15th IAPR International Conference on Machine Vision Applications (MVA)
Nagoya University, Nagoya, Japan, May 8-12, 2017.

A Deep Network Model based on Subspaces: A Novel Approach for Image Classification

Bernardo Bentes Gatto[1], Lincon Sales de Souza[2] & Eulanda M. dos Santos[3]
Federal University of Amazonas, Manaus, Brazil[1,3]
University of Tsukuba, Tsukuba, Japan[2]
{bernardo[1], emsantos[3]}@icomp.ufam.edu.br, lincons@cvlab.cs.tsukuba.ac.jp[2]

Abstract

In this paper, we propose a novel deep neural network based on learning subspaces and convolutional neural network with applications in image classification. Recently, multistage PCA based filter banks have been successfully adopted in convolutional neural networks architectures in many applications including texture classification, face recognition and scene understanding. These approaches have shown to be powerful, with a straightforward implementation that enables a fast prototyping of efficient image classification systems. However, these architectures employ filters based on PCA, which may not achieve high discriminative features in more complicated computer vision datasets. In order to cope with the aforementioned drawback, we propose a Hybrid Subspace Neural Network (HS–Net). The proposed architecture employs filters from both PCA and discriminative filters banks from more sophisticated subspace methods, therefore achieving more representative and discriminative information. In addition, the use of hybrid architecture enables the use of supervised and unsupervised samples, depending on the application, making the introduced architecture quite attractive in practical terms. Experimental results on three publicly available datasets demonstrate the effectiveness and the practicability of the proposed architecture.

1 Introduction

Image classification is one of the central problems in several fields such as pattern recognition, computer vision and image analysis. Since it is an important task for the success of a diverse range of applications including human-computer interaction, image and video retrieval, video surveillance, biometrics and social media networks [1, 2]. Such a complex task may be affected by many factors, like misalignment of the target objects, illumination conditions, occlusions, low image contrast and incorrect camera position. Generally, the categorization process of the input images into training classes may be quite difficult due to the fact that images from the same class might have large variation, making it impractical to create a model to represent this class in a coherent way. In addition, images from different classes may share common structures, increasing the difficulties of the classification task, as the set of common structures may reduce the discriminative ability of the model.

Recently, representation learning has been employed as a competitive alternative to hand-crafted features, such as Gabor features and Local Binary Patterns (LBP) for texture and face classification, and Scale-Invariant Feature Transform (SIFT) and Histogram of Oriented Gradients (HOG) features for object recognition [3, 4]. Learning through deep neural networks has received significant attention due to its impressive improvements over hand-crafted features. A main concept of deep learning is that all relevant information required for recognizing image patterns are contained in hierarchical neural network models through iterative learning of exemplar image patterns. By producing multiple levels of representation through the use of hierarchical models, the higher-level features generate more abstract semantics of the training images, achieving more invariance to intra-class variability.

An example of deep learning architecture is Convolutional Neural Network (CNN) that reached the state-of-the-art performance in various applications [5, 6, 7]. Despite its successful in several applications, the number of parameters to be trained is very large due to the large amount of data to be used, which can lead to a high computational cost, even when using machines equipped with GPU. This high computational complexity required from most of the deep learning architectures prevents some computer vision applications to fully employ the capabilities of deep convolutional networks. In order to solve this issue, several deep learning networks have been proposed based on PCA [8], LDA [9], Gabor and ICA [10] filter banks. For instance, in [8] is proposed a convolutional neural network with no pooling layers, nor active functions and without using back-propagation to learn the weights of the layers. Instead, PCA and LDA are employed to learn and handle the weights of the layers as filter banks in CNN. This approach exhibited performance comparable to the state-of-the-art for several image classification tasks. Other examples include a multilinear discriminant analysis network (MLDANet) [11] for tensor object classification and a discrete cosine transform network (DCTNet) [12] for face recognition.

Recently, Generalized Difference Subspace (GDS) [13] was proposed as a powerful feature descriptor for image classification. The core idea of GDS is based on the assumption that local shape differences among objects provide an efficient approach to represent the objects. GDS is performed by generating a subspace that encapsulates the difference components among all the class subspaces. This method has an impressive ability to evaluate local structures differences between the different class subspaces. Its effectiveness has been proved through extensive experimentation on several tasks, including face recognition and hand shape classification of multi-view images.

