Bag-of-Words and Region-Based Feature Representations in Object Categorization: A Comparative Study

Chih-Fong Tsai[1,*], Ya-Han Hu[2], Ming-Chang Wang[3], Kang Ernest Liu[4]

[1]Department of Information Management, National Central University, Taoyuan, Taiwan.
[2]Department of Information Management, National Chung Cheng University, Chiayi, Taiwan.
[3]Department of Business Administration, National Chung Cheng University, Chiayi, Taiwan.
[4]Department of Agricultural Economics, National Taiwan University, Taipei, Taiwan.

* Corresponding author. Tel.: +886-34227151; email: cftsai@mgt.ncu.edu.tw
Manuscript submitted December 20, 2018; accepted January 25, 2019.
doi: 10.17706/jcp.14.2.93-100

Abstract: The aim of object categorization is to find a given object in an image and the performance of object categorization heavily depends on the extracted features as the image descriptor. In the literature, feature representation can be broadly classified into block/region-based and bag-of-words (BoW) features. However, there is no a comparative study of using these different feature representations over different datasets and different image scales since the image sizes for object recognition are varying from different datasets. Our experimental results using the Corel and PASCAL datasets show that when images contain more complex scenes like Corel images, the block-based feature is a better choice. In addition, the larger the image scales, the better the recognition performance. On the contrary, when images contain fewer objects like PASCAL images, it is better to consider the region-based feature representation. Particularly, reducing the image scale does not degrade the recognition performance; it even shows some level of improvement. On the other hand, although the BoW feature does not perform better than the block/region based features, it shows stable performances over different datasets and different image scales. This indicates that when the chosen dataset contains a large amount of images having various types of contents, which is difficult to decide what features to be extracted, the BoW feature can be extracted as the baseline feature representation.

Key words: Object categorization, feature representation, bag-of-words, segmentation.

1. Introduction

Object categorization has long been regarded as an important and challenging research problem in computer vision. In general, the aim of object categorization is to detect objects in images and determine the object's categories [1]. This task is similar to automatic image annotation (or tagging), where a system assigns keywords representing object categories to images automatically. It is usually used for image retrieval systems that allow users to perform keyword-based queries to search similar images [2], [3].

To achieve object categorization, images must be pre-processed to extract their visual features as the image descriptors. In other words, these feature descriptors are used to represent the image content for categorization. Moreover, the object categorization performance heavily relies on the image descriptors.

In the literature of object categorization and image annotation, there are two widely used feature representation methods, which are region-based and bag-of-words feature representations. In region-

based feature representation, each image is divided or segmented into a number of fixed size grids or object-based regions. Then, some low-level image features, such as color and texture, are then extracted from these regions.

On the other hand, bag-of-words (BoW) feature representation is usually based on tokenizing keypoint-based features, e.g. scale-invariant feature transform (SIFT) [4], to generate a visual-word vocabulary (or codebook). Then, the visual-word vector of an image contains the presence or absence information of each visual word in the image, e.g. the number of keypoints in the corresponding cluster, i.e. visual word.

Most object categorization studies proposing novel approaches are based on one specific feature representation, i.e. either the region-based feature or the BoW feature. For example, to name just a few, Lazebnik *et al.* [5] and Yang *et al*. [6] added spatial and/or contextual information in BoW for better categorization of objects.

On the other hand, for region-based feature representation, Li and Wang [3] introduced the two-dimensional multiresolution hidden Markov model, which summarizes clusters of grid-based feature vectors at multiple resolutions and the spatial relation between the clusters for image annotation. In Barnard *et al.* [2], a machine translation approach is used to learn the mapping between region types and keywords supplied with the images. Since these two types of feature representations are widely used in object categorization, very few compare region-based and keypoint-based BoW features in terms of object categorization (and image annotation). Moreover, since the image sizes for object recognition are varying from different datasets and studies, it is unknown which feature representation method performs better on which image scale.

In Douze *et al.* [7], one very similar work to this paper, the global based GIST feature, GIST with spatial grids, BoW, and BoW with spatial grids were compared in terms of memory usage and retrieval precision over web images. Since local representations can obtain relatively better results for object recognition than global ones, they only considered block-based segmentation and image representation. Moreover, they did not examine different local feature representations over different image scales. Therefore, the contribution of this paper is the first attempt to assess the object categorization performances of using region-based (including blocks) and BoW features over different image scales.

The rest of this paper is organized as follows. Sections 2 and 3 overview the region-based and BoW feature representations respectively. Section 4 presents the experimental setup and results to compare these two types of feature representations over different image scales. Finally, conclusion is provided in Section 5.

