Discriminative Cellets Discovery for Fine-Grained Image Categories Retrieval

Luming Zhang[1], Yi Yang[2], Roger Zimmermann[1]
[1]School of Computing,National University of Singapore
[2]School of Information Technology and Electrical Engineering, University of Queensland

ABSTRACT

Fine-grained image categories recognition is a challenging task aiming at distinguishing objects belonging to the same basic-level category, such as leaf or mushroom. It is a useful technique that can be applied for species recognition, face verification, and *etc.* Most of the existing methods have difficulties to automatically detect discriminative object components. In this paper, we propose a new fine-grained image categorization model that can be deemed as an improved version spatial pyramid matching (SPM). Instead of the conventional SPM that enumeratively conducts cell-to-cell matching between images, the proposed model combines multiple cells into cellets that are highly responsive to object fine-grained categories. In particular, we describe object components by cellets that connect spatially adjacent cells from the same pyramid level. Straightforwardly, image categorization can be casted as the matching between cellets extracted from pairwise images. Toward an effective matching process, a hierarchical sparse coding algorithm is derived that represents each cellet by a linear combination of the basis cellets. Further, a linear discriminant analysis (LDA)-like scheme is employed to select the cellets with high discrimination. On the basis of the feature vector built from the selected cellets, fine-grained image categorization is conducted by training a linear SVM. Experimental results on the Caltech-UCSD birds, the Leeds butterflies, and the COSMIC insects data sets demonstrate our model outperforms the state-of-the-art. Besides, the visualized cellets show discriminative object parts are localized accurately.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis and Sensor Fusion

Keywords

Fine-grained, Categories retrieval, Sparse coding, Cellets, Spatial pyramid

1. INTRODUCTION

In the past a few years, many object recognition models have been proposed for multimedia retrieval and analysis [33, 34, 3]. Most of them focus on discriminative learning for detecting and localizing instances belonging to different basic-level categories (*e.g.*, car, cow, and human). Recently, motivated by the application in areas such as agriculture, medicine, and forestry, fine-grained domain recognition has become a hot research topic [4, 5]. For example, researchers are designing image retrieval models to recognize pests found on and around some farm crops such that their species can be monitored. Thereby, suitable chemicals can be taken to mitigate these pests. However, it is still challenging to deal with fine-grained categorization successfully due to two reasons: 1) the difficulty to automatically discover the arbitrary-shaped object components as the examples shown in Fig. 1; and 2) the dynamic backgrounds, occlusions, and variations in lighting conditions, leading to obstacles to learn a robust fine-grained categorization model.

To solve the above problems, we propose a new fine-grained image categorization model that integrates the spatially adjacent cells from a pyramid into discriminative ones. These spatially adjacent cells are called cellets and describe the discriminative object components in a coarse-to-fine manner, as shown in Figure 2. An overview of the proposed framework is presented as follows. By dividing each image into multi-level cells hierarchically, we construct cellets by connecting spatially adjacent cells to describe object components with different scales. To calculate the similarity

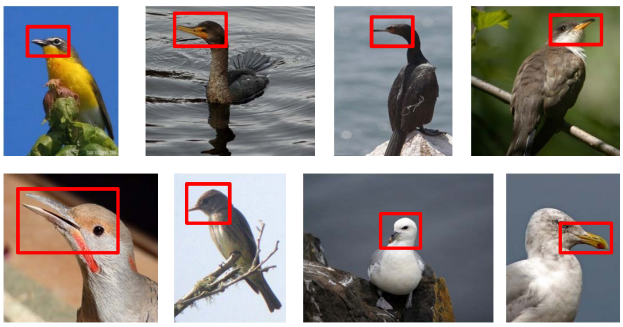

Figure 1: Birds form the Caltech-UCSD data set can be discriminated by their beaks.

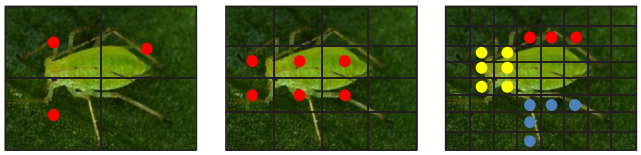

Figure 2: The discriminative cellets hierarchically describe an object. For example, the three different colored cellets reveal different object details.

between images, it is straightforward to match all their corresponding cellets. However, the enumeratively cellet-to-cellet matching is computationally intractable. Toward an efficient matching process, a hierarchical sparse coding algorithm is proposed. The first layer seeks the basis cellets with the same structure; while the second layer finds the basis for cellets with different structures. Afterward, each cellet can be represented by a linear combination of the basis cellets. To accelerate the cellet matching process, we derive the discrimination of a cellet for selecting cellets highly descriptive to the fine-grained categories. Finally, we represent each image by a set of discriminative cellets, which are further converted into an image-level feature vector. These feature vectors are stored for training a linear SVM for fine-grained image categorization.

The main contributions of this work can be summarized as follows: 1) we mine image descriptors highly descriptive to fine-grained categories, based on which an image recognition model is developed, 2) a hierarchical sparse coding algorithm for cellets quantization is introduced, and 3) the first experimental result on a fine-grained data set containing over 5,000 real-world insects is released.

2. RELATED WORK

The proposed fine-grained categorization algorithm is closely related to two research topics in multimedia analysis and pattern recognition: the spatial pyramid matching (SPM) architecture and the fine-grained image categorization.

