Ensemble Partitioning for Unsupervised Image Categorization

Dengxin Dai, Mukta Prasad, Christian Leistner, and Luc Van Gool

Computer Vision Lab., ETH Zürich

Abstract. While the quality of object recognition systems can strongly benefit from more data, human annotation and labeling can hardly keep pace. This motivates the usage of autonomous and unsupervised learning methods. In this paper, we present a simple, yet effective method for unsupervised image categorization, which relies on discriminative learners. Since automatically obtaining error-free labeled training data for the learners is infeasible, we propose the concept of weak training (WT) set. WT sets have various deficiencies, but still carry useful information. Training on a single WT set cannot result in good performance, thus we design a random walk sampling scheme to create a series of diverse WT sets. This naturally allows our categorization learning to leverage ensemble learning techniques. In particular, for each WT set, we train a max-margin classifier to further partition the whole dataset to be categorized. By doing so, each WT set leads to a base partitioning of the dataset and all the base partitionings are combined into an ensemble proximity matrix. The final categorization is completed by feeding this proximity matrix into a spectral clustering algorithm. Experiments on a variety of challenging datasets show that our method outperforms competing methods by a considerable margin.

1 Introduction

Image categorization is an intensely studied field with many applications. In the last decade, advances in representations and supervised learning methods have led to great progress [1,2,3,4]. However, the explosion of available visual data, the cost of annotation, and the user-specific bias of annotations have resulted in an increased focus on learning with less supervision. While powerful methods for semi-supervised [5] or weakly-supervised learning [6,7] have been proposed, unsupervised methods have been studied less.

In this paper, we are interested in automatically discovering image categories from an unordered image collection. This task is hard as objects and scenes can appear with clutter, large variations in pose and lighting, high intra-class variances and low inter-class variances. The recent success of discriminative classifiers in supervised image categorization [1,2,4] suggests that one possible way to progress in unsupervised image categorization is to construct a similar scenario, in which the power of discriminative classifiers can be leveraged. Since obtaining a pure training set automatically is infeasible, in this paper, we

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part III, LNCS 7574, pp. 483–496, 2012.

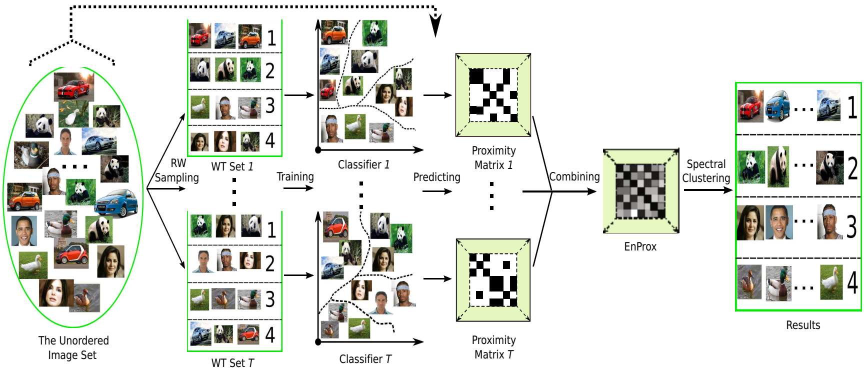

Fig. 1. The pipeline of our method for an image collection comprised of four categories: face, car, panda and duck. The pseudolabels in the weak training (WT) sets are assigned by our random walk sampling scheme. The image collection is partitioned by the classifiers learnt from the WT sets and the results are stored into the binary proximity matrices. All of the binary matrices are then averaged into the ensemble proximity matrix (EnProx) as input for spectral methods. As expected, the partitionings of the WT sets contain errors, but their errors differ, which discovers the underlying truth by ensemble learning techniques.

propose the concept of WT sets (*c.f.* definition in § 2) that can be exploited with supervised learning techniques. Since knowledge carried by a single WT set is limited, we develop a very simple, yet effective random walk (RW) sampling method to create a set of diverse WT sets. The RW sampling is performed in the space of pairwise similarities built up by common distance measures. Samples that are hard to cluster using common distance measures are left out in the sampling step so that the sampled WT sets are not too noisy. In order to learn knowledge contained in these WT sets for unsupervised categorization, we propose the ensemble partitioning framework based on the ensemble learning principle [8].

The remainder of this paper is structured as follows: § 2 reports on related work and introduces our method. In § 3 we make observations that inspired the approach described in § 4. We show improvement over the state-of-the-art through evaluations in § 5. § 6 concludes the paper.

