Object and Feature-Based Scene Classification

Gilbert Kipkirui Langat*, Huang Dong Jun, Ammar Oad, Wilson Kipruto Cheruiyot.

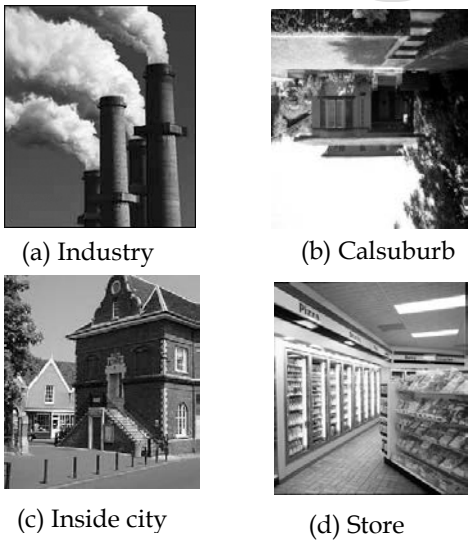
Abstract— Multimedia is a growing technology which nowadays is applied in our daily lives e.g. in entertainment, news bulletin, military, hospital and many areas of applications. Another aspect is the fast growing trends on machine learning, Nowadays machine are taking over from human and are becoming intelligent day in day out by the changing trends on how we do the retrieval and classification gave us a great urge to explore more on how to classify scenes in the field of object recognition.

We propose the use of support vector machine working together with DeCAF [1], this way we will be able to ease the complexity of dealing with big scene categorization. The process of achieving this was by, getting the features of the scene, then do various training by using LSVM (Linear support vector machine) we display our results by using MIT database with several images. Our approach shows it is better than the previous existing ones.

Index Terms— Scene classification, DeCAF, SVM, Image descriptor.

1 INTRODUCTION

Modern day scene classification has become not only important aspect of research but also necessary. But it has a lot of challenges and researchers are trying all they can to counter the challenges and come up with tangible solutions. Mostly these challenges come as a result of scene complexity and also the scene variability. For example, Fig.1 shows some samples from the MIT database, some of are not easy to be differentiated by human because of the great interclass differences.



(a) Industry (b) Calsuburb

(c) Inside city (d) Store

Fig 1.sample images from MIT dataset

A lot of the previous work in scene recognition has been made by designing low-level image features, such as SIFT [2], GIST [3] and CENTRIST [4]. Lazebnik et al. [5] proposed spatial pyramid matching for scene recognition based on SIFT dscriptor, which was better and efficient extension of bag of-words image representation. Oliva and Torralba [3] addressed this problem by GIST feature that could describe the structure characteristics of a scene. Wu et al. [4] proposed CENTRIST to represent the overall structure of scene images by census transform. Quattoni and Torralba [6] explored the problem of modeling scene layout by regions of interest. However, these low-level representations do not perform well for complex scenes due to the lack of scene semantic information. Li et al. [7] proposed an object bank representation to reveal the high-level semantic meanings of images for scene classification. Pandey and Lazebnik [8] regarded the scene as a whole object and applied deformable part models to scene categorization in a weakly supervised manner. Niu et al. [9] proposed a latent topic model for scene recognition by modeling the global spatial layout of different scene elements and the reinforcement of the visual coherence in uniform local regions. Sadeghi and Tappen [10] put forward a representation based on latent pyramidal regions, which can easily capture the discriminative characteristic of scenes. Since Krizhevsky et al. [11] constructed a large and deep convolutional neural network (CNN) on ImageNet, deep convolutional activation feature (DeCAF) [1] had also shown its ability on scene classification. But still, we still have a big challenge in the task of complex scene recognition. Considering that fact, the scene is composed of objects; people are perceived t to recognize scenes by analyzing objects in the said scene. In other words, one can distinguish different scenes by understanding content of the scene. However, current object-based representation methods

- *Gilbert Kipkirui Langat is a master's degree student in school of information science and engineering in central south university, china.*
- *Huang Dong Jun is a Professor in school of information science and engineering in central south university, china.*
- *Ammar Oad is a PhD student in school of information science and engineering in central south university, china.*
- *Wilson Kipruto cheruiyot is aprofessor school of computing Jomo Kenyatta University,kenya*

neglect the structure information of a scene image.

