Extracting Structures in Image Collections for Object Recognition*

Sandra Ebert[1,2,**], Diane Larlus[1,**], and Bernt Schiele[1,2]

[1] Department of Computer Science, TU Darmstadt, Germany
[2] MPI Informatics, Saarbrucken, Germany

Abstract. Many computer vision methods rely on annotated image sets without taking advantage of the increasing number of unlabeled images available. This paper explores an alternative approach involving unsupervised structure discovery and semi-supervised learning (SSL) in image collections. Focusing on object classes, the first part of the paper contributes with an extensive evaluation of state-of-the-art image representations. Thus, it underlines the decisive influence of the local neighborhood structure and its direct consequences on SSL results and the importance of developing powerful object representations. In a second part, we propose and explore promising directions to improve results by looking at the local topology between images and feature combination strategies.

Keywords: object recognition, semi-supervised learning.

1 Introduction

Supervised learning is the de facto standard for many computer vision tasks such as object recognition or scene categorization. Powerful classifiers can obtain impressive results but require a sufficient amount of annotated training data. However, supervised methods have important limitations: Annotation is expensive, prone to error, often biased, and does not scale. Obtaining the required training data representing all relevant aspects of a given category is difficult but key to success for supervised methods. Facing these limitations we argue that the computer vision community should move beyond supervised methods and more seriously tap into the vast collections of images available today.

In particular, we look at the *local structure* of the data (links between images here) in an *unsupervised way*. For larger datasets, this local neighborhood becomes more reliable: Two semantically similar images (belonging to the same class) have a higher probability to be also similar in image representation space for increasing database sizes (see Fig. 1). *Semi-supervised learning* (SSL), the second direction explored here, uses such local neighborhood relations and few labeled images to predict the label of new images. The local structure has in both cases a strong influence on the overall performance of such approaches.

* This work was supported by a Google Research Award.
** The first two authors contributed equally. Names are ordered alphabetically.

K. Daniilidis, P. Maragos, N. Paragios (Eds.): ECCV 2010, Part I, LNCS 6311, pp. 720–733, 2010.

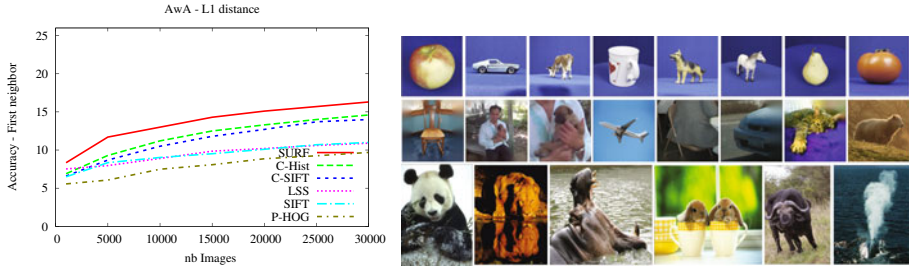

Fig. 1. (Left) For each image, we look at the most similar image (L1 distance NN) to see how often these couples belong to the same class. This increases for larger sets. (Right) Top: ETH, middle: our Cropped PASCAL, bottom: AWA datasets.

This paper is organized in two parts. First, we contribute a study of different representations and SSL algorithms on three image collections of increasing size and difficulty for image categorization (Sec. 3). We show that the results depend on the neighborhood structure induced by object representations and on the graph structure parameters rather than on the particular SSL algorithm employed. We also show that results obtained on the local neighborhood directly transfer to SSL results.

Motivated by these conclusions, the second part of the paper presents different ways of improving the connections between images in the local neighborhood structure. Among the considered strategies: the topology of the dataset is used to refine the existing connections (Sec. 4), and different features are combined (Sec. 5). Results show improvements for both the structure and the SSL predictions on all datasets.

Related Work. The use of large image collections is obviously not a novel idea. [1] directly discovers image clusters, while other approaches aim to globally partition the database in image sets sharing more general concepts [2]. Multi instance learning methods deal with weak or incomplete annotations [3]. Some methods use the web as an external source of information to get many but noisy annotations [4]. Active learning methods aim to identify missing annotations [5]. Finally, attempts are made to make the annotation process more appealing [6]. None of this prior work however systematically analyzes the suitability of today's image and object representations for unsupervised local structure extraction.

Semi-supervised learning (SSL) has been applied to several computer vision problems. Partial labeling of pixels is used as an input for segmentation [7]. Image level annotations are used to find object parts [8]. But only a couple of methods apply SSL to predict labels at the image level from a few annotated images. Of particular interest are [9] using random forests and [10] using boosting, both in an SSL framework. Closer to our work, [11] focuses on graph based propagation algorithms and proposes efficient approximations to scale SSL methods to large datasets. In machine learning, SSL methods have been used with success

for many tasks (e.g. digit recognition, text classification, or speech recognition, see [12] for a survey). Among SSL, graph-based methods play an important role as they concentrate on the local structure of data [13,14]. Most approaches however focus on SSL algorithms rather than on the underlying structure. In this work, we analyze the local neighborhood in detail to improve the performance of SSL graph-based algorithms for image data.

There are very few studies that compare SSL methods on images. [12] contains a single small image set and [15] considers digits and faces only. In both cases, representations are different from commonly used image descriptors for recognition. Therefore, this paper focuses on the important problem of how well standard representations are suited for unsupervised structure discovery as well as SSL and how the structure can be improved such that also SSL can benefit.

2 Datasets and Image Representations

We consider three datasets with increasing number of object classes, number of images, and difficulty. Some of the images are shown in Fig. 1.

