The State of the Art: Object Retrieval in Paintings using Discriminative Regions

Elliot J. Crowley
elliot@robots.ox.ac.uk
Andrew Zisserman
az@robots.ox.ac.uk

Visual Geometry Group
Department of Engineering Science
University of Oxford

Abstract

The objective of this work is to recognize object categories (such as animals and vehicles) in paintings, whilst learning these categories from natural images. This is a challenging problem given the substantial differences between paintings and natural images, and variations in depiction of objects in paintings.

We first demonstrate that classifiers trained on natural images of an object category have quite some success in retrieving paintings containing that category. We then draw upon recent work in mid-level discriminative patches to develop a novel method for reranking paintings based on their spatial consistency with natural images of an object category. This method combines both class based and instance based retrieval in a single framework.

We quantitatively evaluate the method over a number of classes from the PASCAL VOC dataset, and demonstrate significant improvements in rankings of the retrieved paintings over a variety of object categories.

1 Introduction

The question we investigate in this paper is: can paintings containing an object category (e.g. a train or a bird) be retrieved starting from a model learnt from natural images? At first sight, we might not be optimistic since natural images (i.e. everyday photos taken with a camera) and paintings can have very different low level statistics, and paintings vary considerably in depiction style from photo-realistic renderings through particular movements (e.g. impressionism, pointillism – which are almost designed to disrupt the fine scale measurement of local gradients in a HOG or SIFT feature) to more abstract depictions (Fauvism, Cubism).

Apart from the challenge in its own right, this goal of automatically obtaining paintings with a particular object is of much interest to Art Historians who currently find paintings manually or from memory [8, 22, 34]. They can then study the change in the depiction style over time [25], or determine when an object first appeared in paintings.

The problem is essentially one of domain adaptation (also referred to as domain transfer) [12, 24, 31] from natural images to paintings. The problem of generalizing across depiction styles has been studied recently by Wu and Hall [35]. They took the interesting approach of building a multi-layer depiction invariant graph model that was shown to be capable of generalizing to drawings and cartoons in particular. However, a limitation of the method was

© 2014. The copyright of this document resides with its authors.
It may be distributed unchanged freely in print or electronic forms.

that it was restricted in both training and testing to uncluttered Caltech101 [15] style images – where the object of interest fills the image against a uniform background. In our work, both training and testing images are PASCAL VOC style [13] where the object may only occupy a small part of the image, and can be partially occluded, see figure 1. Others [6, 32] have recently considered the problem of using a (single) natural image to retrieve paintings, in particular for the case of a specific building, rather than a class of objects: Shrivastava *et al.* [32] use an Exemplar SVM [26] for retrieval, and Aubry *et al.* [6] improve on this method by employing mid-level discriminative patches (MLDPs) [7, 23, 27, 33] to allow for more variation. We build on and extend the method of [6] from *specific* buildings to *classes* of objects, and overcome its two principal limitations: that the training images must have a very similar pose to the target object, and that the training images have the object (a building in their case rendered from a 3D model) segmented. Others have investigated classification and retrieval in paintings, e.g. the interesting analysis of [9], but have not considered this domain adaptation aspect.

We make the following contributions: (i) we show, somewhat surprizingly, that image classifiers and object detectors learnt from PASCAL VOC images can retrieve paintings containing an object class (section 2); (ii) we introduce a method of re-ranking that is based on spatial consistency of MLDP correspondences (section 3), and show that the precision of low ranked paintings (i.e. the ones that would appear on the first webpage in an image search) can be significantly improved based on how spatially consistent the paintings are with the natural images used to train the classifiers (section 4); and (iii) we compare other methods of training and re-ranking including training from Google images (where the images are more Caltech101 style) and using a DPM [16] detector, and also investigate hybrid re-ranking strategies.

Note, although using spatial consistency to re-rank is standard practice in large scale object instance retrieval [21, 30], using it in this manner for object categories is novel, and contrasts to the spatial consistency implicit in Spatial Pyramids and DPMs.