2 METHODOLOGY

The structure of our proposed scene representation is as illustrated in Fig. 2. By combining the scene models with original object detectors; our proposed scene representation becomes more good and effective on the task of vision recognition. The original object bank [7] vector has a high dimensionality of 44604; such a high dimensional vector is not convenient and conducive for classification due to noise and redundancy. To try and solve this issue, we employ sensible principle component analysis (SPCA) to process each segment of the response vector to get a more compact feature v1. Although object bank carries semantic meaning of scene content, it does not perform well on describing global structure and layout of a scene. To bridge this gap, we trained many scene deformable part models in a weakly supervised manner [8]. These models have a strong power on describing latent structure and texture of a scene. Then they are applied on the corresponding part of each image on different scales. Different from the spatial pyramid matching structure used in object bank, we use a max-pooling scheme on each response map of every scale to get the response vector v2.The two above vectors are concatenated to form our scene representation. Among them, v1 is aimed at describing objects in the scene, while v2 is intended to draw texture structure of the scene. Then many one-vs-rest linear SVM classifiers are trained based on the scene representation to predict the label of a scene.

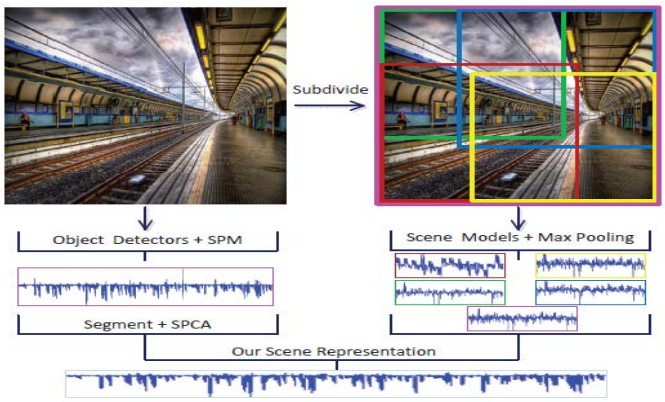


Fig. 2. Framework of our scene representation

2.1 Feature Extraction

We use the object bank method [7] to extract object responses for our scene representation. Object bank has a strong power on describing objects in the scene, where a scene is represented by a response vector of 44604 based on 177 pretrained object detectors. Due to that the above object bank representation is redundant; we can compress it and obtain a compact representation without losing much semantic information. In fact, there are many subspace selection methods [12, 13, 14, and 15] that can be used to alleviate this issue. As a simple dimensionality reduction method, SPCA [16] is a variant of principal component analysis which does define a proper probability model in the data space. So we use SPCA to deal with object bank. Therefore, it is shown by lots of experiments that the classification performance is poor when applying SPCA on the raw object bank vector. But if SPCA is employed upon responses of different detectors separately, it performs even better than the high dimensional data. It is to say, 177 SP-CAs are trained on responses segments of all detectors alone, then concatenating these shorter vectors to form the low dimensional feature. When compressing the object bank vector from 44604 to 4425 in this way, we get a little rise of 3% in the classification performance on MIT Indoor database.

2.2 Feature Extraction by Scene Models

How to discover and describe common visual structure in complicated and cluttered images is a key challenge in scene categorization. In this subsection, we will propose a novel representation to describe the visual structure of a scene.

2.3 Train Scene Deformable Part Models

Deformable part model (DPM) [17] is designed to detect and localize generic objects in a set of images. A DPM consists of a coarse root filter which covers an entire object, a set of part filters that cover parts of the object, and deformation parameters measuring the deviation of the parts from their desired positions relative to the root filter. An object detection hypothesis x is scored with the following function

$$f_\beta(x) = \max_{z \in Z(x)} \beta \cdot \Phi(x, z), \qquad (1)$$

Where β is the vector of model parameters, z are latent values that specify positions of parts relative to root. In order to describe the latent scene structure and the pivotal scene elements, we train DPMs for the scene with the weakly supervised method of [8]. The training process can be divided into two steps:

(1)

For each image I_i (i = 1, 2,..., n) in the original training set T_0, we subdivide it into four rectangular parts by retaining 3/4 of each side along every vertex.