ETH-80 (ETH) [16] contains 3,280 images divided in 8 object classes and 10 instances per class. Each instance is photographed from 41 viewpoints in front of a uniform background. This controlled dataset ensures that a strong local structure exists between images making it a perfect toy dataset for our task.

Cropped PASCAL (C-PASCAL) is based on the PASCAL VOC challenge 2008 training set [17]. Bounding box (BB) annotations are used to extract the objects. Consequently semantic connections between images and SSL predictions on this new dataset can be evaluated in our multi-class protocol. To discard information contained in the aspect ratio of the BB, squared regions (rescaled to 102x102 pixels) are extracted using the larger side of the BB and objects smaller than 50 pixels are discarded. To avoid that a class dominates the evaluation, we subsampled the largest class 'people' from 40% to 16% (the 2nd largest class being 'chair' with $\sim$11%). The set contains 6,175 images of aligned objects from 20 classes but with varying object poses, challenging appearances, and backgrounds.

Animal with Attributes (AWA) [18] is a large and realistic dataset with 30,475 images and 50 classes, without alignment. Objects are located anywhere in the image, in difficult conditions and poses, which complicates the task of finding images containing similar object classes. While being the most challenging dataset in this evaluation, it is the kind of data we are eventually aiming for.

Representations. This paper uses a large spectrum of representations employed by state-of-the art recognition methods [17]. For the first two datasets we consider 7 complementary descriptors: 3 global descriptors (HOG [19], Gist [20], pyramid bag-of-features (P-BoF) [21]), 3 bag-of-features representations (BoF) with different detectors and descriptors, and a texture descriptor (TPLBP [22]). Our HOG implementation uses 9-bins histograms of gradient orientations, locally normalized over contrast, extracted using a dense grid of non-overlapping cells (8x8 pixels). For the Gist scene descriptor we use the code of [20]. P-BoF

features are computed with the implementation of [21]. It extracts patches on 4 different levels and a visual vocabulary of 200 words. Concurrently, we extract bag-of-features representations. We combine Harris (Har-BoF) or Hessian-Affine (Hess-BoF) detectors [23], with SIFT [24] and build visual vocabularies of 10,000 words. We also use C-SIFT based on the code of [25] (Color-SIFT descriptors for Harris points, 2,000 words vocabulary). Finally, the local texture descriptor called Three Patch Local Binary Pattern (TPLBP) [22] considers 3 neighboring patches of size 3×3 arranged in a circle a single bit value for each pixel. For AWA, we use 7 descriptors: the 6 publicly available features [18] (color histograms (C-Hist), Local-Self-Similarity (LSS), Pyramid HOG (P-HOG), bag-of-features representations involving SIFT, color-SIFT (C-SIFT), and SURF descriptors) and the Gist descriptor that we computed additionally.

3 Local Structure and SSL Study

As stated before, this paper looks at two related tasks: *local structure extraction* and the use of this structure for *semi-supervised learning* (SSL). We focus on the question whether today's object class representations are suitable for local structure discovery and how well these observations transfer to SSL.

The following first analyzes neighborhood structures and then compares four different graph-based algorithms for SSL.

Local structure discovery. For all three datasets, we analyze neighborhood structures of different object representations, for the L1 and L2 distance measures[1]. We focus on k-nearest neighbors (k-NN) structures which have better connectivity and lead to more intuitive structures than e.g., ϵ-neighborhood graphs [26]. These properties are also important for SSL algorithms.

Experiments. To evaluate the quality of the k-NN structure for an image, we calculate the percentage of neighbors belonging to the same class as this image. Averaging this percentage over all images results in the overall k-NN structure accuracy. Intuitively, this evaluates how often the k-NN structure connects images from the same class, and how much semantic information it contains.

The left side of Tab. 1 shows L1 and L2 performances for the nearest (1-NN) and the 10 nearest neighbors (10-NN), for all three datasets[2]. First, we see that L1 constantly outperforms L2 for all representations and all datasets.

Also, we observe that results significantly differ between the different representations. Global descriptors like P-BoF or Gist work well for ETH and C-PASCAL as objects are mostly aligned in those databases. Local descriptors are better suited for the more challenging AWA dataset.

Finally, the 1-NN and 10-NN exhibit different behaviors. Some features are more robust for larger numbers of neighbors. For instance for C-PASCAL, Hess-BoF is the third best descriptor when looking at 1-NN structures, with 31.3%,

[1] The χ^2 measure was considered but not reported as it gave similar results as L1.

[2] In all tables for both NN and SSL, best representation per configuration: gray cell (max per column); best configuration: bold numbers (max per line); overall best: red.

Table 1. Quality of the nearest neighbor and the 10 nearest neighbors on the left part, transductive propagation results on the right part, for L1, L2 (Sec. 3) and L2-Context (Sec. 4.1). Only the 3 best descriptors are shown for ETH.

