Submodular Object Recognition

Fan Zhu[†], Zhuolin Jiang[‡], and Ling Shao[†]

[†]Department of Electronic and Electrical Engineering, The University of Sheffield
[‡]Noah's Ark Lab, Huawei Technologies
{fan.zhu, ling.shao}@sheffield.ac.uk, zhuolin.jiang@huawei.com

Abstract

We present a novel object recognition framework based on multiple figure-ground hypotheses with a large object spatial support, generated by bottom-up processes and mid-level cues in an unsupervised manner. We exploit the benefit of regression for discriminating segments' categories and qualities, where a regressor is trained to each category using the overlapping observations between each figure-ground segment hypothesis and the ground-truth of the target category in an image. Object recognition is achieved by maximizing a submodular objective function, which maximizes the similarities between the selected segments (i.e., facility locations) and their group elements (i.e., clients), penalizes the number of selected segments, and more importantly, encourages the consistency of object categories corresponding to maximum regression values from different category-specific regressors for the selected segments. The proposed framework achieves impressive recognition results on three benchmark datasets, including PASCAL VOC 2007, Caltech-101 and ETHZ-shape.

1. Introduction

In recent years, the bag-of-features (BoF) model and its extension, spatial pyramid matching (SPM) [17], have been popular for object recognition. When working with densely sampled pyramid grids and powerful classifiers, BoF and SPM have achieved impressive performance on several object recognition benchmarks including PASCAL VOC 2007 [6] and Caltech-101 [7]. While these densely sampled grids can retain context information, such as spatial layout, for a specific object category, irrelevant background information is also included. To solve this problem, a lot of efforts have been made to leverage segmentation results for better recognition performance. The benefits of incorporating segmentation for recognition lie in two folds: (1) accurate segmentation can enhance the contrast of object boundaries, so that features along the boundaries are more shape-informative; (2) computing features on homogeneous segments improves the signal-to-noise ratio. However, little progress has been achieved due to a lack of reliable segmentation techniques. For example, Nilsback and Zisserman [22] employed segmented images for flower classification. Since only a single segment is considered for an image, and clean segmentations can only be guaranteed for images with simple backgrounds, the performance improvement is not significant when comparing with results from non-segmented images. Unlike approaches that consider only one segment in an image [22], our approach considers multiple segments simultaneously via submodularity. Our approach is based on the recently proposed Constrained Parametric Min Cuts (CPMC) [3] algorithm, which has demonstrated a significant improvement in segmentation. We present a submodular objective function for efficiently selecting discriminative segments from the set of figure-ground hypotheses for object recognition. We learn a scoring (regression) function to each object category with the overlapping observations of each pair of the figure-ground hypothesis and the ground-truth segment. The benefit of regression is exploited for discriminating segments' categories and qualities. Our objective function contains a facility-location term and a discriminative term, where the facility-location term is measured by the total similarities between the selected segments and their group elements and the facility costs for the selected elements, and the discriminative term is measured by the consistency of categories that obtain the maximum regression values on selected segments. Our main contributions are three-fold:

- ⋆ Object recognition is modeled as a facility location problem with the constraint of class purity of selected segments (facility locations), which can be solved by maximizing a submodular function. We provide a new perspective of applying submodularity to the object recognition problem.

- ⋆ Based on its submodularity property, the objective function is solved by an efficient greedy algorithm with the guaranteed performance of at least an $(e-1)/e$-approximation to the optimum.

- ⋆ Our submodular recognition approach achieves state-of-the-art performance on three popular object recognition benchmarks.

2. Related Work

Many recent bottom-up object recognition approaches attempt to use the spatial layouts of objects for better performance. He et al. [13] constructed a Conditional Random Field (CRF) framework on image pixels, where each pixel is assigned to one of a finite set of labels. Both image features and image labels are incorporated into the probabilistic framework. Shotton et al. [28] proposed Textonboost, which incorporates texture, layout and context information for unary classification. By incorporating the unary classifier into a CRF, the spatial interactions between class labels of neighboring pixels are captured to guarantee the smoothness. A major limitation of pixel-level methods is their weak capability for segmenting nearby objects of the same category. Gould et al. [11] and Ladicky et al. [16] addressed such a limitation using rectangular bounding box detection constraints. Rather than using bounding boxes, segment-based or superpixel-based approaches are closer to the ground-truth spatial support. Fulkerson et al. [9] used superpixels as basic units in the recognition framework. To this end, the histogram of local features within each superpixel is used to construct a classifier, which is regularized by aggregating histograms of neighboring superpixels. For segment-level recognition methods, Rabinovich et al. [25] applied a stability heuristic to select a reduced list of segmentations obtained from normalized cuts [27]. For an image I, each segment in the list is regarded as a stand-alone image, and labels from all segments are used to vote for the category of image I. By using a collection of segments for recognition rather than a single segment, more object boundary information can be captured. However, they do not provide a reliable segment selection mechanism for filtering out erroneous segments, and treating the whole collection of segments as a new set of images is too computationally expensive. Carreira et al. [2] presented an object recognition framework based on multiple figure-ground segmentations generated by CPMC, which is the most similar approach to our work. However, our method differs from their approach by the way of selecting the compact and discriminative figure-ground hypotheses. Instead of being ad-hoc as in [2], we apply a constant-factor approximation based on the submodularity.

Submodularity has recently been applied to many computer vision tasks, including clustering [20], segmentation [14]. Liu et al. [20] presented a method that uses the entropy rate of a random walk on a graph for compact and homogeneous clustering. Jiang and Davis [14] solved a facility location problem [10, 18] for salient region detection. The saliency of a region is modeled in terms of its appearance and spatial location, and salient region detection is achieved by maximizing a submodular objective function.

