published in: International Journal of Pattern Recognition and Artificial Intelligence

International Journal of Pattern Recognition and Artificial Intelligence (World Scientific Publishing Company)
– Received: 26 January 2015 –
– Accepted: 22 September 2015 –
– Published: 15 December 2015 –
– Revised Version: 08 February 2016 –

An Iterative Partitioning-based Method for Semi-supervised Annotation Learning in Image Collections

Rene Grzeszick, Gernot A. Fink

Department of Computer Science, TU Dortmund University
Dortmund, North Rhine-Westphalia, 44227, Germany
{*rene.grzeszick,gernot.fink*}*@udo.edu*
http://patrec.cs.tu-dortmund.de

Labeling images is tedious and costly work that is required for many applications, for example, tagging, grouping and exploring of image collections. It is also necessary for training visual classifiers that recognize scenes or objects. It is therefore desirable to either reduce the human effort or infer additional knowledge by addressing this task with algorithms that allow for learning image annotations in a semi-supervised manner. In this paper, a semi-supervised annotation learning algorithm is introduced that is based on partitioning the data in a multi-view approach. The method is applied to large, diverse image collections of natural scene images. Experiments are performed on the 15 Scenes and SUN databases. It is shown that for sparsely labeled datasets the proposed annotation learning algorithm is able to infer additional knowledge from the unlabeled samples and therefore improve the performance of visual classifiers in comparison to supervised learning. Furthermore, the proposed algorithm outperforms other related semi-supervised learning approaches.

Keywords: scene recognition, image classification, semi-supervised learning

1. Introduction

Labeling images is very time consuming manual work. It allows for grouping or tagging images in private, web or stock photo collections, but is also a requirement for organizing large scale image archives which can, for example, be found in many museums and libraries. Furthermore, labeling is an important prerequisite for visual recognition tasks. The recognition of visual scenes or objects relies on large sets of training images that describe the respective problem domain.

In the past, researchers have pointed out that the availability of data is a crucial issue in computer vision[21]. Since then the availability of image databases has improved and large scale image sets, like PASCAL VOC, SUN or ImageNet have been created. However, a key problem remains: despite the availability of web images and crowd sourcing projects, the creation of these large scale datasets is still very time consuming. For example, collecting and labeling the 14 million images in the ImageNet database took five years. Crowd sourcing projects allow for distributing the labeling effort[20], but for the annotator itself the labeling process has not become more efficient. It is worth mentioning that the labeling time is not only dependent on the level of annotation (e.g. outlines, bounding boxes or presence/absence labels) but often also on the number of different categories. For example,

published in: International Journal of Pattern Recognition and Artificial Intelligence

labeling the presence or absence of an object takes about 1 second per object[cf. 24]. For ambiguous multi-class issues the decision making process will obviously take longer than for easy ones. Besides the tremendous amount of work and the costs associated with it there are two more problems. First, in the work of Torralba and Efros it could be shown that image databases are highly biased[27]. The training of a visual classifier will therefore require the creation of a task specific training set. Here, typical details can be the categories to be classified, the viewpoint, the image type or its resolution. Second, in non-academic settings, companies are often required to possess the data and therefore create newly annotated datasets for specific tasks.

Therefore, it is a huge benefit if the problem of labeling can be tackled by algorithms that allow for learning annotations. Such algorithms would either allow to minimize the amount of manual work or to infer more knowledge using additional information from unlabeled samples with the same amount of labeling operations.

In order to efficiently learn annotations, semi-supervised learning algorithms can be used. Common approaches propagate annotations from a few labeled samples to a large set of unlabeled samples. There is a wide variety of algorithms including, for example, mixture models, transductive SVMs and graph-based methods[cf. 33,34]. A major problem of semi-supervised learning is that incorrect matching during the learning phase can lead to degradation in classifier performance. If the model assumptions do not match the problem domain or the initially labeled samples do not describe the problem domain accurately, these algorithms do not work well.

The unlabeled samples that are required to efficiently apply semi-supervised algorithms can be collected as additional samples, e.g. from web sources[cf. 13,25]. Moreover, the noisy labels that are often associated with images from the web can be leveraged for learning annotations in a different dataset or re-ranking those noisy labels[9]. A typical example of a collection of web images with noisy labels is the 80 million tiny image database[26].

In other scenarios all data is collected beforehand, but without any labels. For example, object, person or face recognition often requires tremendous amounts of training data and it is easy to collect this data, but mostly without any labels. In this paper a scenario, where the data is collected beforehand but without any labels is considered. The goal is to not only learn annotations but to train a visual classifier based on the learned annotations.

2. Related Work

Intuitively, semi-supervised learning strategies such as graph-based label propagation introduced by Zhou et al.[30] can be used for labeling a dataset in a semi-supervised manner. The initially labeled dataset that is required for label propagation can be created by random selection or a steered sample selection strategy. Usually, these semi-supervised methods aim toward labeling the complete dataset and do not consider the goal of training a classifier afterwards.

In the work of Ebert et al.[7] it could be shown that active learning is able to improve the quality of semi-supervised learning algorithms on datasets that are collected beforehand. The learning strategies extend the graph-based label propagation[30] and are used

published in: International Journal of Pattern Recognition and Artificial Intelligence

for learning annotations in image collections for visual object recognition. However, the improvements from adding samples by active learning imply that several labels are incorrectly inferred by the label propagation and that either a single feature representation is not necessarily descriptive enough to guarantee a correct propagation or, if possible, a feature selection strategy should be employed for creating the initial labeled dataset that is used for the graph-based label propagation.

A similar problem has been addressed by Richarz et al.[22] in the field of handwritten character recognition. Character images pose a perfect task for semi-supervised annotation learning, since there is a large amount of samples which are very similar if they are from the same writer. In contrast to classical semi-supervised learning scenarios there is neither an initially labeled set, nor a stream from which the data is collected in a fixed order. Instead the samples that need to be labeled are selected from the unlabeled dataset and presented to an annotator, which is similar to the active learning concept presented by Ebert et al.[7] but neither initial labels that can be propagated nor any knowledge about the problem domain are considered. For each handwritten character image different feature representations are computed and an ensemble decision is used for propagating the labels in order to be robust against incorrect matches. The advantage of using different feature representations in ensembles has already been shown by Zhou[31].

