A Testbed for Cross-Dataset Analysis

Tatiana Tommasi(✉) and Tinne Tuytelaars

ESAT-PSI/VISICS - iMinds, KU Leuven, Belgium
ttommasi@east.kuleuven.be

Abstract. Despite the increasing interest towards domain adaptation and transfer learning techniques to generalize over image collections and overcome their biases, the visual community misses a large scale testbed for cross-dataset analysis. In this paper we discuss the challenges faced when aligning twelve existing image databases in a unique corpus, and we propose two cross-dataset setups that introduce new interesting research questions. Moreover, we report on a first set of experimental domain adaptation tests showing the effectiveness of iterative self-labeling for large scale problems.

Keywords: Dataset bias · Domain adaptation · Iterative self-labeling

1 Introduction

In the last two decades computer vision research has lead to the development of many efficient ways to describe and code the image content, and to the definition of several highly performing pattern recognition algorithms. In this evolution a key role was held by different image collections defined both as source of training samples and as evaluation instruments. The plethora of datasets obtained as legacy from the past, together with the modern increasing amount of freely available images from the Internet, pose new challenges and research questions. On one side there is a growing interest for *large scale data* [5,30], i.e. how to mine a huge amount of information and how to use it to tackle difficult problems that were not solvable or not even thinkable before [4]. On the other side there is the *dataset bias* problem [20,32,33]. Every finite image collection tends to be biased due to the acquisition process (used camera, lighting condition, etc.), preferences over certain types of background, post-processing elaboration (e.g. image filtering), or annotator tendencies (e.g. chosen labels). As a consequence the same object category in two datasets can appear visually different, while two different labels can be assigned to the exact same image content. Moreover, not all the datasets cover the same set of classes, thus the definition of what an object "is not" changes depending on the considered collection. The existing curated image datasets were created for a wide variety of tasks, but always with the general purpose of capturing the real visual world. Although each collection ends up covering only a limited part of it, by reorganizing the content of many collections we can define a rich knowledge repository.

L. Agapito et al. (Eds.): ECCV 2014 Workshops, Part III, LNCS 8927, pp. 18–31, 2015.
DOI: 10.1007/978-3-319-16199-0_2

Fig. 1. We show here one image example extracted from each of the 12 datasets (columns) for 7 object categories (rows): mug, bonsai, fire hydrant, car, cow, bottle, horse. The empty positions indicate that the corresponding dataset is not annotated for the considered class.

In this work we discuss the challenges faced when aligning twelve existing image datasets (see Figure 1) and we propose two data setups that can be used both as large scale testbeds for cross-dataset analysis and as a information source for efficient automatic annotation tools.

The rest of the paper is organized as follows. Section 2 gives a brief overview of related works that focused on the dataset bias problem and that proposed domain adaptation solutions. Section 3 introduces the cross-dataset collection, while section 4 reports on the results of a preliminary evaluation of domain adaptation methods over it. We conclude the paper in section 5 pointing to possible directions of future research.

2 Related Work

The existence of several data related issues in any area of automatic classification technology was first discussed by Hand in [17] and [18]. The first sign of peril in image collections was indicated in presenting the Caltech256 dataset [16] where the authors recognized the danger of learning ancillary cues of the image collection (e.g. characteristic image size) instead of intrinsic features of the object categories. However, only recently this topic has been really put under the spotlight for computer vision tasks by Torralba and Efros [33]. Their work pointed out the idiosyncrasies of existing image datasets: the evaluation of cross-dataset performance revealed

that standard detection and classification methods fail because the uniformity of training and test data is not guaranteed.

This initial analysis of the *dataset bias* problem gave rise to a series of works focusing on how to overcome the specific image collection differences and learn robust classifiers with good generalization properties. The proposed methods have been mainly tested on binary tasks (object vs rest) where the attention is focused on categories like *car* or *person* which are common among six popular datasets: SUN, Labelme, Pascal VOC, Caltech101, Imagenet, and MSRC [33]. A further group of three classes was soon added to the original set (*bird*, *chair* and *dog*) defining a total of five object categories over the first four datasets listed before [8,20]. A larger scale analysis in terms of categories was proposed in [28] by focusing on 84 classes of Imagenet and SUN, while a study on how to use weakly labeled Bing images to classify Caltech256 samples was proposed in [1]. Finally the problem of partially overlapping label sets among different datasets was considered in [32].

