Vol.2011-CVIM-175 No.25
2011/1/20

シーンコンテクストスケールを用いた画像分類

姜　有宣[†1]　杉本晃宏[†1]

画像から抽出した特徴量の密度が高いほど高性能な画像分類を実現することができる．テクストンは代表的な高密度の特徴量であり，近年ではテクスチャ解析や一般物体認識にも特徴量として多く用いられ，その有効性が確認されている．しかし，テクストンを抽出する際，テクストンのスケール最適化が考慮されていないため，画像群中でスーケル変化が大きい物体の認識は困難であり，性能低下の原因になる．そこで我々は，予め画像からシーンコンテクストスケールを求め，物体のスーケルによって異なるテキストンを用いる画像分類手法を提案する．提案手法の有効性を確認するため MSRC21 のデータベースを用いた評価実験を行い，従来法に比べ画像分類に対する大幅な精度向上が得られることを確認した．

Image Categorization using Scene-Context Scale

Yousun Kang[†1] and Akihiro Sugimoto[†1]

Densely sampling visual words tend to improve image categorization performance. Textons are representative dense visual words and they have been proven effective in categorizing materials as well as generic object classes. Despite its success and popularity, no prior work has tackled the problem of its scale-optimization for the given image data and the associated object category. We propose scale-optimized textons to learn the best scale for each object in a scene and they are utilized in image categorization. Our textonization process would produce a scale-optimized codebook of visual words thus provide improved image categorization performance. We approach the scale-optimization problem of textons as solving a scene-context scale in each image, which means the effective scale of local context to classify an image pixel in a scene. We perform textonization process using random forests which are powerful tools with high computational efficiency in vision applications. Random forests are efficiently provide both a hierarchical clustering into semantic textons and local classification. In our experiments, we use MSRC21 dataset to assess our method and show that the usage of scale-optimized textons significantly improves the performance of image categorization.

1. Introduction

After Julesz[1)] called textons for the first time, early texton studies were limited by their exclusive focus on artificial texture patterns instead of natural images[2)]. However, recent studies have been proven effective in categorizing materials[3)], various scenes[4)], and generic object classes[5)]. By employing the bag-of-words model[6)], the frameworks using textons as visual words have become a popular and have demonstrated its success in recent years. The bag-of-words model uses a compact histogram representation to record the numbers of occurrences of each visual word in an image. One of the major drawbacks of the bag-of-words model is that it discards the spatial layout of visual words. Lazebnik *et al.*[7)] proposed a spatial pyramid matching technique by utilizing a spatial pyramid image representation. In order to make a codebook including spatial layout of visual words, many works have been presented[8)], however, no prior work has tackled the problem of discard of scale information for the given image data and the associated object category.

For a given large dataset, there are many different scale of the objects present in an image. As shown in Fig. 1, even the objects are treated as same category such as 'face' or 'car', they have quite different scale in a scene. Therefore, the scale information of an object can be a significant cue for recognizing the object in a scene. Kang *et al.* proposed the scene-context scale, which is the effective scale of local context to classify an image pixel in a scene[9)]. They demonstrated the use of the scene-context scale to improve image categorization and semantic segmentation performance[10)]. We approach the scale-optimization problem of textons as solving a scene-context scale in each image pixel. In this paper, we propose the scale-optimized textons using scene-context scale for image categorization. Our method can determine the textons with most appropriate scale for each object in a scene and the class distribution of scale-optimized textons are utilized in bag-of-words model for image categorization.

†1 国立情報学研究所
National Institute of Informatics
⋆1 Contact : Yousun Kang, 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo 101–8430
National Institute of Informatics, phone: (03)4212–2566, email: yskang@nii.ac.jp

Vol.2011-CVIM-175 No.25
2011/1/20

図 1 The objects with different scale in a scene. When the object is recognized in a scene, the scale information of the object should be considered to improve the recognition performance.

The collection of texton are clustered to produce a codebook, typically with the simple but effective k-means, followed by nearest-neighbor assignment. Malik *et al.*[11] analyzed image into texton channels for image segmentation by mapping each pixel to the texton nearest to its vector of bank filter responses. They estabilshed typical textonization procerss such as computation of filter-banks, performing k-means clustering, and nearest-neighbor assignment. Unfortunately, this three stage process is extremely slow and often the most time consuming part of the whole system. Our textonization process is performed using random forests to generate a scale-optimized codebook from multi-scale textons. Random forests are powerful tools with high computational efficiency in vision applications[12].

