LETTER

Incorporating Top-Down Guidance for Extracting Informative Patches for Image Classification

Shuang BAI[†a)], *Nonmember*, **Tetsuya MATSUMOTO**[†], **Yoshinori TAKEUCHI**[†], **Hiroaki KUDO**[†], *and* **Noboru OHNISHI**[†], *Members*

SUMMARY In this letter, we introduce a novel patch sampling strategy for the task of image classification, which is fundamentally different from current patch sampling strategies. A top-down guidance learned from training images is used to guide patch sampling towards informative regions. Experiment results show that this approach achieved noticeable improvement over baseline patch sampling strategies for the classification of both object categories and scene categories.

key words: *patch sampling, informative regions, image classification*

1. Introduction

In image classification, incorporating top-down information is a common strategy. However, to utilize high-level information as top-down guidance often involves constructing a specific model for each class [9], [10]. Consequently, complex algorithms and intensive computation both in model training and testing are necessary. This makes it difficult to extend to a large number of classes, and approaches based on specific class models are sensitive to variations of image contents.

On the other hand, the bag of visual words (BOVW) approach appears to be discriminative and robust for image classification despite of its simplicity [1]. In this framework, first, a codebook of visual words is constructed by applying vector quantization to local image patches extracted from training images. Then, every extracted patch of an image is encoded by assigning it to its nearest visual word in the codebook. In this way, an image is represented as a histogram indicating the frequency of visual words appearing in it. The BOVW framework represents image categories implicitly by the distribution of visual words over the codebook.

Since BOVW was introduced into the field of image classification, it has attracted more attention. As one essential step of this framework, to extract informative patches for creating image representations plays an important role in computation efficiency and classification performance. At first, patches are extracted based on interest point detectors such as Harris-Laplace [2] and Difference of Gaussian [3]. However, in these cases, patches are extracted through a data-driven process. Although they are salient in an individual image, it does not necessarily guarantee that they are informative for image categories. Later, the work in [4] showed that dense sampling is able to outperform interest point detectors, if a large number of patches are used. Nevertheless, this will increase the requirement of computation and memory, accordingly. Therefore, to be practical, procedures must be designed to extract a limited number of informative features for creating image representations, where merely relying on low-level information is not enough.

Recently, approaches are proposed to incorporate high-level information to select a small number of informative patches from all extracted patches. In [5], the point-wise mutual information between each extracted patch and image categories is calculated to evaluate how informative this patch is. In [12], the authors propose to use co-occurrence and spatial information of image patches to construct a contextual saliency measure for each extracted patch. In these approaches, after extracted patches are evaluated, weighted re-sampling is performed based on evaluation results. So patches with high evaluation value are more likely to be selected. Finally, only re-sampled patches are used for creating image representations. Although effective, these methods are not able to increase the computation and memory efficiency, since patch selection is performed after they are extracted. Considerable amount of resources has been spent for extracting these patches and evaluating them.

Approaches which are more closely related to our work are [11], [13]. These methods are designed to determine regions of interest, beforehand. Then, in the patch extraction stage, patch sampling is biased towards determined regions of interest. In [11], loose top-down prior information of each object category is learnt from labeled segments of training images. Then, the top-down information is explored to generate a probabilistic map which indicates the probability for the object of interest to appear within a certain region. [13] proposed to learn object categories against backgrounds and use prior knowledge about where the classifier can detect discriminative features to create a saliency map. However, to build the probability map or saliency map, prior knowledge of specific object classes has to be obtained and stored, which increases human labor and the complexity of these approaches, and makes them sensitive to image contents. Furthermore, these approaches can not be applied to scene image classification.

In this paper, we propose a novel patch sampling strat-

Manuscript received September 29, 2011.
Manuscript revised November 1, 2011.