Toward a general categorization model, SPM is developed that reflects the rough image geometric characteristics. Some researchers [27] have pointed out that the k-means-based codebook in SPM is less effective. Thus, sparse coding [13] and locality preservation [27] techniques are adopted to learned the codebook. Further, to integrate high-level visual features, Li *et al.* proposed object-bank-based SPM [28], wherein an image is described by a scale invariant response map of the pre-defined generic object detectors. In [29], Jia *et al.* adaptively learn the receptive fields for a specific data set. Starting by generating a large number of receptive field candidates, a classifier with structured sparisity is learned for efficiently optimizing the receptive field parameters. In [30], Russakovsky *et al.* proposed object-centric spatial pooling that integrates the object locations in the pooling stage. The key is to learn object detectors using only image-level visual cues. In summary, the pooling components in all these SPM models are based on rectangular receptive fields. Inevitably, these rectangles include large areas irreverent to object components, and thus cannot be applied for fine-grained image categorization.

As the SPM only incorporate coarse visual cues for generic object recognition, models customized for fine-grained categorization have been developed, focusing on discovering tiny discriminative object components. To alleviate the loss of image details in codebook generation, Yao *et al.* [21] represented an image by pooling template matching responses and then designed a bagging mechanism for classification. In [22], Sfar *et al.* proposed to recognize botanical species by combining features related to both basic-level and subordinate-level categories. In [23], Berg *et al.* proposed a grid-level saliency model for fine-grained image categorization. Object parts are aligned and cropped into rectangles, which are then divided into grids and their weights are learned to indicate the discriminative parts. In [24], Deng *et al.* designed a human interactive crowdsourcing system that allows users to localize object parts highly discriminative to fine-grained categories. In [25], Duan *et al.* proposed a fine-grained recognition model that discovers local attributes both discriminative and semantically meaningful, by leveraging the manually annotated object bounding boxes. Further, in [26], a joint object detection and segmentation framework is introduced to localize and normalize an object. Based on it, a state-of-the-art classification model is conducted on the segmented regions for fine-grained categorization.

3. CELLET EXTRACTION

For each image I, we extract a set of D-dimensional local descriptors $\mathbf{X} = [x_1, x_2, \cdots, x_M]^T \in \mathbb{R}^{M \times D}$, wherein x_i denotes the column vector of the i-th local descriptor (*i.e.*, a 128-dimensional SIFT [14] key point), and M denotes the number of local descriptors. To quantize the local descriptors, we use sparse coding to represent each by a linear combination of the basis vectors. In the training stage, we learn a codebook by an alternative optimization:

$$\min_{y_m, \mathbf{B}} \sum\nolimits_m ||x_m - y_m \mathbf{B}||_{l_2}^2 + \lambda_1 ||y_m||_{l_1}, \qquad (1)$$

where $\mathbf{B} \in \mathbb{R}^{R \times D}$ is a codebook with R basis vectors. Normally, codebook $\mathbf{B}$ is an over-complete basis set, *i.e.*, $R > D$.

In the encoding stage, the learned codebook $\mathbf{B}$ is applied to a new set of local descriptors $\mathbf{X}'$ to obtain the encoded local descriptors, *i.e.*,

$$\min_{\mathbf{Y}} \sum\nolimits_m ||x_m - y_m \mathbf{B}||_{l_2}^2 + \lambda_1 ||y_m||_{l_1}, \qquad (2)$$

where $\mathbf{Y} = [y_1, y_2, \cdots, y_{M'}]^T \in \mathbb{R}^{M' \times R}$ is the encoded local descriptors of local descriptor from $\mathbf{X}'$.

To incorporate the spatial information into local descriptors, a spatial pyramid is constructed. In detail, based on the coordinates of the local descriptors, we divide the local descriptors into several cells $\{\mathbf{Y}_{ij}^l\}$, where the script denotes the ij-th cell from the l-th layer spatial pyramid. Formally, the location of cell $\mathbf{Y}_{ij}^l$ is defined as: $\kappa(\mathbf{Y}_{ij}^l) = (l, i, j)$.

Different cells may contain different numbers of local descriptors. Toward a fixed length of feature vector for each cell, the max pooling scheme [6] is applied on each cell:

$$\mathbf{u}_{ij}^l = \xi(\mathbf{Y}_{ij}^l), \qquad (3)$$

where ξ denotes the maximum element on each row of $\mathbf{Y}_{ij}^l$; and $\mathbf{u}_{ij}^l$ denotes a R-dimensional column feature vector.

