Scene Classification by Feature Co-occurrence Matrix

Haitao Lang[1,2], Yuyang Xi[1], Jianying Hu[1], Liang Du[2], Haibin Ling[2]

[1] Department of Physics & Electronics, Beijing Key Laboratory of Environmentally Harmful Chemicals Analysis, Beijing University of Chemical Technology, Beijing, China, 100029

[2] Department of Computer & Information Sciences, Temple University, Philadelphia, USA, 19122

Abstract. Classifying scenes (such as mountains, forests) is not an easy task owing to their variability, ambiguity, and the wide range of illumination and scale conditions that may apply. Bag of features (BoF) model have achieved impressive performances in many famous databases(such as the *15 scene* dataset). A main drawback of the BoF model is it disregards all information about the spatial layout of the features, leads to a limited descriptive ability. In this paper, we use co-occurrence matrix to implant the spatial relations between local features, and demonstrate that feature co-occurrence matrix (FCM) is a potential discriminative character to scenes classification. We propose three FCM based image representations for scenes classification. The experimental results show that, under equal protocol, the proposed method outperforms BoF model and Spatial Pyramid (SP) model and achieves a comparable performance to the state-of-the-art.

1 Introduction

Classifying scenes into semantic categories is a problem of great interest in both research and practice. For example, an online collection of photos needs to be grouped into categories like 'coast', 'highway', and 'office' to support efficient browsing and/or retrieval tasks. At the same time, scene classification is not an easy task owing to their variability, ambiguity, and the wide range of illumination and scale conditions that may apply.

Recently, there is a trend of using low-level image features in classification of imagery data [1–3]. The development and analysis of low-level feature descriptors have been widely considered in the past years. Among the vastly employed methods are the scale-invariant feature transform(SIFT) [4], speeded up robust feature(SURF) [5], histogram of oriented gradients(HOG) [6], gradient location and orientation histogram(GLOH) [7], region covariance matrix(RCM) [8], local binary patterns(LBP) [9] etc. How to organize these local features to construct an robust image descriptor is crucial to the performance of scene classification. The most popular image representation is the bag of features(BoF) [1], which describes an image by the overall distribution of low level features. Traditional

BoF framework equally encodes all local features and does not emphasize any elements with regard to spatial layout. Hence, spatial pyramid (SP) structure representation is often used to extend the global BoF representation. SP model [2] approximates geometric layout of local features by partitioning the image plane into increasingly fine sub-regions. Due to its better performance and simple implementation, it has become a standard procedure for scene classification.

In this paper, we investigate the relationship between spatial layout of local features and the scene categories. It is shown that when using feature co-occurrence matrix(FCM) to map the original scene image from gray space to features distribution space, from a statistical point, there is an explicit difference between scene categories. Based on this observation, we propose to use co-occurrence matrix to extend the orderless BoF representation and construct three FCM based image representations. We evaluate the proposed method and compared it with original BoF model and SP model on 15 scene database with equal experimental protocol. The experimental results show that the proposed method outperforms BoF model and SP model and achieves a comparable performance to the state-of-the-art. The proposed method is a good alternative to image representation for scene classification tasks.

The remaining of this paper is organized as follows. We briefly review the related works on BoF and its extension models and co-occurrence matrix in Section 2. Then we introduce the proposed FCM method and local features used in our work in Section 3. In Section 4, we propose three FCM based image representation methods. The evaluation to our method and comparisons to others are described in Section 5. Finally, we conclude the paper in Section 6.

2 Related Works

2.1 Bag of Features Model and Its Extensions

State-of-the-art methods following the bag of features (BoF) framework mainly contain four steps: (1) local feature extraction and description, (2) feature coding/encoding, (3) feature pooling and (4) classifier learning.

The local features are firstly extracted by densely or randomly sampling, or sparse keypoints detector(such as Harris detector [10], scale and affine invariant detector [11] *etc*). SIFT [4], GLOH [7], HOG [6] *etc.* are usually used to build a descriptor to local interesting points. In "coding" step, a clustering method (such as k-means clustering [12]) is conducted over all descriptors to obtain a vocabulary (codebook). "encoding" procedure deals with how to use one or multiple codes from codebook to represent a new descriptor. Hard voting [13] , soft voting [14] and reconstruction based methods (LCC [15], sparse coding [16] *etc.*) are three typical methods. In pooling step, the quantization indices of all the local features are summarized to form the global image representation. Histogram is a typical average pooling strategy, which sums up all the occurrences of each index throughout the entire image in an orderless manner. Instead of performing averaging operation, max pooling adopts the element wise maximum values of

feature vectors over the whole image as the pooled features. The classifier learning step generally uses the kernel built on matching scores of the global image representations.