[†] The authors are with the Department of Media Sience, Graduate School of Information Science, Nagoya University, Nagoya-shi, 464–8601 Japan.

a) E-mail: baishuang@ohnishi.m.is.nagoya-u.ac.jp

DOI: 10.1587/transinf.E95.D.880

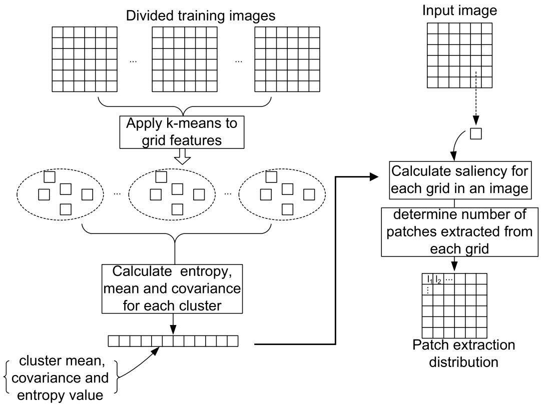
egy by incorporating top-down guidance. This approach can determine informativeness of image regions beforehand. At the same time, no specific class information is needed. Therefore, it is computationally effective and robust for image contents. Moreover, the proposed method can be easily incorporated into other image classification frameworks.

2. Incorporating Top-Down Guidance for Image Patch Sampling

In image classification, the idea to pay more attention to a region that is representative for the category it comes from but difficult to distinguish from some other categories is intuitive. And it is reasonable to use the number of patches extracted from a region to represent the attention paid to it. Therefore, the first step of the proposed method is to design an approach to evaluate the informativeness of image regions for the task of image classification. To do this, just focusing on one image is not enough. Instead, one region in an image should be investigated with respect to other images from both the same category and other categories. The informativeness of an image region should be evaluated statistically. After that, we extract image patches from each region based on the evaluation result. The number of patches extracted from an image region is proportional to its informativeness. The framework of the proposed method is given in Fig. 1.

2.1 Learning Top-Down Guidance for Patch Extraction Based on Training Images

To create a discriminative image representation, image regions that are representative for the category it comes from but difficult to distinguish from other categories need more attention. In information theory, entropy is used to measure the uncertainty associated with a random variable. The more uncertain a variable, the more informative it is. Therefore, in the proposed method, entropy is adopted to evaluate regions in an image. In the following part, we introduce how to utilize entropy values to determine informativeness of image regions.



Divided training images
...
...
Input image
Apply k-means to grid features
Calculate saliency for each grid in an image
determine number of patches extracted from each grid
Calculate entropy, mean and covariance for each cluster
cluster mean, covariance and entropy value
Patch extraction distribution

Fig. 1 Process to determine the distribution of patch extraction in an image. Images are divided into regular grids. Grids from training images are used to learn a top-down guidance by investigating statisitcal property of grid features. Then, the informativeness of grids in each image is evaluated. Finally, different numbers of patches are sampled from each image grid based on its informativeness.

To investigate statistical property of image regions, we first divide each training image into regular $n \times n$ grids, which are described by the SIFT feature [3]. To make the grid SIFT feature easier to handle, we reduce their dimension to 30 using PCA. After we extracted a fixed number of grid features from each training image, we apply k-means to them to obtain K clusters. This set of grid feature clusters are used to evaluate the informativeness of grid features in each image. We calculate a modified entropy value for each cluster over all images of all categories using the equation below

$$\begin{aligned} H(i) &= -\frac{1}{C}\sum_{c=1}^{C}\sum_{m=1}^{M_c} P_{cm}(i)\ln P_{cm}(i) \\ &\quad - \gamma \cdot max\left(-\sum_{c=1}^{C} P_c(i)\ln P_c(i) - \theta_0, 0\right), \qquad (1) \\ P_{cm}(i) &= \frac{\#f_m^i}{\#f_c^i}, \\ P_c(i) &= \frac{\#f_c^i}{\#f^i}, \end{aligned}$$

where $H(i)$ is the modified entropy value calculated for cluster i over all categories, based on training images. C is the number of categories used, and M_c denotes the number of images in category c. $p_{cm}(i)$ is the ratio between $\#f_m^i$ which is the number of features from image m of category c in cluster i and $\#f_c^i$ which is the number of features from category c in cluster i. $P_c(i)$ is the ratio between $\#f_c^i$ and $\#f^i$ which is the number of features in cluster i. γ and θ_0 are constant values.

