Hierarchical Matching with Side Information for Image Classification

Qiang Chen[1], Zheng Song[1], Yang Hua[2], Zhongyang Huang[2],Shuicheng Yan[1]
[1] Department of Electrical and Computer Engineering, National University of Singapore, Singapore
[2] Panasonic Singapore Laboratories, Singapore
{chenqiang,zheng.s,eleyans}@nus.edu.sg, {yang.hua,zhongyang.huang}@sg.panasonic.com

Abstract

In this work, we introduce a hierarchical matching framework with so-called side information for image classification based on bag-of-words representation. Each image is expressed as a bag of orderless pairs, each of which includes a local feature vector encoded over a visual dictionary, and its corresponding side information from priors or contexts. The side information is used for hierarchical clustering of the encoded local features. Then a hierarchical matching kernel *is derived as the weighted sum of the similarities over the encoded features pooled within clusters at different levels. Finally the new kernel is integrated with popular machine learning algorithms for classification purpose. This framework is quite general and flexible, other practical and powerful algorithms can be easily designed by using this framework as a template and utilizing particular side information for hierarchical clustering of the encoded local features. To tackle the latent spatial mismatch issues in SPM, we design in this work two exemplar algorithms based on two types of side information: object confidence map and visual saliency map, from object detection priors and within-image contexts respectively. The extensive experiments over the Caltech-UCSD Birds 200, Oxford Flowers 17 and 102, PASCAL VOC 2007, and PASCAL VOC 2010 databases show the state-of-the-art performances from these two exemplar algorithms.*

1. Introduction

In this work, we focus on image classification according to the objects contained in the images. More specifically, we focus on the classification of complex images which contain objects as well as cluttered background areas. Ideally, different parts of image should serve different roles for the classification. The appearance model of object itself plays a key factor while rich context information from background is helpful for the classification process. However, since the objects may only occupy a small portion of the images, rich context information as well as background noise introduced by the rest area of the image must be well handled in practice. State-of-the-art methods following the bag-of-words (BoW) framework [10] mainly contain three steps: local feature extraction, feature encoding/pooling, and classifier learning. The local features are extracted from the dense grids, or via sparse interest point detection in the images. Feature encoding forms global image representations, e.g. a frequency histogram of visual words, which encodes the local features with a predefined visual dictionary such that the image representation has a comparable unified coordinate. The classifier learning step generally uses the kernel built on matching scores of the global image representations.

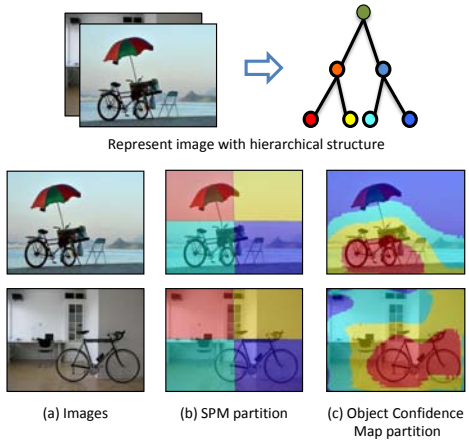

Figure 1. Illustration of the hierarchical matching representation. The local features are pooled according to partition of (b) traditional SPM and (c) the proposed object confidence prior. The figure shows our framework is superior than SPM in object matching across different images. For better viewing of all figures in this paper, please see original color pdf file.

Traditional BoW framework equally encodes all local features and does not emphasize any elements with regard to image layout. Hence, pyramid structure representation is often used to extend the global BoW representation in image classification, e.g. Spatial Pyramid Matching (SPM) [20] for natural scene classification. SPM models global geometric correspondence by partitioning the image plane into increasingly fine sub-regions. The success of SPM comes from the valid assumption that the images with

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

similar scene and geometry layout possibly belong to the same category. However, we argue that this representation is not optimum for object-centered recognition problem. As Figure 1 indicates, the spatial partition based on SPM may have mismatch problem caused by different object locations and scene layouts. In other words, if a prior knowledge, e.g. the possibility of object existence confidence in the image as shown in Figure 1 is acquired, we can construct the representation to match the corresponding object and background more accurately.

To this end, we propose a generalized hierarchical matching framework (GHM), which is capable to integrate different kinds of prior knowledge, including clues of object layout, for enhancing feature matching and towards object-oriented recognition. The prior knowledge, which is called *side information* in this paper, is associated with each local feature vector in image. Using the side information, the image local feature pool can be clustered into cells and further a coarse to fine hierarchical representation can be generated. Since the partition of the cells is guided with side information more semantically concerned, the encoding within each cell tends to be more semantically matchable and thus is expected to achieve better performance. Figure 1 demonstrates an example of how object-level side information is supplied to the proposed GHM framework. The side information of object confidence map can be used as an object-oriented prior for spatial partition of the image local feature pool. Consequently the images represented as hierarchical structures could carry out a coarse to fine matching.

