Subcategory-aware Object Classification

Jian Dong[1], Wei Xia[1], Qiang Chen[1], Jianshi Feng[1], Zhongyang Huang[2], Shuicheng Yan[1]
[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Panasonic Singapore Laboratories, Singapore
{a0068947, weixia, chenqiang, jiashi, eleyans}@nus.edu.sg,{zhongyang.huang}@sg.panasonic.com

Abstract

In this paper, we introduce a subcategory-aware object classification framework to boost category level object classification performance. Motivated by the observation of considerable intra-class diversities and inter-class ambiguities in many current object classification datasets, we explicitly split data into subcategories by ambiguity guided subcategory mining. We then train an individual model for each subcategory rather than attempt to represent an object category with a monolithic model. More specifically, we build the instance affinity graph by combining both intra-class similarity and inter-class ambiguity. Visual subcategories, which correspond to the dense subgraphs, are detected by the graph shift algorithm and seamlessly integrated into the state-of-the-art detection assisted classification framework. Finally the responses from subcategory models are aggregated by subcategory-aware kernel regression. The extensive experiments over the PASCAL VOC 2007 and PASCAL VOC 2010 databases show the state-of-the-art performance from our framework.

1. Introduction

Category level classification based on bag-of-words (BoW) framework [14, 23, 35, 17, 5] has achieved significant advances during the past few years. This framework combines local feature extraction, feature encoding and feature pooling to generate global image representations, and represents each object category with a monolithic model, such as a support vector machine classifier. However, the large intra-class diversities induced by pose, viewpoint and appearance variations [27] make it difficult to build an accurate monolithic model for each category, especially when there are many ambiguous samples. For example, the chair category in Figure 1 includes three obvious subcategories, namely, sofa-like chairs, rigid-material chairs and common chairs. In feature space, these subcategories are essentially far away from each other. Furthermore, the ambiguous sofa-like chairs look more like sofas than common chairs. In this case, representing all chairs with a monolithic model will weaken the model separating capacity and cannot distinguish sofas from chairs. Hence, it is intuitively beneficial to model each subcategory independently. These considerable intra-class diversities and inter-class ambiguities are common in the challenging real world datasets [13, 37], which makes the subcategory mining necessary.

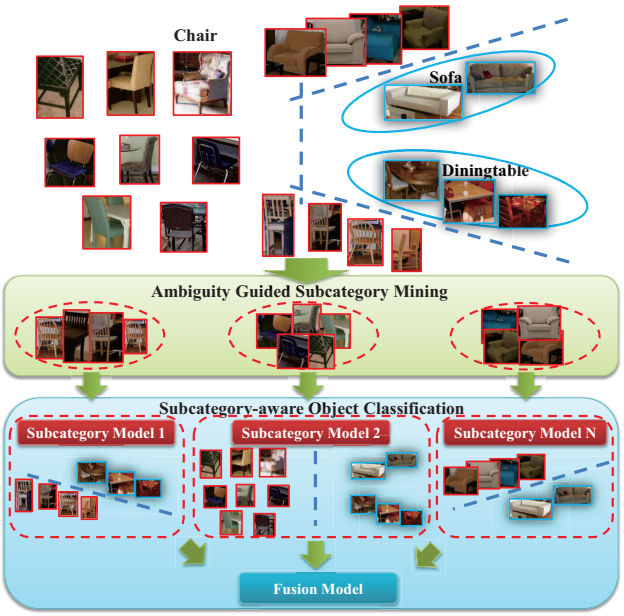

Figure 1: Overview of the proposed ambiguity guided subcategory mining and subcategory-aware object classification framework. For each category, training samples are automatically grouped into subcategories based on both intra-class similarity and inter-class ambiguity. An individual subcategory model is constructed for each detected subcategory. The final classification results are obtained by aggregating responses from all subcategory models.

Clustering all training data of an object category based on intra-class similarity seems to be a natural strategy for subcategory mining, since objects belonging to the same subcategory should intuitively have larger similarity in terms of appearance and shape. However, in the context of generic object classification, subcategories mined with only intra-class visual similarity cues are unnecessary to be

1063-6919/13 $26.00 © 2013 IEEE
DOI 10.1109/CVPR.2013.112

optimal due to the ignorance of valuable inter-class information [8]. More specifically, if the samples are clustered by standard clustering methods, we are unable to utilize the valuable inter-class information to handle the ambiguous samples. Then all ambiguous samples, which often lie near the decision boundary, may be grouped together and preserve the original complicated decision boundary. On the contrary, with the assistance of inter-class information ambiguous samples can be grouped into proper subcategories, which leads to easier subproblems and further improves the overall performance. For instance, chair category and other categories in Figure 1 have non-linear decision boundary. By noting the ambiguous chair sample distribution near the decision boundary, all chairs should be intuitively divided into separate subcategories. The proper split as indicated in Figure 1 will make all subcategories linearly separable from other categories, which is only achievable with the assistance of inter-class information.

