Image Patch Selection Using A Novel Discriminative Saliency Calculation Model

Xiuming Zou[1,2,*], Huaijiang Sun[1], Sai Yang[3]

1. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu, 210094, China
2. School of Physics and Electronic Electrical Engineering, Huaiyin Normal University, Huaian, Jiangsu, 223300, China
3. School of Electrical Engineering, Nantong University, Nantong, Jiangsu, 226019, China

Abstract
This paper addresses local feature selection related to Bag-of-Features (BoF). We present a novel discriminative saliency model to determine most informative patches in dense sampling strategy. Local features after selection are used for classification. We have extensively evaluated our methods on four public benchmark datasets, each types (scene and general object), i.e. Scene 15 and Indoor67, MSRCv2 and Caltech 101. Experimental results demonstrate that our method achieves significant improvement over dense sampling in both classification accuracy and runtime.
Keywords: DENSE SAMPLING, BAG-OF-FEATURE, PATCH SELECTION, DISCRIMINATIVE SALIENCY.

1. Introduction

During the past few years, Bag-of-Features (BoF) approaches[1,2] have gradually became dominant method in image classification. In this setting, BoF first employed some samplers to extract a collection of local descriptors in each image. Further, it quantized each descriptor into a visual word according to a visual codebook, which is created off-line by k-means algorithm. Finally, each image was represented by a histogram of each visual word occurrence number.

The patch sampler is the first important step for any BoF method. Ideally, it should attract the most informative regions that are for classification in each image. The original sampler is the keypoint detectors, which have proven to be very effective in matching applications. But they were not designed to find the most informative patches for image classification[3]. Recently, dense sampling has proven to get a richer description of the images and is currently the state-of-the-art. However, processing the entire image everywhere in detail will face the challenge of a huge amount of irrelevant image patches to be processed. In addition, there is much noise and redundancy in this feature set. Therefore, it is challenging to explore this high-dimensional, noisy, and redundant feature space[4].

Saliency detection, which is closely related to selective processing in human visual system, aims to locate important regions in image has raised much attention recently. Detection of visually salient image regions is useful for applications like object segmentation, adaptive compression, and object recognition[5]. Recent study in Ref. [6] demonstrates that saliency-based bottom-up attention is indeed useful for object recognition. Inspired by the study on saliency calculation, we employ saliency algorithms to find

informative patches in dense sampling. We first segment the input image into regions using dense sampler. Next, each of the patches so obtained is further refined using the measure of discriminative saliency. In this respect, we define discriminative saliency of patches using Locality Sensitive Discriminant Analysis (LSDA).

The rest of the paper is organized as follows. In Section 2, we review the related work on patch selection and saliency calculation. In Section 3, we present our discriminative saliency calculation method to select the patches. Experimental results and comparisons with other methods are presented in Section 4. We conclude our paper in Section 5 with a summary of our observations and some pointers towards future work.

2. Related Work

As for dense sampling, the problems existing in this local feature extraction method can be concluded into the following two aspects. On one hand, many image patches are not discriminative for distinguishing different image classes. On the other hand, as the image patches are highly overlapped in the dense sampling space, it introduces significant redundancy among these features, which increases computation burden in later classification process. However, existing approaches related BoF still lack a good method to thoroughly solve these problems. Recently, Tuytelaars[7] introduced a dense interest sampling method, which is a hybrid scheme combing the strengths of sampling on a regular grid and interest point detection. In specific, it starts from image patches sampled on a regular grid, but then refines their position and scale by optimizing Laplacian criterion. In essence, this method employs an optimization criterion in the interest detection method as the measure of saliency of each patch, and then selects the most informative patches according to the result saliency value. However, just as in interest detectors, the criterion is not designed to select the most informative patches for image classification.