Our contributions are two-fold. First, we propose the Generalized Hierarchical Matching framework for image classification. It gracefully extends the popular pyramid matching work, but further enables us to integrate other semantically useful side information with the flexibility. Second, two novel kinds of side information, i.e. object confidence map and visual saliency map, are introduced to enhance object-oriented image classification tasks based on the proposed GHM framework.

2. Related Work

2.1. Hierarchical Matching

Pyramid structure representation is often used to extend the global BoW representation in image classification, e.g. Spatial Pyramid Matching (SPM) [20] and Pyramid Match Kernel (PMK) [14]. SPM models approximate geometric layout by partitioning the image plane into increasingly fine sub-regions, and due to its better performance and simple implementation, it has become a standard procedure for image classification. However, for object-oriented classification, the increased complexity brought by SPM cannot contribute much to the recognition target because the object may appear in arbitrary position within an image, which thus may reduce the recognition efficiency and bring misalignment issue due to the unpredictable object locations in images.

PMK maps each feature set to a multi-resolution histogram that preserves the individual features' distinctness at the finest level. The histogram pyramids are then compared using a weighted histogram intersection computation, which implicitly defines the correspondence based on the finest resolution histogram cell where a matched pair first appears. It focuses on the mismatch problem caused by inaccurate Vector Quantization in feature encoding procedure. GHM framework well generalizes the SPM and PMK approaches and Section 3.3 will detail their relationship.

2.2. Saliency-guided Object Recognition

The saliency map [16] is a topographically arranged map that represents visual saliency of a corresponding visual scene. The purpose of the saliency map is to represent the conspicuity or "saliency" at every location in the visual field by a scalar quantity and to guide the selection of attended locations, based on the spatial distribution of saliency. Many of these saliency models are based on findings from psychology and neurobiology and explain the mechanisms guiding attention allocation [19, 16]. More recently, a number of models [28, 33] attempt to explain attention based on more mathematically motivated principles. Both types of models tend to rely solely on the statistics of the current test image when it comes to computing the saliency of a point in the image.

Some previous studies attempt to use saliency map as guidance for object recognition. Khan and Weijer [18] use color to guide attention by means of a top-down category-specific attention map. The color attention map is deployed to modulate more shape features from regions within an image that are likely to contain an object instance. Kanan and Cottrell [17] attempt to solve image classification using a biologically-inspired model to approximate the human eye fixations. These fixations are extracted from the feature maps at the sampled location, followed by probabilistic classification and the acquisition of additional fixations. The major difference between the proposed saliency map based GHM algorithm and these methods lies on how to utilize the saliency maps. In other words, GHM attempts to re-partition the features so that the group of features has more meaningful structure and each layer of partition has consistent elements to be matched.

2.3. Region-based Object Recognition

Recently, some work attempts to process the object recognition at the image region level. Andrews and Tsochantaridis [1], Wang and Forsyth [38] explore multiple instance learning respectively to classify images by the highest scored image region. Following this idea, Yakhnenko and Verbeek [41] use a latent-SVM model,

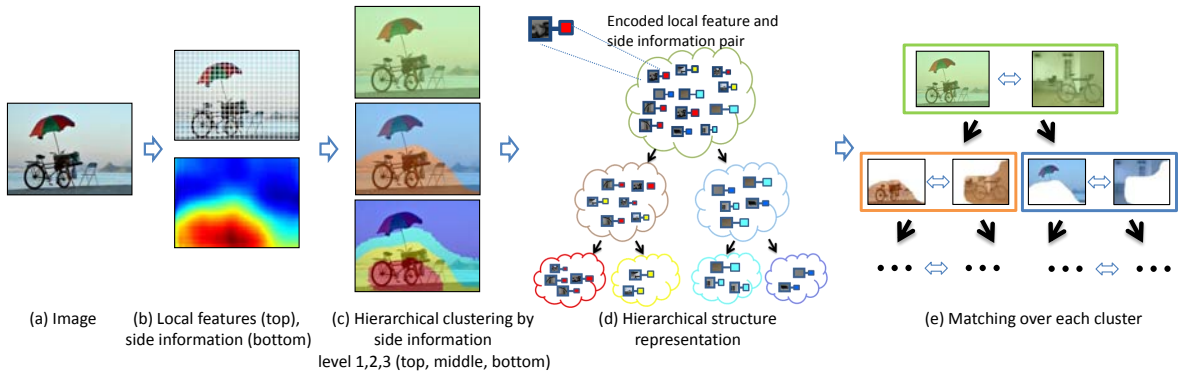

Figure 2. Diagrammatic flowchart of the proposed framework for image classification. The image is along with the (b) local features and side information. (c) The side information is hierarchically clustered to different levels. Different color mask represents different clusters at each level. (d) The encoding is operated on each cluster to form the hierarchical representation. (e) Finally, the matching over each corresponding cluster is performed.

