Unsupervised Robust Feature-based Partition Ensembling to Discover Categories

Roberto J. López-Sastre
GRAM, University of Alcalá, Spain
robertoj.lopez@uah.es

Abstract

The design of novel robust image descriptors is still a formidable problem. Different features, with different capabilities, are introduced every year. However, to explore how to combine them is also a fundamental task. This paper proposes two novel strategies for aggregating different feature-based image partitions to tackle the challenging problem of discovering objects in unlabeled image collections. Inspired by consensus clustering models, we introduce the Aggregated Partition (AP) approach, which, starting from a set of weak input partitions, builds a final partition where the disagreements with the input partitions are optimized. We then generalize the AP formulation and derive the Selective AP, which automatically identifies the subset of features and partitions that further improves the precision of the final partition. Experiments on three challenging datasets show how our methods are able to consistently outperform competing methods, reporting state-of-the-art results.

1. Introduction

The design of novel image descriptors is a topic that has been obsessively drawing the attention of the computer vision research community. Every year, several novel robust features are introduced. Some of them are hand-crafted, while others are directly learned from the data, using, for instance, deep learning. They all come with different capabilities and strengths.

Indeed, this is still a challenging problem. However, in this paper, we argue that it is also a fundamental task to develop novel models which are able to effectively use and combine these descriptors in order to improve the performance of the approaches in which they are involved.

In particular, we address here the formidable problem of discovering objects in unlabeled image collections [3, 16, 28, 32]. Note that the setting of this problem is completely unsupervised: given a set of unlabeled images, the goal is to separate the different classes.

As illustrated in Figure 1, in this work we propose a model to discover object categories based on the power of the ensemble of robust descriptor-based image partitions. Following the motto, "*One for all, all for one*", our hypothesis is that by combining the strengths of the different features we can improve the performance and robustness of the discovery of object categories.

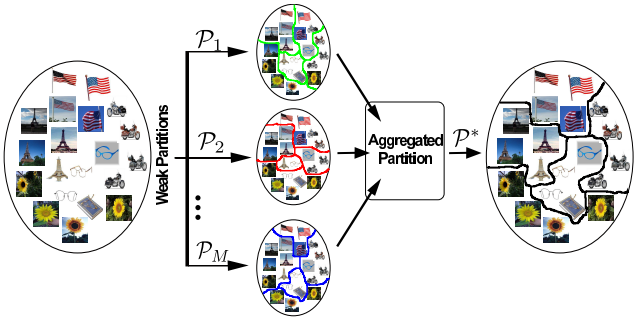

Figure 1. We show an overview of our AP approach. Given a group of images, we start building a set of M weak partitions $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$. For these partitions different descriptors can be used. Our AP model is able to build the partition P^* solving an optimization problem to maximize its purity.

Technically, we draw inspiration from consensus clustering algorithms [11, 14, 31], and propose an unsupervised robust feature-based partition ensembling model: the *Aggregated Partition* (AP) method. Consensus clustering algorithms consider the following problem: given a set of clusterings, find a single and robust clustering that agrees as much as possible with the input weak clusterings. We show in this paper how consensus clustering can be adapted to provide a natural solution for the unsupervised object discovery problem.

See Figure 1, we start from a set of weak input partitions for the unlabeled images, which can be obtained using different features (*e.g.* GIST [24], LBP [23], PHOG [1], LLC [35], BoW [5], etc.) in conjunction with different clustering algorithms (*e.g.* K-mean, Spectral Clustering, etc.). Our approach is able to build a robust final partition, the AP, which better discovers the objects. We formulate the computation of this AP as an optimization problem, where the disagreements of the AP with the input partitions are minimized. Furthermore, we then generalize the AP formula-

tion and derive a novel selective approach, the *Selective AP*, which is able to identify the subset of the provided weak input partitions that further minimizes the conditional entropy of the final partition. All this in a *completely unsupervised* manner.

The key contributions of our work can be summarized as follows: 1) to the best of our knowledge, our AP model (detailed in Section 3.1) is the first one to frame the unsupervised object discovery problem as a consensus clustering based approach; 2) with the formulation of the novel optimization problem for the Selective AP (see Section 3.2), we not only generalize the AP (and any other consensus clustering model), but introduce a new approach, based on a simple weighting mechanism for the input partitions, able to further improve the precision of the final partition, without any supervision; 3) experiments on three challenging and heterogeneous datasets, included in Section 4, show how our methods are able to consistently outperform competing methods, reporting state-of-the-art results; 4) we also provide the first experimental results on the challenging task of unsupervised fine-grained categorization in Section 4.2.3.

2. Related work

This Section reports on related work of object discovery and consensus clustering.

Over the last decade, supervised learning methods for image categorization have led to great progress, *e.g*. see the last benchmark results [7, 8]. These approaches normally need a strong supervision, with a large amount of labeled images so as to learn robust features and train the classifiers or detectors. However, to manually annotate these images is a time-consuming task, with its associated costs. Furthermore, this detailed labeling process is prone to undesirable user-specific biases and errors.

