A Biased Sampling Strategy for Object Categorization

Lei Yang
Xi An Jiaotong University
lyang@aiar.xjtu.edu.cn

Nanning Zheng
Xi An Jiaotong University
nnzheng@mail.xjtu.edu.cn

Jie Yang
Carnegie Mellon University
jie.yang@cs.cmu.edu

Mei Chen
Intel Labs Pittsburgh
Mei.chen@intel.com

Hong Chen
Carnegie Mellon University
hongc@cs.cmu.edu

Abstract

In this paper, we present a biased sampling strategy for object class modeling, which can effectively circumvent the scene matching problem commonly encountered in statistical image-based object categorization. The method optimally combines the bottom-up, biologically inspired saliency information with loose, top-down class prior information to form a probabilistic distribution for feature sampling. When sampling over different positions and scales of patches, the weak spatial coherency is preserved by a segment-based analysis. We evaluate the proposed sampling strategy within the bag-of-features (BoF) object categorization framework on three public data sets. Our technique outperforms other state-of-the-art sampling technologies, and leads to a better performance in object categorization on VOC2008 dataset.

1. Introduction

In object categorization and classification, the problem of appearance variations has been successfully addressed by coding local image patches using statistical representations [5, 24, 28]. One example is the bag-of-features (BoF) method [5], which represents an image as an unordered collection of local features. Despite its simplicity, the BoF representation appears to be discriminative and robust against object appearance variations and occlusions. However, upon close examination, one would realize that the classic BoF representation is applied to the entire scene, rather than the object of interest specifically. When the object is not dominant in the image, or there are other objects and background clutter, or when the background differs considerably within the same category, the classic BoF representation without object prior would fail. Figure 1 shows examples where the same category of objects can have largely different BoF models, and different categories of objects can have a similar BoF model. The main reason is that classic BoF methods apply the uniform sampling strategy to the entire image such that the characteristics of the object to be modeled can be obscured by the background.

In this paper, we propose to improve statistical object class modeling using a biased sampling strategy, which samples patches using a prior distribution over patches of different locations and scales. An image is first over-segmented into patches that preserve weak spatial coherency. The number of random samples on each patch is biased by combining information from both top-down and bottom-up analysis. The bottom-up biologically inspired saliency measures evaluate the extent to which regions stand out from their surroundings, while the top-down process learns simple and loose object class prior information. The top-down and bottom-up information are optimally combined to form the final probabilistic sampling distribution, leading to more samples on the object of interest. The proposed method strikes a balance between random sampling and object classifier based sampling, while at the same time adding weak spatial coherency to global statistical methods.

We demonstrate the efficacy of the proposed method through substantial evaluations. We compare the proposed sampling method with three baseline algorithms within the BoF framework: Harris-Laplace interest point detector, ran-

Figure 1. An illustration of the whole image-based representation sometimes leading to mistakes for object categorization.
We assume each cue is independent to obtain less coupled visual cues such as color, texture, and geometric attributes. Image are described by simple properties from different views using the mean-shift algorithm [4].

Images are preprocessed by segmenting into smaller regions the top-down analysis and bottom-up saliency map. All images are computed for each sampled patch and a histogram-based probabilistic distribution over the patches of different locations and scales. The distribution results from a combination of patch sampling [23, 28], descriptor coding [8, 13, 14], and recognition strategies [5, 7]. This paper intends to address scene modeling problem by improving the component of patch sampling in BoF framework.

The classical patch sampler is a critical component of the statistical-based methods, which can be classified into three groups: sparsely (based on key points or salience measures) [12, 17], densely [1, 11], or randomly [19, 26]. Sparse sampling methods based on interest operators are very good at detecting specific local structures such as edges or corners in a repeatable way. But they do not generalize well and thus often neglect some structures on the objects. Unlike other detection or localization methods which heavily rely on feature selection and complex class modeling, the proposed framework characterizes objects using simple region-based features, loose object class models, and the statistical-based categorization method. With the relaxed class modeling, the worst case is that no discriminative features are discovered and the whole image is sampled under a uniform distribution, leading to random sampling. The remaining process is similar to existing BoF methods. SIFT (Scale Invariant Feature Transform) descriptors are computed for each sampled patch and a histogram-based representation for each image is built for the classification.

