Weakly Supervised Object Localization with Stable Segmentations

Carolina Galleguillos[1], Boris Babenko[1],
Andrew Rabinovich[1], and Serge Belongie[1,2]

[1] Computer Science and Engineering, University of California, San Diego
[2] Electrical Engineering, California Institute of Technology
{cgallegu,bbabenko,amrabino,sjb}@cs.ucsd.edu

Abstract. Multiple Instance Learning (MIL) provides a framework for training a discriminative classifier from data with ambiguous labels. This framework is well suited for the task of learning object classifiers from weakly labeled image data, where only the presence of an object in an image is known, but not its location. Some recent work has explored the application of MIL algorithms to the tasks of image categorization and natural scene classification. In this paper we extend these ideas in a framework that uses MIL to recognize and *localize* objects in images. To achieve this we employ state of the art image descriptors and multiple stable segmentations. These components, combined with a powerful MIL algorithm, form our object recognition system called MILSS. We show highly competitive object categorization results on the Caltech dataset. To evaluate the performance of our algorithm further, we introduce the challenging Landmarks-18 dataset, a collection of photographs of famous landmarks from around the world. The results on this new dataset show the great potential of our proposed algorithm.

1 Introduction

The goal of object categorization is to locate and identify instances of an object category within an image. This task is challenging in real world scenes since objects may vary in scale, position, and viewpoint; in addition, they may be surrounded by background clutter, occluded by other objects, and obscured by poor image quality. To model these sources of variability, traditional approaches to object categorization require large labeled data sets of fully annotated training images. Typical annotations in these "fully" labeled data sets provide masks or bounding boxes that specify the locations, scales, and orientations of objects in each training image. Though extremely valuable, this information is prone to error and is expensive to obtain. Without this information, however, traditional approaches to object categorization tend to learn spurious models of background artifacts, leading to lower accuracy during testing.

Some approaches for object categorization have successfully learned object models from weakly labeled data [1,2,3,4,5]. Weakly labeled training examples indicate which objects of interest are present in training images without specifying the pixels that are associated with them. From weakly labeled examples,

D. Forsyth, P. Torr, and A. Zisserman (Eds.): ECCV 2008, Part I, LNCS 5302, pp. 193–207, 2008.

the existing methods use standard techniques in statistical learning to model the essence of each category. Popular approaches include part-based models [1,6,7], region based methods [2,5] and latent models such as pLSA and LDA, with bag of visual words [3,4,8]. While they excel at exploiting correlations between different image patches, they suffer from computationally expensive inference and background noise that is learned as part of the category model.

Recently, Multiple Instance Learning (MIL) models have been applied to image categorization [9,10]. MIL permits weakly labeled images for training, but avoids the shortcomings of the methods mentioned above. In particular, MIL trains a discriminative classifier, rather than a generative model, which avoids complex inference procedures, and usually results in higher recognition accuracy. Although some of the previous works have applied MIL algorithms to the problem of object categorization, the focus has been on classifying images rather than localizing instances of objects in them.

Following this promising line of work we extend the current frameworks for MIL-based image categorization by adding object localization capabilities and improving image categorization accuracy. The main contribution of this paper is a novel object categorization framework that localizes objects in cluttered, real world scenes. Our method incorporates multiple stable segmentations and Bag-of-Features (BoF) image representation into a MIL framework, see Fig. 1 for an illustration. We demonstrate the efficiency and accuracy of our framework on two databases that present significant intra-class variation: Caltech 4 [11] and a landmark image database, Landmarks-18. The Caltech dataset, although highly popular in the computer vision community, is a rather artificial dataset, where objects often appear in isolation and with uniform backgrounds. The Landmarks-18 dataset on the other hand, is taken directly from common web albums and contains instances of popular landmarks in cluttered scenes with variable viewpoint, weather, and illumination (see Fig. 3).

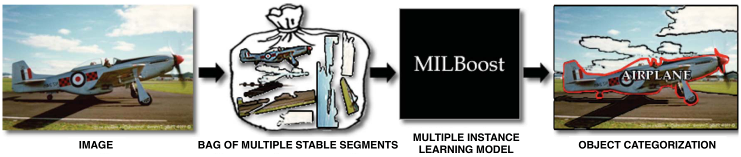


IMAGE
BAG OF MULTIPLE STABLE SEGMENTS
MILBoost
MULTIPLE INSTANCE LEARNING MODEL
AIRPLANE
OBJECT CATEGORIZATION

Fig. 1. An input image containing an airplane is processed through a segmentation-based object recognition engine obtaining a collection of stable segments. The bag of segments is represented as bags of features and then fed into the MIL algorithm. Finally, the model classifies each segment, localizing the object in the image.

