Integrating Randomization and Discrimination for Classifying Human-Object Interaction Activities

Aditya Khosla, Bangpeng Yao and Li Fei-Fei

1 Introduction

Psychologists have shown that the ability of humans to perform basic-level categorization (e.g. cars vs. dogs; kitchen vs. highway) develops well before their ability to perform subordinate-level categorization, or fine-grained visual categorization (e.g. distinguishing dog breeds such as Golden retrievers vs. Labradors) [18]. It is interesting to observe that computer vision research has followed a similar trajectory. Basic-level object and scene recognition has seen great progress [15, 21, 26, 31] while fine-grained categorization has received little attention. Unlike basic-level recognition, even humans might have difficulty with some of the fine-grained categorization [32]. Thus, an automated visual system for this task could be valuable in many applications.

Action recognition in still images can be regarded as a fine-grained classification problem [17] as the action classes only differ by human pose or type of human-object interactions. Unlike traditional object or scene recognition problems where different classes can be distinguished by different parts or coarse spatial layout [16, 21, 15], more detailed visual distinctions need to be explored for fine-grained image classification. The bounding boxes in Figure 1 demarcate the distinguishing characteristics between closely related bird species, or different musical instruments or human poses that differentiate the different playing activities. Models and algorithms designed for basic-level object or image categorization tasks are

Aditya Khosla
MIT, Cambridge, MA, USA, e-mail: khosla@csail.mit.edu

Bangpeng Yao
Stanford University, Stanford, CA, USA, e-mail: bangpeng@cs.stanford.edu

Li Fei-Fei
Stanford University, Stanford, CA, USA, e-mail: feifeili@cs.stanford.edu

An early version of this chapter was presented in [37], and the code is available at http://vision.stanford.edu/discrim_rf/

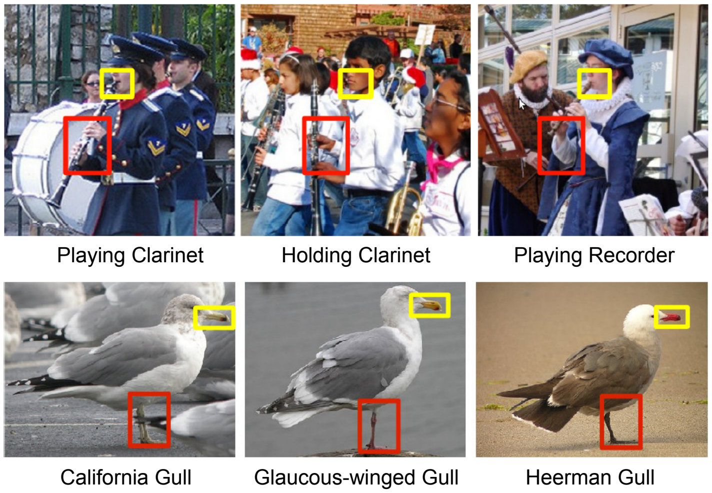


Fig. 1 Human action recognition (top row) is a fine-grained image classification problem, where the human body dominates the image. It is similar to the subordinate object classification problem (bottom row). The red and yellow bounding boxes indicate discriminative image patches for both tasks (manually drawn for illustration). The goal of our algorithm is to discover such discriminative image patches automatically.

often unprepared to capture such subtle differences among the fine-grained visual classes. In this chapter, we approach this problem from the perspective of finding a large number of *image patches* with arbitrary shapes, sizes, or locations, as well as associations between pairs of patches that carry discriminative image statistics [9, 33] (Section 3). However, this approach poses a fundamental challenge: without any feature selection, even a modestly sized image will yield millions or billions of image patches. Furthermore, these patches are highly correlated because many of them overlap significantly. To address these issues, we propose the use of *randomization* that considers a random subset of features at a time.

Specifically, we propose a *random forest with discriminative decision trees* algorithm to discover image patches and pairs of patches that are highly discriminative for fine-grained categorization tasks. Unlike conventional decision trees [8, 4, 2], our algorithm uses strong classifiers at each node and combines information at different depths of the tree to effectively mine a very dense sampling space. Our method significantly improves the strength of the decision trees in the random forest while still maintaining low correlation between the trees. This allows our method to achieve low generalization error according to the theory of random forest [4].

Besides action recognition in still images [33, 11, 12], we evaluate our method on subordinate categorization of closely related animal species [32]. We show that our method achieves state-of-the-art results. Furthermore, our method identifies semantically meaningful image regions[1] that closely match human intuition. Additionally,

[1] We use the terms ‘patches’ and ‘regions’ interchangeably throughout this chapter.

our method tends to automatically generate a coarse-to-fine structure of discriminative image patches, which parallels the human visual system [5].

The remaining part of this chapter is organized as follows: Section 2 discusses related work. Section 3 describes our dense feature space and Section 4 describes our algorithm for mining this space. Experimental results are discussed in Section 5, and Section 6 summarizes this chapter.

2 Related work

Image classification has been studied for many years. Most of the existing work focuses on basic-level categorization such as objects [14, 2, 15] or scenes [26, 13, 21]. In this chapter we focus on two tasks of fine-grained image classification: (1) identifying human-object interaction activities in still images [35, 36, 39, 34], and subordinate-level categorization of animal species [17, 3, 20, 38], which requires an approach that captures the fine and detailed information in images.

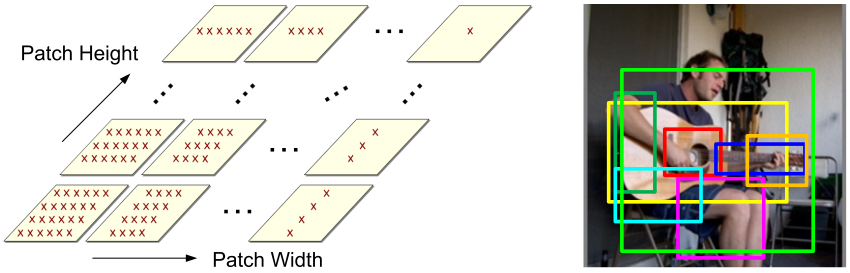
In this chapter, we explore a dense feature representation to distinguish fine-grained image classes. “Grouplet” features [33] have shown the advantage of dense features in classifying human activities. Instead of using the generative local features as in Grouplet, here we consider a richer feature space in a discriminative setting where both local and global visual information are fused together. Inspired by [9, 33], our approach also considers pairwise interactions between image regions.