2.2. Related work

The proposed method is related to some researches in the literature. Several researchers utilized detection and localization based methods for object categorization to avoid modeling the entire scene [3, 22]. These methods, however, are usually computational expensive and sensitive to variations and task changes. Some other researchers tried to improve different stages within the BoF framework to balance robustness and accuracy, such as patch sampling [23, 28], patch description [2, 24], descriptor coding [8, 13, 14], image description [15, 24, 28], and recognition strategies [5, 7]. This paper intends to address scene modeling problem by improving the component of patch sampling in BoF framework.

Figure 2. An overview of the proposed framework.
which are considered to be not salient by their measures. Dense sampling [11] processes every pixel at every scale and captures most information, but it is memory and computation intensive. Nowak et al. demonstrated in [19] that random sampling has shown to provide comparable or even better performance than interest point detection methods and cost less than dense sampling. However, all the above sampling methods reflect generic low level image properties bearing little direct relationship to discriminative power for recognition and are prone to capture the information of scene better than of objects.

Recently, Walther et al. [27] combined the biologically plausible saliency regions with interest point operators to show improved performances. However, it remains unclear whether the extracted regions are indeed salient or yield a good estimate of the object’s size and shape. Several methods for patch sampling are formulated in a top-down class discriminative manner [18, 23]. These methods are actually carried out after feature description stage. Moosmann et al. proposed to use extremely randomized clustering forests to classify patches to be on the objects or the background, which leads to a separate hard problem of building an accurate object part classifier [18]. Parikh et al. proposed a discriminative saliency measure based on class specific co-occurrence and spatial context information between image patches [23]. They used their saliency map to guide the sampling process and produced results at a state-of-the-art level in the classification task.

Moreover, some researchers worked on improving other components or stages within the BoF framework to achieve the same goal, such as an interest point operator with an discriminative codebook or feature selector. We consider these methods of a different line of attack from our approach, yet we make a comparison against their results in our evaluation for completeness.

3. Top-down object class prior map

In order to avoid scene modeling problem, a straightforward way is to localize or segment objects of interest in the scene first, of which are challenging computer vision problems. An alternative way is to localize or segment objects of interest in the scene first, of which are challenging computer vision problems. An alternative is to use bottom-up saliency measure to evaluate the extent to which regions stand out from their surroundings. However, in practice, not all object parts or objects could be unmasked by the saliency measurements. Instead of combining two challenging problems together to solve the scene matching problem or simply use the bottom-up saliency measure, we employ a learning-based method to generate a probabilistic object class map \( O_{det} \), which indicates the probabilities for objects of interest to appear on each patch within candidate regions. Our objective is to obtain a probability map for the regions where an object of interest is likely to appear. This top-down loose class prior map is learnt from training images with object labels. We start from a uniform distribution map and use region-based features to measure probabilities of objects of interest on patches. Therefore, the lower bound of the class prior map is a uniform distribution. In that case, the biased sampling degrades gracefully to random sampling.

3.1. Region representations

In order to predict the probability of the object of interest in a region, we need a way to represent the region. We utilize simple region-based features such as color, texture, and shape, to characterize the basic analysis unit, segment.

For color, we use the hue mode in a region. The hue space is quantified into \( K_c \) clusters.

For texture, we use the algorithm developed by Martin et al. [16] to compute the texton histograms. The texton at each pixel is a vector of responses to 24 filters quantized into 32 textons, and the texton words in a segmented region are then accumulated to form a texton histogram. These histograms can be clustered to create a different vocabulary of region descriptors in size \( K_T \).

