An Incremental Structured Part Model for Image Classification

Huigang Zhang[1], Xiao Bai[1], Jian Cheng[2], Jun Zhou[3], and Huijie Zhao[1]

[1] School of Computer Science and Engineering, Beihang University, Beijing, China
baixiao.buaa@googlemail.com
[2] Institute of Automation Chinese Academy of Sciences, Beijing 100190, China
[3] School of Information and Communication Technology, Griffith University, Nathan, QLD 4111, Australia

Abstract. The state-of-the-art image classification methods usually require many training samples to achieve good performance. To tackle this problem, we present a novel incremental method in this paper, which learns a part model to classify objects using only a small number of training samples. Our model captures the inherent connections of the semantic parts of objects and builds structural relationship between them. In the incremental learning stage, we use high entropy images that have been accepted by users to update the learned model. The proposed approach is evaluated on two datasets, which demonstrates its advantages over several alternative classification methods in the literature.

Keywords: Image classification, semantic parts, structural relationship, incremental learning

1 Introduction

Image classification is one of the most important tasks in computer vision and pattern recognition. A number of methods based on the bag-of-words (BOW) model [1] have been proposed to fulfill this task and have shown to be effective for object and scene classification [2, 3]. The BOW method represents an image as a histogram of its local features. It is robust against spatial translations of features, and has demonstrated decent performance in the whole-image classification. However, the BOW method does not sufficiently characterize the spatial relationship between features. Therefore, it is incapable of capturing structural shapes or locating objects in an image.

Structure based methods extract invariant structures to characterize objects in an image [4]. One popular solution is to use graph structure because graph can be used to represent high level vision information. This property has made the graph based methods capable of bridging the low-level local invariant feature with the high-level vision information in images [5, 6]. More recently, part based models have been proposed [7, 8], which operate on image structure rather than solely extracting discrete features.

Learning frameworks have been introduced to further improve the adaptability of statistical and structural image classification methods. Of particular interest is the spatial pyramid matching (SPM) method [9]. It partitions an image into increasingly finer spatial subregions and computes a histogram of local features from each subregion. The same rationale has been employed by several methods, such as sparse coding for linear spatial pyramid matching (ScSPM) [10] and locality-constrained linear coding (LLC) [11]. Similar work also includes the coarse-to-fine learning framework presented by Li et al [12]. In this work, a novel automatic dataset collecting and model learning approach, OPTIMOL, has been developed to refine online picture selection in an incremental way.

Enlightened by these work, we propose an approach to improve the image classification performance via learning semantic parts of objects and exploring their structural relationship. It includes a feature learning method [13] to enrich the part description, and an incremental framework to iteratively update the learned model. Figure 1 illustrates the framework of this classification method. To validate the effectiveness of this method, we have compared it against several state-of-the-art methods in the literature.

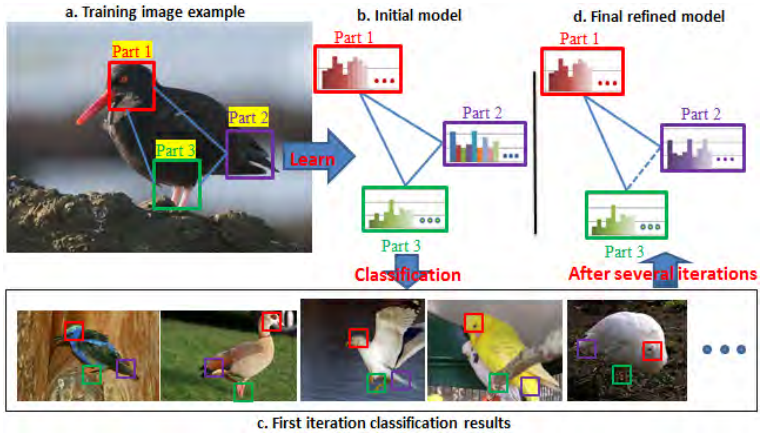

Fig. 1. Framework of the incremental structured part model for image classification. (a) Extracted relevant semantic parts. (b) Training an SVM classifier for each semantic part and building the structured part model. (c) Initial classification results. (d) Iterative model updating using selected images. After several iterations, the model is then updated to a refined model.

The main contribution of this paper is three-fold. Firstly, we propose a part description method that provides abundant mid-level features for image classification. Secondly, the structured part model combines both appearance and structure information of objects in images, which leads to improved classification performance. Thirdly, the incremental learning algorithm can adapt to novel

image features and structures introduced from unseen testing objects. This has greatly reduced the number of training images required.