In recent years, modeling visual saliency[8,9] has raised much interest in computer vision research. A majority of approaches have been proposed. The difference between the various algorithms is in the way they describe the feature and compute distinctness. Color feature is the commonly used feature in saliency calculation. As for computing saliency values, the basic method is the center-surround approach. That is to say, a patch is more different from its surrounding area, the higher saliency value it will be assigned to. According to the size of the surrounding area, these methods can be further divided into two groups. The first is the local one, which calculates saliency by comparing each patch with its local neighborhoods. The second is the global one, the saliency calculation would require comparing each patch with a larger area or even the whole image[10,11].

In specific, Rutishauser et al.[6] used color and luminance features to describe the visual feature of region, and used Euclidean distance to calculate the contrast between each region to its neighbor regions. Cheng et al.[12] computed its saliency value by measuring its color contrast to all other regions in the image. Yan et al.[13] defined the local contrast saliency cue for each region in an image as a weighted sum of color difference from other regions. Achanta et al.[14] determined saliency as the distance between the average feature vector of the pixels of an image sub-region with the average feature vector of the pixels of its neighborhood. Perazzi et al.[15] introduced a parameter to control the influence radius of the uniqueness operator to combine global and local contrast estimation.

However, the above-mentioned approaches are inefficient as it requires numerous path-to-patch distance calculations, which limits their application in patch selection. Margolin et al.[16] proposed to use Principal Components Analysis (PCA) to represent the set of patches of an image and use this representation to determine the distinctness. This method improved calculation efficiency greatly, but it still cannot apply to patch selection directly. Because PCA ignores the local relationship between image patches and saliency calculation does not have the capability to determine which information most distinguishes from others. In order to remedy these problems, we use LSDA to substitute PCA, which represents the set of patches to define discriminative distinctiveness of patches. The patches in dense sampling are finally selected according to its saliency value.

3. Discriminative Saliency in Patch Selection

Let us consider we have a training set $\boldsymbol{S}=\{(I_i, y_i)\}$ i=1,...m of m labeled images, where each image Ii belongs to some image space I and each label yi is in the set $y=\{1,...N\}$. For each image, we consider that a set of patches are extracted and each of them is represented by SIFT descriptor. Let $\boldsymbol{X}$ be a set of D-dimensional local descriptors extracted from an image, i.e. $\boldsymbol{X}_i=[\boldsymbol{x}_1,\boldsymbol{x}_2,...\boldsymbol{x}_M]\in R^{D\times M}$. LSDA seeks the transformation $\boldsymbol{W}_{LSDA}$ to project input data into a subspace Z in which the local neighborhood information, as well as discriminant information of patches can be preserved[17]. The linear transformation $\boldsymbol{W}_{LSDA}$ can be obtained by minimizing an objective function as follows:

$$\min \sum_{ij} (z_i - z_j)^2 W_{w,ij}$$
$$\max \sum_{ij} (z_i - z_j)^2 W_{w,ij} \quad (1)$$

Assume $N(x_i)=\{x_i^1,...x_i^k\}$ be the set of its k nearest neighbors of patch xi, which is split into two subsets, $N_b(x_i)$ and $N_w(x_j)$. $N_w(x_j)$ contains the neighbors sharing the same label with x_i, while $N_b(x_i)$ contains the neighbors having different label. Let G_w be the within-class graph, and G_b be the between-class graph. Let W_w and W_b be the weight matrices of G_w and G_b respectively, which are defined as:

$$W_{b,ij} = \begin{cases} 1 & if\ x_i \in N_b(x_j) or\ x_j \in N_b(x_i) \\ 0 & otherwise \end{cases} \quad (2)$$

$$W_{w,ij} = \begin{cases} 1 & if\ x_i \in N_w(x_j) or\ x_j \in N_w(x_i) \\ 0 & otherwise \end{cases} \quad (3)$$

Let $D_{b,ii}=\sum W_{b,jj}$ be the column sum of W_b. $L_b = D_b - W_b$ is the laplacian matix of Gb G_b. The optimization problem in Eq.(1) can be finally converted to solve a generalized eigenvalue problem as follows:

$$X(\alpha + L_b + (1-\alpha)W_w)X^T\alpha = \lambda X D_w X^T \alpha \quad (4)$$