As we discussed above, the SPM architecture cannot fulfill fine-grained categorization. This is because the cell-to-cell matching includes visual features non-discriminative to object components. To tackle this problem, we propose cellets for spatial pyramid matching. A cellet denotes a set

of spatially adjacent cells $\mathbf{U} = \{u_{ij}^l\}$ associated with their structure, which can be defined as follows:

$$z = [\psi(\mathbf{U}), \phi(\mathbf{U})]^T. \quad (4)$$

The first term $\psi(\mathbf{U}) = \cup_{u \in \mathbf{U}}[u^T]$ denotes a set of spatially

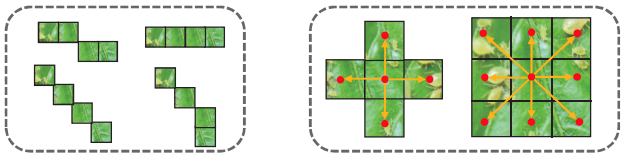

Figure 3: Left: cellets with different structures; Right: spatial relations of cells with different K.

adjacent cells, wherein $\cup[\cdot]$ is a vector concatenation operator. As shown on the left of Figure 3, cellets with the same number of cells may have different structures. These structures are discriminative cues that can contribute to the fine-grained categorization. Thus, we use the second term $\phi(\mathbf{U}) = \text{vec}(\varphi(\mathbf{U}))$ to represent the structure of a cellet. Here, $\text{vec}(\cdot)$ is a row-wise vector stacking operator, and φ is the binary spatial relations between cells in $\mathbf{U}$:

$$\varphi(i,j) = \begin{cases} k & \text{if } \frac{k*\pi}{K} \le \theta(u_i, u_j) \le \frac{(k+1)*\pi}{K} \\ 0 & \text{otherwise} \end{cases}, \quad (5)$$

where $\theta(u_i, u_j)$ denotes the angle between the spatially adjacent cells u_i and u_j and K determines the minimum angle that can be discriminated. We can set K to 4 or 8 depending on different data sets, as shown on the right of Figure 3.

Based on the location of a cell, we define the location of a cellet z by combining the locations of all the cells in z, *i.e.*,

$$\kappa(z) = \cup_{u \in \mathbf{U}}[\kappa(u)]. \quad (6)$$

In summary, the flowchart of generating cellets from an image is presented in Figure 4.

4. HIERARCHICAL CELLECTS ENCODING

Conventional SPM methods cast image categorization as the cellet-to-cellet matching between images. However, there are numerous cellets in a spatial pyramid. Denote $|z|$ as the number of cells that construct cellet z, the number of cellets from the l-th level of spatial pyramid is no less than:

$$H = 4^{l-1} \cdot 4 \cdot 6 \cdot \cdots \cdot (2|z|) = 4^{l-1}|z|!2^{|z|-1}. \quad (7)$$

Thus, given a moderate cellet size, it is intractable to compare pairwise images by enumeratively cellet-to-cellet matching. For fine-grained recognition task, only cellets discriminative to object details should be preserved.

4.1 Hierarchical Sparse Coding of Cellets

To measure the discrimination of a cellet, first we need to measure the similarity between cellets. Because the term $\phi(\mathbf{U})$ in (4) is un-quantized, it is necessary for us to learn a codebook from the training cellets. The codebook allows to represent a cellet by a linear combination of the basis cellets. However, due to the large number of cellets, it is intractable to learn the codebook by employing the training cellets once-for-all, as the standard sparse coding [13].

To solve this problem, a hierarchical sparse coding is proposed that decomposes the encoding task on a large number of cellets into a set of sub-procedures. In the first layer, the basis cellets with the same structure are derived. Particularly, given N training images, we collect cellets with the same structure:

$$\mathbf{Z} = [z_1, z_2, \cdots, z_N], \quad (8)$$

where z_i is the cellet from the i-th training image.

Then, we use sparse coding to find a set of basis cellets:

$$\min_{\mathbf{D}_1} \left\{ \frac{1}{N} \sum_{i=1}^{N} ||z_i - \mathbf{D}_1\alpha_i||_{l_2}^2 + \lambda_2||\alpha_i||_{l_1} \right\}, \quad (9)$$

where $\mathbf{D}_1 \in \mathbb{R}^{S \times T}$ is a codebook learned from $\mathbf{Z}$ and each column of $\mathbf{D}_1$ represents a basis cellet. After the first layer sparse coding, each cellet can be represented by a linear combination of basis cellets with the same structure.

Because the number of basis cellets in $\mathbf{D}_1$ is still large, the second layer sparse coding is proposed to find basis cellets with different structures. Following (7), we obtain H codebooks from the first layer sparse coding:

$$\mathbf{D}_1 = [\mathbf{D}_1^1, \mathbf{D}_1^2, \cdots, \mathbf{D}_1^H], \quad (10)$$

Thereafter, the second layer sparse coding learns a codebook $\mathbf{D}_2$ from $\mathbf{D}_1$, *i.e.*,

$$\min_{\mathbf{D}_2} \left\{ \frac{1}{HT} \sum_{i=1}^{HT} ||d_i - \mathbf{D}_2\beta_i||_{l_2}^2 + \lambda_3||\beta_i||_{l_1} \right\}, \quad (11)$$

where d_i denotes the i-th column of matrix $\mathbf{D}_1$. After the second layer sparse coding, given a new cellet z_{test}, we represent it by a linear combination of the basis cellets:

$$\min_{\gamma} \left\{ ||z_{test} - \mathbf{D}_2\gamma||_{l_2}^2 + \lambda_3||\gamma||_{l_1} \right\}, \quad (12)$$

where $\gamma(z_{test})$ is the sparse representation of cellet z_{test}.

The algorithm of our proposed hierarchical sparse coding of cellets is summarized below.