We extended random forests method into multi-scale texton forests to find the scale-optimized textons for each object in a scene. The multi-scale texton forests can generate different textons according to scale space, where we find the best scale of textons for each category using the scene-context scale. The scene-context scale can be estimated by the entropy of the leaf node in the multi-scale texton forests. For image categorization, we combine the class distributions of estimated scene-context scale at each pixel into bag-of-words model. To assess our framework, we compare the clustering and classification accuracy and the categorization accuracy with that of the state-of-the-art[13]. The results show that our method achieves significantly better classification and categorization accuracy than those of the state-of-the-art.

This paper is organized as follows: Section 2 explains the multi-scale texton forests in detail. Section 3 describes how to combine the scale-optimized textons of each category into the bag-of-words model for image categorization module. Section 4 shows experimental results on performance and our conclusions are presented in the final section.

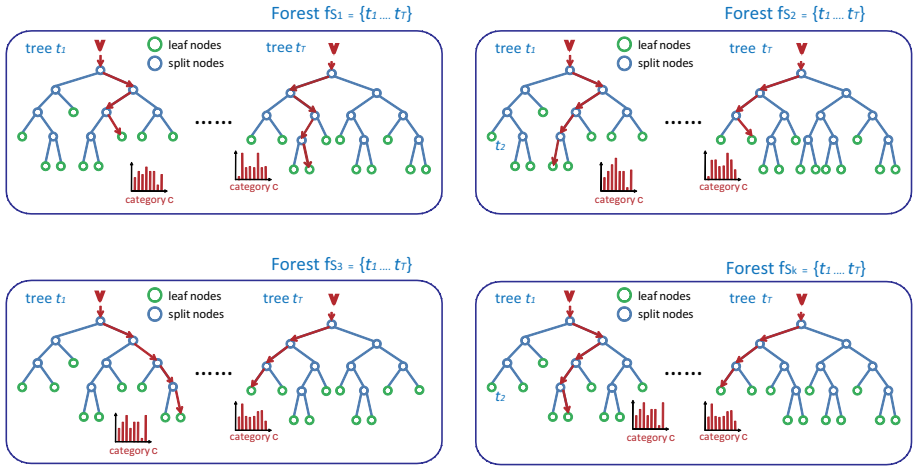


図 2 Multi-scale texton forest. The multi-scale texton forest consists of several random forests with various scale space and each random forest consists of many decision trees with same scale level.

2. Multi-Scale Texton Forests

We extend the texonization process using several random forests to formulate multi-scale texton forests. We employ the semantic texton forests method by Shotton *et al.*[13] and expand their scale level for the multi-scale texton forests. Each random forest has its own scale level and its scale level can expand by increasing the region of interest in multi-scale texton forests. Depending the size of image patches for split functions of a randomize decision tree, the effective region size can be chosen among the multi-scale texton forests with different scale. Therefore, the multi-scale texton forests are randomized decision forests created in different scale space for textonization of an image.

The multi-scale texton forests $\mathcal{F}_{\mathcal{S}}$ consist of several random forests with various scale space $\mathcal{S} = (S_1, ..., S_\tau)$. As shown in Fig. 2, a random forest $\mathcal{F}_{\mathcal{S}_k}$ is a combination of T decision trees at each scale space S_k, where the level of scale is $k = (1, ..., \tau)$. Each random forest $\mathcal{F}_{\mathcal{S}_k}$ achieves an accurate and robust classification by averaging the class distributions over the leaf nodes $L = (l_1, ..., l_T)$ reached for all T decision trees:

Vol.2011-CVIM-175 No.25
2011/1/20

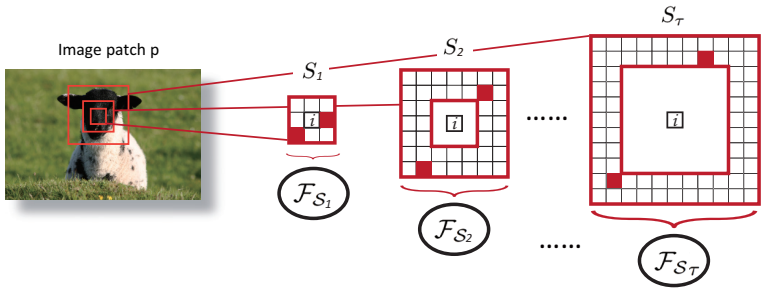


図 3 Dilatation of a region of interest according to scale space S_k. Various sizes of a region of interest are used for node split function in the multi-scale texton forests.

$$P(c|L) = \frac{1}{T}\sum_{t=1}^{T} P(c|l_t), \tag{1}$$

where c is a category label of a pixel. The nodes in the trees efficiently provide a hierarchical clustering into semantic textons with scale-contextual features.