Although convolutional networks based on PCA have been successfully applied in various recognition tasks, PCA filters are not able to efficiently describe high overlapping distributions, which are easily found challenging datasets. In order to deal with the aforementioned drawback, this paper presents a novel object recognition method based on convolutional networks and GDS, called Hybrid Subspace Neural Network (HS–Net). In contrast with convolutional networks based on PCA [8, 10], the filter banks employed by HS–Net are produced by PCA and Difference Subspaces, which preserve the discriminative information among different classes, generating more efficient representations. In addition, HS–Net is able to operate on both labeled and unlabeled data, improving the performance in the presence of large volumes of data. Therefore, our contributions are twofold: (1) We investigate the use of a novel filter bank based on GDS. This filter bank is more powerful than PCA bank filters, as GDS preserves discriminative information, which is not achievable by PCA. (2) We examine the capabilities of PCA and GDS in a deep learning approach in order to exploit supervised and unsupervised data, creating a very flexible framework.

2 Related Work

In this session, we provide a brief review on CNN based-PCA, LDA and variants. This analysis is important in order to clarify the differences between the proposed network and the currenting methods. As well known, learning features directly from the datasets, rather than create complicated techniques has been recognized as a dominant trend to prevent the drawbacks of handcrafted features. Recently, most of the literature have pointed out that deep network architectures can produce higher level features and represent the abstract semantics of the data, decreasing the influence of intra-class variability. Therefore, learning features through the use of deep network architectures provides more invariance to intra-class variability.

Deep architectures based on CNN generally contains the following stages: convolutional filter layer, nonlinear processing layer, and feature pooling layer. In order to initialize the parameter of filter kernels and additive bias, a random schema is employed, which is iteratively updated by stochastic gradient descent (SGD). To cope with the nonlinear processing layer, the ReLu [14] and the Sigmoid functions are applied. Finally, the mean pooling or max pooling are employed in order to decrease the resolution of the feature map. The CNN architecture has been used in several tasks including face recognition [12], object detection [8] and scene understanding.

PCANet [8] is an image classification framework based on CNN, where multistage filter banks are learned from the data as principal components at the local image patch level. In PCANet, the basis vectors of the local covariance matrix are employed as filter banks for convolution and feature extraction, followed by binarization and block-wise histograming. This straightforward deep learning network works surprisingly well in a variety of image classification benchmarks, including handwritten and face recognition datasets, achieving superior performance to the state-of-the-art features.

DCTNet [12] is an alternative to PCANet, which employs Discrete Cosine Transform (DCT) as filter banks instead of PCA. As well known, PCANet is data-dependence hence inflexible. In DCTNet, on the other hand, the filter banks created by DCT achieve a data-independent network, increasing the performance of the network. In order to decrease the computational complexity of the learning stages of the network, 2D DCT is also employed. Besides the low computational complexity, 2D DCT filter banks are independent from data, therefore, generating a learning-free framework. DCTNet has been widely applied to several benchmarks of face databases and have shown performance equivalent or superior to PCANet and LDANet.

In all these examples, the employed techniques can be regarded as CNN architectures based on local multistage filter banks [12]. Therefore, we can introduce more sophisticated subspace methods such as GDS, where the discriminability of features is enhanced with the orthogonalization process of the different class subspaces. This fact encourages us to develop the most sophisticated subspace methods in order to exploits the use of complementary filters. This capability may greatly improve the produced features. In addition, to the best of our knowledge, there is very limited work conducting this type of analysis on subspace methods and deep network architectures.

3 Proposed Method

To understand the flow of the framework procedure, first consider a learning problem with N training images $\{\mathcal{I}_i\}_{i=1}^{N}$, each one with size $m \times n$. Conceptual illustration of the proposed method can be visualized in the Figure 1.