The above observation inspires us to propose an ambiguity guided subcategory mining approach to explore the intrinsic subcategory structure embedded in each category. With subcategory awareness, we can boost category level classification by subcategory-aware object classification (SAOC). As indicated in Figure 1, we split data into subcategories by ambiguity guided subcategory mining and train an individual model for each subcategory. Since the diversities in each subcategory and ambiguities between subcategories and other categories are reduced, more accurate shape-based [9, 16]/appearance-based [33, 24] detectors and foreground classification model [5] can be built, which fits nicely with the state-of-the-art detection assisted classification framework [21, 31]. The final classification results are generated by aggregating subcategory responses through subcategory-aware kernel regression.

The main contributions of this paper are summarized as follows. First, we propose a novel ambiguity guided subcategory mining approach, which gracefully integrates the intra-class similarity and inter-class ambiguity for effective subcategory mining. Second, we provide an effective subcategory-aware object classification framework based on the current detection assisted classification framework [21, 31]. Our ambiguity guided subcategory mining approach can be seamlessly integrated into such framework. Utilizing mined subcategories can improve both detection and classification performance and allow more effective subcategory level interaction in the fusion model. The state-of-the-art classification results on the PASCAL VOC datasets verify the effectiveness of our new framework.

2. Related Work

Many state-of-the-art image classification systems follow the popular local feature extraction-coding-pooling pipeline [14]. First, local features like HOG [9], SIFT [26] and LBP [28] are extracted on the dense grids or sparse interest points. They are then encoded by vector quantization (VQ), locally-constrained linear coding (LLC) [35] or Fisher kernel (FK) [17]. Finally the encoded vectors are pooled together to form the image-level representation [23, 5]. Much research on image classification has been focused on improving this pipeline [35, 17, 4]. Some recent works [31, 21, 24, 33, 30] begin to investigate out of this pipeline. Harzallah et al. [21] introduced the pioneering work for detection and classification contextualization, the extension of which leads to the state-of-the-art results [31, 5, 30]. However, all the above methods train a monolithic model for each category, and there are few works analyzing the data structure embedded in each category. In this work, we show that properly splitting the data into subcategories will boost the performance of the state-of-the-art pipeline.

Object detection [16] is another central problem in object recognition, which is complementary to object classification [31, 21]. As most standard semantic categories do not form coherent visual categories, mixture models are proposed and have become the standard approach for object detection [39, 16]. Early works only investigate heuristics based on meta-data or manual labels such as bounding box aspect ratio [16], object scale [29], object viewpoint [19] and part labels [3] to group the positive samples into clusters. However, each of these methods has its own limitations and ignores other more general intra-class variations such as appearance and shape variance [27, 18]. Malisiewicz et al. [27] handled the intra-class variation by training a separate model for each positive instance, which inevitably reduces the generalization capacity of each model. Some recent works begin to investigate the visual subcategory structure embedded in each category [10, 18, 7, 39, 1, 11], which leads to considerable improvement in object detection performance. In contrast to our method, they either require manual annotation or are fragile to outliers corresponding to highly occluded or strange samples. Furthermore, these methods discard the inter-class information during data grouping, which is critical for object classification.

When the data has a complex non-linear structure, locally adaptive classifiers are usually superior to the use of a single global classifier [32, 22, 8]. Kim and Kittler placed the local classifiers at the clusters obtained by the K-means clustering algorithm [22]. Instead of placing the classifiers based on the data distribution only, Dai et al. [8] proposed a responsibility mixture model that uses the uncertainty associated with the classification at each training sample. Using this model, the local classifiers are placed near the decision boundary where they are most effective. In this work, we borrow the idea of uncertainty piloted classification and propose an ambiguity guided subcategory mining approach under the graph shift [25] framework.