The split nodes in multi-scale texton forests use split functions of image pixels within a region of interest. Each random forest $\mathcal{F}_{\mathcal{S}_k}$ has different set of pixel combinations within a region of interest as shown in Fig. 3. We can increase the scale level k of a random forest by dilatation of a region of interest. At the first scale space S_1, the region of interest R_{S_1} covers whole pixels within a $(d \times d)$ image patch, where the split functions f in $\mathcal{F}_{S_1}$ act on. In next scale space S_2, the region of interest R_{S_2} deals with the pixels within the difference of $(dk \times dk)$ image patch from the region R_{S_1} of a previous scale space S_1. Therefore, the region of interest R_{S_k} increases within a $(dk \times dk) - (d(k-1) \times d(k-1))$ image patch as illustrated in Fig. 3. The number of possible combinations of selecting two pixels inside a region of interest also increases quadratically with respect to the scale level k.

To textonize an image according to scale space, image patches centered at each pixel with various sizes are passed down the multi-scale texton forests resulting in semantic texton leaf nodes $L = (l_1, ..., l_T)$ and the averaged class distribution of each random forest $\mathcal{F}_{\mathcal{S}}\{p(c|L)\}$. The textons generated by each random forest can be extracted in different scales from other forests.

3. Clustering and Classification

In this section, we firstly explain how to estimate the scene-context scale of each image pixel using multi-scale texton forests. Scale-optimized textons can be obtained by finding a scene-context scale in each pixel. Scene-context scale means the effective scale of local context to classify an image pixel in a scene. Secondly, we calcuate the average of class distributions over the leaf nodes at the random forests with estimated scene-context scale. Since the class distributions are calculated at the scale-optimized random forests in each pixel, both clustering and classification guarantee good performance. Finally, we adopt the linear support vector machine (SVM) to classify each category and make a histogram consisting the class distribution to combine bag-of-words model for image categorization.

3.1 Clustering using the scene-context scale

The scene-context scale of each image pixel is obtained by computing the entropies of an image patch in the leaf nodes of each random forest. Since the objects of various size and background/foreground appear together in the image, we should compute scene-context scale per pixel. The confidence of each random forest is computed as the entropies of the class label distribution in leaf nodes. We regard the confidence as the criterion of an optimal scale level to be chosen. At each image pixel, therefore, one scale level with minimum entropy is chosen as the scene-context scale among the multi-scale texton forests.

At first, we compute the entropy $E(I|L)$ of each image patch I at leaf nodes L of a random forest as

$$E(I|L) = -P(c|L) \times \log P(c|L). \tag{2}$$

The entropy $E(I|L)$ can be computed in each random forest $\mathcal{F}_{\mathcal{S}_k}$ with each scale level $k = (1, ..., \tau)$ and we note the entropy of a random forest as $\mathcal{F}_{\mathcal{S}_k}\{E(I|L)\}$. Among the whole scale level $\mathcal{S} = (S_1, ..., S_\tau)$, the one scale level S_i^* is chosen which contains the leaf nodes of a random forest $\mathcal{F}_{S_i}$ with minimum entropy as

$$S_i^* = \arg\min_{S_i}(\mathcal{F}_{\mathcal{S}_i}\{E(I|L)\}). \tag{3}$$

The scene-context scale of an image pixel is the instance S_i^* of the most likely scale

Vol.2011-CVIM-175 No.25
2011/1/20

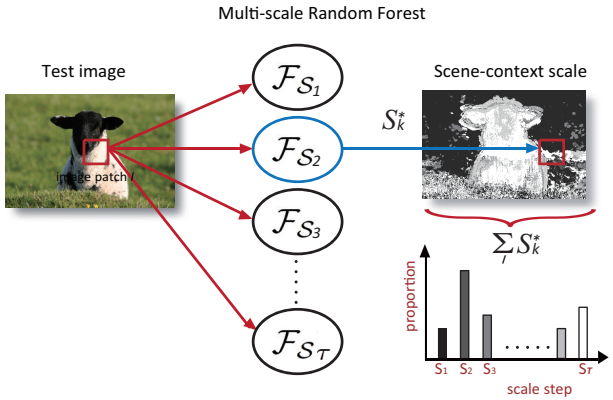


図 4 **Scene-context scale of an images.** Darker pixels correspond to smaller scale, so black pixels represent the first scale level S_1 and white pixels represent the largest scale level S_τ.

