dish with a Landscape
Creator	Johannes Vermeer	Étienne-Maurice Falconet	Rembrandt van Rijn	Frederik van Frytom
Year	1660	1757	1658–1662	1670 – 1685
Type	Painting	Sculpture	Drawing	Saucer
Material	Canvas, Oil paint	White Marble, Copper	Paper, Paint	Faience

Figure 2: Example images from the Rijksmuseum dataset together with their provided descriptions

conservation by investigating automatic prediction of artists, material, techniques and the creation year.

Datasets for evaluation automatic analysis of artistic content are typically dedicated to a single art type, not openly available due to copyright constraints and are often quite small. Examples include paintings where small datasets of 101 [7] and 380 [6] images are not available for download. Recent public datasets increased in size, yet focus on specific artistic subsets such as 4k hieroglyphs [5], 38k vases [3] and 50k sculptures [1]. Here, we introduce a diverse, open, and large artistic dataset with over 110k images from a museum perspective. This provides acces to a real and diverse artistic data as used by conservators and visitors alike.

3. CHALLENGES

The dataset has 112,039 high-quality artwork images recorded in a controlled setting. Images are stored at 300 dpi and the compressed jpeg image file size ranges between 2-5 MB. Unfortunately no UV or IR data is available. Each image has a corresponding xml file containing metadata.

3.1 Challenge 1: Artist Classification

The dataset contains 6,629 artists in total, with high variation in the number of pieces per artist. For example, Rembrandt has 1,384 pieces, and Vermeer has only 4. There are 350 artists with more than 50 pieces, 180 artists have around 100, and 90 artists have 200 pieces.

The challenge is to predict the artist given an image. This is a multi-class problem where each object has a single creator. We measure accuracy with the frequency weighted classification accuracy, *i.e.*, the mean of the confusion matrix diagonal.

3.2 Challenge 2: Categorization into Types

There are 1,824 different art-types in the dataset, where most pieces have 2 types and the most common type with 77,051 pieces is *print making* while the number of *paintings* is 3,593. In figure 3 we show the frequency and the number of types per art piece.

The challenge is a multi-label classification problem where each piece may have one or more types. We measure accuracy with Mean Average Precision (MAP).

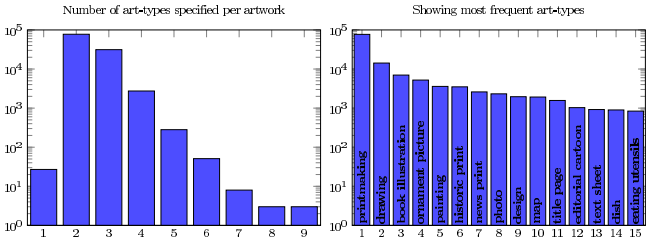

Figure 3: Distribution of art-types

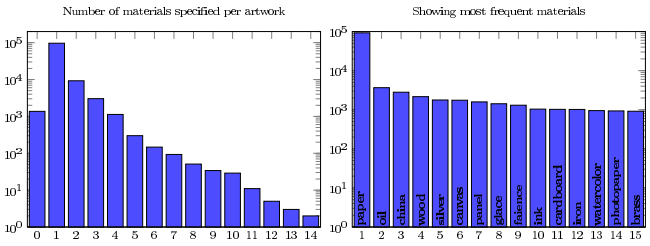

Figure 4: Distribution of used materials

3.3 Challenge 3: Labeling of Materials

In total there are 406 different materials where nearly 100,000 pieces have a single material. The material *paper* is the most common, with over 92,000 pieces. In figure 4 we show the distribution of materials over the dataset.

The challenge is a multi-label classification problem where each piece may have one or more materials. We measure accuracy with Mean Average Precision (MAP).