which scores an image using all regions and associates each region with a latent variable indicating whether the region represents the object of interest or not. The solution takes the classification and foreground estimation into a joint inference framework. Though simpler than our proposed two-step solution, the critical drawback of the joint inference is that it will restrict the source of side information and cannot handle information from too complex sources. Other similar recognition work for image classification also exists. Chai et al. [4] propose to segment the images into foreground and background within co-segmentation scenario to improve image classification performance. Bosch et al. [2] define a Region-Of-Interest in the image and take the maximum response over the coarse image grid as the output of classifier. Comparing to these region-based approaches, the GHM framework aims to utilize all image information including object itself and context from different kinds of sources.

3. Generalized Hierarchical Matching

3.1. Image Classification Flowchart

Figure 2 shows the diagrammatic flowchart for image classification. Each image is expressed as a bag of orderless pairs I, each of which includes a local feature vector x_i encoded as c_i over a visual dictionary, and the side information f_i from priors and/or context, i.e. $I = \{\{x_i, c_i\}, f_i\}_{i=1}^{N}$. The side information is used for hierarchical clustering of the encoded local feature.

Along with the image itself, we may obtain the side information from various sources, e.g. the object confidence map denoting the existence probability of an object from object detector as shown in Figure 2. The side information is quantized into M discrete types. The encoding vectors c_i are assigned into different levels of clusters according to the quantization of side information, and form the hierarchical matching representation. To measure the similarity of two images $I_1 = \{\{x_i^1, c_i^1\}, f_i^1\}_{i=1}^{N_1}$ and $I_2 = \{\{x_i^2, c_i^2\}, f_i^2\}_{i=1}^{N_2}$, a kernel is constructed based on this representation. The kernel could be fed into any popular machine learning algorithm for classification purpose. We detail the GHM representation in the following section.

3.2. Hierarchical Matching Kernel

Assuming there are two images I_1, I_2, we can allocate each pair in I_1, I_2 into a hierarchical structure $G = \{G_1, G_2, ..., G_L\}$, where L is the number of hierarchical levels. Same as in previous hierarchical matching algorithms, only the elements grouped to the same cluster are supposed to match to each other. Hence we quantize all encoded feature vectors into M_l cells at level l, and the corresponding pooling is functioned on each cluster. We explored two ways to construct hierarchical structure. One is to perform hierarchical clustering on single/combined maps. The clustering is operated on the side information of training set. The other one is to design mixed meaningful structure from prior knowledge instead of automatic hierarchical clustering.

Then we can define a cluster kernel through a similarity function, *i.e.* $\kappa_{12}^{jl} = S(I_1, I_2, G_l^j)$, where S is a similarity function based on local feature cluster G_l^j on cell j at level l for images I_1 and I_2, and κ_{12}^{jl} represents the similarity value on cell j at level l. Then the similarity kernel between two images is defined as the weighted sum of similarity values:

$$K_{12} = \sum_{l=1}^{L} \sum_{j=1}^{M_l} w_{jl} \kappa_{12}^{jl}. \quad (1)$$

Similar to other hierarchical methods, it degenerates to a standard BoW when $L = 1, M_l = 1$. It is easy to verify that if the κ^{jl} is a Mercer Kernel, then K is also a Mercer Kernel and thus it can be embedded into any popular kernel-based machine learning algorithm. The kernel weight w_{jl} can be intuitively set or learnt by popular Multiple Kernel Learning (MKL) [29] method.

Table 1. Unified framework of Generalized Hierarchical Matching

Method Name		Side information	Coding method	Similarity function
PMK [14]		Histogram Index	Vector Quantization	Intersection
SPM	General [20]	Location Coordinate	Vector Quantization	Arbitrary
	ScSPM [42]		Sparse Coding	Linear
	ImprovedFV [12]		Fisher Vector Coding	Linear
The proposed GHM		Object Confidence Map, Visual Saliency Map	Arbitrary	

3.3. Generalization and Flexibility

In Table 1, we demonstrate the generalization capability with various configurations of GHM to realize previous hierarchical matching algorithms as well as our proposed object-oriented recognition with new side information.

First, we show that the Pyramid Match Kernel (PMK) [14] is one exemplar of the GHM framework. To encode and match the local feature with more accurate quantization, PMK uses multiple levels of local feature pooling and intersection kernel matching based on Vector Quantization (VQ). The pool of local image features is hierarchically partitioned into clusters according to their histogram indices and the final matching score is defined as weighted sum of all cluster matching scores, which can be straightforwardly explained by our GHM framework. As aforementioned, SPM uses the location coordinate of local features as side information for clustering and it is easily adapted as one special case of the GHM framework. GHM is the general form of PMK and SPM, which use diverse side information respectively.

