Feature Fusion for Efficient Object Classification Using Deep and Shallow Learning

T. Janani and A. Ramanan

***Abstract*—Bag-of-Features (BoF) approach have been successfully applied to visual object classification tasks. Recently, convolutional neural networks (CNNs) demonstrated excellent performance on object classification problems. In this paper we propose to construct a new feature set by processing CNN activations from convolutional layers fused with the traditional BoF representation for efficient object classification using SVMs. The dimension of convolutional features were reduced using PCA technique and the bag-of-features representation was reduced by tailoring the visual codebook using a statistical codeword selection method, in order to obtain a compact representation of the new feature set which achieves increased classification rate while requiring less storage. The proposed framework, based on the new features, outperforms other state-of-the-art approaches that have been evaluated on benchmark datasets: Xerox7, UIUC Texture, and Caltech-101.**

***Index Terms*—Object classification, bag-of-features, convolutional neural network, deep learning, shallow learning.**

I. Introduction

Visual object recognition is a process of identifying a specific object in a digital image or video that has been contributing a major aspect in computer vision and machine learning. In object recognition tasks, deep learning has started to have huge impact on the emergence of large-scale training images in which deep neural networks (e.g., CNNs) show significant advantages compared with shallow models (e.g., SVMs) due to their large learning capacity. Shallow methods start by extracting a representation of the image using handcrafted local image descriptors (e.g., SIFT) and then aggregates such local descriptors into an overall image descriptor by using a pooling mechanism. The emphasis in shallow learning is often on feature engineering and selection while in deep learning the emphasis is on defining the most useful computational graph topology and optimizing hyper parameters correctly.

Bag-of-features (BoF) approach [1] which is a shallow method is a popular image representation scheme used in patch-based visual object recognition over a decade [2]. The underlying idea of such approach is that, in different images, the statistical distribution of the patches is different, which can be effectively exploited for recognition. Recently, convolutional neural networks (CNNs) [3] being as a deep learning method demonstrated excellent performance on large scale visual image classification tasks such as PASCAL visual object classes (VOC) [4] and ImageNet [5] challenges. CNN is a biologically inspired multi-stage architecture composed of convolutional, pooling and fully-connected layers, having very large numbers of parameters that must be efficiently learned from training examples. The way of feature extraction and generalization from simple to complex is much more similar to the processing in the human visual system. The top layers in CNN are sensitive to semantics which is especially critical when objects are occluded or truncated, whereas the intermediate layers are specific to low-level patterns. A deeper network requires a large amount of training data to avoid over-fitting. Deeper networks learn much richer level discriminative features. On the other hand, spatial information is diluted with the depth of the network due to pooling layers of CNNs. In this paper we successfully share the aspects of deep and shallow methods in image representation. In particular, we show that the data augmentation techniques commonly applied to CNN-based methods can also be applied to BoF methods, and result in an analogous performance boost.

The contributions of this work are two-fold. First, we conduct systematic experiments to evaluate the relative performance of the bag-of-features representation with support vector machines (SVMs) [6] and convolutional neural networks with back-propagation neural networks in visual object classification. Second, we improve the overall classification rate of the BoF approach by fusing the convolutional features that are classified using SVMs. Our framework, based on the new features, significantly improves the traditional BoF approach even with a relatively low-dimensional representation.

The remainder of this paper is organised as follows. Section II summarises the standard bag-of-features approach and CNN approach in object classification. Section III summarises various related work that have been used in object recognition. Section IV presents the proposed framework in fusing the BoF representation and convolutional features for efficient object classification. Section V, provides the experimental setup and testing results of the proposed method. Finally Section VI concludes this paper.

II. Background

The construction of the bag-of-features vector h from an image I can be summarised in four steps: (i) keypoints are automatically detected in I using a local invariant detector (e.g., SIFT), (ii) local descriptors pooled from a training set are computed over those regions, (iii) all the descriptors are quantised into codewords by means of a clustering method

Manuscript received August, 2017; revised October 5, 2017.

The authors are with the Department of Computer Science, Faculty of Science, University of Jaffna, Jaffna, Sri Lanka (e-mail: janani22thangavel@ gmail.com, a.ramanan@jfn.ac.lk).

doi: 10.18178/ijmlc.2017.7.5.633

(e.g., K-means), and (iv) counted to construct the BoF representation of the image *I*.