Algorithm 1 Hierarchical Sparse Coding of Cellets

input: A set of training images $\{I_1, I_2, \cdots, I_N\}$, the size of cellet $|z|$, a test cellet z_{test};
output: the sparse representation of cellet z_{test};

1. Extract SIFT descriptors for each image; use sparse coding to encode them based on (1) and (2);
2. Construct spatial pyramid for each image and obtain cellets with size $|z|$ according to (4);
3. Partition the cellets according to their structure;
   for cellets with the i-th structure **do**
   Compute the first layer codebook $\mathbf{D}_1^i$ from (9);
   end for;
   Compute the second layer codebook $\mathbf{D}_2$ from (11);
4. Obtain the sparse representation of z_{test}.

4.2 Selecting Discriminative Cellets

To select cellets descriptive to the fine-grained categories, we need to measure the discrimination of a cellet. Inspired by the Fisher's linear discriminant analysis [1], the measure of a cellet's discrimination can be defined as:

$$d(z) = \frac{\sum_{z'} ||\gamma(z) - \gamma(z)|| \cdot \sigma(z, z')}{\sum_{z'} ||\gamma(z) - \gamma(z')|| \cdot \sigma'(z, z')}, \quad (13)$$

where $\sigma(z, z')$ and $\sigma'(z, z')$ are functions indicating whether cellets z and z' belonging to the same category, *i.e.*, if z and z' belong to different categories, then $\sigma(z, z') = 1$ and $\sigma'(z, z') = 0$; otherwise $\sigma(z, z') = 0$ and $\sigma'(z, z') = 1$. A

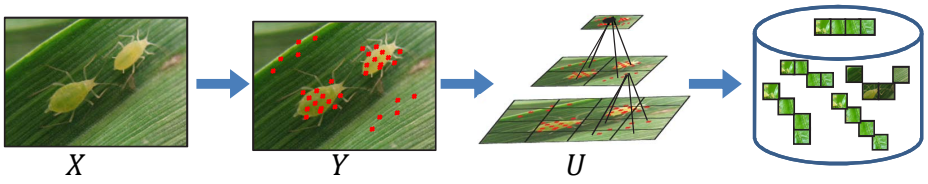

Figure 4: The flowchart of cellets generation (X → Y: sparse coding of local descriptors; Y → U: max pooling of encoded local descriptors in each cell; U → Z: cellets by extracting spatially adjacent cells).

larger $d(z)$ reflects a higher discrimination of cellet z.

In order to compute (13), we have to employ all the training cellets to compare with cellet z, which is highly time-consuming. To accelerate it, following the SPM architecture, only cellets z' with the same location of cellets z are compared. Therefore, (13) can be reformulated as:

$$d'(z) = \frac{\sum_{z' \wedge \kappa(z')=\kappa(z)} ||\gamma(z) - \gamma(z)|| \cdot \sigma(z, z')}{\sum_{z' \wedge \kappa(z')=\kappa(z)} ||\gamma(z) - \gamma(z')|| \cdot \sigma'(z, z')}. \quad (14)$$

It is worth emphasizing another advantage of replacing (13) by (14). In the training stage, the locations of the discriminative cellets are learned implicitly. These learned locations can guide the extraction of the discriminative cellets from the test images. That is to say, there is no need to compute the discrimination of cellets in the test stage, which accelerates the test stage remarkably.

5. THE SVM CLASSIFICATION MODEL

Based on the discrimination measure, for each image, we select fixed number of cellets closely related to its category. These cellets capture the object categorical cues in a coarse-to-fine manner, which can be converted into a vector as:

$$\gamma(\mathbf{Z}^d) = \cup_{z \in \mathbf{Z}^d}[\gamma(z)]. \quad (15)$$

Denoting C as the number of categories, an SVM-based fine-grained classification is carried out by using a pairwise scheme. And $C(C-1)/2$ classifiers are obtained, each of which is trained by images from two different classes. For images from the p-th class and those from the q-th class, we construct a linear SVM classifier as:

$$\min_{\alpha \in \mathbb{R}^{N_{pq}}} \left\{ ||\alpha||^2 + \mathcal{C} \sum_{i=1}^{N_{pq}} l(\alpha, \gamma(\mathbf{Z}_i^d), c_i) \right\}, \quad (16)$$

where $\gamma(\mathbf{Z}_i^d)$ is the feature vector from the i-th training image; $c_i \in \{1, 2, \cdots, C\}$ is the category to the i-th training image; α determines the hyper-plane to separate images in the p-th category from those in the q-th category; and N_{pq} is the number of training images either from the p-th category or from q-th category.

Given a feature vector $\gamma(\mathbf{Z}_{test}^d)$ from a test image, its category (p or q) is predicted by:

$$\text{sgn}(\sum_{i=1}^{N_{pq}} \alpha_i \gamma(\mathbf{Z}_{test}^d) + b). \quad (17)$$

During test, $C(C-1)/2$ times classification will be conducted and the voting rule is utilized for the final decision. That is, each binary classification can be deemed as a voting process, and $\gamma(\mathbf{Z}_{test}^d)$ is assigned to the class with the maximum number of votes.

To summarize the discussion above, we present the pipeline of our proposed model in Algorithm 2.