Table 1 also illustrates that GHM framework can embed any popular coding method with flexibility. The BoW feature encoding approaches such as Sparse Coding [42] and Locality-constrained Linear Coding (LLC) [39] introduce soft assignment for local feature quantization. Fisher encoding[12] and Super vector encoding [44] capture the average first and second order differences between local features and their distribution centres modeled by Gaussian Mixture Models. Most of the coding work include SPM as the spatial pooling step. GHM could also help this step and indicate image coding on well-designed clusters based on provided side information, e.g. object confidence map and visual saliency map which is detailed in next section.

4. Side Information Design

In this section, we design two schemes to construct side information: (1) the object confidence map which reveals the possibility of a local patch containing a object. (2) the visual saliency map which takes advantage of natural image statistic and distinguishes the foreground against the background. Further these two kinds of information, as well as the location coordinate information, can be combined parallelly or hierarchically as side information to reflect meaningful structure for GHM-based image recognition.

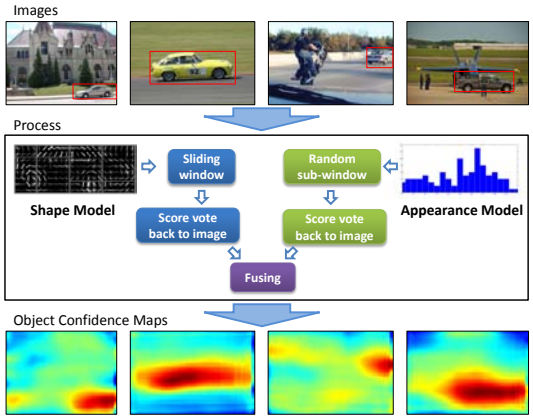

Figure 3. Object confidence map and some examples from car category.

4.1. Object Confidence Map

For object recognition task, it is commonly believed that in traditional well-proposed object recognition datasets, such as CMU PIE face [31] and Caltech-101 [9], most objects are cropped after fine alignment and with little background noises, and such preprocess always leads to much better performance. But it does not work for general object recognition datasets such as Caltech-UCSD Birds [40], PASCAL VOC [7], etc, where no object pre-alignment and cropping is performed. Intuitively the most useful recognition prior for these object-unaligned images is object position. And object position should be extremely beneficial for fine-grained image classification task.

The steps to construct an object confidence map, denoted as GHM Object, is illustrated in Figure 3. For each object category, e.g. car, we train one shape-based and one appearance-based object detectors, 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 in certain way. Instead of constructing the local classifiers on a super-pixel representation as in other work [34, 21], we use square grid samples and sliding-window approach for efficiency consideration.

The shape-based object detection adopts the state-of-the-art part-based model from Felzenszwalb et al. [11] using HOG [6] features. And the appearance-based object detector is trained with BoW features. We use dense SIFT [22] and LBP [26] as local features and the codebook sizes for dense SIFT and LBP are 2000 and 1000 respectively. Each

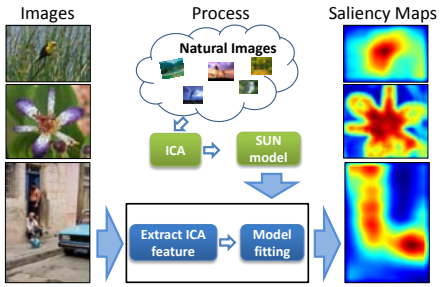

Figure 4. Visual saliency map generation and some examples.

detection sub-window is divided into 3x1 spatial pyramid to provide weak geometry constraint. The BoW histogram is mapped into high-dimension space via Additive Kernel Mapping [37]. This nonlinear transformation guarantees the possibility of using linear classifier for fast detection. We further accelerate the detection by using integral image to construct BoW representation within sub-window. Multiple scale detection is performed in each image and the obtained multi-scale scores are averaged to get final single object confidence map.

4.2. Visual Saliency Map

For some object categories, such as flowers, detection models may perform poorly. We propose another apparent foreground prior on finding visually salient image regions from human attention models and construct saliency maps as side information, denoted as GHM Saliency.

We consider the saliency under the scenario of general visual classification problem. In other words, the saliency information should reflect how human sees the objects against the natural background clutter. For this reason, we use the saliency model SUN (Saliency Using Natural statistics) [43]. This measure of saliency is based on natural image statistics, rather than based on a single test image, providing a straightforward explanation for many search asymmetries observed by humans.

The SUN model illustrated in Figure 4 defines the bottom up saliency as $P(F)^{-1}$, where F indicates the transformed color features through Independent Component Analysis (ICA) [35] on local color patch. Since the components of F have been made largely statistically independent by ICA, SUN models $P(F)$ as the product of unidimensional distributions: $P(F = f) = \prod_i P(f_i)$, where f_i is the ith value of these filter responses at this location. The ICA feature responses to natural images can be fitted very well using Generalized Gaussian Distributions [30], and we obtain the shape and scale parameters for each ICA filter by fitting its response to the ICA training images.