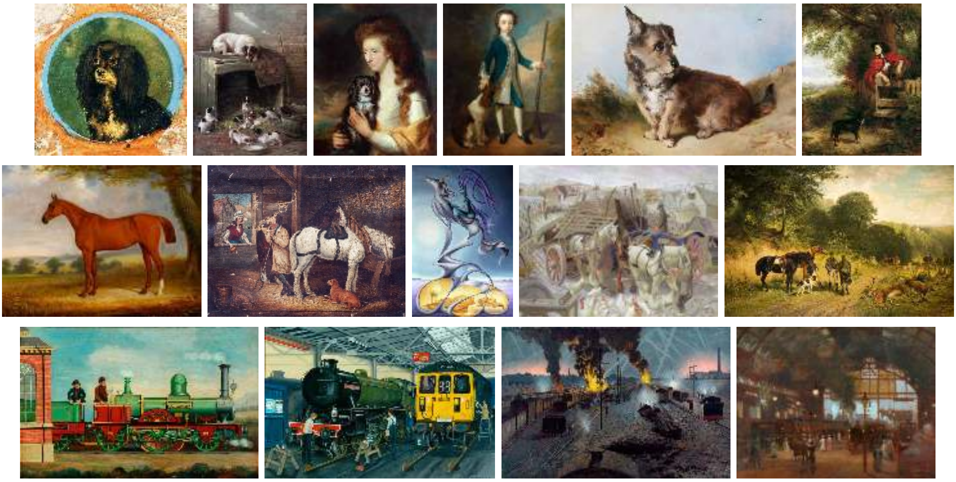


Figure 1: Example class images from the Paintings Dataset. >From top to bottom row: dog, horse, train. Notice that the dataset is challenging: objects have a variety of sizes, poses and depictive styles, and can be partially occluded or truncated.

2 Datasets, evaluation measures, and baseline classifiers

In this section we present the Paintings Dataset, and establish the difficulty of the task by training classifiers on paintings or images to determine the severity of the domain adaptation problem.

Datasets. We construct a Paintings Dataset which is used to assess performance throughout this paper. It is a subset of the publicly available 'Your Paintings' [1] dataset consisting of over 210,000 oil paintings of medium resolution (around 500 pixels in width). 10,000 of these have been annotated as part of the 'Tagger' project [4] whereby members of the public tag the paintings with the objects that they contain. The subset is obtained by searching 'Your Paintings' for annotations and painting titles corresponding to the classes of the PASCAL VOC dataset [13]. With tags and titles complete annotation is assumed in the VOC sense – that each painting has been annotated for all VOC categories – as long as 'people' are ignored, as this particular class has a tendency of appearing frequently without being acknowledged. Thus, the 'person' class is not considered, and also we do not include classes that lack a sufficient number of tags (cat, bicycle, bus, car, motorbike, bottle, potted plant, sofa, tv/monitor). Paintings are included for the remaining classes – aeroplane, bird, boat, chair, cow, dining-table, dog, horse, sheep, train. These are split at random into training, validation and test sets. The statistics are given in table 1, and example class images are shown in figure 1. The URLs for the paintings in this dataset are provided at [3].

For training we use two datasets of natural images. First, PASCAL VOC 2011 [14], and second, a set mined from Google Image Search for the VOC categories used and manually filtered to remove erroneous examples. The reason for using the Google Images dataset is that the images are typically more Caltech101 like than the VOC images, and so provide a possibly easier training scenario. The statistics of these datasets are also given in table 1.

Evaluation measures. To measure performance on the test set for each class we use the precision at rank k (Prec@k), and also the class average of this measure – the mean Prec@k (mPrec@k).

Dataset	Split	Aero	Bird	Boat	Chair	Cow	Dtable	Dog	Horse	Sheep	Train	Total
Paintings Dataset	Train	74	319	862	493	255	485	483	656	270	130	3463
	Val	13	72	222	140	52	130	113	127	76	35	865
	Test	113	414	1059	569	318	586	549	710	405	164	4301
	Total	200	805	2143	1202	625	1201	1145	1493	751	329	8629
PASCAL VOC 2011	Train	331	394	260	555	156	272	634	238	171	274	3285
	Val	340	370	251	555	152	270	654	245	154	271	3262
	Total	671	764	511	1110	308	542	1288	483	325	545	6547
Google Images	Train	90	90	90	90	90	90	90	90	90	90	900
	Val	10	10	10	10	10	10	10	10	10	10	100
	Total	100	100	100	100	100	100	100	100	100	100	1000

Table 1: The statistics for the datasets used in this paper: the number of images containing an instance of a particular class, as well as the total number of images for each subset. For datasets other than the paintings no test set is used.

2.1 Baseline classifiers

We begin by comparing two training regimes using a state of the art Fisher Vector classifier pipeline (details below). The first is by training the classifier using the trainval set of the Paintings Dataset, and the second is by training the classifier on the trainval set of the VOC dataset. In the first case there is no problem of domain adaptation, and to some extent this establishes a 'best case'. The second, addresses the task of this paper by training on natural images. In each case one-vs-the-rest classifiers are learnt for each object class, and then applied to the `Paintings Dataset` test set. For each class this gives a list of paintings ranked by classifier score.