A multiclass classifier performs two separate steps in order to predict the classes of unlabeled images. During training, labeled images using fixed-length feature vector representation, i.e., the BoF are sent to the classifier and used to adapt a statistical decision procedure for distinguishing categories. In testing, the classifier determines which category or categories to assign to the image from its bags-of-features representation.

In CNN, receptive fields slide over an input image and those receptive fields are connected with one unit in the next layer, which yields a feature map. When this feature mapping is over, a convolutional layer is constructed. The idea is that if a feature detector is useful in one part of the image it is likely that it is useful somewhere else, but at the same time it allows each patch of the image to be represented in several ways. Next a pooling layer is used to reduce the neighbouring features from the feature map into single units. This process is repeated for many rounds and eventually arrives to almost scale invariant representation of the input image. This is a very powerful representation since objects can be detected in an image without concerning the spatial information of an object.

III. Related Work

Csurka *et al.* [1] used the Harris affine region detector to identify the keypoints in the images which are then described by SIFT descriptors. A visual codebook was constructed by clustering the extracted features using K-means algorithm. Images are then described by histograms over the learnt codebook. The proposed framework was mainly evaluated on Xerox7 image set. The overall classification rate is reported to be 85% using linear SVMs. A detailed review of the methods for object recognition using the bag-of-features representation can be found in [2].

Chatfield *et al.* [7] proposed different deep architectures: CNN-F, CNN-M and CNN-S, along with a comparison with previous state-of-the-art shallow representations such as the bag-of-features and the improved Fisher vector. The authors have pre-trained the CNN architectures on ILSVRC-2012 dataset (i.e., ImageNet). The CNN-S framework was fine-tuned using PASCAL VOC-2007, VOC-2012, or Caltech-101 as the target data. It has been suggested that CNN learnt on a large dataset such as ImageNet, can be used as a powerful image descriptor applicable to other datasets. In addition, the authors claim that the dimensionality of the CNN output layer can be reduced significantly without having an adverse effect on performance.

Razavian *et al.* [8] conducted a series of experiments for different recognition tasks using the publicly available code and model of the OverFeat [9] network which was trained to perform object classification on ILSVRC13. Authors have used features extracted from the OverFeat network as a generic image representation to tackle the diverse range of recognition tasks. The feature representations are further modified using simple augmentation techniques e.g., jittering. The experimental results are achieved using a linear SVM classifier applied to a feature representation of size 4096 extracted from a layer in the net. The authors strongly suggest that features obtained from deep learning with convolutional nets should be the primary candidate in most visual recognition tasks.

Ng *et al.* [10] have conducted thorough experiments to investigate the performance of features from different layers of CNN and different scales of input test images in instance-level image retrieval. The authors have presented an approach for extracting convolutional features from different layers of the networks, and adopt VLAD encoding to encode features into a single vector for each image. The framework outperforms other VLAD and CNN based approaches. With VLAD encoding on convolutional response, the authors achieve state-of-the-art retrieval results using low dimensional representations on two of the instance image retrieval datasets: OxfordNet and GoogLeNet. They conclude that the intermediate layers of CNN with finer scales produce better results for image retrieval compared to the last layer.

Zheng *et al.* [11] used effective transfer of the convolutional neural network feature in image search and classification. The authors have demonstrated the advantage of using images with a properly large size as input to CNN instead of the conventionally resized one (i.e., 224×224). The authors benchmark the performance of different CNN layers improved by average/max pooling on the feature maps. It has been observed that the Conv5 feature yields better accuracy under such pooling step. They find that the simple combination of pooled features extracted across various CNN layers is effective in collecting evidences from both low and high level descriptors. Following the said techniques, they are capable of improving the state-of-the-art method on a number of benchmark imagesets to a large margin. The datasets are: Holidays, Ukbench, and Oxford5k.The authors have also tested their approach for generic recognition as the case in Caltech-101 and PASCAL VOC-2007 in which the VGGNet is trained on generic ILSVRC classification. The testing results for Conv5 with average/max pooling was reported to be 91.07 and 81.78 on the Caltech and PASCAL VOC-2007 image sets, respectively.

IV. Methodology

Due to the local sensing and weight-sharing of CNN, the convolution layers often contain rich local features and these local features can be combined to the bag-of-features in object classification. The overall framework of the proposed approach is depicted in Fig. 1.