3. Submodular Object Recognition

Our method solves the object recognition problem through the selection of a subset of segments so as to best discover the target object in a query image. Firstly, we apply the CPMC segmentation [3] on each image to produce a set of figure-ground hypotheses in an unsupervised manner. Then we construct a graph G based on the generated figure-ground hypotheses. Since using all segment hypotheses is too computationally expensive and probably produces misleading predictions. Thus, we aim to discover the discriminative segment subset $\mathcal{A}$ of S by iteratively selecting elements of S into $\mathcal{A}$. Object masks are obtained by overlaying selected segments for extracting foreground objects. Finally, a linear classifier is applied for recognizing objects.

3.1. Preliminaries

Submodularity: Let $\mathcal{V}$ be a finite set, $\mathcal{A} \subseteq \mathcal{B} \subseteq \mathcal{V}$ and $a \in \mathcal{V} \backslash \mathcal{B}$. A set function $F : 2^{\mathcal{V}} \rightarrow \mathcal{R}$ is submodular if $F(\mathcal{A} \cup a) - F(\mathcal{A}) \geq F(\mathcal{B} \cup a) - F(\mathcal{B})$. This property is referred to as diminishing returns, stating that adding an element to a smaller set helps more than adding it to a larger set [21].

3.2. Graph-Construction

For an image I, N figure-ground hypotheses $S = \{S_1, S_2, \cdots, S_N\}$[1] are generated by CPMC and the ground-truth segment G_I^k of object category k is provided for the training data. A subset of figure-ground hypotheses is shown in Figure 1(b)∼(f). We construct a graph $G = (\mathcal{V}, E)$ on the segment hypotheses in image I, where the vertices $v \in \mathcal{V}$ are segment hypotheses while the edges $e \in E$ model the pairwise relations between segment hypotheses. The weight w_{ij} assigned to the edge e_{ij} can be computed through Equation 2.

3.3. Salient Segment Selection

Segments are selected according to criteria: 1) In order to obtain the discriminative segments for recognition, we model the segment selection and recognition as a facility location problem; 2) In order to associate the segments with object categories, we train a regressor for each category to predict the overlaps between all the candidate segments and the ground-truth segment. Assuming a category label with the highest regression value is assigned to a segment, we control the purity of segments' category labels in an image during segment selection. The two criteria are satisfied by formulating both the facility location term and the entropy term in the objective function, and then maximizing it based on submodularity.

[1] N is constrained within 100 to limit the computational cost in our work.

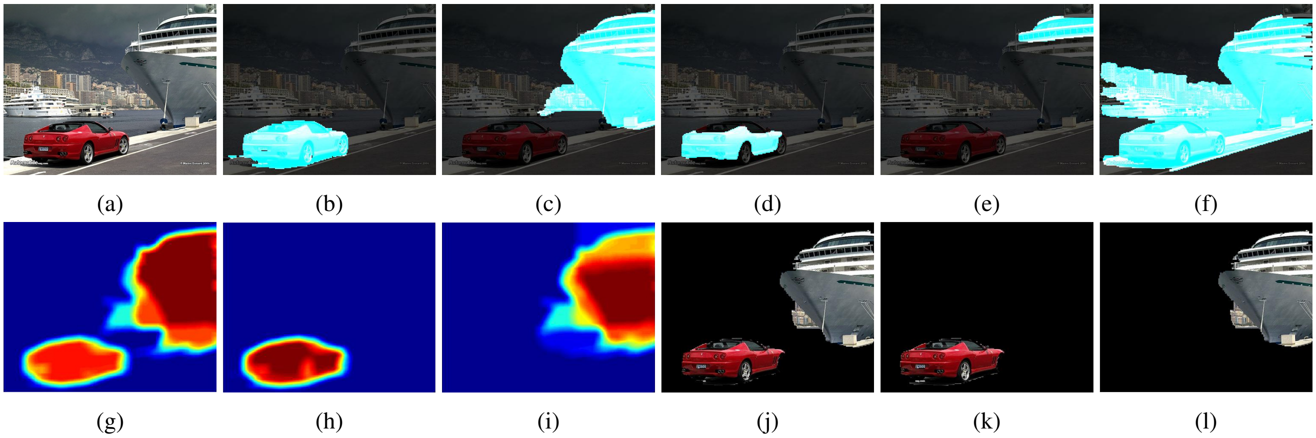


Figure 1: An example of submodular segment selection for the presence/absence classification task. (a): Input image; (b)∼(f): A small subset of figure-ground segments of different qualities generated by CPMC; (g)∼(i): Aggregated confidence of selected segments, where (g) is based on the facility location term only, (h) and (i) are based on both the facility location term and the entropy term; (h) is obtained using both the facility location term and the discriminative term based on the "car" regressor, and (i) is obtained using both the facility location term and the discriminative term based on the "ship" regressor. Since discriminative terms are not included in (g), segments' categories are not considered during the selection. Thus, selected segments cannot focus on a single object (as in (h) or (i)) if an image contains more than one object categories; (j)∼(l): Foreground image regions are covered by region masks, which are obtained by thresholding results of (g)∼(i).