In order to overcome these problems, effective weakly-supervised (*e.g*. [4, 10, 25]) and completely unsupervised [6, 9, 16, 20, 26, 27, 30, 32, 36] object discovery techniques have been proposed.

In [36], we probably find the first unsupervised object discovery approach, which is built employing a constellation model. The use of probabilistic models has been also explored (*e.g*. [27, 30]).

Our models belong to the group of partition based solutions, in which very different approaches have been proposed [6, 9, 16, 20, 26, 32]. In [16], a spectral clustering based method for separating the objects from the background is introduced. In [32], a complete study of both clustering based and probabilistic methods is conducted, along with a clear experimental setup (which we follow in this paper) for enabling further comparisons. Objects are discovered by clustering image contours based on their intrinsic geometric properties, and spatial layouts, in [26]. In [20], the proposed approach leverages knowledge about previously learned categories to enable a more accurate discovery. Recently, the clustering by composition approach [9] has been introduced. It works by detecting statistically significant regions which co-occur between images. The image clusters are then defined as those in which each image can be easily composed using statistically significant pieces from other images in the cluster.

Similar to our idea of using weak input partitions, Dai *et al*. [6] propose the weak training sets. For each of these sets, a discriminative classifier is trained (a Linear SVM) to obtain a base partitioning of the image collection. Then, all these partitions are combined to an ensemble proximity matrix. The final categorization is completed by feeding this proximity matrix into a spectral clustering algorithm. Our method significantly differs from [6]. First, we do not need to train any ensemble of classifiers. We simply start from weak input partitions, which can be easily constructed by clustering methods. Second, while they build the final partition using spectral clustering with the distances encoded in a proximity matrix, as an extra step of their pipeline, we propose a more compact approach, where our final partition is obtained *directly* solving an optimization problem.

To formulate our AP approach for unsupervised object discovery, we leverage existing consensus clustering techniques [11, 14, 31]. While consensus clustering has been previously proposed to improve clustering robustness or to perform clustering of heterogeneous data, in our AP model we explore its benefits in a novel problem: the unsupervised object discovery. We propose to build a robust final partition of the image collection, given a set of initial weak partitions, which can be obtained by a set of different features extracted from the image. Our formulation for the AP follows a consensus clustering approach where, instead of maximizing the average mutual information of the final partition with all the input partitions [31], we choose the criterion in [14], which consist in minimizing the number of disagreements between the final and the weak input partitions. It is in the Selective AP, where we go further and generalize the formulation of the consensus clustering. Our novel Selective AP introduces an approach which is able to identify the subset of weak input partitions that minimizes the conditional entropy of the final partition. Specifically, we propose a weighting approach for the partitions, where the weights assigned encode the contribution of the input partitions to the final partition. Our experiments reveal that the Selective AP systematically outperforms the AP.

Finally, simply note that consensus clustering and stable clustering [13, 18, 33] are very different techniques, although sometimes in the literature stable clustering is named consensus clustering (*e.g*. [21]). Both families of approaches can be used as model selection strategies for clustering (*e.g*. for automatically selecting the number of clusters). However, essentially, the problems of consensus clus-

tering and stable clustering are different. For the clustering stability, one first selects a clustering algorithm X, and then the idea is to use perturbed versions of the input data, to find the best solution using a stability metric. This clustering stability models have some applied limitations, which have been recently highlighted in [29]. On the other hand, in the consensus clustering problem that we follow in this paper: a) the data is fixed, so no perturbations are needed; b) the final partition is not based on a stability criterion, but in a consensus criterion; c) we are not restricted to use only one clustering algorithm, actually, our models can simultaneously work with different clustering algorithms and features.

3. Unsupervised Object Discovery

3.1. Aggregated Partition

If we are given a set of N unlabeled images $\mathcal{S} = \{I_1, I_2, \ldots, I_N\}$, where each image belongs to one of the K predefined categories, our goal is then to separate the different object classes. We consider that each image I_i can be characterized by a set of F different feature types, $\{\mathbf{f}_i^{(1)}, \mathbf{f}_i^{(2)}, \ldots, \mathbf{f}_i^{(F)}\}$, where $\mathbf{f}_i^{(f)} \in \mathbb{R}^{D_f}$, and D_f is the dimensionality of feature type f. So, for each feature type f, we can build the set $\mathcal{F}_f = \{\mathbf{f}_1^{(f)}, \mathbf{f}_2^{(f)}, \ldots, \mathbf{f}_N^{(f)}\}$, containing the features for all the images in $\mathcal{S}$. As we show in the experiments, our models generalize to any image representation, from Bag of Words (BoW) [5], to descriptors such as GIST [24].

With these F different feature sets $\mathcal{F}_f$, we can proceed to build the set of *weak* input partitions used by our AP approach. We define the set of partitions $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$, where M is the number of weak input partitions, each of which groups the images using a particular clustering algorithm. Note that our approach generalizes to any clustering algorithm and feature combination, so $M \neq F$. For example, we can build a weak partition, using as features the concatenation of BoW and GIST descriptors, and the clustering algorithm K-means.