The former part of Eq. (1) represents the distribution of a cluster i over all images of all categories. When a cluster distributes uniformly over all images of all categories, it obtains its highest value. The later part represents the distribution of a cluster over all categories. It is used for penalizing clusters that distribute uniformly across many categories. Since when a cluster distributes across many categories, it may represent background information. Therefore, clusters that distribute uniformly over many images of a small number of categories will have higher $H(i)$ value. In this case, this cluster contains representative information of a small number of categories, for which their regions are similar in appearance and need to be checked in detail to distinguish them.

By employing the above procedure, the modified entropy value of each cluster is calculated and recorded. At the same time, we represent each cluster by using the mean value and covariance of all grid features in this cluster, which are calculated

$$\overline{\mu}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \qquad (2)$$

$$S_i = \frac{1}{n_i - 1} \sum_{j=1}^{n_i} (x_{ij} - \overline{\mu}_i)(x_{ij} - \overline{\mu}_i)^T \tag{3}$$

where $\overline{\mu}_i$ is the mean value of cluster i, and x_{ij} is grid feature in cluster i, while S_i is the covariance of features in cluster i. All the above attained values are recorded together with their corresponding cluster for use in later stage.

2.2 Image Patch Extraction Distribution Calculation

After we constructed the set of grid feature clusters and calculated the values associated with each cluster, we use these information to determine the informativeness of each grid in an image. For an image divided into regular grids, each grid of this image is represented as a SIFT feature as in the previous section. To determine how informative a grid in an image is, we first calculate the Mahalanobis distance between a given grid feature and each cluster,

$$d_{ij}^2 = (x_j - \overline{\mu}_i)^T S_i^{-1} (x_j - \overline{\mu}_i), \tag{4}$$

where d_{ij} is the Mahalanobis distance between image grid feature j and cluster i. x_j is the j_{th} regular grid feature extracted from an image, $\overline{\mu}_i$ and S_i are the mean value and covariance of cluster i. After the distances between a regular grid feature and each cluster are calculated, its N nearest clusters are preserved to calculate its informativeness. That is only clusters with distances $d_{ji_1}, d_{ji_2}, \cdots, d_{ji_N} < \forall d_{ji_k}$ $(1 \le i_k \le K,\ i_k \ne i_1, i_2, \cdots, i_N)$ are preserved. Then using the N preserved nearest clusters, we calculate a saliency value for this grid feature:

$$sal(x_j) = \frac{1}{N} \sum_{i=i_1}^{i_N} \exp(-\alpha \cdot d_{ij}) \exp(\beta \cdot H(i)), \tag{5}$$

where d_{ij} is the Mahalanobis distance between grid feature x_j and its preserved nearest cluster i, α and β are constant values. $H(i)$ is the modified entropy of cluster i. By utilizing this equation, the closer a grid feature to a cluster in Mahalanobis distance, and the higher entropy value for this preserved cluster to have, the higher saliency value for this grid feature to have. Higher saliency value indicates that the grid feature is more informative for the task of image classification.

Based on the saliency value of each grid feature in an image, the number of patches extracted from each grid is determined. Extracted patches in this stage are used to create image representations. Assume L patches are going to be extracted from the whole image, then, the distribution of extracted patches in an image is proportional to the distribution of saliency values of grids. The number of patches extracted from a regularly divided grid i is calculated by

$$l_i = L \cdot \frac{sal(x_i)}{\sum_{g=1}^{n \times n} sal(x_g)}, \tag{6}$$

where l_i is the number of image patches extracted from grid i, and $\sum_{g=1}^{n \times n} sal(x_g)$ is the saliency value normalization over all grids in this image.