Let the column vector $\alpha_1,\alpha_2,... \alpha_d$ be the solutions of Eq.(4), ordered according to their eigenvalues, $\lambda_1>...>\lambda_d$. Thus, local descriptors in each image are projected in the subspace as follows:

$$x_i \rightarrow z_i = W_{LSDA}{}^T x_i \quad (5)$$

where $W_{LSDA}=[\alpha_1,\alpha_2,... \alpha_d]$ is a $D\times d$ matrix, z_i is a d-dimensional vector. Thus, the saliency of each patch is defined as l_1 norm of z_i, which is expressed as:

$$S(x_i) = \sum_{j=1}^{d} |z_{id}|_1 \quad (6)$$

where z_{id} is dth dimensional value of x_i in LSDA subspace. After all patches are processed, we sort all patches according to their saliency values, we search for the L most distinct patches in the image for later classification.

4. Experiment

4.1. Image Dataset

To validate our proposed method, we carried out several experiments on four public benchmark datasets. They are scene 15, indoor 67, MSRCv2 and Caltech101, which respectively belong to two types of scenes and general objects. The scene datasets include scene 15 and indoor 67. And the general object datasets include MSRCv2 and Caltech101.

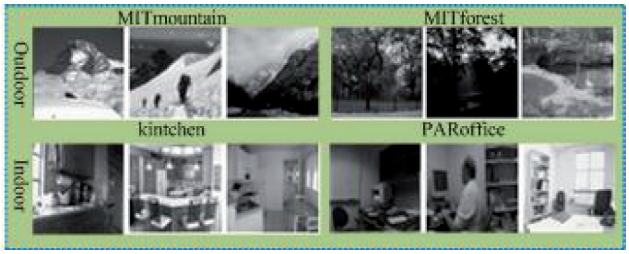

Figure 1. Example images of Scene 15 dataset.

Scene 15: Scene 15 dataset contains a total of 4485 images which spread over 15 categories varying from outdoor scenes like mountains and forests to indoor scenes like kitchens and offices. Each category has between 200 to 400 grey-level images, and the average image size is about 300×250 pixels. In Fig. 1, we show example images of indoor scenes and outdoor scenes. We follow the experimental setup used in Ref. [18], 100 images per category are randomly sampled as training and the remaining images as testing. Meantime, we randomly select 100 images from training set as validation set. In particular, we repeat the evaluation three times, and then report the average results and the corresponding standard deviation.

Indoor 67: Indoor 67 dataset has a total of 15620 images respectively belong to 67 different indoor scene categories. The minimum resolution in the smaller axis of all images is 200 pixels. The numbers of images per category are different, but there are at least 100 images in each class. As shown in Fig. 2, all the images can be further organized into the 5 big scene groups. They are store, home public spaces, leisure, and working places. Following the standard set up in Ref. [19], 80 images per category are used for training and 20 images for testing, whose partition is provided on the dataset website.

Figure 2. Example images of Indoor 67 dataset.

MSRCv2: MSRCv2 dataset[20] consists of 15 object categories and per category contains 30 images. We choose nine categories out of fifteen. They are cow, airplane, faces, cars, bikes, signs, sheep and chair. Fig. 3 gives some example images. For our experiments, we randomly select 15 images per category for training and 15 images per category for testing. At the same time, we randomly select 20 images from training set as validation set. In particular, we repeat it three times, and then report the average classification accuracy and the corresponding standard deviation.

Figure 3. Example images of MSRCv2 dataset.

Caltech 101: Caltech101 dataset[21] contains 9144 images in total from 102 different categories, including 101 object categories and one additional background category. The number of images per category ranges from 31 to 800, the resolution of most images is about 300×300. In Fig. 4, we show some examples of the dataset. In our experiments, we randomly select 30 images per category for training and test on the rest. Meantime, 500 images are selected randomly from training set as validation set. Similarly, we repeat it three times, and then report the average classification accuracy and the corresponding standard deviation.

Figure 4. Example images of Caltech101 dataset.