The per class Prec@k results are given in table 2. It can be seen that for all classes there is a drop in performance when training on VOC images compared to paintings. The drop depends on the class, for example it is small for boat and large for cow. It is surprizing that the performance drop is not higher for vehicle classes considering that these have evolved significantly from their earlier forms in paintings to those in modern natural images. The explanation is probably that there are still key discriminative elements present in both cases, for example most images and paintings of boats will still contain masts and water irrespective of the time period.

Overall, there is a drop in mPrec@k from 0.98 (paintings) to 0.66 (VOC images) at $k = 5$, and from 0.91 to 0.63 at $k = 20$, i.e. a significant difference. A similar drop in mPrec@k also occurs for classifiers trained on Google images. This performance drop is also reflected in the mean Average Precision (mAP, as used in VOC, not included in the table) where the mAP drops from 0.59 for classifiers trained on paintings to 0.36 for classifiers trained on VOC images.

Implementation details: The top performing classification pipeline of [10] is used, with the implementation available from the website [2]: RootSIFT [5] features are extracted at multiple scales from each image. These are reduced using PCA to 80-D and augmented with (x,y) co-ordinates. The features are encoded with a 512 component GMM to form a 83,968D Fisher Vector [29] for each image. For each class a Linear-SVM is trained in a one-vs rest manner to rank the test images.

TrainSet	k	Aero	Bird	Boat	Chair	Cow	Dtab	Dog	Horse	Sheep	Train	Mean
`Paint`	5	1.00	1.00	1.00	1.00	1.00	1.00	0.80	1.00	1.00	1.00	0.98
VOC	5	0.80	0.40	1.00	0.60	0.00	1.00	0.40	1.00	0.60	0.80	0.66
Google	5	1.00	0.20	0.40	0.20	0.20	0.60	0.80	1.00	0.60	1.00	0.60
`Paint`	10	1.00	1.00	1.00	1.00	0.80	0.90	0.80	1.00	1.00	1.00	0.95
VOC	10	0.80	0.40	1.00	0.50	0.10	0.90	0.40	0.90	0.50	0.90	0.64
Google	10	0.90	0.10	0.70	0.10	0.30	0.50	0.50	0.90	0.70	0.80	0.55
`Paint`	20	0.95	0.90	1.00	0.95	0.75	0.90	0.70	0.95	1.00	1.00	0.91
VOC	20	0.65	0.35	0.95	0.55	0.20	0.70	0.60	0.85	0.50	0.95	0.63
Google	20	0.65	0.20	0.80	0.10	0.30	0.40	0.35	0.85	0.60	0.85	0.51

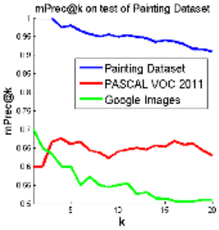


Table 2: Prec@k on the test set of the `Paintings Dataset` using classifiers learnt from different training sets. Notice the large gap in performance between the classifiers trained on paintings and those trained on natural images.

3 Spatial consistency using discriminative patches

This section describes our method for establishing and measuring spatial consistency between objects in the training images and objects in the paintings. The consistency scores obtained by this method will be used in section 4.1 for re-ranking the low rank (i.e. high classification score) paintings for each class.

The method proceeds in three stages: (i) a set of MLDPs are generated for each image in the VOC training dataset for a class (e.g. for trains); (ii) classifiers are learnt from these patches and applied as sliding window detectors to find the highest scoring regions in the paintings, leading to a set of putative correspondences; Lastly, (iii), a RANSAC [18]-style algorithm is used to select a subset of these correspondences that are spatially consistent, and each painting is scored based on this subset. We use MLDPs here for two reasons: first, because they can be obtained with minimal supervision; and second, since an MLDP covers only part of an object (rather than all of it), they are more tolerant to viewpoint and within-class variation.

3.1 Obtaining discriminative patches

Aubry *et al.* [6] provide a fast method for choosing a set of MLDPs and ranking the discriminability of these regions in an image: if q is a descriptor of an image region, μ the mean of those descriptors in a dataset, and Σ the covariance then the discriminability $|\phi(q)|$ can be measured as $|\phi(q)|^2 = (q-\mu)^T\Sigma^{-1}(q-\mu)$. This describes how the patch differs from the mean of the dataset in a whitened space.

Here we use the HOG descriptor [11], and obtain the D most discriminative patches for each training image. This forms the set of MLDPs for each image. Some examples of high scoring regions for PASCAL VOC 2011 are shown in figure 2.