For geometric measures, we employ moment invariants [9] which define a set of region properties that can be used for shape classification and part recognition. For each region, we obtain a 7-dimension vector to describe its shape. Vectors of all regions would be clustered into a vocabulary of region descriptors of size \( K_s \).

In summary, for each region in an image, we compute features of a color mode (1-dimension), a texton histogram (32-dimension), and a shape descriptor (7-dimension). By assuming independence of these features, we could build vocabularies of each feature separately. In this paper, we choose \( K_c = K_T = K_s = 100 \) to quantify all feature sets.

3.2. Region score by the classification model

Let \( F \) represent the region-based feature. \( F_i \) is one of the entries in a certain vocabulary. To generate the object class prior map, we need to learn scores for each region-based feature. The scoring method used in this work is similar to the method by Pantofaru et al. [21]. We assign a score to each feature which indicates how well it discriminates between the object and background based on the image labels. Let \( O \) indicate the presence of the positive object class in an image, and \( \tilde{O} \) the absence of the object in an image. We can define \( \tilde{R} \) as the posterior belief in \( O \) given \( F_i \) (assuming that \( P(O) = P(\tilde{O}) \)):

\[
\tilde{R}(F_i) = P(O|F_i) = \frac{P(F_i|O)}{P(F_i|O) + P(F_i|\tilde{O})} \in [0,1].
\]  (1)

For \( \tilde{R} \), a score of 0 implies a negative image, a score of 1 implies a positive image, and a score of 0.5 is uninformative. In our algorithm, we choose

\[
R(F_i) = \exp(\tilde{R}(F_i) - 0.5).
\]  (2)
After the learning process, we could obtain three feature vocabularies and their corresponding score tables. Given a new image, we generate its segmentation, compute features for each region, assign them to their nearest neighbors in color $C_i$, texture $T_i$, and shape $SH_i$ vocabularies, and compute three corresponding scores for each region $R(C_i)$, $R(T_i)$, and $R(SH_i)$. The top-down object class prior map could be computed by

$$O_{td}(x, y) = N(R(C_i)R(T_i)R(SH_i)), \forall (x, y) \in \text{Region}_i.$$  \hfill (3)

$N(\bullet)$ is a normalization operator.

### 4. Bottom-up biological inspired saliency map

The bottom-up saliency map is commonly used to assist localization of objects. There are many ways to build a saliency map. In this research, We adopt the method in [10] to obtain a biological inspired saliency map. The method is able to select highly salient points and pre-attentive, low-level feature descriptors for these points. The procedure of the method is summarized below.

The method identifies salient points by computing seven center-surround features: image intensity contrast, red/green and blue/yellow double opponent channels, and seven center-surround features: image intensity contrast, the method is summarized below.

The method identifies salient points by computing seven center-surround features: image intensity contrast, red/green and blue/yellow double opponent channels, and four orientation contrasts. Center-surround operations are implemented as a difference between the images at two scales: the image is first interpolated to the finer scale, and then subtracted point by point from the image at the previous scale.

The input image $X$ is sub-sampled into a Gaussian pyramid, and each pyramid level is decomposed into channels for red ($R$), green ($G$), blue ($B$), yellow ($Y$), intensity ($I$) and local orientation ($O_B$). If $r$, $g$ and $b$ are the red, green and blue values of the color image, normalized by the image intensity $I$, then $R = r - (g + b)/2$, $G = g - (r + b)/2$, $B = b - (r + g)/2$, and $Y = r + g - 2|(r - g) + b|$ (negative values are set to zero). Local orientations $O_B$ are obtained by applying Gabor filters to the images in the intensity pyramid $I$. From these channels, center-surround “feature maps” are constructed and normalized:

$$F_{I,c,s} = N(|I(c) \ominus I(s)|),$$

$$F_{RG,c,s} = N(|(R(c) - G(c)) \ominus (R(s) - G(s))|),$$

$$F_{BY,c,s} = N(|(B(c) - Y(c)) \ominus (B(s) - Y(s))|),$$

$$F_{O,B,c,s} = N(|O_B(c) \ominus O_B(s)|),$$

where $\ominus$ denotes the across-scale difference between two maps at the center $(c)$ and the surround $(s)$ levels of the respective feature pyramids. The feature maps are summed over the center-surround combinations using across-scale addition $\oplus$, and the sums are normalized again:

$$\bar{F}_{I,c,s} = N(F_{I,c,s}), \forall l \in L_I \cup L_C \cup L_O$$ \hfill (5)

with $L_I = I$, $L_C = \{RG, BY\}$, and $L_O = \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$.