4.3. Side Information Combination

The nature of the GHM framework enables us to flexibly combine side information from various sources. One straight way to combine the side information is parallel information fusion, e.g. the spatial location information and the saliency map coupling as $f = \{f_{location}, f_{saliency}\}$ collaboratively. The clustering over this combination aims to consider the geometric constraint and saliency information so that each of the sub-cluster in the image contains equal amount of salient area. We denote this parallel combination as GHM LocSaliency.

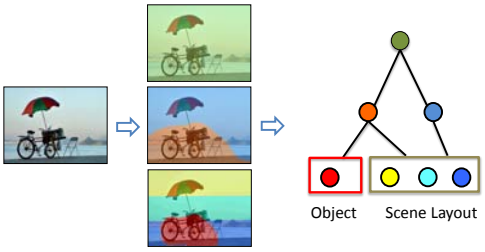

Figure 5. Combine object confidence map and spatial layout into one GHM. Level 2 is clustered according to object confidence map. Level 3 is designed for foreground matching and scene layout matching.

Another feasible solution for side information combination is to design mixed hierarchical structure. Most natural images (e.g. those from PASCAL VOC dataset) contain large amount of background area, which in fact supplies rich contexts for the recognition of certain object categories e.g. sky for aeroplane/bird, urban scene for various vehicles. This motivates us to design a configuration which simultaneously matches foreground objects and background scenes. The background confidence can be simply obtained from the foreground object confidence with reversed process, i.e. small object confidence map value meaning higher possibility of background. The spatial layout is proved to be useful for the recognition of background scenes [20]. We design a 3 level hierarchical structure with combined side information: the whole image as level 1, object confidence map is used in level 2 as the foreground confidence map, and the small value denoting the background area will be further utilized in level 3 to construct the 3×1 spatial layout modeling the background scene as shown in Figure 5. We denote this hierarchical combination as GHM ObjHierarchy.

In summary, we propose two useful resources of side information to fit into proposed GHM framework for image classification, i.e. the object confidence map and the visual saliency map. We further propose to associate the side information from multiple resources, either through simple parallel combination or via sophisticated hierarchical design to reflect the semantic complexity in real image recognition task.

5. Experiments

5.1. Datasets and Metric

We evaluate our proposed Generalized Hierarchical Matching framework on several popular datasets, the re-

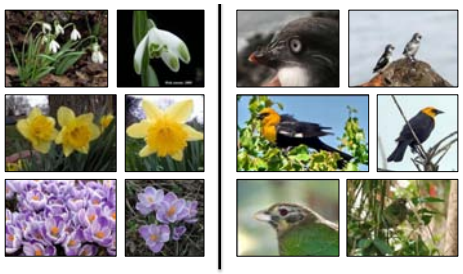

Figure 6. Sample images from Oxford Flowers 17 and CUB 200. The images in the same row belong to the same category.

cently released Caltech-UCSD Birds 200 (CUB-200) [40], the Oxford Flowers 17 (Flowers 17) [24] and 102 (Flowers 102) [25], and the PASCAL Visual Object Challenge (VOC) datasets [7].

The CUB-200 contains 200 bird categories and 6033 images in total. It is created to enable the study of subordinate categorization. The Flowers 17 [24] dataset contains 17 different flower species with 80 images per category. The dataset provides three different data splits with each including 60 training and 20 test images. The Flowers 102 [25] dataset includes 8289 images divided into 102 categories with 40 to 250 images per category. We use the provided data split with 20 images per-category for training and the rest for testing. Figure 6 shows some examples of the Oxford Flowers and CUB-200 images. It can be seen that these two fine-category classification datasets are very challenging due to the large intra variances.

The PASCAL Visual Object Challenge (VOC) datasets [7] are widely used for many image understanding tasks and provide a common evaluation platform for both object classification and detection. We use PASCAL VOC 2007 and 2010 datasets for experiments. VOC 2007 and VOC 2010 datasets contain 9,963 and 21,738 images respectively. 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 Average Precision* (mAP) complying with the PASCAL challenge rules.

5.2. Experimental Details

Baseline Configuration: For CUB-200, Flowers17 and Flowers102 datasets, the local features used for the image recognition are RGB color moment and dense SIFT descriptors. The implementation of dense SIFT is based on VLFeat [36] using multiple scales setting (spatial bins are set as 4, 6, 8, 10) with step 4. We use the improved Fisher vector coding [12] with SPM setting which has demonstrated the superiority over other coding methods in a fair setting [5]. The size of Gaussian Mixture Model in Fisher vector coding is set to 256 for these two features separately. One-vs-All SVM is learnt for each category using the representation generated by GHM and returns the class with the maximum score over all the image classifiers. The SPM is with typical setting, 3 levels are used, $1 \times 1, 2 \times 2, 4 \times 4$ spatial separation.