To overcome the loss of spatial information in original BoF model, Lazebnik *et al.*[2] propose the spatial pyramid matching (SPM) model. The image is subdivided at three different levels of resolution. For each level of resolution, the features falling in each sub-region (bin) are counted. Finally, each spatial histogram is weighted according to:

$$\kappa^L(X,Y) = \mathcal{I}^L + \sum_{\ell=0}^{L-1} \frac{1}{2^{L-\ell}} (\mathcal{I}^\ell - \mathcal{I}^{\ell+1}) \tag{1}$$

The success of spatial pyramid representation comes from the valid assumption that the images with similar scene and geometry layout possibly belong to the same category. While due to there exists large intra-class variation of same scene categories as well as significant inter-class similarity between different scene categories. In many visual classification tasks, the spatial distribution of discriminative information is non uniform. Thus different parts of image should serve different roles for scene classification. Sharma *et al.* [17] use the saliency maps to weight the corresponding visual features improves the discriminative power of the image representation. Chen *et al.* [18] introduces so-called side information (i.e., prior knowledge such as clues of object layout) for image classification based on BoF representation. 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.

2.2 Co-occurrence Matrix

Co-occurrence matrix is essentially a two-dimensional histogram in which the $(i,j)th$ element of the matrix $\mathbf{M}$ is the frequency of *event* i co-occurs with *event* j. Here *"event"* can be a pixel value and also can be a specific low level feature of image. In texture classification community, gray level co-occurrence matrix (GLCM) is firstly introduced by Haralick [19]. A GLCM is specified by the relative frequencies $\mathbf{M}(i,j,d,\theta)$ in which two pixels, separated by distance d , occur in a direction specified by the angle θ, one with gray level i and the other with gray level j:

$$\begin{aligned}\mathbf{M}(i,j,d,\theta) = \sum_{p=1}^{I_y}\sum_{q=1}^{I_x} I(p,q)=i \quad and \quad I(p+d_y(\theta), q+d_x(\theta))=i \\ if \quad \theta=0, \quad d_y=0 \quad and \quad |d_x|=d \\ if \quad \theta=45, \quad d_x=d_y=d \quad or \quad d_x=d_y=-d \\ if \quad \theta=90, \quad d_x=0 \quad and \quad |d_y|=d \\ if \quad \theta=135, \quad d_x=-d, \quad d_y=d \quad or \quad d_x=d, \quad d_y=-d \end{aligned} \tag{2}$$

where I_y and I_x represents the row and column number of the image I, $I(p,q)$ is pixel gray value in $p-th$ row and $q-th$ column. i and j is the gray level with the maximum H.

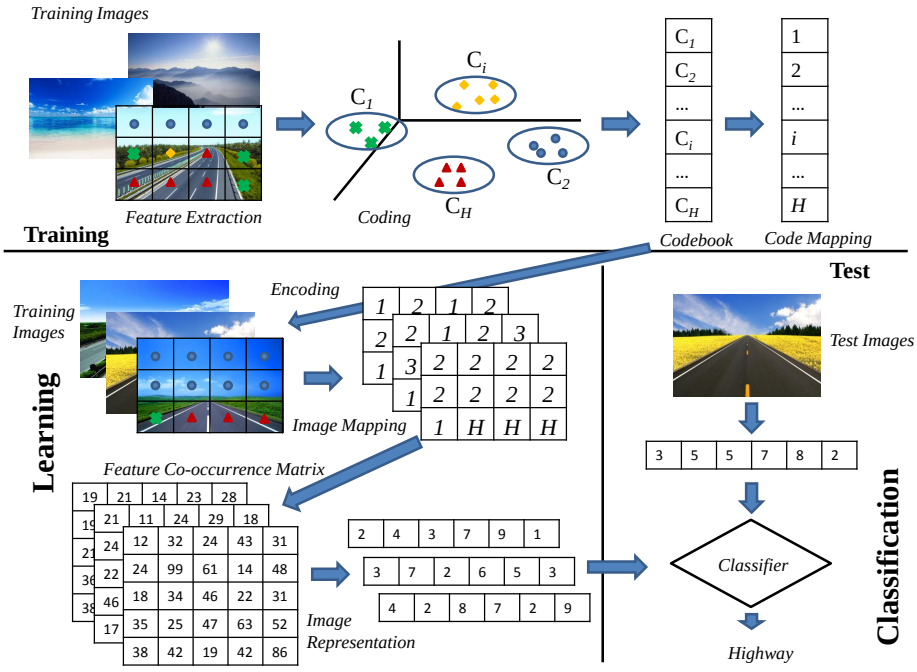

Fig. 1. Toy example of FCM based scene classification framework.