2 Related Work

2.1 Multiple Instance Learning

The MIL problem was first introduced by Dietterich *et al.* [12] for the problem of drug discovery. In this domain it is desired to predict properties of a drug molecule using the molecule's shape as an input to the classifier. Each molecule, however, can take on multiple shapes, and it is not known during training which shape is responsible for certain properties of the training molecules. Formally, traditional supervised learning requires training data $\{(x_1, y_1), ..., (x_N, y_N)\}$, $x_i \in \mathcal{X}, y_i \in \mathcal{Y}$ where $\mathcal{X}$ is the input space and $\mathcal{Y}$ is the output space. On the other hand, MIL is able to learn from training data of the form $\{(X_1, y_1), ..., (X_N, y_N)\}, X_i = \{x_{i1}, x_{i2}...\}, x_{ij} \in \mathcal{X}, y_i \in \mathcal{Y}$. For example, in the drug discovery problem each X_i is a molecule, and each x_{ij} is one particular shape of that molecule. The MIL problem is defined only for binary classification, so we will assume that $\mathcal{Y} = \{1, 0\}$. In this setting X_i is an unordered set of inputs (often called a "bag"), and the bag label y_i follows the rule $y_i = \max_j(y_{ij})$. Notice that although true instance labels y_{ij} are assumed to exist, the learning algorithm does not have access to them during training. The goal of a MIL algorithm is then to learn a classifier function $H : x \rightarrow \{0, 1\}$, that acts on instances. Various algorithms have been proposed for solving this problem [12,13,14], and in this paper we chose the MilBoost algorithm by Viola *et al.* [14].

2.2 MIL and Image Categorization

In recent years MIL algorithms have attracted the attention of the computer vision community because they provide a way of training classifiers with weakly labeled data. These models have tried to address various problems such as scene classification, image annotation, and image and object categorization. In natural scene classification, several models have successfully classified images into predefined semantic concepts (categories) using MIL. For example, Maron *et al.* applied the Diverse Density (DD) algorithm to the problem of natural scene classification [15]. Trying to solve the same problem, Zhou [16] introduced MIML, where each training example is associated with not only multiple instances but also multiple class labels. Both methods consider classification on a bag (image) level only, and do not take advantage of the instance classifier returned by a MIL algorithm. Similarly, in image annotation, MI-SVM [17] and ASVM-MIL [18] algorithms use variations of the popular SVM algorithm modified to solve MIL. In the problem of image categorization many MIL approaches have been shown to outperform traditional supervised object categorization models. The DD-SVM [19] model uses the DD algorithm to select prototypes and an SVM to classify bags in the prototypes' space. Bi *et al.* [20] and MILES [9] embed bags into a feature space defined by instances and use a 1-norm SVM to construct bag classifiers. Recently, the results of ConMIL [10] showed that modeling interdependencies between instances can improve accuracy in instance and bag classification.

With respect to object categorization, ConMIL and MILES have achieved competitive results relative to traditional approaches. In these methods, object categorization is framed as binary classification which tries to separate object instances from background clutter. Although these algorithms achieve good performance on an image level, their models often capture parts of the background in the positive images. While the backgrounds of positive images provide clues in image classification (*e.g.* an airplane will often co-occur with a sky background), models that capture this information would have trouble in correctly localizing the objects of interest.

3 Multiple Instance Learning Using Stable Segmentations

The problem of learning an object classifier from weakly labeled data can be elegantly framed as multiple instance learning. During training it is known for each image whether a certain object category is present, but the exact location of that object is unknown. If we split an image $\mathcal{I}_i$ into J multiple regions or segments $\{s_{i1}, s_{i2} ..., s_{iJ}\}$, we can assume that one of the segments contains the object of interest (we will discuss different strategies for doing this shortly). For each image we are given a category label $y_i = \{c_1, c_2, ..., c_C\}$; however, since the MIL problem is defined only for binary classification, we will train our classifiers in a one versus all manner. If we define $y_{ik} \in \mathbf{1}(y_i = c_k)$ to be a binary label indicating the presence of category k in image i, we can train C different classifiers. For each category k, we train a classifier $H^k : s \to \{1, 0\}$ using the training data set $\{(\mathcal{I}_1, y_{ik}), ...\}$. In practice, since our problem is multi-class it is more useful for us to also obtain the probability of the segment containing an object category k, $p(c_k|s)$. The boosting algorithm for MIL developed in [14] provides us with an effective way of learning these functions, and in the section below we briefly review this algorithm.

3.1 MilBoost

The MilBoost algorithm developed by Viola et al. in [14] uses the gradient boosting framework of Friedman [21]. The classifier learned by a boosting framework has the form $H^k(s) = \sum_{t=1}^{T} \alpha_t^k h_t^k(s)$ where each h_t^k is a weak classifier and α_t^k is a scalar weight. We use a simple decision stump as the weak classifier as is done in much of the boosting literature [22,23][1]. To get a binary label from this classifier we could use $\text{sign}(H^k)$, but recall that we would also like to retrieve a probability. Instead, we use the sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$ to define this probability as follows:

$$p(c_k|s) = \sigma\Big(\sum_{t=1}^{T} \alpha_t^k h_t^k(s)\Big). \tag{1}$$

The loss function we optimize is the binomial log likelihood over bags:

$$\mathcal{L}^k(H^k) = -\sum_i \Big(y_{ik} \log(p_{ik}) + (1 - y_{ik}) \log(1 - p_{ik})\Big), \tag{2}$$

where $p_{ik} = p(c_k|\mathcal{I}_i)$ is the probability that image i contains an object from category k. Note that it is impossible to compute the likelihood over segments

[1] Using a decision stump as a weak classifier also results in feature selection during training.