We use a random forest framework to identify discriminative image regions. Random forests have been used successfully in many vision tasks such as object detection [2], segmentation [27] and codebook learning [24]. Inspired from [28], we combine discriminative training and randomization to obtain an effective classifier with good generalizability. Our method differs from [28] in that for each tree node, we train an SVM classifier from one of the randomly sampled image regions, instead of using AdaBoost to combine weak features from a fixed set of regions. This allows us to explore an extremely large feature set efficiently.

A classical image classification framework [31] is *Feature Extraction* → *Coding* → *Pooling* → *Concatenating*. *Feature extraction* [23] and better *coding* and *pooling* methods [31] have been extensively studied for object recognition. In this work, we use discriminative feature mining and randomization to propose a new feature *concatenating* approach, and demonstrate its effectiveness on fine-grained image categorization tasks.

3 Dense sampling space

Our algorithm aims to identify fine image statistics that are useful for fine-grained categorization. For example, in order to classify whether a human is playing a guitar or holding a guitar without playing it, we want to use the image patches below the



(a) (b)

Fig. 2 Illustration of the proposed dense sampling space. (a) We densely sample rectangular image patches with varying widths and heights. The regions are closely located and have significant overlaps. The red × denote the centers of the patches, and the arrows indicate the increment of the patch width or height. (b) Illustration of some image patches that may be discriminative for "playing-guitar". All those patches can be sampled from our dense sampling space.

human face that are closely related to the human-guitar interaction (Figure 2(b)). An algorithm that can reliably locate such regions is expected to achieve high classification accuracy. We achieve this goal by searching over rectangular image patches of arbitrary width, height, and image location. We refer to this extensive set of image regions as the *dense sampling space*, as shown in Figure 2(a). This figure has been simplified for visual clarity, and the actual density of regions considered in our algorithm is significantly higher. We note that the regions considered by spatial pyramid matching [21] is a very small subset lying along the diagonal of the height-width plane that we consider. Further, to capture more discriminative distinctions, we also consider interactions between pairs of arbitrary patches. The pairwise interactions are modeled by applying concatenation, absolute of difference, or intersection between the feature representations of two image patches.

However, the dense sampling space is very huge. Sampling image patches of size 50×50 in a 400×400 image every four pixels leads to thousands of patches. This increases many-folds when considering regions with arbitrary widths and heights. Further considering pairwise interactions of image patches will effectively lead to trillions of features for each image. In addition, there is much noise and redundancy in this feature set. On the one hand, many image patches are not discriminative for distinguishing different image classes. On the other hand, the image patches are highly overlapped in the dense sampling space, which introduces significant redundancy among these features. Therefore, it is challenging to explore this high-dimensional, noisy, and redundant feature space. In this work, we address this issue using randomization.

```
foreach tree t do
    - Sample a random set of training examples $\mathscr{D}$;
    - SplitNode($\mathscr{D}$);
    if needs to split then
        i. Randomly sample the candidate (pairs of) image regions (Section 4.2);
        ii. Select the best region to split $\mathscr{D}$ into two sets $\mathscr{D}_1$ and $\mathscr{D}_2$ (Section 4.3);
        iii. Go to SplitNode($\mathscr{D}_1$) and SplitNode($\mathscr{D}_2$).
    else
        Return $P_t(c)$ for the current leaf node.
    end
end
```

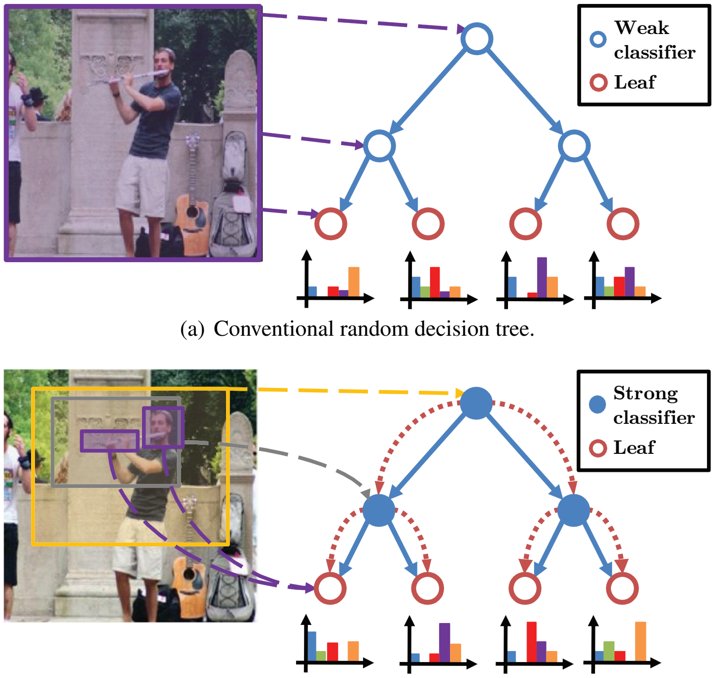
Algorithm 1: Overview of the process of growing decision trees in the random forest framework.

4 Discriminative random forest

In order to explore the dense sampling feature space for fine-grained visual categorization, we combine two concepts: (1) *Discriminative training* to extract the information in the image patches *effectively*; (2) *Randomization* to explore the dense feature space *efficiently*. Specifically, we adopt a random forest [4] framework where each tree node is a discriminative classifier that is trained on one or a pair of image patches. In our setting, the discriminative training and randomization can benefit from each other. We summarize the advantages of our method below:

- The random forest framework allows us to consider a subset of the image regions at a time, which allows us to explore the dense sampling space efficiently in a principled way.
- Random forest selects a best image patch in each node, and therefore it can remove the noise-prone image patches and reduce redundancy in the feature set.
- By using discriminative classifiers to train the tree nodes, our random forest has much stronger decision trees. Further, because of the large number of possible image regions, it is likely that different decision trees will use different image regions, which reduces the correlation between decision trees. Therefore, our method is likely to achieve low generalization error (Section 4.4) compared with the traditional random forest [4] which uses weak classifiers in the tree nodes.