2 Incremental Structured Part Model

The proposed approach is a combination of both statistical and structural pattern recognition methods. It is based on the observation that different parts of objects in the same class normally share similar spatial relationship. For example, all birds have beaks, legs, and tails, and they follow similar spatial layout. Therefore, we only need to recognize these three parts and model their spatial relationship in order to distinguish birds from other objects (Figure 1(a)).

2.1 Semantic Part Learning

We commence by semantic part learning which allows the treatment of each part as mid-level semantic attribute. We first define the part classes that are important to object classification, then image patches for these parts are manually selected from the training set. From each of these patches, SIFT, texture, color, and edge direction features are extracted. The SIFT features [14] are extracted in a grid-based manner, while the texture descriptors [15] are computed at each pixel using a set of filter banks. To extract the color feature, we use the LAB values [16] of densely sampled pixels. Edges are generated via standard Canny edge detector [17]. Using the bag-of-words model, these four types of features are quantized into vectors with 1000, 256, 128 and 8 dimensions, respectively, and are concatenated into a vector of length 1392.

Using the part feature vectors, a multi-class support vector machines (SVMs) can be learned. Let M be the number of part classes, x_n denotes the n-th training sample and y_n denotes its part class label. The multi-class SVM generates an M-dimensional weight vector $\{w_m^*\}_{m=1}^M$, with one weight for each class. Let W denote a matrix whose columns are w_m. To estimate W, we minimize the following loss function:

$$W^* = \arg\min \sum_n \sum_{t=1}^{M} d(w_t^T x_n, y_{nt}) + \gamma \sum_m \|w_m\|_2^2 \tag{1}$$

where $\gamma \geq 0$ is a tradeoff parameter that regularizes the model complexity, and is set to 0.8 by threefold cross-validation. $d(\cdot\,,\cdot)$ is the loss function.

After solving this optimization problem, we get a semantic part classifier. When an unlabeled image is given, this classifier can be applied to detect relevant parts in the image. In the next section, we explore the structural relationship between these parts.

2.2 Structured Part Model Matching

In this step, we effectively arrange the semantic parts in a deformable configuration to represent an object. The structure model here is inspired by the pictorial structure method [7].

Given an image, let $p_i(l_i)$ be a function measuring the degree of part similarity when part v_i is placed at location l_i. Let $p_{ij}(l_i, l_j)$ be a function measuring the degree of deformation when part v_i is placed at location l_i and part v_j is placed at location l_j. We define the problem of matching a structured part model to an image as a statistical function to be maximized

$$L^* = \arg\max_L (\sum_{i=1}^{n} p_i(l_i) + \lambda \sum_{(v_i,v_j)\in E} p_{ij}(l_i, l_j)) \tag{2}$$

This function maximizes the sum of the matching probabilities $p_i(l_i)$ of each individual part and the deformation similarities $p_{ij}(l_i, l_j)$ for connected pairs of parts. Therefore, it can be decomposed into two equations as follows:

$$L_1^* = \arg\max_L \sum_{i=1}^{n} p_i(l_i) \tag{3}$$

$$L_2^* = \arg\max_L \sum_{(v_i,v_j)\in E} p_{ij}(l_i, l_j) \tag{4}$$

where Eq. 3 is a standard part model and Eq. 4 is a structure model. λ is a parameter that adjust the contribution from the part model and the structure model. It leads to the extension of [7] to a more flexible setting and is self-adaptive through the incremental process to be described later.

We use a sliding window method to detect parts in an unseen image and to compute $p_i(l_i)$. This is achieved by searching the testing images at three different scales, i.e., 0.7, 1, 1.3 times the reference part scale (50×50 pixels), respectively. Using the learned multi-class SVM classifier, we can compute the probability of these candidate patches by fitting a sigmoid function to the original SVM decision values [18]. To compute $p_{ij}(l_i, l_j)$, we use the same method as [7] to calculate the degree of deformation, and fit it to $(0, 1]$ via an exponential function.

The proposed structured part model is robust to missing parts in an image. In Eq. 2, even if one or two p_i is incorrect, high probability still can be achieved on parts from object in the same class due to the contribution from the structure model.