Together with the growing awareness about the characteristic signature of each existing image set, the related problem of *domain shift* has also emerged. Given a source and target image set with different marginal probability distributions, any learning method trained on the first will present lower performance on the second. In real life settings it is often impossible to have full control on how the test images will differ from the original training data and an adaptation procedure to remove the domain shift is necessary. An efficient (and possibly unsupervised) solution is to learn a shared representation that eliminates the original distribution mismatch. Different methods based on subspace data embedding [11,13], metric [29, 31] and vocabulary learning [27] have been presented. Recently several works have also demonstrated that deep learning architectures may produce domain invariant descriptors through highly non-linear transformation of the original features [6]. Domain adaptation algorithms have been mostly evaluated on the Office dataset [29] containing 31 office-related object categories from three domains. A subset of the Caltech256 dataset was later included defining a setting with 10 classes and four different data sources [13].

Despite their close relation, visual domain and dataset bias are not the same. Domain adaptation solutions have been used to tackle the dataset bias problem, but domain discovery approaches have shown that a single dataset may contain several domains [19] while a single domain may be shared across several datasets [15]. Moreover, the domain shift problem is generally considered under the covariate shift assumption with a fixed set of classes shared by the domains and analogous conditional distributions. On the other hand, different image datasets may contain different object classes.

Currently the literature misses a standard testbed for large scale cross-dataset analysis. We believe that widening the attention from few shared classes to the whole dataset structures can reveal much about the nature of the biases, and on the effectiveness of the proposed algorithmic solutions. Moreover it allows to extend the experience gained by years of research on each image collection to the others. Finally, the use of multiple sources has proven to be beneficial in reducing the domain shift and improve transfer learning for new tasks [26].

3 A Large Scale Cross-Dataset Testbed

In this section we describe the steps taken to define the proposed large scale cross-dataset testbed. We start with a brief description of the considered image datasets (section 3.1) and we give an overview of the merging process (section 3.2), presenting two data setups (section 3.3).

3.1 Collection Details

We focus on twelve datasets that were created and used before for object categorization.

ETH80 [23] was created to facilitate the transition from object identification (recognize a specific given object instance) to categorization (assign the correct class label to an object instance never seen before). It contains 8 categories and 10 toy objects for every category. Each object is captured against a blue background and it is represented by 41 images from viewpoints spaced equally over the upper viewing hemisphere.

Caltech101 [10] contains 101 object categories and was the first large scale collection proposed as a testbed for object recognition algorithms. Each category contain a different number of samples going from a minimum of 31 to a maximum of 800. The images have little or no clutter with the objects centered and presented in a stereotypical pose.

Caltech256 [16]. Differently from the previous case the images in this dataset were not manually aligned, thus the objects appear in several different poses. This collection contains 256 categories with a minimum of 80 and a maximum of 827 images.

Bing [1] contains images downloaded from the Internet for the same set of 256 object categories of the previous collection. Text queries give as output several noisy images which are not removed, resulting in a weakly labeled collection. The number of samples per class goes from a minimum of 197 to a maximum of 593.

Animals with Attributes (AwA) [22] presents a total of 30475 images of 50 animal categories. Each class is associated to a 85-element vector of numeric attribute values that indicate general characteristics shared between different classes. The animals appear in different pose and at different scales in the images.

a-Yahoo [9]. As the previous one, this dataset was collected to explore attribute descriptions. It contains 12 object categories with a minimum of 48 and a maximum of 366 samples per class.

MSRCORID [24]. The Microsoft Research Cambridge Object Recognition Image Database contains a set of digital photographs grouped into 22 categories spanning over objects (19 classes) and scenes (3 classes).

PascalVOC2007 [7]. The Pascal Visual Object Classes dataset contain 20 object categories and a total of 9963 images. Each image depicts objects in realistic scenes and may contain instances of more than one category. This dataset

was used as testbed for the Pascal object recognition and detection challenges in 2007.

SUN [34] contains a total of 142165 pictures[1] and it was created as a comprehensive collection of annotated images covering a large variety of environmental scenes, places and objects. Here the objects appears at different scales and positions in the images and many of the instances are partially occluded making object recognition and categorization very challenging.