Algorithm 2 Cellet-Encoded Spatial Pyramid for Fine-grained Image Categorization

//training stage
input: A set of training images $\{I_1, I_2, \cdots, I_N\}$ labeled by the fine-grained image categories;
output: Discriminative cellets, a multi-class linear SVM;
1. For image I_i, obtain its cellets $\{z_i^1, z_i^2, \cdots, z_i^H\}$; Represent each cellet by a set of basis cellets by Alg. 1;
2. For each cellet, compute its discrimination from (14); select the discriminative ones; and represent training image I_i by a feature vector $\gamma(\mathbf{Z}_i^d)$ according to (15);
3. Train a multi-class linear SVM based on (16).
//test stage
input: A test image I_{test}; **output**: The category of I_{test};
1. Obtain the discriminative cellets $\mathbf{Z}_{test}^d$ of image I_{test} based on the location of the discriminative cellets;
2. Compute the feature vector $\gamma(\mathbf{Z}_{test}^d)$; classify it using the trained linear SVM classifier.

5.1 Time Complexity Analysis

The time consumption of our approach is as follows: In the training stage, Step 1 and Step 3 respectively contain one sparse coding; the time consumption of the linear SVM training in Step 4 is $\mathcal{O}(N)$; and Step 2 contains H times sparse coding. In our experiment, H ranges from 100 to 5000, reflecting that the training time consumption is largely determined by the efficiency of sparse coding. Practically, there are many off-the-shelf efficient sparse coding solvers, such as that proposed by Lee *et al.* [13]. Thus, the training time consumption is acceptable. Different from the training phase, the test stage can be carried out rapidly. This is because the time complexity of Step 1 and Step 2 are $\mathcal{O}(|\mathbf{Z}^d|)$ and $\mathcal{O}(1)$ respectively, where $|\mathbf{Z}^d|$ denotes the number of the selected discriminative cellets.

6. EXPERIMENTAL RESULTS

This section justifies the effectiveness of the proposed algorithm based on three experiments. The first experiment compares the proposed cellet-encoded SPM with the other SPM variants and fine-grained categorization models. The second experiment test the influence of different parameter settings. Last but not least, we visualize the discovered cellets, which illustrates the high accuracy of our approach.

Toward a comprehensive evaluation, we complied a new data set termed COSMIC insects, containing insects from 15 categories. The 15 categories are listed in Table 1. In addition, we also experiment on the Caltech-UCSD birds [31] and the Leed butterflies [32]. The experiments are carried out on a computer equipped with an Intel Xeon X5482 CPU and 8GB RAM. All the comparative algorithms are implemented on the Matlab 2011 platform.

Table 1: Statistics of the COSMIC insects data set

Aphids	Armyworm	Bollworm	Colorado.	DBM	FleaBeetle	Jassides	LeafRoller
340	533	444	446	437	237	136	219
Mealybugs	Planthopper	Serpentine.	StinkBug	Thrips	WhiteFly	WhiteGrub	
575	144	202	447	222	362	445	

Table 2: Comparison of categorization accuracies on the three data sets

	Method	Caltech-UCSD	Leeds	COSMIC
30% train	SPM	35.4%	31.6%	42.2%
	SC-SPM	38.9%	32.4%	44.4%
	LLC-SPM	38.7%	31.9%	43.2%
	Our	41.9%	36.4%	45.7%
50% train	SPM	40.1%	36.4%	63.2%
	SC-SPM	44.3%	39.1%	63.8%
	LLC-SPM	44.1%	37.7%	64.5%
	Our	47.8%	41.3%	65.2%

6.1 Comparison with the Existing Methods

In this subsection, we first compare our method with the conventional SPM [2] and its two variants: SC-SPM [6], LLC-SPM [27]. The Matlab codes of all the three compared methods are publicly available[1]. The parameter settings of the compared methods are as follows. For SPM, SC-SPM and LLC-SPM, we construct a three level spatial pyramid; then we extract over one million SIFT descriptors from 16×16 patches computed over a grid with spacing of 8 pixels from all training images. Finally, a codebook with size 256 is generated by k-means [12] clustering on these SIFT descriptors. For each of the three data sets, 30% and 50% images are used respectively for training, while the rest are for testing. We report the categorization accuracy in Table 2. As can be seen, the proposed method outperforms SPM and its two variants, since the discovered cellets are more descriptive to object parts than the conventional cells.

In addition, we compare our method with four existing fine-grained categorization models that are proposed by Yao *et al.* [21], Berg *et al.* [23], Duan *et al.* [25], and Angelova *et al.* [26] respectively. We implement all the four algorithms because their codes are unavailable. Different proportion of training images are used by selecting 10% to 90% training images. As shown in Figure 5, the proposed approach beats all the compared methods. Further, the per-category accuracy on the Caltech-UCSD birds is presented in Figure 6. We compare our approach with Duan *et al.*'s approach, which is the second best in Figure 5. As can be seen, for most categories, the proposed method outperforms Duan *et al.*'s approach significantly.

6.2 Effects of Different Parameter Settings

In this subsection, we study the influence of different parameter settings on the three aforementioned data sets. Particularly, we first set the default values of the parameters as detailed in Table 3. Then, we tune one of the parameters and report the categorization accuracy correspondingly. For convenience, the codebook size and the regularization parameter of the dual sparse codings are set to be equal.

1) We report the performance of our approach with different codebook sizes in Table 4. As seen, by increasing the codebook size from 128 to 256, 512 and 1024, the recognition accuracy improves dramatically. But the improvement becomes smaller when the codebook size is larger than 512.