3.3.1 Facility-Location Term

We model the problem of selecting the discriminative segment set among all the segments in an image as the facility location problem [10, 18]. It can be considered as the set of locations for opening facilities. Let $N_{\mathcal{A}}$ denote the number of open facilities. With the constraint K, the combinatorial formulation of the facility location problem can be applied:

$$\max_{\mathcal{A}} \mathcal{H}(\mathcal{A}) = \sum_{i\in\mathcal{V}} \max_{j\in\mathcal{A}} w_{ij} - \sum_{j\in\mathcal{A}} \phi_j \qquad (1)$$
$$s.t. \quad \mathcal{A} \subseteq S \subseteq \mathcal{V}, N_{\mathcal{A}} \leq K$$

where w_{ij} denotes the pairwise relationship between a group element v_i (considered as clients) and a potential group center vertex v_j (considered as facilities), and the cost ϕ_j of opening a facility is fixed to δ. Submodularity of the overall profit $\mathcal{H}$ has been proved in [10, 18].

The first term in (1) encourages the element v_i has the largest value with its assigned group center. It favors the selected segment v_j (group center) to well represent or be similar to its clients (group elements) so that the final selected set $\mathcal{A}$ can be representative. The weight w_{ij} of each edge e_{ij} is computed as:

$$w_{ij} = \mathcal{K}(v_i, v_j) + O(v_i, v_j), \qquad (2)$$

where $\mathcal{K}(v_i, v_j)$ denotes the chi-squared distance $exp(-\gamma\chi^2(v_i, v_j))$ on histogram features of any pair of group elements, and $O(v_i, v_j)$ denotes the 'union-over-intersection' overlap measurement of the same pair of group elements:

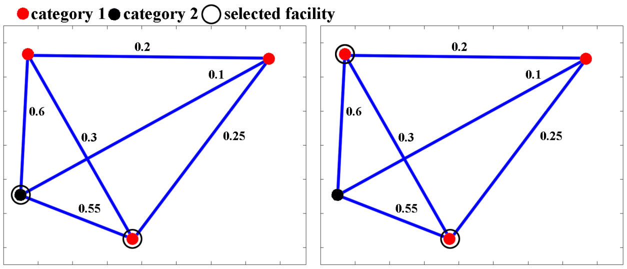
$$O(v_i, v_j) = \frac{|v_i \bigcap v_j|}{|v_i \bigcup v_j|}. \qquad (3)$$

If w_{ij} is computed only based on the overlap measurement, the facility location term will pursue segments that have highest overlap values with neighbouring segments, so that segments with large background coverage are preferably selected. Including the chi-squared distance on segments' histogram features can effectively avoid such a problem. The second term penalizes on extraneous facilities. When the gain obtained by introducing a new segment to the histogram is offset by the cost of opening such a facility, $\mathcal{A}$ will stop growing. Hence this selected A is representative (i.e., central) and compact (i.e., diversified).

3.3.2 Discriminative Term

We enforce a class-purity constraint to boost the discriminativity power of the selected $\mathcal{A}$. The discriminative term is based on the class purity constraint, which can be obtained through the learned segment regressor of each category.

These segments are represented by the spatial pyramid descriptors [29]. The object category contained in image I is $k \in \{1, 2, \cdots, m\}$, and we need to learn m scoring functions $f_1(S_i), f_2(S_i), \cdots, f_m(S_i)$ for each object category. Each function is defined on the score set O, which is computed by the overlaps between a segment S_i and the ground-truth segment G_I^k of category k in an image I using the



Obj value = 2.45

(a)

Obj value = 4.8

(b)

Figure 2: Segment selection based on (a) the facility location term, and (b) both the facility location term and the entropy term. Dots with different colors denote vertices from different categories, and a black circle denotes a selected segment. The number next to an edge is the weight w_{ij} of two vertices defined in Equation 2, and objective function values of two selection results are shown under the box. By including the entropy term, we observe that the category consistency of selected segments is considered as well as the saliency of each segment. Thus, (b) is preferred.

Algorithm 1 Submodular Object Recognition

```
**Input:** $I$, $S$, $f_k(\cdot)$, $\hat{W}$, $K$ and $\tau$.
**Output:** $\mathcal{A}$, $\mathcal{M}$, $k^*$.
**Initialization:** $\mathcal{A} \leftarrow \emptyset$, $\mathcal{R}_{a_i} \leftarrow 0$, $\mathcal{M} \leftarrow 0$.
**loop**
        $a^* = \underset{\{\mathcal{A}\cup a\}\in\mathcal{V}}{\arg\max}\ \mathcal{C}(\mathcal{A}\cup\{a\}) - \mathcal{C}(\mathcal{A})$
    **if** $\mathcal{C}(\mathcal{A}\cup\{a\}) \leq \mathcal{C}(\mathcal{A})$ or $N_{\mathcal{A}} > K$ **then**
        break;
    **end if**
    $\mathcal{A} \leftarrow \mathcal{A}\cup\{a^*\}$, $\mathcal{R}_{a^*} \leftarrow 0$
    $\mathcal{M} \leftarrow \mathcal{M} + f_k(a^*) * a^*$
    **for** $\forall i \in \mathcal{V}\setminus\mathcal{A}$ **do**
        $\mathcal{R}_{a_i} = \mathcal{R}_{a_i}^{\text{new}}$
    **end for**
**end loop**
**for** each pixel $I_{ij}$ in $\mathcal{M}$ **do**
    **if** $I_{ij} > \tau$ **then**
        $\mathcal{M}(I_{ij}) =$ foreground
    **else**
        $\mathcal{M}(I_{ij}) =$ background
    **end if**
**end for**
Integrate the final mask $\mathcal{M}$ with the SPM framework, and obtain a
    global representation $x_I$ for image $I$.
Obtain the category $k^* = \arg\max_k(l = \hat{W}x_I)$.
```

