This ICCV workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Beyond Photo-Domain Object Recognition: Benchmarks for the Cross-Depiction Problem

Hongping Cai
Department of Computer Science
University of Bath, UK
H.Cai@bath.ac.uk

Qi Wu
School of Computer Science
University of Adelaide, Australia
qi.wu01@adelaide.edu.au

Peter Hall
Department of Computer Science
University of Bath, UK
pmh@bath.ac.uk

Abstract

The cross-depiction problem is that of recognising visual objects regardless of whether they are photographed, painted, drawn, etc. *It introduces great challenge as the variance across photo and art domains is much larger than either alone. We extensively evaluate classification, domain adaptation and detection benchmarks for leading techniques, demonstrating that none perform consistently well given the cross-depiction problem. Finally we refine the DPM model, based on query expansion, enabling it to bridge the gap across depiction boundaries to some extent.*

1. Introduction

Humans are able to recognise objects in an astonishing variety of forms. The same is not true of computers. Even the very best classification and detection algorithms exhibit a significant drop in performance when presented with images that are not photographic. Figure 1 provides a hint of the reason for such a performance drop. It shows the distribution of art features is visually distinct from that of photo features. This wide variation is a property of all visual classes we have tested, and underpins the intuition that the underlying difficulty in the cross-depiction problem is the seemingly unbounded number of distinct depictive styles.

This *cross-depiction* problem is under-explored, interesting. Advancing this area would provide a significant boost to current applications such as image search over a database. More importantly, a solution to the cross-depiction problem forces us to consider ways to represent objects that are more general than appearance-based approaches currently used.

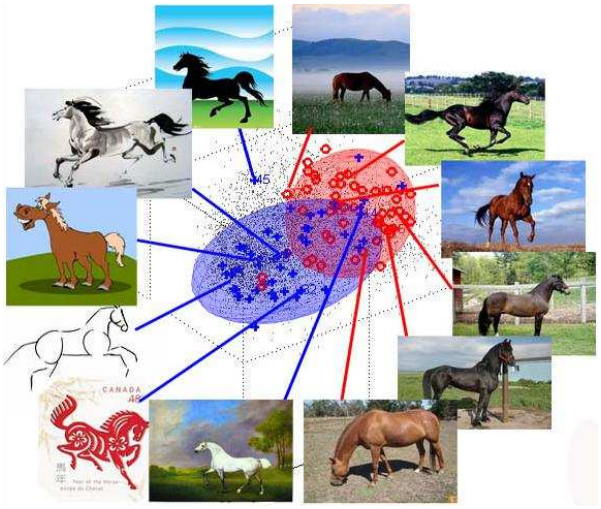

Figure 1. Example images and their distribution in the BoW-SIFT feature space. The art domain (blue) and photo domain (red) of 'horse' distribute differently, partially overlapped. The art features tend to spread wider than the photo features, which is consistent with its higher variation of visual appearance.

The cross-depiction problem is beginning to receive attention in the computer vision community. However, this nascent field suffers from a record of baseline performance of algorithms. Due to the very different distribution of photo and art domains, it is natural to resort to the domain adaptation techniques. However, how much the existing domain adaptation methods could help is unclear. Accordingly this paper makes three contributions.

1. We confirm the intuition by experiment that the variance across photo and art domains is much larger than the conventional cross domain problem.

2. We benchmark leading recognition (both shallow and deep representations) and detection methods, state-of-the-art domain adaptive methods for cross-depiction task, showing none perform well.

3. We argue that cross depiction expansion is of value in bridging the 'cross-depiction gap'.

2. Related Work

Cross-depiction problems are comparatively less well-explored. Most existing researches on non-photograph recognition usually consider specific styles. For example, [18, 21] searched a database of photographs based on a sketch query; edge-based HoG is efficient. The most recent work [33] designed a specific deep network for sketch recognition which even outperforms the human. As annotation is a human intensive and time-consuming task, [28] did the first research on training a pedestrian model from video-game images, then used it for pedestrain detection in real-world images, i.e., photographs. Others [5, 6] explored object retrieval over oil paintings with the model trained on photographs. They concluded that classifiers trained on photographs have quite success in retrieving paintings. However, such success might not happen if consider all possible image styles, which is what this paper focuses on.

It is natural to seek domain adaptation methods to address the cross-depiction problem. We would not give a comprehensive review of this active field. The readers are referred to the recent two excellent surveys [14, 24] about visual domain adaptation.