2 Unsupervised Image Categorization

2.1 Related Work

Several unsupervised image categorization techniques have been proposed. Fergus *et al.* [9] modeled objects as constellations of visual parts and estimated parameters using the expectation-maximization algorithm for unsupervised recognition. Sivic *et al.* [10] proposed using aspect models to discover object categories from an unordered image collection. Aspect models were originally developed for content analysis in textual document collections, but they perform very well for images as well [10]. Later on, [11] used Hierarchical Latent Dirichlet Allocation to automatically discover object class hierarchies. For scene category discovery, Dai *et al.* [12] proposed to combine information projection and clustering sampling. All these methods share assumptions about the sample distribution. Image categories, nevertheless, are arranged in complex and widely diverging shapes, making the design of explicit models difficult. An alternative strand, which is more versatile in handling structured data, builds on similarity-based methods. Frey and Dueck [13] applied the affinity propagation algorithm [14] for unsupervised image categorization. Grauman and Darrell [15] developed partially matching image features to compute image similarity and used spectral methods for image clustering. The main difficulty of this strand is how to measure image similarity as the semantic level goes up. For a more detailed survey, we refer readers to [16].

The method closest to ours is that of Gomes *et al.* [17]. They also propose a framework for learning discriminative classifiers while clustering the data, called regularized information maximization. Yet, the method still significantly differs from ours. In particular, [17] finds a partitioning of images that results in a 'good' discriminative classifier, which is evaluated by three criteria: class separation, class balance and classifier complexity. Our method however, more closely follows the training-test paradigm; that is, sample a series of weak training sets, learn knowledge from these sets and classify the whole dataset with this learned knowledge. Our method also shares similarities with the method of Lee and Grauman [18]. They proposed to use *curriculum learning* (CL) [19] for unsupervised object discovery. CL imitates the learning process of people; that is, using a pre-defined learning schedule, it starts from learning the easiest concepts first and graduately increases the complexity of concepts [19]. Besides technical differences, the main benefit of our work compared to [19] is that we do not rely on pre-selected learning schedules.

As we already stated, our method builds on the ensemble learning principle [8]. Ensemble methods build a committee of (usually weak) base learners that overall performs better than its individual parts. Thereby, they benefit from the strength and diversity of their base learners. Popular ensemble methods that have been successfully applied in computer vision are boosting [20], bagging [21] and random forests [22]. There has been previous work using ensemble methods for clustering *e.g.* [23,24]. The main difference of our work to [23,24] is that we focus on high-level categories. [23,24] make basic assumptions about the underlying distributions, which are hard to obtain for our task and, hence, makes previous methods only applicable in rather low-level clustering tasks. In contrast, we propose to learn category distinctions from sampled training data. We also differ from previous approaches in the way we combine our base learners. In particular, [24] used a boosting framework and [23] used hierarchical clustering, while our method is based on spectral techniques.

2.2 Notation and Overview

In this section, we first give the definition of the weak training set and then give an overview of our method.

Definition 1. *Given a categorization task $\mathbb{E}$, a weak training set is a labeled dataset $\mathcal{B}$ which has severe deficiencies to be a good training set, e.g. some categories may be absent, some may be impure, and/or some may be with bias, but training on it can lead to a weak classifier $\phi(.)$ which performs better than random guessing for the task $\mathbb{E}$.*

Our dataset $\mathcal{D}$ consists of N images, each belonging to one of C categories. The task is to discover these categories. Each image is represented by a d-dimensional feature vector $f_n, n \in \{1 \dots N\}$. Assume that the ensemble partitioning learns knowledge from T WT sets in total. For the t^{th} trial, the WT set $\mathcal{D}_t = \{\mathbf{s}_t, \hat{\mathbf{c}}_t\}$ is a set of sampled images $\mathbf{s}_t$ and their pseudolabels $\hat{\mathbf{c}}_t$, which *imitate* the role of labeled training data in supervised frameworks. The unordered image collection $\mathcal{D}$ is partitioned according to the discriminative classifier $\phi(.)_t$ learnt from the WT set $\mathcal{D}_t$. Classification from each trial t creates a unique partitioning of the data $\mathcal{D}$. Since pseudolabels cannot be mapped into correspondence across different trials, the classification agreement between image pairs in each trial t is stored as a binary proximity matrix $\mathtt{A}^t$, where $a^t_{ij} = 1$ implies that images $i, j \in \{1 \dots N\}^2$ belong to the same class. The final proximity matrix $\mathtt{A}$, is averaged over all the T binary matrices: $\mathtt{A} = \frac{1}{T}\sum_t \mathtt{A}^t$. The final categorization results are obtained by feeding $\mathtt{A}$ to a spectral clustering algorithm. The whole pipeline of our method is sketched in Fig. 1. Though simple, our method outperforms competing methods by a considerable margin on various challenging datasets.

The main contribution of this paper is a novel approach to unsupervised image categorization. In particular, we (i) propose the concept of WT sets for image categorization, (ii) design a random-walk based sampling scheme to generate them, and (iii) propose the ensemble partitioning framework to learn the category distinctions for unsupervised image categorization.

3 Observations

In this section, we share some insights with the reader that help to motivate our approach and understand why it is working better. We raise two questions: First, how much does the impurity of the training data in WT sets influence the convergence of the learning? Second, given a standard distance metric, how does one approach the problem of getting good WT sets in an unsupervised manner?