Let I_i1, I_i2, I_i3, I_i4 represent the part containing the upper left, upper right, lower right, and lower left vertex individually. Then a series of images I_{1m}, I_2m, ..., I_{nm} form a new training set T_m

(m = 1, 2, 3, 4).

(2)

With each training set, one scene model is trained for each

class using images from other class as negative samples. Specially, the square root filter is restricted to have at least

40% overlap with the entire image. Each model has 2 components.

2.3.1 Reverse Spatial Pyramid Scheme

For us to compute the object vector from object detectors, they are evaluated at every location at multiple scales. Then these response maps are divided into grids (1 × 1, 2 × 2, 4× 4 grids) in a spatial pyramid manner, and the maximum score in each grid is retained to form the final representation. But for scene deformable part models, they are intended to represent a whole scene, so it is unreasonable to do like that and each component has 8 part filters, which has been proved to be the most suitable configuration for describing

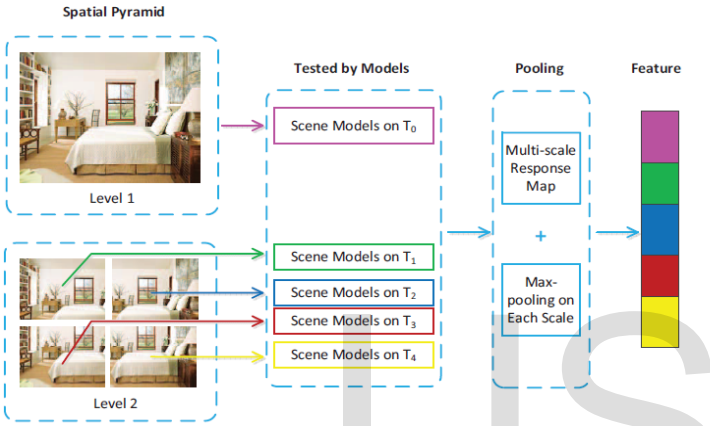


fig. 3 display of reverse spatial pyramid scheme

Latent scene structure. However, a"reverse spatial pyramid scheme" is proposed to settle this issue. A three level spatial pyramid is applied to object detectors, while a two level reverse spatial pyramid is employed for scene detectors, shown in the Fig. 3. Given an image I_j, for the first level, scene models trained on T_0 are tested on each location different scales on the entire image, and the maximum response on different scales is kept to form the vector for the first level. For the second level, the image is firstly split into four equal parts in a spatial pyramid manner. Then detectors trained on T_m are tested on the corresponding part I_{jm} (m = 1, 2, 3, 4), and the maximum response on different scales is kept to make up the final feature. Various features could describe different views of a scene, so multiview methods [18, 19] may be used for scene classification.

2.4 Classification by SVM

As shown in the Fig. 2 above, our proposed detector-based representation (denoted by DBR) is generated by combining object response vector v1 and scene response vector v2. Each response vector is normalized respectively, and then stacked together to give the final compact and powerful representation. The classification is done with one-vs-all linear SVMs which are trained to sperate each class from the rest classes. Given a test image, it is assigned with the label of the classifier with the highest response.