Algorithms usually assume that the training and test data are drawn from the same distribution. This assumption may be breached in real-world applications, leading to domain adaptation (DA) methods. Both sampling geodesic flow (SGF) [13] and geodesic flow kernel (GFK) [12] use intermediate subspaces on the geodesic flow connecting the source and target domain. SGF and GFK assume that the best path between source and target domains is the geodesic curve (shortest path). However, this curve does not provide the necessary flexibility to model the domain shift. In contract, the Sampling Spline Flow (SSF) [2] uses more complex curves (e.g. splines) on the manifold to interpolate between multiple sources and the target domain. Instead of building a set of intermediate representations, [10] learns a linear transformation function that align the source subspace coordinate system to the target one, by using a subspace alignment (SA) approach. Both SA and GFK can be unified in the frame of subspace distribution alignment (SDA) [27]. It aligns the distributions as well as the subspace bases.

Besides the above feature augmentation-based approaches, dictionary-learning methods [23, 32] and domain re-sampling methods [11, 23] can also help to bridge the domains. [32] not only learns a common dictionary to encode the domain-shared features, but also learn a set of domain-specific dictionaries to model the domain shift. [11] attempts to bridge the two domains using feature instances, called landmarks, which are distributed similarly to the target domain. [23] incrementally learn the dictionary by augmenting the source data with the so-called supportive samples in the target domain. The supportive samples are chose if they are close to the source domain and could reduce the domain mismatch. Our proposed algorithm, DPM with a cross-depiction expansion paradigm, also belongs to domain re-sampling methods. The expanded set belongs to the target domain but is similar to the source, which enables the detection models retrained with this set would generalise better to the target domain.

The above methods [2, 10, 11, 12, 23, 32] yield state-of-the-art performance on the standard cross-domain dataset [26], i.e., photographs acquired under different environmental conditions, at different times, or from different viewpoints – none have been tested on the cross-depiction problem. DA is a promising direction to address cross-depiction when one regards artwork and photographs as belonging to different domains. However, cross-depiction possess more challenges, as we demonstrate in Table 1 that even the largest divergence in the cross-domain datasets is still not comparable with the photo-art divergence.

3. Divergence of Cross-depiction Data

Photo-Art-50 [31] is a recently released cross-depiction dataset. It contains 50 objects, 90 to 138 images for each object with approximately half photos and half art images.

To visualise how the photos and artworks are distributed, we display a few horse images in Fig 1. Differences between photo and art images are significant. One may notice that the art domain exhibits larger diversity than photos in the visual appearance. Such diversity is demonstrated with its larger variance in the feature space as shown in Fig 1. This explains why the classification performance on art domain is inferior to the photo domain in Sec. 4.

K-L Divergence: In order to discover how much statistical difference exists in the feature space, we compute the symmetric Kullbeck-Liebler divergence between art and photo feature distributions. Each image is represented as a 5000-d BoW histogram with dense SIFT descriptors, as described in Sec. 4.1. We project these features to 10-d subspace using PCA, and approximate the distributions using Gaussian densities respectively. The symmetrical K-L divergence are then computed.

Table 1 illustrates the K-L divergences between photo and art images in Photo-Art-50 [31]. We also compute the K-L divergences for domain pairs [12, 26] under different photographic conditions for comparison. Three pairs have large diversity, which is consistent with the observation in

Cross-domain datasets [12, 26]					Photo-Art-50 [31]
C-A	C-D	A-W	D-A	D-W	Photo-Art
0.079	0.271	0.239	0.292	0.047	**0.466**

Table 1. Comparison of K-L divergence between domain pairs. C - Caltech-256, A - Amazon, W - WebCam, D - DSLR.

[12]. However, even the largest K-L divergence in the cross-domain dataset is still not comparable with the photo-art divergence. This clearly tells that the photo-art distributions differ much more than distributions of photos capturing in different conditions.

4. Classification Benchmarks

We evaluate both the shallow and deep representations for the classification on the Photo-Art-50 dataset.

4.1. Bag-of-Words

We use Lazebnik *et al.*'s version of **BoW**, namely, spatial pyramid [20], implemented with the VLfeat [29]. Each image is coded by a 5000-bin histogram (1000 codewords, 2 pyramid levels). A one-versus-all linear SVM classifier is then trained on a χ^2-homogeneous kernel map [30].