'union-over-intersection' measurement. Specifically, each O_i is computed as:

$$O_i(S_i, G_I^k) = \frac{|S_i \bigcap G_I^k|}{|S_i \bigcup G_I^k|}. \tag{4}$$

Thus, each figure-ground segment S_i is associated with its regression score O_i through $f_k(S_i) = O_i(S_i, G_I^k)$. Since a segment usually overlaps with more than one ground-truth segments when training each $f_k(S_i)$, it can have different regression values for a segment when training the scoring function of different categories. If category k does not appear in image I, all the segments in image I are considered as having no overlap with category k. A simple linear Support Vector Regression is applied to learn each scoring function $f_k(S)$ by regressing on the score set $\{O\}$ against S for all images in the training set[2]. During testing, the scoring function which results in the highest regression value determines the category of a query segment S_i, i.e., the category of S_i is computed by $y_i = \arg\max_k f_k(S_i)$.

The entropy is governed by the probability distribution of category labels that exist in $\mathcal{A}$, and it measures the consistency of the labels of selected segments. Note that the probability $p(j)$ is not calculated by counting the number of segments that contribute to each category, but by directly using the category label of each selected segment. The definition of the entropy term is given by:

$$\mathbb{E}(\mathcal{A}) = -\sum_{j\in\mathcal{A}} p(j)\log p(j), \tag{5}$$

with

$$p(j) = \frac{\arg\max_k f_k(S_j)}{\sum_{j\in\mathcal{A}} \arg\max_k f_k(S_j)}, \tag{6}$$

where the numerator denotes the object category of segment j, and the denominator sums all values on numerators to guarantee the convolution of the probability distribution equals to one. To each candidate segment, its category is assigned by the scoring function that achieves the highest regression value. By maximizing $\mathbb{E}(\mathcal{A})$, we encourage the selected segment set $\mathcal{A}$ to possess homogeneous category labels, which reduce negative effects from erroneous regressors. The maximum value of $\mathbb{E}(\mathcal{A})$ is reached when $p(i) = p(j), \forall i, j \in \mathcal{A}$, i.e., all segments in $\mathcal{A}$ come from the same object category. Note that different ways for including the entropy term into the objective function are used for the multi-category classification task (i.e., the Caltech-101 dataset) and the presence/absence classification[3] task (i.e., the PASCAL 2007 and the ETHZ-shape dataset dataset). For multi-category classification, regressors of all categories are used during the segment selection process, and a segment's category is allocated by the regressor that possesses the highest regression value. For the presence/absence classification task, only a single regressor

[2] The regressors are first trained only using the ground-truth segments, after which all candidate segments are fed into the regressors for classification. The miss-classified segments are then added to the training segments for re-training the regressors. Considering the high computational cost caused by huge amounts of segment hypotheses, we adopt the hard negative example mining strategy to refine the training as in [2].

[3] Following [29, 23, 4, 5, 12, 1], the presence/absence classification is to predict presence/absence of an example of that class in the test image.

of the query category is considered, and a segment's category is allocated as '1' if the regression value is above 0.5, and '2' otherwise. When the query category changes, different segments are selected with respect to a different regressor. Thus, our method can well handle the presence/absence classification problem, where images contain more than one objects. The proof of monotonicity and submodularity of $\mathbb{E}(\mathcal{A})$ is given in the Appendix section.

Figure 1(j)$\sim$(l) show the segment selection results for an example of the presence/absence classification task. Without the entropy term as in Figure 1(j), the facility location term only favors representative segments. When the entropy term is included, category-specific segments are selected. When detecting object category 'car' segments with high regression scores of the "car" regressor are prefered (shown in Figure 1(k)). Figure 2 illustrates how the facility location term and the entropy term contribute to a selection.

3.4. Optimization

We can combine the facility-location term and the discriminative term into a unified objective function:

$$\begin{aligned}\max_{\mathcal{A}} \mathcal{C}(\mathcal{A}) &= \max_{\mathcal{A}} \mathcal{H}(\mathcal{A}) + \lambda \mathbb{E}(\mathcal{A}) \\ &= \max_{\mathcal{A}} \sum_{i \in \mathcal{V}} \max_{j \in \mathcal{A}} c_{ij} - \sum_{j \in \mathcal{A}} \phi_j \\ &\quad -\lambda \sum_{j \in \mathcal{A}} p(j) \log p(j)\end{aligned} \tag{7}$$

Direct maximization of $\mathcal{C}(\mathcal{A})$ is an NP-hard problem [10]. The submodularity of (7) is preserved by taking non-negative linear combinations of the two submodular terms $\mathcal{H}(\mathcal{A})$ and $\mathbb{E}(\mathcal{A})$. Utilizing this property, (7) can be efficiently solved via a greedy algorithm [10] [21]. The segment set $\mathcal{A}$ is initialized with $\emptyset$, and a segment $a^* \in \mathcal{A} \setminus S$ that leads to the largest marginal gain $\mathcal{R}_{a^*}$ at each iteration is iteratively added to $\mathcal{A}$. The algorithm updates the marginal gain of the selected segment with 0 and the remaining facility assignments in $\mathcal{V}$ by $\mathcal{R}^{\text{new}}_{a_i^*}$ at every iteration. $\mathcal{A}$ stops absorbing new segments when the desired number of segments is reached or the gain decreases. The constraint on the number of open facilities induces a simple uniform matriod $\mathcal{U} = (S, \mathcal{I})$, where $\mathcal{I}$ is the collection of subsets $\mathcal{A} \subseteq S$, which satisfies that the number of open facilities $N_{\mathcal{A}}$ is less than K. Maximization of a submodular function with a uniform matroid constraint yields a $(1-1/e)$-approximation [21]. Hence our approach provides a performance-guarantee solution.