because the labels for these are unknown during training. Finally, we need to define the image probability p_{ik} in terms of the probabilities of its segments. Ideally we would define this as $p_{ik} = \max_j p(c_k|s_{ij})$. Since the boosting framework uses gradient descent to learn, however, this definition would cause problems due to the non differentiable max operator. Instead Viola et al. suggest using the Noisy-OR model as follows:

$$p(c_k|\mathcal{I}_i) = 1 - \prod_j \big(1 - p(c_k|s_{ij})\big). \tag{3}$$

Having all of these terms defined, we can now use the gradient boosting framework to learn each weak classifier h_t^k in a greedy fashion. Given an incomplete classifier $H_{t-1}^k(s) = \sum_{l=1}^{t-1} \alpha_l^k h_l^k(s)$ we seek to add one more weak classifier and its corresponding weight to optimize the overall loss function:

$$(\alpha_t^k, h_t^k) = \underset{(h,\alpha)}{\operatorname{argmin}} \Big(\mathcal{L}^k(H_{t-1}^k + \alpha h)\Big). \tag{4}$$

To achieve this, Viola et al. follow Friedman's suggestion of viewing the boosting procedure as a gradient descent in function space (where the value of H^k for every training instance corresponds to a dimension). In this sense, we would like to add a weak classifier h_t^k that is along the direction of the gradient

$$w_{ij} = \left.\frac{\partial \mathcal{L}^k}{H^k(s_{ij})}\right|_{H_{t-1}^k}. \tag{5}$$

Unfortunately, we cannot move in arbitrary directions in function space because we are limited by the class of weak learners we have chosen. Therefore, we would like to choose a weak classifier which moves in a direction that is as close as possible to this gradient:

$$h_t^k = \underset{h}{\operatorname{argmin}} \sum_{ij} h(s_{ij}) w_{ij}. \tag{6}$$

Finally, we can determine α_t by doing a simple line search.

3.2 Region Extraction

In MIL an image is divided into segments or regions and each region is represented by a high dimensional feature vector. Existing MIL-based approaches have adopted a variety of techniques for partitioning an image, including blocks, patches and single segmentations. One simple convention is to use a single non overlapping grid of 4×4 blocks [9,19,20]. In order to obtain representative regions of the possible objects in the scene, this block segmentation is followed by K-means clustering of the feature vectors extracted from the blocks. The number of clusters depends on the the number of objects that typically appear in the scene, introducing a model order selection problem. Other MIL-based approaches [9,10] extract salient regions using Kadir's detector [24]. This allows them to compare their object categorization results to non-MIL-based methods. Salient regions are detected over different locations and scales and then cropped from the image and rescaled into an image patch of size 11×11 pixels. Partitioning an image into blocks or patches often breaks an object into several pieces or puts different objects into a single patch. Alternatively, image segmentation is a way to decompose an image into a collection of regions that hopefully correspond to objects. The methods in [15,17] use the blobworld representation of [25] in which an image is segmented into a set of regions, each characterized by color, texture and shape descriptors. Other aproaches [10,18] obtain meaningful image regions using JSEG [26] or NCut [27] segmentation algorithms. Each image is typically segmented into ten or fewer regions, and only segments bigger than a certain threshold are kept. However, as described in [28], there usually does not exist a single correct segmentation of an image, but rather a collection of potentially meaningful image segmentations. Thus, using just a single segmentation may hinder recognition due to splitting or merging errors.

The idea of using multiple segmentation has recently emerged [3,5,29,30,31] in the area of object recognition. Segmentations are computed resulting in a bag (or soup) of segments, with the hope that a subset of them will capture adequate object boundaries. Multiple stable segmentations have been shown to produce competitive results in object categorization [29,32]. In this work we advocate their use as a substrate for MIL-based object categorization.

3.3 Multiple Stable Segmentations

In order to extract more adequate image regions for our system, we compute multiple stable segmentations [28]. The method of multiple stable segmentations uses stability as a heuristic for a particular set of parameters, cue weightings and a model order. For each choice of parameters for cue combinations $\boldsymbol{p}$ and number of segments q, the image is segmented using Normalized Cuts [27,33]. The segmentation is considered stable if small perturbations of the image do not yield substantial changes in the segmentation. The image is perturbed and segmented T times and the following score is evaluated:

$$\Phi(q, \boldsymbol{p}) = \frac{1}{n - \frac{n}{q}} \left(\sum_{i=1}^{n} \sum_{j=1}^{T} \delta_{ij} - \frac{n}{q} \right), \text{where } \delta_{ij} = \begin{cases} 1 & \text{if } i = \text{j} \\ 0 & \text{otherwise} \end{cases}. \tag{7}$$

Here n is the number of pixels and δ_{ij} is equal to 1 if the i-th pixel is mapped to a different segment in the j-th perturbed segmentation, and zero otherwise. Thus Φ is a properly normalized[2] measure of the probability of a pixel to change label due to a perturbation of the image. Segmentations with a high stability score are retained. Notice that, in general, there may exist several stable segmentations for an image.

[2] In particular Φ ranges in $[0, 1]$ and it is not biased towards a particular value of q.

3.4 MILSS Framework

Our Multiple Instance Learning framework using multiple Stable Segmentations (MILSS) presents a novel approach for object categorization that combines popular elements from previous work in object recognition with a MIL framework. Multiple stable segmentations [28] provide a spatial grouping of pixels into regions that increase the chances of extracting meaningful segments for MIL. They are memory efficient compared to extracting a large number of patches and they provide localization capability to our framework.