Two methods for learning annotations are presented by Richarz et al.[22]: in the first approach, the feature representations are clustered so that one label has to be requested for each cluster in a given representation by judging all samples. In the second approach, randomly drawn samples are labeled and used as query samples for a retrieval task. Samples that receive the same annotation in different feature representations are assigned a label. It could be shown that the manual labeling effort for characters written by a single person can be reduced to less than 1%. The advantage of the clustering based method is that it forms comparably large partitions of the data and is therefore especially suitable to identify groups of samples that can easily be associated with one class. However, samples that are more difficult to assign to one specific class, e.g. those on class borders, will not be labeled by this approach although they carry information that might be crucial for recognition tasks with a high intra class variability. The main advantage of the retrieval based approach is that only one label needs to be requested for a sample in all feature representations. The retrieval lists that are used for propagating the labels allow for either a very coarse or a very fine grained analysis of the feature space. It is, however, very difficult to determine an appropriate parameterization. Thi especially applies for feature representations that are not necessarily normalized and a problem domain where no training data is available.

In this paper, a new iterative partitioning-based learning approach is presented that allows to overcome the shortcomings of these approaches. It combines key-principles of joint, active and ensemble learning: A cluster based ensemble learning approach that partitions the dataset is combined with a refinement step. It uses a multi-view cluster and distance evaluation in order to select samples in regions of the feature space that are difficult to recognize within the ensemble. 1) Both clustering and refinement step are combined in one joint learning method that iteratively refines the partitioning of the feature space. 2) All labels are requested from a human in the loop using sample selection that is either

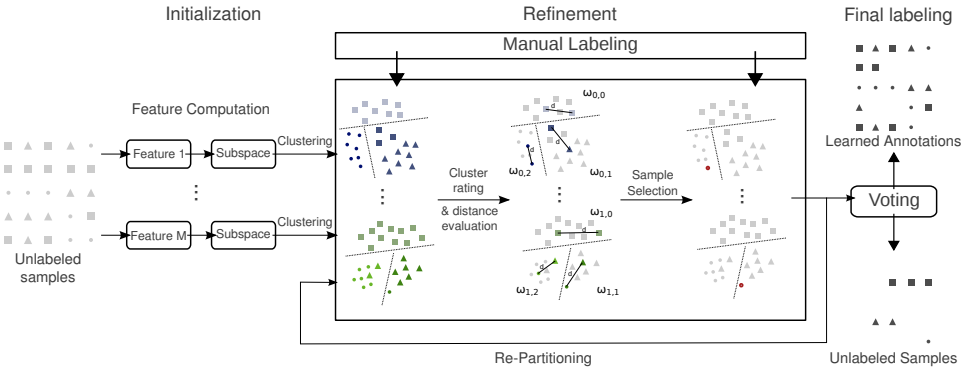

Fig. 1. Overview of the iterative partitioning-based learning approach. The method is initialized by clustering the data independently in multiple feature representations which is then refined based on a multi-view cluster scoring and distance evaluation. Using manual labeling for a few representatives, labels for are inferred for the unlabeled samples, resulting in a partially labeled sample set.

based on the clustering or the retrieval of difficult samples. The system becomes more informed the more samples are labeled. 3) It is assumed that either the propagation of labels will introduce some errors or that labels given by the human in the loop are not always completely accurate. The cluster ensemble introduces multiple views on the data and allows for being more robust against such errors by a voting before propagating the data to the remaining unlabeled samples.

The method is applied to collections of natural scene images, which are much more complex than, for example, the handwritten character images and pose a difficult task for semi-supervised learning. An extensive evaluation is given, comparing the proposed approach to supervised learning, the annotation learning methods from by Richarz et al.[22] and semi-supervised label propagation by Zhou et al.[30]. It will be shown that for the same amount of labeling operations the proposed method outperforms supervised learning on sparsely labeled datasets and also shows better performance than the other semi-supervised learning approaches.

3. Iterative partitioning-based learning

In the following section the new iterative partitioning-based learning approach is presented. It is illustrated in Fig. 1 and consists of three main steps:

1) The learning algorithm is initialized by computing different feature representations and reducing their dimensionality by computing a subspace representation. These representations are then clustered independently.

2) The cluster partitions are then refined by iteratively adding new partitions in regions where little to no knowledge about the samples can be inferred from the clustering process. Here, a multi-view cluster and distance evaluation is computed for finding these regions.

published in: International Journal of Pattern Recognition and Artificial Intelligence

3) The centroids are manually labeled and used in order to infer labels for the unlabeled samples, resulting in a partially labeled sample set. Samples where no class can be assigned with high certainty remain unlabeled.

The learned image annotations are then used for training a visual classifier. The main goal of the semi-supervised learning approach is to improve the recognition rate of this classifier so that it requires only a low number of annotated samples and performs better than a classifier that is trained in a supervised manner.

Initialization

First, M feature representations are computed in order to implement an ensemble that offers different views on the data. Typical feature representations, like Bag-of-Features, LBP histograms or the GIST, are very high dimensional and not necessarily ideal for creating partitions as well as for computing meaningful cluster scores[cf. 19]. A lower dimensional representation, which is non-sparse, is desirable since cluster evaluation methods basically measure the overlap between different cluster partitions. It is also more efficient when clustering large data collections.

There are two alternative approaches for creating a lower dimensional subspace. On the one side transformations that compute new features by computing linear combinations of a given feature representation and uncovering a latent structure. On the other side feature selection methods which attempt to find the most relevant dimensions of a given feature representation. Feature selection methods are often used in the context of subspace clustering. They assume that not all dimensions carry information that is relevant for clustering the data and that interpretability is a crucial issue[cf. 19]. However, for typical image classification representations that include a feature learning step it can be assumed that all dimensions are relevant. For example, in Bag-of-Features representations the features are learned from the data by clustering local feature descriptors in order to obtain a codebook of meaningful representatives. Some of these representatives might be correlated and it is therefore more appropriate to compute a transformation[19].