3.1 Representation by Patches

First, the images are divided into smaller patches, so consider a patch size of the form $k_1 \times k_2$. The jth patch of the ith image is represented as the vector $\boldsymbol{x}_{i,j} \in \mathbb{R}^{k_1 k_2}$. The collection of all overlapping patches mapped around each pixel is represented as the matrix: $\boldsymbol{X}_i = [\boldsymbol{x}_{i,1}, \boldsymbol{x}_{i,2}, ..., \boldsymbol{x}_{i,mn}] \in \mathbb{R}^{k_1 k_2 \times mn}$.

Suppose that for a K classes problem the cth class holds N_c images. The collection of all images in one class is represented as: $\boldsymbol{X}_c = [\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_{N_c}] \in \mathbb{R}^{k_1 k_2 \times N_c mn}$.

3.2 Difference Subspace Filtering

The first step in filtering stage is to Principal Component Analysis (PCA) to be executed for each class. In terms of the filter space $\mathbb{R}^{k_1 k_2}$, and assuming a class dimensionality d, it can be said that the objective of PCA is to find the class subspace $\boldsymbol{E}_c$ that minimizes the approximation error, i.e.: $min \left\| \boldsymbol{X}_c - \boldsymbol{E}_c \boldsymbol{E}_c^T \boldsymbol{X}_c \right\|_F^2, s.t. \boldsymbol{E}_c \in \mathbb{R}^{k_1 k_2 \times d}$.

The solution for $\boldsymbol{E}_c$ is to calculate the eigenvectors of $\boldsymbol{X}_c \boldsymbol{X}_c^T$. Once equipped with all class subspaces, the sum matrix $\boldsymbol{A}$ is yielded by the covariance matrices of each class subspace:

$$\boldsymbol{A} = \sum_{c=1}^{C} \boldsymbol{E}_c \boldsymbol{E}_c^T \tag{1}$$

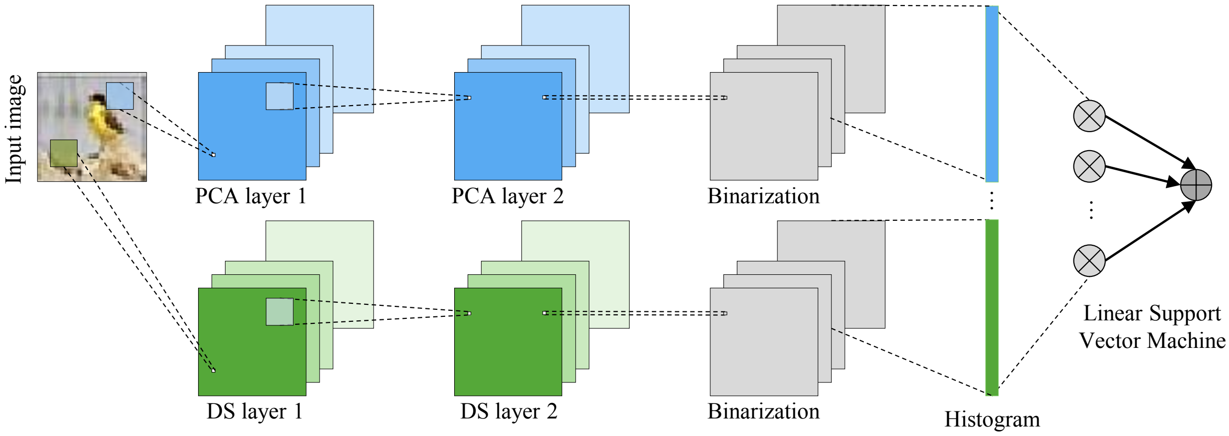

Figure 1. Conceptual illustration of the proposed method. The HS–Net employs two distinct filters that works in complementary directions. The input image is represented by its difference subspace features and PCA features, improving the produced feature robustness, as the difference subspace features enforces discriminability among different image classes. In order to reduce the high dimensionality of the features and increase rotation invariance, the proposed method is followed by binarization and histogramming. The classification is performed by using Linear Support Vector Machine.