The optimization process can be accelerated by using the submodularity property of the objective function. Instead of recomputing the gain for adding every segment $a \in \mathcal{V} \setminus \mathcal{A}$, which requires $|\mathcal{V}| - |\mathcal{A}|$ evaluations for the gain $\mathcal{C}(\mathcal{A})$, we use the lazy evaluation form from [19]. The pseudocode of submodular object recognition framework is given in Algorithm 1 , where the submodular optimization process is given between line $3 \sim 20$.

3.5. Segmentation Mask Construction

The final mask $\mathcal{M}$ is obtained by overlaying all segments in $\mathcal{A}$ while taking account of the confidence score $f_k(a_j)$ of each segment $a_j \in \mathcal{A}$. An adaptive threshold $\tau = 0.6 \times N_{\mathcal{A}}$ is applied to $\mathcal{M}$ to filter out pixels with low confidence scores.

3.6. Classification

We integrate the final mask $\mathcal{M}$ with the SPM framework [29] for object representation. For each image, $\mathcal{M}$ is applied to mask and zero pad the original image. By discarding regions that fall outside the mask $\mathcal{M}$, the image representation x_I is computed as in the SPM framework. Then we use the multivariate ridge regression model to train a linear classifier $\hat{W}$:

$$\hat{W} = \arg\max_{W} \|H - WX\|_2^2 + \varphi \|W\|_2^2, \tag{8}$$

where X is the training data, H is the class label matrix of X, and W denotes classifier parameters. This yields the solution $\hat{W} = HX^T(XX^T + \varphi \mathcal{Z})^{-1}$, with $\mathcal{Z}$ being an identity matrix. For a test image I, we first compute its representation x_I and then estimate its class label vector $l = \hat{W} x_I$, where $l \in R^m$. Its label is the index i corresponding to the largest element in l.

4. Experiments

We evaluate our submodular object recognition approach on three popular benchmarks, including Caltech-101 [7], PASCAL VOC 2007 [6] and ETHZ-shape [8]. For all three datasets, we compute the dense SIFT features on each image. Regressors are trained based the ground-truth segmentations provided with the training data. For all the experiments, we evaluate our approach by either using the facility location term ("FL") only or using both the facility location term and the entropy term ("FL"+"EN").

4.1. PASCAL VOC 2007

We extensively evaluate the effectiveness of our approach on the PASCAL VOC 2007 dataset, as the ground-truth of the testing data is released. The PASCAL VOC 2007 dataset contains $9,963$ images from 20 visual object categories, and the dataset is evenly split to "trainval" and "test" parts. Following typical settings in [29, 23, 4, 5, 12], we conduct experiments on the "trainval" and "test" splits. In our algorithm, we train the regressors according to the overlap observations between each figure-ground hypothesis and the ground-truth segmentation of an object category. Since the ground-truth segmentations are only available for those images provided in the segmentations challenge, we train the regressors only based on images with provided ground-truth segmentation in the "trainval" split. We show

Table 1: Average precisions (APS) of each object category achieved by the baseline method and our proposed methods on the PASCAL VOC 2007 dataset.

Methods	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	Avg
Yang [29]	74.8	65.2	50.7	70.9	28.7	68.8	78.5	61.7	54.3	48.6	51.8	44.1	76.6	66.9	83.5	30.8	44.6	53.4	78.2	53.5	59.3
Florent [23]	75.7	64.8	52.8	70.6	30.0	64.1	77.5	55.5	55.6	41.8	56.3	41.7	76.3	64.4	82.7	28.3	39.7	56.6	79.7	51.5	58.3
Harzallah [12]	77.2	69.3	56.2	66.6	45.5	68.1	83.4	53.6	58.3	51.1	62.2	45.2	78.4	69.7	86.1	52.4	54.4	54.3	75.8	62.1	63.5
Qiang [4]	76.7	74.7	53.8	72.1	40.4	71.7	83.6	66.5	52.5	57.5	62.8	51.1	81.4	71.5	86.5	36.4	55.3	60.6	80.6	57.8	64.7
Dong [5]	82.2	**83.0**	58.4	76.1	56.4	77.5	88.8	69.1	62.2	61.8	64.2	51.3	85.4	80.2	**91.1**	48.1	61.7	67.7	86.3	70.9	71.1
FL	81.2	82.2	56.7	73.5	56.2	76.5	88.5	67.8	58.0	60.1	61.7	48.1	85.1	77.8	89.3	45.5	60.6	64.4	84.3	69.2	69.3
FL+EN	**83.7**	82.5	**63.3**	**77.3**	**58.0**	**80.2**	**89.4**	**68.8**	**63.1**	**63.7**	**67.4**	**53.5**	**86.4**	**82.7**	90.5	**48.4**	**62.0**	**67.9**	**87.2**	**71.5**	**72.4**

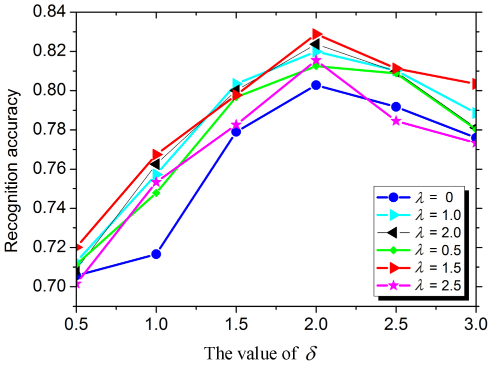


Figure 3: Effects of parameter selection of λ and δ on the recognition performance on the Caltech-101 dataset when using 30 training examples per category. The horizontal axis denotes different values of δ, while lines with different colors denote different λ values.

the results achieved by both "FL" and "FL+EN" in Table 1. We calculate the average precisions (APs) for each object category using both approaches, and compare with state-of-the-art approaches [29, 23, 4, 5, 12]. As can be observed, the "FL+EN" approach outperforms all other approaches.