Here, using a Singular Value Decomposition is proposed (also referred to as LSI[3]). It is computed in each feature representation m and the dimensionality of the data is reduced:

$$\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^T = \mathrm{svd}(\boldsymbol{X_m}) \tag{1}$$

where $\boldsymbol{X}_m$ is approximated by $\hat{\boldsymbol{X}}_m$ by choosing the Z dimensions with the largest singular values from $\boldsymbol{\Sigma}$ and $\hat{\boldsymbol{x}}_{i,m}$ denotes the i^{th} feature in representation m. These subspaces are often referred to as topic spaces in the context of Bag-of-Words representations. It has been shown that up to a certain point topic space transformations allow for creating lower dimensional feature representations without loosing the descriptiveness [cf. 1,3].

The partitioning of the feature spaces is then initialized by applying clustering in each topic space independently. Note that any method that allows for partitioning the data in an unsupervised manner is applicable, e.g. Lloyd's[14], MacQueen's[15] or spherical k-means clustering[4]. It is however important that the clustering allows for determining a meaningful centroid. In every feature representation m, K partitions $\boldsymbol{\mathcal{Q}_{k,m}}$ of the data are created.

published in: International Journal of Pattern Recognition and Artificial Intelligence

Ideally, each partition is assigned a label based on its centroid. If the centroid does not match a sample, the sample closest to the centroid is annotated. Using these annotations a label matrix Y is created that assumes that the labels are propagated to all samples within the partition. Hence, the number of manual annotations is reduced to $M \cdot K$. For the i^{th} sample the labels assigned in the different feature representations are then represented as C-dimensional binary vectors

$$[Y_{i,m,1}, \ldots, Y_{i,m,C}]^T \in \{0,1\}, \quad m = 1, \ldots, M \tag{2}$$

where C is the number of classes. An entry of the vector is 1 if the label of the sample's partition equals c and 0 otherwise. Note that for an unknown number of classes the size of the vector can be iteratively increased. The initial regions consist mostly of samples that are easy to distinguish based on the feature representations. However, for very diverse classification problems such as natural scene recognition the majority of samples are rather difficult to assign to a class, because of the high intra class variability and ambiguities. Therefore, it is interesting to further explore the feature space by iteratively refining the partitioning.

Refinement

In order to find regions where it is useful to refine the partitioning two basic assumptions are made. First, partitions that are not well separated from other partitions are not meaningful. Second, the uncertainty is higher if a sample is far away from its centroid since the labeling is based on the partitions' centroid.

The first condition can be evaluated by measures such as the Dunn Index[6], the Silhouette Score[23] or the Davies Bouldin measure[2]. Here, a modified version of the Dunn Index that uses the preliminary label information for evaluating the partitions is proposed:

$$\mathrm{dunn}(k,m) = \frac{d_{\mathrm{inter}}(\mathcal{Q}_{k,m}, m)}{d_{\mathrm{intra}}(\mathcal{Q}_{k,m})} \tag{3}$$

where d_{intra} represents the intra partition distance and d_{inter} represents the distance between two partitions that have different labels:

$$d_{\mathrm{intra}}(\mathcal{Q}) = \frac{1}{|\mathcal{Q}|} \sum_{\boldsymbol{a} \in \mathcal{Q}} \max_{\boldsymbol{b} \in \mathcal{Q}} d(\boldsymbol{a}, \boldsymbol{b}) \tag{4}$$

$$d_{\mathrm{inter}}(\mathcal{Q}, m) = \frac{1}{|\mathcal{Q}|} \sum_{\boldsymbol{a} \in \mathcal{Q}} \min_{\boldsymbol{b} \notin \mathcal{Q};\, \boldsymbol{Y}_{a,m,:} \neq \boldsymbol{Y}_{b,m,:}} d(\boldsymbol{a}, \boldsymbol{b}) \tag{5}$$

Hence, partitions that are very close and assigned the same label do not influence the Dunn Index negatively. For example, the samples of one class might be represented by a few partitions. If these partitions are close together they are ignored with respect to the distance d_{inter}. In order to combine the knowledge from the different feature representations, the label vector for a sample is multiplied with the respective Dunn scores

published in: International Journal of Pattern Recognition and Artificial Intelligence

$$\lambda_{\mathrm{dunn}}(i) = \max_{c} \frac{1}{M} \sum_{m} \mathrm{dunn}(q(\hat{\boldsymbol{x}}_{i,m}), m) \cdot \boldsymbol{Y}_{i,m,:} \tag{6}$$

where $q(\cdot)$ assigns a sample to its partition $\boldsymbol{\mathcal{Q}_{k,m}}$. By computing the maximum of the vector sum, samples that belong to partitions with high Dunn scores and where the different feature representations agree on a label will be assigned a high combined score. Samples that belong to partitions with low Dunn scores that do not agree on a label will be assigned a low combined score. This combination can be seen as an active learning process that incorporates context information by taking into account the neighborhood of a sample in form of its respective partition and reliable knowledge from the already labeled samples in the different feature representations.

The second assumption is evaluated by a distance function that is computed in all representations:

$$\lambda_{\mathrm{dist}}(i) = \frac{1}{M} \sum_{m} \frac{1}{1 + d(\hat{\boldsymbol{x}}_{i,m}, q(\hat{\boldsymbol{x}}_{i,m}))} \tag{7}$$

Here, d can be any kind of distance function, e.g. the Euclidean or the cosine distance. Both terms are then combined in a target function that is evaluated for all samples:

$$i^* = \operatorname*{argmin}_{i} \lambda_{\mathrm{dunn}}(i) + \lambda_{\mathrm{dist}}(i) \tag{8}$$

The sample with the minimum rating is chosen to be labeled by an annotator and added as an additional centroid. The Dunn term causes the minimization to focus on regions where the ensemble does not agree and the partitions are not well separated. The distance term focuses on samples that are farther away from their respective centroids. As a result the samples computed by this refinement strategy will focus on either outliers in regions where no knowledge can be inferred or on class borders where the ensemble disagrees. The idea of the selection is very similar to the informativeness in SVM based semi-supervised learning approaches[cf. 32].