		NN quality						SSL results					
		L1		L2		L2-ctxt		L1		L2		L2-ctxt	
	Features	k=1	k=10	k=1	k=10	k=1	k=10	acc	var	acc	var	acc	var
ETH	C-SIFT	**96.6**	89.0	80.5	63.1	92.8	82.7	**89.0**	0.6	60.9	2.0	83.7	1.4
	Gist	**93.6**	**85.4**	92.9	83.5	**93.6**	84.8	83.1	1.4	82.5	1.0	**84.5**	0.8
	HOG	**96.9**	**88.6**	95.5	86.2	**96.9**	**88.6**	84.5	1.8	83.3	1.7	**86.8**	1.3
C-PASCAL	C-SIFT	**32.6**	**19.6**	24.2	13.9	30.1	17.8	**24.0**	0.4	16.6	2.2	20.5	0.4
	Gist	30.8	24.3	29.5	23.5	**31.8**	**24.9**	**28.4**	0.3	27.5	0.4	28.1	0.8
	HOG	27.3	21.4	22.0	17.3	**34.6**	**26.8**	19.2	2.3	13.9	2.4	**28.8**	1.6
	Har-BoF	**28.7**	**16.2**	17.8	10.4	25.1	13.6	**20.1**	0.5	13.1	2.7	15.7	0.5
	Hess-BoF	**31.3**	**17.9**	20.1	11.2	26.7	15.2	**21.6**	0.7	15.3	2.3	16.5	1.1
	P-BoF	**28.5**	**22.3**	24.1	17.7	24.1	17.7	**28.4**	0.9	20.2	1.2	20.6	0.9
	TPLBP	**33.5**	**26.2**	26.9	20.4	26.9	20.4	**29.5**	0.9	20.4	2.0	20.5	1.9
AWA	C-Hist	**14.5**	**9.2**	9.8	6.6	12.2	8.3	8.4	0.2	5.7	0.1	**8.5**	0.2
	C-SIFT	14.2	9.2	12.2	8.0	**15.4**	**10.4**	8.0	0.2	7.0	0.1	**10.3**	0.2
	Gist	12.1	8.1	12.0	8.1	**15.0**	**10.3**	7.2	0.2	7.4	0.2	**10.9**	0.1
	LSS	10.9	7.8	8.3	6.3	**11.7**	**8.1**	6.9	0.1	5.5	0.3	**8.2**	0.3
	PHOG	**9.7**	6.7	7.7	5.6	8.9	**7.0**	6.3	0.2	5.4	0.1	**7.0**	0.1
	SIFT	11.0	8.1	10.4	7.6	**12.4**	**8.8**	7.7	0.2	7.3	0.3	**9.2**	0.2
	SURF	**16.4**	10.6	11.7	8.0	14.3	**10.7**	9.0	0.1	6.8	0.3	**10.4**	0.2

but loses almost half of the performance when considering 10-NN (17.9%). On the contrary, P-BoF gives poor results for 1-NN but is more robust for 10-NN (22.3%). When considering SSL results, we will refer mainly to the 10-NN structures, as graphs are using k-NN structures with large enough values of k.

Semi-supervised learning. We use the previously studied k-NN structure and few labels in a graph and analyze several SSL methods for the object recognition problem. These methods build a graph (X, Y) where the nodes $X = \{X_l, X_u\}$ represent images and $Y = \{Y_l, Y_u\}$ are the labels. (X_l, Y_l) are labeled images and (X_u, Y_u) are unlabeled images. A graph is represented by an adjacency matrix W built from the k-NN structure. The degree of each node is $d_{ii} \leftarrow \sum_j w_{ij}$ and defines the diagonal matrix D. Here, we evaluate *non-symmetric* (directed) graphs. We do not evaluate fully connected graphs due to their computational complexity and memory requirements. We also considered weighted graphs but found that performance did not improve significantly.

Graph-based methods distribute labels from labeled to unlabeled nodes. In our experiments, we compare four methods covering a broad range of possible strategies. These methods are designed for binary problems, and expandable to multi-class problems with n classes, by splitting them into n one-versus-all binary problems, that share the same graph structure. All algorithms follow the same pattern. First, labels are initialized, with Y_l taking values in $\{1, -1\}$ and elements of Y_u set to 0 resulting in $\hat{Y}^{(0)}$. Then labels are updated iteratively $\hat{Y}^{(t+1)} \leftarrow L\hat{Y}^{(t)}$ for a certain number of iterations[3]. This part differs for each method, and is briefly described below.

[3] Typically a small number of iterations is used to avoid over-fitting.

Gaussian Fields Harmonic Functions (GFHF) [14] uses a transition probability matrix $L = D^{-1}W$ to propagate the labels. Original labels cannot change.
Quadratic Criterion (QC) [27] is a variant of the previous method allowing the original labels to change, which can help for ambiguous representations. It also introduces a regularization term for better numerical stability.
Local Global Consistency (LGC) [13] uses a normalized graph Laplacian $L = D^{-\frac{1}{2}}WD^{-\frac{1}{2}}$ instead of transition probabilities. The initial labels are also allowed to change, but in a regularized way. A parameter α (set to 0.5) regularizes the modifications to limit overwritten labels and weights how much newly predicted labels are trusted compared to original ones, $\hat{Y}^{(t+1)} \leftarrow \alpha L\hat{Y}^{(t)} + (1-\alpha)\hat{Y}^{(0)}$.
Discrete Regularization (DR) [28] incorporates local graph properties by looking at the degree of two neighboring nodes. An additional cost function reduces the influence of nodes with many connections.

Experiments. We apply these algorithms to all datasets and focus on the following aspects: the differences between the 4 SSL algorithms, between the different representations, and the influence of the local structure on SSL results.

We evaluate transductive results (i.e. prediction for the remaining labels) with 10% labeled data, for all datasets on their different representations for L1 and L2. All experiments randomly select 5 sets of labeled data, and produce mean and variance of the overall multi-class accuracy on unlabeled data only (transductive results). For comparison, we also use the supervised k-NN classifier, based on the same representations and measures. Some representative results[2], which illustrate our main findings, are summarized in Fig. 2 and in Tab. 1 (right).

i) *Graph structure and different algorithms.* The first experiment varies the number of neighbors k in the graph, for the different algorithms. Fig. 2 (left) shows the obtained performances for TPLBP, which performed best for C-PASCAL in the case of local structures. As we can see, the number of neighbors k is a crucial parameter and an optimal value exists. This value appeared to be dataset- and SSL algorithm dependent. A minimum number is required to perform reasonably. Too small k values result in a graph with disconnected components, where no information is propagated and some images are not classified.