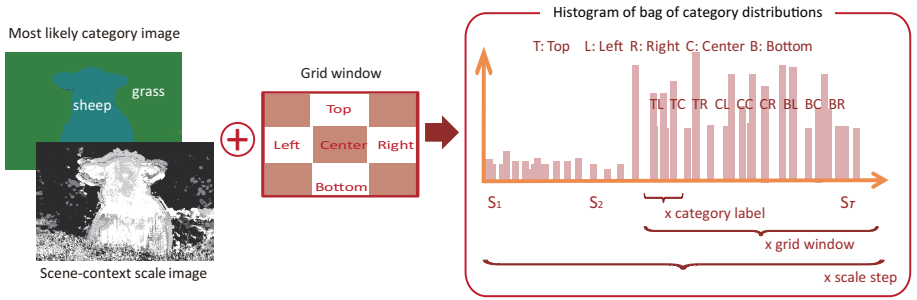


図 5 **Histogram of bag of category distributions.** The dimension of a histogram is number of grid window times number of category times total scale level

from whole scale level as shown in Fig. 4. At scene-context scale S_i^*, the category distributions $\mathcal{F}_{\mathcal{S}_i}\{p(c|L)\}$ are available for local classification and we use the category distributions consist in the histogram of a bag-of-words model. The goal of this clustering process is that we divide an image into coherent regions and simultaneously infer the class label of each region.

3.2 Classification using bag-of-words model

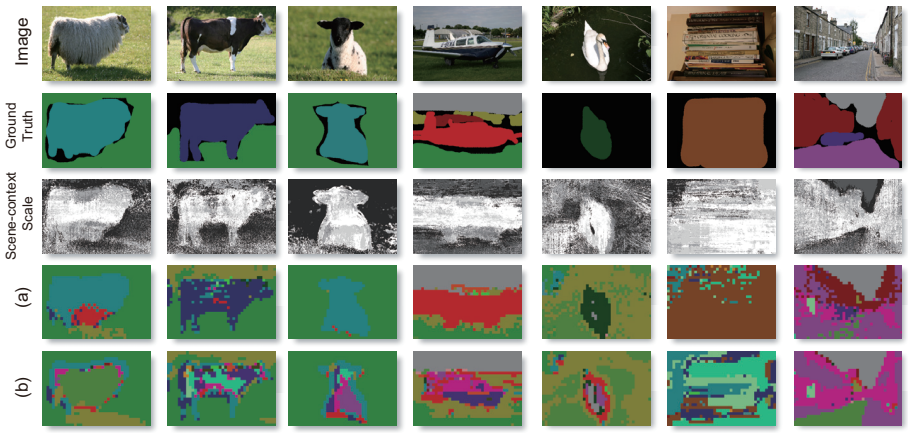
We use a bag-of-words model computed across the whole image for image categorization. Since the bag-of-words models discard spatial layout, we use a local grid window as shown in Fig. 5. The local grid window consists of nine sub-grid such as Top-Left (TL), Top-Center (TC), Top-Right (TR), Center-Left (CL), Center-Center (CC), Center-Right (CR), Bottom-Left (BL), Bottom-Center (BC), and Bottom-Right (BR). We make the histograms which consist of the class distributions at estimated scene-context scale over the whole image. To learn layout and context information automatically, we use class distributions at estimated scene in a local grid window. The scene-context scale S_k^* is chosen by using the entropy of class distribution and the class distributions consist in a histogram computed from nine grid windows from top-left (TL) to bottom-right (BR). S_1 are first chosen covering about $(d \times d)$ the pixel area. We concatenated histograms consisting of the class distributions of a scene-context scale among from S_1 to S_τ. Therefore, we make the histogram of the localized bag-of-words model using the most likely category $c_i^* = \arg\max_{c_i} P(c_i|L)$, and the most likely scene-context scale $S_i^* = \arg\min_{S_i}(\mathcal{F}_{\mathcal{S}_i}\{E(I|L)\}$ as illustrated in Fig. 4. Finally, the normalized histogram with grid windows is used as a feature vector for image categorization.

We employ the non-linear SVM algorithm to select discriminative features of the bag-of-word model. Multi-class classification is done with LibSVM[14] trained using the one-versus-all rule : a classifier is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest response.