2. Region-Based Feature Representation

Before extracting local region features of images, image segmentation must be performed in order to segment an image into numbers of regions. In general, there are two approaches to segment an image into local contents, which are block- and object-based segmentation.

(a) 5-block (b) 5-region

Fig. 1. Image segmentation.

The first and the simplest approach is to divide an image into a fixed number of (non) overlapping grids or blocks whose size and shape can be either the same or different. Fig. 1(a) shows an example of the tiling scheme where four sub-images represent four quadrants of the image and one represent the center sub-image [8]. Some other block-based segmentation methods are based on a fixed pixel resolution, say 5×5, as a grid where an image is composed of a fixed number of grids. Please refer to Tsai and Hung [9] for a survey of different tiling schemes.

The second approach for image segmentation is to divide the image into homogenous objects/regions (such as horse and car) using some region segmentation algorithms. Fig. 1(b) shows an example of using the Normalized Cuts algorithm [10] to perform 5-region based segmentation.

However, it is very difficult to achieve accurate region-based segmentation in an image, especially when images contain less distinctive objects [11] and the resulting homogenous regions may not represent meaningful objects of interest. In other words, under- and/or over-segmentation usually occurs in broad domains of general images.

On the other hand, although a simple partition (i.e. the block-based approach) does not generate perceptually meaningful regions or objects, it is a computationally efficient way of representing the global features of the image at a finer resolution [12].

In the literature, these two types of segmentation methods have both been used for the object categorization task. For example, Tsai *et al.* [8]; Boutell *et al.* [13]; and Wang *et al.* [14] used block-based segmentation, and Akbas and Yarman Vural [15]; Fan *et al.* [16] used region-based segmentation.

After an image is segmented into local contents, some visual image features can be extracted from each of the sub-images, such color and texture. Since the objects in images are vary, there is no general agreement about what kinds of features are better to represent different image contents. However, color and texture are the two most extracted features from local blocks or regions for object categorization [9].

3. Bag-of-Words Feature Representation

The bag-of-words (BoW) model is a well-known and popular method for document representation in information retrieval. It is based on a dictionary, and each document containing some words from the dictionary is regarded as a bag. In particular, the word orders are treated the same under this model. The first work that applied the BoW model to the field of image and video retrieval was conducted by Sivic and Zisserman [17].

In the BoW feature representation approach, it creates a corpus of image (region) features. As a result, *term-document-matrix* of document retrieval becomes *feature-image-matrix* of image retrieval, which means that this visual-word representation is analogous to the bag-of-words representation of text documents. In image annotation, it can be embedded in such a vector space representation, so that annotated images are modeled by concatenated feature vectors of word (i.e. keyword) and image features (i.e. visual features).

Briefly speaking, images are represented by sets of low-level features (e.g. keypoint descriptors). Particularly, sparse and dense features are two possibilities of defining interest points and scales for feature extraction. In sparse features, interest points are detected at local extremas in the difference of Gaussian pyramid. Then, a position and scale are automatically assigned to each point. On the other hand, in dense features, interest points are defined at evenly sampled grid points. Feature vectors are then computed based on three different neighborhood sizes, i.e. at different scales, around each interest point.

Next, vector quantization (VQ) is performed to cluster the feature descriptors into a large number of clusters (usually based on the *k*-means clustering algorithm), and then encode each keypoint by the index

of the cluster to which it belongs. In this case, each cluster is regarded as a visual word (or codeword) that represents a specific local pattern shared by the keypoints in that the number of clusters is the size of the vocabulary.

In the BoW model, it can be defined as follows. Given a training dataset D containing n images represented by $D = d_1, d_2, ...,$ and d_n where d is the extracted visual features. A specific unsupervised learning algorithm, such as k-means, is used to group D based on a fixed number of visual words W (or categories) represented by $W = w_1, w_2, ...,$ and w_V where V is the cluster number. Then, we can summarize the data in a $V \times N$ co-occurrence table of counts $N_{ij} = n(w_i, d_j)$, where $n(w_i, d_j)$ denotes how often the word w_i occurred in an image d_i.

4. Experiments

4.1. Experimental Setup

Two related datasets are chosen for the experiments. The first one is based on the Corel dataset used in Duygulu *et al.* [18][1]. It is composed of 4500 training and 500 testing images. In addition, each image belongs to a specific category and each category contains 100 images. Therefore, there are 50 categories in this dataset.

The second dataset is based on PASCAL VOC (Visual Object Classes) Challenge 2008[2], which contains 20 object categories. There are 2111 and 2221 images for training and validation respectively, which are used for classifier training and testing.