A. Bag-of-Features

SIFT [12] is a method that extracts distinctive features from gray-value images, by filtering images at multiple scales and patches of interest that have sharp changes in local image intensities. It turns images into a collection of local feature vectors which are invariant to translation, scaling, and rotation. The features are located at maxima and minima of a difference of Gaussian functions applied in scale space. Next, the descriptors are computed based on eight orientation histograms at a 4×4 sub-region around the interest point, resulting in a 128 dimensional vector. A visual codebook is then constructed by a vector quantisation technique that groups similar keypoint descriptors together that are detected in training images. Each group is represented by the learnt cluster centres referred to as ‘visual words’ or ‘codewords’. Each group is represented by the learnt cluster centres referred to as ‘visual words’ or ‘codewords’. The size of the codebook is the number of clusters obtained from the clustering technique. Each interest keypoint of an image in the dataset is then quantised to its closest codeword in the codebook, such that it maps the entire patches of an image in to a fixed-length feature vector of frequency histograms known as bag-of-features representation, i.e., the visual codebook model treats an image as a distribution of local features.

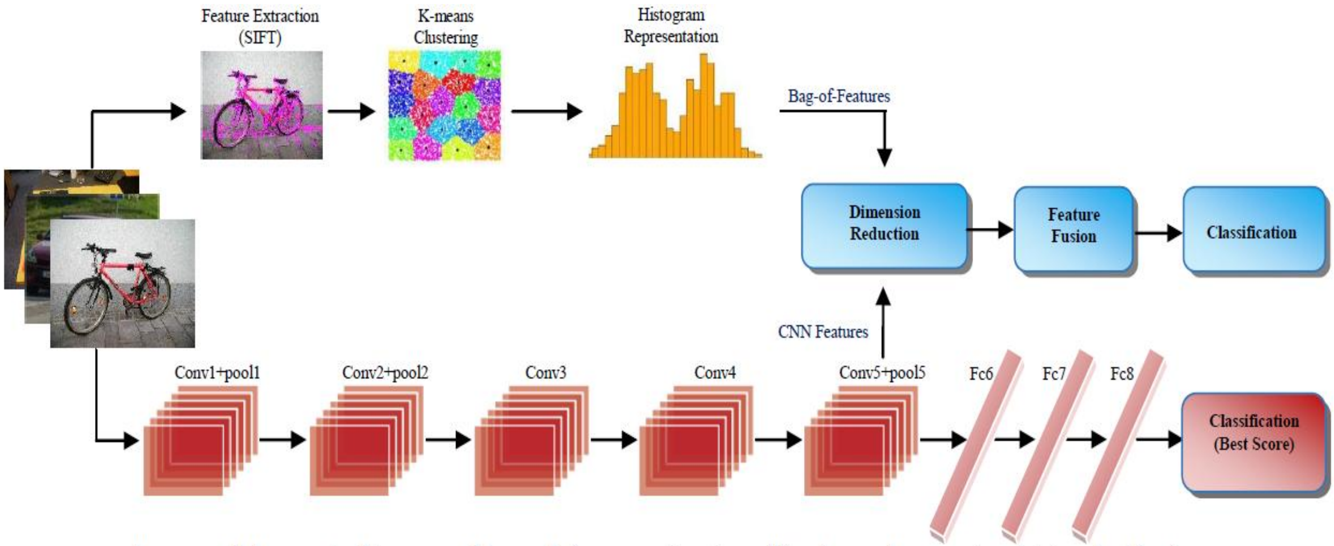

Fig. 1. Overall framework of the proposed feature fusion approach to the traditional Bag-of-Features based object classification.

B. CNN Features

Recently, SIFT-based approaches have been substantially outperformed by the introduction of the latest generation of CNNs in object classification. Recent works in object classification use the last hidden layer of the deep CNN which ignores the local features extracted by the lower layers. In this regard, limited experiments were carried out in this paper to test the relative performance of CNN features that are obtained from intermediate layers up to the fully connected layers and it has been observed that the pooled Conv5 feature, with much lower dimensionality, shown to yield increased accuracy. It was demonstrated in [13] that fine-tuning a pre-trained CNN on the target data can significantly improve the performance. In this work the fine-tuned CNN at the pooled Conv5 leads to a $6\times6\times512$ dimensional image representation. This representation was further reduced around ten-fold using PCA method without loss of performance in classification.