4. Experimental Results

This section presents our experimental results for image categorization using scale-optimized texton. To assess the utility of the scene-context scale and multi-scale texton forests, we compare the classification accuracy with that of conventional semantic texton forests method[13] without using scale-optimized texton. We evaluate our algorithm using challenging MSRC21 segmentation dataset that includes a variety of objects such as building, grass, tree, cow, sheep, sky, aeroplane, water, face, car, bike, flower, sign, bird, book, chair, road, cat, dog, body, boat. We use the standard train/test splits such

Vol.2011-CVIM-175 No.25
2011/1/20



	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	global	class average
(a)	12	90	43	60	79	90	89	36	89	28	34	65	19	6	64	14	46	42	22	36	49	53.0	48.3
(b)	2	89	49	43	39	84	36	45	74	30	58	66	28	3	17	6	57	19	14	17	28	50.2	38.4

図 6 **Clustering and classification results using scale-optimized textons.** Above : (a) Classification result with using scale-optimized textons based on scene-context scale. (b) Classification result based on single-scale semantic texton forests[13] Below: Classification accuracies (percent) over the whole dataset, without-(b), and with-(a), the scale-optimized textons. Our new highly efficient scale-optimized textons achieve a significant improvement on previous work (b).

as 256 images for training, 257 images for test, and remaining 59 images for validation, and the hand-labeled ground truth to train the classifiers.

Before presenting categorization accuracy, let us show the clustering and classification results using scale-optimized texton. The multi-scale texton forests provide both a hierarchical clustering into semantic textons and local classification in various scale space. We separately train the forests in different scale space. To train the multi-scale texton forest, we prepared six scale steps $\mathcal{S} = (S_1, ..., S_6)$ and an initial image patch size is (15×15). Therefore, the size of image patches for split function f is $(15k \times 15k)$ at each scale step S_k. A randomized decision forest $\mathcal{F}_\mathcal{S}$ has the following parameters : $T = 5$ trees, maximum depth $D = 10$, $500k$ feature tests and 10 threshold tests per split,

	building	grass	tree	cow	sheep	sky	aeroplane	water	face	car	bicycle	flower	sign	bird	book	chair	road	cat	dog	body	boat	class average
(a) RBF	**97**	**98**	**98**	**100**	**100**	**99**	**100**	**98**	**99**	**98**	**100**	**100**	**100**	**99**	**100**	**100**	**94**	**99**	**100**	**97**	**96**	**98.7**
(b) PMK	90	90	73	91	93	94	100	95	77	90	100	96	100	94	96	84	78	98	97	74	93	90.6
(c) State-of-art [14]	64	86	75	86	92	90	74	66	64	88	72	84	70	53	90	67	67	57	36	64	77	72.8

図 7 **Image categorization results on MSRC21 datasets**. Categorization accuracies (percent) over the whole dataset. Scale-optimized texton achieves a improvement on previous work.

and 0.25 of the data per tree, resulting in approximately 500 leaves per tree. Training the randomized decision forest on the MSRC dataset took only 10 minutes at each scale step.

At test time, the most likely category in the averaged category distribution gives the clustering and classification results for each pixel as shown in Fig. 6. Clustering and local classification performance is measured as both the class average accuracy (the average proportion of pixels correct in each category) and the global accuracy (total proportion of pixels correct). Fig. 6 shows the results of the clustering and local classification based on scene-context scale. We estimate the scene-context scale per image pixel using multi-scale texton forests as shown in the third row of Fig. 6. Since each image pixel has the category distribution at the scene-context scale, we can infer the most likely category $c_i^* = \arg\max_{c_i} P(c_i|L)$ of leaf nodes $L = (l_1, ..., l_T)$ for each pixel i as shown in Fig. 6(a). On the other hand, Fig. 6(b) shows the results of the state-of-art[13] without using scale-optimized textons based on single-scale semantic texton forests. The single-scale semantic texton forests used the same parameter of the multi-scale texton forests with the first scale level $\mathcal{F}_{\mathcal{S}_1}$.

As shown in Fig. 6, a pixel level classification based on the local distributions $P(c|L)$ gives poor, but still good performance. The global classification accuracy without scale-optimized texton gives 50.2% and the result with using scale-optimized texton based on scene-context scale gives 53.0%. In particular, significant improvement can be observed most of the classes except some classes: tree, water, car, bicycle, sign and road. It should seem that they have not influence on scene-context scale. Across the whole MSRC21 dataset, using the scale-optimized textons achieved a class average perfor-

mance of 48.3%, which is better than the 38.4% of (b) as shown in the table of Fig. 6. Therefore, we can see that the proposed scale-optimized textons can be powerful and effective visual words of bag-of-words model and they can produce a scale-optimized codebook for image clustering.