3 Feature Co-occurrence Matrix

3.1 Feature Co-occurrence Matrix

The framework to use FCM method to conduct scene classification is shown in Fig. 1. In training phase, we first use standard BoF method to construct a codebook. Then we build a mapping from codes(visual words) to numerical indexes, *i.e.* we assign a specific number to each code. The maximum number is the size of the vocabulary. By this way, we can calculate the co-occurrence relationship between codes simply. In learning phase, low level features are extracted from each image firstly, then are encoded according to the codebook. Then the image pixel value where the feature is extracted is replaced by the number which represent the corresponding code. By this way, an image is mapped from the gray/color space to code index space. After that, we compute the FCM according to equation 2. Once the frequency of each code transition is computed, a normalization is conducted to $\mathbf{M}(i, j, d, \theta)$ based on:

$$\mathbf{M}^*(i, j, d, \theta) = \frac{\mathbf{M}(i, j, d, \theta)}{\sum_{m=0}^{H} \sum_{n=0}^{H} \mathbf{M}(m, n, d, \theta)} \tag{3}$$

where H denotes the size of codebook.

Due to the similarity between two matrix is hard to evaluate, we need transform the FOM to a vector as the image representation. In this paper, we introduce three methods to conduct this operation. The details is described in Section 4.

3.2 Discrimination of FCM for Scenes Classification

We investigate the discriminative ability of FCM to different scene categories. As an instance, we plot the mean FCM of four scenes from 15 scene dataset in Fig. 2. From the heat maps, we find when describing scenes with FCMs, there are distinct statistical differences between categories. Based on this observation, we argue that FCM is a potential discriminative features to classify scenes.

4 FCM based Image Representation

We propose three strategies to build an image representation based on FCM.

4.1 Image Representation by Unfolding FCM

A simple method to construct an image representation with FCM is unfolding the matrix to a vector. Considering the dictionary capacity up to hundreds of thousands of levels, we use PCA to reduce the dimension of original unfolded vector to a reasonable level, e.g. 256 etc. Then the compact vectors extracted from four FCMs are concatenated in a single feature vector as the final image representation.

Table 1. Haralick's Statistical Properties of GLCM

Name	Formula
Angular	$f_1 = \sum_i \sum_j \mathbf{M}(i,j)^2$
Contrast	$f_2 = \sum_{n=0}^{H-1} n^2 \sum_{i=1}^{H} \sum_{j=1}^{H} \mathbf{M}(i,j)$
Correlation	$f_3 = \frac{\sum_i \sum_j (i,j)\mathbf{M}(i,j) - \mu_x \mu_y}{\sigma_x \sigma_y}$
Variance	$f_4 = \sum_i \sum_j (i-\mu)^2 \mathbf{M}(i,j)$
Inverse difference moment	$f_5 = \sum_i \sum_j \frac{1}{1+(i-j)^2} \mathbf{M}(i,j)$
Sum average	$f_6 = \sum_{k=2}^{2H} k\mathbf{M}(k)$
Sum variance	$f_7 = \sum_{k=2}^{2H} [(k-f_6)^2 \mathbf{M}_{x+y}(k)]$
Sum entropy	$f_8 = -\sum_{k=2}^{2H} \mathbf{M}_{x+y}(k) \log[\mathbf{M}_{x+y}(k)]$
Entropy	$f_9 = -\sum_i \sum_j \mathbf{M}(i,j) \log[\mathbf{M}(i,j)]$
Difference variance	$f_{10} = \sum_{k=0}^{H-1} [k - \sum_{l=0}^{H-1} l\mathbf{M}_{x-y}(l)]^2 \mathbf{M}_{x-y}(k)$
Difference entropy	$f_{11} = -\sum_{k=0}^{H-1} \mathbf{M}_{x-y}(k) \log[\mathbf{M}_{x-y}(k)]$
Measure of correlation 1	$f_{12} = \frac{f_9 + \sum_{i=1}^{H} \sum_{j=1}^{H} \mathbf{M}(i,j) \log[\mathbf{M}(i)\mathbf{M}(j)]}{\max(HX, HY)}$
Measure of correlation 2	$f_{13} = \sqrt{1 - \exp[2(\sum_{i=1}^{H} \sum_{j=1}^{H} \mathbf{M}(i)\mathbf{M}(j) \log[\mathbf{M}(i)\mathbf{M}(j)] + f_9)]}$
Maximal correlation coefficient	$f_{14} = \sqrt{\langle (\sum_{k=1}^{H} \frac{\mathbf{M}(i,k)\mathbf{M}(j,k)}{\mathbf{M}_x(i)\mathbf{M}_y(j)}) \rangle_2}$

