

Modelling Visual Objects Invariant to Depictive Style (Supplementary Material)

BMVC 2013 Submission # 103

This document is a supplementary material of our paper. In the original paper, the results of classification accuracy are shown in the table 1. This document provides more details of all the experiment cases, specifically, the confusion matrix of the classification experiment will be given. The results produced by two state of art method will also be provided. These comparison results give the evidence that *it is possible to learn models of object classes that generalise across depictive styles, in the sense that it is possible to learn a model using one style but classify objects depicted in other styles, by using our proposal method*, which is claimed in our paper.

1 Training on Photos Alone

In this experiment, only photos (real object photos)¹ are trained, with different numbers. And we test on photos and artwork separately. Three methods are used.

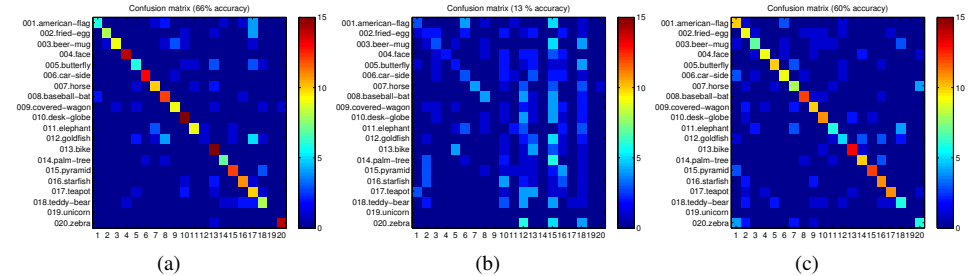


Figure 1: Train on 3 photos each class, test on 15 photos each class, using (a): Dense SIFT [9]. (b): Structure Only [9]. (c): Proposal Method

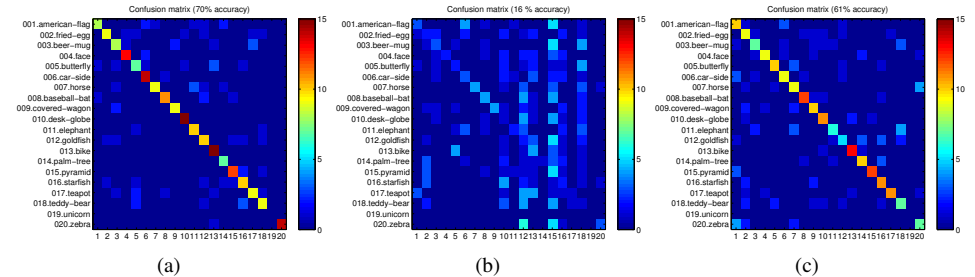


Figure 2: Train on 5 photos each class, test on 15 photos each class, using (a): Dense SIFT [9]. (b): Structure Only [9]. (c): Proposal Method

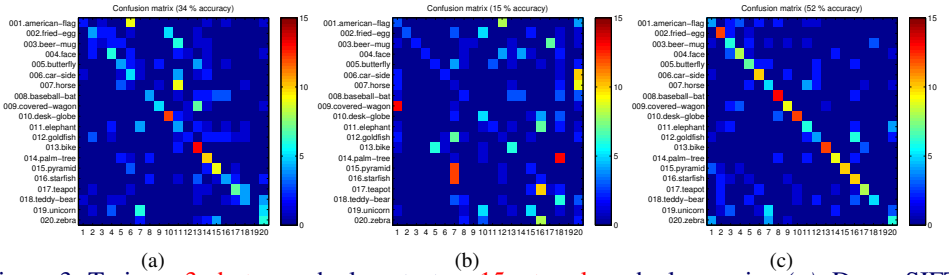


Figure 3: Train on **3 photos** each class, test on **15 artwork** each class, using (a): Dense SIFT [■]. (b): Structure Only [■]. (c): Proposal Method