Given this set of weak input partitions $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$, our objective is to build a final partition, *i.e*. the *Aggregated Partition (AP)* P^*, which betters discovers the objects in the image collection $\mathcal{S}$ (see Figure 1). For doing so, we propose to leverage consensus clustering techniques [11, 14, 31], which consider an equivalent problem: given a set of clusterings, seek a final clustering that shares the most information with the original clusterings.

We proceed to formulate the computation of our AP $\mathcal{P}^*$ as a consensus clustering problem. To build the optimal combined clustering, we adopt the criterion in [14], which consist in minimizing the number of disagreements between the final and the input partitions.

So, following [14], we formulate an optimization problem where, given a set of M weak input image partitions $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$, the objective is to build the partition P^* that minimizes the total number of disagreements with the M *input* partitions,

$$\arg\min_{P^*} \frac{1}{M} \sum_{m=1}^{M} d(P^*, P_m)\,, \tag{1}$$

where $d(P^*, P_m)$ measures the dissimilarity between image partitions P^* and P_m as the number of pairs of images on which the two partitions disagree. Note that solving Eq. (1), we do not impose any constraint on the number of clusters of P^*, *i.e*. the number of categories discovered is automatically determined by the optimization process.

The objective function in Eq. (1) represents a difficult combinatorial optimization problem, where an exhaustive search results unapproachable. In order to solve Eq. (1), we proceed to reduce the problem of building the AP to a graph partitioning problem. We first have to transform the given cluster labels for the images into a suitable graph representation, where the edge weights encode the dissimilarities in Eq. (1). We consider the images $I_i \in \mathcal{S}$ as vertices of a connected, undirected and weighted graph denoted by $\mathcal{G} = (\mathcal{S}, \mathbf{E})$, where $\mathbf{E}$ is a symmetric non-negative affinity matrix $\mathbf{E} = [e_{ij}] \in \mathbb{R}^{N \times N} (e_{ii} = 0)$. We define the weight $e_{ij} \in [0, 1]$ of edge (i, j) as the fraction of weak partitions that assign the pair of images I_i and I_j into different clusters. The AP can be now formulated as the partition P^* that minimizes the following function,

$$\arg\min_{P^*} \sum_{l(I_i,P^*)=l(I_j,P^*)} e_{ij} + \sum_{l(I_i,P^*)\neq l(I_j,P^*)} (1 - e_{ij})\,, \tag{2}$$

where, $l(I_i, P^*)$ represents the cluster label assigned by the partition P^* to the image I_i. This way, if the AP $\mathcal{P}^*$ places images I_i and I_j in the same cluster, it will disagree with $M \times e_{ij}$ of the original partitions, while if they are separated into different clusters, $\mathcal{P}^*$ will disagree with the remaining $M \times (1 - e_{ij})$. In our implementation, Eq. (2) is solved following the *ALGCOMPLETE* graph partitioning algorithm introduced in [2].

3.2. Selective AP

If there is no a priori information about the *relative importance* of the individual input partitions, then a reasonable goal for the AP is to seek a clustering that shares the most information with *all* the original clusterings. However, this is not optimal. This becomes evident when one inspects the purity of the different input partitions: some of them discover the objects better than others. With the Selective AP we introduce an approach, generalizing the AP, able to identify the subset of the provided input partitions which contribute to obtain a final partition whose conditional entropy

is further minimized, improving the results of the standard AP. Specifically, we propose a weighting approach for the input partitions, where the weights encode the relative importance of the input partitions to the final partition.

Let us first define some notations. Given a partition P_i, we define the weighting operation wP_i, where $w \in \mathbb{N}$, as obtaining a set of w copies of partition P_i, *i.e.* $wP_i = \{P_{i_1}, \ldots, P_{i_w}\}$. With this weighting mechanism we are able to discard a partition, *i.e.* when $w = 0$, or to increase its relative importance by introducing multiple copies of it into our approach.

The Selective AP is again a fully unsupervised approach which can be formulated as follows. As for the AP, we are given the set of M weak image partitions $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$. The Selective AP algorithm starts iterating. At each iteration t, one of the weak partitions in $\mathcal{P}$ is chosen, and the algorithm *assumes* that it encodes the ground truth labels. We identify this partition as $P_t \in \mathcal{P} : t \in \{1, M\}$. We fix the number of iterations to the number of weak input partitions, this way we use all the input partitions. So, for iteration t, we build the set of partitions $\mathcal{P}^{(t)} = \mathcal{P} \setminus P_t$. Note that the size of set $\mathcal{P}^{(t)}$ is $M - 1$.

The objective of the algorithm is now to identify the combination of weights $\mathbf{w}^{(t)} \in \mathbb{N}^{M-1}$, for the partitions in $\mathcal{P}^{(t)}$, that generates the AP $P^{*(t)}$ whose conditional entropy is minimum, considering P_t as the ground truth partition. This objective is formulated as an optimization problem.