Abbreviations:

$\mathbf{M}(i,j)$: (i,h)th entry in $\mathbf{M}$, H: dimension of $\mathbf{M}$, μ is the mean of μ_x and μ_y

$\mathbf{M}(x)_i = \sum_{j=1}^{H} \mathbf{M}(i,j)$, $\mathbf{M}(y)_j = \sum_{i=1}^{H} \mathbf{M}(i,j)$, $\mu_x = \sum_{i=1}^{H} i\mathbf{M}_x(i)$, $\mu_y = \sum_{i=1}^{H} i\mathbf{M}_y(i)$

$\sigma_x = \sqrt{\sum_{i=1}^{H}} \mathbf{M}_x(i)(i-\mu_x)^2$, $\sigma_y = \sqrt{\sum_{i=1}^{H}} \mathbf{M}_y(i)(i-\mu_y)^2$

$\mathbf{M}_{x+y}(k) = \sum_{i=1}^{H} \sum_{j=1, i+j=k}^{H} \mathbf{M}(i,j)$, $\mathbf{M}_{x-y}(k) = \sum_{i=1}^{H} \sum_{j=1, |i-j|=k}^{H} \mathbf{M}(i,j)$

$\langle \cdot \rangle$ denotes 2nd largest eigenvalue

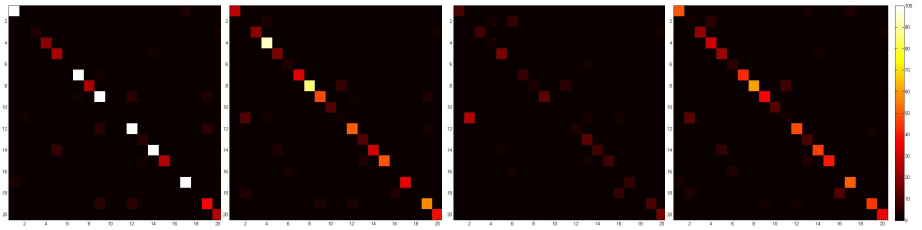

Fig. 2. FCMs of four scene categories, from left to right is coast, bedroom, forest, and kitchen respectively. SIFT is used as low level feature. The codebook size is 200. To obtain better visualization, we segment a subregion, i.e., the up-left corner of whole FCM with the area of 20×20.

4.2 Image Representation by Properties of FCM

The second method is use the properties contained in the co-occurrence matrices to construct the image representation. In this paper, we use Haralick' [19] 14 statistical properties computed from the co-occurrence matrices i.e., (1)angular second moment, (2)contrast, (3)correlation, (4)sum of squares, (5)inverse difference moment, (6)sum average, (7)sum variance, (8)sum entropy, (9)entropy, (10)difference variance, (11)difference entropy, (12-13)two information measures of correlation, and (14)maximal correlation coefficient. Once the properties extracted from four directional FCMs, we concatenate them in a single image representation vector.

4.3 Image Representation by Singular Value of FCM

In this method, we conduct a singular value decomposition (SVD) to FCM. Formally, the SVD of an $H \times H$ real feature co-occurrence matrix M is a factorization of the form:

$$\mathbf{M} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^{*} \tag{4}$$

where $\mathbf{\Sigma}$ is an $m \times m$ rectangular diagonal matrix with nonnegative real numbers on the diagonal. The diagonal entries $\mathbf{\Sigma_{i,i}}$ of $\mathbf{\Sigma}$ are known as the singular values of $\mathbf{M}$. The non-zero singular values of $\mathbf{M}$ are the square roots of the non-zero eigenvalues of both $\mathbf{M}^{*}\mathbf{M}$ and $\mathbf{M}\mathbf{M}^{*}$. To construct a image descriptor, we combine all diagonal entries of $\mathbf{\Sigma}$ corresponding to four FCMs.