2.3 Coarse-to-fine Updating

Given a very small number of training images of an object class, our algorithm learns the optimal structured part model L^* that best describes this class using the steps introduced above. Now we introduce a coarse-to-fine process to iteratively update L^*, which further improves the robustness of the proposed method.

We randomly separate testing images into several batches and feed them sequentially into the system. Each batch is treated as an iteration. Our incremental process is performed when a new batch comes in. It continuously classifies the

images while learning a more robust model. On each image batch, we compute the probability that the current optimal structured part model matches the images using Eq. 2. The model update is dependent on the image matching results. Images with low matching probability are discarded, while the rest are divided into two sets based on the entropy value generated from the following equation

$$H(I) = -\sum_i p_i \ln p_i - \lambda \sum_E p_{ij} \ln p_{ij} \tag{5}$$

According to Shannon's entropy theory, Eq. 5 relates to the amount of uncertainty of an image I. High entropy indicates high uncertainty of an image, which, in turn, suggests possible new structures. Thus, we choose those images with high probability and high entropy for model updating. Images with high likelihood and low entropy are classified to be positive images. The model updating follows the method introduced in the previous two subsections. It allows refinement of the part classifiers and the corresponding structure model.

At the same time, the weight parameter λ is updated iteratively to make the learned model more robust. In each iteration, the image probabilities are calculated using L_1^* and L_2^*. This can be achieved by setting λ to 0 and 100 (a large enough number) respectively. Let φ_i= {x|x be an image belongs to the positive part using L_i^*}; φ= {x|x is an image belongs to the positive part using L^*}; con_i represents the contribution of model L_i^* to L^*. Then

$$con_i = \frac{\#\{\varphi_i \cap \varphi\}}{\#\{\varphi\}}, i = 1, 2 \tag{6}$$

$$\lambda = \frac{con_2}{con_1} = \frac{\#\{\varphi_2 \cap \varphi\}}{\#\{\varphi_1 \cap \varphi\}} \tag{7}$$

Eq. 7 determines the weights of the part model and the structure model. By calculating λ in each batch, more refined model can be achieved. The proposed coarse-to-fine framework is an iterative process that continuously classifies an image data set with high accuracy while learning a more robust object model. We summarize the steps of our algorithm in Algorithm 1.

3 Experimental Results

We evaluate the performance of the proposed incremental structured part model on two widely used datasets, Caltech-256 [19] and Pascal VOC 2007 [20], and show that only a small number of training images is required for the proposed model (Section 3.1). We also compare our method with other classification methods such as the model by Gritfin et al. [19], ScSPM [10], and LLC [11].

The Caltech-256 dataset contains 30,607 images in 256 categories, with each class containing at least 80 images. The Pascal VOC 2007 dataset consists of 9,963 images from 20 classes. Objects in this dataset reside in cluttered scenes with a high degree of variation in viewing angle, illumination and object appearance. Before the experiments, each image is resized to less than 300×300 pixels with the aspect ratio unchanged. We used all classes in these two datasets for the experiments.

Algorithm 1 Incremental Structured Part Model for Classification

Input: Set of N positive images (N is a small number), set of novel unlabeled images, part number n, and weight λ=1.
Output: Set of classified positive images, and the final Structured Part Model
Initialize Manually select n parts in each training image
Repeat
  Learn Calculate the features of each part in the latest input images and train SVM models. (Sec. 2.1)
  Learn the Structured Part Model. (Sec. 2.2)
  Classify Classify images using the current Structured Part Model. (Sec. 2.2)
  Incremental Use the images with high probability and high entropy for model updating. (Sec. 2.3)
until User satisfied or images exhausted

3.1 Incremental structured part model evaluation

In the first experiment, we randomly chose 5, 10, 15, 20, 25 and 30 training images per class respectively to validate the effectiveness of the proposed method. We consider three baselines to compare our system with: 1) a standard part model L_1^* as in Eq. 3, 2) a structure model L_2^* of Eq. 4, and 3) our structured part model without a coarse-to-fine process. The results are shown in Figure 2. It can be seen that our incremental structured part model outperforms the the baselines by nearly 10 percent. The proposed model is very stable on both datasets when different training sizes are used. At the 5% level, our method achieves classification accuracies that are nearly 10 and 20 percent higher than the alternatives, respectively.