The partitions are then updated by re-assigning all samples to the centroids including the one added by minimizing eq. 8. The update can be computed by the same update rule as used in the clustering process, i.e. for k-means problems each sample is assigned to the closest centroid yielding a new Voronoi tesselation. Changing the partitioning also includes an update of the label matrix $\boldsymbol{Y}$. Finally, new Dunn scores are computed using the assignment of the samples to the respective centroids. This can efficiently be solved by updating only those scores that are influenced by the re-partitioning: the distance d_{intra} can only change for those partitions that formerly contained samples that are assigned to the newly added partition. The distance between the partitions is more difficult, as it contains the constraint that for all samples $\boldsymbol{a}$ of a partition $\mathcal{Q}$ the nearest sample $\boldsymbol{b}$ needs to have a different label. Given that only partitions with labels of a set of classes $\mathcal{C}$ have changed by the re-partitioning, including the newly assigned label, it is possible to constrain the computation. If a partition has not been changed by the refinement, then its distance d_{inter} cannot change if its label $c \notin \mathcal{C}$. The nearest sample $\boldsymbol{b}$ with a different label would be

the same, no matter what the actual label of $\boldsymbol{b}$ is. In the evaluations these constraints were able to improve the computation time of the Dunn scores by $50\% - 80\%$, depending on the number of partitions and the number of classes in the dataset. For a large number of partitions the re-computation of the distances between the partitions could be further constrained by applying the triangle inequality[8] on the centroids. Otherwise the process can also be speed-up by computing approximate Nearest Neighbors[16].

Typically, the partitions that are computed by adding new centroids in the refinement steps are rather small. It will, however, be shown that they contain valuable information for training a visual recognizer.

The process is iterated until either all scores are sufficiently high or a fixed number of labeling operations have been performed. This allows to reduce the number of further manual annotations to K' as the labels for the additional centroids in the partitioning step can be propagated through all feature representations.

Final label propagation

In a final voting step the previously computed label matrix $\boldsymbol{Y}$ that has been updated during the refinement of the partitions is used to assign labels to the so far unlabeled samples of a dataset. Applying majority voting for the i^{th} sample results in an ensemble decision for a specific class label

$$y_i^{max} = \begin{cases} \underset{c}{\operatorname{argmax}} \sum_{m=1}^{M} Y_{i,m,c} & \text{if } \max\limits_{c} Y_{i,m,c} > \frac{M}{2} \\ -1 & \text{otherwise ,} \end{cases} \tag{9}$$

where -1 is a rejection class. Annotations are learned only for samples where the class membership is determined with a majority agreement. If the goal is a very precise labeling (e.g. tagging images) other voting schemes, like unanimity voting, might be considered. Here, the samples that were assigned a label by the semi-supervised learning algorithm will later on be used for training a visual classifier and it is beneficial to label more images although a few errors might be introduced by the majority voting. Samples where no majority was observed among the different setups are rejected by the learning process. This is an important difference to several semi-supervised learning algorithms that aim toward labeling all samples and therefore tend to make mistakes in regions of the dataset where not enough knowledge can be inferred from the manually labeled samples.

4. Evaluation

The evaluation of the method was performed on two different datasets. First, suitable features and parameters were determined on the 15 Scenes database[12]. Using the model assumptions that were derived from these experiments, the method was applied to the SUN scene recognition database[29]. To the best of our knowledge our approach and the works from Ebert et al.[7] and Fergus et al.[9] are the only efforts applying semi-supervised learning methods to train a visual classifier on such large scale image collections.

published in: International Journal of Pattern Recognition and Artificial Intelligence

Implementation details

In the following Lloyd's k-Means algorithm[14] is used for the clustering as it is widely recognized as a standard technique. As a result, the distance evaluation that is used for both conditions of the refinement step is based on the Euclidean distance measure. Accordingly, The re-partitioning is based on the Euclidean distance as well so that a new Voronoi Tesselation is formed in every iteration. In the following the effect of different topic space sizes, the improvement achieved by the refinement and varying numbers of labels in the refinement step will be evaluated.

4.1. *15 Scenes database*

The experimental setup is based on the evaluation protocol for supervised classification that has been introduced for the SUN database[29]: 200 samples per class (= 3,000 in total) were randomly chosen and considered unlabeled for evaluating the semi-supervised approaches. The remaining $1,485$ samples were used for testing the classifier. In all experiments the number of manual labeling operations is used as a free parameter. The unlabeled samples are then labeled by either manually labeling a fixed number of samples (supervised learning) or a semi-supervised learning approach. The labeling is simulated, assuming a perfect annotation for each sample (i.e. the partition centroids). For the supervised learning the samples are randomly drawn from the training set. Finally an SVM is trained based on the samples that were labeled in a semi-supervised manner. A five-fold crossvalidation was performed for all experiments.

4.1.1. *Feature selection*

Most importantly, for the semi-supervised approaches the features must represent the different categories well, which can be evaluated by a supervised classification experiment. In addition, it is desirable for the multi-view approach to have diverse representations of the data. Hence, common local and global image descriptors are considered. In this work features derived from Deep Neural Networks[(cf. 11)] are not considered. Although they achieve state-of-the-art results, training or just adapting them to a given task is typically done in a supervised manner and therefore requires tremendous amounts of data which is counteracting the proposed approach.

A description of the feature types and the results of a supervised experiment using 200 labeled samples per class is shown in Table 1. The classification is performed using an SVM. These results can also be seen as an upper bound for the results that are achievable in any semi-supervised setting. For the local image descriptors SIFT shows the best results. Spatial Visual Vocabularies are computed from the local SIFT descriptors. They are then aggregated in a Bag-of-Features representation[10]. With respect to the performance there is only a slight difference between the xy- and the radial-tiling, as well, as the very similar SIFT and HOG features. For the global descriptors the GIST descriptor yields the best results and even outperforms the LBP histograms on this task. An exhaustive evaluation of feature combinations has been performed, yielding that a combination of the two different

published in: International Journal of Pattern Recognition and Artificial Intelligence