Office [29]. This dataset contains images of 31 object classes over three domains: the images are either obtained from the Amazon website, or acquired with a high resolution digital camera (DSLR), or taken with a low resolution webcam. The collection contains a total of 4110 images with a minimum of 7 and a maximum of 100 samples per domain and category.

RGB-D [21] is similar in spirit to ETH80 but it was collected with a Kinect camera, thus each RGB image is associated to a depth map. It contains images of 300 objects acquired under multiple views and organized into 51 categories.

Imagenet [5]. At the moment this collection contains around 21800 object classes organized according to the Wordnet hierarchy.

3.2 Merging Challenges

There are two main challenges that must be faced when organizing and using at once all the data collections listed before. One is related to the alignment of the object classes and the other is the need for a shared feature representation.

Composing the datasets in a single corpus turned out to be quite difficult. Even if each image is labeled with an object category name, the class alignment is tricky due to the use of different words to indicate the very same object, for instance *bike* vs *bicycle* and *mobilephone* vs *cellphone*. Sometimes the different nuance of meaning of each word are not respected: *cup* and *mug* should indicate two different objects, but the images are often mixed; *people* is the plural of *person*, but images of this last class often contain more than one subject. Moreover, the choice of different ontology hierarchical levels (*dog* vs *dalmatian* vs *greyhound*, *bottle* vs *water-bottle* vs *wine-bottle*) complicates the combination. Psychological studies demonstrated that humans prefer entry-level categories when naming visual objects [25], thus when combining the datasets we chose "natural" labels that correspond to intermediate nodes in the Wordnet hierarchy. For instance, we used *bird* to associate humming bird, pigeon, ibis, flamingo, flamingo head, rooster, cormorant, ostrich and owl, while *boat* covers kayak, ketch, schooner, speed boat, canoe and ferry. In the cases in which we combine only two classes we keep both their names, e.g. *cup & mug*.

In the alignment process we came across a few peculiar cases. Figure 2 shows samples of three classes in Imagenet. The category *chess board* does not exist at

[1] Here we consider the version available in December 2013 at http://labelme.csail.mit.edu/Release3.0/Images/users/antonio/static_sun_database/ and the list of objects reported at http://groups.csail.mit.edu/vision/SUN/.

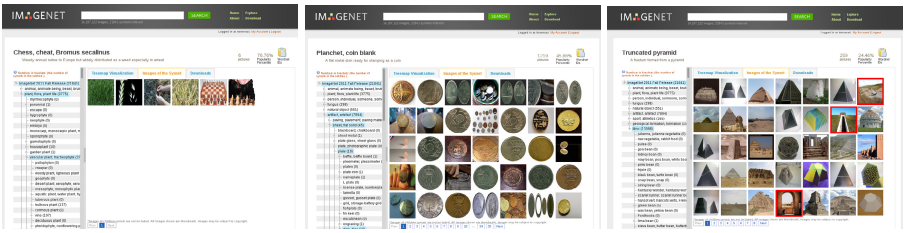

Fig. 2. Three cases of Imagenet categories. Left: some images in class *chess* are wrongly labeled. Middle: the class *planchet* or coin blank contains images that can be more easily labeled as *coin*. Right: the images highlighted with a red square in the class *truncated pyramid* do not contain a pyramid (best viewed in color and with magnification).

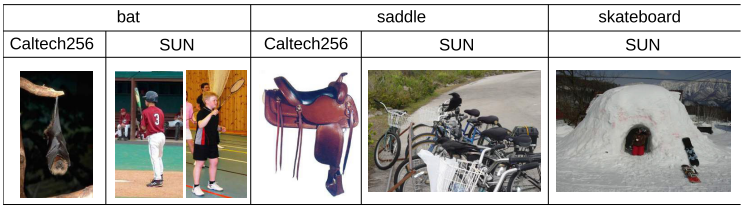

Fig. 3. Three categories with labeling issues. The class *bat* has different meanings both across datasets and within a dataset. A *saddle* can be a seat to ride a horse or a part of a bicycle. A *skateboard* and a *snowboard* may be visually similar, but they are not the same object.

the moment, but there are three classes related to the word *chess*: chess master, chessman or chess piece, chess or cheat or bromus secalinus (we use "or" here to indicate different labels associated to the same synset). This last category contains only few images but some of them are not correctly annotated. The categories *coin* and *pyramid* are still not present in Imagenet. For the first, the most closely related class is *planchet* or *coin blank*, which contains many example of what would be commonly named as a coin. For the second, the most similar *truncated pyramid* contains images of some non-truncated pyramids as well as images not containing any pyramids at all. In general, it is important to keep in mind that several of the Imagenet pictures are weakly labeled, thus they cannot be considered as much more reliable than the corresponding Bing images. Imagenet users are asked to clean and refine the data collection by indicating whether an image is a typical or wrong example.