For PASCAL VOC 2007 [7] datasets, we use only dense SIFT feature with the Fisher vector coding to make it comparable with other popular works. We also conduct the experiments with "heavy" setting to obtain state-of-the-art performance for PASCAL VOC 2010 dataset. For local features, we extract dense SIFT, HOG, color moment and LBP features in a multi-scale setting. Typically, the number of local features for each image is around 30K for SIFT, 5K-10K for others. This is critical in feature coding to produce non-sparse representation. All these features are also encoded with improved Fisher vector coding. One-vs-All SVM is learnt and the performance is evaluated by AP.

Side Information Generation: We implement the proposed two kinds of side information: (1) the supervised object confidence map and (2) the unsupervised visual saliency map. The two detectors used to generate object confidence map are trained with PASCAL VOC images. For part-based model [11], the HOG and LBP features are used for object description and the number of part models for each object category is set to 8. For appearance-based approach, we sample 4000 sub-windows with different size and scale and perform the BoW based object detector on these sub-windows. We construct the hierarchical structure with three-level clusters, each of which includes 1, 2, 4 nodes respectively on the training images. For each class, we sample the responses from the positive images and the same number of negative images and get various cluster centers with clustering process. Finally each local feature is assigned to the nearest center at each level.

For the saliency map generation, we follow the SUN [43] framework and adopt the ICA filters model from [17]. These filters are learned with the images from the McGill color image dataset [27]. For the following experiments, we use this setting unless otherwise stated: three-level clusters for hierarchical structure, each of level with 1, 2, 4 nodes respectively. The clustering is operated on single image but not cross dataset since we find that the saliency map values for different images are not comparable.

The weight w_{jl} is intuitively set without fine tuning: the higher confidence cluster has higher weight within each level and the weights are normalized to have unit sum for each level.

5.3. Exp1: Caltech-UCSD Birds 200

We first evaluate our methods on the newly released Caltech-UCSD Bird 200 dataset and show that the visual saliency map and the object confidence map are very helpful for the fine categorization problem. The dataset is extremely challenging, and its authors report only 19% recognition accuracy [3] when using ground truth masks. The

Table 2. Performance comparison on Caltech-UCSD Birds 200. The proposed methods lead to the highest recognition accuracy.

Methods	Recognition Acc.
Chai et al. [4]	17.0
Branson et al. [3]	19.0
BoW Baseline	15.2
FVSPM	15.0
GHM Saliency	**18.1**
GHM Object	**19.2**

recognition performance is listed on Table 2 (using the suggested 20 training images per class split). Chai et al. [4] first segment the image into foreground and background and then extracted feature on the foreground. We also implement the Fisher vector coding with SPM (FVSPM) [12].

For this fine-grained categorization problem, the spatial layout has no exact meaning for different fine classes since most of classes share the same background. We propose to use saliency map (GHM Saliency) and the object confidence map (GHM Object) as a guidance to partition the images into different levels. The object confidence map is obtained by performing the "bird" detector trained from VOC 2010 datasets. Both of the results are much better than FVSPM. The results show that the unsupervised saliency performs very well on this dataset and the object confidence map gives strong support for separating the foreground and background so that fine-grained categorization is possible.

5.4. Exp2: Oxford Flowers 17 and 102

We compare our proposed GHM method with other state-of-the-art results on Oxford Flowers datasets. Gehler and Nowozin [13] adopt multiple feature combination method. Kanan and Cottrell [17] use the same saliency map as ours. Chai et al. [4] use segmentation to get the foreground area which is current leading method in this dataset. It is almost impossible to train a "flower" detector for this dataset, on the other hand, the saliency map shows strong evidence over this datasets: most of the flowers are within the salient foreground area of the images. So we evaluate the GHM with saliency map performance and its combination with spatial information. The recognition performances on Oxford Flowers 17 and 102 are listed on Table 3.

The GHM with the saliency map (GHM Saliency) achieves comparable performance with FVSPM. It shows that the saliency map is comparable prior for object recognition with the weak geometric alignment at these two datasets. It is worth noting that for these two datasets, we use compact representation. i.e. 3 levels of saliency map with total 1+2+4=7 cells compared with 21 cells in SPM. We also use the parallel combination design of side information by using saliency map together with spatial information (GHM LocSaliency). The side information is designed as $f = \{f_{location}, f_{saliency}\}$. Then a 2 level GHM with $1 \times 1, 2 \times 2$ setting is constructed. The results show the additional improvement over the single channel of side information with very compact representation.

Table 3. Performance comparison on Oxford Flowers datasets.