Once the correct parameters for the graphs are chosen, there are surprisingly small differences between SSL methods. For instance, L1 numbers vary between 29.5% for LGC and 24.6% for QC. This emphasizes our claim that the structure is more important than the algorithms. LGC is more stable across experiments and the QC method tends to achieve lower results. This algorithm allows to change the original labels but has no regularization parameter like LGC, leading to many changes in the original labels, and accuracy drops for large k values. Finally, all SSL results outperform the best k-NN result (k=80) of 23.5%, showing the benefits of the unlabeled data in the classification process.

In the remainder, we use LGC [13], as it exhibited stable results across multiple settings, and its best settings determined from our parameter study.

ii) *Image representations.* As before in the NN study, we notice large differences between image representations (Fig. 2 (right) for C-PASCAL and Tab. 1 for ETH

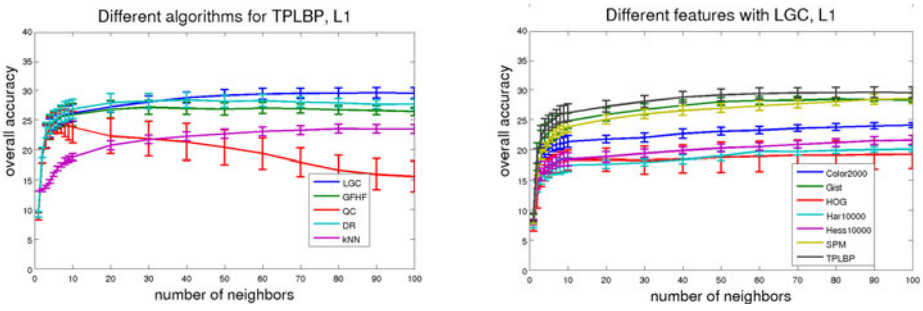

Fig. 2. For C-PASCAL. Mean/Variance of overall accuracy on unlabeled data only for LGC, GFHF, QC, and DR and k-NN (left). Different features with LGC (right).

and AWA). For C-PASCAL[4], accuracies vary from 19.2% for HOG to 29.5% for TPLBP and for AWA from 6.3% for P-HOG to 9% for SURF, with L1. Second, we observe the existing gap between the representations and the task. Our toy dataset ETH exhibits good values, meaning that for datasets with an obvious underlying structure (same objects and views), it is accurately extracted and used by the propagation algorithm. We were able to obtain satisfying results even with minimal supervision. For the more challenging C-PASCAL and AWA datasets, numbers are more disappointing. We can conclude that today's image representations are still not rich enough for building good semantic structures.

iii) *Transfer from neighborhood structure.* Tab. 1 shows that the results[2] of the 10-NN structures (left side) transfers directly to the SSL performances (right side) for each dataset, including the observed semantic gap. 10-NN performance is more consistent with the SSL algorithms as the latter need a minimum number of connections to propagate the labels. Note that 1-NN structure quality is always higher than SSL results because it gives only an intuition on the probability for an image to transfer the correct label to its first neighbor. We could reach this number for about 50% of the images labeled.

Summary. In this section, we studied the local neighborhood structure and its influence on different SSL algorithms. We observed that the parameters that determine the local neighborhood structure (image representations, value of k, etc.) result in larger differences in performance than the particular choice of the SSL algorithm. ETH presents high quality neighbors and the semantic structure of the dataset is captured accurately. For more realistic datasets, like C-PASCAL and AWA, the quality of neighbors is disappointing and underline the existing gap between considered categories' appearance and today's computer vision representation. This limitation also transfers to SSL results.

[4] Note that the non-balanced C-PASCAL (dominated by the well recognized person class) shares similar observations with our C-PASCAL, across all experiments, with higher overall numbers. For instance here, L1 varies from 27.4% for HOG and 47.1% for TPLBP.

4 Improving the Local Structure between Images

The previous section showed that local neighborhood structures capture some semantic information between images, but still a gap exists between the connections in a structure and the object categories, leaving ample room for improvement. Also, we showed that this structure has a stronger influence on the results than the SSL method itself. Therefore, this and the following section consider different directions to improve the quality of the connections between images. We want to improve the local structure without any learning involved, trying again to move away from supervised methods. Our goal is to build an improved *unsupervised local neighborhood structure*, which consequently can be generic, and does not depend on the considered SSL problem.

In the following we explore to which extent the neighborhood structure can be improved without labels, using only topological information of the dataset itself and look at its influence on (i) the local structure itself, and (ii) the SSL results. Sec. 4.1 considers context measures as an improvement over standard measures and Sec. 4.2 shows the benefit of symmetric relations between neighbors.

4.1 Context Measures

We consider the contextual measure, proposed in [29] for the image retrieval task. This context measure is applied for L2[5] to our problem.

Principle. When trying to decide if two images are close, the answer is often given for a given context. We do not only look at the images themselves, but also at the surrounding images. This is the intuition behind the contextual measure [29], that computes the distance from a first descriptor p to another descriptor q in the context of u using: $L2_{ctxt}(q,p|u) = argmin_{0\leq\omega\leq1}\{||q-(\omega p+(1-\omega)u)||_2\}$. The context vector u is obtained by computing the mean vector of the l nearest neighbors (l=100 in our experiments) of p in the collection.