4.2. Caltech-101 Dataset

The Caltech-101 dataset [7] contains $9,144$ images from 102 classes (101 object classes and a 'background' class). The ground-truth segmentations are provided in this dataset. We train a codebook with 2048 bases, and choose 4×4, 2×2 and 1×1 sub-regions for SPM. Following the common experimental protocol, randomly selected 5, 10, 15, 20, 25, 30 samples per category are used for training, and remaining samples are for testing. We repeat the experiments 10 times and the final results are reported as the average of each run. We compare our results with state-of-the-art approaches [29, 15, 26, 2] in Table 2. We also show the results of "BS" and "GT", which denote results produced by using only the best segments[4] and ground-truth segments, respectively. The high performance of "BS" and "GT" proves our motivation that recognition performance can be improved by segmentation.

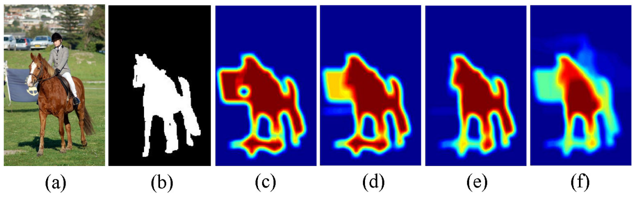


Figure 4: Effects of parameter selection of δ on the aggregated confidence of selected segments $\mathcal{M}$. (a): Input image; (b): Ground truth segment; (c)$\sim$(f): The aggregated confidence of selected segments when the penalty cost $\delta = 3, 2.5, 2, 0.5$, respectively. The color denotes different confidence values (red: high, blue: low). In case of too few segments are selected as in (c), the aggregated confidence does not have accurate coverage of the object. The coverage of the aggregated confidence is improved in (d) when more segments are selected. (e) has the most accurate coverage. $\mathcal{A}$ can "over-select" segments if we reduce the penalty term. In (f), the aggregated confidence focuses on a small central region of the object as too many segments are selected.

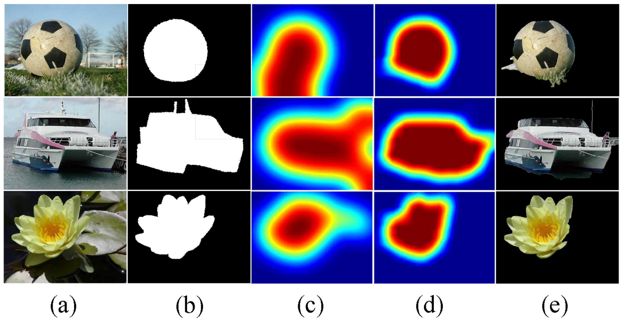


Figure 5: Examples of aggregated confidence maps of selected segments on images from Caltech-101 dataset. (a): Input images; (b): Ground truth object segmentations; (c): Aggregated confidence of selected segments using "FL" method only; (d): Aggregated confidence of selected segments using "FL"+"EN" method; (e): Foreground objects based on masks generated by the results of (d) through adaptive thresholds.

We randomly select 30 images as training data, and eval-

[4] For an image I, the best segment is a segment that has the largest overlap with the ground-truth segment G_I.

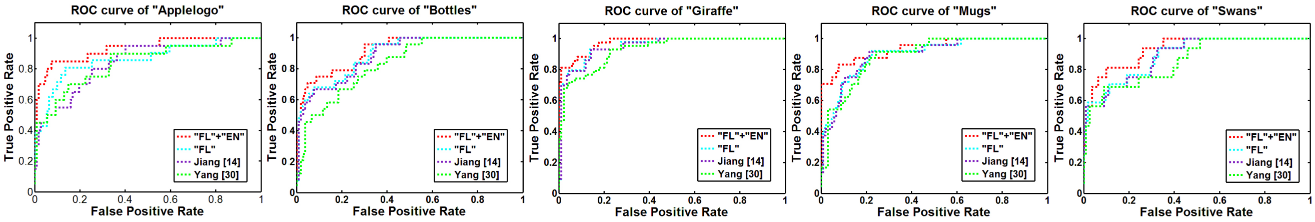
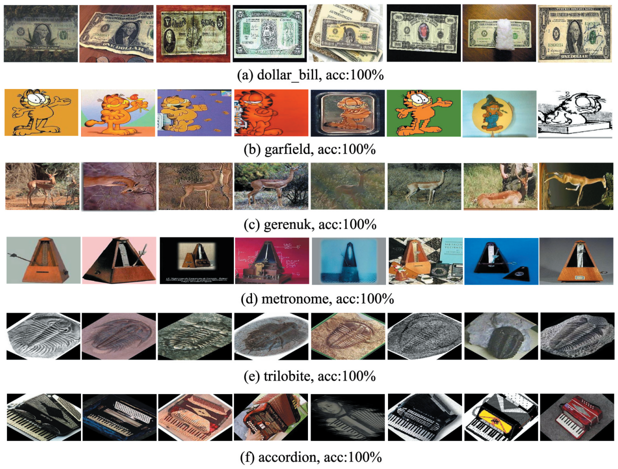


Figure 6: ROC curves of our approaches (“FL” and “FL” +“ EN”) and state-of-the-art approaches on the all five categories of the ETHZ Shape Classes dataset.