We noticed that the word *bat* usually indicates the flying mammal except in SUN where it refers to the baseball and badminton bat. A *saddle* in Caltech256 is the supportive structure for a horse rider, while in SUN it is a bicycle seat. Tennis shoes and sneakers are two synonyms associated to the same synset in Imagenet, while they correspond to two different classes in Caltech256. In SUN, there are two objects annotated as skateboards, but they are in fact two snowboards. Some

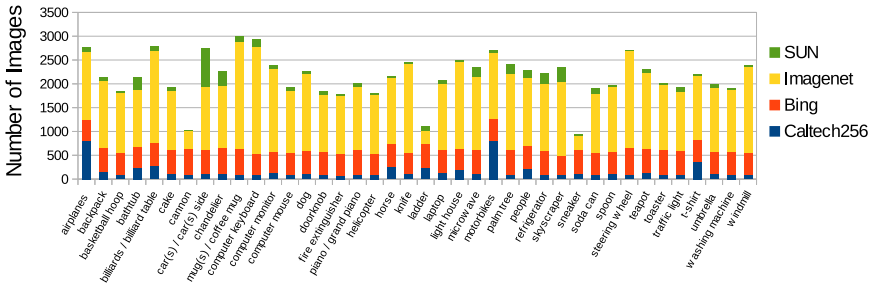

Fig. 4. Stack histogram showing the number of images per class of our cross-dataset dense setup

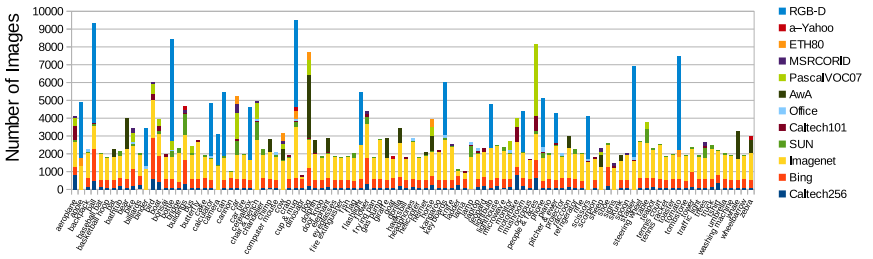

Fig. 5. Stack histogram showing the number of images per class of our cross-dataset sparse setup (best viewed in color and with magnification)

examples are shown in Figure 3. We disregarded all these ambiguous cases and we do not consider them in the final combined setups.

Although many descriptors have been extracted and evaluated separately on each image collection, the considered features usually differ across datasets. Public repositories with pre-calculated features exist for Caltech101 and Caltech256, Bing and Caltech256, and for a set of five classes out of four datasets[2]. Here we consider the group of twelve datasets listed in the previous section and extracted the same feature from all of them defining a homogeneous reference representation for cross-dataset analysis.

3.3 Data Setups and Feature Descriptor

Dense set. Among the considered datasets, the ones with the highest number of categories are Caltech256, Bing, SUN and Imagenet. In fact the last two are open

[2] Available respectively at http://files.is.tue.mpg.de/pgehler/projects/iccv09/, http://vlg.cs.dartmouth.edu/projects/domainadapt/, http://undoingbias.csail.mit.edu/

collections progressively growing in time. Overall they share 114 categories: some of the 256 object categories are missing at the moment in Imagenet but they are present in SUN (e.g. desk-globe, fire-hydrant) and vice-versa (e.g. butterfly, pram). Out of this shared group, 40 classes (see Figure 4) contain more than 20 images per dataset and we selected them to define a dense cross-dataset setup. We remark that each image in SUN is annotated with the list of objects visible in the depicted scene: we consider an image as a sample of a category if the category name is in the mentioned list.