Experiments. Tab. 1 summarizes the results[2] obtained for the L2-context measure, in comparison with the L1 and L2 measures considered in our previous study. From this table we can make the following observations. First, context measure yields a consistent improvement to the L2 measure. For 1-NN, this improvement represents about 9% on average for ETH, almost 5% on average for C-PASCAL and 2.5% for AWA. The same applies for the SSL results: we note 11% improvement for ETH, about 3% for C-PASCAL and for AWA, on average. Sparse vectors (e.g. Hess-BoF or Har-BoF) benefit the most. Again we observe (cf. Tab. 1) the consistency between the NN quality and the corresponding SSL results, already underlined in the previous section's study.

Interestingly, L2-ctxt brings L2 to the level of L1 and sometimes outperforms it. Context measures are a promising direction and one could expect further improvement from the context version of L1. As no closed-form solution exists for L1, this new measure will be difficult to scale to very large datasets. Therefore, we consider a different strategy which scales more easily in the following.

[5] A closed-form solution is available for L2, making computations faster.

Table 2. Quality of the nearest and the 10 nearest neighbors, chosen with distance- or a rank-based strategy, for AWA

		1-NN						10-NN					
		Dist			Rank			Dist			Rank		
	Features	L1	L2	L2-cxt	L1	L2	L2-cxt	L1	L2	L2-cxt	L1	L2	L2-cxt
	C-Hist	14.5	9.8	12.1	**16.8**	12.1	12.9	9.2	6.6	8.3	10.6	7.9	8.5
	C-SIFT	14.1	12.1	15.3	**19.1**	14.8	16.6	9.1	8.0	10.4	11.6	9.4	10.6
AWA	Gist	12.1	11.9	15.0	14.8	14.4	**15.2**	8.0	8.0	10.2	9.7	9.7	10.2
	LSS	10.8	8.2	11.7	**14.6**	10.4	12.3	7.8	6.2	8.0	9.5	7.1	8.1
NN quality	PHOG	9.7	7.6	8.8	**12.2**	9.3	9.8	6.7	5.6	6.9	7.9	6.2	7.0
	SIFT	11.0	10.3	12.3	**12.5**	11.5	**12.5**	8.0	7.6	8.8	8.9	8.4	8.9
	SURF	16.3	11.7	14.3	**23.0**	14.2	15.8	10.5	8.0	10.7	14.2	8.9	10.8

Table 3. Left: non-symmetric graphs on rank-based structures. Middle: symmetric graphs on distance-based structures. Right: symmetric graph and rank based structures. The improvement obtained in comparison to Tab. 1 is shown in the gain column.

	ETH - SSL results																	
	Rank, non sym						Dist, sym						Rank, sym					
	L1		L2		L2-cxt		L1		L2		L2-cxt		L1		L2		L2-cxt	
Feat.	acc	gain	acc	gain	acc	gain	acc	gain	acc	gain	acc	gain	acc	gain	acc	gain	acc	gain
CSIFT	**91.5**	+2.5	79.5	+18.6	84.6	+0.9	90.9	+1.8	74.1	+13.2	84.5	+0.8	91.3	+2.3	82.4	+21.5	84.7	+1.0
Gist	**84.9**	+1.8	83.2	+0.7	84.6	+0.1	83.8	+0.7	83.0	+0.5	84.1	-0.3	84.5	+1.5	83.2	+0.7	84.4	-0.1
HOG	87.4	+2.9	86.8	+3.5	87.6	+0.8	87.4	+2.9	**88.1**	+4.8	86.1	-0.7	87.3	+2.9	87.8	+4.4	87.1	+0.3
	C-PASCAL - SSL results																	
CSIFT	**24.8**	+0.8	18.8	+2.3	20.4	0.0	24.1	0.0	19.6	+3.1	20.7	+0.3	24.3	+0.2	20.8	+4.3	20.8	+0.3
Gist	30.8	+2.4	30.3	+2.7	29.1	+1.0	30.4	+2.0	30.0	+2.5	29.0	+1.0	**32.0**	+3.6	31.4	+3.9	29.5	+1.4
HOG	29.6	+10.4	24.4	+10.4	30.3	+1.5	29.5	+10.3	26.8	+12.9	30.4	+1.6	**32.2**	+13.0	29.5	+15.6	31.0	+2.1
Har	**20.7**	+0.6	14.9	+1.8	15.4	-0.3	20.3	+0.2	16.5	+3.5	16.1	+0.4	20.5	+0.4	16.7	+3.7	16.0	+0.3
Hess	**23.2**	+1.7	16.2	+0.9	16.6	+0.1	22.4	+0.9	17.4	+2.2	17.4	+0.9	23.1	+1.5	17.8	+2.5	17.4	+0.9
P-BoF	**29.9**	+1.4	23.9	+3.7	23.5	+2.8	29.4	+1.0	23.8	+3.6	22.5	+1.9	**29.9**	+1.4	25.3	+5.1	23.9	+3.3
TPLBP	32.7	+3.2	29.4	+9.0	28.6	+8.2	32.0	+2.5	28.1	+7.7	26.7	+6.2	**33.8**	+4.3	30.9	+10.5	29.5	+9.0
	AWA - SSL results																	
C-Hist	11.0	+2.6	7.6	+1.9	8.9	+0.4	10.8	+2.4	8.6	+2.9	9.1	+0.6	**11.2**	+2.8	8.7	+2.9	8.9	+0.5
CSIFT	11.0	+3.0	8.9	+1.9	10.8	+0.5	12.8	+4.7	11.5	+4.5	12.0	+1.7	**13.0**	+5.0	11.3	+4.3	11.8	+1.5
Gist	10.1	+2.7	10.1	+2.7	11.2	+0.3	10.8	+3.4	11.0	+3.6	**11.5**	+0.6	11.1	+3.7	11.1	+3.7	11.2	+0.4
LSS	9.4	+2.6	6.9	+1.4	8.5	+0.3	10.9	+4.0	9.1	+3.6	9.4	+1.2	**11.1**	+4.3	9.0	+3.5	9.4	+1.2
PHOG	7.5	+1.2	6.1	+0.7	7.2	+0.2	**9.4**	+3.2	8.0	+2.5	8.1	+1.0	**9.4**	+3.1	7.9	+2.5	8.1	+1.1
SIFT	9.5	+1.8	8.9	+1.6	9.8	+0.6	**10.4**	+2.6	9.6	+2.3	9.9	+0.7	10.0	+2.3	9.4	+2.1	9.9	+0.7
SURF	13.7	+4.7	8.0	+1.2	10.6	+0.3	16.3	+7.2	12.8	+6.1	13.4	+3.0	**16.7**	+7.7	12.9	+6.1	13.4	+3.0