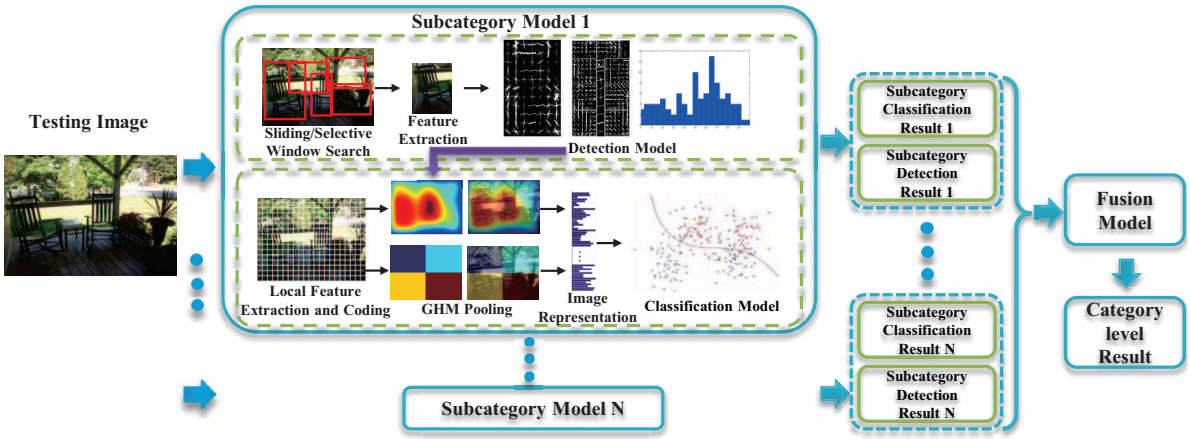

Figure 2: Diagrammatic flowchart of the proposed subcategory-aware object classification framework. Given a testing image, they are first processed by each learnt subcategory model including detection and classification models. Then the responses from all subcategory models are fed into the fusion model to generate the final category level classification results.

3. Subcategory-aware Object Classification

Figure 2 shows the diagrammatic flowchart of our SAOC framework. We will first introduce each component of the framework and emphasize how subcategory mining fits into each step later.

As shown in Figure 2, the whole framework consists of three models - detection, classification and fusion models. For detection, each subcategory is characterized by one shape-based sliding window detector [16, 38] and one appearance-based selective window detector [34, 33], respectively. The usage of two detectors is to guarantee both high precision and high recall on object detection since none of the detectors can achieve this alone and they complement each other. For classification, we follow the state-of-the-art pipeline [5] and train a classifier for each subcategory individually. Since the background is cluttered and many of the concerned object classes may co-occur in a single image, detection confidence maps are employed as the side information for Generalized Hierarchical Matching (GHM) pooling proposed in [5]. The fusion model mainly aims to: (1) boost the classification performance by complementary detection results, (2) utilize the context of all categories for reweighting, and (3) fuse the subcategory level results into final category level results. All of these are achieved by kernel regression. First, we construct a middle level representation for each training/testing image by concatenating classification scores and the leading two detection scores from each subcategory model. The final category level classification results are then obtained by performing Gaussian kernel regression on this representation. Without sophisticated models and complicated postprocessing [12, 31], our subcategory-aware kernel regression is very efficient and still performs well experimentally.

Subcategory awareness, which benefits each model separately and then boosts the overall performance of the framework, plays a critical role in extending current detection assisted classification framework. 1) The subcategory information can be used to initialize both detection and classification models to better handle the rich intra-class diversities in challenging datasets. Less diversity in each subcategory will lead to a simpler learning problem, which can be better characterized by current state-of-the-art models, such as the Deformable Part based Model (DPM) for detection and the foreground BoW models involved in GHM. 2) The subcategory awareness will lead to more effective fusion models. First, subcategory awareness allows us to model the subcategory level interaction. For example, occluded chairs and sitting persons often occur together and should boost the classification scores of each other. On the contrary, unoccluded chairs and pedestrians are independent and should not boost each other. However, these two different cases cannot be differentiated in the category level. Only by subcategory awareness can such underlying correlation be captured effectively. Second, the subcategory awareness is able to reduce the false boosting caused by ambiguity. For example, diningtables often appear together with common chairs, which leads to mutual boosting in classification. Sofas and diningtables are independent and should not boost each other. If sofas are misclassified as chairs, the dinningtable scores may be incorrectly boosted and lead to false alarms on diningtables in category level interaction. With subcategory awareness, the response of diningtable will not be boosted as there is no boosting correlation between the sofa-like chairs and diningtables.

4. Ambiguity Guided Subcategory Mining

In this section, we will introduce how to find the subcategories by our ambiguity guided subcategory mining approach as illustrated in Figure 3. Before digging into details, we first summarize the notations used in this work. For a classification problem, a training set of M samples are given and represented by the matrix X =

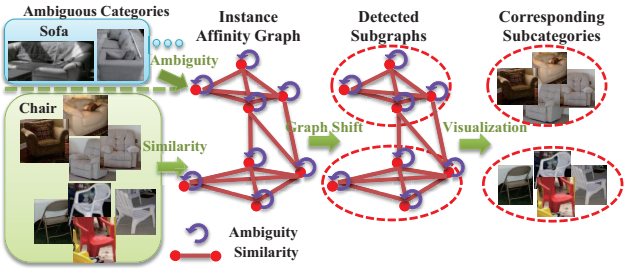

Figure 3: Ambiguity guided subcategory mining approach. First instance affinity graph is built by combining both intra-class similarity and inter-class ambiguity. Then dense subgraphs are detected within the affinity graph by performing graph shift. Each detected dense subgraph corresponds to a certain subcategory.