For the general features intensity, color and orientation, the contributions of the sub-features are linearly summed and normalized once more to yield “conspicuity maps”:

$$C_I = F_I, C_c = N(\sum_{l \in L_C} F_l), C_o = N(\sum_{l \in L_o} F_l).$$ \hfill (6)

All conspicuity maps are combined into one saliency map:

$$S = \frac{1}{3} \sum_{k \in I,C,O} C_k.$$ \hfill (7)

In this research, we combine the saliency map with segments $\{\text{Region}_i\}$, thus enforce weak spatial co-coherency.

$$S_{bu}(x, y) = \max_{(x,y) \in \text{Region}_i} S(x, y), \forall (x, y) \in \text{Region}_i.$$ \hfill (8)

### 5. Probabilistic distribution generation

We optimally combine both top-down and bottom-up information to formulate the final sampling bias. The two maps computed from the previous sections are combined by searching for the optimal weight which makes the salient regions in the final map more compact. Let the bottom-up saliency map be $S_{bu}$ and the top-down class prior map be $O_{td}$, the probability distribution $T$ over regions of different locations can be represented as:

$$T = \alpha S_{bu} + (1 - \alpha)O_{td},$$ \hfill (9)

where $\alpha$ is an image dependent parameter. Binarizing $T$ gives $n$ separate region masks $\{B_i| i = 1, \ldots, n\}$. The compactness of an image is the sum of the compactness of all regions in $T$ (corresponding to the region masks). The compactness of a region is evaluated by formulating it as a probability density function and computing its variances:

$$T^i_{com} = T^i_{varX} + T^i_{varY},$$ \hfill (10)

where

$$T^i_{varX} = \sum_{(x,y) \in B_i} \frac{(x - T^i_{meanX})^2}{\sum_{(x,y) \in B_i} T(x,y)} T(x,y)$$

and

$$T^i_{meanX} = \sum_{(x,y) \in B_i} x T(x,y) \frac{1}{\sum_{(x,y) \in B_i} T(x,y)}$$

are the spatial variances and means of the object probabilities in the $x$ direction, respectively. The spatial variances and means in the $y$ direction can be defined similarly. $T(x,y)$ denotes the likelihood of existing objects of interest at location $(x,y)$. We search for the optimal $\alpha$ between $[0, 1]$ at 50 possible values. Figure 3 gives several examples of the map combination process. The final map $T$ is normalized to $[0,1]$. In this way, we can ensure all regions including regions with zero probability of object existence to have a chance to be sampled.
Figure 3. Examples of results from combing the bottom-up saliency map and top-down class prior map. From left to right each row shows the original image, the bottom-up saliency map \( S_{bu} \), the top-down class prior map \( O_{ct} \), and the final 2D probability map \( T \) respectively.

We further convert the 2D object locations probability distribution into a 3D representation with \( x \) position, \( y \) position, and scale being the three dimensions. To attain computation efficiency, we use integral images of \( T \) to compute the saliency of regions with arbitrary scale and on arbitrary positions of the image

\[
I_b(x, y) = \sum_{x' < x, y' < y} T(x', y').
\]

The 3D probability density function (PDF) could be approximated and normalized from:

\[
P(x, y, \text{scale}) = I_b(x + \frac{\text{scale}}{2}, y + \frac{\text{scale}}{2}) + I_b(x - \frac{\text{scale}}{2}, y - \frac{\text{scale}}{2}) - I_b(x + \frac{\text{scale}}{2}, y - \frac{\text{scale}}{2}) - I_b(x - \frac{\text{scale}}{2}, y + \frac{\text{scale}}{2})
\]

We then can use this 3D PDF to bias samples over image patches.