Choice of feature may be crucial to the cross depiction problem. Thus we test a collection of features, as follows:

SIFT [22] is a 128-d vector created by stacking 8-bin orientation histograms on 4×4 cells. Geometric Blur (**GB**) [1] describes local regions by geometrically blurring oriented edge maps. It is able to match object parts with very different appearance in two images. Self-similarity descriptors (**SSD**) [3] measure local self-similarity patterns by correlating a tiny local patch within a larger local region. It computes local correlations of patches rather than pixel values, and performance well at matching similar objects invariant to depictive styles. Histogram of Oriented Gradient (**HOG**) [7] is a vector of normalised histograms from tiled block regions. The gradients in HOG are quantised into 9 orientations and 4 cell sizes. Unlike standard HOG which extracts the descriptor on the original image map, **edge-HOG** [17] computes HOGs over edge maps.

4.2. Fisher Vector

Instead of counting the codewords occurrence in BoW, Fisher Vector (**FV**) records the statistic information of local features inside each cluster.

The FV of an image is the stacking of the mean and covariance deviation vectors for each of the K clusters in the Gaussian mixture. Like BoW, spatial pyramid is also applied in this experiment. Then, a one-versus-all linear SVM classifier is trained on the Fisher vectors obtained from all training images.

4.3. Deep Representation

Deep convolutional neural networks (**CNNs**) [19] have shown clear performance boost in image classification. We follow the network architecture in [19]. It consists of 5 convolutional layers and 3 fully-connected layers. The 4096-d CNN features, whose model is pre-trained on ILSVRC2012 dataset, are first extracted on all images. They are then fed to learn a one-vs-all SVM classifier with chi-squared kernel.

4.4. Discussion

Comparing different local descriptors in the BoW framework in Table 2, SIFT wins all training-test combinations except 'Photo-Art' setting. Surprisingly, though SSD is designed for matching a common 'shape' regardless of their appearance, it performs poorly in classification on both same domain and different domains. EdgeHOG outperforms the standard HOG when art images are involved, which is consistent with the observation of [8, 17]. This may also explain the good performance of BoW-GB which also computes the descriptor on the edge map. When testing on the art domain, BoW-GB performs competitively and even outperforms BoW-SIFT when training on photo domain. This might result from the fact that edges posses some invariance across photo and artworks.

Consistent with the observation in [25], FV-SIFT outperforms BoW-SIFT by 2-3% in all 'train-test' settings. In spite of such an improvement, FV still suffers from significant performance drop in the condition of different training and test depiction domains.

Not surprisingly, the pre-trained CNN features obtain large gain in all the settings. Especially, when the testing set is photo, it can reach above 90% correct rate. In contrast, the performance of testing on artworks drops. This results from the fact that CNNs are pre-trained on a photo dataset, which has limited generalisation ability.

Both shallow and deep representations share the same trend: All methods show a significant drop when trained on one depiction style and tested on another. The most difficult one is the 'train-on-photo-test-on-art' setting. It can be explained by the degree of variation in the features as evidenced in Table 1. This tells that they do not generalise well across depictive styles.

5. Domain Adaptive Benchmarks

In dealing with mismatched distributions between the training set and the test set, domain adaptive methods [10, 12, 13, 15, 16, 26] have shown clear benefits. However, all these methods have been tested only on datasets containing photographs with different capture conditions, so we test them on the cross-depiction classification task.

Intuitively, the distribution between photographs and artworks would have a greater variability. This intuition has

model		BoW					FV	CNN
train	test	SIFT	GB	SSD	HOG	edgeHOG	SIFT	Pre-trained
Photo	Photo	83.69 ± 0.6	76.83±1.4	66.48±1.3	72.40±0.8	70.04±1.0	**87.42±0.5**	96.95±0.3
A+P	Photo	80.38 ± 1.1	71.94±1.1	57.85±0.9	64.67±1.4	63.25±1.3	**83.53±0.7**	96.23±0.5
Art	Photo	63.93 ± 1.1	59.90±0.8	38.89±1.6	42.45±1.1	50.13±1.4	**65.67±0.5**	90.50±0.7
Art	Art	74.25 ± 1.1	72.05±1.4	49.03±1.4	55.13±0.6	59.55±0.6	**76.74±0.5**	89.24±0.5
A+P	Art	69.47 ± 1.1	67.08±0.6	45.27±2.1	49.87±1.0	56.07±2.0	**72.82±1.0**	87.13±1.2
Photo	Art	43.78 ± 0.6	**50.42±1.4**	31.16±1.0	28.99±1.4	39.91±1.6	47.35±1.2	72.54±1.3