$[x_1, x_2, \ldots, x_M] \in \mathbb{R}^{d\times M}$. The class label of x_i is $c_i \in \{1, 2, \ldots, N_c\}$, where N_c is the number of classes. We also denote the number of samples belonging to the cth class by n_c, and the corresponding index set of samples by π_c.

4.1. Similarity Modeling

In this work, we define the appearance similarity as the Gaussian similarity between classification features ($\exp\{-||x_i - x_j||^2/\delta^2\}$), where δ^2 is the empirical variance of x. Though it is a common similarity metric for object classification, appearance similarity only is not enough for our SAOC framework, as in SAOC classification and detection are closely integrated. Subcategory mining only based on appearance similarity may lead to poor detectors, which in turn harms the overall performance. Hence detection and classification feature spaces ought to be taken into count simultaneously for similarity calculation.

The HOG based sliding window methods are the dominant approaches for object detection, which concatenate all the local gradients to form the window representation. These grid based HOG representations roughly capture object shapes and thus are sensitive to highly cluttered backgrounds and misalignments. Directly computing distance in concatenated HOG feature space often leads to poor results due to image misalignments [27]. To better measure the shape similarity between samples, we train a separate Exemplar-SVM detector[27, 20] for each positive sample. The misalignments can thus be partially handled by sliding the detector. The calibrated detection scores are defined as the pair-wise shape similarity.

The final instance similarity is defined by fusing the appearance similarity and pair-wise shape similarity. More specifically, we denote the appearance similarity as $S(A)_{i,j}$ and the pair-wise shape similarity as $S(P)_{i,j}$. Both $S(A)$ and $S(P)$ are normalized to $[0, 1]$. The final instance similarity is defined as $S_{i,j} = S(A)_{i,j} \times S(P)_{i,j}$.

4.2. Ambiguity Modeling

As discussed above, inter-class information is crucial for object classification. Dai et al. [8] have shown that placing local classifiers near the decision boundary instead of based on the data distribution only leads to better performance. This is intuitive as even there are many subcategories spreading separately in the feature space, if none of subcategories are close to samples of other categories, a single classifier may be enough to correctly classify all these subcategories. On the contrary, if some subcategories are near the decision boundary, separate classifiers should be trained for these ambiguous subcategories. Otherwise the ambiguous subcategories may decrease the classification performance of categories near the decision boundary.

As ambiguity is critical for object classification, subcategory mining should be guided by ambiguity instead of only relying on intra-class data distribution. Before introducing how to combine sample similarity and ambiguity into a unified framework, we need to first explicitly define the ambiguity measure. Here, we consider the L-nearest neighbours[1] of a particular sample x_i. If most of its neighbours share the same class label as x_i, the classification of x_i should be easy. Otherwise, x_i will be ambiguous and likely to be classified incorrectly. We thus define the ambiguity $A(x_i)$ of a training sample x_i as:

$$A(x_i) = \frac{\sum_{j\in N_i^L, j\notin \pi_{c_i}} S_{i,j}}{\sum_{j\in N_i^L} S_{i,j}}, \tag{1}$$

where N_i^L is the index set of the L-nearest neighbours of x_i. From the definition, a large $A(x_i)$ means that the neighbouring samples are likely to be of different classes, and hence the classification of x_i is more uncertain. On the contrary, a small $A(x_i)$ indicates that more neighbouring samples share the same class label of x_i. Note that computing the ambiguity relies on not only the intra-class information but also the inter-class formation. The ambiguity will be high for those training samples lying close to the decision boundary, and thus such samples should be more likely to form a separate subcategory.

4.3. Subcategory Mining by Graph Shift

Intuitively, the subcategory mining algorithm is expected to satisfy the following three properties. (1) It should be compatible with graph representation. Many similarity metrics are defined based on pair-wise relation, such as our pair-wise shape similarity, hence only graph based algorithms can directly utilize this pair-wise information. (2) It is able to utilize the informative inter-class ambiguities. Clustering methods based on only intra-class data distribution may fail to detect the ambiguous subcategories on the decision boundary and lead to subcategories imperfect for classification. Hence the expected algorithm should be able to adaptively cluster the data guided by ambiguity. (3) It should be robust to outliers. Some samples, such as highly occluded or strange images, may not belong to any subcategory. Methods insisting on partitioning all the input data

[1] In the experiments, we simply use $L = n_c/10$ for the cth class.

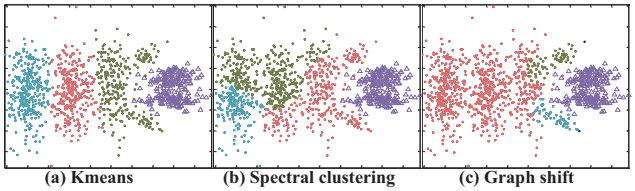

Figure 4: The subcategory mining results on synthetic data from kmeans, spectral clustering and graph shift. Here, triangles ($\triangle$) and dots ($\cdot$) represent samples from two different categories, respectively. Dots are split into subcategories, and different colors represent different subcategories. Kmeans and spectral clustering cluster the dots relying on only intra-class information, which leads to non-linearly separable subcategories from triangles. However, by utilizing the inter-class information, all three subcategories mined by the ambiguity guided graph shift are linearly separable from triangles, which is desired for classification. For better viewing, please see original colour pdf file.

into coherent groups without explicit outlier handling may fail to find the true subcategory structure.