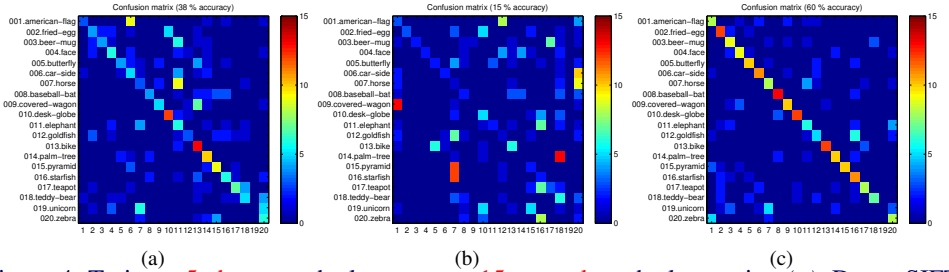


Figure 4: Train on **5 photos** each class, test on **15 artwork** each class, using (a): Dense SIFT [■]. (b): Structure Only [■]. (c): Proposal Method

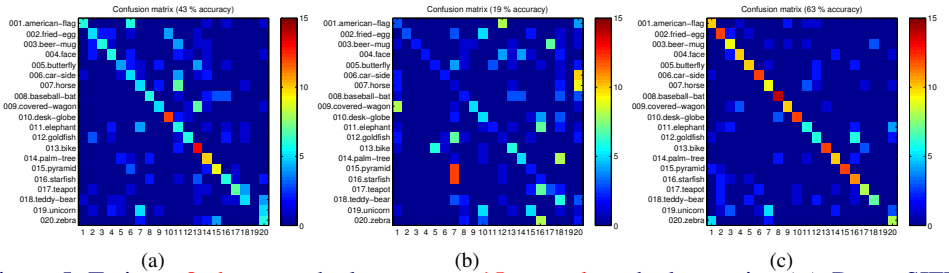


Figure 5: Train on **8 photos** each class, test on **15 artwork** each class, using (a): Dense SIFT [■]. (b): Structure Only [■]. (c): Proposal Method

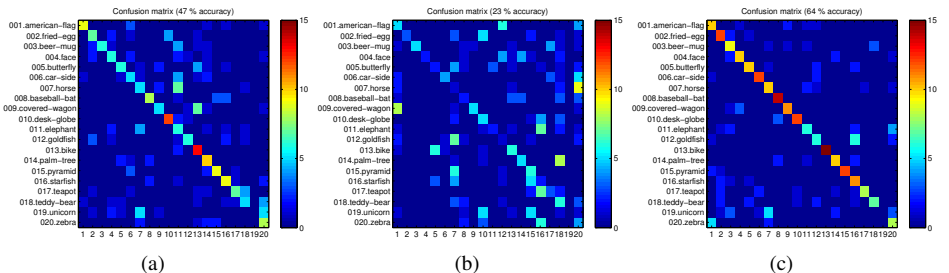


Figure 6: Train on **10 photos** each class, test on **15 artwork** each class, using (a): Dense SIFT [■]. (b): Structure Only [■]. (c): Proposal Method

2 Training on Artwork Alone

In this experiment, only Artwork (paintings, line drawings .etc) are trained, with different numbers. And we test on photos and artwork separately. Three methods are used.

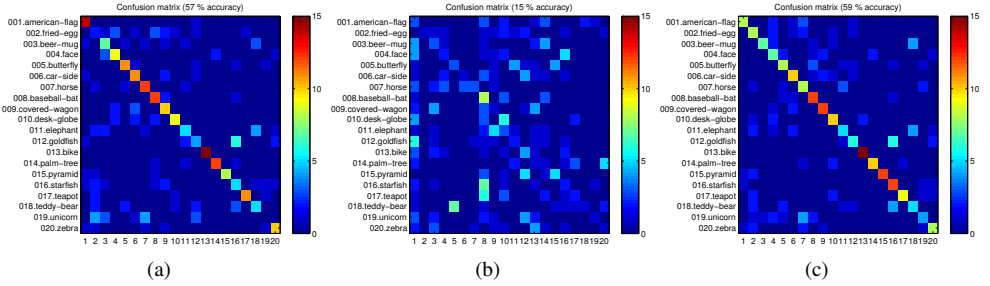


Figure 7: Train on 3 artwork each class, test on 15 artwork each class, using (a): Dense SIFT []. (b): Structure Only []. (c): Proposal Method