4.2. Experimental Setup

For our experimental validation, we performed the following default settings throughout the experiments. All the experiments are implemented in Matlab7.9. The PC we used has an Intel i3 2.1GHz CPU and 6 GB RAM. And all the four image datasets described above are used in each experiment. The SIFT features are extracted using dense sampling strategies, which is carried out by using the VLFeat library provided in Ref. [22]. Before feature extraction, all images are resized with reserved aspect ratio to no more than 300×300 pixels. We use k-means clustering to generate visual words from the resulting local features of all training images. We experiment with visual vocabularies of size 1500. Local descriptors are coded by hard assignment[23] to the nearest visual word, and use max pooling[24] method to yield an image-level representation for each image. We use SVM with linear Mercer kernel for classifying images implemented by LibSVM, and use the built-in one-versus-one approach for multi-class classification. The parameter C is determined for each SVM and the values of around C=10 typically give the best results.

4.3. Experimental Results

4.3.1. Redundancy in dense sampling

It is necessary to investigate the redundancy in dense sampling, i.e. how the patch numbers in dense sampling affect classification performance, which is controlled by the patch and step size. Thus, we implement the experiment under three conditions.

Firstly, we fix the patch size to be 8 pixels, and vary the step in {2,4,6,8,10,12,14,16} pixels. Fig.5 plots mean classification accuracies for the different datasets. From the result of Fig.5, we can see that step of 2 pixels does not get the highest classification in all datasets, which is the densest point. In the second condition, we fix the patch size to be 16 pixels, and vary the step in {2,4,6,8,10,12,14,16,18,20} pixels.

The classification accuracies of different datasets under different step are shown in Fig.6. From the result of Fig.6, we can see that except MSRCv2 dataset, step of 2 pixels does not get the highest classification in other datasets, which is the densest point. In the third condition, the patch size and step is set to be the same, which is varied in{4,6,8,10,12,14,16} pixels. The classification result is in Fig.7 shows that in all datasets, step of 4 pixels does not get the highest classification in other datasets, which is the densest point. From the above analysis, we can get the following conclusions: there indeed exists redundancy in dense sampling.

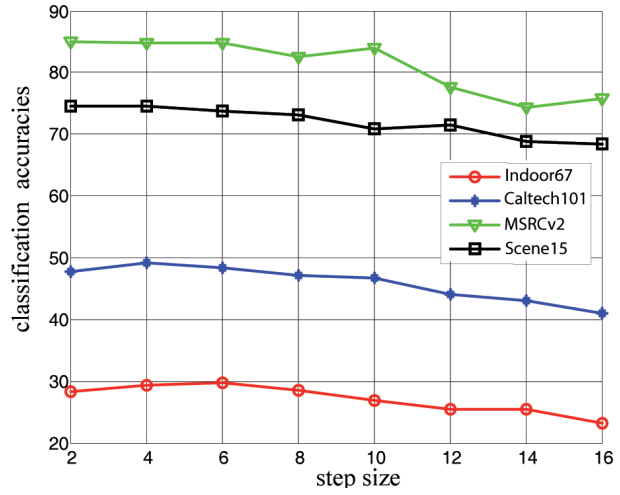

Figure 5. Classification accuracies of different step size when patch size is set to 8.

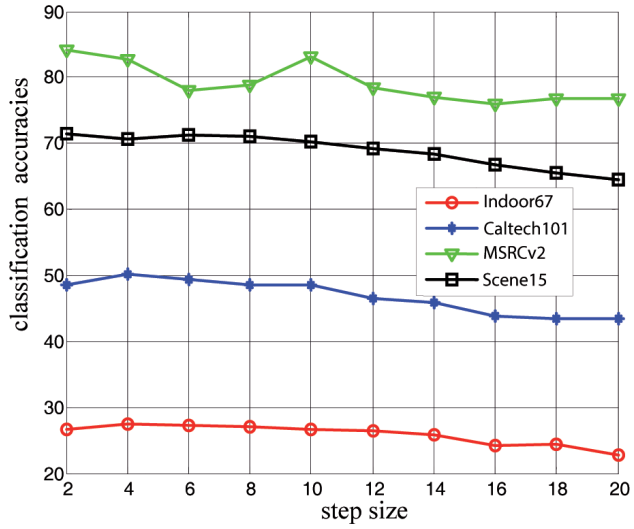

Figure 6. Classification accuracies of different step size when patch size is set to 16.