The traditional partition methods, such as k-means and spectral clustering methods, are not expected to always work well for subcategory mining due to their insisting on partitioning all the input data and inability to integrate the inter-class information. Hence we need a more effective algorithm satisfying the above three properties. The graph shift algorithm [25], which is efficient and robust for graph mode seeking, appears to be particularly suitable for our subcategory mining problem as it directly works on graph, allows one to extract as many clusters as desired, and leaves the outlier points ungrouped. More importantly, the ambiguity can be seamlessly integrated into the graph shift framework. The graph shift algorithm shares the similar spirit with mean shift [6] algorithm and evolves through iterative expansion and shrink procedures. The main difference is that mean shift operates directly on the feature space, while graph shift operates on the affinity graph. The simulation results for comparing our ambiguity guided graph shift (AGS) with kmeans and spectral clustering are provided in Figure 4, from which we can see that our AGS can lead to subcategories more suitable for boosting classification.

Formally, we define an individual graph $G = (V, A)$ for each category. $V = \{v_1, \ldots, v_n\}$ is the vertex set, which represents the positive samples for the corresponding category. A is a symmetric matrix with non-negative elements. The diagonal elements of A represent the ambiguity of the samples while the non-diagonal element measures the similarity between samples. The modes of a graph G are defined as local maximizers of graph density function $g(y) = y^T Ay, y \in \Delta^n$, where $\Delta^n = \{y \in R^n : y \geq 0 \text{ and } ||y||_1 = 1\}$. More specifically, in this paper sample similarity and ambiguity are integrated and encoded as the edge weights of a graph, whose nodes represent the instances of the specific object category. Hence subcategories should correspond to those strongly connected subgraphs. All such strongly connected subgraphs correspond to large local maxima of $g(y)$ over simplex, which is an approximate measure of the average affinity score of these subgraphs.

Since the modes are local maximizers of $g(y)$, to find these modes, we need to solve following standard quadratic optimization problem (StQP) [2]:

$$\begin{aligned} &\text{maximize} \;\; g(y) = y^T Ay \\ &\text{subject to} \quad y \in \Delta^n. \end{aligned} \tag{2}$$

Replicator dynamics, which arises in evolutionary game theory, is the most popular method to find the local maxima of StQP (2). Given an initialization $y(0)$, corresponding local solution y^* of StQP (2) can be efficiently computed by the discrete-time version of first-order replicator equation, which has the following form:

$$y_i(t+1) = y_i(t)\frac{(Ay(t))_i}{y(t)^T Ay(t)}, i = 1, \ldots, n. \tag{3}$$

It can be observed that the simplex Δ^n is invariant under these dynamics, which means that every trajectory starting in Δ^n will remain in Δ^n . Moreover, it has been proven in [36] that, when A is symmetric and with non-negative entries, the objective function $g(y) = y^T Ay$ strictly increases along any non-constant trajectory of Eqn. (3), and its asymptotically stable points are in one-to-one correspondence with strict local solutions of StQP (2). One of the main drawbacks of replicator dynamics is that it can only drop vertices and be easily trapped in any local maximum. The graph shift algorithm provides a complementary neighbourhood expansion procedure to expand the supporting vertices. The replicator dynamics and the neighbourhood expansion procedure thus have complementary properties, the combination of which leads to better performance.

Like mean shift algorithm, the graph shift algorithm starts from an individual sample and evolves towards the mode of G. The samples reaching the same mode are grouped as a cluster. Each large cluster corresponds to one subcategory, while small clusters usually result from noises and/or outliers.

5. Experiments

5.1. Datasets and Metrics

We validate the proposed framework on the challenging PASCAL Visual Object Challenge (VOC) datasets [13], which provide a common evaluation platform for object classification and detection. VOC 2007 and 2010 datasets, which contain 9,963 and 21,738 images respectively, are used for experiments. The two datasets are divided into "train", "val" and "test" subsets. We conduct our experiments on the "trainval" and "test" splits. The employed evaluation metric is Average Precision (AP) and mean of

Figure 5: Visualization of our ambiguity guided subcategory mining results for bus and chair category on VOC 2007. Each row on the left shows one mined subcategory. Images on the right are detected as outliers.

Average Precision (mAP). We follow the standard PASCAL VOC comp1 test protocol for classification and PASCAL VOC comp3 test protocol for detection.