If $\mathbf{w}^{(t)} = (w_1, \ldots, w_{M-1})$, and $\mathcal{P}^{(t)} = \{P_1^{(t)}, P_2^{(t)}, \ldots, P_{M-1}^{(t)}\}$, we define the following weighted set of partitions

$$< \mathbf{w}^{(t)}, \mathcal{P}^{(t)} >= \{w_1 P_1^{(t)}, \ldots, w_{M-1} P_{M-1}^{(t)}\} . \quad (3)$$

We denote by $|\mathbf{w}^{(t)}|$ the total length of the weighted set $< \mathbf{w}^{(t)}, \mathcal{P}^{(t)} >$. For each weighted set $< \mathbf{w}^{(t)}, \mathcal{P}^{(t)} >$, we can define its associated AP $P^{(t)*}$ as

$$\arg\min_{P^{(t)*}} \frac{1}{M-1} \sum_{m=1}^{M-1} w_m d(P^{(t)*}, P_m^{(t)}) . \quad (4)$$

The optimization problems defined by equations (4) and (1) are equivalent, so the same solver is used now for Equation (4).

The final step of the iteration consist in finding the optimal weight vector $\mathbf{w}^{(t)}$. Formally, this is described by the following optimization problem,

$$\arg\min_{\mathbf{w}^{(t)}} H(P_t|P^{(t)*}) + |\mathbf{w}^{(t)}| \text{ s. t. } 0 < w_m < M . \quad (5)$$

$H(P_t|P^{(t)*})$ is the conditional entropy measured considering the *ground truth* category labels in P_t and the labels obtained by the AP $P^{(t)*}$. We follow the classical formulation proposed in Information Theory: $H(X|Y) = \sum_{y \in Y} p(y) \sum_{x \in X} p(x|y) log(\frac{1}{p(x|y)})$. It gives us the average amount of uncertainty that remains in the true class $P^{(t)*}$ given the instances estimated topic or cluster label P_t. The term $|\mathbf{w}^{(t)}|$ of Eq. (5) can be considered as a L_1 sparsity term, which causes most weights w_m to be 0, providing a compact set $< \mathbf{w}^{(t)}, \mathcal{P}^{(t)} >$. Eq. (5) is solved using a genetic algorithm for integer optimization. Note that for each weight vector evaluated by Eq. (5), a partition $P^{(t)*}$ has to be obtained solving Eq. (4). In summary, in each iteration t, the optimal weight vector $\mathbf{w}^{(t)}$ is computed solving Equation (5), which considers partition $\mathcal{P}_t$ as the ground truth.

Algorithm 1 Selective AP

Require: Weak input partitions $\mathcal{P} = \{P_1, P_2, \ldots, P_M\}$
Ensure: Definitive partition P^*
1: **for** t from 1 to M **do**
2: Select partition P_t
3: Build $\mathcal{P}^{(t)} = \mathcal{P} \setminus P_t$
4: Solve Equation (5) to compute each $\mathbf{w}^{(t)}$
5: **end for**
6: Build matrix $\hat{\mathbf{W}}$
7: Compute $\mathbf{w}^*$ using Equation (6)
8: Build the optimized weighted set $< \mathbf{w}^*, \mathcal{P} >$
9: Obtain the final partition P^* solving Eq. (7)
10: **return** P^*

Once all the iterations have finished, the weight vectors $\mathbf{w}^{(t)}$ have to be consolidated. First, all weight vectors $\mathbf{w}^{(t)}$ are combined into an ensemble weight matrix $\hat{\mathbf{W}}$, of size $M \times M$. Each row in $\hat{\mathbf{W}}$ is an *augmented* weight vector $\hat{\mathbf{w}}^{(t)}$, which incorporates the weights in $\mathbf{w}^{(t)}$, and inserts a zero weight at column t, which corresponds to the selected partition $\mathcal{P}_t$ during the iteration: $\hat{\mathbf{w}}^{(t)} = (w_1, \ldots, w_{t-1}, 0, w_{t+1}, \ldots, w_{M-1})$. Matrix $\hat{\mathbf{W}}$ is used to consolidate all the weights received by each of the weak input partitions.

In order to obtain the final weight vector $\mathbf{w}^* \in \mathbb{N}^M$ from $\hat{\mathbf{W}}$, we solve the following equation,

$$\mathbf{w}_i^* = \arg\max_i \text{hist}(\hat{\mathbf{W}(i,:)}) , \quad (6)$$

where $\text{hist}(\hat{\mathbf{W}(i,:)})$ is the histogram of weights values for the column i of matrix $\hat{\mathbf{W}}$. Intuitively, Equation (6) gives us the most voted integer weight, *i.e.* the consensus weight, for each weak input partition after the iterations. Once this final weight vector $\mathbf{w}^*$ has been obtained, the last step of our Selective AP algorithm consists in computing the definitive partition P^*, solving Eq. (1), but now using the optimized weighted set $< \mathbf{w}^*, \mathcal{P} >$, as follows,

$$\arg\min_{P^*} \frac{1}{|\mathbf{w}^*|} \sum_{m=1}^{|\mathbf{w}^*|} d(P^*, P_m) . \quad (7)$$

We summarize the whole optimization process in Algorithm 1. Note that for the especial case in which $\mathbf{w}^* = (1, 1, \ldots, 1)$, the Selective AP is equivalent to the AP.