2) Then we report the recognition accuracy under different regularization parameter of sparse coding. More specifically, we choose the regularization parameter from $[0, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}]$ and report the categorization accuracy in Table 5. As can be seen, the highest categorization is always achieved when the regularization parameter is 10^{-3} or 10^{-2}.

3) Next, we report recognition accuracy with different size of cellets (under 30% training samples on each data set). As seen from Figure 7, when the cellet size is tuned from 1 to 10, the categorization accuracy increases moderately and steadily, but the time consumption increased sharply. This observation demonstrates that in practice, we should choose an optimal cellet size.

4) Then, we present the categorization accuracy with dif-

[1]http://www.ifp.illinois.edu/ jyang29/

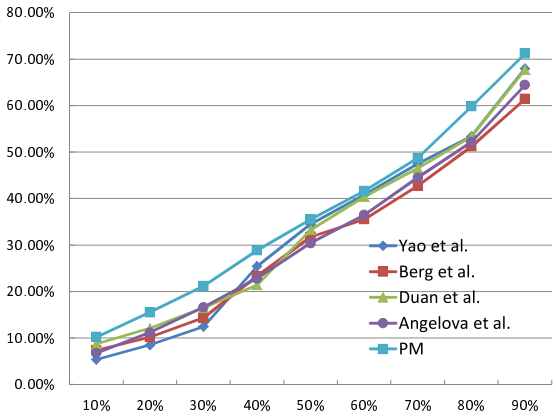

Figure 5: Comparative results with four state-of-the-art fine-grained categorization models (PM: the proposed method).

Table 3: The default parameter settings

Parameter	Caltech-UCSD	Leeds	COSMIC
Codebook size	256	256	256
Regular. para.	10^{-1}	10^{-1}	10^{-1}
Cellet size	4	4	4
Value of K	4	4	8
Pyramid level	3	3	4

Table 4: Comparison of recognition accuracy on the Caltech-UCSD, the Leeds and our own data set

	Code. size	128	256	512	1024
30% train	Catletch-UCSD	41.2%	41.9%	42.4%	42.6%
	Leeds	35.8%	36.4%	37.1%	37.4%
	COSMIC	34.5%	39.5%	40.3%	41.2%
50% train	Caltech-UCSD	47.1%	47.8%	48.4%	48.6%
	Leeds	41.5%	41.3%	41.8%	42.0%
	COSMIC	54.5%	62.1%	54.5%	65.1%

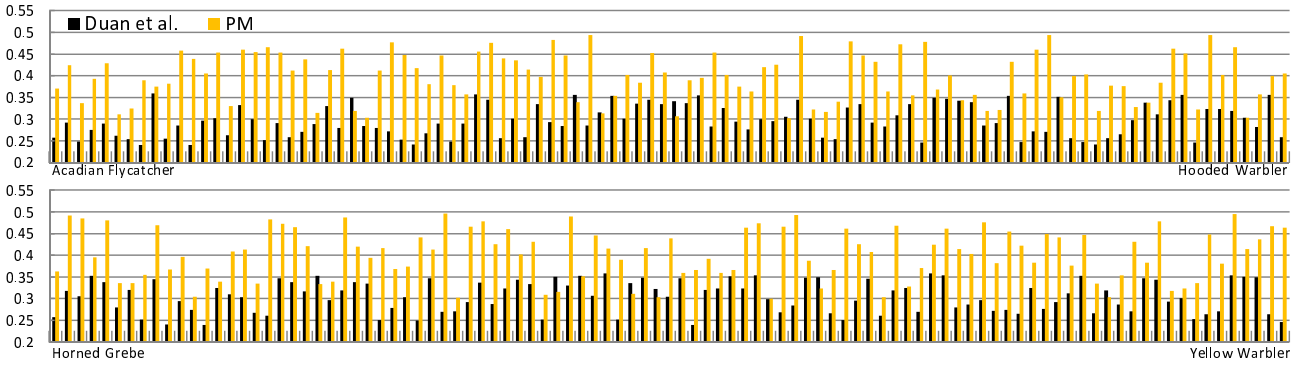

Figure 6: Comparative per-category accuracies of the 200 categories from the Caltech-UCSD birds (the categories are in alphabetical order).

Table 5: Performance under different regularization parameter

	Regu. Para.	0	10^{-4}	10^{-3}	10^{-2}	10^{-1}
30% train	Caltech-UCSD	34.5%	38.7%	41.9%	41.1%	40.1%
	Leeds	27.6%	34.6%	36.4%	35.4%	34.3%
	COSMIC	35.6%	38.7%	40.3%	37.2%	36.9%
50% train	Caltech-UCSD	37.6%	42.4%	45.1%	47.8%	47.1%
	Leeds	32.8%	39.2%	38.4%	41.3%	37.7%
	COSMIC	55.7%	58.7%	59.3%	61.2%	61.9%

Table 6: Performance under different regularization parameter

	K	4	8
30% train	Caltech-UCSD	41.9%	43.2%
	Leeds	36.4%	37.7%
	COSMIC	39.7%	41.2%
50% train	Caltech-UCSD	47.8%	50.1%
	Leeds	41.3%	43.8%
	COSMIC	57.6%	64.7%

Table 7: Performance under different pyramid level

	L	1	2	3	4
30% train	Caltech-UCSD	26.4%	36.9%	41.9%	43.3%
	Leeds	21.2%	31.4%	36.4%	38.4%
	COSMIC	28.5%	32.2%	37.2%	41.7%
50% train	Caltech-UCSD	30.2%	35.1%	41.9%	43.5%
	Leeds	26.2%	3.06%	36.4%	38.4%
	COSMIC	47.4%	54.3%	57.5%	59.8%

ferent values of K, the minimum angle that can be discriminated. As shown from Table 6, the categorization accuracy increases slightly from $K = 4$ to $K = 8$ on the Leeds and the Caltech-UCSD data sets.