In order to improve instance classification, we use the bag of features model (BoF) [11] to capture appearance information. Recently, the BoF image representation has found widespread application in object categorization due to its simplicity and efficiency. To represent an image segment as a BoF, we first detect salient regions in the segment and compute a feature vector for each region. These feature vectors are then mapped to a vocabulary of "visual words" which are computed using vector quantization. The BoF representation of an image segment is then a histogram of these visual words (often referred to as a signature).

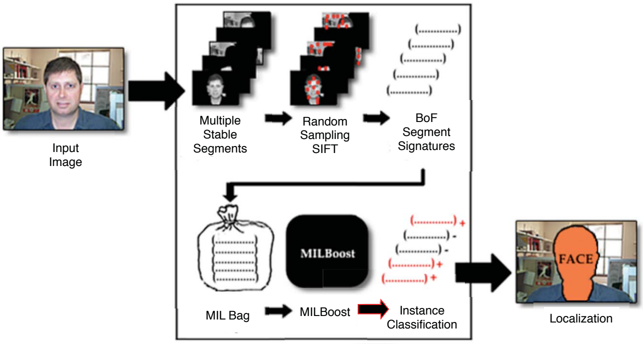


Fig. 2. An object is recognized by the MILSS framework. An input image containing a face is partitioned into a collection of stable segments. Then a BoF approach computes SIFT [34] descriptors in a random fashion on each segment. A signature is computed for each segment and the resulting bag of signatures is fed into the MilBoost model. MilBoost classifies each signature (instance) and the bag, resulting in the localization of the face within the image and the classification of the image as a whole.

We combine multiple segmentations and the BoF representation with the MilBoost framework [14] which performs feature selection during training and allows rapid segment and image classification at runtime. Figure 2 shows each step of our categorization model. Next we address, in detail, how the image segments and their signatures are used for object categorization.

Classification. Given an image $\mathcal{I}_i$ we compute q stable segmentations resulting in multiple segments $\{s_{i1}, s_{i2}..., s_{iJ}\}$. For each segment s_{ij} we compute a BoF signature, with each signature corresponds to an instance of the bag. A segment s_{ij} is classified as follows:

$$y_{ij} = \underset{k}{\operatorname{argmax}}\, p(c_k|s_{ij}), \tag{8}$$

where $p(c_k|s_{ij})$ is the probability of the segment s_{ij} belonging to the category c_k, defined by Eq. 1. We classify an image $\mathcal{I}_i$ as proposed by [29]:

$$y_i = \underset{k}{\operatorname{argmax}} \sum_{j=1}^{J} p(c_k|s_{ij}). \tag{9}$$

Localization. The task of object localization generally corresponds to placing a bounding box, or preferably the actual object outline, around the object within the image. Since our framework uses segments for categorization, we utilize segment boundaries that yield highest recognition score in order to describe object locations [1]. For evaluating our localization performance for an image $\mathcal{I}_i$ classified overall as y_i and segment labels y_{ij}, we look for segments with labels such that $y_{ij} = y_i$. Then we check for overlapping segments and return the first n unique segment boundaries, with $n \ll J$.

4 Experimental Results

To evaluate the MILSS framework, we compare our approach to the state-of-the-art methods in object categorization. Existing MIL-based approaches often use the COREL dataset to evaluate their models for image categorization. However, since we concentrate on object categorization, the performance of our approach is evaluated on Caltech 4 and a new dataset Landmarks-18.

4.1 Caltech 4 Dataset

Caltech 4 [11] is a well established dataset and is a standard benchmark for object categorization. Although simple, we utilize this dataset as a means of comparison with Mil-based methods. Following the experimental set up of [9,10], we perform a category versus background classification. Table 1(a) presents the results of categorization accuracy for our method. Results are compared to existing MIL-based image categorization models [9,10] and a non-MIL-based approach of [6]. The presented results are competitive with the rest of the algorithms. The average categorization accuracy for MILSS as well as ConMIL is 98%; while MILES is 97% and Bar-Hillel *et al.*'s algorithm is 93%. Note that the highest performance is achieved in the Airplanes category given that the stable segmentations were able to separate the background from the objects accurately. In a second experiment, we include the Leopard class for comparing our method to existing algorithms [8,11] in a multi-class setting.

Table 1(b) reports accuracy for multi-class object categorization. Instead of considering a background category, images belonging to each category acted as negative examples for models trained on the other categories. We compare our method to existing non-MIL-based object recognition frameworks: the dependent Hierarchical Dirichlet process (DHDP) [8] and constellation of parts model

Table 1. (a) Comparison of categorization results between our framework, MIL-based models [9,10] and a traditional object categorization approach [6] for Caltech 4 categories. Results in **bold** indicate the highest performance for each category. (b) MILSS Confusion matrix between the four categories for multi-class object recognition.