6. Evaluations

6.1. Experiment design

We carried out two groups of experiments. The first group was on Graz-02 and VOC05 to compare the proposed biased sampling method with other state-of-the-art sampling techniques. We chose these two data sets because they are popular data sets used in evaluating patch sampling techniques [18, 19, 23]. Three baselines of sampling techniques were selected: (1) an interest-point based saliency measure using the Harris of Laplace detector; (2) the random patch sampling method [19]; and (3) the method from [23] that measures co-occurrence class specific content saliency information and uses that in the direct sampling process. We use SIFT descriptors for the sampled patches and construct a codebook of feature descriptors using standard \( K \)-means clustering with \( K = 1000 \). Descriptors were coded by hard assignment to the nearest codebook centers, yielding a histogram of codeword counts for each image. Binary one against other SVM is used for classification. In addition, we also compared the performance of the proposed sampling method with some state-of-the-art class dependent feature selection methods [13, 18, 20], even they are for different components in the BoF framework.

The second group of experiments uses the up-to-date VOC2008 dataset to further demonstrate the efficacy of the proposed biased sampling component and evaluate its impact on the performance of object categorization systems. We combined the proposed sampling method with a top performing system in both VOC2007 and VOC2008 competitions [25]. To illustrate the potential of the proposed sampling component, and to make experiments comparable, we kept the framework and parameter setting as reported in [25], except that we used the proposed biased sampling component instead of Harris-Laplace detector together with dense sampling in the system reported in [25].

6.2. Experimental results on ”Graz-02”

Graz-02 contains 4 categories: bikes, cars, persons, and background. The average ratio of the size of the object versus the size of image is: 0.22 for bikes, 0.17 for persons, and 0.09 for cars. We used the setting in [20] in our experiments. We took 150 images of the object category as positive images and 150 of the counter-classes as negative images for training, and 300 images half belonging to the category and half not for testing. For each category, we measure the performance by ROC curves, summarized by the classification accuracy at the equal error rate (EER) point.

Figure 4 shows the ROC curves for each test on Graz-02: person vs. \( N \), bike vs. \( N \), and car vs. \( N \). Table 1 reports the categorization results measured in ROC-ERR of sampling methods on the three categories of this dataset. For all the baseline techniques, we fixed the number of sampling patches to be 1000, except for Harris-Laplace detector. In Harris-Laplace detection, the cornerness threshold would control the number of samples. In the experiment, we fixed a small one to ensure enough informative patches are sampled. The proposed biased sampling method performs the best among the four algorithms for each category. The classification accuracy for category person improved by about 10% over the best result of the three baselines. Compared to cars and bikes, person could appear in dramatically different backgrounds, challenging those techniques that model

<table>
<thead>
<tr>
<th>Method</th>
<th>person</th>
<th>car</th>
<th>bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.902</td>
<td>0.871</td>
<td>0.851</td>
</tr>
<tr>
<td>D.Parikh et al. [23]</td>
<td>0.810</td>
<td>0.824</td>
<td>0.784</td>
</tr>
<tr>
<td>Random sampling in [19]</td>
<td>0.785</td>
<td>0.772</td>
<td>0.777</td>
</tr>
<tr>
<td>Harris-Laplace detector</td>
<td>0.742</td>
<td>0.824</td>
<td>0.764</td>
</tr>
</tbody>
</table>

Table 1. Classification rates at EER for GRAZ-02
Figure 4. The ROC curves of the binary classification results for each category in Graz-02.

Figure 5. The average classification accuracy changes with the number of sampling patches for different sampling methods.

Figure 6. An illustration of results from different sampling methods. Each row from left to right shows sampling results from: random sampling, Harris-Laplace detector, D.Parikh et al.’s context saliency-based sampling, and the proposed biased sampling.