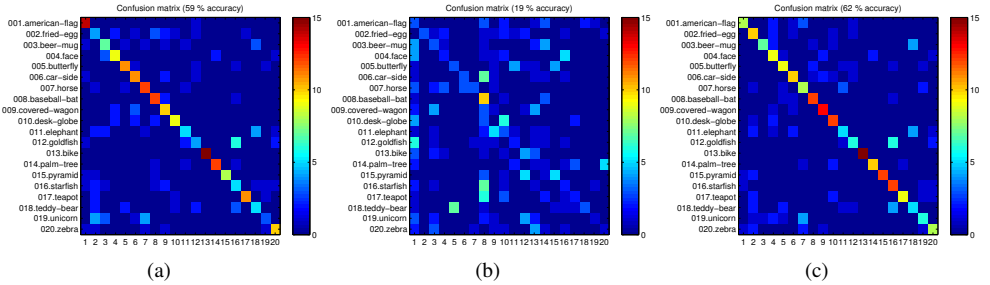


Figure 8: Train on 5 artwork each class, test on 15 artwork each class, using (a): Dense SIFT []. (b): Structure Only []. (c): Proposal Method

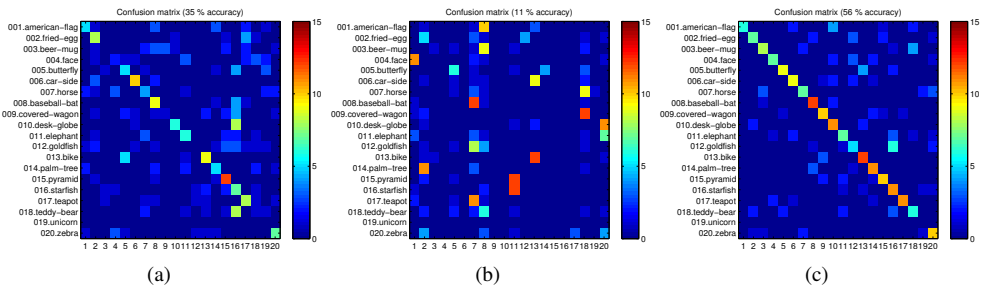


Figure 9: Train on 3 artwork each class, test on 15 photos each class, using (a): Dense SIFT []. (b): Structure Only []. (c): Proposal Method

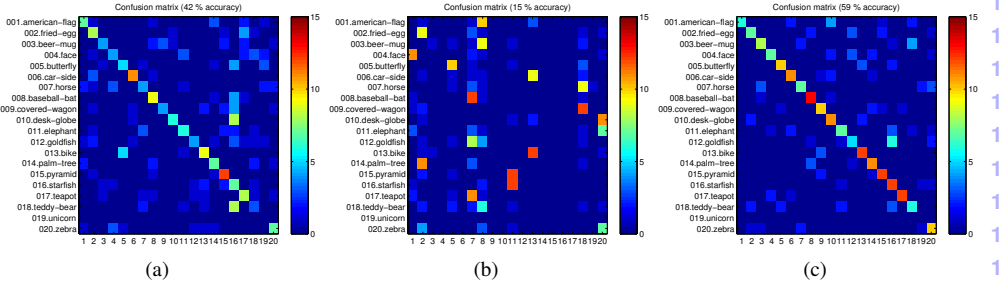


Figure 10: Train on 5 artwork each class, test on 15 photos each class, using (a): Dense SIFT [■]. (b): Structure Only [■]. (c): Proposal Method