5) Last but not least, we report the categorization accuracy with different value of L, the pyramid level, wherein the values is tuned from 1 to 4. As can be shown in Table 7, the recognition accuracy increases dramatically when L increases, but the time consumption of our approach increases significantly also.

6.3 Visualization of Discriminative Cellets

A unique property of our approach is the "transparency" of the fine-grained visual features extraction process. As shown in Figure 8, we visualize the most discriminative cellets from the Catletch-UCSD birds, the COSMIC insects, and the Leeds butterflies data sets. Due to space limitation, only cellets from the last layer spatial pyramid are shown, reflecting the most detailed object discriminative components. As can be seen, the discriminative cellets from different categories have significantly different appearances and structures, which demonstrates the effectiveness of our proposed model. Further, the selected cellets contain little background information, reflecting that our model is robust to the dynamic backgrounds.

7. CONCLUSIONS

This paper proposes a novel fine-grained image categorization framework. By introducing cellet to represent the spatial layout of an image, we cast fine-grained categorization as the matching between cellets from pairwise images. Then, we develop a hierarchical sparse coding algorithm that represents each cellet by a linear combination of the basis cellets. Finally, the discrimination of cellet is derived for selecting a few discriminative cellets for fine-grained categorization. Experimental results thoroughly demonstrate the advantage of the proposed model.

The training time of the proposed method is moderately large, while the testing time consumption is very small. Recently, the cloud computing technique has become a hot research area. Inspired by this concept, the training stage of the proposed method can be computed in a distributed way, *e.g.*, based on a workstation. And the test stage can be calculated based on a the cell phone. Therefore, the training and testing can both be carried out efficiently.

8. ACKNOWLEDGMENT

Figure 8: The visualized discriminative cellets.

This research has been supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

9. REFERENCES

[1] Jieping Ye, Least squares linear discriminant analysis, *in Proc. of ICML*, pages 1087–1093, 2007.

[2] Svetlana Lazebnik, Cordelia Schmid and Jean Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. *in Proc. of CVPR*, pages 2169–2178, 1006.

[3] Yue Gao, Jinhui Tang, Richang Hong, Shuicheng Yan, Qionghai Dai, Naiyao Zhang, Tat-Seng Chua, Camera Constraint-Free View-Based 3D Object Retrieval, *IEEE T-IP*, 21(4), pages: 2269–2281, 2012.

[4] Yue Gao, Meng Wang, Zhengjun Zha, Qi Tian, Qionghai Dai, Naiyao Zhang, Less is More: Efficient 3D Object Retrieval with Query View Selection, *IEEE T-MM*, 11(5), pages: 1007–1018, 2011.

[5] Yue Gao, Meng Wang, Rongrong Ji, Xindong Wu, Qionghai Dai, 3-D Object Retrieval With Hausdorff Distance Learning, *IEEE T-IE*, 61(4), pages: 2088–2098, 2014.

[6] Jianchao Yang; Kai Yu; Yihong Gong; Huang, T. Linear spatial pyramid matching using sparse coding for image classification. *in Proc. of CVPR*, pages: 2169–2178, 2009.

[7] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang and Yihong Gong. Locality-constrained Linear Coding for Image Classification. *in Proc. of CVPR*, 2010.

[8] Harchaoui Z. and Bach, F. Image Classification with Segmentation Graph Kernels. *in Proc. of CVPR*,pages:

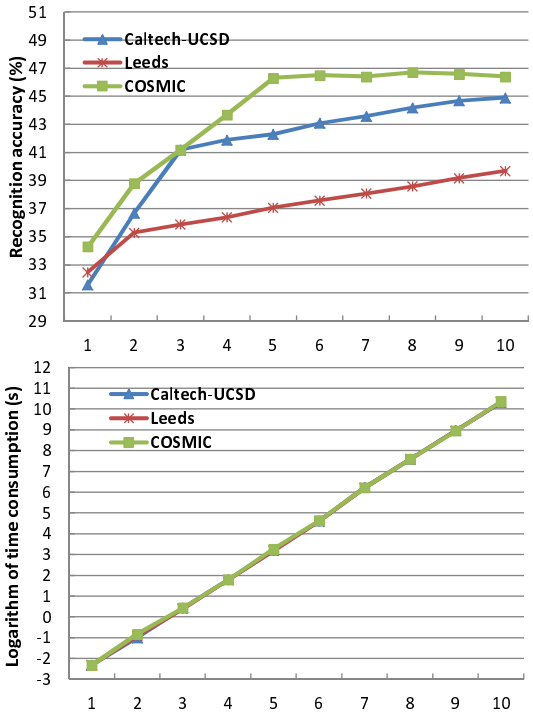

Figure 7: Top: categorization accuracy v.s. size of cellets; bottom: Logarithm of time consumption v.s. size of cellets.