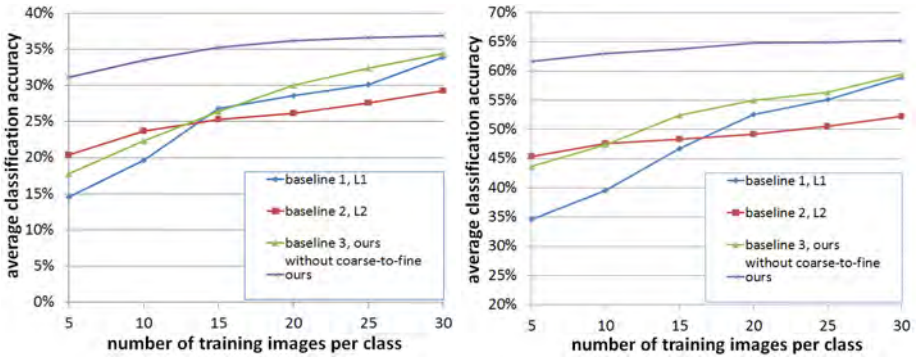

Fig. 2. The average classification results of all the categories in the Caltech-256 dataset (*left*) and Pascal VOC 2007 dataset (*right*), when different training sizes is used.

The reason that our model can achieve good performance under small number of training images is due to the effect of the coarse-to-fine process. By choosing

those images with high entropy, large amount of novel information can be acquired for model updating. The effect of the incremental process is three-fold. Firstly, it can refine the multi-SVM part model. As illustrated in Figure 1, part 2 of the first model is actually a coarse model, as marked by bars in different colors. After several incremental iterations, this model is well refined, which is represented by bars in the same color. Secondly, this process can refine the structural model both in shape and in edge relationships. Take the last model in Figure 1 for example, the dotted line between part 2 and 3 shows that this relationship should be week compared with others, because it's changes in accordance with different birds' postures. Thirdly, the iteration refines the parameter λ in Eq. 2, which leads to a refined global model.

Figure 3 shows some example images with high classification accuracy in the Caltech-256 dataset. We have also tracked those image data with missing parts. The results show that most of them can be classified correctly, which proves the robustness of the proposed method.

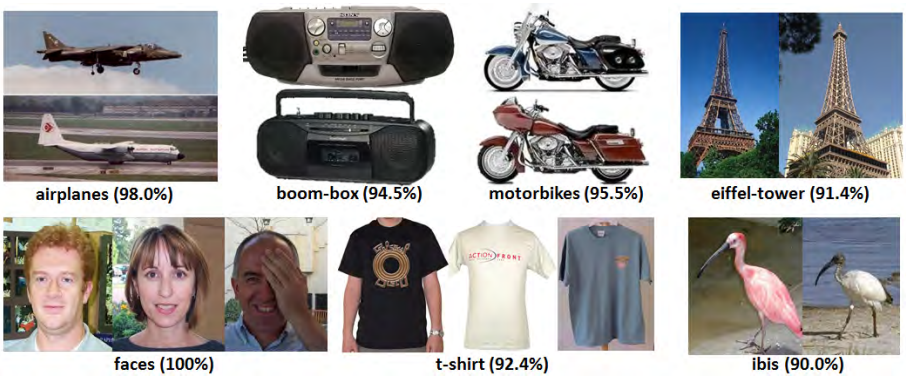

Fig. 3. Example images from categories with high classification accuracy in the Caltech-256 dataset. The percentages in the brackets represent the corresponding classification accuracy.

3.2 Comparison with other classification methods

In this experiment, we first compared the proposed method with several state-of-the-art classification methods on the Pascal dataset. The classification performance is evaluated using the Average Precision (AP) measure. It computes the area under the Precision/Recall curve, in which higher score means better performance. Table 1 shows the classification accuracy on all 20 classes compared against several other classification methods [11, 21, 20]. Our method has achieved the highest accuracy in most classes, especially those with similar shapes such as *bicycle and motorbike*, *cat and dog*, *cow and sheep*. The results show that our

Table 1. Image classification results on Pascal VOC 2007 dataset.