(a)

	Airplanes	Cars	Faces	Motorbikes
Training data	400	400	218	400
MILSS	**1**	.971	.976	.972
ConMIL [10]	.992	**.984**	.976	**.987**
MILES [9]	.980	.945	**.995**	.967
Bar-Hillel [6]	.897	.977	.917	.931

(b)

	A	F	L	M
Training data	400	218	100	400
Airplanes (A)	**.98**	.00	.01	.01
Faces (F)	.01	**.99**	.00	.00
Leopards (L)	.05	.01	**.93**	.01
Motorbikes (M)	.01	.00	.01	**.97**

[1]. As shown in Table 2(a), MILSS reports an average recognition accuracy of 97% while DHDP reports 98%. Looking closely at the categories, MILSS outperforms DHDP in three out of four of them. The Leopards category seems to be the most challenging for our framework, since it contains fewer images than the rest of the categories (100 for training and 100 for testing). In order to improve these results we could easily augment our training set with images from public repositories, as manual labeling is not required.

Table 2. (a) Results of multiple object categorization models for four Caltech categories. We compare our results to those of non MIL-based models. Results in **bold** indicate the highest performance for each category. (b) Average localization results of MILSS for four categories of Caltech.

(a)

	Airplanes	Faces	Leopards	Motorbikes	Mean
Training data	400	218	100	400	
MILSS	**.977**	**.986**	.927	.971	0.965
DHDP [8]	.961	.978	**1**	.967	0.976
Fergus [1]	.888	.862	-	**.977**	0.909

(b)

	MILSS
Airplanes	.932
Faces	.902
Leopards	.891
Motorbikes	.859
Mean	.896

In a multi-class setting, localization accuracy of MILSS is 90% on the Caltech 4 dataset. Our localization results are presented in detail in Table 2 (b). To quantify the accuracy of object localization we adopt the methodology of [1] and consider the overlap $\alpha = \frac{B \cap B_{gt}}{B \cup B_{gt}}$. Note that our method may be at a disadvantage in cases where the objects' contour areas B are smaller than the ground truth bounding box B_{gt}; thus it is difficult to make a direct comparison with the results in [1]. Since our method localizes objects using segment boundaries, the location and extent of the object is captured more precisely than those with bounding boxes, see Fig. 6.

4.2 Landmark Database

With the increasing popularity of digital photography and the user's desire to share their pictures in web albums, recognition of destinations and landmarks

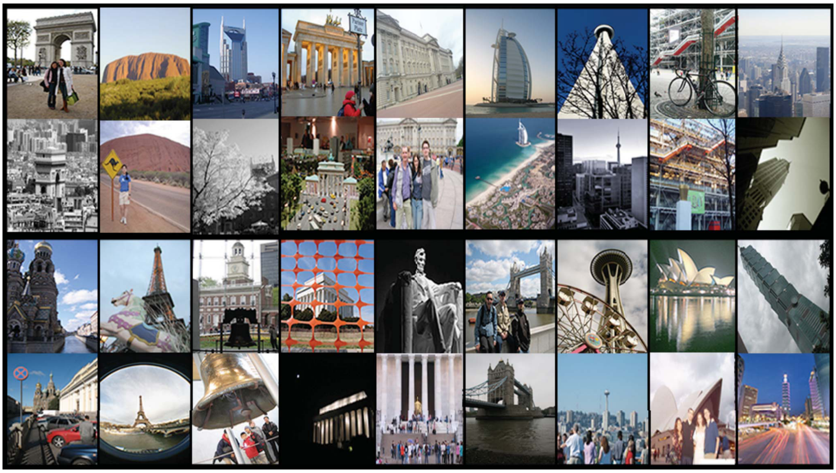
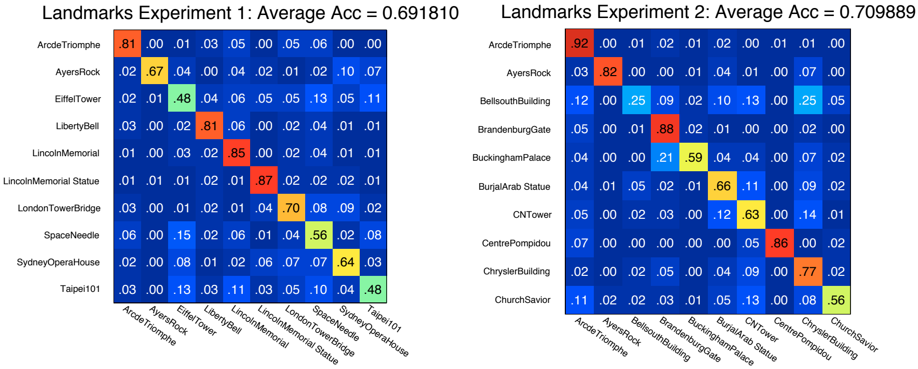


Fig. 3. Landmarks-18 Dataset. Two examples are shown per landmark and each row shows 9 categories. **Top row**: Arc de Triomphe, Ayres Rock, Bellsouth Building, Brandenburg Gate, Buckingham Palace, Burjal Arab, CN Tower, Centre Pompidou and Chrysler Building. **Bottom row**: Church Savior Spilled Blood, Eiffel Tower, Liberty Bell, Lincoln Memorial, Lincoln Memorial Statue, London Tower Bridge, Space Needle, Sydney Opera House and Taipei 101.

has become an interesting problem. Recognizing objects in real world images is a challenging task, as images are presented at a variety of viewpoints, scales, and illuminations; noise, background clutter, and occlusions also make the problem more difficult. Since photo-sharing sites are a vast resource of weakly labeled image data, we easily gather large datasets to evaluate our framework.