1–8, 2007.

[9] Xi Zhou, Na Cui, Zhen Li, Feng Liang, and Thomas S. Huang. Hierarchical Gaussianization for Image Classification. *In Proc. of ICCV*, pages: 1971–197,2009

[10] Jianxin Wu James M. Rehg. Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. *In Proc. IEEE ICCV*, pages:630–637,2009

[11] J. C. van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. M. Smeulders. Kernel codebooks for scene categorization. *In Proc. ECCV*, pages:696–709,2008

[12] Richard O. Duda, Peter E. Hart and David G. Stork: Pattern Classification. Wiley-Interscience, 2000.

[13] Honglak Lee, Alexis Battle, Rajat Raina, and Andrew Y. Ng. Efficient sparse coding algorithms. *in Proc. of NIPS*, 2006.

[14] J. Porway, K. Wang and B. Yao and S.C. Zhu, Scale-invariant shape features for recognition of object categories. *in Proc. of ICCV*, 2004.

[15] Zaïd Harchaoui, Francis Bach, Image Classification with Segmentation Graph Kernels, *in Proc. of ICCV*, pages: 1–8, 2007.

[16] Nino Shervashidze, S V N Vishwanathan, Tobias Petri, Kurt Mehlhorn, Karsten Borgwardt, Efficient Graphlet Kernels for Large Graph Comparison, *in Proc. of AISTATS*, pages: 488–495, 2009.

[17] Yakov Keselman, ven Dickinson, Generic Model Abstraction from Examples, *IEEE T-PAMI*, 27(7), pages: 1141-1156, 2005.

[18] M. Fatih Demirci, Ali Shokoufandeh, Yakov Keselman, Lars Bretzner, Sven Dickinson, Object Recognition as Many-to-Many Feature Matching, *IJCV*, 69(2), pages: 203–222, 2006.

[19] Pedro F. Felzenszwalb, Daniel P. Huttenlocher, Pictorial Structures for Object Recognition, *IJCV*, 61(1), pages: 55–79, 2005.

[20] Yong Jae Lee, Kristen Grauman, Object-Graphs for Context-Aware Category Discovery, *in Proc. of CVPR*, pages: 346-358, 2009.

[21] Bangpeng Yao, Gary Bradski, Li Fei-Fei, A Codebook-Free and Annotation-Free Approach for Fine-Grained Image Categorization, *in Proc. of CVPR*, pages: 3466–3473, 2012.

[22] Asma Rejeb Sfar, Nozha Boujemaa, Donald Geman, Vantage Feature Frames For Fine-Grained Categorization, *in Proc. of CVPR*, pages: 835–842, 2013.

[23] Thomas Berg, Peter N. Belhumeur, POOF: Part-Based One-vs.-One Features for Fine-Grained Categorization, Face Verification, and Attribute Estimation, *in Proc. of CVPR*, pages: 955–962, 2013.

[24] Jia Deng, Jonathan Krause, Li Fei-Fei, Fine-Grained Crowdsourcing for Fine-Grained Recognition, *in Proc. of CVPR*, pages: 580–587, 2013.

[25] Kun Duan, Devi Parikh, David Crandall, Kristen Grauman, Discovering Localized Attributes for Fine-grained Recognition, *in Proc. of CVPR*, pages: 3474–3481, 2013.

[26] Anelia Angelova, Shenghuo Zhu, Efficient Object Detection and Segmentation for Fine-grained Recognition, *in Proc. of CVPR*, pages: 811–818, 2013.

[27] Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, Yihong Gong, Locality-constrained Linear Coding for Image Classification, *in Proc. of CVPR*, pages: 3360–3367, 2010.

[28] Li-Jia Li, Hao Su, Eric P. Xing, Li Fei-Fei, Object Bank: A High-Level Image Representation for Scene Classification and Semantic Feature Sparsification, *in Proc. of NIPS*, pages: 1378–1386, 2010.

[29] Yangqing Jia, Chang Huang, Trevor Darrell, Beyond Spatial Pyramids: Receptive Field Learning for Pooled Image Features, *in Proc. of CVPR*, pages: 3370–3377, 2012.

[30] Olga Russakovsky, Yuanqing Lin, Kai Yu, Li Fei-Fei, Object-Centric Spatial Pooling for Image Classification, *in Proc. of ECCV*, pages: 1–15, 2012.

[31] Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Caltech-UCSD Birds 200, *California Institute of Technology*, CNS-TR-2010-001, 2010.

[32] Josiah Wang, Katja Markert, Mark Everingham, Object-Centric Spatial Pooling for Image Classification, *in Proc. of BMVC*, pages: 1–11, 2009.

[33] Yue Gao, Meng Wang, Dacheng Tao, Rongrong Ji, Qionghai Dai, 3D Object Retrieval and Recognition with Hypergraph Analysis, *IEEE T-IP*, 21(9), pages: 4290–4303, 2012.

[34] Yue Gao, Meng Wang, Zhengjun Zha, Jialie Shen, Xuelong Li, Xindong Wu, Visual-Textual Joint Relevance Learning for Tag-Based Social Image Search, *IEEE T-IP*, 22(1), pages: 363–376, 2013.