We used block and region based feature extraction methods to represent local image features respectively. In particular, each image is resized into five different resolutions, which are 256×256, 128×128, 64×64, 32×32, and 16×16 respectively. Next, for a specific image scale, each image was segmented into 5 blocks [8] and 5 regions respectively. For region based segmentation, the Normalized Cuts algorithm [10] is used to segment each image into 5 regions.

Since color and texture features are the two widely extracted features from segmented regions [9], the HSV color and three levels of Daubechies-4 wavelet texture features are extracted from each segmented blocks or regions. Finally, each block or region is represented by the mean value of these low-level features [8]. As a result, each block or region is represented by 19 dimensional visual features. Therefore, 95 dimensional features (i.e. 19 low-level features $\times$ 5 blocks/regions) are used as the image descriptors for image classification.

On the other hand, to extract the BoW feature from these two datasets, the dense based BoW feature was extracted for image representation since many related studies have shown its superiority over the sparse based BoW feature, such as Bosch *et al.* [19]. Therefore, each image was first of all segmented into 16×16 grids over the 256×256 image scale. The keypoints are detected by difference of Gaussian (DoG) detector and described by SIFT descriptor[3] [4]. Note that other descriptors, such as GIST and SURF are not compared since many above motioned related studies use SIFT as the baseline descriptor for object categorization. On average, there are about 700 to 1500 keypoints detected from each image.

Next, the k-means clustering algorithm was used to generate the visual vocabularies. In addition, the vocabulary size was set to 500. Consequently, each image is represented by 500 BoW features. More

[1] http://kobus.ca/research/data/eccv_2002/index.html
[2] http://pascallin.ecs.soton.ac.uk/challenges/VOC/voc2008/
[3] http://www.cs.ubc.ca/~lowe/keypoints/

specifically, the term weighting schemes are based on TF (term frequency) and TF-IDF (inverse document frequency) for comparisons. However, since there are four smaller image scales, the vocabulary sizes were set differently. That is, they are set to 250, 125, 62, and 31 for 128×128, 64×64, 32×32, and 16×16 image resolutions respectively.

Moreover, each image was segmented into 8×8, 4×4, 2×2 grids for 128×128, 64×64, and 32×32 image resolutions respectively. Note that there is no segmentation over the 16×16 image scale.

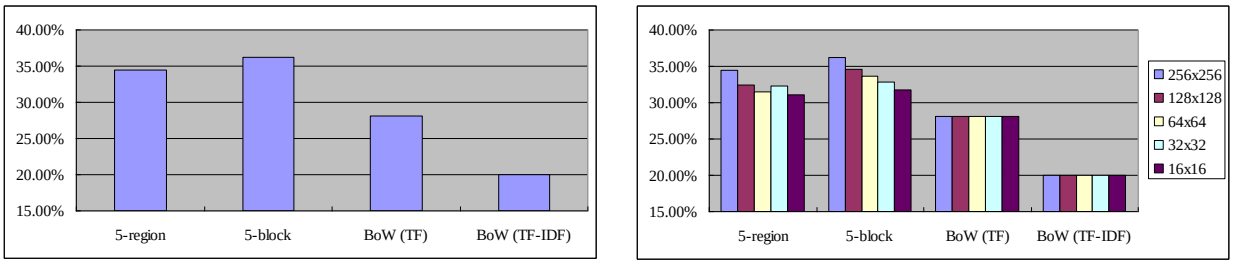
It should be noted that we use these two feature representation methods for the comparison because they are the most widely used baseline methods in region-based and BoW feature representation methods.

To evaluate the classification performances of different feature representations, support vector machines (SVM) are constructed and compared. In addition, radial basis function (RBF) is used as the kernel function, and several different gamma values ('γ') are examined in order to obtain the best SVM, which provides the best classification performance. In particular, we set γ to 0, 0.1, 0.5, 1, 2, and 2.5 for comparisons.

4.2. Results on the Corel Dataset

Fig. 2(a) and 2(b) show the classification accuracy by block/region-based and BoW feature representations over the 256×256 and different scales, respectively. Fig. 2(a) indicates that 5-block based feature representation performs best. On the other hand, the BoW feature representations by TF and TF-IDF cannot make SVM provide comparable performance with 5-block/region-based feature representations.

On the other hand, similar to Fig. 2(a), the 5-block based feature representation performs best no matter what image scale is used, but the performance degrades when the image scale becomes smaller. It is interesting that the BoW features using TF or TF-IDF make SVM perform almost the same over different image scales, i.e. 28.04% to 28.06% by TF and 20.04% by TF-IDF.