Sparse set. A second setup is obtained by searching over all the datasets for the categories which are shared at least by four collections and that contain a minimum of 20 samples. We allow a lower number of samples only for the classes shared by more than four datasets (i.e. from the fifth dataset on the images per category may be less than 20). These conditions are satisfied by 105 object categories in Imagenet overlapping with 95 categories of Caltech256 and Bing, 89 categories of SUN, 35 categories of Caltech101, 17 categories of Office, 18 categories of RGB-D, 16 categories of AwA and PascalVOC07, 13 categories of MSRCORID, 7 categories of ETH80 and 4 categories of a-Yahoo. The histogram in Figure 5 shows the defined sparse set and the number of images per class: the category *cup & mug* is shared across nine datasets, making it the most popular one.

Representation. Dense SIFTs are among the most widely used features in several computer vision tasks, thus we decided to use this descriptor and we adopted the same extraction protocol proposed in the Imagenet development kit[3] by running their code over the twelve considered datasets. Each image is resized to have a max size length of no more than 300 pixels and SIFT descriptors are computed on 20x20 overlapping patches with a spacing of 10 pixels. Images are further downsized (to 1/2 and 1/4 of the side length) and more descriptors are computed. We publicly release both the raw descriptors and the Bag of Words (BOW) representation. We used the visual vocabulary of 1000 words provided with the mentioned kit: it was built over the images of the 1000 classes of the ILSVRC2010 challenge[4] by clustering a random subset of 10 million SIFT vectors.

4 A First Experimental Evaluation

Given the wide variability within and between the considered collections the defined setups can be used for several tasks. Some of the datasets come with extra side information (e.g. attributes, point clouds, bounding boxes) and this opens many possibilities for the evaluation of different (transfer) learning methods across the datasets. Here we kick the experimental analysis off with an initial study on domain adaptation methods.

[3] www.image-net.org/download-features
[4] http://www.image-net.org/challenges/LSVRC/2010/

Subspace Methods. Subspace domain adaptation approaches presented high performance in the unsupervised setting where the labeled source data are used to classify on unlabeled target samples. In particular the LANDMARK method proposed in [14] was indicated as a reliable technique to overcome the dataset bias [12]. This approach consists of three steps. (1) A subset of the source data is selected by choosing the samples that are distributed most similarly to the target. This process is executed by solving a quadratic programming problem and it is repeated to consider different similarity levels among the domains by changing the bandwidth σ_q of a Gaussian RBF kernel. (2) Each data subset works then as an auxiliary source to learn a domain invariant representation with the GFK algorithm [13]. Thus, for each sample $\mathbf{x} \in \mathcal{R}^D$ and each scale q we obtain a mapping $\Phi_q(\mathbf{x}) \in \mathcal{R}^d$ to a subspace with $d < D$. (3) Finally a classification model is learned on the auxiliary data by combining the different obtained representations with a multi-kernel SVM. Overall the method needs several parameters: a threshold to binarize the solution of the quadratic problem and identify the landmarks in the source, a set of σ_q values and the subspace dimensionality d.

The GFK algorithm represents each domain in a d dimensional linear subspace and embeds them onto a Grassmann manifold. The geodesic flow on the manifold parametrizes the path connecting the subspaces and it is used to define the mapping Φ to a domain invariant feature as mentioned above.

Self-labeling. Instead of subselecting the source, a different domain adaptation approach can be defined by subselecting and using the target samples while learning the source model. This technique is known as self-labeling [2,3] and starts by annotating the target with a classifier trained on the source. The target samples for which the source model presents the highest confidence are then used together with the source samples in the following training iteration.

When dealing with large scale problems self-labeling appears much more suited than the LANDMARK method. The main reason is in the high computational complexity of solving a quadratic problem over a source set with thousands of samples and repeating this operation several times over different similarity scales among the domains.

We consider here a naïve multiclass self-labeling method and we indicate it as SELF LAB in the following. A one-vs-all SVM model is trained on the source data with the C parameter chosen by cross validation. At each iteration the model is used to classify on the target data and the images assigned to every class are ranked on the basis of their output margin. Only the images with a margin higher than the average are selected and sorted by the difference between the first and the second higher margin over the classes. The top samples in the obtained list per class are then used in training with the pseudo-labels assigned to them in the previous iteration. In this way the sample selection process comes as a side-product of the classification together with a re-ranking of the SVM output margins. Moreover this approach directly exploits the multiclass nature of the domains which is generally disregarded when focusing only on how to reduce the mismatch among their marginal distributions.