The eigenvectors of $\boldsymbol{AA}^T$ actually form the sum subspace [13], which is the direct sum of a Principal Subspace $\boldsymbol{P}$, and the General Difference Subspace $\boldsymbol{G}$:

$$eig(\boldsymbol{AA}^T) = \boldsymbol{P} \oplus \boldsymbol{G} \tag{2}$$

This relation means that the null space is the only intersection between $\boldsymbol{P}$ and $\boldsymbol{G}$, so assuming the dimensionality of the principal subspace is D, the dimensionality of the GDS must be at most $L = D - k_1k_2$. Thus, we can say that $\boldsymbol{G} \in \mathbb{R}^{k_1k_2 \times L}$.

Each basis vector of the GDS will be a filter in the network, in such a way that a dimensionality L_s is also the number of filters in the layer s. By using these concepts, the definition of a GDS filter can be realized as: $\boldsymbol{W}_l^s = map_{k_1 \times k_2}(\boldsymbol{g}_l) \in \mathbb{R}^{k_1 \times k_2}, l = 1, 2, ..., L_s$, where $map_{k_1 \times k_2}$ is a function that maps the lth basis vector $\boldsymbol{g}_l$ to a matrix $\boldsymbol{W} \in \mathbb{R}^{k_1 \times k_2}$, which expresses the main variation between classes in such a way that looks to maximize their difference, by highlighting local shapes on its projected images. Then, the output of the stage is the operation:

$$\mathcal{I}_i^l \doteq W_l^s * \mathcal{I}_i \tag{3}$$

where $*$ refers to a 2D convolution with zero-padding in the boundary. That makes $\mathcal{I}_i^l$ have the same size of $\mathcal{I}_i$. Note that the output of one stage of GDS Filtering produces L_sN images ($i = 1, ..., N$ and $l = 1, ..., L_s$). And similar to DNN and PCANet, multiple filtering stage architectures can be created, by feeding the produced images as input to a new stage. In general, a Z layers filtering system produces $N_Z = L_1L_2...L_Z$ images for each of the N images, so in total N_ZN images are produced. For a high number of layers, the number of filter indexes will increase, so for generalization purposes the representation of images produced by Z layers will be $\{\mathcal{I}_i^z\}_{z=1}^{N_Z}$. The next steps are the Hashing and Histogram procedures, which are the same techniques employed by [8].

The vector f_i is a column vector in a sparse matrix of observations f, which is used to train a classifier. In our architectures we have used support vector machines (SVM). The hyper-parameters of the HS–Net include the filter size k_1, k_2, the number of filters in each stage $L_1, L_2, ..., L_Z$, the number of stages Z, the block size for the histogram, and the dimensionality of the class subspaces and principal subspace.

4 Experimental Evaluation

In order to evaluate the performance of the HS–Net, we use LFW dataset [15]. LFW dataset consists of 13233 images of faces collected from the web. The faces were detected using Viola-Jones face detector and cropped. In addition, 1680 of all 5749 individuals have two or more distinct photos in the dataset. LFW dataset is a specially challenging dataset because it was designed for studying the problem of unconstrained face recognition.

For object recognition, we use CIFAR-10 [16] dataset that consists of $50,000$ training and $10,000$ test images. In CIFAR-10 dataset there are 10 classes, namely airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. The large variability in scale, viewpoint, illumination, and background clutter poses a significant challenge for classification.

We also use NYU Depth V1 dataset [17], which was collected by the New York University. The dataset includes depth information which contains both geometric information and distance of objects. NYU Depth V1 dataset consists of 2347 pairs of images grouped into seven categories, including bathroom, bedroom, bookstore, cafe, kitchen, living room, and office.

In our experiments, we set all networks to two layers as in our experiments employing more than two layers does not significantly affect the performance of the methods. In addition, we set the filter size $k = 5 \times 5$ for each layer and the number of filter for each layer is $P_1 = P_2 = 8$.

Table 1. Accuracy of HS–Net compared to the PCANet, LDANet and DCTNet.