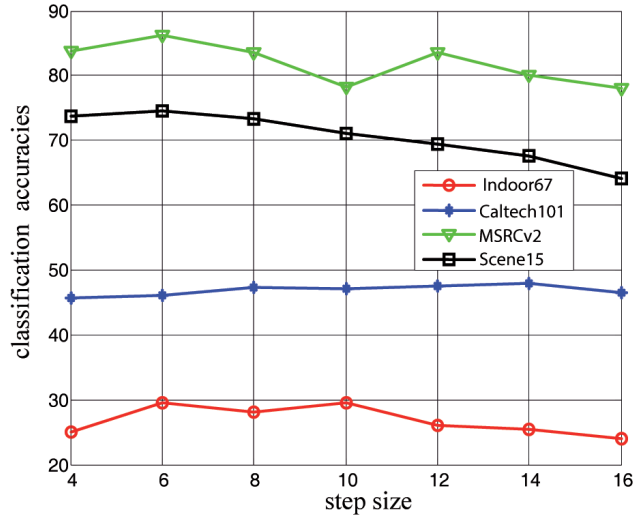

Figure 7. Classification accuracies when step size and patch are same.

4.3.2. Patch selection

In order to validate the efficiency of our method, we compared classification performance of selection method using our discriminative saliency with that not using any selection method. These two methods are denoted as DSS and NUAS respectively. During dense sampling, the patch size is set to be 8 pixels, the step is set to be{2,4,6,8} pixels. The validation set is used to optimize the number of L in selection. All the classification results are summarized in Table 1. We can see that in all datasets, the classification accuracies increase after selecting the patches using our proposed method. Thus, our method is very efficient to reduce the redundancy in dense sampling.

Table 1. Classification accuracy(%) comparison for patches in dense sampling with using our selection method

step size	method	Scene 15	Indoor 67	MSRCv2	Caltech 101
2 pixels	NUAS	73.65(0.40)	28.36	82.47(1.54)	47.53(0.27)
	DSS	74.36(0.31)	29.55	83.95(1.54)	49.03(0.21)
4 pixels	NUAS	73.23(0.09)	29.33	84.44(1.28)	50.17(1.39)
	DSS	74.18(0.77)	29.78	86.42(1.54)	50.17(1.39)
6 pixels	NUAS	73.85(0.39)	29.78	78.52(3.39)	47.57(0.39)
	DSS	73.97(0.40)	29.78	81.98(4.93)	48.22(0.52)
8 pixels	NUAS	72.88(0.64)	28.51	80.99(1.13)	46.37(0.15)
	DSS	73.07(0.72)	29.55	84.69(3.08)	47.06(0.39)

T In addition to the classification performance of these methods, we also compared their speed performance. The denotation of both methods is the same as that in the above part. The item includes training time and the time for classifying one test images. The calculation of the later is as follows, we recorded the total time for classifying all the test images, then divided it by the number of the test images, hence obtained the average processing time for each testing image. For scene15, Caltech 101, and MSRCv2, we ran this for 3 rounds and calculated the average result, and just once in Indoor67. Table 2 reports average running time of DSS and NUAS methods. We can see that training and test time are all reduced after using our discriminative saliency method. Therefore, our method can reduce the computational time complexity.