	Flowers 17	Flowers 102
Methods	Recognition Acc.	
Gehler and Nowozin [13]	88.5 ± 3.0	–
Kanan and Cottrell [17]	–	72.8
Chai et al. [4]	90.4 ± 2.3	80.0
FVSPM	93.0 ± 1.7	82.0
GHM Saliency	**93.1 ± 1.8**	**82.3**
GHM LocSaliency	**93.5 ± 1.5**	**82.6**

5.5. Exp3: VOC 2007 and VOC 2010

We evaluate our proposed method on PASCAL VOC 2007 and VOC 2010 dataset. The classification results on VOC 2007 are listed on Table 4. INRIA [23] is the winner of VOC 2007 and uses multiple kernel learning to balance the weight of different features. LLC [39] is the popular state-of-the-art feature coding method. We follow coding method in FisherVec [12] which results in mAP 58.3%. Our baseline FVSPM (mAP 60.6%) achieves higher performance than FisherVec approach, since more dense SIFT features with smaller step for one image is extracted. All these methods report much lower mAP than the leading score in [32] which uses "heavy" setting. Also note that the object classes in this dataset are conflicted with the saliency map assumption since many of the concerned classes and object instances in VOC are not at the foreground area, e.g. bottle, chair, tv. So we mainly use the object confidence map for each class and encode the features with GHM. The results (GHM Object) show mAP +3% absolute improvement over the baseline method using SPM. The prior of object confidence map is much stronger than the spatial layout for object-oriented classification.

VOC images contain large amount of background area which provides rich context information for recognition of certain objects. This also leads us to design a configuration which simultaneously matches foreground objects and background contexts. We design the mixed hierarchical structure setting with combined side information as proposed in Sec. 4.3. The significant performance improvement from mAP 60.6% (by FVSPM) to 64.7% (by GHM ObjHierarchy) demonstrates the effectiveness of this hierarchical structure of mixed spatial layout and object confidence modeling.

We also compare our method with the current leading approach [32] on PASCAL VOC 2010 with "heavy" setting. We adopt the Context SVM method with its configuration which combines the object detection and classification in a context-aware scenario, but generate the representation with GHM ObjHierarchy. The classification results on VOC 2010 are listed in Table 5. The final results of GHM ObjHierarchy outperform the leading scores in VOC 2010 challenge. The usage of hierarchical object and scene layout side information provides great gain for this classifi-

Table 4. Classification results (AP in %) on VOC 2007. The proposed GHM Object and GHM ObjHierarchy outperform the baseline methods.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
INRIA [23]	**77.5**	63.6	**56.1**	71.9	33.1	60.6	78.0	58.8	53.5	42.6	54.9	45.8	77.5	64.0	85.9	36.3	44.7	50.6	79.2	53.2	59.4
LLC [39]	74.8	65.2	50.7	70.9	28.7	68.8	78.5	61.7	54.3	48.6	51.8	44.1	76.6	66.9	83.5	30.8	44.6	53.4	78.2	53.5	59.3
FisherVec [12]	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
FVSPM	75.8	68.1	51.6	71.6	30.0	69.4	78.9	61.9	50.7	50.6	55.5	45.8	79.2	69.1	84.6	31.9	49.9	53.1	79.7	54.4	60.6
GHM Object	77.0	73.5	51.8	71.1	37.1	70.8	82.3	63.4	52.0	55.2	60.9	49.9	80.7	71.2	86.0	36.3	53.8	59.8	79.6	**57.8**	63.5
GHM ObjHierarchy	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**

Table 5. Classification results (AP in %) on VOC 2010. The proposed GHM ObjHierarchy outperforms the state-of-the-art performance.

	plane	bike	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	motor	person	plant	sheep	sofa	train	tv	mAP
NLPR [8]	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 [8]	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 [32]	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	**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**

cation task.

6. Conclusions and Future Work

In this work, we introduced a generalized hierarchical matching (GHM) framework for image classification task. This general and flexible scheme allows us to embed any useful side information into the image recognition framework. We also presented two novel exemplar approaches for side information generation towards object-oriented recognition, i.e. object confidence map and visual saliency map. Extensive experimental results clearly demonstrated the proposed GHM together with designed varieties of side information could achieve state-of-art performance on diverse and popular image recognition datasets. In future, we shall further explore more semantically meaningful side information and new approach for combining different types of side information.