Observation 1: Ensemble learning can learn the essence of categories from a series of diverse WT sets. We examined this idea in supervised image categorization. Given the ground truth data divided as training and test sets: $\mathcal{D} = \{\mathcal{D}^{\text{train}}, \mathcal{D}^{\text{test}}\}$, (i) we artificially synthesized a set of WT sets $\mathcal{D}^{\text{train}}_t, t = 1, \dots, T$ from training data $\mathcal{D}^{\text{train}}$, and (ii) ensemble learning was then performed on these sets and its performance on test data classification was measured.

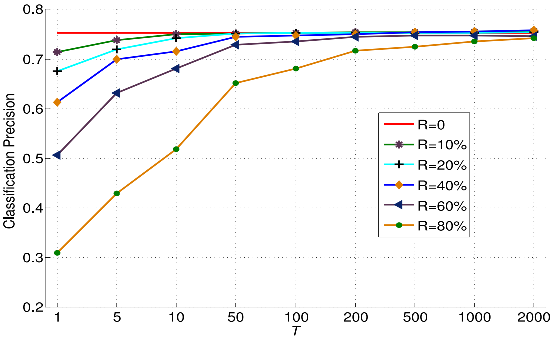

Fig. 2. Classification accuracies of ensemble learning on the 15-Scene dataset [1], for varying training label noise R and number of WT sets T. High precision can be learnt for noise percentage as high as 80% given enough WT sets (Best viewed in color).

In order to guarantee the diversity of the training sets (for ensemble learning), each WT set $\mathcal{D}_t^{\text{train}}$ is formed by randomly taking 30% of the images from $\mathcal{D}^{\text{train}}$, and randomly re-assigning labels of a fixed percentage R of these. Hence, $R = 0$ corresponds to the upper performance bound as every sample is assigned its true label. A classifier is trained for each of these WT sets. At test time, each of these classifiers returns the category label of each image in $\mathcal{D}^{\text{test}}$. The winning label is the mode of the results returned by all the classifiers. Fig. 2 evaluates this for the 15-Scene dataset [1]. Linear SVMs are used as the classifiers with a concatenation of GIST [25], PHOG [2] and LBP [26] as input. When the label noise percentage R is low, the classification precision starts out high and levels quickly with T, as one would expect. But interestingly, for R even as high as 80%, the classification precision, which starts low, converges to a similarly high precision given more WT sets T (≈ 500). This shows that it is possible to learn the essence of image categories even from very weak (extremely mislabelled) training sets, given a sufficient number with diverse deficiencies. This is crucial for our task, because WT sets are much easier to get than pure training data in unsupervised settings.

Observation 2: Given a standard distance metric, how does one approach the problem of getting good (precise and diverse) WT sets without supervision? An ideal image representation along with a powerful metric should ensure that all images of the same category are more similar to each other than to those of other categories. In order to examine it, we tabulate how often an image is from the same category as its k^{th}-nearest neighbor. We refer to the frequency as label co-occurrence probability $p(k)$. $p(k)$ is averaged across images and labels in the dataset. Various features and distance metrics were tested: GIST [25] with Euclidean distance, PHOG [2] and LBP [26] with the χ^2 distance.

Fig. 3 shows the results on the 15-Scene dataset [1]. The results reveal that using the distance metric in conventional ways (*e.g.* clustering by K-means and spectral methods) will result in very noisy training sets, because the label co-occurrence probability $p(k)$ drops very quickly with k. However, sampling in the very close neighborhood of a given image is likely to generate more instances of the same category, whereas sampling in very far-away space can gather samples of new categories. This suggests that samples along with a few very close neighbors,

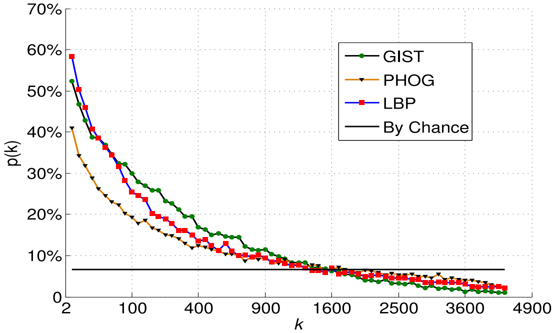

Fig. 3. The label co-occurrence probability $p(k)$ calculated on the 15-Scene dataset (4485 images in total) using GIST [25] with Euclidean distance, and PHOG [2] and LBP [26] with the χ^2 distance (Best viewed in color).

namely "compact" image clusters, can form a training set for a single class, and a set of such image clusters which scatter far away from each other in feature space can serve as a good WT set for our task. Furthermore, sampling in this way provides the chance of creating a large number of diverse WT sets, due to the small size of the sampled WT set. This will be further examined in § 4.1.

4 Our Approach

In this section, we introduce our approach for unsupervised image category discovery. It has two interleaved parts: 1) the creation of the weak training (WT) sets by random walk (RW) sampling, and 2) learning the ensemble proximity matrix $\mathtt{A}$ for categorization.