4.2 Ranking and Symmetry

Here we explicitly look at the distribution of neighbors and try to build a more intuitive and more evenly distributed structure. In particular, we would like to emphasize the symmetric relations between images when building the local neighborhood structure. First, we propose a new neighbor selection procedure to emphasize symmetric relations using the "rank as neighbor". Second, we consider symmetry within the SSL-algorithm during graph propagation.

Improving the structure using ranking. In Sec. 3, for a particular distance measure and representation, we extracted *distance-based neighbors*, i.e. we look for the k images with the smallest distances to a given image, and use these images to build our local neighborhood structure.

We propose a new way of selecting neighbors, so called *rank-based neighbors*, that refines the notion of distances by emphasizing symmetric relations between images. Intuitively, we connect two images which both have the other image as

one of their nearest neighbors. More formally, we choose rank neighbors of image i as follows. We compute a first set of (distance-based) neighbors for i, and keep the one having the smallest score according to $sc(d_i, d_j) = \tau_j(d_i) + \tau_i(d_j)$, where d_i, d_j are the descriptors of image i and image candidate j, and $\tau_i(d_j)$ encodes the NN rank of descriptor d_j as (distance-based) neighbor of image i. In practice, we consider only the l nearest neighbors of image i as candidates, and $\tau_j(d_i)$ is replaced by $\tau'_j(d_i) = \min(\tau_j(i), l)$. l is used to narrow the search, and only needs to be large enough (here l=800). Note that even though the function $sc(d_j, d_j)$ is symmetric, rank-based neighborhood is not a symmetric relation.

Improving the graph by using symmetric relations. Sec. 3 considered *non-symmetric* (directed) graphs. Here we also look at *symmetric* (undirected) graphs. They consider incoming as well as outgoing links for propagation. There is a similar intuition behind *symmetric* graphs and rank-based NN as they both enforce more symmetric interaction between images. In the case of rank-based NN, a new structure is proposed which is potentially more effective, while for symmetric graphs, the influence of images that are too often selected as a neighbor is reduced within the existing structure.

Experiments. The study is divided in two parts. First, we look at the gain obtained by the structure between images using ranking, and then we study the improvement brought by both the ranking and the symmetry for the SSL results.

i) *Rank-based local structure.* In terms of NN quality, the rank strategy brings a consistent improvement. Over the different image descriptors, when looking at the first neighbor (1-NN quality), ETH gains 1.6% in average for L1, 9.5% for L2 and 1.9% for L2-ctxt by using the rank-based strategy. For C-PASCAL, we get 2.4% improvement for L1, 5.5% for L2 and 1.9% for L2-ctxt in average. For AWA (the results are shown in Tab. 2), the 1-NN quality for SURF is improved from 16.3% to 23% and the 10-NN quality goes from 10.5% to 14.5%. As a remark, few images were very often chosen as neighbors with the standard distance-based strategy leading to an unbalanced and unadapted structure. We observed that with the rank-based structures, this phenomena is highly reduced, which is a direct consequence of the improvement of the local structure.

Finally, L2-ctxt benefits the least from the rank NN strategy. Our intuition is that looking at the local neighborhood through the context vector allows to select more intuitive and symmetric connections.

ii) *Rank structure and symmetry for SSL results.* Tab. 3 can be directly compared with the right part of Tab. 1 and shows the following results.

First, we see for the non-symmetric case the same kind of improvement with rank-based structure as in the NN study. Tab. 1 presents results for a structure built with distance-based neighbors and a non-symmetric strategy. When comparing it with the first column of Tab. 3, we see in the gain column, that the rank brings an improvement of up to 18.6% for ETH (L2 and C-SIFT), up to 10.4% for C-PASCAL (HOG for both L1 and L2) and 4.7% for AWA (for SURF).