In this paper we introduce a new dataset called Landmarks-18, consisting of 18 different categories of landmarks, provided by Google Research and collected from public web albums. Landmarks-18 captures much more significant intra-class variability than standard benchmark datasets for object recognition. Figure 3 demonstrates the diversity of landmarks in the dataset while Fig. 5(b) provides the statistics of the dataset.

Here we performed two different multi-class categorization experiments on Landmarks-18. Each experiment considers 10 different categories, where images in each category were divided randomly into 80%/20% for training and testing respectively. Experiments were performed with 5-fold cross validation to obtain statistically relevant average categorization results. Figure 4 shows confusion matrices for both experiments. The results show that Landmarks-18 is much more difficult for categorization than Caltech 4, due to the challenging characteristics of its images and the larger number of classes. Despite this, MILSS achieves high categorization accuracy in both experiments. The outcome of both experiments indicate that Eiffel Tower, Taipei101, and Bellsouth Building are the most challenging categories. The main source of low recognition accuracy isnbetween visually similar categories such as Bellsouth Building vs. Chrysler



Landmarks Experiment 1: Average Acc = 0.691810

	ArcdeTriomphe	AyersRock	EiffelTower	LibertyBell	LincolnMemorial	LincolnMemorial Statue	LondonTowerBridge	SpaceNeedle	SydneyOperaHouse	Taipei101
ArcdeTriomphe	.81	.00	.01	.03	.05	.00	.05	.06	.00	.00
AyersRock	.02	.67	.04	.00	.04	.02	.01	.02	.10	.07
EiffelTower	.02	.01	.48	.04	.06	.05	.05	.13	.05	.11
LibertyBell	.03	.00	.02	.81	.06	.00	.02	.04	.01	.01
LincolnMemorial	.01	.00	.03	.02	.85	.00	.02	.04	.01	.01
LincolnMemorial Statue	.01	.01	.01	.02	.01	.87	.02	.02	.02	.01
LondonTowerBridge	.03	.00	.01	.02	.01	.04	.70	.08	.09	.02
SpaceNeedle	.06	.00	.15	.02	.06	.01	.04	.56	.02	.08
SydneyOperaHouse	.02	.00	.08	.01	.02	.06	.07	.07	.64	.03
Taipei101	.03	.00	.13	.03	.11	.03	.05	.10	.04	.48

(a)

Landmarks Experiment 2: Average Acc = 0.709889

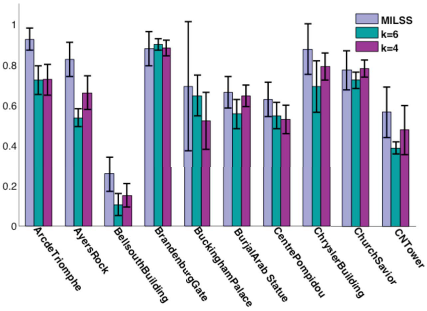
	ArcdeTriomphe	AyersRock	BellsouthBuilding	BrandenburgGate	BuckinghamPalace	BurjalArab Statue	CNTower	CentrePompidou	ChryslerBuilding	ChurchSavior
ArcdeTriomphe	.92	.00	.01	.02	.01	.02	.00	.01	.01	.00
AyersRock	.03	.82	.00	.00	.01	.04	.01	.00	.07	.00
BellsouthBuilding	.12	.00	.25	.09	.02	.10	.13	.00	.25	.05
BrandenburgGate	.05	.00	.01	.88	.02	.01	.00	.00	.02	.00
BuckinghamPalace	.04	.00	.00	.21	.59	.04	.04	.00	.07	.02
BurjalArab Statue	.04	.01	.05	.02	.01	.66	.11	.00	.09	.02
CNTower	.05	.00	.02	.03	.00	.12	.63	.00	.14	.01
CentrePompidou	.07	.00	.00	.00	.00	.00	.05	.86	.00	.02
ChryslerBuilding	.02	.00	.02	.05	.00	.04	.09	.00	.77	.02
ChurchSavior	.11	.02	.02	.03	.01	.05	.13	.00	.08	.56

(b)

Fig. 4. Confusion matrices of categorization accuracy for the Landmark-18 dataset. (a) Experiment 1; (b) Experiment 2.

Building. For this dataset we were unable to compare our results to other MIL-based categorization systems as code was not available.

To evaluate the importance of the multiple stable segmentations within MILSS, we also experimented with two different single segmentations ($q = 4$ and $q = 6$) using Normalized Cuts [27]. Figure 5 (a) shows the average categorization accuracy for each method using 5-fold cross validation. With multiple stable segmentations categorization performance is improved in almost all categories. The average categorization accuracy for $q = 4, 6$ and multiple segmentations is 58.3%, 61.8% and 71.0% respectively. The total number of segmentations extracted from an image plays an important role in categorization accuracy. As noted by others, as the number of segments per image increases, so does the



(a)

Category	n	Category	n
ArcdeTriomphe	146	ChurchSavior	109
AyresRock	113	EiffelTower	194
BellsouthBuild	107	LibertyBell	175
BrandenburgG	166	LincolnMem	198
BuckinghamP	87	LincolnMStatue	200
BurjalArab	158	LondonTower	195
CNTower	160	SydneyOHouse	186
CPompidou	71	SpaceNeedle	219
ChryslerBuild	204	Taipei101	176

(b)

Fig. 5. (a) Three different types of region extraction: two single segmentations with number of segments equal to 4 and 6, and multiple stable segmentations. The average categorization accuracy for $q = 4, 6$ and multiple segmentations is 58.3%, 61.8% and 71.0% respectively. Multiple stable segmentations outperform (on average) all the other methods. (b) shows the statistics of Landmarks-18 database.

chance of having a segment that represents the object accurately [29,30]. We believe that multiple stable segmentations provide a way of gathering the most meaningful segments, as is reflected in our results.