Table 2. Run time (s) comparison for patches in dense sampling with using our selection method

step size	learning stage	method	Scene 15	Indoor 67	MSRCv2	Caltech 101
2 pixels	training	NUAS	2840.00 (386.22)	16014	209.93 (22.02)	6539.1 (168.72)
		DSS	2305.2 (279.12)	11210	111.57 (10.24)	4713.5 (96.81)
	test	NUAS	1.93 (0.27)	2.79	1.62 (0.18)	2.36 (0.18)
		DSS	1.65 (0.34)	1.83	0.81 (0.06)	1.99 (0.28)
4 pixels	training	NUAS	717.44 (106.09)	3267.5	56.89 (5.75)	1979.4 (114.66)
		DSS	565.86 (27.13)	2352.6	31.22 (7.18)	1585.6 (146.07)
	test	NUAS	0.47 (0.03)	0.65	0.42 (0.04)	0.76 (0.12)
		DSS	0.44 (0.07)	0.52	0.23 (0.05)	0.49 (0.06)
6 pixels	training	NUAS	313.45 (10.71)	1453.1	27.24 (6.63)	695.64 (83.15)
		DSS	259.13 (14.84)	1073.7	12.55 (0.52)	589.73 (39.70)
	test	NUAS	0.21 (0.01)	0.29	0.19 (0.03)	0.28 (0.08)
		DSS	0.17 (0.01)	0.21	0.09 (0.01)	0.20 (0.02)
8 pixels	training	NUAS	287.04 (17.59)	710.29	13.52 (1.14)	428.77 (74.24)
		DSS	153.05 (12.87)	556.72	7.25 (0.72)	339.55 (30.70)
	test	NUAS	0.12 (0.01)	0.13	0.11 (0.01)	0.16 (0.04)
		DSS	0.11 (0.02)	0.11	0.05 (0.01)	0.12 (0.01)

4.3.3. Comparison with related method

Lastly we report the performance of our algorithm compared with the work reported in Ref.[7]. We implemented this algorithm, since there is no code publicly available for this method. During dense sampling, the patch size is set to be 8 pixels, the step is set to be 2 pixels. All the comparison results are summarized in Table 3. The run time is the time for classifying one test images, whose calculation is the same to that in Part 4.3.2. As the method in Ref.[7] does not consider the discriminative information of patches, from Table 3 we can see that our proposed method obtains higher accuracy except MSRCv2 dataset. The run time of our method is comparative to that in Ref. [7] in most datasets .

Table 3. Classification accuracy (%) and run time (s) comparison with related method

Performance index	method	Scene 15	Indoor 67	MSRCv2	Caltech 101
accuracy	our method	74.36 (0.31)	29.55	83.95 (1.54)	49.03 (0.21)
	method in[7]	73.84 (0.87)	28.73	85.43 (1.86)	48.82 (0.71)
run time	our method	1.65 (0.34)	1.83	0.81 (0.06)	1.99 (0.28)
	method in[7]	1.46 (0.06)	1.82	1.71 (0.02)	1.83 (0.13)

5. Conclusion and Future Work

We presented a novel method of finding discriminative patches in images using saliency calculation, which is easy to be implemented. This approach has been evaluated on four benchmark image datasets, each types (scene and general object). The experimental results show that our approach is able to reduce the redundancy in dense sampling and superior to another related method in terms of both precision and runtime. The extensions and vairations on this basic scheme include applying more efficient discriminative saliency method to solve redundancy problem in dense sampling and validating the efficiency of our selection method in other improved BoF models.

Acknowledgements

This work is supported by the Key University Science Research Project of Jiangsu Province, China (No. 15KJA460004).

References

1. B. Fernando, E. Fromont and D. Muselet, M. Sebban. Discriminative feature fusion for image classification, Proceedings of the 25th International Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 3434-3441.
2. L. J. Cao, R. R. Ji and Y. Cao, Y.Yang. Weakly supervised sparse coding with geometric consistency pooling, Proceedings of the 25th International Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012, pp. 3578-3585.
3. E. Nowak, F. Jurie and B. Triggs, Sampling strategies for Bag-of-features image classification, Proceedings of the 9th European Conference on Computer Vision, Graz, Austria, 2006, pp. 490-503.
4. B. P. Yao, A. Khosla and F. F. Li, Combining randomization and discrimination for fine-grained image categorization, Proceedings of