4.1 Creating WT Sets via Random Walk Sampling

A WT set $\mathcal{D}_t = \{\mathbf{s}_t, \hat{\mathbf{c}}_t\}$ consists of a set of $\mathbf{s}_t \in \{1 \dots N\}^Q$ images of size Q sampled from $\mathcal{D}$, where $Q = CK$ and K is the *ideal* number of training images for each of the C classes. $\hat{\mathbf{c}}_t \in \left\{1 \dots \hat{C}_t\right\}^Q$ are pseudo labels, assigned during the creation of WT set $\mathcal{D}_t$. Note that $\hat{C}_t$ refers to the number of empirical categories estimated in the t^{th} trial and does not necessarily have to reflect the true C.

Recall that as an insight from § 3, in the creation process of WT sets and with respect to a standard distance measure, images that get assigned the same pseudo labels $\hat{\mathbf{c}}$ should be very close to each other and images with different pseudo labels far away, respectively. Additionally, ensemble learning theory teaches us that weak learners should have low bias but high variance to form a powerful combined classifier [8]. We use strong discriminative learners (see below) to achieve the first requirement and we incorporate randomness into the sampling process to ensure the second one [21,22].

To this end, we adopt a random walk [27] based sampling method. In detail, each image is taken as a node and the weight of the edge between nodes i and j is defined as

$$w_{i,j} = \begin{cases} 0 & if\ i = j \\ \exp\{-\frac{\text{dist}(\mathbf{f}_i, \mathbf{f}_j)^2}{\sigma^2}\} & \text{otherwise} \end{cases} \tag{1}$$

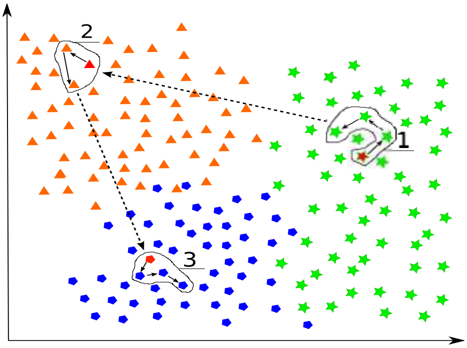

Fig. 4. A toy example of random walk sampling in an image collection comprised of images from three categories. Solid lines indicate the process of exploiting neighborhoods. Dash lines indicate the process of exploring the space for new categories. The isolated red samples indicate the 'far' samples explored. The numerical numbers denotes the pseudolabels. (Best viewed in color).

where $\mathrm{dist}(\cdot,\cdot)$ denotes a common distance measure and σ is a scale parameter. Besides the N nodes from individual images, we take the current WT set $\mathcal{D}_t$ as the $(N+1)^{th}$ node in order to provide the possibility of jumping to a far-away node from all visited nodes. Its weight to node $j \in \{1, ..., N\}$ is defined as $w_{N+1,j} = \mathrm{argmax}_{\forall i \in \mathbf{s}_t} w_{i,j}$ and to itself it is 0. The maximization is used to guarantee that nodes far away from node $(N+1)$ are far away from all the visited nodes. Note that node $N+1$ and its weights need to be updated when a new node is visited: a new image is added into $\mathcal{D}_t$. When $\mathcal{D}_t = \emptyset$, its weights are uniformly distributed over all images. Having all the weights, we define the transition matrix to be

$$\mathtt{P} = [p_{i,j}]_{(N+1)\times(N+1)} = \left[\frac{w_{i,j}}{\sum_k w_{i,k}}\right]_{(N+1)\times(N+1)} . \tag{2}$$

Obtaining each WT set $\mathcal{D}_t$ involves sampling for Q images (visiting Q nodes) and assigning pseudolabels $\hat{\mathbf{c}}_t$ to them. Starting with $\mathcal{D}_t = \{\}$, it is built up using the following scheme: at each instant either jump to a neighboring node i of the previously visited node j to exploit the regions already explored, or jump to a node i that is far away from all the visited nodes, to explore the space of new categories. The corresponding image is added to $\mathbf{s}_t$. The pseudolabels are generated as follows: every time a 'far' node is explored, a new class label is assigned to it. If a node i neighboring node j is selected, the class label of image j is assigned to it. The above scheme is formally described in Algo. 1 and a toy example is illustrated in Fig. 4. For the handling of node $N+1$, the reader is referred to Algo. 1.

4.2 Ensemble Partitioning

We now explore the use of the WT sets acquired in § 4.1 for image categorization. In particular, we train a discriminative classifier for each trial t: $\phi_t(\cdot) \in \left\{1 \ldots \hat{C}_t\right\}$ for each WT set $\mathcal{D}_t$. The full data $\mathcal{D}$ is then classified by ϕ_t to obtain a binary proximity matrix $\mathtt{A}^t$, where $a^t_{ij} = 1$, if $\phi_t(\mathbf{f}_i) = \phi_t(\mathbf{f}_j)$, and 0 otherwise. The binary proximity matrices over multiple trials are combined into an ensemble proximity matrix $\mathtt{A} = (1/T)\sum_t \mathtt{A}^t$. We employ an SVM-based $\hat{C}_t$-class classifier