An overview of the random forest framework we use is shown in Algorithm 1. In the following sections, we first describe this framework (Section 4.1). Then we elaborate on our feature sampling (Section 4.2) and split learning (Section 4.3) strategies in detail, and describe the generalization theory [4] of random forest which guarantees the effectiveness of our algorithm (Section 4.4).



(a) Conventional random decision tree.

(b) The proposed discriminative decision tree.

Fig. 3 Comparison of conventional random decision trees with our discriminative decision trees. Solid blue arrows show binary splits of the data. Dotted lines from the shaded image regions indicate the region used at each node. Conventional decision trees use information from the entire image at each node, which encodes no spatial or structural information, while our decision trees sample single or multiple image regions from the dense sampling space (Figure 2(a)). The histograms below the leaf nodes illustrate the posterior probability distribution $P_{t,l}(c)$ (Section 4.1). In (b), dotted red arrows between nodes show our nested tree structure that allows information to flow in a top-down manner. Our approach uses strong classifiers in each node (Section 4.3), while the conventional method uses weak classifiers.

4.1 The random forest framework

Random forest is a multi-class classifier consisting of an ensemble of decision trees where each tree is constructed via some randomization. As illustrated in Figure 3(a), the leaf nodes of each tree encode a distribution over the image classes. All internal nodes contain a binary test that splits the data and sends the splits to its children nodes. The splitting is stopped when a leaf node is encountered. An image is classified by descending each tree and combining the leaf distributions from all the trees. This method allows the flexibility to explore a large feature space effectively because it only considers a subset of features in every tree node.

Each tree returns the posterior probability of an example belonging to the given classes. The posterior probability of a particular class at each leaf node is learned as the proportion of the training images belonging to that class at the given leaf node. The posterior probability of class c at leaf l of tree t is denoted as $P_{t,l}(c)$. Thus, a test image can be classified by averaging the posterior probability from the leaf node of each tree:

$$c^* = \arg\max_{c} \frac{1}{T} \sum_{t=1}^{T} P_{t,l_t}(c), \tag{1}$$

where c^* is the predicted class label, T is the total number of trees, and l_t is the leaf node that the image falls into.

In the following sections, we describe the process of obtaining $P_{t,l}(c)$ using our algorithm. Readers can refer to previous works [4, 2, 27] for more details of the conventional decision tree learning procedure.

4.2 Sampling the dense feature space

As shown in Figure 3(b), each internal node in our decision tree corresponds to a single or a pair of rectangular image regions that are sampled from the dense sampling space (Section 3), where the regions can have many possible widths, heights, and image locations. In order to sample a candidate image region, we first normalize all images to unit width and height, and then randomly sample (x_1, y_1) and (x_2, y_2) from a uniform distribution $U([0,1])$. These coordinates specify two diagonally opposite vertices of a rectangular region. Such regions could correspond to small areas of the image (e.g. the purple bounding boxes in Figure 3(b)) or even the complete image. This allows our method to capture both global and local information in the image.

In our approach, each sampled image region is represented by a histogram of visual descriptors. For a pair of regions, the feature representation is formed by applying histogram operations (e.g. concatenation, intersection, etc.) to the histograms obtained from both regions. Furthermore, the features are augmented with the decision value $\mathbf{w}^{\mathrm{T}}\mathbf{f}$ (described in Section 4.3) of this image from its parent node (indicated by the dashed red lines in Figure 3(b)). Therefore, our feature representation combines the information of all upstream tree nodes that the corresponding image has descended from. We refer to this idea as "nesting". Using feature sampling and nesting, we obtain a candidate set of features, $\mathbf{f} \in \mathbb{R}^n$, corresponding to a candidate image region of the current node.

Implementation details. Our method is flexible to use many different visual descriptors. In this work, we densely extract SIFT [23] descriptors on each image with a spacing of four pixels. The scales of the grids to extract descriptors are 8, 12, 16, 24, and 30. Using k-means clustering, we construct a vocabulary of codewords[2]. Then, we use Locality-constrained Linear Coding [31] to assign the descriptors to

[2] A dictionary size of 1024, 256, 256 is used for PASCAL action [11, 12], PPMI [33], and Caltech-UCSD Birds [32] datasets respectively.

codewords. A bag-of-words histogram representation is used if the area of the patch is smaller than 0.2, while a 2-level or 3-level spatial pyramid is used if the area is between 0.2 and 0.8 or larger than 0.8 respectively. Note that all parameter here are empirically chose. Using other similar parameters will lead to very similar results.

During sampling (step i of Algorithm 1), we consider four settings of image patches: a single image patch and three types of pairwise interactions (concatenation, intersection, and absolute of difference of the two histograms). We sample 25 and 50 image regions (or pairs of regions) in the root node and the first level nodes respectively, and sample 100 regions (or pairs of regions) in all other nodes. Sampling a smaller number of image patches in the root can reduce the correlation between the resulting trees.

4.3 Learning the splits

In this section, we describe the process of learning the binary splits of the data using SVM (step ii in Algorithm 1). This is achieved in two steps: (1) Randomly assigning all examples from each class to a binary label; (2) Using SVM to learn a binary split of the data.

Assume that we have C classes of images at a given node. We uniformly sample C binary variables, $\mathbf{b}$, and assign all examples of a particular class c_i a class label of b_i. As each node performs a binary split of the data, this allows us to learn a simple binary SVM at each node. This improves the scalability of our method to a large number of classes and results in well-balanced trees. Using the feature representation $\mathbf{f}$ of an image region (or pairs of regions) as described in Section 4.2, we find a binary split of the data:

$$\begin{cases} \mathbf{w}^{\mathrm{T}}\mathbf{f} \leq 0, \text{go to left child} \\ \text{otherwise, go to right child} \end{cases}$$

where $\mathbf{w}$ is the set of weights learned from a linear SVM.