the 24th International Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA ,2011, pp.1577-1584.
5. R. Achanta, S. Hemami and F. Estrada, S Susstrunk. Frequency-tuned salient region detection, Proceedings of the 22nd International Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009, pp. 1597-1604.
6. U. Rutishauser, D. Walther, and C. Koch, P.Perona, Is bottom-up attention useful for object recognition? Proceedings of the 17th International Conference on Computer Vision and Pattern Recognition, Washington, DC, 2004, pp.37-44.
7. T. Tuytelaars, Dense Interest Points, Proceedings of the 23rd International Conference on Computer Vision and Pattern Recognition, 2010, pp.2281-2288.
8. T. Liu, Z. J. Yuan and J. Sun, J D Wang, N N Zheng, X Tang, H Y Shum. Learning to detect a salient object, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.33 , Issue 2,2011, pp.353-367.
9. M. Wang, J. Konrad and P. Ishwar, Image saliency: from intrinsic to extrinsic context, Proceedings of the 24th International Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA , 2011, pp.417-424.
10. K. Y. Chang, T. L. Liu and H. T. Chen, S H Lai. Fusing generic objectness and visual saliency for salient object detection, Proceedings of the 13th International Conference on Computer Vision, Barcelona, Spain, 2011, pp.914-921.
11. Y. F. Ma and H. J. Zhang, Contrast-based image attention analysis by using fuzzy growing, Proceedings of 11th ACM International Conference on Multimedia, New York, USA, 2003, pp.374-381.
12. M. M. Cheng, G. X. Zhang and N. J. Mitra, X L,Huang,S M Hu. Global contrast based salient region detection, Proceedings of the 24th International Conference on Computer Vision and Pattern Recognition, Colorado Springs, USA , 2011, pp.409-416.
13. Q. Yan, L. Xu and J. P. Shi, J Y Jia, Hierarchical saliency detection, Proceedings of the 26th International Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013, pp.1155-1162.
14. R. Achanta, F. Estrada and P. Weils, Salient region detection and segmentation, Proceedings of the 6th International Conference on Computer Vision Systems, Santorin, Greece, 2008, pp.66-75.
15. F. Perazzi, P. Krahenbuhl and Y. Pritch, A Hornung, Saliency filters: contrast based filtering for salient region detection, Proceedings of the 25th International Conference on Computer Vision and Pattern Recognition, Providence, USA, 2012, pp.733-740.
16. R. Margolin, A. Tal and L. Zelnik-Manor, What makes a patch distinct? Proceedings of the 26th International Conference on Computer Vision and Pattern Recognition, Portland, USA, 2013, pp.1139-1146.
17. D. Cai, X. F. He and K. Zhou, Locality sensitive discriminant analysis, Proceedings of international Conference on Artificial Intelligence, Vancouver, 2007, pp.1713-1726.
18. S. Lazebnik, C. Schmid and J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, Proceedings of the 19th International Conference on Computer Vision and Pattern Recognition, New York, USA, 2006, pp. 2169-2178
19. A. Quattoni and A. Torralba, Recognizing indoor scenes, Proceedings of the 22nd Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009, pp. 413-420.
20. Y. Zhang, T. Chen, Efficient kernels for identifying unbounded-order spatial features, Proceedings of the International Conference on Computer Vision and Pattern Recognition, Miami, USA, 2009, pp. 1762-1769.
21. F. F. Li, R. Fergus and P. Perona, Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories, Computer Vision and Image Understanding,Vol.106, Isuue 1, 2007, pp.59-70.
22. A. Vedaldi and B. Fulkerson, An open and portable library of computer vision algorithm, http: // www. vlfeat. org/,2012.
23. Y. G. Jiang, C. W. Ngo and J. Yang, Towards optimal bag-of-features for object categorization and semantic video retrieval, Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, 2007, pp.494-501.
24. Y. L. Boureau, F. Boch, Y. LeCun and J. Ponce, Learning mid-level features for recognition, Proceedings of the 23rd International Conference on Computer Vision and Pattern Recognition, San Francisco, USA, 2010, pp. 1-8.