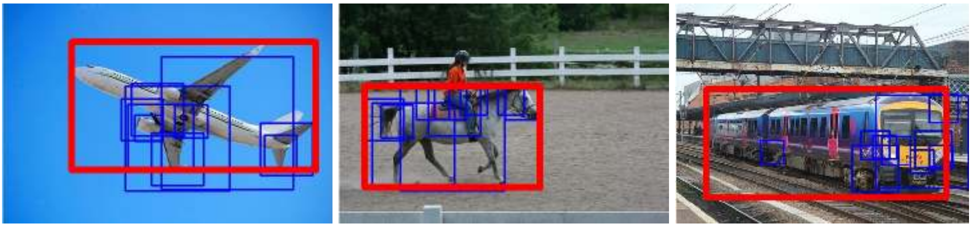


Figure 2: A subset of discriminative regions (blue) overlapping with VOC ROIs (red). Notice that informative areas of the objects are picked out such as a horse's head, and even within the ROI no indiscriminate background patches are selected.

Implementation details. Each VOC training image is annotated with a Region of Interest (ROI) for each object instance in that image. Candidate square shaped regions that overlap with the ROI are extracted from each image (and its left-right flipped version) at 3 scales per octave. For each of these a contrast-sensitive 5×5 HOG descriptor with 8×8 pixel cells is formed using the implementation of [16] resulting in a 775-D vector. μ and Σ are obtained from the training set using the method of [19] with a window size of 20 pixels. Squares are ranked and selected according to $|\phi(q)|$. Low gradient regions are ignored. Non-maximal suppression is performed using an intersection over union of 0.5 between squares as a threshold and the top D squares are retained.

3.2 Putative correspondences using MLDPs

The correspondences between the set of MLDPs in an image and a painting are established by using the patch as a detector. A Linear Discriminate Analysis (LDA) classifier [19], w, is learnt for each MLDP (i.e. the discriminative squares) in the natural image as $w = \Sigma^{-1}(q-\mu)$. LDA allows for efficient training of detectors without the need to mine for hard negatives, greatly cutting down the time required for training.

Each MLDP is used as a sliding-window detector in the manner of [16], and the highest scoring detection window on the painting is recorded. This gives a provisional correspondence (x_1, y_1, x_2, y_2, s) where (x_1, y_1) is the centre of the discriminative region used to train the classifier, (x_2, y_2) is the centre of the highest scoring detection window and s is the scale change between the two windows. The set of MLDPs creates a set of provisional correspondences between the regions used to train the classifiers in the natural image and the regions corresponding to the highest scoring detections in the painting.

3.3 Enforcing spatial consistency between correspondences

Given the set of provisional correspondences between a training set image and a painting, we now obtain a subset of these that are spatially consistent. This is achieved by fitting a linear spatial mapping. Correspondences that do not agree with this mapping are considered erroneous and removed, this enforces spatial consistency between correspondences. The mapping is a restricted similarity homography [20] that allows the object in the natural image to be uniformly scaled and translated to the painting but not rotated as:

$$\begin{bmatrix} x_2 \\ y_2 \end{bmatrix} = \begin{bmatrix} s & 0 \\ 0 & s \end{bmatrix} \begin{bmatrix} x_1 \\ y_1 \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix} \tag{1}$$

The best mapping is obtained using a RANSAC-style approach: for an image-painting pair each provisional correspondence (x_1, y_1, x_2, y_2, s) can be used to form an estimation of the mapping (1). The number of other correspondences within a scale and distance threshold of this mapping are considered to be inliers and tallied. Each correspondence is evaluated exhaustively and the mapping that produces the highest number of inliers is assumed to be the best mapping. These inliers are then used to compute an affine homography, which allows for rotation and shearing, and the number of inliers is re-estimated to provide a score. In the following section this score will be used to re-rank images.

Results. Figure 3 shows example image-painting pairs before and after enforcing spatial consistency. It can be seen that the combination of discriminative patches (that are able to ignore 'background' regions) together with the spatial consistency is able to overcome the problem of background clutter – i.e. other objects and 'stuff' in the paintings. The method is able to match class instances despite significant scale changes, and also to match parts of objects when there is partial occlusion.

4 Experiments

In this section we demonstrate that rankings obtained with the baseline classifiers of section 2.1 can be improved by re-ranking the high scoring paintings using spatially consistent sets of MLDPs. We also compare using a DPM for re-ranking, and training on the Google Image set instead of on PASCAL VOC images.

4.1 Discriminative patch re-ranking using PASCAL VOC images

We start from the `Paintings Dataset` rankings obtained by the classifiers trained on the VOC trainval images, and investigate how the re-ranking performance is affected by the number of MLDPs, D, used for each training image; and by the number, N, of paintings that are re-ranked (i.e. only the top N classifier ranked paintings are considered). We observe the effect of varying the parameters N, D on mPrec@k for low k's to determine what provides the best performance at low ranks.