4. Experiments

4.1. Experimental Setup

4.1.1 Datasets

We report experimental results for three challenging problems: 1) unsupervised scene discovery with the 15-Scenes dataset [19]; 2) unsupervised object discovery with the Caltech-256 database [17]; and 3) unsupervised fine-grained categorization with the Caltech-UCSD Birds-200-2011 dataset [34]. With the Caltech-256, we follow the experimental setup described by Tuytelaars *et al*. in [32] for unsupervised object discovery. We use the first subset of 20 categories proposed in [32]. For the rest of datasets we use all the images provided. The 15-Scenes dataset contains 15 scene categories with both indoor and outdoor environments, 4485 images in total. The Caltech-UCSD Birds-200-2011 dataset organizes 11788 images in 200 subcategories of birds.

4.1.2 Features and clustering algorithms

Note that our approaches do not impose any constraint either on the feature type or on the clustering algorithm to be used to build the weak input partitions. Furthermore, any combination of features and partition algorithms can be integrated.

With respect to the clustering algorithms, we evaluate K-means and Spectral Clustering (SC) [22]. We choose these two algorithms due to their excellent performance in unsupervised object discovery: they report state-of-the-art results in [32].

With respect to the features, we integrate: GIST [24], LBP [23], PHOG [1], LLC [35], BoW [5] and Spatial Pyramids of BoW (SP-BoW) [19]. Specifically, when using the GIST descriptors all images are first scaled to a size of 256×256. For the LBP features we use the uniform version. The PHOG is built from a two-layer pyramid and computing the derivatives in 8 directions. For the LLC representation, SIFT descriptors are extracted from patches densely located by every 8 pixels. During LLC processing, we train a codebook with 1024 bases, and only the approximated LLC is used (the number of neighbors is set to 5 with the shift-invariant constraint).

For the Caltech-256, we use the original BoW representation provided in [32][1]. We choose the one that obtained the best results: a BoW with a vocabulary size of 3000 (using SIFT descriptors). We refer to this representation as BoW-3000. For the 15-Scenes dataset we use a SP-BoW representation. We follow a dense sampling strategy: SIFT descriptors of 16×16 pixel patches computed over a grid with spacing of 8 pixels. A BoW of 200 visual words (with L_2 normalization) is computed. Then, we construct the SP-BoW representation using a pyramid of two levels.

We also report results combining different features (*e.g*., $P_{\text{BoW+GIST}}$, which means that we use BoW and GIST features to obtain the partitions).

4.1.3 Experimental settings

We follow the experimental validation for unsupervised object discovery proposed by Tuytelaars *et al*. [32]. An input database is given, which is composed of images belonging to a fixed number of categories. The dataset provides ground truth information against which the results obtained by the unsupervised methods can be evaluated quantitatively.

In a nutshell, the experiments with our approach consist of the following steps. We select a particular dataset and compute M different weak input partitions P_m. During clustering, we fix K for all the input partitions. The AP P^* is then built solving the optimization problems described, and evaluated against the ground truth data. In [6] and [32], the only parameter that is assumed to be known in advance is the number of categories which are present in each dataset. However, our approach is fully unsupervised: we do not fix the number of categories to be discovered. This is automatically identified by the optimization process of both the AP and the Selective AP models.

Note that we do not perform any parameter tuning during the experiments. Each experiment is repeated 10 times, and its average performance is reported. We use the evaluation metrics previously proposed by others: in the Caltech-256 the conditional entropy (CE) [32], and in the 15-Scenes the purity (P) proposed in [6]. For the new problem we propose with Caltech-UCSD Birds-200-2011, we choose to report both CE and P.

Note that our AP approach may discover more clusters than is known in the ground truth. Both the P and CE, get better and better when the number of discovered objects increases. But this is due to over-fitting rather than discovering a good partition. In order to solve this problem, we strictly follow the evaluation protocol described in [32]. It uses an oracle to assign each discovered topic to its best known class, and then evaluates the resultant assignments using P or CE. By doing so, we can establish a fair comparison between all the approaches.

[1] All image representation used in [32] can be downloaded from `http://homes.esat.kuleuven.be/~tuytelaa/unsupervised.html`

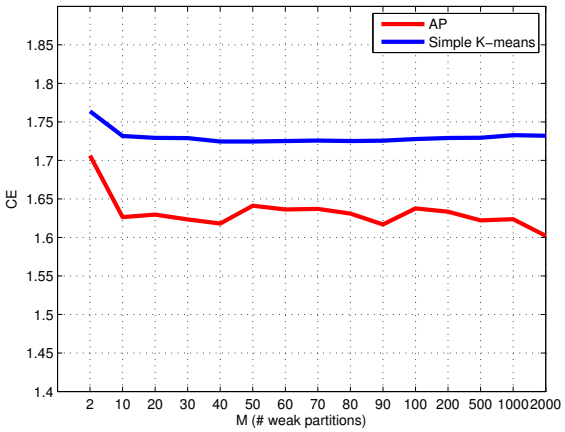

Figure 2. CE (the lower the better) of the discovered categories in the Caltech-256 as a function of M.