Algorithm 1. RW Sampling in t^{th} trial

```
Input:
– Image data $\mathcal{D}$, feature representation matrix $\mathtt{F}$
– Initialize training set $\mathbf{s}_t = \{any(1 \ldots N)\}$ and $\hat{\mathbf{c}}_t = \{1\}$
– Initialize transition matrix $\mathtt{P}$
for $m = 1 \rightarrow Q - 1$ do
    1. With probability $(Q - C)/Q$ exploit the neighborhood:
    Jump to a neighboring node $i^* \in \{1, ...N\} \backslash \mathbf{s}_t$ with probability $\frac{\mathtt{P}(\mathbf{s}_{tm}, i^*)}{\sum_{j \in \{1, \ldots N\} \backslash \mathbf{s}_t} \mathtt{P}(\mathbf{s}_{tm}, j)}$, $c^* = \hat{\mathbf{c}}_{tm}$ OR
    2. With probability $C/Q$ explore new space:
    Jump to a ‘far’ node $i^* \in \{1, ..., N\} \backslash \mathbf{s}_t$ with probability $\frac{1 - \mathtt{P}(N+1, i^*)}{\sum_{j \in \{1, \ldots N\} \backslash \mathbf{s}_t} 1 - \mathtt{P}(N+1, j)}$,
    assign a new category label to $c^*$.
    Add sample $i^*$ to $\mathbf{s}_t$ and pseudolabel $c^*$ to $\hat{\mathbf{c}}_t$.
    Update transition matrix $\mathtt{P}$.
end for
```

Algorithm 2. Ensemble Partitioning

```
Input:
– the image dataset $\mathcal{D}$, feature representation $\mathtt{F}$.
– Initialize proximity matrix $\mathtt{A} = 0$.
for $t = 1 \rightarrow T$ do
    0. Initialize the current proximity matrix $\mathtt{A}^t = 0$.
    1. Sample $\mathcal{D}_t = \{\mathbf{s}_t, \hat{\mathbf{c}}_t\}$ using Algo. 1.
    3. Train classifier $\phi_t(\cdot) \in \left\{1 \ldots \hat{C}_t\right\}$ on $\mathcal{D}_t$.
    4. Partitioning the entire data $\mathcal{D}$ according to:
    $a^t_{ij} = \Delta(\phi_t(\mathbf{f}_i), \phi_t(\mathbf{f}_j)), \forall \{i, j\} \in \{1 \ldots N\}^2$,
end for
$\mathtt{A} = \frac{1}{T} \sum_t \mathtt{A}^t$.
Obtain object categorization by spectral clustering on $\mathtt{A}$.
```

ϕ_t in each trial as they usually work well for most features types. Additionally, they have also been shown to generalize well even when trained only on a few training samples [28], which is an important characteristic as we keep the size of the individual WT sets relatively small. Given a collection of WT sets, the ensemble partitioning is summarized in Algo. 2. Final categorization can be completed by feeding $\mathtt{A}$ to any similarity-based clustering algorithm. In this work, we choose spectral clustering due to its simplicity and popularity. The whole pipeline of Fig. 1 has now been explained.

5 Experiments

In this section, we elaborate upon the datasets, evaluation criteria and experimental settings for the evaluation of the proposed approach.

Evaluation Datasets: We tested our method on three datasets: 15-Scene [1], Caltech-101 [29], and a 35-Compound dataset. The 15-Scene dataset contains 15 scene categories in both indoor and outdoor environments. Each category has 200 to 400 images, and they are of size 300 × 250 pixels on average. The Caltech-101 dataset [29] contains 101 object categories and is one of the most popular datasets for supervised object recognition. The large number of categories, combined with the uneven category distribution (from 31 to 800 images per category), poses a real challenge for unsupervised categorization. In order to evaluate our method on a more general image dataset, we composed a 35-Compound data set by mixing up the 15 scene categories [1] with the 20 object categories selected from Caltech 256. The 20 object categories are: American flag, fern, French horn, leopards 101, pci card, tombstone, airplanes 101, diamond ring, fire extinguisher, ketch 101, mandolin, rotary phone, Pisa tower, face easy 101, dice, fireworks, killer whale, motorbikes 101, roulette wheel, and zebra.

Features: The following features are considered in our experiments: f_1 =SIFT [30], f_2 = GIST [25], f_3 =PHOG [2], f_4 = LBP [26], f_5 = GIST + PHOG, f_6 = GIST + LBP, f_7 = PHOG + LBP, and f_8 = GIST + PHOG + LBP, where + means a simple concatenation. GIST features are computed on resized images of 256 × 256 pixels and other features are computed on the original images. For PHOG, we computed the derivatives in 8 directions and used a two-layer pyramid. For LBP, the uniform LBP was used. For SIFT, we used Hessian detectors and k-means to get 300 centroids for the bag-of-feature representation.