5.2. Ambiguity Guided Subcategory Mining Results

It has been shown that models trained by "clean" subsets of images usually perform better than trained with all images [39]. The importance of "clean" training data suggests that it is critical to cluster training data into "clean" subsets and remove outliers simultaneously. Figure 5 displays our subcategory mining results for bus and chair categories. Each row on the left side shows one discovered subcategory while right side images are detected as outliers and left ungrouped.

For the bus category, the first 3 subcategories correspond to 3 different views of buses. This is mainly due to the discriminative pair-wise shape similarity for different views of buses, as the Exemplar-SVM works well for the categories with common rigid shapes. We note the shape and appearance of the last subcategory show much larger diversity than other subcategories. Though these images are not very similar to each other, the strong ambiguity with the person category still guides them to form a separate subcategory.

For chairs, there are no common rigid shapes as buses and the shapes of various chairs are very diverse, which leads to much noisier pair-wise shape similarity. Hence the subcategory mining results should be the combination effects of both appearance similarity and shape similarity, which can be observed from the discovered subcategories. Some subcategories may not have common shapes, but have similar local patterns. For example, chairs of the 2nd subcategory all have the stripe-like patterns. We note again the last detected subcategory looks like sofas. Besides being different from other chair subcategories, the ambiguity with sofa is also one of the main reasons that these images form a separate subcategory.

5.3. Subcategory Mining Method Comparison

We extensively evaluate the effectiveness of different subcategory mining approaches on the VOC 2007 dataset, as the ground-truth of its testing set is released. To allow direct comparison with other popular works [17, 4, 5], we only implement a simplified SOAC framework. More specifically, we choose the state-of-the-art FVGHM [5] as the classification pipeline (dense SIFT feature [26] with FK coding [17] plus GHM pooling [23, 5]) and the customized DPM [15] as object detector. The only difference between customized DPM and the standard DPM is the model initialization. DPM-spectral, DPM-GS and DPM-AGS replace the aspect ratio based initialization with spectral clustering,

Table 1: Classification results (AP in %) comparison for different subcategory mining approaches on VOC 2007. For each category, the winner is shown in **bold** font.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
FV [17]	75.7	64.8	52.8	70.6	30.0	64.1	77.5	55.5	55.6	41.8	56.3	41.7	76.3	64.4	82.7	28.3	39.7	56.6	79.7	51.5	58.3
FVGHM [5]	76.7	74.7	53.8	72.1	40.4	71.7	83.6	66.5	52.5	57.5	62.8	51.1	81.4	71.5	86.5	36.4	55.3	60.6	80.6	57.8	64.7
FVGHM-CTX	78.5	80.0	54.9	71.9	55.4	75.1	87.1	67.2	58.4	60.3	60.0	47.3	83.0	76.3	90.5	44.9	59.6	63.2	83.5	68.9	68.3
FVGHM-CTX-spectral	81.2	82.1	56.7	73.5	56.2	76.5	88.5	67.8	58.0	60.1	61.7	48.1	85.1	77.8	90.7	45.5	60.6	64.4	84.3	69.2	69.4
FVGHM-CTX-GS	81.8	82.3	**58.5**	74.1	**56.5**	77.2	88.7	68.4	59.4	61.5	63.0	49.8	84.9	80.0	**91.3**	47.7	61.3	65.9	85.7	70.8	70.4
FVGHM-CTX-AGS	**82.2**	**83.0**	58.4	**76.1**	56.4	**77.5**	**88.8**	**69.1**	**62.2**	**61.8**	**64.2**	**51.3**	**85.4**	**80.2**	91.1	**48.1**	**61.7**	**67.7**	**86.3**	**70.9**	**71.1**

Table 2: Detection results (AP in %) comparison for different subcategory mining approaches on VOC 2007.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
E-SVM [27]	20.8	48.0	7.7	14.3	13.1	39.7	41.1	5.2	11.6	18.6	11.1	3.1	44.7	39.4	16.9	11.2	22.6	17.0	36.9	30.0	22.7
MC [18]	33.4	37.0	**15.0**	15.0	22.6	43.1	49.3	**32.8**	11.5	**35.8**	17.8	**16.3**	43.6	38.2	29.8	11.6	**33.3**	23.5	30.2	39.6	29.0
DPM [15]	28.9	59.5	10.0	15.2	25.5	49.6	57.9	19.3	22.4	25.2	23.3	11.1	56.8	48.7	41.9	12.2	17.8	33.6	45.1	41.6	32.3
DPM-spectral	32.9	60.3	9.6	15.9	29.2	52.6	58.1	21.6	21.1	24.6	26.1	10.8	58.2	48.1	37.6	11.9	21.5	35.3	48.6	43.1	33.4
DPM-GS	34.3	60.7	11.4	17.5	29.9	53.0	**58.9**	23.7	22.9	25.8	30.3	12.6	60.8	**49.2**	**42.6**	**13.3**	22.9	37.0	50.2	45.4	35.1
DPM-AGS	**34.7**	**61.4**	11.5	**18.6**	**30.0**	**53.8**	58.8	24.7	**24.7**	26.8	**31.4**	13.8	**61.4**	**49.2**	42.2	12.9	23.9	**38.5**	**50.8**	**45.5**	**35.7**