Table 1. Classification rate results (%) on the Office-Caltech dataset. Here A,C,D,W stand respectively for Amazon, Caltech, Dslr, Webcam and e.g. A-C indicates the source:A, target:C pair. The results of NO ADAPT, GFK and LANDMARK are reported from [14]. The last column contains the average results per row. Best results per column in bold.

	A-C	A-D	A-W	C-A	C-D	C-W	W-A	W-C	W-D	AVG
NO ADAPT [14]	41.7	41.4	34.2	51.8	54.1	46.8	31.1	31.5	70.7	44.8
GFK [14]	42.2	42.7	40.7	44.5	43.3	44.7	31.8	30.8	75.6	44.0
LANDMARK [14]	**45.5**	**47.1**	**46.1**	**56.7**	**57.3**	49.5	**40.2**	35.4	75.2	**50.3**
SELF LAB	43.6	43.3	45.8	55.8	41.4	**53.2**	39.9	**36.1**	**82.8**	49.1

In our implementation we considered the target selection at a single scale by using a simple linear SVM, but it can also be extended to multiple scales considering non-linear kernels. For the experiments we set the number of iterations and the number of selected target samples per class respectively to 10 and 2. In this way a maximum of 20 target samples per class are used to define the training model.

A First Test on the Office-Caltech Dataset. Up to now the Office-Caltech dataset is the most widely used testbed for domain adaptation with its 4 domains and 10 shared object classes. The images of this collection were released together with SURF BOW features and subspace domain adaptation methods showed particularly high performance over it. To have a sanity check on the performance of SELF LAB we run it on this dataset following the setup used in [14].

In Table 1 we show the performance of SELF LAB, reporting the results presented in [14] as baselines. Here NO ADAPT corresponds to learning only on the source data for training. We can see that the proposed naïve self-labeling approach performs better than NO ADAPT and GFK on average, and it is only slightly worse than LANDMARK, despite being less computationally expensive. On the downside SELF LAB has only a minimal safeguard against negative transfer (that can be improved by better thresholding the SVM output margins or tuning the number of iterations), and it suffers from it in the Caltech-Dslr case (C-D), but GFK seems to have a similar behavior that affects all the cases with Caltech as source. Overall this analysis indicates the value of SELF LAB as a useful basic domain adaptation method.

A Larger Scale Evaluation. We repeat the evaluation described before on three of the datasets in the proposed dense set: Imagenet, Caltech256 and SUN. We leave out Bing, postponing the study of noisy source/target domains for future work. We consider the SIFT BOW features and 5 splits per dataset each containing respectively 5534 images for Imagenet, 2875 images for SUN and 4366 images for Caltech256 over 40 classes. Every split is then equally divided in two parts for training (source) and test (target). We use linear SVM for NO ADAPT with the best C value obtained by cross-validation on each source. The same C value is then used for SVM in combination with the GFK kernel and we tune the subspace dimensionality on the target reporting the best obtained

Table 2. Average classification rate results (%) over 5 splits for cross-dataset classification on 40 classes of three datasets: I, C, S stands respectively for Imagenet, Caltech256 and SUN. With *ss* we indicate the column containing the source-to-source results; see the text for the definition of *drop*. Best average source-to-target and drop results in bold.

	NO ADAPT					GFK				LANDMARK				SELF LAB			
	ss	I	C	S	*drop*	I	C	S	*drop*	I	C	S	*drop*	I	C	S	*drop*
I	30.9	-	22.1	13.0	43.0	-	24.8	13.4	38.2	-	24.7	13.0	39.2	-	24.0	13.5	39.2
C	48.9	18.9	-	9.8	70.6	19.1	-	10.5	69.7	17.5	-	9.9	70.6	21.4	-	11.7	66.1
S	29.5	9.4	7.4	-	71.5	9.3	8.0	-	70.5	9.1	8.7	-	71.9	11.2	9.2	-	65.4
AVG		13.5			61.7		14.2		59.5		13.6		60.6		**15.0**		**56.9**

results. A similar approach[5] is adopted to choose the subspace dimensionality for LANDMARK while the C value for the multi-kernel SVM is optimized over the source through a cross-validation between the landmark and non-landmark samples (see [14] for more details). For both GFK and LANDMARK we use the original implementation provided by the authors. A rough idea about the computational complexity of the methods can be obtained by their running time on a modern desktop computer (2.8GHz cpu, 4Gb of ram, 1core): for a single Caltech-SUN split and fixing d=100, LANDMARK needs 1695s for the source sample selection process on one scale. GFK kernel calculation and the subsequent source training and target classification run in 7s, while SELF LAB performs 10 iterations in 110s.