C. Feature Fusion and Compact Representation

Weighting method and connecting method are two commonly used feature fusion methods in machine learning. Assigning the weight value for each feature is the crucial part as different combinations of weights will result in different outcomes. It will be a problem to determine the optimal weight combination among the various combinations. In this way, connecting method is used in this paper for feature fusion. In the training stage,

i) The fine-tuned CNN is used to extract the pooled Conv5 image features from the training set. Then principal components analysis (PCA) is performed to reduce the dimension of the features extracted by the fine-tuned CNN. PCA is a linear dimensionality reduction technique that transforms a number of correlated variables into un-correlated variables called principal components. The goal of principal component analysis is to embed the data into a linear subspace of lower dimensionality describing as much of the variance in the data as possible.

ii) For constructing category-specific visual codebooks, SIFT features were extracted from each of the training set of a category and clustered independently using K-means algorithm. The category-specific codebooks were then concatenated into a global codebook. The codewords of the learnt global codebook then serves to construct a histogram for representing an image. Then a statistical codeword selection technique based on within-category confidences is used to reduce the dimension of the bag-of-features representation. The visual codebook constructed using K-means algorithm often contains lot of noisy and/or ambiguous codewords. Different category images can have similar kind of codewords and categorization should be based on histogram values of codewords. A high variance histogram value of a codeword interrupts the categorisation process, i.e., it makes the classifier difficult to classify visual object categories. In order to build a compact and discriminant codebook from an initially build larger codebook we reformulate it using within-category confidence. Within-category confidence is calculated by analysing a within-category variance of i^{th} codeword. The within-category confidence of the i^{th} codeword is represented as follows:

$$C_{\text{within},i} = \frac{1}{\sum_{j=1}^{n} \text{var}(h_{ij})}$$

where h_{ij} is the i^{th} codeword value of each image belonging to the j^{th} object in the BoF histogram domain, var(h_{ij}) is a variance of the h_{ij} and n is the number of object categories.

In the testing stage, the testing images will also go through feature extraction and dimensionality reduction to get the ultimate feature. Then, these features: CNN and BoF are fused using the connecting method and put into the trained SVM classifier to generate the object class label of the testing images.

D. Classifier

Support vector machine has proven to be successful in solving many image categorization problems. In this work multiclass classification is performed with a linear SVM trained using the one-versus-all (OVA) rule. OVA-SVM is learned to separate each class from the rest, and a test image is assigned the label of the classifier with the highest response. The trade-off regulariser and loss parameter C in linear SVM is determined from a range of values [2^{-14}, 2^{-13}, ..., 2^{1}, 2^{2}] using a separate validation subset of the data. We used the $\text{SVM}^{\text{light}}$ package [6] in our experiments.

We use the VGG-M implementation for the CNN that comprises eight learnable layers, five of which are convolutional and the last three are fully-connected. The input image size is 224×224. Fine-tuning of CNN was carried out using the VGG-M framework that was trained on ILSVRC.

E. Evaluation Criteria

Mean Average Precision (mAP) is taken as the evaluation metric for all experiments conducted in this paper. For obtaining unbiased precision-recall curve estimates when the classes are imbalanced, we compute the classification output based on binormal model [14].

V. Experimental Setup and Testing Results

A. Dataset

1) Xerox7: The dataset [1] contains 1776 images from seven categories: Faces, buildings, trees, cars, phones, bikes and books. The object poses are highly variable and there is a significant amount of background clutter, some of which belongs to the other categories making the classification task fairly challenging.
2) UIUCTex: The dataset [15] contains 25 texture classes with 40 images per class. It has surfaces whose texture is mainly due to albedo variations (e.g. wood and marble), 3D shape (e.g. gravel and fur), as well as a mixture of both (e.g. carpet and brick). It also has significant viewpoint changes, uncontrolled illumination, arbitrary rotations, and scale differences within each class.
3) Caltech-101: The dataset [16] consists of a total of 9,146 images, split between 101 different object categories, as well as an additional background/clutter category. Each object category contains between 31 and 800 images. Common and popular categories such as faces tend to have a large number of images than others.

The experimental setup was such that, for all the image sets listed above, we used 70% for training and 30% for testing.

B. Testing Results

The testing results presented in Table I, shows that fusing CNN features obtained at pooled Conv5 with BoF features significantly improves the classification rate except the smallest improvement comes from UIUCTex. It is because the recognition accuracy on the datasets is already high enough (98.86%). All the results presented in Table I is without reducing the dimension of the features that were extracted by the BoF and pre-trained CNN.

The testing results presented in Table II, shows the classification rate on the fused features after performing dimensionality reduction technique as mentioned in section IV(C). The K of K-means in constructing category-specific codebooks for the BoF approach is fixed to the values 20, 40, and 100 for the Caltech-101, UIUCTex, and Xerox7 datasets, respectively. The BoF approach leads to an $N \times K$ dimensional image representation, where N is the number of categories in each of the datasets.