5.1 Experimental Setup

Our Method: We evaluated our approach against competing methods for the datasets and features described above. The number of categories C is assumed to be known. Although the transition matrix $\mathtt{P}$ of § 4.1, Algo. 1, the classifier ϕ_t of § 4.2 and Algo. 2 use the same notation for their features, speed-accuracy tradeoffs may cause different features to be optimal at each stage. Since GIST features have proven very effective for holistic image categorization with the Euclidean distance [25], we adopted them for building $\mathtt{P}$ of Algo. 1, where $\text{dist}(\cdot,\cdot)$ in Eq. 1 is the Euclidean distance between features. For training the classifiers, $f_2 - f_8$ were considered [1]. For the evaluation and comparison across all the datasets and features, we used the same set of parameters: $T = 1000$ and $K = 9$. The influence of these two parameters on the performance of our method was also tested on the three datasets, where a set of T's in the range of $[1, 1000]$ and a set of K's in the range of $[3, 30]$ were evaluated. The σ in Eq. 1 is set automatically by using a self-tuning method proposed in [31], which has already shown promising performance for similarity scaling.

Competing Methods: We compared our method to well-known clustering methods: k-means and spectral clustering. For k-means, four features $f2 - f4$ and $f8$ were tried. For spectral clustering, the same features are tried: $f2$ with the Euclidean distance measure, $f3$ and $f4$ with the χ^2 distance measure, and $f8$

[1] The reason why we did not consider $f1$ is to avoid severe over-fitting, given the small size of the WT sets and the high dimension of $f1$.

with the Mahalanobis distance measure with diagonal covariances (the off diagonal elements were set to 0). We selected these combinations just because the χ^2 distance measure is superior for histogram-based features, and the Mahalanobis distance measure with diagonal covariances is capable of learning feature weights. We also compared our method to the PLSA-based method of [10], the affinity propagation method (AP) of [14], and the regularized information maximization method (RIM) of [17]. $f1$ is used as the input for the PLSA-based method. Spatial pyramid matching (SPM) [1] is used as the input for the RIM as it was used in [17]. The AP used the same input as the spectral clustering. For the implementation, we used the authors' code directly [2] [3] [4]. Since the number of categories cannot be set directly in the RIM and the AP – the input parameters are searched to yield the right number of categories – we only tested these two methods on the 15-Scene dataset.

Baselines for Sampling: To investigate the performance of our random walk sampling scheme, we create two baselines for obtaining the WT sets. Everything else in Algo. 2 remains as before. GIST features are also used for the baselines.

1. **Baseline 1:** In the t^{th} trial, run k-means on the whole dataset to cluster images into C groups. The WT set $\mathcal{D}_t$ was created by adding all images and their labels to $\mathbf{s}_t$ and $\mathbf{c}_t$, resp.
2. **Baseline 2:** In the t^{th} trial, we run k-means on a bootstrap subset of the whole dataset. For a fair comparison to the WT sampling, CK images were randomly selected for each subset with $K = 9$. Again, the WT set $\mathcal{D}_t$ was created by adding all images and their labels to $\mathbf{s}_t$ and $\mathbf{c}_t$, resp.

5.2 Results and Discussion

We follow [12,16] to use the purity (the bigger the better) for performance evaluation. First, let us compare our method to existing methods. Table 1 lists the average and variance of the purity over 10 runnings of all the methods on 15-Scene, Caltech-101 and the 35-Compound dataset. The table shows that our method outperforms all the competing methods by a substantial margin on all the three datasets while having comparable variance. The superiority of our method can be attributed to its discriminative learning ability, which gives more weight to relevant variables and rejects the irrelevant ones. This can be proven by simply examining how the performance of our method improves when providing more features to the classifier. This learning ability is crucial for image and object categorization, as so far no single feature can describe all image categories very well. The classifier trained on each individual WT set may yield over-fitting, but this can be offset by the classifiers trained on other WT sets as the sets are mutually diverse.

[2] http://www.robots.ox.ac.uk/~vgg/software

[3] http://vision.caltech.edu/~gomes/software.html

[4] http://www.psi.toronto.edu/index.php?q=affinity%20propagation

Table 1. Categorization results on 15-Scene, Caltech-101 and the 35-Compound dataset. SClustering, EnPar, G, P, L are the shorthands of spectral clustering, ensemble partitioning, GIST, PHOG, and LBP respectively. $\mathtt{A}(f)$ means that feature f is used for training the classifier of our method. $\mathtt{A}_1$ and $\mathtt{A}_2$ indicate that the WT sets are obtained by the baseline 1 and baseline 2 respectively.