Table 2. Categorisation performance on the Photo-Art-50, with 30 images per category for training. Average correct rates are reported by running 5 rounds with random training-test split. 'A+P' stands for a mixture training set of 15 photo images and 15 art images.

been verified by the higher K-L divergency than standard cross-domain problem in Sec. 3. It is unclear if the current domain adaptive methods can handle such high diversity between photos and artworks. To find the answer, two state of the art methods are evaluated:

Geodesic Flow Kernel (**GFK**) [12] models the source domain $\mathcal{S}$ and target domain $\mathcal{T}$ with lower dimensional linear subspaces and embeds them onto a Grassmann manifold. The geodesic flow is parameterized as a curve between these two subspaces on the manifold. See Gong *et al* for mathematical details [12]. Following Gong *et al* [12], we generate two variants of GFK kernels: **GFK_PCA** and **GFK_LDA**. GFK_PCA means that the original features are projected onto the 49 dimensional subspace with PCA on each domain. In contrast, GFK_LDA replaces PCA with supervised dimension reduction method – linear discriminant analysis (LDA) – on the source domain. As LDA takes label information into account in the training stage, it possesses more discriminability for classification. Subspace Alignment (**SA**) [10] projects each source domain $\mathcal{S}$ and target domain $\mathcal{T}$ to its respective subspace. Then, a linear transformation function is learned to align the source subspace coordinate system to the target one.

Other than the original 5000-d BoW-SIFT features (**OrigFeat**) as described in Sec. 4.1, we also compare GFK and SA with another two no-domain-adaptation methods, the projected features with PCA bases from the source domain (**PCA_S**) and from the target domain (**PCA_T**), respectively. For the classifier, we implement both the Nearest Neighbour (1-NN) and the linear SVM. (Different from using chi-square kernel of SVM as in Sec. 4.1, here we use linear SVM for fair comparison.)

Discussion: Fig. 2 compares domain adaptation methods with no-adaptation methods. PCA_T always produce higher accuracies than PCA_S, due to the better approximation of the distribution in the target domain. Using NN, GFK_LDA performs the best. However, the gain of GFK_LDA with either classifier is very little compared with PCA_T. Regarding SVM, the original feature surprisingly outperforms all the the other projected features, even the domain adaptive methods. In addition, we also replace the BoW-SIFT with the 4096-d CNN features as the original features for domain adaptation and observe higher accuracies, but the performance rank remains the same.

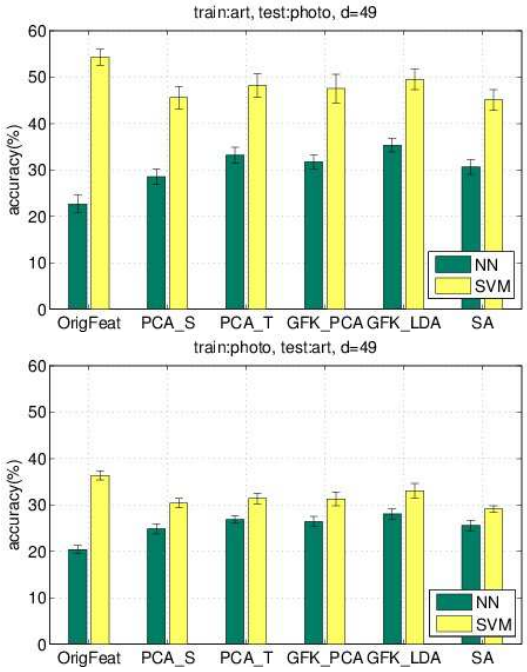

Figure 2. Classification accuracies without (OrigFeat, PCA_S and PCA_T) and with (GFK_PCA, GFK_LDA, SA) domain adaptive methods on Photo-Art-50. 'OrigFeat' means classifying with the original 5000-bin BoW-SIFT histograms. Except OrigFeat, the rest methods are with 49-d projected features.

Different from the effectiveness in conventional domain adaptive problem, GFK [12] and SA [10] fail in the cross-depiction problem. Since the main difference between artworks and photos originates in the local textures, it may cause the image presentations to differ too much, This difference leads to either the case that no such smooth manifold exists or that the two subspaces are located too far apart on the manifold. Negative effects might occur with direct domain adaptation in such situations.