4.3 Implementation Details

The stability based image segmentation was implemented using Normalized Cuts [27,35]. Five iterations, combining brightness and texture cues with $\boldsymbol{p} = \{0.4, 0.5, 0.6, 0.7\}$ were used to sample the parameter space. For the categorization experiments done for Caltech and Landmarks-18, we computed 5 different

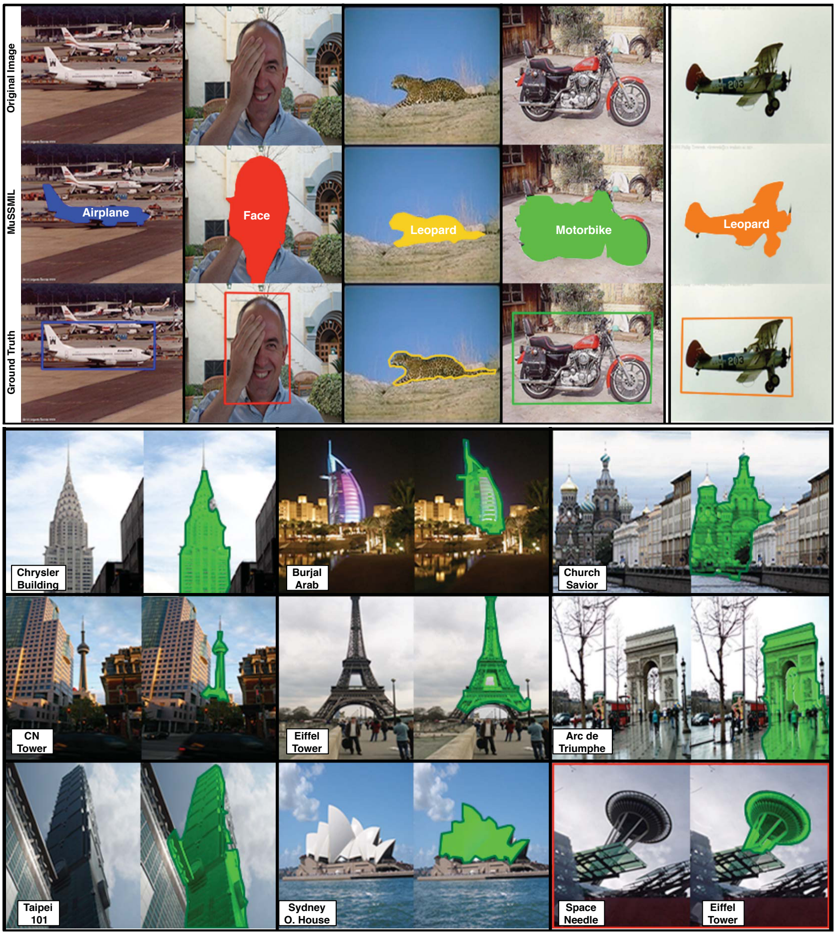


Fig. 6. Top image: Examples of Caltech test images. First three columns correspond to successful image categorization and localization of objects in the scene. Last column correspond to a false positive. **Bottom row**: Examples of Landmarks-18 test images. Green segments represent the image region with the highest probability of being the landmark. Images enclosed by a red rectangle correspond to a false positive.

segmentations with $q = 2, \ldots, 6$ with a total of 20 segments per image. Computing a single segmentation takes about 20-30 seconds per image. For the BoF model we computed 5000 random SIFT [34] features at multiple scales (from 12 pixels up to the full image size) for each image segment. Visual words are obtained computing a hierarchical K-means with $K = 17$ and three levels. The computation of SIFT descriptors and signatures takes about 1 second per segment in a MATLAB/C implementation. Constructing the vocabulary tree takes 40-50 minutes for ten categories. Training time for MilBoost on four Caltech categories takes about 1 day using 500 weak classifiers. Using ten categories of Landmarks-18 MilBoost take less than a day of training using 200 weak classifiers. Classification of all test images for ten categories is done in 0.5 seconds. All above operations were performed on a Pentium 2.8 GHz.

5 Conclusions and Future Work

In this paper we proposed a novel framework for image categorization and localization of objects in real world scenes using weakly labeled data. Our performance is highly competitive with current MIL-based and traditional approaches for image and object categorization. We showed that multiple stable segmentations extracted suitable regions for the MIL problem, thus increasing performance in categorization and permitting accurate localization capabilities. We tested our framework on Caltech 4 and Landmarks-18 datasets, obtaining high accuracy in object categorization tasks. As future work, we want to explore new methods to scale our object categorization framework to a larger number of categories and handle multiple objects in the scene.