From the above procedure, we evaluate the informativeness of grids in an image. Based on evaluation result of each grid feature, more patches are extracted from more informative regions, while fewer patches are extracted from less informative regions.

3. Experiments and Results

In this section, experiments are designed to evaluate the proposed method. After image patches are extracted based on the proposed method, we test its performance following traditional bag of visual words procedure. A codebook of 1000 visual words are created by applying k-means to image patches extracted from training images. An image is represented by assigning patches from it to their nearest visual words in the codebook. At last, we use SVM classifier to classify image representations, where RBF kernel is adopted. We used LIBSVM [6] in this work.

In the experiments, datasets Caltech 256 [7] and SceneClass13 [8] are used. For dataset Caltech 256, 10 object classes are selected randomly with each class containing around 100 images, and for dataset ScnceClass13, all 13 classes of scene images are used with each class containing 100 images. In experiments, each class is divided into two parts with 50 images for training and the other 50 images for testing. Each evaluation experiment is performed three times under different training and testing sets division. The average is taken as the final result. We use average classification accuracy to measure the performance of patch sampling strategies. The proposed method is compared with interest point detectors [2], [3], random sampling and other feature sampling strategies utilizing top-down information [5], [12]. For object classes, saliency map learning [13] is also compared. Figure 2 shows some sample images giving the distribution of extracted patches based on the proposed method.

We first evaluate the proposed methed under different image divisions 8 × 8, 10 × 10, 12 × 12. After that, the proposed method under image division 10 × 10 is compared

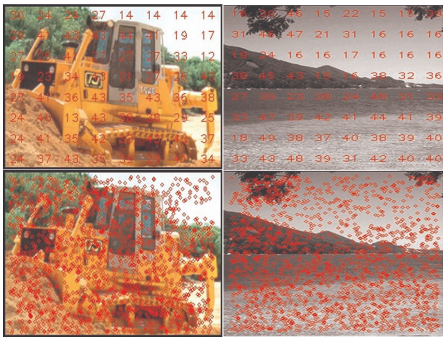


Fig. 2 Sample images indicating the distribution of extracted patches in an image. In the top images, the number in red color is the number of patches extracted from the corresponding regular grid. In the bottom images, the red dots indicate positions where image patches are extracted.

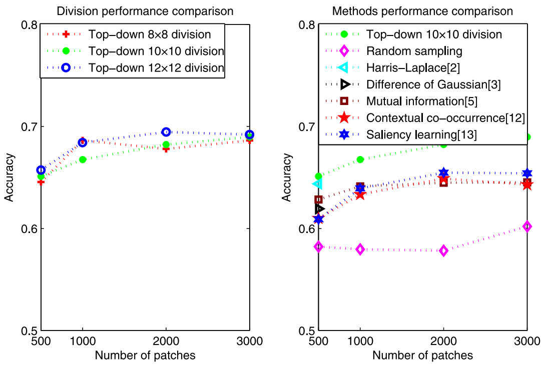


Fig. 3 Performance evaluation of the proposed method on different image divisions and its comparison with baseline patch extraction strategies on dataset Caltech 256.

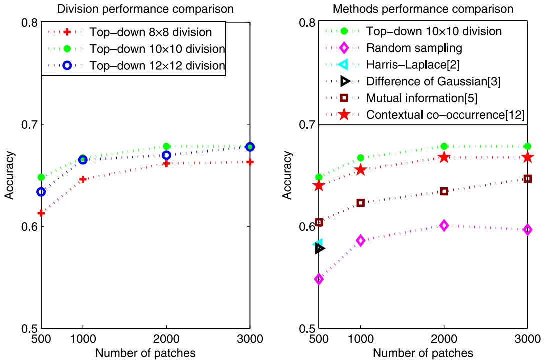


Fig. 4 Performance evaluation of the proposed method on different image divisions and its comparison with baseline patch extraction strategies on dataset SceneClass 13.

with other baseline patch sampling strategies. Experiment results on dataset Caltech256 are given in Fig. 3. And experiment results on dataset SceneClass13 are given in Fig. 4. In the experiments, interest point detectors are based on optimized parameters. Because the numbers of extracted interest points from images are usually small, we plot the interest point detectors on the patch number of 500 in Figs. 3 and 4.