graph shift, ambiguity guided graph shift mining results, respectively. As detection assisted classification has become a standard approach for classification on PASCAL VOC. We augment FVGHM with detection context information and utilize the resulting FVGHM-CTX as the starting point to evaluate different subcategory mining methods. Dense SIFT is extracted using multiple scales setting (spatial bins are set as 4, 6, 8, 10) with step 4. The size of Gaussian Mixture Model in FK is set to 256. For GHM [5], we construct the hierarchical structure with three-level clusters, each of which includes 1, 2, 4 nodes respectively. The subcategory number is determined by the expansion size of the graph shift algorithm. Here the expansion size is decided by cross-validation, and the subcategory number is generally from 2 to 5. Spectral clustering, the representative graph based partition method, is chosen for comparison. We extensively evaluate spectral clustering with the cluster number from 2 to 5 and report the best results.

The detailed classification results are shown in Table 1. It can be concluded from the table that: 1) subcategory awareness does improve the performance of current detection assisted classification framework, and even with the naive spectral clustering, we can still boost the state-of-the-art classification performance; 2) our ambiguity guided graph shift approach is effective for subcategory mining and the resulting subcategories can obviously improve the classification performance; and 3) ambiguity is informative for subcategories mining and with the assistance of sample ambiguity, the graph shift algorithm can obtain better results for 17 out of 20 categories.

As object detection is an inseparable component of our SAOC framework, we also show the intermediate detection results in Table 2. Besides standard DPM, we add two more baselines, which also use the multiple components/models for object detection [18, 27]. When compared with other leading techniques in subcategory based detection, our method obtains the best results for most categories, achieving superior performance on categories with rigid shape or high ambiguity. We note the MC [18], which requires manually labelling the pose of each image, performs quite well on articulated categories. The inferior performance of our ambiguity guided mining framework on articulated categories is mainly due to the limited discriminative ability of current similarity metric.

5.4. Comparison with the State-of-the-arts

In this section we compare the performance of our SAOC framework with the reported state-of-the-art results on the VOC 2010 dataset. To obtain the state-of-the-art performance, we conduct the experiments with more complicated setting. For classification, we extract dense SIFT, HOG, color moment and LBP features in a multi-scale setting. All these features are encoded with VQ, LLC and FK [4] and then pooled by GHM. The pooling results are concatenated to form the final image representation. For object detection, we train one shape-based detector and one appearance-based object detector for each object category. The augmented DPM [38, 31] employing both the HOG and LBP features is adopted as the shape-based model. For appearance-based approach [34, 33], we sample 4000 sub-windows of different sizes and scales, and perform the BoW based object detector on these sub-windows. The number of subcategories is also determined by cross-validation as mentioned above.

The comparison results are presented in Table 3, from which it can be observed that our proposed method outperforms the competing methods on all 20 object categories. We note that all the leading classification methods combine object classification and object detection to achieve higher accuracy. However, most of the previous methods simply fuse the outputs of a monolithic classification model and a monolithic detection at category level. This limitation prevents them from grasping the informative subcategory structure and the interaction among the subcategories. By effectively employing the subcategory structure, we can further improve the state-of-the-art performance by 2.1%. Note that our methods can significantly improve the perfor-

Table 3: Classification results from our complete framework with comparison to other leading methods on VOC 2010.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
NLPR [12]	90.3	77.0	65.3	75.0	53.7	85.9	80.4	74.6	62.9	66.2	54.1	66.8	76.1	81.7	89.9	41.6	66.3	57.0	85.0	74.3	71.2
NEC [12]	93.3	72.9	69.9	77.2	47.9	85.6	79.7	79.4	61.7	56.6	61.1	71.1	76.7	79.3	86.8	38.1	63.9	55.8	87.5	72.9	70.9
ContextSVM [31]	93.1	78.9	73.2	77.1	54.3	85.3	80.7	78.9	64.5	68.4	64.1	70.3	81.3	83.9	91.5	48.9	72.6	58.2	87.8	76.6	74.5
GHM ObjHierarchy [5]	94.3	81.3	77.2	80.3	56.3	87.3	83.8	82.2	65.8	73.7	67.0	75.9	82.3	86.5	92.0	51.7	75.1	63.3	89.9	77.3	77.2
Our Method	**95.9**	**83.2**	**79.0**	**84.0**	**57.5**	**91.4**	**84.3**	**83.4**	**70.2**	**75.1**	**68.9**	**78.2**	**85.4**	**88.4**	**92.8**	**52.4**	**78.5**	**67.8**	**93.0**	**77.4**	**79.3**

mance of rigid categories (bus, train) and ambiguous categories (sofa, chair). When measured with object detection, we can achieve the performance of 37.1% compared to the state-of-the-art results of 36.8 % [12], which is obtained by much more complicated detection models than ours. As our framework focuses on classification, detailed detection results are omitted due to the space limitation.