4.1.4 Competing methods

We compare our methods with the state-of-the-art approaches in [6, 12, 15, 30, 32]. Additionally, we compare the performance of the AP P^* with the performance of the weak input partitions P_m, which define our baselines. The AP performance must be always better than this baseline to corroborate our hypothesis. This comparison is a very challenging yardstick, because, as it is concluded in [32], when dealing with images with a single object category, the partitions obtained with simple clustering-based methods systematically report the best results.

4.2. Results

4.2.1 Unsupervised Object Discovery: Caltech-256 Experiments

Let us start using the Caltech-256 for examining the influence of our unique input parameter M, *i.e.* the number of input weak partitions, on the performance of our method. We compare the CE reported by the AP, with the mean CE obtained by all the weak input partitions. For this evaluation, to build the input partitions, we use K-means (with random initialization) and the BoW-3000 features. Figure 2 shows that the CE decreases pretty fast with M, and then stabilizes quickly (for $M > 10$). Therefore, the AP approach shows a considerable robustness against its parameter. We can also conclude that the AP actually benefits from incorporating multiple input partitions. Figure 2 also reveals that the AP always outperforms the simple clustering methods.

After this study of the influence of parameter M, we simply fix $M = 100$, *i.e.* we report results always using 100 input partitions (*e.g.* $100 \times \mathcal{P}_{\text{BoW}}$ means that we use 100 weak input partitions, using BoW features). For the baselines, we report the average of the CE for the 100 partitions.

Table 1. CE (the lower the better) of the AP and Selective AP approaches and baseline methods in the Caltech-256.

	AP	**Selective AP**	**Baseline**
Features	SC	SC	SC
$100 \times P_{\text{BOW}}$	1.8	1.8	1.8
$100 \times P_{\text{GIST}}$	1.78	1.77	1.83
$100 \times P_{\text{PHOG}}$	2.0	1.74	2.04
$100 \times P_{\text{LLC}}$	1.92	1.81	1.89
$100 \times P_{\text{LBP}}$	2.42	2.37	2.47
$100 \times P_{\text{BOW+GIST+PHOG+LLC+LBP}}$	1.51	**1.43**	1.52

Table 2. Comparison with the state-of-the-art.

Methods	**Caltech-256** - (CE)	**15-Scenes** - (P)
LDA [32]	1.99	–
NMF [32]	2.00	–
L2-LEM-χ^2[32]	1.58	–
AP (ours)	1.51	54%
Selective AP (ours)	**1.43**	54%
PLSA [30]	–	29.34%
RIM [15]	–	38.40%
AP [12]	–	44.24%
EnPar [6]	–	**61.49**%

The best CE in [32] is obtained when SC (L2-LEM-χ^2) is used, so, for our evaluation in this benchmark, we also report the results of our methods when SC input partitions are used.

We proceed to evaluate the performance of the AP for all the features proposed and their combinations. Table 1 shows all the quantitative results. First, our experiments show that the performance of both the AP and the Selective AP is always better than the baselines, for all the features. This confirms the convenience of using our models instead of single clustering based methods. We also observe that the *best performance has been obtained combining multiple features to build the input partitions*. Furthermore, our results outperform the state-of-the-art results reported in [32] (see Table 2). Finally, Table 1 also reveals that the Selective AP always outperforms the results reported by the AP. This is remarkable, confirming that the selective strategy designed, based on the weighting of input partitions, is able to automatically penalize those partitions that do not contribute to improve the final solution.

We additionally show qualitative results for our best approach in Figure 3. Both the CE and the number of images assigned to each topic discovered are shown in the first row. Interestingly, our approach seems to be able to discover a finer granularity than expected, *e.g.* splitting motorbikes with uncluttered background from motorbikes with cluttered background, or airplanes on the ground and airplanes in the sky. Some discovered topics obtain a CE < 0.2.

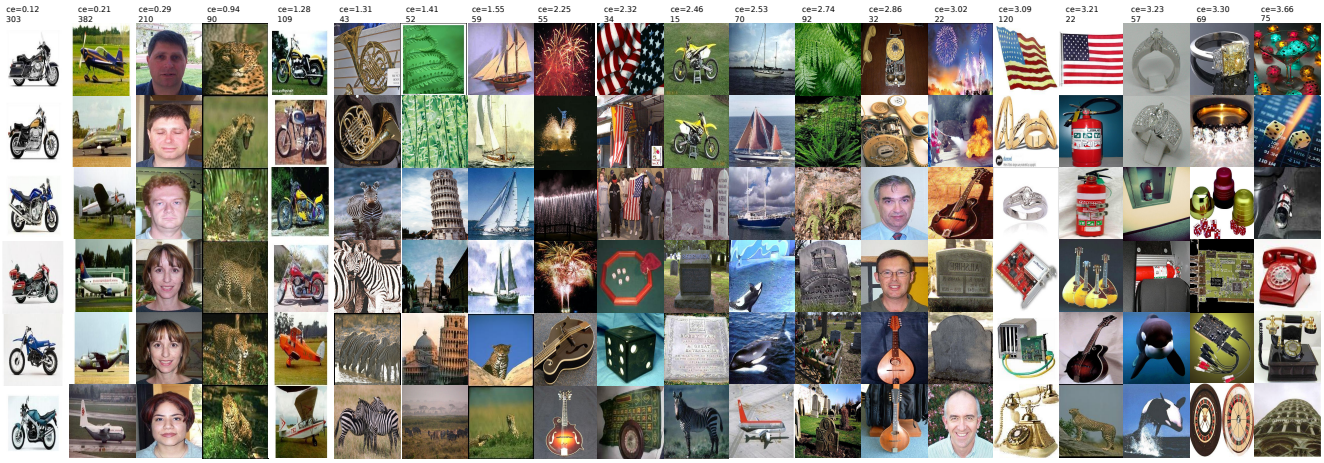

Figure 3. Qualitative results for the Selective AP method on the Caltech-256. 6 random images for each of the 20 categories discovered.