2.5 Combined with DeCAF

DeCAF [1] is the responses on the 6-th layer of CNN, which is an effective representation for scene classification. In order to utilize the complementary of DBR and DeCAF, we use a same simple method as [8] to combine their classifier scores. Specifically, each feature gives a set of one-vs-all classifiers for each of the n scene categories. If an image gets scores (s1... sn) from one of these classifier set, then the confidence that the image

$$\exp(s_i)/(\sum_{j=1}^{n}\exp(s_j)), \qquad (2)$$

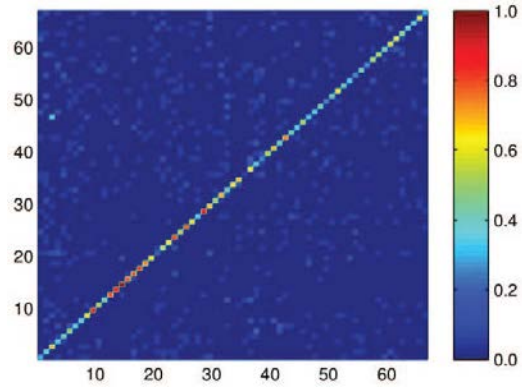


Fig. 4. The confusion matrix of all categories on MIT Indoor database

Fig.4. the confusion matrix of all categories on MIT Indoor database, classes are sorted in alphabetical order. The classification rates for each class are shown by color of the grid along the diagonal. The grid in the ith row and jth column shows the percentage of images from classes i which were misclassified as class j. To get the combined confidence for class i, we multiply the respective confidence of them, and assign the test image to the class which has the highest confidence value.

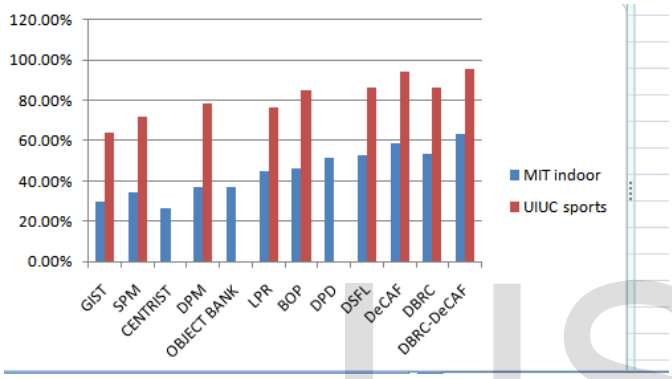
3 RESULTS

In this section, we will perform scene classification on the MIT Indoor database [6] and UIUC Sports database [20] to evaluate the effectiveness of our scene representation. As we all know, they are both universal benchmarks for the task of complex scene classification. Following the standard settings [6, 20]: on MIT Indoor, we use 80 training images and 20 testing images for each category; on UIUC Sports, we randomly selected 70 images per class as training images and 60 images per class as testing images.

3.1 DISCUSSION

To show the performance of our proposed scene representation DBR, we evaluate it by performing scene classification on the MIT Indoor and UIUC Sports database. For simplicity, we refer to the DBR-based classification method as DBRC, and the method combining DBR and DeCAF as DBRC DeCAF. Fig. 4 illustrates the confusion matrix which displays error rate of

mis-categoried images between the MIT scene categories. Confusion often occurs between classes cluttered by many small objects or classes without a general shape. Due to the effectiveness of numerous scene and object detectors, our DBRC performs obviously well on scenes with a common global layout (such as elevator, cloister, and corridor) and scenes composed by some typically objects (such as class room, waiting room, and inside subway). Despite our DBRC could achieve good performance on challenging scene classification datasets, it will take a long time to learn scene models on a new dataset. Fortunately, these scene models can be learned without disturbing each other. So multi-threaded parallel technology and distributed computer system can be used to reduce training time greatly. To further verify the effectiveness of DBRC, we compare it with 11 representative classification methods.



Graph 1. Shows the results of scene classification

databases by our DBRC and the 11 methods. It can be seen from the graph 1. that, we can see our proposed method outperform the Object Bank method by over 16% on MIT Indoor and 10% on UIUC Sports. This is mainly benefit from the effectiveness of these well-trained scene detectors which have been employed in a reverse spatial pyramid manner. Our method can achieve 53.58% accuracy on MIT Indoor and 86.25% on UIUC Sports which are superior with most of the popular approaches and comparable with schemes based on convolution neural networks. Considering DeCAF is powerful on describing the semantic information of a scene image, we combine it with our DBRC which can capture the structure information. In this way, we can achieve a satisfactory classification performance with an accuracy rate of 63.21% on MIT Indoor and 95.41% on UIUC Sports.