As a result of image categorization, we obtained the accuracy of 21 categories as shown in Fig. 7. We compare the class average of our method using radial basis function (RBF) kernel and pyramid match kernel (PMK)[15] to the state-of-art[13]. We confirmed that a per-category kernel K_c shows improvement of each category from experiments and the RBF kernel improves on the PMK. As can be seen, the proposed method using the scale-optimized textons gives significantly better results than state-of-art and improves performance for all classes.

5. Conclusion

This paper presented a new framework for image categorization using scale-optimized textons. We estimated the scene-context scale from multi-scale texton forest which consist of several random forests with various scale level. The scale-optimized textons of each object can integrate the class distribution into bag of textons method. In experiments, we confirmed that the usage of scale-optimized textons significantly improves the performance of image categorization.

謝辞 This work was in part supported by JST, CREST.

参 考 文 献

1) B. Julesz, Textons, the elements of texture perception and their interactions. *Nature*, 290:91–97, 1981.
2) S. Zhu, C. Guo, Y. Wang, and Z. Xu, What are Textons? *Int. Journal of Computer Vision*, 62(1):121–143, 2005.
3) M. Varma and A. Zisserman. A Statistical Approach to Texture Classification from Single Images. *Int. Journal of Computer Vision*, 62(1):61–81, 2005.
4) S. Battiato, G. Farinella, G Gallo, and D. Ravi. Spatial Hierarchy of Textons Distributions for Scene Classification In *Proc. 15th Int. Multimedia Modeling Conf. on Advances in Multimedia Modeling*, LNCS 5371, pages 333–343, 2009.
5) J. Winn, A. Criminisi, and T. Minka. Categorization by learned universal visual dictionary. In *Proc. Int. Conf. on Computer Vision*, pages 2:1800–1807, 2005.
6) G. Csurka, C. Bray, C. Dance, L. Fan. Visual categorization with bags of keypoints. In *Proc. ECCV Workshop on Statistical Learning in Computer Vision*, 2004.
7) S. Lazebnik, C. Schmid, and J. Ponce. Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2006.
8) J. Yang, K. Yu, Y. Gong, and T.S. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2009.
9) Y. Kang and A. Sugimoto. Random Forests Based Image Categorization Using Scene-Context Scale. In *Proc. of 13th Meeting on Image Recognition and Understanding* , 2010.
10) Y. Kang, H. Nagahashi, and A. Sugimoto. Semantic Segmentation and Object Recognition using Scene-Context Scale. In *Proc. of Pacific-Rim Symposium on Image and Video Technology*, 2010.
11) J. Malik, S. Belongie, T. Leung, and J. Shi. Contour and Texture Analysis for Image Segmentation. *Int. Journal of Computer Vision*, 43(1):7–27, 2001.
12) L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
13) J. Shotton, M. Johnson, and R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2008.
14) C. Chang and C. Lin. LIBSVM: a library for support vector machines. Software available at `http://www.csie.ntu.edu.tw/libsvm`, 2001.
15) K. Grauman and T. Darrell. The pyramid match kernel: Discriminative classification with sets of image features. In *Proc. Int. Conf. on Computer Vision*, 2005.
16) J. Wu, and J.M. Rehg. Beyond the Euclidean distance: Creating effective visual codebooks using the histogram intersection kernel. In *Proc. Int. Conf. on Computer Vision*, 2009.
17) M. Johnson. Semantic Segmentation and Image Search. *Phd Thesis*, University of Cambridge, 2008
18) K. Barnard, K. Yanai, M. Johnson, and P. Gabbur. Cross Modal Disambiguation. In Toward Category-Level Object Recognition, *Lecture Notes in Computer Science*, vol.4170 pages 248-259. Springer Berlin/Heidelberg, 2006.
19) M. Johnson and R. Cipolla. Improved Image Annotation and Labelling Through Multi-Label Boosting. In *Proc. of British Machine Vision Conf.*, September 2005.
20) J. Zhang, M. Marszalek, S. Lazebnik, and C. Schmid. Local features and kernels for classificaiton of texture and object categories: A comprehensive study. *Int. Journal of Computer Vision*, 73(2):213–38, 2007.