We evaluate each binary split that corresponds to an image region or pairs of regions with the information gain criteria [2], which is computed from the complete training images that fall at the current tree node. The splits that maximize the information gain are selected and the splitting process (step iii in Algorithm 1) is repeated with the new splits of the data. The tree splitting stops if a pre-specified maximum tree depth has been reached, or the information gain of the current node is larger than a threshold, or the number of samples in the current node is small.

4.4 Generalization error of random forests

In [4], it has been shown that an upper bound for the generalization error of a random forest is given by

$$\rho(1-s^2)/s^2, \tag{2}$$

where s is the strength of the decision trees in the forest, and ρ is the correlation between the trees. Therefore, the generalization error of a random forest can be reduced by making the decision trees stronger or reducing the correlation between the trees.

In our approach, we learn discriminative SVM classifiers for the tree nodes. Therefore, compared to the traditional random forests where the tree nodes are weak classifiers of randomly generated feature weights [2], our decision trees are much stronger. Furthermore, since we are considering an extremely dense feature space, each decision tree only considers a relatively small subset of image patches. This means there is little correlation between the trees. Therefore, our random forest with discriminative decision trees algorithm can achieve very good performance on fine-grained image classification, where exploring fine image statistics discriminatively is important. In Section 5.5, we show the strength and correlation of different settings of random forests with respect to the number of decision trees, which justifies the above arguments. Please refer to [4] for details about how to compute the strength and correlation values for a random forest.

5 Experiments

In this section, we first evaluate our algorithm on two fine-grained image datasets: actions of people-playing-musical-instrument (PPMI) [33] (Section 5.1) and a subordinate object categorization dataset of 200 bird species [32] (Section 5.2). Experimental results show that our algorithm outperforms state-of-the-art methods on these datasets. Further, we use the proposed method to participate the action classification competition of the PASCAL VOC challenge, and obtain the winning award in both 2011 [11] and 2012 [12]. Detailed results and analysis are shown in Section 5.3 and Section 5.4. Finally, we evaluate the strength and correlation of the decision trees in our method, and compare the result with the other settings of random forests to show why our method can lead to better classification performance (Section 5.5).

Method	BoW	Grouplet [33]	SPM [21]	LLC [31]	Ours
mAP (%)	22.7	36.7	39.1	41.8	**47.0**

Table 1 Mean Average Precision (% mAP) on the 24-class classification problem of the PPMI dataset. The best result is highlighted with bold fonts.

5.1 People-Playing-Musical-Instruments (PPMI)

The people-playing-musical-instrument (PPMI) data set is introduced in [33]. This data set puts emphasis on understanding subtle interactions between humans and objects. Here we use a full version of the dataset which contains twelve musical instruments; for each instrument there are images of people playing the instrument and holding the instrument but not playing it. We evaluate the performance of our method with 100 decision trees on the 24-class classification problem. We compare our method with many previous results[3], including bag of words, grouplet [33], spatial pyramid matching (SPM) [21], locality-constrained linear coding (LLC) [31]. The grouplet method uses one SIFT scale, while all the other methods use multiple SIFT scales described in Section 4.2. Table 1 shows that we significantly outperform the a various of previous approaches.

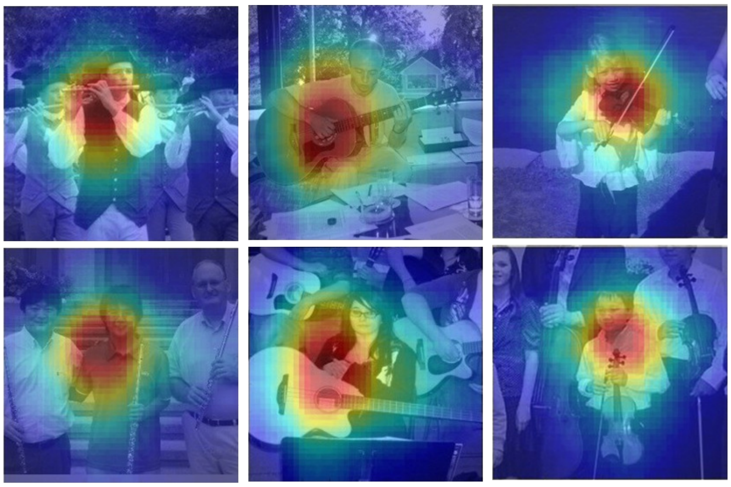
Table 2 shows the result of our method on the 12 binary classification tasks where each task involves distinguishing the activities of playing and not playing for the same instrument. Despite a high baseline of 89.2% mAP, our method outperforms by 2.9% to achieve a result of 92.1% overall. We also perform better than the grouplet approach [33] by 7%, mainly because the random forest approach is more expressive. While each grouplet is encoded by a single visual codeword, each node of the decision trees here corresponds to an SVM classifier. Furthermore, we outperform the baseline methods on nine of the twelve binary classification tasks. In Figure 4, we visualize the heat map of the features learned for this task. We observe that they show semantically meaningful locations of where we would expect the discriminative regions of people playing different instruments to occur. For example, for flute, the region around the face provides important information while for guitar, the region to the left of the torso provides more discriminative information. It is interesting to note that despite the randomization and the algorithm having no prior information, it is able to locate the region of interest reliably.

Furthermore, we also demonstrate that the method learns a coarse-to-fine region of interest for identification. This is similar to the human visual system which is believed to analyze raw input in order from low to high spatial frequencies or from large global shapes to smaller local ones [5]. Figure 5 shows the heat map of the area selected by our classifier as we consider different depths of the decision tree. We observe that our random forest follows a similar coarse-to-fine structure. The average area of the patches selected reduces as the tree depth increases. This shows that the classifier first starts with more global features or high frequency features to

[3] The baseline results are available from the dataset website:
`http://ai.stanford.edu/~bangpeng/ppmi`

Instrument	BoW	Grouplet [33]	SPM [21]	LLC [31]	Ours
Bassoon	73.6	78.5	84.6	85.0	**86.2**
Erhu	82.2	87.6	88.0	89.5	**89.8**
Flute	86.3	95.7	95.3	97.3	**98.6**
French horn	79.0	84.0	93.2	93.6	**97.3**
Guitar	85.1	87.7	**93.7**	92.4	93.0
Saxophone	84.4	87.7	89.5	88.2	**92.4**
Violin	80.6	93.0	93.4	**96.3**	95.7
Trumpet	69.3	76.3	82.5	86.7	**90.0**
Cello	77.3	84.6	85.7	82.3	**86.7**
Clarinet	70.5	82.3	82.7	84.8	**90.4**
Harp	75.0	87.1	92.1	**93.9**	92.8
Recorder	73.0	76.5	78.0	79.1	**92.8**
Average	78.0	85.1	88.2	89.2	**92.1**

Table 2 Comparison of mean Average Precision (% mAP) of the results obtained by different methods on the PPMI binary classification tasks of people playing and holding different musical instruments. Each column shows the results obtained from one method. The best results are highlighted with bold fonts.