5 Experiments

5.1 Databset

Scene 15 [1]is a dataset containing 15 scene categories, e.g. 'coast', 'beach', 'office', with 4485 images. The task is multi-class classification with the dataset split into 100 random images per class for training and the rest for testing.

5.2 Local Visual Features

Low level feature has significant effect on total performance of algorithm. In this paper, we evaluate and compare three features, raw gray value, SIFT [4] and LBP [9]. For color images, we first transform it to gray image, then use above algorithms to build descriptors to all sampling points in a single gray channel. For SIFT, we use the square root of normalized 128 dimensional descriptor. For LBP, a standard 256-bin histogram is used as a feature descriptor.

5.3 Experimental Protocol

Here we present some details of our experiments. In our experiments, the local features are extracted by dense sampling, the sampling interval in both row and column direction is set to 8 pixel. For gray feature, we use the average of neighboring 4×4 pixels around the sampling point as the descriptor. For SIFT, we use neighboring 16×16 pixels to describe the sampling point. While for LBP, we use neighboring 3×3 pixels to build the descriptor. In "coding" step, k-means clustering is used to construct codebook, while in"encoding" step, minimizing Euclidean distance based hard voting method is used. In this paper, the size of codebook is fixed as 256. For "pooling" step, we compare the proposed method with two baselines, i.e., original BoF model and SP model(three levels). To compute FCM, we use $0, 45, 90, 135$ four transition directions and $d = 1$ as step interval. When image is represented by unfolding FCMs, we use PCA to reduce the dimension of descriptor from $4 \times 200 \times 200$ to 256. When using FCM's properties to represent the image, we build a 4×14 dimensional descriptor. When using SVD to FCM, we build a 4×200 dimensional descriptor. RBF kernel based SVM [20] is used to learn classifier. We conduct 15 binary one-vs-rest classification problems to solve multi-class task. All experiments are repeated 10 times and the mean and standard deviation are reported.

5.4 Results

The experimental results is listed in Tab. 5.3. For gray feature, our method Unfold+FCM achieves 51.1% improving the better baseline(SP) by 17.1%. For SIFT and LBP features, our method SVD+FCM achieves 80.2% and 82.7% improving the better baseline(SP) by 2.3% and 3.7% respectively. We also compare

[1] http://www-cvr.ai.uiuc.edu/ponce_grp/data/

Table 2. Evaluation of different methods

~~Feature~~	Baseline		FCM		
	BoF	**SP**	**Unfold**	**Property**	**SVD**
gray	29.8 ± 0.8	34.0 ± 078	51.1 ± 0.6	33.5 ± 0.4	49.8 ± 0.7
SIFT	67.3 ± 0.4	77.9 ± 0.6	73.3 ± 0.9	54.8 ± 0.5	80.2 ± 0.9
LBP	74.3 ± 0.6	79.0 ± 0.5	80.9 ± 0.7	57.3 ± 0.9	82.7 ± 0.8

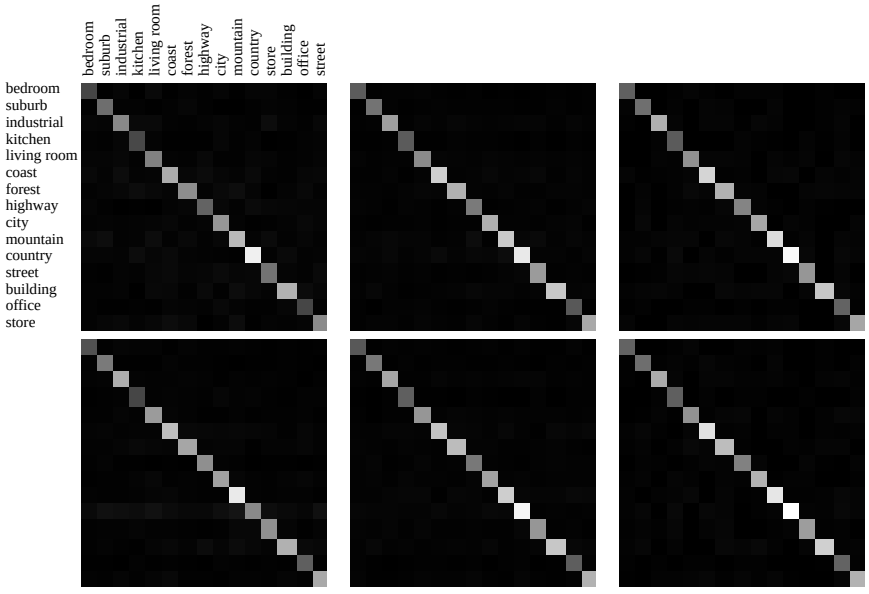

Fig. 3. Confusion matrix for evaluated methods. The top row from left to right are BoF+SIFT, SP+SIFT and SVD+SIFT respectively. The bottom row from left to right are BoF+LBP, SP+LBP and SVD+LBP respectively.