Name	Description	Recognition rate
SIFT (xy tiling)	SIFT descriptors extracted on a dense grid with step size of 5px and bin sizes of $4, 6, 8, 10$px. The descriptors are quantized into a codebook of $1,000$ Spatial Visual Words[10] that introduce a 2×2 xy tiling (comparable to a pyramid[12], but with adaptive codebooks for each tile). This Bag-of-Features histogram is then represented by square rooted frequencies[28].	$82.1 \pm 1.1\%$
SIFT (radial tiling)	The same descriptors and representation as above, but radial-tiling with two circles.	$82.6 \pm 0.7\%$
HOG (xy tiling)	A histogram of Oriented Gradient descriptors extracted on a 5px grid with 3×3 regions of 8px size and 12 different orientations. The HOG features are also represented in a Bag-of-Features histogram with 2×2 xy tiling as has been done for the SIFT features.	$80.4 \pm 1.3\%$
LBP Histograms	A histogram of rotation invariant Local Binary Patterns[17]. At each pixel 12 comparison points are chosen on a circle with a radius of one. A pyramid scheme is built that computes an LBP histogram for 3×3 tiles and a histogram of the complete image is derived using max pooling.	$68.8 \pm 1.1\%$
SURF (xy tiling)	SURF descriptors extracted on a grid with step size of 5px. The SURF features are also represented in a Bag-of-Features histogram 2×2 xy tiling as has been done for the SIFT features.	$68.0 \pm 0.7\%$
GIST	Spatial Envelope representation[5,18]. The descriptor is computed using three channels (RGB), three scales and 12, 12 & 4 orientations.	$72.0 \pm 1.0\%$
Color/Intensities	RGB/gray scale intensities represented by mean, std. dev. and a color histogram with 16 bins per channel.	$25.7 \pm 1.4\%$
Tiny Img	Images scaled down to 32×32 pixels[26].	$18.4 \pm 3.2\%$

Table 1. Description and recognition rates (with std. dev.) of different feature representations (top: local, bottom: global) that are evaluated in an supervised experiment on the 15 Scenes dataset. For each class 200 labeled samples are used for training. Hence, 3,000 samples in total. Note that this experiment can also be seen as an upper baseline for the recognition rate of the given feature representation.

published in: International Journal of Pattern Recognition and Artificial Intelligence

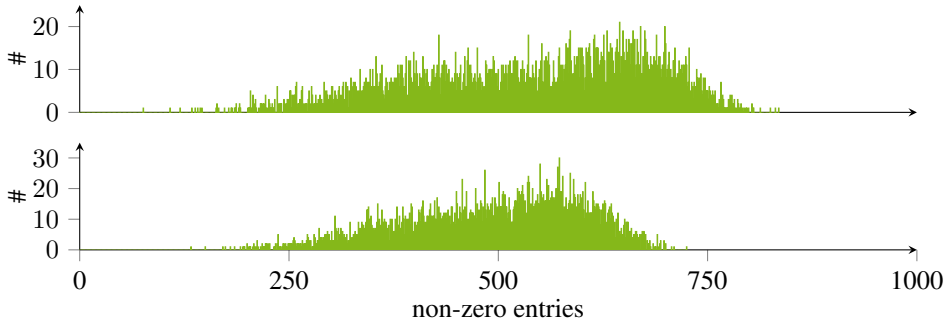

Fig. 2. Number of non-zero entries for Bag-of-Feature representations with a vocabulary size of 1,000 in the 15 Scenes dataset: (Top) SIFTxy (Bottom) SIFTrad.

Bag-of-Feature representations using SIFT features, as well as the GIST feature performs best. Hence, for the further experiments this setup is considered. At most three different feature representations have been chosen, as the number of labels in the initial clustering step increases with the number representations. Furthermore, at some point multiple feature representations either introduce redundancies or, if the feature representations are not very descriptive, noise is introduced. Note that a representation for the proposed method could also consist of a combination of different feature types, for example, by simple feature stacking.

4.1.2. *Topic space*

The size of the topic space is estimated by evaluating the sparsity of the dataset. Figure 2 shows that for a vocabulary size of 1,000 Visual Words as few as 70 dimensions are non-

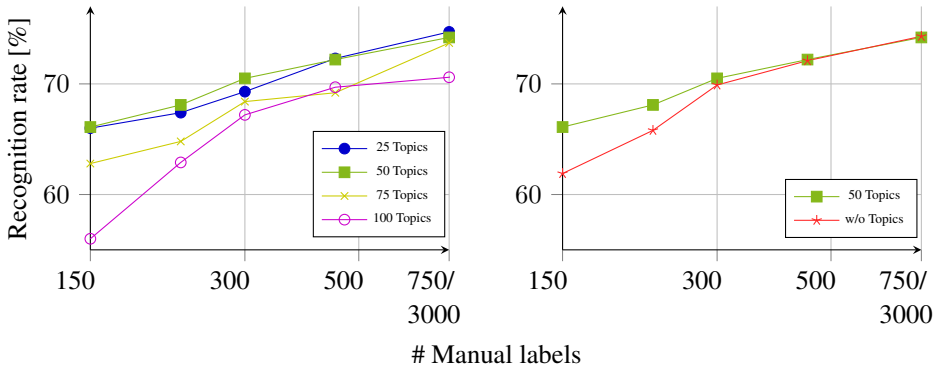

Fig. 3. Recognition rates of an SVM that is trained on samples that are labeled in a semi-supervised manner by the proposed iterative partition-based method on the 15 Scenes database. (Left) Different sizes for the topic space that is computed by LSI are compared. (Right) The topic space is compared to running the method without the dimensionality reduction by LSI.

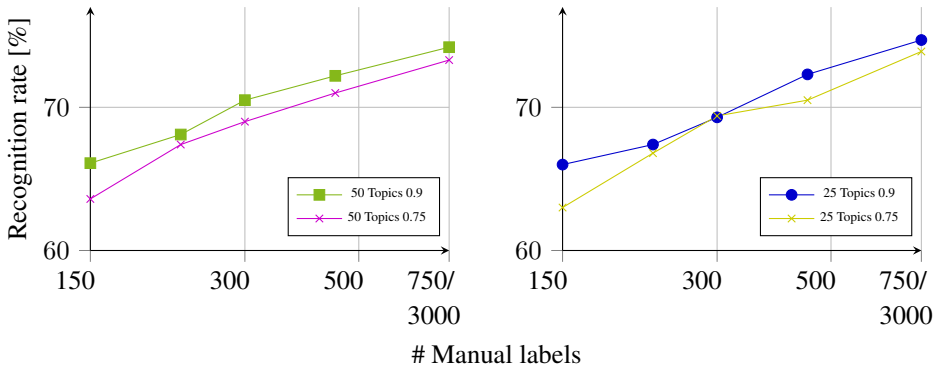

Fig. 4. Recognition rates of an SVM that is trained on samples that are labeled in a semi-supervised manner by the proposed iterative partition-based method on the 15 Scenes database. (Left) Initializations using 75% and 90% of the labeling operations for 50 topics computed by LSI. (Right) Initializations using 75% and 90% of the labeling operations for 25 topics computed by LSI.