Category	aero	bicyc	bird	boat	bottle	bus	car	cat	chair	cow
PASCAL 07 Best [20]	**77.5**	63.6	56.1	71.9	33.1	60.6	78	58.8	53.5	42.6
LLC [11]	74.8	65.2	50.7	70.9	28.7	68.8	**78.5**	61.7	54.3	48.6
Su [21]	76.2	66.4	**59.2**	70.3	35.4	63.6	79.4	62.4	59.5	47.9
ours	77.1	**73.0**	54.8	**75.2**	**37.2**	**70.3**	72.4	**65.7**	**60.6**	**50.8**
Category	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv
PASCAL 07 Best [20]	54.9	45.8	77.5	64.0	85.9	36.3	44.7	50.9	79.2	53.2
LLC [11]	51.8	44.1	76.6	66.9	83.5	30.8	44.6	53.4	78.2	53.5
Su [21]	**58.8**	44.9	**78.3**	67.4	**87.9**	32.9	46.9	**53.8**	78.6	**58.9**
ours	57.5	**49.3**	75.7	**72.9**	77.2	**42.1**	**47.9**	51.5	**80.6**	58.6

semantic part model is capable of extracting features and their structural relationships in order to distinguish similar objects. We also tested the method on Caltech-256 dataset, in which we used 5, 15, and 30 training images per class. Detailed results are shown in Table 2. It suggests that our method leads the performance with a small number of training images.

Table 2. Image classification results on Caltech-256 dataset.

Algorithms	5 training	15 training	30 training
Gritfin et al. [19]	18.40	28.30	34.10
ScSPM [10]	-	27.73	34.02
LLC [11]	-	34.36	**41.19**
ours	**31.15**	**35.22**	36.87

4 Conclusion

In this paper we have proposed a novel incremental structured part model for image classification. This method first builds image classification models by incorporating both advantages from semantic parts and their structural relation description. Then an incremental framework is employed to refine the model iteratively, which makes the proposed method more robust. This method requires only a small number of training images to achieve good classification performance. Future work will explore the use of hierarchical segmentations to find the semantic parts at the training stage. We will also investigate other features to train the part classifier.

References

1. Fei-Fei, L., Fergus, R., Torralba, A.: Recognizing and learning object categories. In: ICCV Short Course. (2005)

2. Yang, L., Jin, R., Sukthankar, R., Jurie, F.: Unifying discriminative visual codebook generation with classifier training for object category recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR'08 (2008) 1–8
3. Fergus, R., Fei-Fei, L., Perona, P., Zisserman, A.: Learning object categories from google's image search. In: IEEE International Conference on Computer Vision. Volume 2., ICCV'05 (2005) 1816–1823
4. Bunke, H., Sanfeliu, A.: Syntactic and structural pattern recognition: theory and applications. Volume 7. World Scientific Pub Co Inc (1990)
5. Xiao, B., Hancock, E., Wilson, R.: Graph characteristics from the heat kernel trace. Pattern Recognition **42**(11) (2009) 2589–2606
6. Wilson, R., Hancock, E., Luo, B.: Pattern vectors from algebraic graph theory. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(7) (2005) 1112–1124
7. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. International Journal of Computer Vision **61**(1) (2005) 55–79
8. Felzenszwalb, P., Girshick, R., McAllester, D., Ramanan, D.: Object detection with discriminatively trained part-based models. IEEE Transactions on Pattern Analysis and Machine Intelligence **32**(9) (2010) 1627–1645
9. Grauman, K., Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features. In: IEEE International Conference on Computer Vision. Volume 2., ICCV'05 (2005) 1458–1465
10. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR'09 (2009) 1794–1801
11. Wang, J., Yang, J., Yu, K., Lv, F., Huang, T., Gong, Y.: Locality-constrained linear coding for image classification. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR'10 (2010) 3360–3367
12. Li, L., Fei-Fei, L.: Optimol: Automatic online picture collection via incremental model learning. International Journal of Computer Vision **88**(2) (2010) 147–168
13. Farhadi, A., Endres, I., Hoiem, D., Forsyth, D.: Describing objects by their attributes. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR'09 (2009) 1778–1785
14. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2) (2004) 91–110
15. Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal of Computer Vision **62**(1) (2005) 61–81
16. Wyszecki, G., Stiles, W.: Color Science: Concepts and Methods, Quantitative Data and Formulae. Wiley, New York (1982)
17. Canny, J.: A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence (6) (1986) 679–698
18. Platt, J., et al.: Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. Advances in large margin classifiers **10**(3) (1999) 61–74
19. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. (2007)
20. Everingham, M., Van Gool, L., Williams, C., Winn, J., Zisserman, A.: The pascal visual object classes challenge 2007 (voc2007) results. (2007)
21. Su, Y., Allan, M., Jurie, F.: Improving object classification using semantic attributes. In: Proceedings of the British Machine Vision Conference, BMVC'10 (2010) 26–1