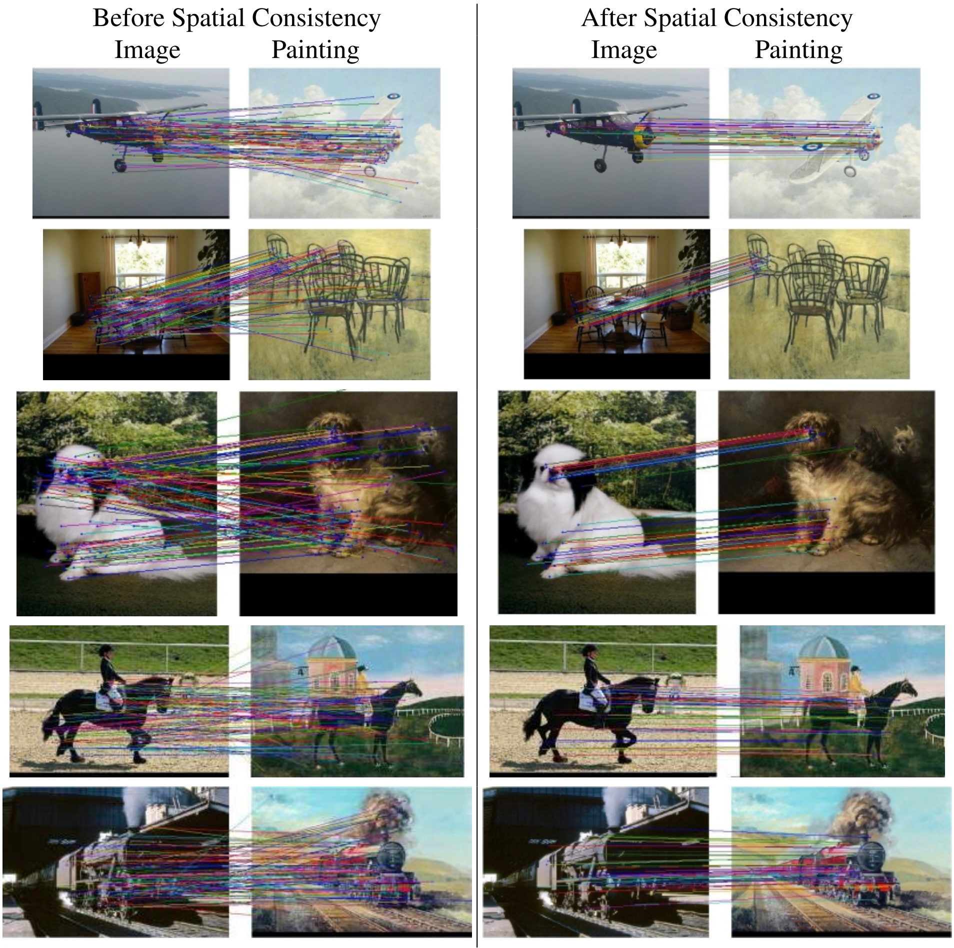
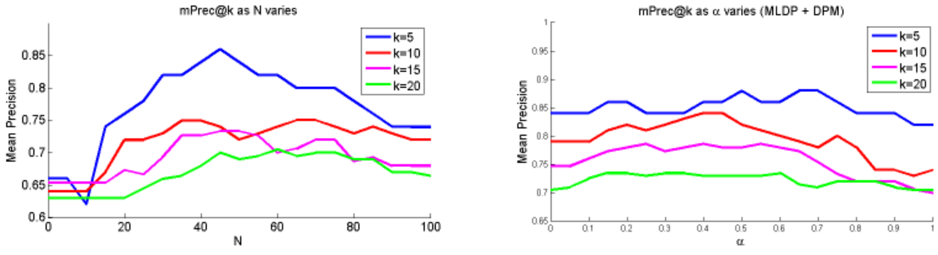


Figure 3: Image-painting pair correspondences before (left) and after (right) computing a spatially consistent subset. Note, that the MLDP correspondences are able to generalize slightly over viewpoint, intra-class differences, and between natural images and paintings.

The effect of varying N on mPrec@k for low k is shown in figure 4(a) for fixed $D = 100$. Initially, precision increases with N, but as N gets too large (for example exceeds the number of positives ranked well by the classifier), then the performance decreases, as the scoring provided by the classifier rankings are then of no benefit (if N is equal to the number of paintings, then this disregards the initial ranking). In general, mPrec@k increases with D as there needs to be sufficient patches to cover all the salient areas of the object in the image – though eventually there is insufficient increase to warrant the extra computation. In the following we set $N = 60$ & $D = 100$ to achieve high Prec@k at low ranks.