Methods	Input	15-Scene	Caltech-101	35-Compound
		Purity (%)	Purity (%)	Purity (%)
K-means	G	39.44 (1.01)	28.71 (0.30)	34.32 (0.39)
K-means	P	29.99 (0.72)	30.35 (0.26)	26.71 (0.47)
K-means	L	29.01 (0.66)	29.66 (0.31)	33.06 (0.34)
K-means	G+P+L	31.50 (1.08)	35.51 (0.35)	35.28 (0.67)
Sclustering	Euclidean(G)	43.48 (1.04)	33.48 (0.52)	41.12 (0.32)
Sclustering	χ^2(P)	29.32 (0.89)	35.77 (0.82)	33.16 (0.73)
Sclustering	χ^2(L)	34.09 (0.96)	34.71 (0.71)	44.49 (0.60)
Sclustering	Mahalanobis(G+P+L)	42.05 (1.60)	37.72 (0.64)	48.51 (0.85)
PLSA [10]	SIFT	29.34 (1.21)	26.57 (0.81)	28.21 (0.79)
RIM [17]	SPM[1]	38.40 (1.04)	–	–
AP [14]	Euclidean(G)	40.71 (0.98)	–	–
AP [14]	χ^2(P)	32.14 (0.73)	–	–
AP [14]	χ^2(L)	35.23 (0.82)	–	–
AP [14]	Mahalanobis(G+P+L)	44.24 (1.01)	–	–
EnPar	$\mathtt{A}$(G)	46.31 (0.88)	35.79 (0.60)	45.14 (0.66)
EnPar	$\mathtt{A}$(P)	44.31 (0.81)	36.20 (0.58)	41.78 (0.59)
EnPar	$\mathtt{A}$(L)	46.44 (0.91)	37.16 (0.64)	53.18 (0.71)
EnPar	$\mathtt{A}$(G+P)	56.27 (0.98)	39.37 (0.79)	53.17 (0.68)
EnPar	$\mathtt{A}$(G+L)	55.41 (0.95)	39.60 (0.78)	55.01 (0.72)
EnPar	$\mathtt{A}$(P+L)	56.10 (1.01)	40.40 (0.83)	56.60 (0.81)
EnPar	$\mathtt{A}$(G+P+L)	**61.49** (1.08)	**42.31** (0.88)	**58.34** (0.76)
EnPar	$\mathtt{A}_1$(G+P+L)	47.52 (0.84)	35.76 (0.55)	42.70 (0.65)
EnPar	$\mathtt{A}_2$(G+P+L)	55.43 (0.98)	39.90 (0.82)	52.12 (0.74)

Secondly, let us compare the three ways of creating the WT sets. From Table 1, we can see that the RW sampling outperforms the two baselines substantially. This superiority can be attributed to the good diversity and accuracy of the WT sets generated by the RW sampling, obviously fulfilling the requirements of successful ensemble learning [8]. The accuracy comes from the underlying principle of the RW sampling: with a high probability choosing the most separable images and leaving out the ambiguous images, which will be handled after abstracting knowledge by the max-margin learning. The great diversity comes from the random jumping by the RW sampling. The possibility that two WT sets are the same or very similar is very low. The superiority of baseline 2 over baseline 1 results from the higher diversity of the WT sets that baseline 2 generated. Using the bootstrap technique to increase the diversity of training sets and in turn boost overall performance, has been widely accepted for supervised learning. Our experimental results confirm that this diversity is also very important for unsupervised scenarios like ours.

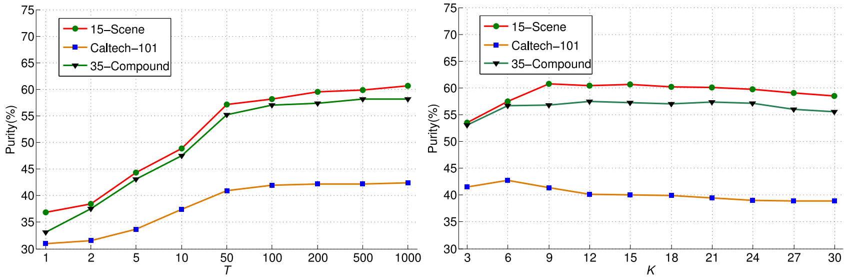

Fig. 5. The purity of discovered categories as a function of T and K

One may argue that the WT sets generated by the RW sampling are not precise and the precision of some individual sets may be affected by the outliers in the dataset. This is true, but our method demands far less from a WT set than a supervised categorization method does from its training set – the label assignment of the WT sets only needs to be better than random guessing. The imprecision of one WT set can be offset by other WT sets since their deficiencies are different. This is analogous to the weak learner of ensemble learning methods such as bagging and boosting – each weak learner could be imprecise, but the ensemble is precise and with good generality. Also, our categorization framework is very flexible and not tied to the way the WT sets are created. Any method that can generate good WT sets can be used in this framework. The RW sampling is a case in point. An interesting thing is that even using the WT set created by the two casual baselines, the ensemble partitioning framework can generate more precise results than competing methods.

Furthermore, let us examine the influence of the parameters T and K on the performance of our method. Fig. 5 shows the evaluation results over a variety of values for T and K. From the curves of T in Fig. 5, it is evident that the purity increases pretty fast with T but then also stabilizes quickly to a good performance. That is, our method benefits from adding more WT sets and can provide accurate results when a fairly large number of WT sets are provided. This is quite similar to how bagging [21] and random forests [22] behave with different numbers of training sets. For parameter $K = Q/C$, we find that very small and very large values (*e.g.* 3 and 30 in our case) degrade the performance. A very small K leads to insufficient training samples per category and a very large K results in very noisy training sets. But as Fig. 5 shows, our method is not overly sensitive to the choice of K: there are large ranges to select appropriate values from. Its choice can be based on the experience of choosing training data for supervised categorization.