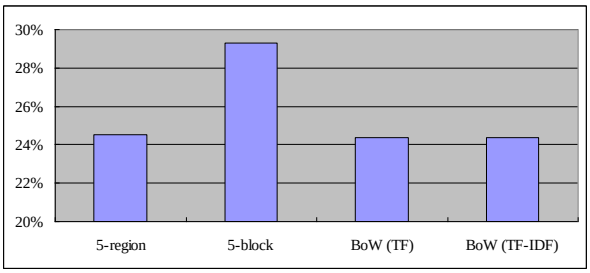
(a) 256×256 scale (b) different scales

Fig. 2. Classification accuracy over corel.

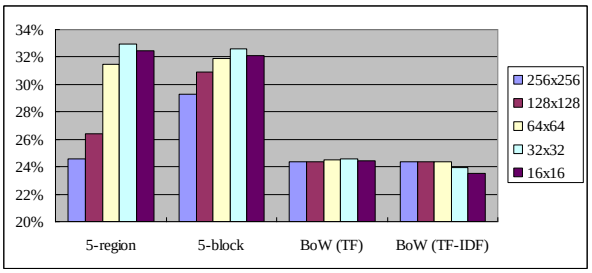
4.3. Results on the PASCAL Dataset

Fig. 3(a) and 3(b) show the classification accuracy of block/region-based and BoW feature representations over the 256×256 and different scales, respectively. Similar to the Corel dataset, for the the 256×256 scale, the 5-block based feature representation performs best. However, the performances of SVM using the 5-region based and BoW features are not significantly different.

For the results of different scales, interestingly, the classification accuracy increases when the image scale reduces by the 5-block/region based features. More specifically, the image scale by 32×32 is the most suitable resolution for region-based feature extraction, which makes SVM provide 32.91% accuracy. On the other hand, for the BoW feature the effect of using different image scales on classification accuracy is very small, which is the same as using the Corel dataset. In particular, SVM provides around 24.36% to 24.56% by TF and 23.55% to 24.38% by TF-IDF. This shows the stable characteristic of the BoW feature representation for different domains of images.



(b) 256×256 scale



(b) different scales

Fig. 3. Classification accuracy over PASCAL.

When using the block/region based features, we obtain different results over the two different datasets. That is, the block based feature extracted from the 256×256 scale of Corel images can allow SVM to provide the highest rate of accuracy, i.e. 36.2%, but the region based feature extracted from 32×32 PASCAL images makes SVM provide the highest rate of accuracy, i.e. 32.9%. This may be because PASCAL images are used for object recognition, which usually contain only one instance of an object per image. Therefore, a relative small image scale can still provide a good level of object information. However, Corel images usually contain more complex scenes leading to the more suitability of using the block based feature representation for this dataset and reducing the image scale will results in less scenery information.

5. Conclusion

In this paper, we examine the object categorization performances by using block/region-based and bag-of-words (BoW) feature representations over different image datasets and image scales. The experimental results indicate that the block based image feature is a better representation method for object recognition when images contain complex scenes. In this feature representation and scenery images, a larger image resolution can possess complete scenery information of images, which leads to better recognition performance. For the region based feature, it is good at recognizing objects when images contain fewer objects. Furthermore, reducing the image scale will not degrade the recognition performance; using the region based feature representation even shows some level of improvement. On the other hand, although the BoW feature does not perform better than the block/region based features, it shows very stable performances over different datasets and different image scales.

Therefore, when the image dataset contain a large amount of images having various types of contents that are difficult to determine which feature representation should be used; the BoW feature is a good choice for object categorization. In addition, using color SIFT descriptors [20] could further improve the categorization performance. The experimental results also imply that the future work for better object recognition could be based on the combination of block/region based and BoW features in a more sophisticated way.

References

[1] Pinz, A. (2005). *Object Categorization Foundations and Trends in Computer Graphics and Vision,* 255-353.

[2] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research, 3,* 1107-35.

[3] Li, J., & Wang, J. Z. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25,* 1075-1088.

[4] Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision, 60,* 91-110.

[5] Lazebnik, S., Schmid, C., & Ponce, J. (2006). Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition,* 2169-2178.

[6] Yang, L., Zheng, N., & Yang, J. (2011). A unified context assessing model for object categorization. *Computer Vision and Image Understanding, 115,* 310-322

[7] Douze, M., Jegou, H., Singh, H., Amsaleg, L., & Schmid, C. (2009). Evaluation of GIST descriptors for web-scale image search. *Proceedings of the International Conference on Image and Video Retrieval* (pp. 1-8).