3.4 Challenge 4: Estimating Creation Year

The artworks in the dataset date from biblical times, through the medieval period and until the late 19th century. The majority of the artworks is dated between 1500 and 1900. Some of the works, mostly the more recent ones, have an exact year of creation, yet many of them have an interval of years in which the artwork is estimated to be created. In figure 5 we show the distribution of artworks over time.

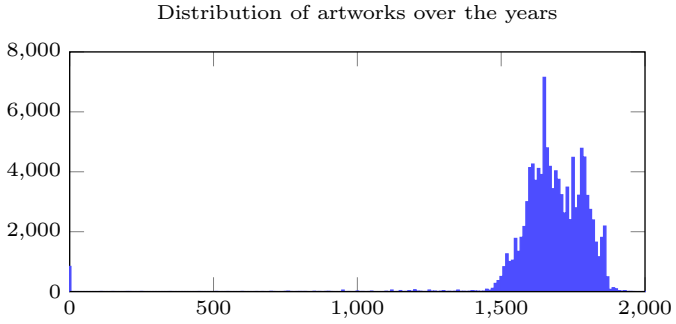

Figure 5: Distribution of years

4. BASELINE EXPERIMENTS

4.1 Image features

Images are encoded using state-of-the-art Fisher vectors (FV) [12], which aggregates local SIFT descriptors in a global feature vector. Two different SIFT features were extracted, namely intensity and opponent color SIFT [13]. Using PCA, both SIFT types are reduced to 96 dimensions. We use a codebook with $k = 16$ elements, and extract FV w.r.t. to their mean and variance. This yields a 3,072 dimensional feature vector for each SIFT type per image.

4.2 Experimental Setup

For all experiments we use the same experimental setup. The dataset is divided randomly into three separate sets, the train set (78,427 images, 70%), is used to train classification and regression models, the validation set (11,204 images, 10%), is used to tune the hyper parameters of the models, and the test set (22,408 images, 20%), is used to measure the performance. For all experiments we use liblinear SVM [4].

4.3 Challenge 1: Artist Classification

We learn 1-vs-Rest linear SVM classifiers for all creators that have an artwork in all sets of the data and at least 10 artworks in the testset. This results into 374 different creators, all other images are grouped together into an unknown class u. At evaluation time, each artwork is assigned to a single artist. Performance is measured as the weighted mean class accuracy (MCA). This measure ensures that the classification performance of an artist with only a few works accounts as much as an artists with thousands of pieces.

In Table 1 we show our main results. First, we observe that the performance difference between the two SIFT features is limited to a maximum of 2% MCA. Second, when removing the large u class, which results in evaluating on about 60% of the test set, the performance improves significantly. This is understandable, since this u class contains a diverse set of artworks from many different creators.

In the table we show also results for using a subset of the artists, where we have ranked all artists according to the number of artworks in the test set, and select the most productive ones. We observe that the MCA increases when we focus on more productive artists.

In Figure 6 we show the confusion matrix when using the opponent SIFT features evaluated for the top 100 creators. We observe a strong diagonal, which indicates a high MCA (76.3%). By removing the diagonal, we observe that the confusion is centred around the most-productive artists.

#	%	intensity	opponent
374+u	100.0	51.0	52.8
374	59.1	65.5	66.5
300	55.5	67.6	68.7
200	48.7	71.2	72.1
100	36.8	75.7	76.3

Table 1: Artist classification results, the mean class accuracy (MCA) for different number of artist (#) using a subset of the test set (%). Artists are ordered based on productivity.

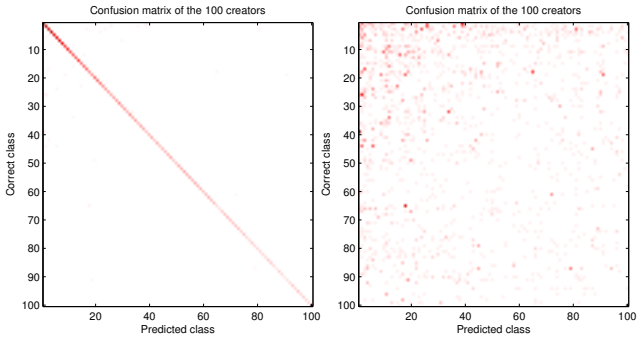

Figure 6: Confusion matrix for the top 100 artists, (*left*) the strong diagonal indicates a high mean class accuracy, (*right*) removing the diagonal shows the intra-artist confusion more clearly.