Table 2: Recognition accuracies using spatial pyramid features on the Caltech-101 dataset. “BS” and “GT” denote results produced by using only the best ranked segments and ground-truth segments, respectively.

Method	5	10	15	20	25	30
Yang [29]	49.84%	57.26%	62.75%	68.78%	71.12%	73.72%
Jiang [15]	54.00%	63.10%	67.70%	70.50%	72.30%	73.60%
Shaban [26]	54.01%	63.86%	68.70%	71.58%	73.73%	75.07%
Carreira [2]	60.90%	–	74.70%	–	–	81.90%
GT	68.71%	77.67%	81.33%	84.49%	86.73%	88.34%
BS	63.95%	72.03%	77.66%	79.69%	82.24%	83.27%
FL	59.81%	68.45%	73.90%	76.98%	78.96%	80.28%
FL+EN	63.29%	71.47%	76.43%	78.26%	81.03%	83.18%

uate our approach when different values of the entropy term weight λ and the penalty cost δ are selected. As shown in Figure 3, the best performance is achieved when $\lambda = 1.5$ and $\delta = 2$. If λ is set to 0, the performance degrades since segments' purity is not considered. On the other hand, if λ is too large, pursuing segments' purity while considering less on their visual saliency is harmful to the performance. The performance is more sensitive to the penalty cost δ. If δ is large, the cost of opening a new facility can easily exceed the profit that the system can benefit from such a facility. Consequently, only one or a few facilities can be selected. In general, there does not exist a single correct segmentation for an image, so that the performance is weakened when recognition is performed on too few segments within an image (as shown in Figure 4(c) and Figure 4(d)). A small δ can lead to a large collection of segments being selected. Thus, the intersection of selected segments is concentrated on a small image region (as shown in Figure 4(f)), and much object information is discarded. As a result, recognition performance significantly degrades. Figure 5 demonstrates results of aggregated confidence maps of selected segments and resulting foreground objects, and Figure 7 shows example images from classes with high classication accuracy of the Caltech-101 dataset.


(a) dollar_bill, acc:100%

(b) garfield, acc:100%

(c) gerenuk, acc:100%

(d) metronome, acc:100%

(e) trilobite, acc:100%

(f) accordion, acc:100%

Figure 7: Example images from classes with high classication accuracy of the Caltech-101 dataset.

Table 3: Average precisions (APS) of each object category on the ETHZ shape classes dataset.

Methods	Apple	Bottles	Gira	Mugs	Swans	Avg
Yang [29]	83.79	83.13	92.77	89.16	85.75	86.92
Jiang [15]	84.11	88.71	94.74	89.64	88.91	89.22
FL	86.40	89.50	95.04	89.72	89.16	89.96
FL+EN	**93.18**	**91.71**	**97.43**	**93.61**	**92.96**	**93.78**

4.3. ETHZ Shape Classes

The ETHZ Shape Classes dataset [24] contains 255 images from 5 shape categories, including “Applelogo”, “Bottles”, “Giraffes”, “Mugs”, “Swans”, and object ground-truth outlines are provided for all images. Following the PASCAL classification criterion, for each of the 5 categories, we predict presence/absence of an example of that class in the test image. The dataset is evenly split into training and testing sets and performance is averaged over 5 random splits. Performance comparisons between our approaches (“FL” and “FL” +“EN”) and approaches in [29, 15] are given in Table 3. It can be observed that the proposed “FL”+“EN” significantly outperforms other methods. The ROC curves of our approaches and approaches in [29, 15] for the all five categories are shown in Figure 6.

5. Conclusions

We have proposed a greedy object recognition approach based on submodularity. Discriminative segments are selected by maximizing a submodular function, which can be viewed as a facility location problem with the constraint of class purity of selected segments. Segments' categories are estimated by regressors trained within each category. The objective function is optimized by a highly efficient greedy algorithm. Experimental results on three public benchmarks indicate that our method outperforms state-of-the-art recognition techniques.

We plan to apply our approach to object detection tasks. Since our approach can efficiently select potential segments so as to discover the target object, it is superior to traditional sliding window based detection approaches.

References

[1] Y. Bengio, P. Lamblin, D. Popovici, and H. Larochelle. Greedy layer-wise training of deep networks. In *NIPS*, 2007.

[2] J. Carreira, F. Li, and C. Sminchisescu. Object recognition by sequential figure-ground ranking. *International Journal of Computer Vision*, 98(3):243–262, 2012.

[3] J. Carreira and C. Sminchisescu. Constrained parametric min-cuts for automatic object segmentation. In *CVPR*, 2010.

[4] Q. Chen, Z. Song, Y. Hua, Z. Huang, and S. Yan. Hierarchical matching with side information for image classification. In *ICCV*, 2012.

[5] J. Dong, W. Xia, Q. Chen, J. Feng, Z. Huang, and S. Yan. Subcategory-aware object classification. In *ICCV*, 2013.

[6] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2007 (VOC2007) Results. http://www.pascal-network.org/challenges/VOC/voc2007/workshop/index.html.