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] S. Andrews and I. Tsochantaridis. Support vector machines for multiple-instance learning. In *NIPS*, 2003. 2
[2] A. Bosch, A. Zisserman, and X. Munoz. Image classification using random forests and ferns. In *CVPR*, 2007. 3
[3] S. Branson, C. Wah, F. Schroff, B. Babenko, P. Welinder, P. Perona, and S. Belongie. Visual Recognition with Humans in the Loop. In *ECCV*, 2010. 6, 7
[4] Y. Chai, V. Lempitsky, and A. Zisserman. BiCoS: A Bi-level Co-Segmentation Method for Image Classification. In *ICCV*, 2011. 3, 7
[5] K. Chatfield, V. Lempitsky, A. Vedaldi and A. Zisserman. The devil is in the details: an evaluation of recent feature encoding methods. In *BMVC*, 2011. 6
[6] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005. 4
[7] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The Pascal Visual Object Classes (VOC) Challenge. *IJCV*, 2010. 4, 6
[8] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman. The PASCAL Visual Object Classes Challenge 2010 (VOC2010) Results. http://www.pascal-network.org/challenges/VOC/voc2010/workshop/index.html. 8
[9] L. Fei-Fei and R. Fergus. One-Shot learning of object categories. *TPAMI*, 2006. 4
[10] L. Fei-Fei and P. Perona. A bayesian hierarchical model for learning natural scene categories. In *CVPR*, 2005. 1
[11] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object Detection with Discriminatively Trained Part-Based Models. *TPAMI*, 2010. 4, 6
[12] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale Image Classification. In *ECCV*, 2010. 4, 6, 7, 8
[13] P. Gehler and S. Nowozin. On feature combination for multiclass object classification. In *ICCV*, 2009. 7
[14] K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *ICCV*, 2005. 2, 4
[15] S. Ito and S. Kubota. Object Classification Using Heterogeneous Co-occurrence Features. In *ECCV*, 2010.
[16] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 1998. 2
[17] C. Kanan and G. Cottrell. Robust classification of objects, faces, and flowers using natural image statistics. In *CVPR*, 2010. 2, 6, 7
[18] F. S. Khan and J. van de Weijer. Top-down color attention for object recognition. In *CVPR*, 2009. 2
[19] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Hum Neurobiol*, 1985. 2
[20] S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *CVPR*, 2006. 1, 2, 4, 5
[21] F. Li, J. Carreira, and C. Sminchisescu. Object recognition as ranking holistic figure-ground hypotheses. In *CVPR*, 2010. 4
[22] D. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 4
[23] M. Marszalek, C. Schmid, H. Harzallah, and J. van de Weijer. Learning object representations for visual object class recognition. In *Visual Recognition Challange workshop, ICCV*, 2007. 7, 8
[24] M.-E. Nilsback and A. Zisserman. A Visual Vocabulary for Flower Classification. In *CVPR*, 2006. 6
[25] M.-E. Nilsback and A. Zisserman. Automated Flower Classification over a Large Number of Classes. In *ICVGIP*, 2008. 6
[26] T. Ojala, M. Pietikäinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *PR*, 1996. 4
[27] A. Olmos and F. Kingdom. McGill Calibrated Colour Image Database. 6
[28] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson. Top-down control of visual attention in object detection. In *ICIP*, 2003. 2
[29] A. Rakotomamonjy, F. Bach, S. Canu, Y. Grandvalet. simpleMKL. In *JMLR*, 2008. 3
[30] K.-S. Song. A globally convergent and consistent method for estimating the shape parameter of a generalized Gaussian distribution. *TIT*, 2006. 5
[31] T. Sim, S. Baker, and M. Bsat The CMU Pose, Illumination, and Expression Database. *TPAMI*, 2003. 4
[32] Z. Song, Q. Chen, Z. Huang, Y. Hua, and S. Yan. Contextualizing object detection and classification. In *CVPR*, 2011. 7, 8
[33] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 2006. 2
[34] K. van de Sande, J. Uijlings, T. Gevers, and A. W. M. Smeulders. Segmentation as Selective Search for Object Recognition. In *ICCV*, 2011. 4
[35] J. H. van Hateren and A. van der Schaaf. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society B: Biological Sciences*, 1998. 5
[36] A. Vedaldi and B. Fulkerson. VLFeat: An open and portable library of computer vision algorithms. http://www.vlfeat.org/, 2008. 6
[37] A. Vedaldi and A. Zisserman. Efficient Additive Kernels via Explicit Feature Maps. *TPAMI*, 2011. 5
[38] G. Wang and D. Forsyth. Joint learning of visual attributes, object classes and visual saliency. In *ICCV*, 2009. 2
[39] J. Wang, J. Yang, K. Yu, F. Lv, and T. Huang. Locality-constrained linear coding for image classification. In *CVPR*, 2010. 4, 7, 8
[40] P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona. Caltech-UCSD birds 200. Technical report, California Institute of Technology, CNS-TR-2010-001, 2010. 4, 6
[41] O. Yakhnenko and J. Verbeek. Region-Based Image Classification with a Latent SVM Model. Technical report, INRIA, 2011. 2
[42] J. Yang, K. Yu, and Y. Gong. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009. 4
[43] L. Zhang, M. Tong, T. Marks, H. Shan, and G. W. Cottrell. SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 2008. 5, 6
[44] X. Zhou, K. Yu, and T. Zhang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010. 4