4.2.2 Unsupervised Scene Discovery: 15-Scenes Experiments

The unsupervised discovery of scenes is a very challenging problem. Only Dai *et al*. [6] have previously reported results for this problem. Here we follow their experimental setup, using the 15-Scenes dataset [19].

We report in Table 3 the results obtained by our approaches for all the features. Our best purity is of 54%, for a Selective AP built with $100 \times P_{\text{SP-BOW+GIST+PHOG+LLC+LBP}}$ weak partitions. Note that this time we report our results using the K-means clustering algorithm for the computation of the input partitions. Using SC and the same features we obtain a slightly worse performance of 53%. In general, the results and conclusions are consistent with the ones previously reported using the Caltech-256: a) We again observe that both the AP and the Selective AP approaches always outperform the baselines; b) The Selective AP obtains better results than the AP for most of the features, although their best performances coincide in this dataset.

We now compare our performance with state-of-the-art results. Table 2 reveals that our models outperform the competing methods [30, 12, 15]. The margin of improvement is considerable. The best results for this dataset have been reported in [6], where: a) the number of categories to be discovered is fixed; and b) parameter tuning is used (both manual and using [38]) for the winner approach, based on an ensemble of 1000 classifiers. However, recall that our models are fully unsupervised (we even do not fix the number of categories) and that we avoid any kind of parameter tuning. In spite of this generality, our methods markedly outperform the previous approaches, and closely compete with [6].

We finally show qualitative results for our best configuration in Figure 4, where both the purity and the number of images assigned to each topic are shown in the first row. We observe that there is not too much confusion splitting between indoor and outdoor scenes, and that some topics obtain a purity $> 80\%$.

Table 3. Purity (the higher the better) of the AP and the Selective AP approaches and baseline methods in the 15-Scenes dataset.

	AP	**Selective AP**	**Baseline**
Features	K-means	K-means	K-means
$100 \times P_{\text{SP-BOW}}$	0.31	0.29	0.29
$100 \times P_{\text{GIST}}$	0.43	0.43	0.41
$100 \times P_{\text{PHOG}}$	0.29	0.30	0.29
$100 \times P_{\text{LBP}}$	0.29	0.29	0.27
$100 \times P_{\text{LLC}}$	0.51	0.52	0.49
$100 \times P_{\text{SP-BOW+GIST+PHOG+LLC+LBP}}$	**0.54**	**0.54**	0.52

4.2.3 Unsupervised fine-grained categorization: Caltech-UCSD Birds-200-2011 experiments

Finally, we introduce in this paper a new challenging problem which has not been previously explored, to the best of our knowledge: *i.e*. the unsupervised fine-grained categorization problem. We propose the following experimental setup using the Caltech-UCSD Birds-200-2011 dataset [34]. This database organizes 11788 images in 200 subcategories of birds. Any method has to organize the images in a *fully unsupervised manner*. All images must be used in the evaluation, to report both the CE and P of the final partition. We here report the mean performance after 10 trials of a Random Assignment (RA) strategy, where each image is randomly assigned to a subcategory with uniform probability. RA provides a sanity check in that other methods should always perform better.

To establish further comparisons, we report the performance of our best method, the Selective AP. To compute the weak input partitions, we have decided to use the K-means algorithm. It is computationally efficient, and its performance is similar to the one reported by SC partitions. For the features, this time we only use the LLC representation.

Figure 4. Qualitative results for the AP on the 15-Scenes. We show 6 random images for each of the categories discovered.

Table 4. CE and P for the unsupervised subcategory discovery problem in the Caltech-UCSD Birds-200-2011 dataset.

	Selective AP		**Baseline**		**RA**	
Features	CE	Purity	CE	Purity	CE	Purity
$100 \times \mathcal{P}_{\text{LLC}}$	**4.8**	**7.3** (%)	5.5	6.6 (%)	5.6	4.3 (%)

The rest of features, while effective for generic object categorization, result in a large loss of finer details that are important for differentiating fine-grained object classes, as it is concluded in [37].

Results are shown in Table 4. Our Selective AP outperforms the results reported by both the baseline and RAN. Our main conclusion with respect to the challenging unsupervised fine-grained categorization problem is that there is still room for improvement for fully unsupervised methods. An important contribution of this paper is to introduce this analysis for establishing further comparisons, with the clear experimental setup proposed.