Results. Prec@k curves for selected classes before and after re-ranking are given in the first row of figure 5, and Prec@k for all classes is shown in table 3. Notice that the performance at low ranks is improved by MLDP re-ranking for almost every class; this is because for most classes an object in a painting will strongly resemble the same object in one of the natural images for that class, differing only by scale and translation with minimal rotation, allowing consistent regions to be located using MLDPs. Consider a cow; it is usually an unrotated rectangular entity, rarely seen from above – there is little variety in its pose so it is very likely that for a painting of a cow there will be a similar natural image. This also applies to isolated parts of more deformable objects; although the body of a dog is highly deformable, its face is not and will be consistent between some natural image-painting pairs. The top ranked paintings after re-ranking for selected classes are displayed in figure 6.

The one class that does not improve is dining-table. This class is highly prone to variety, a dining-table can be seen from many angles, is often covered with other objects, and frequently is heavily occluded. There is very little consistency between natural image-painting pairs.



(a) (b)

Figure 4: mPrec@k as (a) N varies, and (b) as α varies, where α controls the MLDP vs DPM score weighting of the hybrid re-ranking scheme.

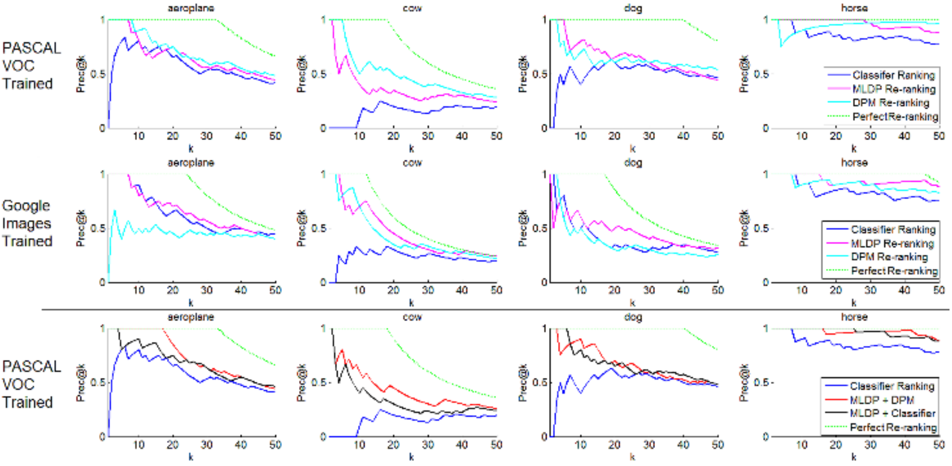


Figure 5: Prec@k on the test set of the `Paintings Dataset` for training on VOC images (top row), Google images (middle row); and VOC images (bottom row) with hybrid re-ranking. The green curves show the perfect re-ranking of the top N classified paintings for each class.

4.2 DPM re-ranking using PASCAL VOC 2011

Here, we return to the rankings obtained with the baseline classifiers of section 2.1, and re-rank using a Deformable Part Model (DPM) [16] object category detector, instead of a set of MLDPs. DPMs excel at finding spatially consistent object regions, and thus provide a natural comparison to the MLDP method.

Ranking	k	Aero	Bird	Boat	Chair	Cow	Dtab	Dog	Horse	Sheep	Train	Mean
MLDP	5	1.00	0.60	1.00	0.60	0.60	0.40	1.00	1.00	1.00	1.00	0.82
Classifier	5	0.80	0.40	1.00	0.60	0.00	1.00	0.40	1.00	0.60	0.80	0.66
MLDP	10	0.80	0.40	1.00	0.70	0.40	0.50	0.80	1.00	0.80	1.00	0.74
Classifier	10	0.80	0.40	1.00	0.50	0.10	0.90	0.40	0.90	0.50	0.90	0.64
MLDP	15	0.67	0.47	1.00	0.60	0.33	0.53	0.73	1.00	0.67	1.00	0.70
Classifier	15	0.73	0.47	0.93	0.60	0.20	0.80	0.53	0.87	0.47	0.93	0.65
MLDP	20	0.75	0.50	1.00	0.55	0.35	0.65	0.65	1.00	0.65	0.95	0.71
Classifier	20	0.65	0.35	0.95	0.55	0.20	0.70	0.60	0.85	0.50	0.95	0.63

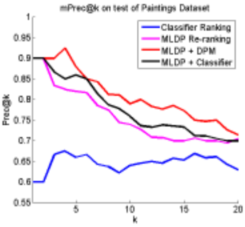


Table 3: Prec@k on the test set of the `Paintings Dataset` before and after MLDP re-ranking using VOC training images. MLDP improves Prec@k in almost all instances. The plot shows that hybrid scoring schemes (section 4.4) improve the precision even further.