(a) flute (b) guitar (c) violin

Fig. 4 (a) Heat map of the dominant regions of interest selected by our method for playing flute on images of playing flute (top row) and holding a flute without playing it (bottom row). (b,c) shows similar images for guitar and violin, respectively. The heat maps are obtained by aggregating image regions of all the tree nodes in the random forest weighted by the probability of the corresponding class. Red indicates high frequency and blue indicates low frequency.

discriminate between multiple classes, and finally zeros in on the specific discriminative regions for some particular classes.

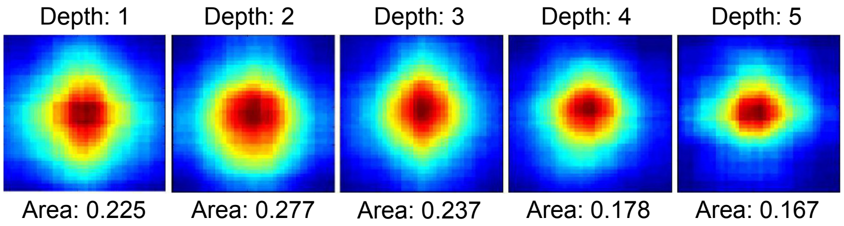


Fig. 5 Heat map for "playing trumpet" class with the weighted average area of selected image regions for each tree depth. Please refer to Figure 4 for how the heat maps are obtained.

Method	MKL [3]	LLC [31]	Ours
Accuracy	19.0%	18.0%	**19.2%**

Table 3 Comparison of the mean classification accuracy of our method and the baseline results on the Caltech-UCSD Birds 200 dataset. The best performance is indicated with bold fonts.

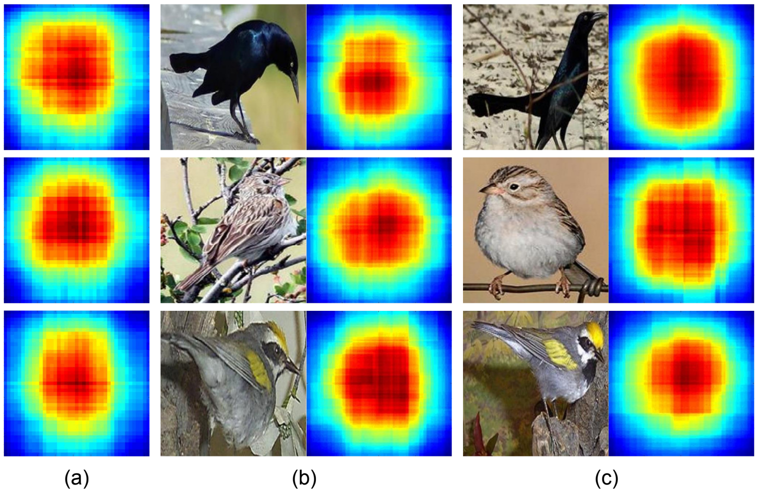


Fig. 6 Each row represents visualizations for a single class of birds (from top to bottom): boat tailed grackle, brewer sparrow, and golden winged warbler. For each class, we visualize: (a) Heat map for the given bird as described in Figure 4; (b,c) Two example images of the corresponding bird and the distribution of image patches selected for the specific image.

5.2 Caltech-UCSD Birds 200 (CUB-200)

The Caltech-UCSD Birds (CUB-200) dataset contains 6,033 annotated images of 200 different bird species [32]. This dataset has been designed for subordinate image categorization. It is a very challenging dataset as the different species are very closely related and have similar shape/color. There are around 30 images per class with 15 for training and the remaining for testing. The test-train splits are fixed (provided on their website).

The images are cropped to the provided bounding box annotations. These regions are resized such that the smaller image dimension is 150 pixels. As color provides important discriminative information, we extract C-SIFT descriptors [29] in the same way described in Section 4.2. We use 300 decision trees in our random forest. Table 3 compares the performance of our algorithm against the LLC baseline and the state-of-the-art result (multiple kernel learning (MKL) [3]) on this dataset. Our method outperforms LLC and achieves comparable performance with the MKL approach. We note that [3] uses multiple features e.g. geometric blur, gray/color SIFT, full image color histograms etc. It is expected that including these features can further improve the performance of our method. Furthermore, we show in Figure 6 that our method is able to capture the intra-class pose variations by focusing on different image regions for different images.

5.3 PASCAL VOC 2011 Action Classification

The recent PASCAL VOC challenges incorporated the task of recognizing actions in still images. The images describe ten common human activities: "Jumping", "Phoning", "Playing a musical instrument", "Reading", "Riding a bicycle or motorcycle", "Riding a horse", "Running", "Taking a photograph", "Using a computer", and "Walking". Each person that we need to classify is indicated by a bounding box and is annotated with one of the nine actions they are performing. There are also humans performing actions that do not belong to any of the ten aforementioned categories. These actions are all labeled as "Other".

We participated the competition using the method proposed in this chapter, and won the winning award in both 2011 [11][4] and 2012 [12][5]. We introduce the details of our results in the 2011 challenge [11] in the rest of this subsection. Section 5.4 will cover our results in the 2012 challenge [12].

There are around 2,500 training/validation images and a similar number of testing images in the 2011 dataset. As in [7], we obtain a foreground image for each person by extending the bounding box of the person to contain 1.5× the original size of the bounding box, and resizing it such that the larger dimension is 300 pixels. We also resize the original image accordingly. Therefore for each person, we have a "person image" as well as a "background image". We only sample regions from the foreground and concatenate the features with a 2-level spatial pyramid of the background. We use 100 decision trees in our random forest.