4.4 Challenge 2: Categorisation into Types

For this task we want to predict the relevant types for an artwork. While the annotations contain information about types and subtypes, we consider this task as an image annotation task where all types are treated equally. Therefore, we train 1-vs-Rest linear SVM classifiers for all types which have an artwork in all sets of the data and at least 10 artworks in the testset. At evaluation time we measure the performance in terms of Mean Average Precision (MAP) and image Mean Average Precision (iMAP). Where MAP is a retrieval measure which evaluates per concept the ranking of images, iMAP is an annotation measure which evaluates per image the ranking of the concepts.

The results are shown in Table 2. First, we observe, again, that the opponent SIFT features yield a higher performance than the intensity based features. Second, reducing the number of types, by focusing on more frequent used types, does improve just slightly the MAP performance. Furthermore, in decreases the iMAP performance, which indicates that even among the very frequent types a few mistakes are made in the ranking, in a smaller set of used types these account heavier in the performance measure. Reducing the number of types only marginally influence the percentage of the dataset used, indicating that over 92% of the dataset has a annotation from the 25 most frequent types.

4.5 Challenge 3: Labeling of Materials

For this task we want to predict the relevant types for an materials. Similar to the type prediction task, we consider this as an image annotation task and measure the performance with MAP and iMAP. Again, we train 1-vs-Rest linear SVM classifiers for all materials which have an artwork in all sets of the data and at least 10 artworks in the testset.

In Table 3, we show our main results, and the conclusions are similar to the type-prediction challenge.

#	%	intensity MAP	intensity iMAP	opponent MAP	opponent iMAP
103	100.0	67.1	90.4	67.3	91.4
75	96.3	67.5	90.1	68.8	91.1
50	94.9	69.2	89.4	70.6	90.4
25	92.2	70.8	87.6	71.8	88.6

Table 2: Type prediction results in MAP and iMAP, evaluated using different number of types (#), representing a subset of the test set (%).

#	%	intensity MAP	intensity iMAP	opponent MAP	opponent iMAP
81	100.0	48.6	94.1	53.3	94.7
75	98.5	50.3	94.2	54.7	94.7
50	98.0	61.8	94.1	56.8	94.6
25	96.1	58.8	93.5	63.9	93.9

Table 3: Material labeling results in MAP and iMAP, evaluated using different number of materials (#), representing a subset of the test set (%).

4.6 Challenge 4: Estimating Creation Year

For this task the goal is to estimate the year of creation. We consider this as a regression task and learn a max-margin regressor using liblinear [4]. From the dataset we use all images for which the provided date interval is < 100 years. As a baseline we compare to the *mean year predictor*, which just assigns the mean creation year of all artworks (1676).

In Table 4 we report the performance measured in (i) the rooted mean squared loss, (ii) the mean absolute loss, and (iii) the interval accuracy. The latter evaluates the percentage of artworks for which the estimated year was ± 50 years from the ground truth year. Finally, we also report the Spearman's rank correlation coefficient.

To gain more insight in the performance, we performed a second set of experiments using the opponent sift FV. For the 15 most frequent types, we evaluate the performance of (i) the learned regressor on all data (SR), (ii) the mean year of this type (YM), and (iii) a regressor learned specifically for this type (TR). The results in Table 5 show that learning a regressor per type is beneficial. However, it requires to know the type of an artwork. Therefore we also evaluate the predicted type regressor (PTR), where we first estimate the type using SVM and then use the type regressor for the estimated type. We observe that for most types, and for the average over all images, this strategy yields better performance than using a single regressor.