Table 2. A comparison of object categorization performance among our method and some feature selection techniques.

<table>
<thead>
<tr>
<th>Method</th>
<th>person</th>
<th>car</th>
<th>bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.902</td>
<td>0.871</td>
<td>0.851</td>
</tr>
<tr>
<td>A. Opelt et al. [20]</td>
<td>0.700</td>
<td>0.689</td>
<td>0.765</td>
</tr>
<tr>
<td>F. Moosmann et al. [18]</td>
<td>-</td>
<td>0.799</td>
<td>0.844</td>
</tr>
</tbody>
</table>
6.3. Experiments on the VOC2005

We chose VOC2005 specifically because it has two test sets whereas such a division no longer exists in later Pascal datasets. The test set 2 has been freshly collected for the Pascal challenge and therefore is not expected to have the same background distribution as the training data. We chose to use this dataset to prove the proposed method’s ability for modeling objects independent of datasets. VOC2005 contains four categories (motorbikes, bicycles, people and cars). It has a 684 training + val set, a 689 image test set 1, and a harder 1282 image test set 2. We conducted experiments on both test sets and show results for four binary classification problems on each test set.

We used the train + val data as training set, and chose test 1 as test data. The classification accuracy at EER is shown in table 3. We again fixed the number of sampling patches to be 1000 for all the baseline techniques except for Harris-Laplace detector, for which the same threshold in section 6.2 was used. We obtained equitable or better performance compared to the best reported results in [7].

The differentiation of our biased sampling strategy lies in its aptness at object modeling rather than scene modeling. The success of whole-image based statistical methods rests on modeling the scene where objects of interest appear. However, in reality, the appearances of background scenes usually vary more dramatically than the object itself. For test 1, most images came from the same collector and have similar background distribution as the training set. Therefore, to validate the effectiveness of ‘object modeling’ by our biased sampling strategy, we also performed experiments on the VOC2005 test 2. Table 4 shows that the biased sampling method outperforms all other three sampling methods by at least 3 percent for each category. Our method has proved to be superior at modeling objects rather than scenes and is dataset independent.

<table>
<thead>
<tr>
<th>Method</th>
<th>motorbike</th>
<th>person</th>
<th>car</th>
<th>bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.978</td>
<td>0.935</td>
<td>0.954</td>
<td>0.956</td>
</tr>
<tr>
<td>D.Parikh et al. [23]</td>
<td>0.978</td>
<td>0.921</td>
<td>0.951</td>
<td>0.941</td>
</tr>
<tr>
<td>Random sampling [19]</td>
<td>0.941</td>
<td>0.890</td>
<td>0.927</td>
<td>0.891</td>
</tr>
<tr>
<td>Harris-Laplace detector</td>
<td>0.803</td>
<td>0.859</td>
<td>0.852</td>
<td>0.856</td>
</tr>
<tr>
<td>Challenge Best [7]</td>
<td>0.977</td>
<td>0.917</td>
<td>0.961</td>
<td>0.930</td>
</tr>
</tbody>
</table>

Table 3. Classification rates at EER on VOC2005 test set 1

<table>
<thead>
<tr>
<th>Method</th>
<th>motorbike</th>
<th>person</th>
<th>car</th>
<th>bike</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours</td>
<td>0.845</td>
<td>0.867</td>
<td>0.831</td>
<td>0.830</td>
</tr>
<tr>
<td>D.Parikh et al. [23]</td>
<td>0.813</td>
<td>0.763</td>
<td>0.798</td>
<td>0.777</td>
</tr>
<tr>
<td>Random sampling [19]</td>
<td>0.735</td>
<td>0.705</td>
<td>0.701</td>
<td>0.711</td>
</tr>
<tr>
<td>Harris-Laplace detector</td>
<td>0.690</td>
<td>0.672</td>
<td>0.647</td>
<td>0.669</td>
</tr>
<tr>
<td>Challenge Best [7]</td>
<td>0.798</td>
<td>0.719</td>
<td>0.720</td>
<td>0.728</td>
</tr>
</tbody>
</table>

Table 4. Classification rates at EER on VOC2005 test set 2

6.4. Experiments on the VOC2008

To assess the impact of the proposed biased patch sampling method on object categorization performances, we conduct experiments on VOC2008 Challenge dataset which contains 20 classes. In table 5, we list some results from the challenge 2008’s winner [6] (only the first ten classes due to space limit).