Classification results measured by mean Average Precision (mAP) are shown in Table 4. Our method achieves the best result on six out of the ten actions. Note that we achieved this accuracy based on only grayscale SIFT descriptors, without using any other features or contextual information like object detectors.

[4] A summary of the results in 2011 PASCAL challenge is in
http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2011/workshop/index.html.

[5] A summary of the results in 2012 PASCAL challenge is in
http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2012/workshop/index.html.

Action	CAENLEAR	CAENLEAR	NUDT	NUDT	Ours
	DSAL	HOBJ_DSAL	CONTEXT	SEMANTIC	
Jumping	62.1%	**71.6%**	65.9%	66.3%	66.0%
Phoning	39.7%	**50.7%**	41.5%	41.3%	41.0%
Playing instrument	60.5%	**77.5%**	57.4%	53.9%	60.0%
Reading	33.6%	37.8%	34.7%	35.2%	**41.5%**
Riding bike	80.8%	86.5%	88.8%	88.8%	**90.0%**
Riding horse	83.6%	89.5%	90.2%	90.0%	**92.1%**
Running	80.3%	83.8%	**87.9%**	87.6%	86.6%
Taking photo	23.2%	25.1%	25.7%	25.5%	**28.8%**
Using computer	53.4%	58.9%	54.5%	53.7%	**62.0%**
Walking	50.2%	59.2%	59.5%	58.2%	**65.9%**

Table 4 Comparison of the mean Average Precision of our method and the other approaches in the action classification competition of PASCAL VOC 2011. Each column shows the result from one method. The best results are highlighted with bold fonts. We omitted the results of MISSOURI_SSLMF and WVU_SVM-PHOW, which did not outperform on any class, due to space limitations.

Figure 7 shows the frequency of an image patch being selected by our method. For each activity, the figure is obtained by considering the features selected in the tree nodes weighted by the proportion of samples of this activity in this node. From the results, we can clearly see the difference of distributions for different activities. For example, the image patches corresponding to human-object interactions are usually highlighted, such as the patches of bikes and books. We can also see that the image patches corresponding to background are not frequently selected. This demonstrates our algorithm's ability to deal with background clutter.

5.4 PASCAL VOC 2012 Action Classification

The action classification competition of the 2012 PASCAL VOC challenge [12] contains more than 5,000 training/validation images and a similar number of testing images, which is an increase of around 90% in size over VOC 2011. We use our proposed method with two improvements: (1) combining multiple features, and (2) greedy tree selection. We describe these in greater detail below. The results are shown in Table 5. In 2012 we had only one competitor (DPM_RF_SVM), and our method outperformed this approach on eight of the ten action classes. Further, comparing "Ours 2012" with "Ours 2011", we observe that combining multiple features and using a tree selection approach[6] improves the performance by 6% mAP.

Combining multiple features: Besides the SIFT image descriptor [23] used in the 2011 challenge, we also consider four other descriptors: HOG2x2 [6], color naming [30], local binary pattern [25], and object bank [22]. These features are

[6] These approaches were specifically developed for the 2012 PASCAL VOC challenge and have not been tested on other datasets but we expect similar performance improvements on them.

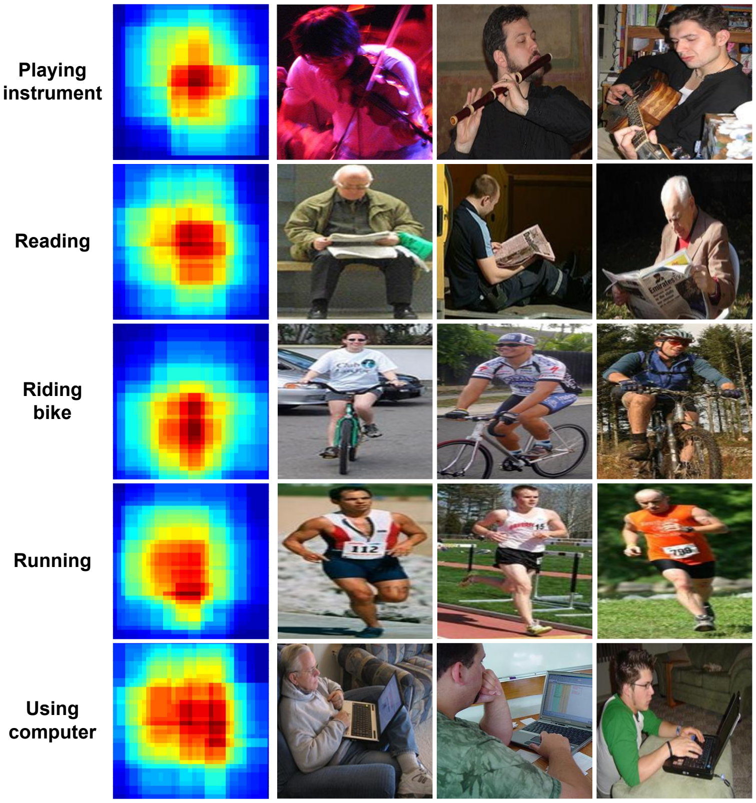


Fig. 7 Heat maps that show distributions of frequency that an image patch is selected in our method. Please refer to Figure 4 for an explanation on how the heat maps are obtained.

extracted in a similar manner to [19]. For HOG2x2 and color naming features, we use a dictionary size of 1024 and 256 respectively. For object bank features, we train the deformable parts-based model (DPM) [15] on the 20 object categories in PASCAL VOC. We build decision trees for each feature independently. Then, we train a linear SVM on the class histograms obtained using the different features to obtain the final output.