From the above results, we can see that for both datasets Caltech 256 and SceneClass13, the proposed method has demonstrated superior performance over other baseline methods. In the case of object category dataset, interest point detector gained better performance than other baseline sampling strategies, when the number of extracted patches is small. And it is close to the result of the proposed method. However, when the number of patches extracted from each image increases, the performance of the proposed method also increases accordingly. Finally, it outperformed the interest point detector noticeably. While for the scene categories, the proposed method demonstrated much better performance than the baseline methods except the method based on contextual co-occurrence which has similar performance with the proposed method. Furthermore, the performance of the proposed method also increases as the number of extracted patches increases. The obtained results demonstrated the effectiveness of the proposed method.

4. Discussion

From the results shown in Figs. 3 and 4, it is noticeable that the results obtained from image division of 12 by 12 and division of 10 by 10 are better than the division of 8 by 8. The reason is that grids in fine scales can be evaluated in more detail, so that information can be represented more specifically. However, this process also increases the computational burden. At the same time, when the image grid gets smaller, it also becomes unstable. Therefore, the division of images should be a tradeoff between computation cost and classification performance.

5. Conclusion

In this letter, we have proposed a novel top-down guidance mechanism to guide patch extraction towards informative image regions for image classification. In our algorithm, the informativeness of each image region is evaluated by utilizing entropy of image grid feature clusters. Then, based on the evaluation result, the number of image patches extracted from each image region is determined. We compared the proposed method with interest point detectors, random sampling and other patch sampling strategies which explored top-down information on both object category dataset and scene category dataset. The results demonstrated the superiority of the proposed method.

References

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," ECCV International Workshop on Statistical Learning in Computer Vision, 2004.

[2] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," Int. J. Comput. Vis., vol.60, no.1, pp.63–86, 2004.

[3] D.G. Lowe, "Distinctive image features from scale-invariant key points," Int. J. Comput. Vis., vol.60, no.2, pp.91–110, 2004.

[4] E. Nowak, F. Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," Proc. ECCV, pp.490–503, 2006.

[5] F. Xu and Y.J. Zhang, "Feature selection for image categorization," Proc. ACCV, pp.653–662, 2006.

[6] C.C. Chang and C.J. Lin, "LIBSVM: A library for support vector machines," 2001. Software available at http://www.csie.ntu.edu.tw/˜cjlin/libsm.

[7] G. Griffin, A.D. Holub, and P. Perona, The Caltech-256, Caltech Technical Report, 2006.

[8] F. Li and P. Perona, "A hierarchical model for learning natural scene categories," Proc. CVPR, pp.524–531, 2005.

[9] O. Chum and A. Zisserman, "An exemplar model for learning object classes," Proc. CVPR, pp.1–8, 2007.

[10] R. Fergus, P. Perona, and A. Zisserman, "Object class recognition by unsupervised scale-invariant learning," Proc. CVPR, pp.264–271, 2003.

[11] L. Yang, N. Zheng, and H. Cheng, "A biased sampling strategy for object categorization," Proc. ICCV, pp.1141–1148, 2009.

[12] D. Parikh, C.L. Zitnick, and T. Chen, "Determining patch saliency using low-level context," Proc. ECCV, pp.446–459, 2008.

[13] F. Moosmann, D. Larlus, and F. Jurie, "Learning saliency maps for object categorization," ECCV International Workshop on the Representation and Use of Prior Knowledge in Vision, 2006.