5. CONCLUSIONS

In this paper we introduced the Rijksmuseum dataset and proposed four visual recognition challenges. The dataset contains over 110k different artworks with rich annotations, and is available for download.

6. ACKNOWLEDGMENTS

This research is supported by the STW STORY project and the Dutch national program COMMIT.

7. REFERENCES

[1] R. Arandjelović and A. Zisserman. Name that sculpture. In *ICMR*, 2012.

	squared	absolute	interval	Spearman
mean	173.8	89.2	36.8	-
intensity	163.0	74.4	49.0	0.544
opponent	161.3	72.4	51.1	0.569

Table 4: Estimating the year of creation, evaluated using the squared loss, the absolute loss, the interval accuracy, and Spearman's rank correlation.

type	SR	YM	TR	PTR
printmaking	102.2	117.9	99.3	104.1
book illustration	69.9	86.1	62.6	69.8
drawing	108.4	111.8	98.3	81.8
ornament picture	135.5	132.6	121.0	136.8
historic print	63.1	78.6	45.7	88.8
news print	59.7	44.1	28.5	48.6
design	108.0	116.0	95.5	74.7
map	54.8	42.6	33.9	30.4
painting	100.8	93.4	78.2	79.5
title page	52.9	51.7	34.0	49.8
photo	283.2	267.9	219.5	233.7
editorial cartoon	58.1	55.9	35.9	44.7
title print	57.3	64.4	49.0	51.3
dish	210.8	219.1	181.8	189.2
saucer	151.4	146.4	134.3	51.8
mean	112.7	122.5	102.4	105.8

Table 5: Year estimation per type, evaluated using squared loss, see text for details

[2] M. Barni, A. Pelagotti, and A. Piva. Image processing for the analysis and conservation of paintings: opportunities and challenges. *IEEE Sig. Proc. Mag.*, 22:141–144, 2005.

[3] E. J. Crowley and A. Zisserman. Of gods and goats: Weakly supervised learning of figurative art. In *British Machine Vision Conference*, 2013.

[4] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. Liblinear: A library for large linear classification. *JMLR*, 9:1871 – 1874, 2008.

[5] M. Franken and J. C. van Gemert. Automatic egyptian hieroglyph recognition by retrieving images as texts. In *ACM Multimedia*, 2013.

[6] K. Ivanova and P. Stanchev. Color harmonies and contrasts search in art image collections. In *MMEDIA*, 2009.

[7] C. Johnson, E. Hendriks, I. Berezhnoy, E. Brevdo, S. Hughes, I. Daubechies, J. Li, E. Postma, and J. Wang. Image processing for artist identification: Computerized analysis of vincent van gogh's painting brushstrokes. *IEEE Signal Processing Magazine*, 25:37 – 48, 2008.

[8] M. Levoy et al. The digital Michelangelo project: 3D scanning of large statues. In *Computer graphics and interactive techniques*, 2000.

[9] C. Li and T. Chen. Aesthetic visual quality assessment of paintings. *Selected Topics in Signal Processing*, 3(2):236–252, 2009.

[10] J. Li and J. Z. Wang. Studying digital imagery of ancient paintings by mixtures of stochastic models. *Image Processing, IEEE Transactions on*, 13(3):340–353, 2004.

[11] S. Lyu, D. Rockmore, and H. Farid. A digital technique for art authentication. *Proceedings of the National Academy of Sciences of the United States of America*, 101(49):17006–17010, 2004.

[12] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek. Image classification with the fisher vector: Theory and practice. *IJCV*, 2013.

[13] K. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Trans. PAMI*, 32(9):1582–1596, 2010.

[14] J. van Gemert. Exploiting photographic style for category-level image classification by generalizing the spatial pyramid. In *ICMR*, 2011.