[7] L. Fei-Fei, R. Fergus, and P. Perona. One-shot learning of object categories. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(4):594–611, 2006.

[8] V. Ferrari, F. Jurie, and C. Schmid. From images to shape models for object detection. *International Journal of Computer Vision*, 87(3):284–303, 2010.

[9] B. Fulkerson, A. Vedaldi, and S. Soatto. Class segmentation and object localization with superpixel neighborhoods. In *ICCV*, 2009.

[10] R. D. Galvão. Uncapacitated facility location problems: contributions. *Pesquisa Operacional*, 24(1):7–38, 2004.

[11] S. Gould, R. Fulton, and D. Koller. Decomposing a scene into geometric and semantically consistent regions. In *ICCV*, 2009.

[12] H. Harzallah, F. Jurie, and C. Schmid. Combining efficient object localization and image classification. In *ICCV*, 2009.

[13] X. He, R. S. Zemel, and M. A. Carreira-Perpinán. Multiscale conditional random fields for image labeling. In *CVPR*, 2004.

[14] Z. Jiang and L. S. Davis. Submodular salient region detection. In *CVPR*, 2013.

[15] Z. Jiang, Z. Lin, and L. Davis. Label consistent k-svd: Learning a discriminative dictionary for recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(11):2651–2664, 2013.

[16] L. Ladicky, C. Russell, P. Kohli, and P. H. Torr. Associative hierarchical crfs for object class image segmentation. In *ICCV*, 2009.

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

[18] N. Lazic, I. Givoni, B. Frey, and P. Aarabi. Floss: Facility location for subspace segmentation. In *ICCV*, 2009.

[19] J. Leskovec, A. Krause, C. Guestrin, C. Faloutsos, J. VanBriesen, and N. Glance. Cost-effective outbreak detection in networks. In *SIGKDD*, 2007.

[20] M.-Y. Liu, R. Chellappa, O. Tuzel, and S. Ramalingam. Entropy-rate clustering: Cluster analysis via maximizing a submodular function subject to a matroid constraint. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36(1):99–112, 2013.

[21] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

[22] M.-E. Nilsback and A. Zisserman. A visual vocabulary for flower classification. In *CVPR*, 2006.

[23] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*. 2010.

[24] T. Quack, V. Ferrari, B. Leibe, and L. Van Gool. Efficient mining of frequent and distinctive feature configurations. In *ICCV*, 2007.

[25] A. Rabinovich, S. Belongie, T. Lange, and J. M. Buhmann. Model order selection and cue combination for image segmentation. In *CVPR*, 2006.

[26] A. Shaban, H. R. Rabiee, M. Farajtabar, and M. Ghazvininejad. From local similarity to global coding: An application to image classification. In *CVPR*, 2013.

[27] J. Shi and J. Malik. Normalized cuts and image segmentation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 22(8):888–905, 2000.

[28] J. Shotton, J. Winn, C. Rother, and A. Criminisi. Textonboost for image understanding: Multi-class object recognition and segmentation by jointly modeling texture, layout, and context. *International Journal of Computer Vision*, 81(1):2–23, 2009.

[29] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009.

Appendix

In this section, we give the proof of monotonicity and submodularity of the entropy term in Equation (5).

A. Monotonicity Proof of $\mathbb{E}(\mathcal{A})$

We prove $\mathbb{E}(\mathcal{A})$ is monotonically increasing by showing $\mathbb{E}(\mathcal{A} \cup a) \geq \mathbb{E}(\mathcal{A})$, for all $a \in \mathcal{V} \backslash \mathcal{A}$ and $A \subseteq \mathcal{V}$, $\rho_a(\mathcal{A}) \geq 0$, where $\rho_a(\cdot)$ is the marginal gain when adding element a.

$$\begin{aligned}
&\mathbb{E}(\mathcal{A} \cup a) - \mathbb{E}(\mathcal{A}) \\
&\quad = -\sum_{j \in \mathcal{A} \cup a} p(j) \log p(j) + \sum_{j \in \mathcal{A}} p(j) \log p(j) \\
&\quad = -p(a) log p(a) \geq 0
\end{aligned} \tag{9}$$

B. Submodularity Proof of $\mathbb{E}(\mathcal{A})$

We prove $\mathbb{E}(\mathcal{A})$ is a submodular function using the diminishing returns definition by showing that for any $\mathcal{A} \subseteq \mathcal{B} \subset \mathcal{V}$, and $a \in \mathcal{V} \setminus \mathcal{B}$, $\rho_a(\mathcal{A}) \geq \rho_a(\mathcal{B})$.

$$\begin{aligned}
&\mathbb{E}(\mathcal{B} \cup a) - \mathbb{E}(\mathcal{B}) \\
&\quad = -\sum_{j \in \mathcal{B}, a \in \mathcal{V}} p(j,a) \log p(j,a) + \sum_{j \in \mathcal{B}} p(j) \log p(j) \\
&\quad = -\sum_{j \in \mathcal{B}, a \in \mathcal{V}} p(j,a) \log p(j,a) + \sum_{j \in \mathcal{B}, a \in \mathcal{V}} p(j,a) \log p(j) \\
&\quad = \sum_{j \in \mathcal{B}, a \in \mathcal{V}} p(j,a) \log \frac{p(j)}{p(j,a)} \\
&\quad = \mathbb{E}(a|\mathcal{B}) \leq \mathbb{E}(a|\mathcal{A}) = \mathbb{E}(\mathcal{A} \cup a) - \mathbb{E}(\mathcal{A})
\end{aligned} \tag{10}$$