6. Conclusions and Future Work

In this paper, we proposed an ambiguity guided subcategory mining and subcategory-aware object classification framework for object classification. We modeled the subcategory mining as a dense subgraph seeking problem. This general scheme allows us to gracefully embed intra-class similarity and inter-class ambiguity into a unified framework. The subcategories, which correspond to the dense subgraphs, can be effectively detected by the graph shift algorithm. Ambiguity guided subcategory mining results are then seamlessly integrated into the subcategory-aware detection assisted object classification framework. Extensive experimental results on both PASCAL VOC 2007 and VOC2010 clearly demonstrated the proposed framework achieved the state-of-the-art performance.

In the future, we plan to further explore whether our ambiguity guided subcategory mining can be extended for object segmentation and also develop a more efficient and scalable version of current framework to handle bigger data.

Acknowledgment

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

References

[1] O. Aghazadeh, H. Azizpour, J. Sullivan, and S. Carlsson. Mixture component identification and learning for visual recognition. In *ECCV*. 2012.

[2] I. M. Bomze. Branch-and-bound approaches to standard quadratic optimization problems. *J. of Global Optimization*, 2002.

[3] L. D. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. In *ECCV*, 2010.

[4] K. Chatfield, V. Lempitsky, and A. Vedaldi. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011.

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

[6] D. Comaniciu and P. Meer. Mean shift: A robust approach toward feature space analysis. *TPAMI*, 2002.

[7] J. Dai, J. Feng, and J. Zhou. Subordinate class recognition using relational object models. In *ICPR*, 2012.

[8] J. Dai, S. Yan, X. Tang, and J. T. Kwok. Locally adaptive classification piloted by uncertainty. In *ICML*, 2006.

[9] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[10] S. K. Divvala, A. A. Efros, and M. Hebert. How important are "deformable parts" in the deformable parts model? In *ECCV Workshops*, 2012.

[11] S. K. Divvala, A. A. Efros, and M. Hebert. Object instance sharing by enhanced bounding box correspondence. In *BMVC*, 2012.

[12] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results.

[13] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal visual object classes (voc) challenge. *IJCV*, 2010.

[14] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005.

[15] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4.

[16] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *TPAMI*, 2010.

[17] J. S. Florent Perronnin and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *ECCV*, 2010.

[18] C. Gu, P. A. Arbeláez, Y. Lin, K. Yu, and J. Malik. Multi-component models for object detection. In *ECCV*, 2012.

[19] C. Gu and X. Ren. Discriminative mixture-of-templates for viewpoint classification. In *ECCV*, 2010.

[20] H. Hajishirzi, M. Rastegari, A. Farhadi, and J. Hodgins. Understanding of proffesional soccer commentaries. In *UAI*, 2012.

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

[22] T.-K. Kim and J. Kittler. Locally linear discriminant analysis for multimodally distributed classes for face recognition with a single model image. *TPAMI*, 2005.

[23] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006.

[24] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010.

[25] H. Liu and S. Yan. Robust graph mode seeking by graph shift. In *ICML*, 2010.

[26] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[27] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011.

[28] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 1996.

[29] D. Park, D. Ramanan, and C. Fowlkes. Multiresolution models for object detection. In *ECCV*, 2010.

[30] O. Russakovsky, Y. Lin, K. Yu, and L. Fei-Fei. Object-centric spatial pooling for image classification. In *ECCV*, 2012.

[31] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2011.

[32] M. Toussaint and S. Vijayakumar. Learning discontinuities with products-of-sigmoids for switching between local models. In *ICML*, 2005.

[33] K. E. A. van de Sande, J. R. R. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as selective search for object recognition. In *ICCV*, 2011.

[34] A. Vedaldi, V. Gulshan, M. Varma, and A. Zisserman. Multiple kernels for object detection. In *ICCV*, 2009.

[35] J. Wang, J. Yang, K. Yu, F. Lv, and T. Huang. Locality-constrained linear coding for image classification. In *CVPR*, 2010.

[36] J. Weibull. *Evolutionary game theory*. MIT press, 1997.

[37] J. Xiao, J. Hays, K. A. Ehinger, A. Oliva, and A. Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *CVPR*, 2010.

[38] L. Zhu, Y. Chen, A. L. Yuille, and W. T. Freeman. Latent hierarchical structural learning for object detection. In *CVPR*, 2010.

[39] X. Zhu, C. Vondrick, D. Ramanan, and C. Fowlkes. Do we need more training data or better models for object detection? In *BMVC*, 2012.