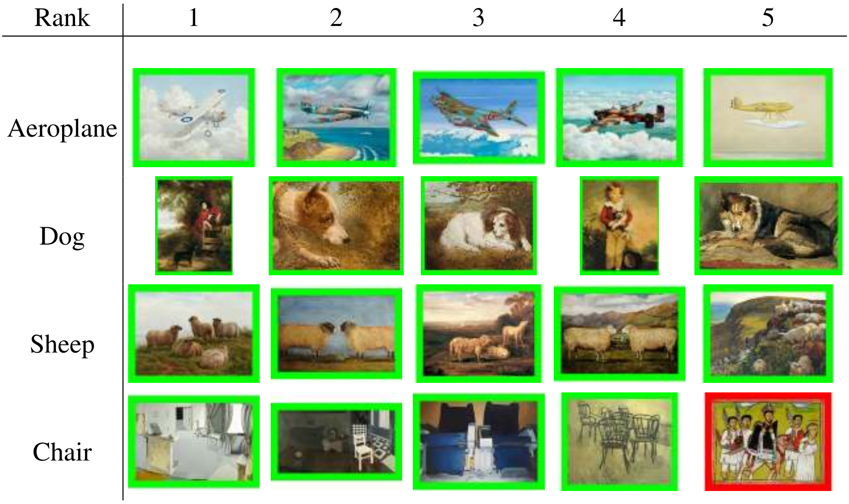


Figure 6: Top 5 ranked paintings after re-ranking using MLDP for various classes. A green border indicates a correct classification and a red border an incorrect one. Note that in the case of chair a hybrid score has been used (section 4.4).

For each class, a DPM is learnt using class ROIs as positive examples and other regions as negative examples as in [16]. Each DPM has 6 components and 8 parts. These are then applied to the top N ranked test set paintings of the `Paintings Dataset` in a sliding window cascade [17] and the score corresponding to the highest detection is recorded. The paintings are then re-ranked by this score. $N = 60$ is used to allow for direct comparison with the MLDP re-ranking of section 4.1.

Results. The Prec@k curves for DPM re-ranking for selected classes are also given in the first row of figure 5. DPM re-ranking performs better than MLDP re-ranking for objects that appear in the most generic poses; for example, a cow is usually either at front or side profile. For such classes object instances will strongly resemble one of the DPMs components – generalized from many training examples. However, for objects that assume many different poses like dog, MLDP re-ranking proves more successful as each dog only has to have a part (e.g. face) in common with a natural image rather than an entire pose in common with many.

4.3 Re-ranking using Google Images

Here, learning classifiers and re-ranking are performed using the Google images. The classifiers are again learnt in a one-vs-the-rest manner. There are three key differences between this and VOC; (i) the images are less cluttered with more centred objects like those in Caltech101, (ii) there are much fewer images, (iii) no ROI is provided, so the entire image is used as the ROI. MLDP re-ranking and DPM re-ranking are performed in the same manner as in sections 4.1 and 4.2 where both MLDP extraction and DPM training are performed on Google images. DPMs have been trained previously using the entire image as the ROI, but for scene classification [28].

The Prec@k curves for selected classes for both MLDP and DPM re-ranking are given in the second row of figure 5. MLDP re-ranking generally outperforms DPM re-ranking. This is because with a small training set it is difficult for a DPM to generalize the poses of an object, whereas for MLDP it is simply required that there exists an image resembling the pose of a painting. Note that MLDPs are able to localize an object even without the correct ROI being provided.

4.4 Hybrid re-ranking strategies

MLDPs and DPMs succeed in different scenarios; a DPM will often find the entirety of an object, whereas MLDP will find salient parts, (face, legs). A combination of these two measures can provide a good understanding of what an object is. A simple linear weighting is used to combine their scores as $\alpha A + (1-\alpha)B$, where A is the number of MLDP inliers and B is the DPM score (both normalized to lie between 0 & 1). Figure 4(b) illustrates the change in mPrec@k as α varies. The Prec@k curves when $\alpha = 0.7$ for certain classes are given in figure 5. Notice that performance is particularly high for aeroplane, this is because the DPM and MLDP re-ranking are both able to compensate for each other when one makes a mistake – for example, MLDP mapping a small part of a plane to a boat will be nullified by a low DPM score on that boat. The Prec@k when B in the above weighting is changed to the original classifier score is also given in figure 5, and the top ranked paintings for chair for this weighting are shown in figure 6. The mPrec@k for both hybrid schemes can be seen in the plot beside table 3.

5 Conclusion

In this paper we have opened up the possibility of easily learning to recognize objects in paintings starting from natural images of the objects. We have also shown that spatial consistency of discriminative patches between the paintings and natural images from the classifier training set can be used to improve the precision of low rank results.