5. Conclusion

This paper proposes novel strategies to perform a robust feature-based partition ensembling for discovering object categories. From image partitions, obtained with different descriptors and clustering algorithms, our AP model is able to find a more robust and precise partition, using consensus clustering techniques. Overall, our solutions demonstrate that by aggregating the different feature-based partitions, we are able to combine their strengths to improve the accuracy and robustness of the object discovery.

We conclude that the AP models offer promising solutions for the unsupervised object discovery problem. Their performance has been thoroughly evaluated on a variety of challenging datasets and features. The results obtained confirm that our methods are able to consistently outperform the baselines, reporting state-of-the-art results. We publicly release the code for the AP methods to reproduce the results in the paper.

Acknowledgements

This work has been supported by the project TEC2013-45183-R (Ministry of Economy and Competitiveness), and the projects of the DGT SPIP2014-1468 and SPIP2015-1809.

References

[1] A. Bosch, A. Zisserman, and X. Muñoz. Image classification using random forests and ferns. In *ICCV*, 2007.

[2] M. Charikar, V. Guruswami, and A. Wirth. Clustering with qualitative information. In *IEEE Symposium on Foundations of Computer Science*, pages 524–533, 2003.

[3] M. Cho, S. Kwak, C. Schmid, and J. Ponce. Unsupervised object discovery and localization in the wild: Part-based matching with bottom-up region proposals. In *CVPR*, 2015.

[4] R. G. Cinbis, J. Verbeek, and C. Schmid. Multi-fold MIL training for weakly supervised object localization. In *CVPR*, 2014.

[5] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004.

[6] D. Dai, M. Prasad, C. Leistner, and L. Van Gool. Ensemble partitioning for unsupervised image categorization. In *ECCV*, 2012.

[7] J. Deng, W. Dong, R. Socher, L.-J. Li, and L. Fei-Fei. ImageNet: A large-scale hierarchical image database. In *CVPR*, 2009.

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes (VOC) challenge. *IJCV*, 88(2):303–338, 2010.

[9] A. Faktor and M. Irani. Clustering by composition – unsupervised discovery of image categories. In *ECCV*, 2012.

[10] R. Fergus, Y. Weiss, and A. Torralba. Semi-supervised learning in gigantic image collections. In *NIPS*, 2009.

[11] X. Z. Fern and C. E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *ICML*, 2004.

[12] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.

[13] C. Galleguillos, B. Babenko, A. Rabinovich, and S. Belongie. Weakly supervised object localization with stable segmentations. In *ECCV*, 2008.

[14] A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1):4, 2007.

[15] R. Gomes, A. Krause, and P. Perona. Discriminative clustering by regularized information maximization. In *NIPS*, 2010.

[16] K. Grauman and T. Darrell. Unsupervised learning of categories from sets of partially matching image features. In *CVPR*, 2006.

[17] G. Griffin, A. Holub, and P. Perona. Caltech-256 object category dataset. Technical Report 7694, Caltech, 2007.

[18] T. Lange, M. Braun, V. Roth, and J. Buhmann. Stability-based validation of clustering solutions. In *NIPS*, 2002.

[19] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[20] Y. J. Lee and K. Grauman. Object-graphs for context-aware category discovery. In *CVPR*, 2010.

[21] S. Monti, P. Tamayo, J. Mesirov, and T. Golub. Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning*, 52(1):91–118, 2003.

[22] A. Y. Ng, M. I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. In *NIPS*, 2001.

[23] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE TPAMI*, 24(7):971–987, 2002.

[24] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *IJCV*, 42(3):145–175, 2001.

[25] M. Pandey and S. Lazebnik. Scene recognition and weakly supervised object localization with deformable part-based models. In *ICCV*, 2011.

[26] N. Payet and S. Todorovic. From a set of shapes to object discovery. In *ECCV*, 2010.

[27] J. Philbin, J. Sivic, and A. Zisserman. Geometric lda: A generative model for particular object discovery. In *BMVC*, 2008.

[28] S. Schulter, C. Leistner, P. M. Roth, and H. Bischof. Unsupervised object discovery and segmentation in videos. In *BMVC*, 2013.

[29] Y. Senbabaoğlu, G. Michailidis, and J. Z. Li. Critical limitations of consensus clustering in class discovery. *Scientific Reports*, 4(6207), 2014.

[30] J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman. Discovering objects and their location in images. In *ICCV*, 2005.

[31] A. Strehl and J. Ghosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Machine Learning Research*, 3:583–617, 2002.

[32] T. Tuytelaars, C. Lampert, M. Blaschko, and W. Buntine. Unsupervised object discovery: A comparison. *IJCV*, 88(2):284–302, 2009.

[33] U. von Luxburg. Clustering stability: An overview. *Foundations and Trends in Machine Learning*, 2(3):235–274, 2010.

[34] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, Caltech, 2011.

[35] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[36] M. Weber, , M. Welling, and P. Perona. Towards automatic discovery of object categories. In *CVPR*, 2000.

[37] B. Yao, G. Bradski, and L. Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *CVPR*, 2012.

[38] L. Zelnik-Manor and P. Perona. Self-tuning spectral clustering. In *NIPS*, 2004.