Greedy tree selection: Figure 8 illustrates our algorithm. We use training images to train N decision trees independently, and then select the best subset of decision trees based on the validation performance in a greedy manner. We build the forest from decision trees in a sequential manner: first, we evaluate the performance of all individual decision trees on held-out validation data. Then, we select the tree that maximizes the validation performance. This results in a forest with 1 decision tree. We then evaluate the validation performance when we add one more tree from the remaining set of $N-1$ trees and pick the tree that maximizes performance. We repeat this process for N trees, and select the best subset as the first $S \leq N$ trees that maximize the validation performance ($S = 3$ in Figure 8). A greedy method (or

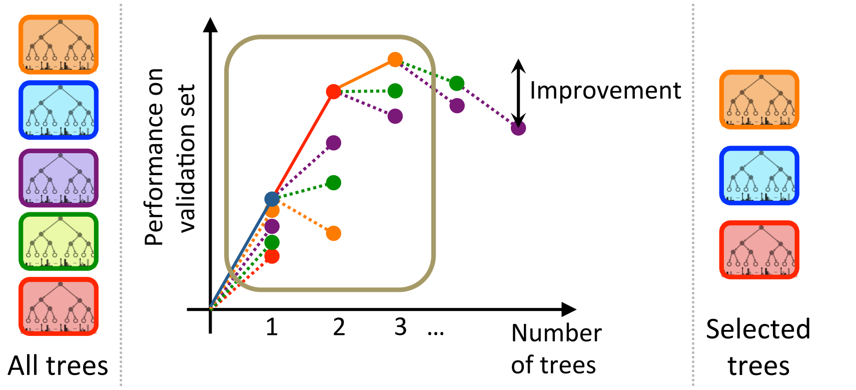


Fig. 8 Illustration of the process of greedy tree selection described in Section 5.4. **Left:** Initially, we start with all the independently trained trees. **Middle:** Then, we measure their performance on the validation data, one at a time. We select the tree with the highest validation performance in the first step (blue), and then choose from the remaining trees in the second step and so on. Overall, the improvement obtained by tree selection is indicated in the figure. **Right:** Selected trees that maximize validation performance.

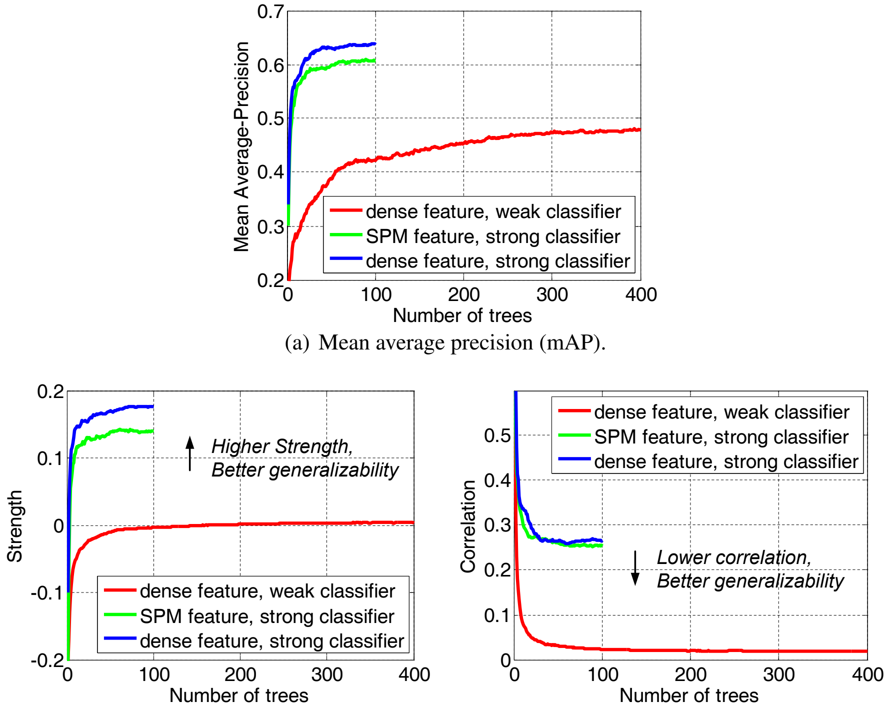
Action	DPM_RF_SVM	Ours 2011	Ours 2012
Jumping	73.8%	71.1%	**75.7%**
Phoning	**45.0%**	41.2%	44.8%
Playing instrument	62.8%	61.9%	**66.6%**
Reading	41.4%	39.3%	**44.4%**
Riding bike	93.0%	92.4%	**93.2%**
Riding horse	93.4%	92.5%	**94.2%**
Running	**87.8%**	86.1%	87.6%
Taking photo	35.0%	31.3%	**38.4%**
Using computer	64.7%	60.4%	**70.6%**
Walking	73.5%	68.9%	**75.6%**

Table 5 Comparison of the mean Average Precision of our method and the other approaches in the action classification competition of PASCAL VOC 2012. "Ours 2011" indicates our approach used for the 2011 challenge. The best results are highlighted with bold fonts.

another approximation) is required as there are too many possible subsets of trees (2^N) to enumerate exhaustively. The idea of tree selection has also been explored in prior works [1].

5.5 Strength and correlation of decision trees

We compare our method against two control settings of random forests on the PASCAL action dataset. Here we use the PASCAL VOC 2010 dataset [10] where there are fewer images than that on 2011 to make our experiments easier to conduct.



(a) Mean average precision (mAP).

(b) Strength of the decision trees.

(c) Correlation between the decision trees

Fig. 9 Comparison of different random forest settings. (a) We compare the classification performance (mAP) obtained by our method dense feature, strong classifier with two control settings. Please refer to Section 5.5 for details of these settings. (b,c) We also compare the strength of the decision trees learned by these approaches and correlation between these trees (Section 4.4), which are highly related to the generalization error of random forests.

- *Dense feature, weak classifier*: For each image region or pairs of regions sampled from our dense sampling space, replace the SVM classifier in our method with a weak classifier as in the conventional decision tree learning approach [8, 4], i.e. randomly generating 100 sets of feature weights and select the best one.
- *SPM feature, strong classifier*: Use SVM classifiers to split the tree nodes as in our method, but the image regions are limited to that from a 4-level spatial pyramid.

Note that all other settings of the above two approaches remain unchanged as compared to our method (as described in Section 4). Figure 9 shows that on this dataset, a set of strong classifiers with relatively high correlation can lead to better performance than a set of weak classifiers with low correlation. We can see that the performance of random forests can be significantly improved by using strong classifiers in the nodes of decision trees. Compared to the random forests that only sample spatial pyramid regions, using the dense sampling space obtains stronger trees without significantly increasing the correlation between different trees, thereby improving the classification performance. Furthermore, the performance of the random forests using discriminative node classifiers converges with a small number of decision trees, indicating that our method is more efficient than the conventional random forest approach. In our experiment, the two settings and our method need a similar amount of time to train a single decision tree.