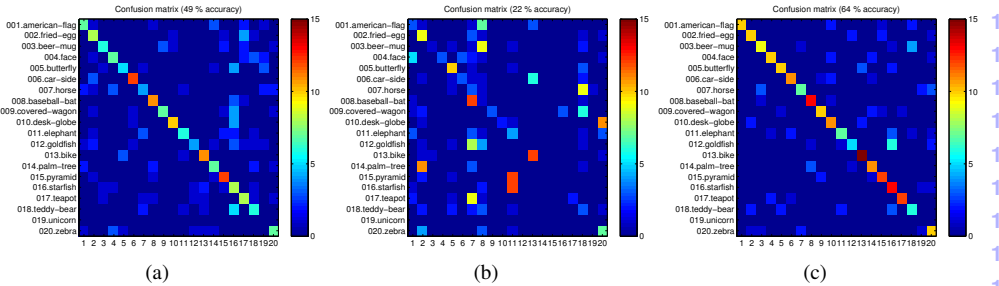


Figure 11: Train on 8 artwork each class, test on 15 photos each class, using (a): Dense SIFT [■]. (b): Structure Only [■]. (c): Proposal Method

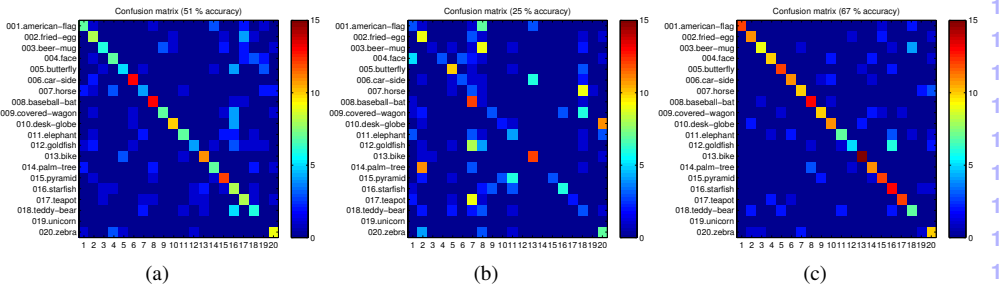


Figure 12: Train on 10 artwork each class, test on 15 photos each class, using (a): Dense SIFT [■]. (b): Structure Only [■]. (c): Proposal Method

3 Training a Mixture

In this experiment, both artwork (paintings, line drawings .etc) and photos are trained as a mixture, with different numbers. And we test on photos and artwork separately. Three methods are used.

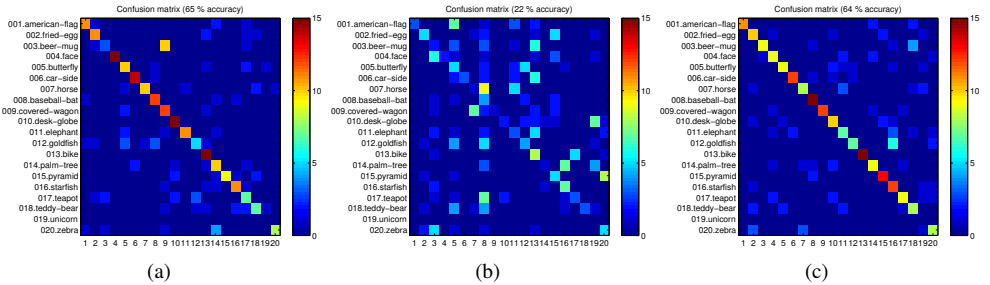


Figure 13: Train on 3 artwork+3 photos each class, test on 15 photos each class, using (a): Dense SIFT [14]. (b): Structure Only [14]. (c): Proposal Method