We adopted a system from [25], which is based on the BoF framework and has obtained good results in both VOC2007 and 2008 challenges. Its average precision by each category is listed in table 5 as UvASoft5ColorSift (method’s name in VOC Challenge 2008). In our experiment, we only changed the sampling method from dense sampling plus Harris-Laplace interest point detection to the biased sampling method, and kept the other components and parameters the same. For each image, we fixed the number of sampling patches to 2000. Note that this number is smaller than that of the original UvASoft5ColorSift. Average Precision by each category is then computed and the results by integrating the proposed biased sampling component into UvASoft5ColorSift are showed in Table 5. Our system achieves higher median average precision rate on the twenty classes than UvASoft5ColorSift, while being comparable or even superior than the winner of VOC Challenge 2008. Based on the results, it is evident that the proposed sampling strategy led the BoF based methods toward a superior performance in object categorization on VOC2008 dataset even with fewer sampled patches.

7. Conclusions

We have proposed a novel sampling strategy for statistics-based object modeling. Compared to the random or interest point sampling methods, the proposed method focuses more samples on the objects of interest based on a prior distribution coming from a combination of a prior map learned from the training data and a bottom-up saliency map. Compared to the template or structure-based object modeling methods, the proposed method does not need a precise and robust class model or precise predictions of object location. We obtain the probability distribution of objects of interest over different locations and scales of patches by optimally combining the bottom-up biologically inspired saliency information and loose top-down class prior information for each image. The information from two directions would complement each other to steer the sampling strategy to obtain superior object models. We have evaluated the proposed algorithm on three challenging datasets and compared it with several state-of-the-art algorithms within the BoF framework. Experiments have demonstrated that the proposed method outperforms other state-of-the-art sampling technologies, and leads to a better performance in object categorization on VOC2008 dataset.
Acknowledgements

This research was partially supported by China Scholarship Council, the State Key Program of National Natural Science of China (Grant No.60635050), and NIH under Genes, Environment and Health Initiative of US (Grant No.U01HL91736). The third Author was partially supported by National Science Foundation of USA.

References


<table>
<thead>
<tr>
<th>Challenge 2008 winner[6]</th>
<th>0.795</th>
<th>0.543</th>
<th>0.614</th>
<th>0.648</th>
<th>0.300</th>
<th>0.521</th>
<th>0.595</th>
<th>0.594</th>
<th>0.489</th>
<th>0.336</th>
<th>0.569</th>
</tr>
</thead>
<tbody>
<tr>
<td>UvASoft5ColorSift [25]</td>
<td>0.797</td>
<td>0.521</td>
<td>0.615</td>
<td>0.655</td>
<td>0.291</td>
<td>0.465</td>
<td>0.583</td>
<td>0.574</td>
<td>0.482</td>
<td>0.279</td>
<td>0.548</td>
</tr>
<tr>
<td>The proposed biased sampling +UvASoft5ColorSift</td>
<td><strong>0.819</strong></td>
<td>0.541</td>
<td><strong>0.615</strong></td>
<td><strong>0.664</strong></td>
<td>0.299</td>
<td>0.519</td>
<td><strong>0.607</strong></td>
<td><strong>0.594</strong></td>
<td>0.487</td>
<td><strong>0.361</strong></td>
<td><strong>0.577</strong></td>
</tr>
</tbody>
</table>

Table 5. Comparisons of average precision by method and by class (we only list the first 10 classes) on VOC2008.