Databases	PCANet	LDANet	DCTNet	HS–Net
CIFAR-10 [16]	78.67 ± 2.11	78.33 ± 2.19	77.13 ± 2.33	**80.11 ± 1.93**
LFW dataset [15]	85.20 ± 1.46	85.67 ± 1.87	84.20 ± 1.93	**86.78 ± 1.39**
NYU Depth V1 [17]	81.59 ± 1.55	80.20 ± 1.67	79.33 ± 1.71	**82.79 ± 1.47**

5 Conclusions and Future Directions

In this paper, we presented a new image classification framework for face recognition, object recognition and scene understanding, namely Hybrid Subspace Neural Network. In order to show a flexibility of the proposed method, we perform experiment evaluation on LFW, CIFAR-10 and NYU Depth V1 dataset. We showed that by employing PCA and DS filters on HS–Net we could improve the classification accuracy of the method. In addition, we introduced the concept of DS filters, which efficiently extracts high discriminant features, improving the features produced by the proposed method. This fact motivated us to investigate the relationship between the number of supervised features employed by DS filter banks and the number of unsupervised features employed by PCA filter banks. The proposed method has the advantage of efficiently make use of diverse source of features, depending on the application. Therefore, we experimentally showed that the resulting framework is competitive with existing methods while making use of supervised and unsupervised features.

For future work, we will investigate how to automatically select the number of basis vectors employed by PCA and DS filter banks. An interesting direction would introduce a new analysis involving all the eigenvectors of PCA and DS to select the most discriminative filters. Another important avenue is to develop a tensor version of the HS–Net in order to deal with video analysis, as gesture recognition and action recognition. As is well known, the performance of subspace-based methods highly depends on the data distribution. Hence, by using mechanisms where the relative importance of different data distributions can be analyzed separately one may perform a weighted combination of the filter banks, improving the quality of the features produced by the network.

References

[1] W. Zhang, M. L. Smith, L. N. Smith, and A. Farooq, “Gender and gaze gesture recognition for human-computer interaction,” *Computer Vision and Image Understanding*, vol. 149, pp. 32–50, 2016.

[2] Y. Li, R. Wang, Z. Huang, S. Shan, and X. Chen, “Face video retrieval with image query via hashing across euclidean space and riemannian manifold,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4758–4767.

[3] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[4] N. Dalal and B. Triggs, “Histograms of oriented gradients for human detection,” in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR’05)*, vol. 1. IEEE, 2005, pp. 886–893.

[5] D. Ciregan, U. Meier, and J. Schmidhuber, “Multi-column deep neural networks for image classification,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.* IEEE, 2012, pp. 3642–3649.

[6] C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning hierarchical features for scene labeling,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[7] Y. Sun, Y. Chen, X. Wang, and X. Tang, “Deep learning face representation by joint identification-verification,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[8] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, and Y. Ma, “Pcanet: A simple deep learning baseline for image classification?” *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[9] M. Dorfer, R. Kelz, and G. Widmer, “Deep linear discriminant analysis,” *arXiv preprint arXiv:1511.04707*, 2015.

[10] C.-Y. Low, A. B.-J. Teoh, and C.-J. Ng, “Multi-fold gabor filter convolution descriptor for face recognition,” in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2016, pp. 2094–2098.

[11] R. Zeng, J. Wu, L. Senhadji, and H. Shu, “Tensor object classification via multilinear discriminant analysis network,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, 2015, pp. 1971–1975.

[12] C. J. Ng and A. B. J. Teoh, “Dctnet: A simple learning-free approach for face recognition,” in *2015 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA).* IEEE, 2015, pp. 761–768.

[13] K. Fukui and A. Maki, “Difference subspace and its generalization for subspace-based methods,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 11, pp. 2164–2177, 2015.

[14] V. Nair and G. E. Hinton, “Rectified linear units improve restricted boltzmann machines,” in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 807–814.

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, “Labeled faces in the wild: A database for studying face recognition in unconstrained environments,” Technical Report 07-49, University of Massachusetts, Amherst, Tech. Rep., 2007.

[16] A. Krizhevsky and G. Hinton, “Learning multiple layers of features from tiny images,” 2009.

[17] N. Silberman and R. Fergus, “Indoor scene segmentation using a structured light sensor,” in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on.* IEEE, 2011, pp. 601–608.