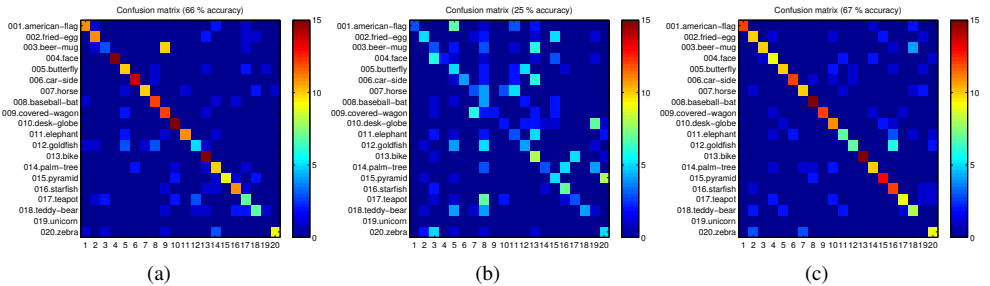


Figure 14: Train on 5 artwork+5 photos each class, test on 15 photos each class, using (a): Dense SIFT [14]. (b): Structure Only [14]. (c): Proposal Method

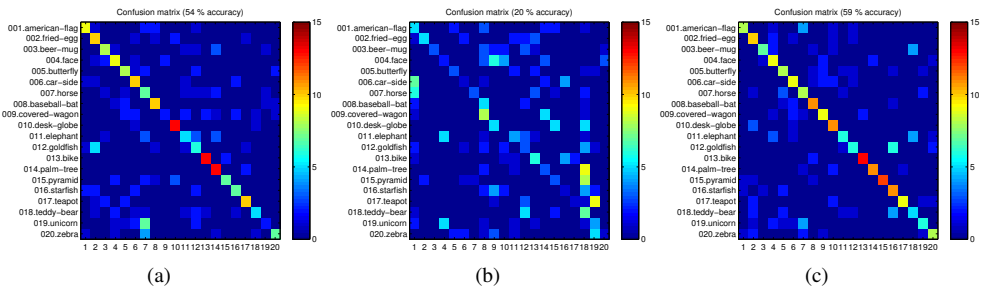


Figure 15: Train on 3 artwork+3 photos each class, test on 15 artwork each class, using (a): Dense SIFT [14]. (b): Structure Only [14]. (c): Proposal Method

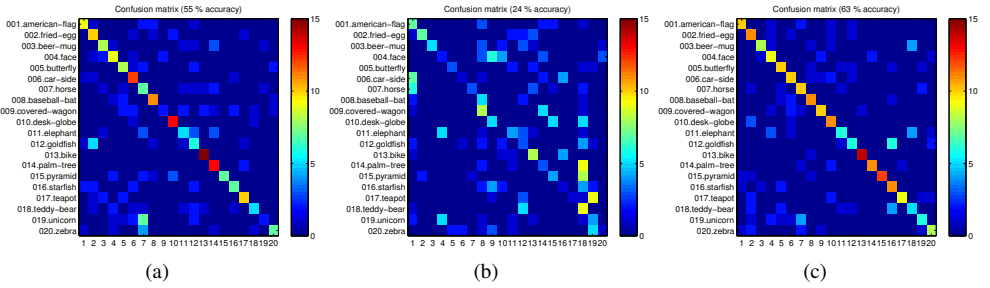


Figure 16: Train on 5 artwork+5 photos each class, test on 15 artwork each class, using (a): Dense SIFT [■]. (b): Structure Only [■]. (c): Proposal Method

4 Performance table and figure

case 1: Training	3a	5a	3p	5p
case 1: Testing	15a	15a	15p	15p
Dense SIFT [■]	57%	59%	66%	70%
Structure Only [■]	15%	19%	13%	16%
Proposed Method	59%	62%	60%	61%

case 2: Training	3p	5p	8p	10p	3a	5a	8a	10a
case 2: Testing	15a	15a	15a	15a	15p	15p	15p	15p
Dense SIFT [■]	34%	38%	43%	47%	35%	42%	49%	51%
Structure Only [■]	15%	15%	19%	23%	11%	15%	22%	25%
Proposed Method	52%	60%	63%	64%	56%	59%	64%	67%

case 3: Training	3a	5a	3p	5p
case 3: Testing	30m	30m	30m	30m
Dense SIFT [■]	46%	50%	50%	54%
Structure Only [■]	13%	16%	14%	16%
Proposed Method	58%	61%	56%	61%

case 4: Training	6m	10m	6m	10m	case 4: Training	6m	10m
case 4: Testing	15p	15p	15a	15a	case 4: Testing	30m	30m
Dense SIFT [■]	65%	66%	54%	55%	Dense SIFT [■]	60%	61%
Structure Only [■]	22%	25%	20%	24%	Structure Only [■]	21%	24%
Proposed Method	64%	67%	59%	63%	Proposed Method	62%	65%