6. DPM with Cross-Depiction Expansion

Deformable part model (DPM) [9] performs remarkably well in the context of object detection due to its capability of modelling variations both in appearance and non-rigid

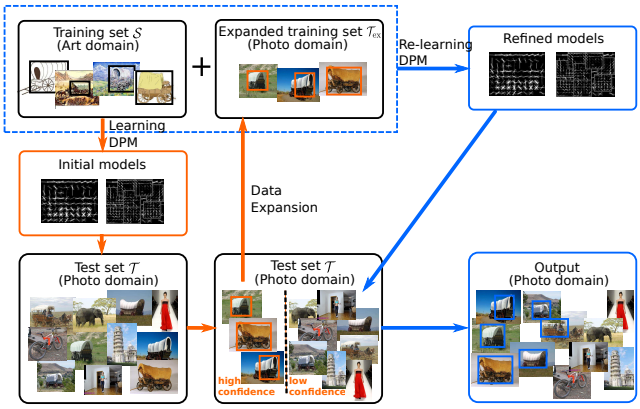

Figure 3. The pipeline of DPM with cross-depiction expansion.

deformation. However, it has never been tested in artworks. Therefore, we evaluate its cross-depiction ability on Photo-Art-50 given annotated bounding boxes in the training set.

6.1. Recall of DPM

DPM [9] models an object with a star structure, *i.e.*, a root filter plus a set of parts. Given the location of the root and the relative location of n parts, the score of the star model is the sum of responses of the root filter and parts filters, minus the displacement cost. The score function for an example x is defined as $f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z)$ where $z = (p_0, ..., p_n)$ is a latent vector denoting the location, and set $Z(x)$ the possible latent values for x. β denotes a vector of model parameters and Φ is a concatenation of HOG features and part displacement features. The parameters are then solved with latent SVM.

6.2. DPM-CDE

In query expansion [4], the most similar documents/images of the original query are reissued as new queries. Due to additional information from the expanded queries, the retrieval performance is significantly improved. We borrow this idea for the cross-depiction detection. The pipeline is illustrated in Fig. 3. We first train the DPM model for each object category in the source domain $\mathcal{S}$. Then we apply the models on the target domain $\mathcal{T}$. A confidence set $\mathcal{T}_{ex} \subset \mathcal{T}$ is picked from the target domain for training expansion. Statistically, these expansion data can be treated as the overlap area of the two domains, as shown in Fig. 1. We re-learn the DPM model on the expanded training set $\mathcal{S} \cup \mathcal{T}_{\text{ex}}$. Then this refined DPM model, named **DPM-CDE**, is used in the detection task.

The principle for adding a test image x to $\mathcal{C}$ is: *DPM responses to one certain object class is distinctively strong.*

$$\mathcal{C} = \{x \in \mathcal{T} | s_1(x) > \theta_1 \wedge s_1(x) - s_2(x) > \theta_2\} \quad (1)$$

with $s_1(x)$ the highest DPM score, and $s_2(x) \leq s_1(x)$ the second highest score; θ_1, θ_2 are user-specified parameters to threshold the best score and margin respectively. We found $\theta_1 = -0.8$ and $\theta_2 = 0.1$ to be a good trade-off between minimising false positives (5%) and including appropriate number of expanded data (around 580 images in $\mathcal{C}$).

train	test	DPM	DPM-CDE
Photo	Photo	0.957	–
Art	Photo	0.798	**0.843**
Art	Art	0.839	–
Photo	Art	0.727	**0.783**

Table 3. Comparison of mean average precision (mAP) on Photo-Art-50, 30 images per object for training.

Discussion: Table 3 compares the detection performance on Photo-Art-50 with DPM and DPM-CDE. Compared with those 'train on A, test on A', the performance of standard DPM in 'train on B, test on A' condition significantly drops. However, this performance gap is shortened when the DPM model is re-learned on the expanded training set $\mathcal{S} \cup \mathcal{T}_{\text{ex}}$. It is shown that DPM-CDE improves the performance by around 5%, which demonstrates that the expanded set does capture new information in the target domain and helps to refine the models according to the target domain.

7. Conclusion

The cross-depiction problem is an important new challenge for computer vision research. We have demonstrated that the feature distributions of the photographic and art domains differ more strongly than those in conventional domain adaptation research.

Our classification and detection benchmark experiments on this dataset show that all state-of-the-art methods exhibit a drop in performance, given the cross-depiction problem. This drop is explained by the wide diversity of features. This also leads to the fact that domain adaptation could hardly solve this problem.