TABLE I: Classification Rate of the Proposed Method Compared to the Standard Bag-of-Features Approach without Dimensionality Reduction of Features

Image Set	**BoF**	**BoF+CNN**
Xerox7	78.13	90.28
UIUCTex	98.86	99.17
Caltech-101	76.24	97.65

TABLE II: Classification Rate of the Proposed Method with Dimensionality Reduction of Features

Image Set	**Dimension**					**Rate**
	BoF	**C_{within}**	**CNN**	**PCA**	**Total**	
Xerox7	700	**506**	18432	**1242**	**1748**	91.67
UIUCTex	1000	**852**	18432	**699**	**1551**	99.17
Caltech-101	2020	**1011**	18432	**6075**	**7086**	97.43

VI. Conclusion

In this paper we have demonstrated that the performance of bag-of-features representations can be significantly improved by adopting convolutional features with much lower dimensionality. For the sake of the reduction of computational complexity, principal component analysis for CNN features and statistical codeword selection technique are introduced to reduce the feature dimension. Finally, the features of the reduced dimension obtained from CNN and BoF are fused using the connecting method. The fused features are then input into SVMs to accomplish the visual object classification. Experimental results on three different benchmark image datasets show that the proposed method has significantly improved the classification accuracy compared to the traditional bag-of-features approach.

References

[1] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, “Visual categorization with bags of keypoints,” in *Proc. Workshop on Statistical Learning in Computer Vision*, ECCV, 2004, pp. 1-22.

[2] A. Ramanan and M. Niranjan, “A review of codebook models in patch-based visual object recognition,” *Journal of Signal Processing Systems*, Springer New York, vol. 68, pp. 333-352, 2011.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet classification with deep convolutional neural networks,” *Advances in Neural Information Processing Systems*, pp. 1097-1105, 2012.

[4] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, “The PASCAL visual object classes (VOC) challenge,” *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303-338, 2010.

[5] J. Deng, W. Dong, R. Socher *et al*., “Imagenet: A large-scale hierarchical image database,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248-255.

[6] T. Joachims, “Text categorization with support vector machines: Learning with many relevant features,” in *Proc. the European Conference on Machine Learning*, 1998, pp. 137–142.

[7] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, “Return of the devil in the details: Delving deep into convolutional nets,” in *Proc. British Machine Vision Conference (BMVC)*, 2014.

[8] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, “CNN features off-the-shelf: An astounding baseline for recognition,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops,* 2014, pp. 806–813.

[9] P. Sermanet, D. Eigen, X. Zhang *et al.*, “Overfeat: Integrated recognition, localization and detection using convolutional networks,” in *Proc. the International Conference on Learning Representations*, 2014.

[10] J. Y-H. Ng, F. Yang, and L. S. Davis, “Exploiting local features from deep networks for image retrieval,” in *Proc. IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2015, pp. 53–61.

[11] L. Zheng, Y. Zhao, S. Wang, J. Wang, and Q. Tian, “Good practice in CNN feature transfer,” *arXiv preprint arXiv:1604.00133*, 2016.

[12] D. Lowe, “Distinctive image features from scale-invariant keypoints,” *International Journal of Computer Vision (IJCV)*, vol. 60, no. 2, pp. 91-110, 2004.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR),* 2014, pp. 580-587.

[14] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, “The Binormal Assumption on Precision-Recall Curves”, in *Proc. International Conference on Pattern Recognition (CVPR),* 2010, pp. 4263-4266.

[15] S. Lazebnik, C. Schmid, and J. Ponce, “A Sparse texture representation using local affine regions,” *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 27, pp. 1265-1278, 2005.

[16] P. Perona *et al.*, “Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories,” *Computer Vision and Image Understanding*, vol. 106, no. 1, pp. 59-70, 2007.

Janani Thangavel is an associate software engineer at WSO$_2$ Solutions, Sri Lanka. She received her B.Sc. Special in computer science in 2017 from the University of Jaffna, Sri Lanka. Her research interest includes pattern recognition and visual object classification.

Amirthalingam Ramanan is a senior lecturer at the Department of Computer Science at University of Jaffna, Sri Lanka. He received his B.Sc. special in computer science in 2002 from the University of Jaffna, Sri Lanka and his PhD in computer science from the University of Southampton, United Kingdom in 2010. His research interests are in the algorithmic and applied aspects of Machine Learning and Computer Vision.