[8] Tsai, C. F., McGarry, K & Tait, J. (2006). CLAIRE: A modular support vector image indexing and classification system. *ACM Transactions on Information Systems, 24,* 353-279.

[9] Tsai, C. F., & Hung C. (2008). Automatically annotating images with keywords: A review of image annotation systems. *Recent Patents on Computer Science, 1,* 55-68.

[10] Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 22,* 888-905.

[11] Smith, J. R., & Chang, S. F. (1996). Visualseek: A fully automated content-based image query system. *Proceedings of ACM Conference on Multimedia*, 87-98.

[12] Long, F., Zhang, H., & Feng, D. (2003). Fundamentals of content-based image retrieval. In D. D. Feng, W. C. Siu, & H. Zhang (Eds.), *Multimedia Information Retrieval and Management – Technological Fundamentals and Applications* Springer-Verlag, Germany.

[13] Boutell, M., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition, 37,* 1757-1771.

[14] Wang, C., Blei, D., & Fei-Fei, L. (2009). Simultaneous image classification and annotation. *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition* (pp. 1903-1910).

[15] Akbas, E., & Vural, F. T. Y. (2007). Automatic image annotation by ensemble of visual descriptors. (pp. 1-8).

[16] Fan, J., Gao, Y., Luo, H., & Jain, R. (2008). Mining multilevel image semantics via hierarchical classification. *IEEE Transactions on Multimedia, 10,* 167-187.

[17] Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. *Proceedings of International Conference on Computer Vision* (pp. 1470-1477).

[18] Duygulu, P., Barnard, K., Freitas, N., & Forsyth, D. (2002). Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. *Proceedings of European Conference on Computer Vision* (pp. 97-112).

[19] Bosch, A., Zisserman, A., & Munoz, X. (2006). Scene classification via pLSA. *Proceedings of European Conference on Computer Vision* (pp. 517-530).

[20] Burghouts, G. J., & Geusebroek, J. M. (2009). Performance evaluation of local colour invariants. *Computer Vision and Image Understanding, 113,* 48-62.

Chih-Fong Tsai received a Ph.D at School of Computing and Technology from the University of Sunderland, UK in 2005. He is now a professor at the Department of Information Management, National Central University, Taiwan. His current research focuses on data mining and machine learning. He has published more than 70 professional publications where some were published in prestigious journals including: ACM transactions on information systems, ACM transactions on management information systems, decision support systems, European Journal of Operational Research, IEEE transactions on systems, man and cybernetics – Part C: Applications and reviews, information processing & management, information systems, Journal of Systems and Software, Journal of the Association for Information Science and Technology, and pattern recognition. He received the highly commended award (emerald literati network 2008 awards for

excellence) from online information review ("A Review of Image Retrieval Methods for Digital Cultural Heritage Resources") and the award for top 10 cited articles in 2008 from expert systems with applications ("Using Neural Network Ensembles for Bankruptcy Prediction and Credit Scoring").

Ya-Han Hu is currently a professor of Department of Information Management at National Chung Cheng University, Taiwan. He received a Ph.D degree in information management from National Central University of Taiwan in 2007. His current research interests include text mining and information retrieval, clinical decision support systems, and recommender systems. His research has appeared in information & management, decision support systems, Journal of the American Society for information science and technology, IEEE transactions on systems, man, and cybernetics, International Journal of Information Management, artificial intelligence in medicine, applied soft computing, computers in human behavior, data & knowledge engineering, expert systems, knowledge-based systems, information systems and e-business management, Journal of Information Science, Journal of Clinical Epidemiology, methods of information in medicine, online information review, and Journal of Systems and Software.

Ming-Chang Wang is currently a professor of Department of Business Administration at National Chung Cheng University, Taiwan. He received a Ph.D degree in Department of Finance from National Sun Yat-sen University in 2007. His current research interests include financial technology, big-data investment, market microstructure and corporate governance. His research has appeared in European Financial Management, The quarterly review of economics & finance, international review of economics and finance and Asia-Pacific Journal of Financial Studies.

Kang Ernest Liu is currently a professor of Department of Agricultural Economics at National Taiwan University, Taiwan. He received a Ph.D degree in agricultural, environmental, and development economics from the Ohio State University in 2003. His current research interests include consumption economics, experimental economics, tourism economics, data mining, and applied microeconometrics. His research has appeared in Japan and the world economy, Singapore economic review, small business economics, China agricultural economic review, agricultural economics, and applied economics.