Next, we observe a similar, but often smaller, improvement when comparing non-symmetric graphs (right part of Tab. 1) with symmetric graphs (middle column of Tab. 3). Descriptors improving significantly with the new rank structure

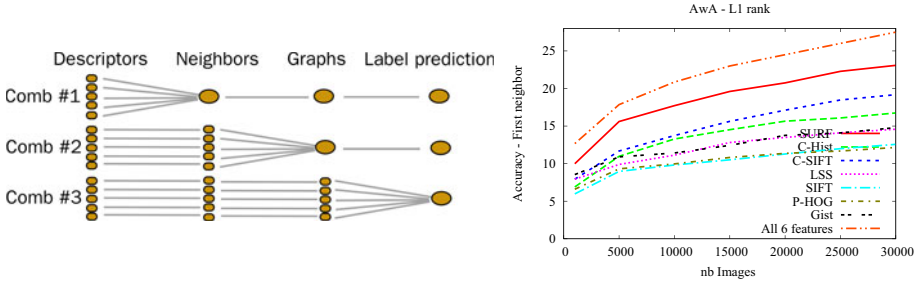

Fig. 3. Left: SSL combination strategies. Right: 1-NN rank-based structure quality for single features and their combination on AWA

also benefit the most from the symmetric graphs. Finally, when combining the two new strategies (symmetric graphs using a rank structure, shown on the right part of Tab. 3), we obtain similar or even better results than both previous strategies. As a more general comment, rank methods (non-symmetric or symmetric) combined with L1 give nearly always the best performance and bring significant overall improvement. For C-PASCAL, TPLBP's accuracy increases from 29.5% to **33.8**%. AWA benefits the most as the SURF descriptor improves from 9.0% accuracy to **16.7**% with a symmetric graph and a rank-based structure.

5 Combination

In the previous section, we have seen that we can significantly improve the local structure of the image collection for a given image representation and a given measure. In the following we will combine different features.

Feature combination has become an active area of research in the last years, and the supervised framework allows to learn the different feature contributions using the labels. Recent work [30] showed that simply averaging kernels already gives a good improvement. Consequently, both our related tasks - building the NN structure and predicting labels with SSL - should benefit from the combination of several image representations. In this section, we only look at the second task, i.e. image categorization using SSL algorithms.

Principles. Three graph combinations are considered (illustrated Fig. 3).

Combination #1 assumes that the combination is done on the structure level. The different features are used to compute a single k-NN local structure. Averaging all single feature distances builds a single list of *distance* NN. A multi-feature *rank score* builds a single list of *rank* NN. This score is calculated by $sc_{comb}(i,j) = \sum_{m \in Features} sc(d_i^m, d_j^m)$ where d_j^m is the m^{th} representation of image j, and $sc(d_i^m, d_j^m)$ is the single feature rank score (see in Sec. 4.2).

Combination #2 builds a graph for each feature, and forms one combined graph, the *union* graph, whose edges are the union of edges of each graph.

Combination #3 builds as many graphs as features and propagates labels in each graph. All propagation results are combined and yield the final label.

Table 4. Accuracy on unlabeled data only and gain compared to the best single feature for C-PASCAL (left) and AWA (right). Results are proposed for the rank based structures, for both symmetric and non-symmetric graphs.

C-PASCAL - SSL results							
		rank			rank sym		
Features	Comb	av	var	gain	av	var	gain
Gist+	#1	34.0	0.8	0.0	**34.8**	0.7	+0.1
P-BoF+	#2	35.6	0.9	+2.9	**36.4**	0.8	+2.6
TPLBP	#3	35.8	1.0	+3.1	**36.7**	0.9	+2.9
all	#1	34.9	0.8	+0.9	**35.7**	0.3	+0.9
	#2	**37.2**	0.6	+4.5	36.8	0.5	+3.0
	#3	**38.0**	0.7	+5.3	37.9	0.8	+4.1

AWA - SSL results							
		rank			rank sym		
Features	Comb	av	var	gain	av	var	gain
C-Hist+	#1	15.1	0.1	+1.4	**17.7**	0.2	+1.0
C-SIFT+	#2	**16.7**	0.1	+3.0	17.8	0.1	+1.1
SURF	#3	17.7	0.2	+4.0	**19.1**	0.2	+2.4
all	#1	17.3	0.3	+3.6	**20.1**	0.3	+3.3
	#2	18.1	0.2	+4.4	**19.2**	0.1	+2.5
	#3	19.9	0.2	+6.2	**21.8**	0.2	+5.1

Table 5. Successive improvements of the local structure: best single feature with distance (top) and with rank structure and symmetric graphs (middle), and best feature combination (bottom) - for the first neighbor quality (left), and the SSL results (right).

	1-NN quality			SSL-results		
strategy	ETH	C-PASCAL	AWA	ETH	C-PASCAL	AWA
single feature	96.9	34.6	16.4	89.0	29.5	9.0
single feature + rank	97.6	38.3	23.0	91.3	33.8	16.7
multiple features + rank	98.5	45.5	27.5	94.0	38.0	21.8

Combinations #2 and #3 use multiple graphs. Each graph can either be built from distance or rank based local structures.

Experiments. For the C-PASCAL and AWA datasets, we combine all descriptors and the 3 best performing ones. Due to space constraints, Tab. 4 only shows the SSL results[2], for L1 and for the 2 most promising strategies from the previous section, namely non-symmetric and symmetric graphs, on rank local structures. Each combination setting is considered for the 3 different combination strategies. Transductive accuracy is presented together with the gain in comparison to the best single feature within the combination.

As a first and expected conclusion, the combination of different features improves the SSL results in all settings. For C-PASCAL, the setting with all features improves the best single feature result by 5.3% reaching an accuracy of 38%. Also the AWA dataset benefits by 5.1% when combining all 7 descriptors reaching 21.8%. Second, there are only small differences between the combination methods, but combination #3 generally gives the best results.