Table 1: Classification accuracy for different cases. From top to bottom, left to right: (a) single domain task, (b) single cross depiction task, (c) single to mixture depiction task, and (d) mixture cross depiction task. The character 'p' is 'photos', 'a' is 'art' and 'm' is 'mixture'. The deeper the color, the better the performance. The numbers of images in the table are *per-class* figures, the rates are averaged over 20 classes. In total our test used 800 images, including our extension to CalTech 256.

From the results, it is clear shown that our proposed method performs better than both traditional bag-of-words method and structure only method in terms of cross-depiction generalisation. This is most obvious when training in one depiction and testing in another: the accuracy of our method is nearly 20 percent points higher than the method using dense SIFT, and nearly 50 percentage points higher than structure alone. The traditional BoW method is superior by up to 9% when photographs are used for both training and testing – which is the

¹ Please note there are not unicorns in the real photos and in the world

kind of result we anticipated since BoW models are tuned to low-variance features whereas we set out to allow for wider variation. The low score of the structure-only method may be explained by our use of more complex structures than the original [1].

The case of training on both photographs and artwork is interesting. When photographs are the test case BoW and our method perform about equally, but our method performs the better when artwork is used to test. It is notable that “artwork” in fact covers a broad variety of depictions. This result suggests that word formation inside BoW is biased towards “photographic words”, where we would expect the densest concentration of features. This is underlined by the fall in performance of BoW when the number of artworks in the training set rises. Our method is very stable to different test cases, around 60%, in all cases, whereas BoW is more variable in performance.

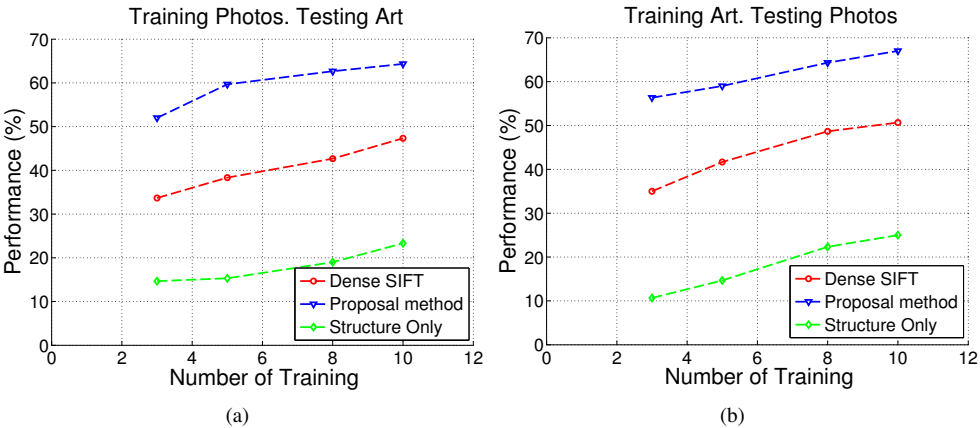


Figure 17: Performance trend when using different numbers of training images in case 2, the single cross depiction task. (a) Photos to artwork classification task. (b) Artwork to photos classification task. Our method (the blue one) outperforms the other two methods obviously.

References

- [1] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms, 2008.
- [2] Bai Xiao, Song Yi-Zhe, and Peter Hall. Learning invariant structure for object identification by using graph methods. *Computer Vision and Image Understanding*, 115(7): 1023–1031, 2011.