As a first attempt of addressing this problem, we retrained DPM on an expanded training set. It provided clear performance benefits, which implies that the cross-depiction expansion is a simple but effective way of narrowing the gap between photo and art domains.

References

[1] A. C. Berg and J. Malik. Geometric blur for template matching. In *CVPR*, 2001. 3

[2] R. Caseiro, J. F. Henriques, P. Martins, and J. Batista. Beyond the shortest path : Unsupervised domain adaptation by sampling subspaces along the spline flow. In *CVPR*, June 2015. 2

[3] K. Chatfield, J. Philbin, and A. Zisserman. Efficient retrieval of deformable shape classes using local self-

similarities. In *Workshop on Non-rigid Shape Analysis and Deformable Image Alignment, ICCV*, 2009. 3

[4] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. Total recall: Automatic query expansion with a generative feature model for object retrieval. In *ICCV*, 2007. 5

[5] E. Crowley and A. Zisserman. In search of art. In *ECCV Workshop: VisArt*, 2014. 2

[6] E. Crowley and A. Zisserman. The state of the art: Object retrieval in paintings using discriminative regions. In *BMVC*, 2014. 2

[7] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, volume 2, pages 886–893, 2005. 3

[8] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa. Sketch-based image retrieval: Benchmark and bag-of-features descriptors. *IEEE Trans. Visualization and Computer Graphics*, 17(11):1624–1636, 2011. 3

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010. 4, 5

[10] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *ICCV*, 2013. 2, 3, 4

[11] B. Gong, K. Grauman, and F. Sha. Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In *ICML*, pages 222–230, 2013. 2

[12] B. Gong, Y. Shi, F. Sha, and K. Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *CVPR*, pages 2066–2073, 2012. 2, 3, 4

[13] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, volume 0, pages 999–1006, 2011. 2, 3

[14] R. Gopalan, R. Li, and R. Chellappa. Unsupervised adaptation across domain shifts by generating intermediate data representations. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(11):2288–2302, 2014. 2

[15] R. Gopalan, R. Li, V. Patel, and R. Chellappa. Domain adaptation for visual recognition. *Foundations and Trends in Computer Graphics and Vision*, 2015. 3

[16] J. Hoffman, T. Darrell, and K. Saenko. Continuous manifold based adaptation for evolving visual domains, 2014. 3

[17] R. Hu, M. Barnard, and J. P. Collomosse. Gradient field descriptor for sketch based retrieval and localization. In *ICIP*, pages 1025–1028, 2010. 3

[18] R. Hu and J. Collomosse. A performance evaluation of gradient field hog descriptor for sketch based image retrieval. *CVIU*, 117(7):790–806, 2013. 2

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1097–1105, 2012. 3

[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, volume 2, pages 2169–2178, 2006. 3

[21] Y. Li, Y.-Z. Song, and S. Gong. Sketch recognition by ensemble matching of structured features. In *BMVC*. Citeseer, 2013. 2

[22] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Intl. Journal of Computer Vision*, 60(2):91–110, 2004. 3

[23] B. Lu, R. Chellappa, and N. M. Nasrabadi. Incremental dictionary learning for unsupervised domain adaptation. In *BMVC*, 2015. 2

[24] V. M. Patel, R. Gopalan, R. Li, and R. Chellappa. Visual domain adaptation: A survey of recent advances. *IEEE Signal Process. Mag.*, 32(3):53–69, 2015. 2

[25] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010. 3

[26] K. Saenko, B. Kulis, M. Fritz, and T. Darrell. Adapting visual category models to new domains. In *ECCV*, pages 213–226, 2010. 2, 3

[27] B. Sun and K. Saenko. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, 2015. 2

[28] D. Vázquez, A. M. López, J. Marín, D. Ponsa, and D. G. Gomez. Virtual and real world adaptationfor pedestrian detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(4):797–809, 2014. 2

[29] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008. 3

[30] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010. 3

[31] Q. Wu, H. Cai, and P. Hall. Learning graphs to model visual objects across different depictive styles. In *ECCV*, 2014. 2, 3

[32] H. Xu, J. Zheng, and R. Chellappa. Bridging the domain shift by domain adaptive dictionary learning. In *BMVC*, 2015. 2

[33] Q. Yu, Y. Yang, Y.-Z. Song, T. Xiang, and T. Hospedales. Sketch-a-net that beats humans. In *BMVC*, 2015. 2