Summary. If we look back at the different improvements of the local neighborhood structure we proposed in this paper, the absolute gain for each dataset is summarized in Tab. 5. In particular, SSL results are enhanced from 89% to **94**% for ETH, from 29.5% to **38**% for C-PASCAL and in the case of AWA we doubled the performance from 9% up to **21.8**% without any label. This underlines the assumption that the structure matters more than the SSL algorithm and that the structure can be improved in an unsupervised manner.

We believe that these encouraging results will be more pronounced for larger datasets. Compared to Fig. 1, Fig. 3 shows that both i) the ranking structure strategy and ii) the combination of features benefit more for larger datasets.

6 Conclusions

This paper explored ways of using the large amount of available image data in order to overcome inherent problems of supervised approaches. In particular, we consider methods which rely less on supervised classifiers and more on the structure of the data itself, namely the unsupervised construction of a local structure between images and the use of this structure in a SSL framework.

An important conclusion of our study is that the local structure – induced by the employed image representation, the distance measure and the number of nearest neighbors considered – matters more than the SSL algorithm. Zhu made this claim [31] together with the remark that there is only little work on the structure itself. In that sense, our study contributes to a better understanding of such structures for the tasks of object recognition and image categorization.

It is worth noting that the results obtained for the NN analysis directly translate into the corresponding performances of SSL algorithms. We indeed observed that the right set of parameters (image representation, distance measure and strategy to use it) can literally predict the SSL accuracy. On the more negative side, the overall performance obtained by the SSL algorithms is far from being satisfactory. This fits our intuition that unsupervised local structure contains some semantic information, but that the current object representations are not powerful enough for realistic datasets without supervised learning and discriminant classifiers.

To overcome these limitations we proposed different directions to improve the local structure of the dataset without any label and consequently improve the SSL results. In particular, we showed the benefits of contextual measures, symmetric relations between images, and feature combinations. Overall, a 12.8% accuracy improvement was obtained for the realistic AWA dataset without using any supervision for building the local structure.

As a conclusion, using large image collections and unsupervised local structure construction in combination with SSL algorithms is a promising direction. A generic structure can be built independently of the task, and then combined with different sets of labels. This structure can be improved by considering more suitable and complementary object and image representations, combining them, and using the information contained on the image collection topology.

References

1. Chum, O., Perdoch, M., Matas, J.: Geometric min-hashing: Finding a (thick) needle in a haystack. In: CVPR (2009)
2. Sivic, J., Russell, B., Zisserman, A., Freeman, W., Efros, A.: Unsupervised discovery of visual object class hierarchies. In: CVPR (2008)
3. Vijayanarasimhan, S., Grauman, K.: Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In: CVPR (2008)
4. Li, L.-J., Fei-Fei, L.: Optimol: Automatic online picture collection via incremental model learning. IJCV (2009)
5. Vijayanarasimhan, S., Grauman, K.: Multi-level active prediction of useful image annotations for recognition. In: NIPS (2008)

6. Russell, B., Torralba, A., Murphy, K., Freeman, W.: LabelMe: A database and web-based tool for image annotation. IJCV (2008)
7. Verbeek, J., Triggs, B.: Scene segmentation with CRFs learned from partially labeled images. In: NIPS, vol. 20 (2008)
8. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. IJCV (2007)
9. Leistner, C., Saffari, A., Santner, J., Bischof, H.: Semi-supervised random forests. In: ICCV (2009)
10. Saffari, A., Leistner, C., Bischof, H.: Regularized multi-class semi-supervised boosting. In: CVPR (2009)
11. Fergus, R., Weiss, Y., Torralba, A.: Semi-supervised learning in gigantic image collections. In: NIPS (2009)
12. Chapelle, O., Schölkopf, B., Zien, A. (eds.): Semi-Supervised Learning. MIT Press, Cambridge (2006)
13. Zhou, D., Bousquet, O., Lal, T.N., Weston, J., Schölkopf, B.: Learning with local and global consistency. In: NIPS (2004)
14. Zhu, X., Ghahramani, Z., Lafferty, J.: Semi-supervised learning using gaussian fields and harmonic functions. In: ICML (2003)
15. Liu, W., Chang, S.: Robust multi-class transductive learning with graphs. In: CVPR (2009)
16. Leibe, B., Schiele, B.: Analyzing appearance and contour based methods for object categorization. In: CVPR (2003)
17. Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., Zisserman, A.: The PASCAL VOC Challenge 2008 Results (2008)
18. Lampert, C., Nickisch, H., Harmeling, S.: Learning to detect unseen object classes by between-class attribute transfer. In: CVPR (2009)
19. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: CVPR (2005)
20. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV (2001)
21. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
22. Wolf, L., Hassner, T., Taigman, Y.: Descriptor based methods in the wild. In: ECCV (2008)
23. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. IJCV (2004)
24. Lowe, D.: Distinctive image features from scale-invariant keypoints. IJCV (2004)
25. van de Sande, K., Gevers, T., Snoek, C.: Evaluating color descriptors for object and scene recognition. PAMI (2010)
26. Hein, M., Maier, M.: Manifold denoising. In: NIPS (2006)
27. Bengio, Y., Delalleau, O., Le Roux, N.: Label propagation and quadratic criterion. In: [12]
28. Zhou, D., Huang, J., Schöhkopf, B.: Learning from labeled and unlabeled data on a directed graph. In: ICML (2005)
29. Perronnin, F., Liu, Y., Renders, J.: A family of contextual measures of similarity between distributions with application to image retrieval. In: CVPR (2009)
30. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: CVPR (2009)
31. Zhu, X.: Semi-supervised learning literature survey. Technical report, UW (2005)