4. CONCLUSION AND FUTURE SCOPE

In our work we were able to find a detector-based for easy way of scene classification and it proved to be better than the previous methods in place. Our proposed method was able to capture general structure and objects within the scene. In this regard our contribution towards scene classification has been seen as reliable especially when dealing with more complicated and ambiguous scenes.our future work will be application of multiview methods to improve on the scene classification.

5. REFERENCES

[1] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang,E. Tzeng, and T. Darrell, "DeCAF: A deep convolutional activation feature for generic visual recognition," in Proc. of International Conf. Machine Learning, 2014, pp. 647–655.

[2] D. G. Lowe, "Distinctive image features from scale invariant key points," International journal of computer vision, vol. 60, no. 2, pp. 91–110, 2004.

[3] A. Oliva and A. Torralba, "Building the gist of a scene: The role of global image features in recognition," Progress in brain research, vol. 155, pp. 23–36, 2006.

[4] J.Wu and J. M. Rehg, "CENTRIST: A visual descriptor for scene categorization," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 33, no. 8, pp. 1489–1501, 2011.

[5] 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, vol. 2, pp. 2169–2178.

[6] A. Quattoni and A. Torralba, "Recognizing indoor scenes," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2009, pp. 413–420.

[7] L. J. Li, H. Su, F. F. Li, and E. P. Xing, "Object bank: A high-level image representation for scene classification & semantic feature sparsification," in Advances in Neural Information Processing Systems, 2010, pp. 1378–1386.

[8] M. Pandey and S. Lazebnik, "Scene recognition and weakly supervised object localization with deformable part-based models," in Proc. IEEE International Conf. on Computer Vision, 2011, pp. 1307–1314.

[9] Z. Niu, G. Hua, X. Gao, and Q. Tian, "Context awaretopic model for scene recognition," in Proc. IEEE Conf.Computer Vision and Pattern Recognition, 2012, pp. 2743–2750.

[10] F. Sadeghi and M. F. Tappen, "Latent pyramidal regions for recognizing scenes," in Proc. European Conf. Computer Vision, 2012, pp. 228–241.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolution neural networks," in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[12] D. Tao, X. Li, X. Wu, and S. Maybank, "Geometric mean for subspace selection," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 31, no. 2, pp.260–274, 2009.

[13] D. Tao, X. Tang, X. Li, and X. Wu, "Asymmetric bagging and random subspace for support vector machine based relevance feedback in image retrieval," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 28, no. 7, pp. 1088–1099, 2006.

[14]] D. Tao, X. Li, X. Wu, and S. Maybank, "General averaged divergence analysis," in Proc. IEEE International Conf. Data Mining, 2007, pp. 302–311.

[15]] D. Tao, X. Li, X. Wu, and S. Maybank, "General tensor discriminant analysis and gabor features for gait recognition," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 29, no. 10, pp. 1700–1715, 2007.

[16] S. Roweis, "EM algorithms for PCA and SPCA," in Advances in Neural Information Processing Systems, 1998, pp. 626–632.

[17] P. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 32, no. 9, pp. 1627– 1645, 2010.

[18] C. Xu, D. Tao, and C. Xu, "Large-margin multi-view information bottleneck," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 36, no. 8, pp. 1559–1572,2014.

[19] C. Xu, D. Tao, and C. Xu, "A survey on multi-view learning," Computing Research Repository, vol. abs/1304.5634, 2013.

[20] L. Li and F. Li, "What, where and who? Classifying events by scene and object recognition," in Proc. IEEE International Conf. on Computer Vision, 2007, pp. 1–8.

[21] M. Juneja, A. Vedaldi, C. Jawahar, and A. Zisserman, "Blocks that shout: Distinctive parts for scene classification," in Proc. IEEE Conf. Computer Vision and Pattern Recognition, 2013, pp. 923–930.