zero. Taking also possible correlations into account, setups with $25, 50, 75$ and 100 topics are evaluated. Assuming that the clustering partitions the feature space in a meaningful manner, 10% of the labeling operations are used for refining the partitions. The number of labeling operations that are used for the refinement are further investigated in the next section. Fig. 3 shows that the the best results can be achieved with a low number of topics. Furthermore, it is shown that in comparison to an approach without using topics the recognition rate can be improved. However, as the number of labels increases, the feature space becomes well explored forming a set of smaller partitions. Hence, the advantage of the compactness of the feature space decreases with an increasing number of labeling operations.

4.1.3. *Initialization*

The number of labeling operations that need to be assigned to the iterative refinement are further investigated by evaluating the two best performing topic space sizes. In both cases 10% and 25% of labeling operations are used for the refinement. The results are shown in Fig. 4. It can be seen that a certain amount of labeling operations is required in the initial clustering in order for the method to be successful. Both configurations perform better when using 10% of all labeling operations in the refinement. Furthermore, the comparison to a pure cluster based method (CBL)[22] in the next section will show the necessity of refining the partitions. In the further experiments the size of the topic space is set to 50 and 10% of the labeling operations are used for refining the partitions.

4.1.4. *Method comparison*

In order to evaluate the performance of the iterative partitioning-based learning approach (PBL) it is compared to supervised classifier training and three semi-supervised ap-

published in: International Journal of Pattern Recognition and Artificial Intelligence

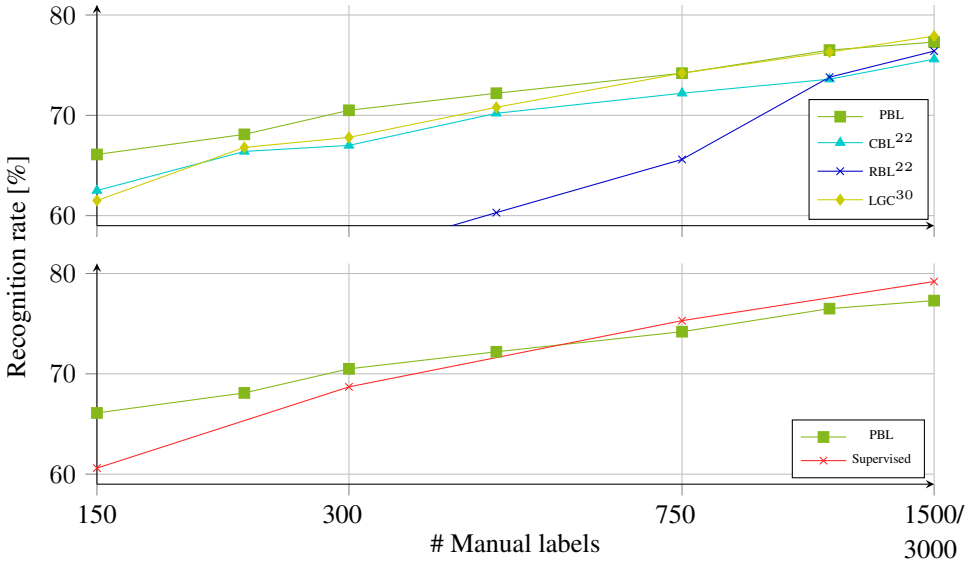

Fig. 5. Recognition rates of an SVM that is trained on samples that are labeled in a semi-supervised manner on the 15 Scenes database. (Top) Comparison with cluster and retrieval based labeling and semi-supervised label propagation with local and global consistency. (Bottom) Comparison with supervised classifier training.

proaches: the cluster and retrieval based labeling algorithms (CBL & RBL)[22] and graph based semi-supervised label propagation with local and global consistency (LGC)[30]. For the supervised classifier training a given number of samples is randomly chosen and labeled. For LGC these samples are used for initialization. Note that for all approaches there are the same training and test sets and no further unlabeled samples. This is a difference to other semi-supervised setups (e.g. the work of Ebert et al.[7]) where the semi-supervised approaches use an additional pool of unlabeled data, which however introduces a bias that favors semi-supervised approaches. In contrast our setting is more difficult since the semi-supervised algorithms have no prior knowledge about the classes or the sample distribution, whereas even a random drawing will roughly resemble a uniform sample distribution. In all cases the SIFT features with 2×2 xy tiling are used for training the final classifier.

The results are shown in Fig. 5. The proposed partition based learning approach outperforms the other semi-supervised algorithms and performs very well when the dataset is only sparsely annotated. Only with an increasing number of samples LGC is able to show similar performance when the initialization works well enough to propagate labels to all samples. In comparison with supervised classifier training there is a break-even point from which on supervised learning is more useful than semi-supervised learning. It can be observed at roughly 25% labeled samples within the training set. This is a good result considering that most semi-supervised setups work with a much larger set of unlabeled samples since it is very easy to obtain additional unlabeled samples.

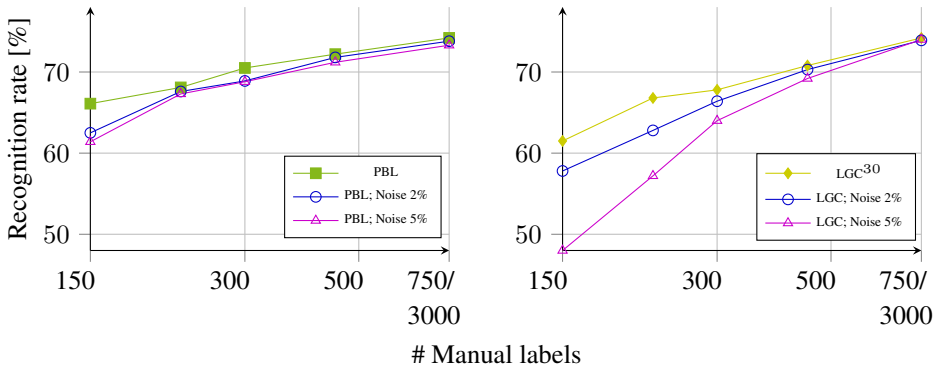

Fig. 6. Recognition rates of an SVM that is trained on samples that are labeled in a semi-supervised manner on the 15 Scenes database. Different percentages of noise have been added by randomly swapping the correct labels. (Left) Proposed iterative partition-based method. (Right) Semi-supervised label propagation with local and global consistency.