our method with two state-of-the-art methods, [17](85.5%) and [21](88.1%). The former use the saliency maps to weight the corresponding visual features and the latter combines 14 different low level features to improve the discriminative power of the image representation. While our method use a simple framework and a single feature(LBP) achieves a comparable performance. The confusion matrix is shown in Fig. 3

6 Conclusion

In this paper, we demonstrate that FCM is a potential discriminative feature to classify scenes. The experimental results show the proposed method outperforms the original BoF model and its popular extension SP model. The proposed method achieves comparable performance to the state-of-the-art on 15 scene dataset. There still is a lot of potential to improve its performance when considering the follows. The first is the size of a codebook which is controlled by the number of keypoint clusters in the clustering process. The second is that we can use the proposed framework to combine multiple complementary features to improve the performance. In future work, we will optimize the proposed method from above two aspects.

Acknowledgement. This work was supported in part by Beijing Higher Education Young Elite Teacher Project under Grants YETP0514, and NSFC under Grants 61471024. Ling was supported in part by the US NSF Grants IIS-1218156 and IIS-1350521.

References

1. Fei-Fei, L., Perona, P.: A bayesian hierarchical model for learning natural scene categories. In: Computer Vision and Pattern Recognition, IEEE Conference on. Volume 2., Ieee (2005) 524–531
2. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: Computer Vision and Pattern Recognition, IEEE Conference on. Volume 2., Ieee (2006) 2169–2178
3. Vogel, J., Schiele, B.: Semantic modeling of natural scenes for content-based image retrieval. International Journal of Computer Vision **72** (2007) 133–157
4. Lowe, D.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60** (2004) 91–110
5. Bay, H., Tuytelaars, T., Van Gool, L.: Surf: Speeded up robust features. In: Computer Vision–ECCV 2006. Springer (2006) 404–417
6. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Computer Vision and Pattern Recognition, IEEE Conference on. Volume 1. (2005) 886–893
7. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. Pattern Analysis and Machine Intelligence, IEEE Transactions on **27** (2005) 1615–1630
8. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. Computer Vision C ECCV 2006 (2006) 589–600

9. Ojala, T., Pietikäinen, M., Mäenpää, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. Pattern Analysis and Machine Intelligence, IEEE Transactions on (2002) 971–987
10. Harris, C., Stephens, M.: A combined corner and edge detector. In: Alvey vision conference. Volume 15., Manchester, UK (1988) 50
11. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. International journal of computer vision **60** (2004) 63–86
12. Hamerly, G., Elkan, C.: Learning the k in a¿ means. Advances in neural information processing systems **16** (2004) 281
13. Csurka, G., Dance, C., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: Workshop on statistical learning in computer vision, ECCV. Volume 1. (2004) 1–2
14. van Gemert, J.C., Geusebroek, J.M., Veenman, C.J., Smeulders, A.W.: Kernel codebooks for scene categorization. In: Computer Vision–ECCV 2008. Springer (2008) 696–709
15. Yu, K., Zhang, T., Gong, Y.: Nonlinear learning using local coordinate coding. In: Advances in neural information processing systems. (2009) 2223–2231
16. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, IEEE (2009) 1794–1801
17. Sharma, G., Jurie, F., Schmid, C.: Discriminative spatial saliency for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3506–3513
18. Chen, Q., Song, Z., Hua, Y., Huang, Z., Yan, S.: Hierarchical matching with side information for image classification. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE (2012) 3426–3433
19. Haralick, R.M., Shanmugam, K., Dinstein, I.H.: Textural features for image classification. Systems, Man and Cybernetics, IEEE Transactions on (1973) 610–621
20. Chang, C., Lin, C.: Libsvm: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST) **2** (2011) 27
21. J. Xiao, J. Hays, K.E.A.O., Torralba., A.: Sun database: Large scale scene recognition from abbey to zoo. In: Computer Vision and Pattern Recognition, 2010. CVPR 2010. IEEE Conference on, IEEE (2010) 1–8