Finally, though additional time is needed for training (most unsupervised categorization methods need no training), our method is still quite efficient. The efficiency comes from two factors: 1) Training linear SVMs is very efficient by using the optimized package[5]; and 2) the performance of our method stabilizes

[5] www.csie.ntu.edu.tw/~cjlin/liblinear

quickly with respect to T as Fig. 5 shows. The training time on a Core i5 2.80 GHz desktop PC are: 3.67 minutes for 15-Scene (4485 images in total), 66.18 minutes for Caltech-101 (8251 images in total), and 12.33 minutes for the 35-Compound dateset (8671 images in total). Furthermore, our method is inherently parallelizable and can take advantage of multi-core processors.

6 Conclusion

This paper has tackled the hard task of unsupervised image categorization. To this end, we leveraged the power of discriminative learning and ensemble methods. We presented the concept of weak training sets and proposed a new yet simple sampling scheme, denoted as RW sampling, to generate them. For each weak training set, a discriminative classifier is trained to get a base partitioning of the unordered image collection and then all the base partitionings are combined to a strong ensemble proximity matrix that can be incorporated into spectral clustering. In the experiments on several challenging datasets, we showed that our method is able to consistently outperform the state-of-the-art methods. The study poses multiple interesting questions for future research: Is it possible to design more sophisticated methods for the creation of the WT sets, rather than relying on the common distance metric as we do? What is the way of extending the method to handle noisy web data? How could we apply the method to weakly-supervised image categorization?

Acknowledgements. The authors gratefully acknowledge support from SNF project Aerial Crowds and the ERC Advanced Grant VarCity.

References

1. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR (2006)
2. Bosch, A., Zisserman, A., Muoz, X.: Image classification using random forests and ferns. In: ICCV (2007)
3. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR (2008)
4. Xiao, J., Hays, J., Ehinger, K.A., Oliva, A., Torralba, A.: Sun database large-scale scene recognition from abbey to zoo. In: CVPR (2010)
5. Fergus, R., Weiss, Y., Torralba, A.: Semi-supervised learning in gigantic image collections. In: NIPS (2009)
6. Deselaers, T., Alexe, B., Ferrari, V.: Localizing Objects While Learning Their Appearance. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part IV. LNCS, vol. 6314, pp. 452–466. Springer, Heidelberg (2010)
7. Blaschko, M.B., Vedaldi, A., Zisserman, A.: Simultaneous object detection and ranking with weak supervision. In: NIPS (2010)
8. Dietterich, T.G.: Ensemble Methods in Machine Learning. In: Kittler, J., Roli, F. (eds.) MCS 2000. LNCS, vol. 1857, pp. 1–15. Springer, Heidelberg (2000)
9. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR (2003)

10. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: ICCV (2005)
11. Sivic, J., Russell, B.C., Zisserman, A., Freeman, W.T., Efros, A.A.: Unsupervised discovery of visual object class hierarchies. In: CVPR (2008)
12. Dai, D., Wu, T., Zhu, S.C.: Discovering scene categories by information projection and cluster sampling. In: CVPR (2010)
13. Dueck, D., Frey, B.J.: Non-metric affinity propagation for unsupervised image categorization. In: ICCV (2007)
14. Frey, B.J., Dueck, D.: Clustering by passing messages between data points. Science 315, 972–976 (2007)
15. Grauman, K., Darrell, T.: Unsupervised learning of categories from sets of partially matching image features. In: CVPR (2006)
16. Tuytelaars, T., Lampert, C.H., Blaschko, M.B., Buntine, W.: Unsupervised object discovery: A comparison. IJCV 88, 284–302 (2009)
17. Gomes, R., Krause, A., Perona, P.: Discriminative clustering by regularized information maximization. In: NIPS (2010)
18. Lee, Y.J., Grauman, K.: Learning the easy things first: Self-paced visual category discovery. In: CVPR (2011)
19. Bengio, Y., Louradour, J., Collobert, R., Weston, J.: Curriculum learning. In: ICML (2009)
20. Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: a Statistical View of Boosting. The Annals of Statistics 28, 337–407 (2000)
21. Breiman, L.: Bagging predictors. ML 24, 123–140 (1996)
22. Breiman, L.: Random forest. ML 45, 5–32 (2001)
23. Leisch, F.: Bagged clustering. Working Paper Series (1999)
24. Saffari, A., Bischof, H.: Clustering in a boosting framework. In: CVWW (2007)
25. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV 41, 145–175 (2001)
26. Ojala, T., Pietikinen, M., Menp, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 24, 971–987 (2002)
27. Harel, D., Koren, Y.: On Clustering Using Random Walks. In: Hariharan, R., Mukund, M., Vinay, V. (eds.) FSTTCS 2001. LNCS, vol. 2245, pp. 18–41. Springer, Heidelberg (2001)
28. Malisiewicz, T., Gupta, A., Efros, A.A.: Ensemble of exemplar-svms for object detection and beyond. In: ICCV (2011)
29. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: WGMBV (2004)
30. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60, 91–110 (2004)
31. Zelnik-Manor, L., Perona, P.: Self-tuning spectral clustering. In: NIPS (2004)