4.1.5. *Performance of labeling*

With respect to the labeling error it can be said that both the number of learned labels as well as the ratio of correctly labeled samples increase with the number of manual labeling operations. For 1500 of 3000 possible labeling operations, as shown at the right-most point of the evaluation in Fig. 5, the iterative partition-based learning approach labels $80.9 \pm 0.8\%$ of the samples with $88.0 \pm 0.6\%$ of the learned labels being correct. This also emphasizes why multiple feature representations are necessary, as the best single view (SIFT radial) labels only $64.5 \pm 0.7\%$ of the samples correctly.

4.1.6. *Robustness toward labeling noise*

The multi-view concept does not only allow for rejecting samples that cannot be labeled with sufficient reliability, but also adds some robustness toward labeling noise. Figure 6 shows the recognition rate of the proposed approach compared to two evaluations where artificial labeling noise has been introduced. Therefore, 2% and 5% of the labels have been swapped by randomly replacing them with another label, which seems a realistic range for the number of errors that can be made by a human annotator.

It can be seen that the results of the proposed method converge very fast toward the recognition rates without labeling errors and that increasing the noise hardly makes any difference. For the single view based LGC on the other hand it can be seen that increasing the noise strongly influences the recognition rates, especially if the number of manual labeling operations is low.

4.2. *SUN database*

In the following experiments the proposed method is evaluated on the SUN scene recognition database. With 397 categories and more than 100,000 images the SUN database is the

published in: International Journal of Pattern Recognition and Artificial Intelligence

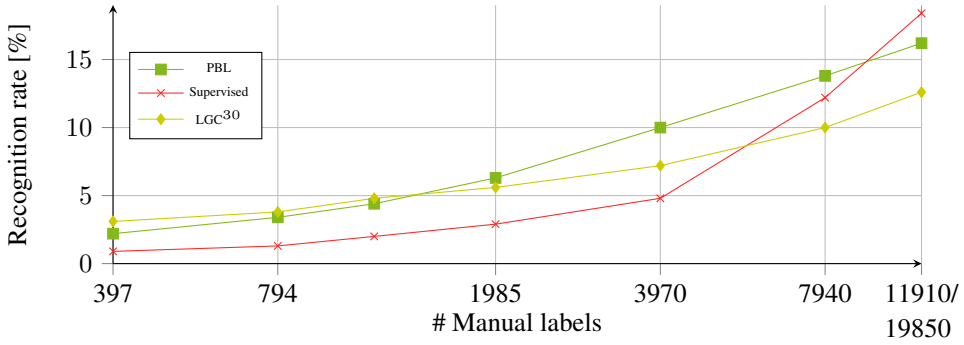

Fig. 7. Recognition rates of an SVM on the 397 scene categories of the SUN database. The proposed approach is compared to supervised classifier training and LGC.

largest benchmark for natural scene images. For each category 50 samples are randomly chosen for training and the remaining ones for testing so that for every category at least 50 samples remain in the test set. Hence, the training set contains $19,850$ samples in total. The results for a five-fold crossvalidation are shown in Fig. 7.

Note that the performance of the supervised training approach is very low for a standard classification method[cf. 29], which demonstrates that the benchmark is an extremely challenging task. Consequently, semi-supervised learning on this dataset is extremely difficult, too. Nevertheless, the results show that semi-supervised learning is also beneficial for challenging tasks. More sophisticated features or feature encodings would be able to improve the recognition of supervised as well as semi-supervised learning[cf. 29].

The semi-supervised learning methods perform much better since the low number of samples that are available when training a classifier in a supervised manner is not sufficient to distinguish the 397 categories. Here, the iterative partitioning-based learning approach shows good performance compared to supervised learning. Even with roughly 45% of the training set being labeled the classifier trained on the samples that have been annotated by the partitioning-based learning approach shows better recognition rates than the supervised learning. The method is also able to outperform LGC, which emphasizes that the refinement of the partitioning in regions without thorough knowledge about the samples is especially useful for training a classifier on samples that are learned in a semi-supervised manner.

4.3. *Discussion*

While the evaluation showed that the semi-supervised approach also performs well on challenging tasks that can hardly be solved by supervised methods, a main limitation of the approach are the underlying feature representations. If these do not distinguish the classes well it is impossible for any clustering method to find homogeneous clusters that contain samples from the same class and that do not introduce errors when inferring the label from the centroid. Better features will represent samples from the same class in compact regions

published in: International Journal of Pattern Recognition and Artificial Intelligence

of the feature space that are easier to identify. It will therefore be of interest to integrate features derived from Convolutional Neural Networks so that adapting the Network to a given task is solved jointly with the semi-supervised learning.

It should furthermore be noted is that the approach requires datasets that are completely unlabeled and that are typically too large to be labeled completely manually. The proposed approach is like most semi-supervised learning methods only beneficial if the dataset is only sparsely labeled and a sufficient amount of unlabeled samples is available for learning.

Given such a dataset, the proposed approach describes a general framework that can be combined with different feature representations that are suitable for the task at hand. Although different representations will result in feature spaces with different attributes these can be accounted for using an appropriate distance measure and clustering algorithm.

5. Conclusion

In this paper a novel iterative partitioning based learning approach has been introduced. The method is initialized by clustering the data independently in multiple, dimensionality reduced, feature representations. The cluster partitions are then refined based on a multi-view cluster and distance evaluation. The refinement of the feature space focuses on regions without thorough knowledge about the samples and, therefore, allows to find samples that are particularly interesting for training a visual classifier. Using manual labeling for a few representatives, labels are inferred for the unlabeled samples. The approach is robust against labeling errors due to an ensemble of multiple feature representations and context information that is included in the sample selection by taking into account the cluster assignment as well as top down knowledge from previously labeled samples. Samples for which no label can be inferred with high certainty remain unlabeled, yielding a partially labeled sample set.