We show the obtained recognition rate results in Table 2. The table is divided in four parts, each for one of the considered four methods (NO ADAPT, GFK, LANDMARK and SELF LAB). Here the training and the test datasets are respectively specified in each row and column. We indicate with *ss*, *st* the source-to-source and source-to-target results. The classification performance drop among them is $drop = (ss - st) * 100/ss$. In the last row of the table we present both the average drop value for each method and the average source-to-target results. The obtained accuracy confirms the existence of the dataset bias which is particularly severe when passing from object-centric (Imagenet and Caltech) to scene images (SUN). The considered domain adaptation methods appear only minimally effective to alleviate it indicating the difficulty of the task. SELF LAB shows here the best advantage with respect to NO ADAPT.

Although a more in-depth analysis is needed, these preliminary results already give an indication of how the domain adaptation and dataset bias scenario may change when we consider large scale problems. In general a large amount of

[5] The original feature dimensionality is 1000 and for GFK we tuned the subspace dimensionality in d=[10,20,30, ...,500]. On average over all the source-target combinations the GFK performance increases with d and reaches a plateau for $d > 200$. For LANDMARK the source sample selection threshold is chosen in [0.0001, 0.0005, 0.001] and for time constraints we restricted the range for the subspace dimensionality to two values d=[100,300]. The source and target domains are compared at five scales q=[-2, -1, 0, 1, 2] with $\sigma_q = 2^q \sigma_0$ where σ_0 is equal to the median distance over all pairwise data points.

data calls for methods able to deal efficiently with them. Moreover, a high number of images per class together with high intra-class variability may reduce the mismatch among the corresponding marginal data distributions. However, the relation among the classes in two datasets can still be different. This pushes towards discriminative approaches able to deal with differences in the conditional distributions of the data.

5 Conclusions

In this paper we discussed the challenges faced when aligning twelve existing image datasets and we proposed two data setups that can be used as testbed for cross-dataset analysis. We extracted dense SIFT descriptors from the images and we created a useful feature repository for future research. We consider this as the first step of a wider project (official webpage: https://sites.google.com/site/crossdataset/) that will continue by both extracting new features and running several cross-dataset classification tests. The preliminary experimental analysis presented here has already indicated the value of self-labeling as a possible baseline for this task. Besides offering new challenges to domain adaptation methods, the proposed setups introduce also new research questions on how to deal with different forms of weak supervision or whether it is possible to transfer attributes and depth information across datasets. Moreover, it may be interesting to understand if the dataset alignment could be done automatically.

To conclude, we believe that exploring the common aspects and the specific characteristics of each collection will not only reveal more about the dataset bias problem, but it will also allow to improve the generalization capabilities of learning methods and mitigate the need for manual image annotation.

Acknowledgments. The authors acknowledge the support of the FP7 EC project AXES and FP7 ERC Starting Grant 240530 COGNIMUND.

References

1. Bergamo, A., Torresani, L.: Exploiting weakly-labeled web images to improve object classification: a domain adaptation approach. In: NIPS (2010)
2. Bruzzone, L., Marconcini, M.: Domain adaptation problems: A dasvm classification technique and a circular validation strategy. IEEE Trans. PAMI **32**(5), 770–787 (2010)
3. Chen, M., Weinberger, K.Q., Blitzer, J.: Co-training for domain adaptation. In: NIPS (2011)
4. Deng, J., Berg, A.C., Li, K., Fei-Fei, L.: What Does Classifying More Than 10,000 Image Categories Tell Us? In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 71–84. Springer, Heidelberg (2010)
5. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: ImageNet: A Large-Scale Hierarchical Image Database. In: CVPR (2009)
6. Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T.: Decaf: A deep convolutional activation feature for generic visual recognition. arXiv preprint arXiv:1310.1531 (2013)