There are at least two extensions that require further investigation: Firstly, in the case of classes like a horse, where the object is portrayed in a consistent manner, a discriminative region will likely appear at the same orientation on many other horses. On the other hand an aeroplane frequently undergoes a rotation, so a discriminative region on one aeroplane would exist in a rotated form on another, and thus the correspondence would be missed (as the MLDP detector is not rotation invariant). Future work will involve exploring the benefits of extracting discriminative patches from rotated images. Secondly, each painting is currently matched only to a single training image. However, a painting could instead be matched to multiple training images, e.g. where one part of an object has a strong match to one image, and another part to another image – a Frankenstein approach.

Acknowledgements. Funding for this research is provided by the EPSRC and ERC grant VisRec no. 228180. We are grateful for the comments provided by the anonymous reviewers and to Josef Sivic for his insights during discussions.

References

[1] BBC – Your Paintings. http://www.bbc.co.uk/arts/yourpaintings/.

[2] Encoding methods evaluation toolkit. http://www.robots.ox.ac.uk/~vgg/software/enceval_toolkit/.

[3] Visual Geometry Group Art Research. http://www.robots.ox.ac.uk/~vgg/research/art/.

[4] Your Paintings tagger. http://tagger.thepcf.org.uk/.

[5] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, 2012.

[6] M. Aubry, B. Russell, and J. Sivic. Painting-to-3D model alignment via discriminative visual elements. In *ACM Transactions of Graphics*, 2013.

[7] Y. Aytar and A. Zisserman. Enhancing exemplar svms using part level transfer regularization. In *Proc. BMVC.*, 2012.

[8] J. Burke. Nakedness and other peoples: Rethinking the italian renaissance nude. *Art History*, 36(4):714–739, 2013.

[9] G. Carneiro, N. P. da Silva, A. Del Bue, and J. P. Costeira. Artistic image classification: an analysis on the printart database. In *Proc. ECCV*, 2012.

[10] K. Chatfield, V. Lempitsky, A. Vedaldi, and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *Proc. BMVC.*, 2011.

[11] N. Dalal and B Triggs. Histogram of Oriented Gradients for Human Detection. In *Proc. CVPR*, volume 2, pages 886–893, 2005.

[12] H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *J. Artif. Intell. Res.(JAIR)*, 26:101–126, 2006.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[14] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011). http://www.pascal-network.org/challenges/VOC/voc2011/, 2011.

[15] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *IEEE CVPR Workshop of Generative Model Based Vision*, 2004.

[16] P. Felzenszwalb, D. McAllester, and D. Ramanan. A discriminatively trained, multiscale, deformable part model. In *Proc. CVPR*, 2008.

[17] P. Felzenszwalb, R. Girshick, and D. McAllester. Cascade object detection with deformable part models. In *Proc. CVPR*, 2010.

[18] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. ACM*, 24(6):381–395, 1981.

[19] B. Hariharan, J. Malik, and D. Ramanan. Discriminative decorrelation for clustering and classification. In *Proc. ECCV*, 2012.

[20] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521540518, second edition, 2004.

[21] H. Jégou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *Proc. ECCV*, 2008.

[22] R. Juan. The turn of the skull: Andreas Vesalius and the early modern memento mori. *Art History*, 35(5):958–975, 2012.

[23] M. Juneja, A. Vedaldi, C. V. Jawahar, and A. Zisserman. Blocks that shout: Distinctive parts for scene classification. In *Proc. CVPR*, 2013.

[24] B. Kulis, K. Saenko, and T. Darrell. What you saw is not what you get: Domain adaptation using asymmetric kernel transforms. In *Proc. CVPR*, 2011.

[25] Y. Lee, A. Efros, and M. Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *Proc. ICCV*, 2013.

[26] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-SVMs for object detection and beyond. In *Proc. ICCV*, 2011.

[27] S. Naderi Parizi, J. Oberlin, and P. Felzenszwalb. Reconfigurable models for scene recognition. In *Proc. CVPR*, 2012.

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

[29] F. Perronnin, J. Sánchez, and T. Mensink. Improving the Fisher kernel for large-scale image classification. In *Proc. ECCV*, 2010.

[30] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proc. CVPR*, 2007.

[31] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *Proc. ECCV*, 2010.

[32] A. Shrivastava, T. Malisiewicz, A. Gupta, and A. Efros. Data-driven visual similarity for cross-domain image matching. *ACM Transaction of Graphics*, 2011.

[33] S. Singh, A. Gupta, and A. A. Efros. Unsupervised discovery of mid-level discriminative patches. In *Proc. ECCV*, 2012.

[34] J. Woodall. Laying the table: The procedures of still life. *Art History*, 35(5):976–1003, 2012.

[35] Q. Wu and P. Hall. Modelling visual objects invariant to depictive style. In *Proc. BMVC.*, 2013.