The capabilities of the method have been demonstrated in an extensive evaluation. Suitable parameters for the method have been determined on the 15 Scenes database. The method has then been evaluated on the SUN scene recognition database. It has been demonstrated that the proposed approach outperforms similar semi-supervised learning algorithms. Furthermore, for sparsely labeled datasets the method shows better recognition rates than supervised classifier training. On the SUN dataset the proposed semi-supervised method performs better than the supervised case until more than 45% of the training set are labeled. Such scenarios are very realistic since it is easy to obtain unlabeled data for several tasks and the method could even be combined with approaches that distribute the labeling effort like crowd sourcing projects.

published in: International Journal of Pattern Recognition and Artificial Intelligence

References

1. A. Bosch, A. Zisserman, and X. Munoz, "Scene classification via pLSA," *Proc. European Conference on Computer Vision (ECCV)*, pp. 517–530, 2006.
2. D. L. Davies and D. W. Bouldin, "A cluster separation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 2, pp. 224–227, 1979.
3. S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman, "Indexing by latent semantic analysis," *Journal of the American society for information science*, vol. 41, no. 6, pp. 391–407, 1990.
4. I. S. Dhillon and D. S. Modha, "Concept decompositions for large sparse text data using clustering," *Machine learning*, vol. 42, no. 1-2, pp. 143–175, 2001.
5. M. Douze, H. Jégou, H. Sandhawalia, L. Amsaleg, and C. Schmid, "Evaluation of gist descriptors for web-scale image search," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2009, p. 19.
6. J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Journal of Cybernetics*, vol. 3, no. 3, pp. 32–57, 1973.
7. S. Ebert, M. Fritz, and B. Schiele, "Semi-supervised learning on a budget: scaling up to large datasets," in *Proc. Asian Conference on Computer Vision (ACCV)*. Springer, 2013, pp. 232–245.
8. C. Elkan, "Using the triangle inequality to accelerate k-means," in *ICML*, vol. 3, 2003, pp. 147–153.
9. R. Fergus, Y. Weiss, and A. Torralba, "Semi-supervised learning in gigantic image collections," in *Advances in neural information processing systems*, 2009, pp. 522–530.
10. R. Grzeszick, L. Rothacker, and G. A. Fink, "Bag-of-Features Representations using Spatial Visual Vocabularies for Object Classification," in *Proc. Int. Conf. on Image Processing (ICIP)*, 2013.
11. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.
12. S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 2169–2178.
13. Y. Liu, D. Xu, I.-H. Tsang, and J. Luo, "Textual query of personal photos facilitated by large-scale web data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 5, pp. 1022–1036, 2011.
14. S. Lloyd, "Least squares quantization in {PCM}," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.
15. J. MacQueen, "Some methods for classification and analysis of multivariate observations," in *Proc. of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, 1967, pp. 281–297.
16. M. Muja and D. G. Lowe, "Scalable nearest neighbor algorithms for high dimensional data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, 2014.
17. T. Ojala, M. Pietikäinen, and T. Mäenpää, "Gray scale and rotation invariant texture classification with local binary patterns," in *Proc. European Conference on Computer Vision (ECCV)*. Springer, 2000, pp. 404–420.
18. A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," *Progress in brain research*, vol. 155, p. 23, 2006.
19. L. Parsons, E. Haque, and H. Liu, "Subspace clustering for high dimensional data: a review," *ACM SIGKDD Explorations Newsletter*, vol. 6, no. 1, pp. 90–105, 2004.
20. P. Perona, "Vision of a Visipedia," *Proceedings of the IEEE*, vol. 98, no. 8, pp. 1526–1534, 2010.
21. J. Ponce, T. L. Berg, M. Everingham, D. A. Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, B. C. Russell, A. Torralba, and Others, "Dataset issues in object recognition," in

published in: International Journal of Pattern Recognition and Artificial Intelligence

Toward category-level object recognition. Springer, 2006, pp. 29–48.
22. J. Richarz, S. Vajda, R. Grzeszick, and G. A. Fink, "Semi-Supervised Learning for Character Recognition in Historical Archive Documents," *Pattern Recognition, Special Issue on Handwriting Recognition*, vol. 47, no. 3, pp. 1011–1020, 2014.
23. P. J. Rousseeuw, "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis," *Journal of computational and applied mathematics*, vol. 20, pp. 53–65, 1987.
24. O. Russakovsky, A. L. Bearman, V. Ferrari, and L. Fei-Fei, "What's the point: Semantic segmentation with point supervision," *ArXiv e-prints*, 2015.
25. F. Schroff, A. Criminisi, and A. Zisserman, "Harvesting image databases from the web," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 4, pp. 754–766, 2011.
26. A. Torralba, R. Fergus, and W. T. Freeman, "80 million tiny images: A large data set for non-parametric object and scene recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 11, pp. 1958–1970, 2008.
27. A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2011, pp. 1521–1528.
28. A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman, "Multiple kernels for object detection," in *Proc. IEEE Int. Conf. on Computer Vision (ICCV)*. IEEE, 2009, pp. 606–613.
29. J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba, "Sun database: Large-scale scene recognition from abbey to zoo," in *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2010, pp. 3485–3492.
30. D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf, "Learning with local and global consistency," *Advances in neural information processing systems*, vol. 16, no. 16, pp. 321–328, 2004.
31. Z.-H. Zhou, "When semi-supervised learning meets ensemble learning," *Frontiers of Electrical and Electronic Engineering in China*, vol. 6, no. 1, pp. 6–16, 2011.
32. Z.-H. Zhou and R. Jin, "Active Learning by Querying Informative and Representative Examples," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, no. 10, pp. 1936–1949, 2014.
33. X. Zhu, "Semi-supervised learning literature survey," *Technical report, Computer Science, University of Wisconsin-Madison*, 2006.
34. X. Zhu and A. B. Goldberg, *Introduction to Semi-Supervised Learning*. Morgan & Claypool Publishers, 2007.