Acknowledgments. Special thanks to Hartmut Neven and Hartwig Adam from Google for providing the Landmarks-18 dataset used in this paper. This work was funded in part by NSF Career Grant #0448615, the Alfred P. Sloan Research Fellowship, NSF IGERT Grant DGE-0333451 and a Google Research Award.

References

1. Fergus, R., Perona, P., Zisserman, A.: Weakly supervised scale-invariant learning of models for visual recognition. IJCV 71(3), 273–303 (2007)
2. Opelt, A., Fussenegger, M., Auer, P.: Generic object recognition with boosting. PAMI 28(3), 416–431 (2006)
3. Russell, B., Efros, A., Sivic, J., Freeman, W., Zisserman, A.: Using multiple segmentations to discover objects and their extent in image collections. In: CVPR (2006)
4. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering object categories in image collections. In: CVPR (2005)
5. Todorovic, S., Ahuja, N.: Extracting subimages of an unknown category from a set of images. In: CVPR (2006)
6. Bar-Hillel, A., Hertz, T., Weinshall, D.: Object class recognition by boosting a part-based model. In: CVPR (2005)

7. Crandall, D., Huttenlocher, D.: Weakly supervised learning of part-based spatial models for visual object recognition. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 16–29. Springer, Heidelberg (2006)
8. Wang, G., Zhang, Y., Fei-Fei, L.: Using dependent regions for object categorization in a generative framework. In: CVPR (2006)
9. Chen, Y., Bi, J., Wang, J.: MILES: Multiple-instance learning via embedded instance selection. PAMI 28(12), 1931–1947 (2006)
10. Qi, G., Hua, X., Rui, Y., Mei, T., Tang, J., Zhang, H.: Concurrent multiple instance learning for image categorization. In: CVPR (2007)
11. Fergus, R., Perona, P., Zisserman, A.: Object class recognition by unsupervised scale-invariant learning. In: CVPR (2003)
12. Dietterich, T.G., Lathrop, R.H., Perez, L.T.: Solving the multiple-instance problem with axis parallel rectangles. AAAI, Menlo Park (1997)
13. Andrews, S., Hofmann, T., Tsochantaridis, I.: Multiple instance learning with generalized support vector machines. AAAI, Menlo Park (2002)
14. Viola, P., Platt, J.C., Zhang, C.: Multiple instance boosting for object detection. In: NIPS, vol. 18 (2006)
15. Maron, O., Ratan, A.: Multiple-instance learning for natural scene classification. In: ICML (1998)
16. Zhou, Z., Zhang, M.: Multi-instance multi-label learning with application to scene classification. In: NIPS, vol. 19 (2007)
17. Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: NIPS, vol. 15 (2002)
18. Yang, C., Dong, M., Hua, J.: Region-based image annotation using asymmetrical support vector machine-based multi-instance learning. In: CVPR (2006)
19. Chen, Y., Wang, J.: Image categorization by learning and reasoning with regions. JMLR 5, 913–939 (2004)
20. Bi, J., Chen, Y., Wang, J.: A sparse support vector machine approach to region-based image categorization. In: CVPR (2005)
21. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29(5), 1189–1232 (2001)
22. Freund, Y., Schapire, R.E.: A decision-theoretic generalization of on-line learning and an application to boosting. JCSS 55, 119–139 (1997)
23. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
24. Kadir, T., Brady, M.: Saliency, scale and image description. IJCV 45 (2001)
25. Carson, C., Belongie, S., Greenspan, H., Malik, J.: Blobworld: image segmentation using expectation-maximization and its application to image querying. PAMI 24(8), 1026–1038 (2002)
26. Deng, Y., Manjunath, B.: Unsupervised segmentation of color-texture regions in images and video. PAMI 23(8), 800–810 (2001)
27. Shi, J., Malik, J.: Normalized cuts and image segmentation. PAMI 22(8), 888–905 (2000)
28. Rabinovich, A., Lange, T., Buhmann, J., Belongie, S.: Model order selection and cue combination for image segmentation. In: CVPR (2006)
29. Rabinovich, A., Vedaldi, A., Belongie, S.: Does image segmentation improve object categorization? UCSD Technical Report CSE CS2007-0908 (2007)
30. Malisiewicz, T., Efros, A.: Improving spatial support for objects via multiple segmentations. BMVC (2007)

31. Roth, V., Ommer, B.: Exploiting low-level image segmentation for object recognition. In: Franke, K., Müller, K.-R., Nickolay, B., Schäfer, R. (eds.) DAGM 2006. LNCS, vol. 4174, pp. 11–20. Springer, Heidelberg (2006)
32. Rabinovich, A., Vedaldi, A., Galleguillos, C., Wiewora, E., Belongie, S.: Objects in context. In: ICCV (2007)
33. Malik, J., Belongie, S., Shi, J., Leung, T.: Textons, contours and regions: Cue integration in image segmentation. In: ICCV (1999)
34. Lowe, D.: Object recognition from local scale-invariant features. In: ICCV (1999)
35. Cour, T., Benezit, F., Shi, J.: Spectral segmentation with multiscale graph decomposition. In: CVPR (2005)