Additionally, we show the effectiveness of random binary assignment of class labels (Section 4.3) when we train classifiers for each tree node. Here we ignore this step and train a one-vs-all multi-class SVM for each sampled image region or pairs of regions. In this case C sets of weights are obtained when there are C classes of images at the current node. The best set of weights is selected using information gain as before. This setting leads to deeper and significantly unbalanced trees, and the performance decreases to 58.1% with 100 trees. Furthermore, it is highly inefficient as it does not scale well with increasing number of classes.

6 Summary

In this chapter, we propose a random forest with discriminative decision trees algorithm to explore a dense sampling space for fine-grained image categorization. Experimental results on subordinate classification and activity classification show that our method achieves state-of-the-art performance and discovers much semantically meaningful information. One direction for future work is to extend the method to allow for more flexible regions where their location can vary from image to image. Furthermore, it would be interesting to apply other classifiers with analytical solutions such as Linear Discriminant Analysis to speed up the training procedure[7].

References

1. Simon Bernard, Laurent Heutte, and Sébastien Adam. On the selection of decision trees in random forests. In *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*, pages 302–307. IEEE, 2009.
2. A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2007.
3. Steve Branson, Catherine Wah, Boris Babenko, Florian Schroff, Peter Welinder, Pietro Perona, and Serge Belongie. Visual recognition with humans in the loop. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.
4. Leo Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

[7] **Acknowledgements:** L.F-F. is partially supported by an NSF CAREER grant (IIS-0845230), an ONR MURI grant, the DARPA VIRAT program and the DARPA Mind's Eye program. B.Y. is partially supported by the SAP Stanford Graduate Fellowship, and the Microsoft Research Fellowship. A.K. is partially supported by the Facebook Fellowship.

5. Charles A. Collin and Patricia A. McMullen. Subordinate-level categorization relies on high spatial frequencies to a greater degree than basic-level categorization. *Perception & Psychophysics*, 67(2):354–364, 2005.
6. N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
7. Vincent Delaitre, Ivan Laptev, and Josef Sivic. Recognizing human actions in still images: A study of bag-of-features and part-based representations. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2010.
8. T.G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine Learning*, 40:139–157, 2000.
9. Genquan Duan, Chang Huang, Haizhou Ai, and Shihong Lao. Boosting associated pairing comparison features for pedestrian detection. In *Proceedings of the Workshop on Visual Surveillance*, 2009.
10. M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results, 2010.
11. M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2011) Results, 2011.
12. M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2011 (VOC2012) Results, 2012.
13. L. Fei-Fei and P. Perona. A Bayesian hierarchical model for learning natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005.
14. Li Fei-Fei, Rob Fergus, and Antonio Torralba. Recognizing and learning object categories. Short Course in the IEEE International Conference on Computer Vision, 2009.
15. P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminantly trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32:1627–1645, 2010.
16. R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2003.
17. Aharon Bar Hillel and Daphna Weinshall. Subordinate class recognition using relational object models. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2007.
18. Kathy E. Johnson and Amy T. Eilers. Effects of knowledge and development on subordinate level categorization. *Cognitive Development*, 13(4):515–545, 1998.
19. Aditya Khosla, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Memorability of image regions. In *Advances in Neural Information Processing Systems (NIPS)*, Lake Tahoe, USA, December 2012.
20. Aditya Khosla, Bangpeng Yao, Nityananda Jayadevaprakash, and Li Fei-Fei. Novel dataset for fine-grained image categorization. In *First Workshop on Fine-Grained Visual Categorization, IEEE Conference on Computer Vision and Pattern Recognition*, Colorado Springs, CO, 2011.
21. S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.
22. L.-J. Li, H. Su, E. Xing, and L. Fei-Fei. Object bank: A high-level image representation for scene classification and semantic feature sparsification. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2010.
23. David G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
24. Frank Moosmann, Bill Triggs, and Frederic Jurie. Fast discriminative visual codebooks using randomized clustering forests. In *Proceedings of the Conference on Neural Information Processing Systems (NIPS)*, 2007.
25. T. Ojala, M. Pietikainen, and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. In *Proceedings of the IEEE International Conference on Pattern Recognition (ICPR)*, 1994.

26. A. Oliva and A. Torralba. Modeling the shape of the scene: a holistic representation of the shape envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.
27. Jamie Shotton, Matthew Johnson, and Roberto Cipolla. Semantic texton forests for image categorization and segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.
28. Zhuowen Tu. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2005.
29. K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek. Evaluating color descriptors for object and scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(9):1582–1596, 2010.
30. J. van de Weijer, C. Schmid, J. Verbeek, and D. Larlus. Learning color names for real-world applications. *IEEE Transactions on Image Processing*, 18(7):1512–1523, 2009.
31. Jinjun Wang, Jianchao Yang, Kai Yu, Fengjun Lv, Thomas Huang, and Yihong Gong. Locality-constrained linear coding for image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
32. Peter Welinder, Steve Branson, Takeshi Mita, Catherine Wah, Florian Schroff, Serge Belongie, and Pietro Perona. Caltech-UCSD birds 200. Technical Report CNS-TR-201, Caltech, 2010.
33. B. Yao and L. Fei-Fei. Grouplet: A structured image representation for recognizing human and object interactions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
34. B. Yao and L. Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.
35. B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and Li Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2011.
36. B. Yao, A. Khosla, and L. Fei-Fei. Classifying actions and measuring action similarity by modeling the mutual context of objects and human poses. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2011.
37. B. Yao, A. Khosla, and L. Fei-Fei. Combining randomization and discrimination for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.
38. Bangpeng Yao, Gary Bradski, and Li Fei-Fei. A codebook-free and annotation-free approach for fine-grained image categorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.
39. Bangpeng Yao and Li Fei-Fei. Action recognition with exemplar based 2.5D graph matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2012.