7. Everingham, M., Gool, L.V., Williams, C.K., Winn, J., Zisserman, A.: The Pascal Visual Object Classes (VOC) Challenge. IJCV 88(2) (2010)
8. Fang, C., Xu, Y., Rockmore, D.N.: Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias. In: ICCV (2013)
9. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: CVPR (2009)
10. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Comput. Vis. Image Underst. **106**(1), 59–70 (2007)
11. Fernando, B., Habrard, A., Sebban, M., Tuytelaars, T.: Unsupervised visual domain adaptation using subspace alignment. In: ICCV (2013)
12. Gong, B., Sha, F., Grauman, K.: Overcoming dataset bias: An unsupervised domain adaptation approach. In: NIPS Workshop on Large Scale Visual Recognition and Retrieval (2012)
13. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: CVPR (2012)
14. Gong, B., Grauman, K., Sha, F.: Connecting the dots with landmarks: Discriminatively learning domain-invariant features for unsupervised domain adaptation. In: ICML (2013)
15. Gong, B., Grauman, K., Sha, F.: Reshaping visual datasets for domain adaptation. In: NIPS (2013)
16. Griffin, G., Holub, A., Perona, P.: Caltech 256 object category dataset. Tech. Rep. UCB/CSD-04-1366, California Institue of Technology (2007)
17. Hand, D.J.: Classifier Technology and the Illusion of Progress. Stat. Sci. **21**, 1–15 (2006)
18. Hand, D.J.: Academic obsessions and classification realities: ignoring practicalities in supervised classification. In: Classification, Clustering, and Data Mining Applications, pp. 209–232 (2004)
19. Hoffman, J., Kulis, B., Darrell, T., Saenko, K.: Discovering Latent Domains for Multisource Domain Adaptation. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 702–715. Springer, Heidelberg (2012)
20. Khosla, A., Zhou, T., Malisiewicz, T., Efros, A.A., Torralba, A.: Undoing the Damage of Dataset Bias. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part I. LNCS, vol. 7572, pp. 158–171. Springer, Heidelberg (2012)
21. Lai, K., Bo, L., Ren, X., Fox, D.: A large-scale hierarchical multi-view rgb-d object dataset. In: ICRA (2011)
22. Lampert, C.H., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between class attribute transfer. In: CVPR (2009)
23. Leibe, B., Schiele, B.: Analyzing appearance and contour based methods for object categorization. In: CVPR (2003)
24. Microsoft: Microsoft Research Cambridge Object Recognition Image Database. http://research.microsoft.com/en-us/downloads/b94de342-60dc-45d0-830b-9f6eff91b301/default.aspx (2005)
25. Ordonez, V., Deng, J., Choi, Y., Berg, A.C., Berg, T.L.: From large scale image categorization to entry-level categories. In: ICCV (2013)
26. Patricia, N., Caputo, B.: Learning to learn, from transfer learning to domain adaptation: A unifying perspective. In: CVPR (2014)

27. Qiu, Q., Patel, V.M., Turaga, P., Chellappa, R.: Domain Adaptive Dictionary Learning. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part IV. LNCS, vol. 7575, pp. 631–645. Springer, Heidelberg (2012)
28. Rodner, E., Hoffman, J., Donahue, J., Darrell, T., Saenko, K.: Towards adapting imagenet to reality: Scalable domain adaptation with implicit low-rank transformations. CoRR abs/1308.4200 (2013)
29. Saenko, K., Kulis, B., Fritz, M., Darrell, T.: Adapting Visual Category Models to New Domains. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 213–226. Springer, Heidelberg (2010)
30. Sanchez, J., Perronnin, F.: High-dimensional signature compression for large-scale image classification. In: CVPR (2011)
31. Tommasi, T., Caputo, B.: Frustratingly easy nbnn domain adaptation. In: ICCV (2013)
32. Tommasi, T., Quadrianto, N., Caputo, B., Lampert, C.H.: Beyond Dataset Bias: Multi-task Unaligned Shared Knowledge Transfer. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 1–15. Springer, Heidelberg (2013)
33. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR (2011)
34. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database: Large-scale scene recognition from abbey to zoo. In: CVPR (2010)