DISCRIMINATIVE CANONICAL CORRELATION ANALYSIS NETWORK FOR IMAGE CLASSIFICATION

Bernardo B. Gatto, Eulanda M. dos Santos

Institute of Computing (ICOMP), Federal University of Amazonas, Brazil
Email: {*bernardo, emsantos*}*@icomp.ufam.edu.br*

ABSTRACT

Recently, convolutional neural network has attracted an increasing amount of attention in machine learning and computer vision areas, improving the performance of several related applications. Currently, many deep learning network architectures such as principal component analysis network (PCANet), linear discriminant analysis network (LDANet) and canonical correlation analysis network (CCANet) have been proposed for object and face classification. These architecture solutions have demonstrated high efficiency, with a simple implementation, providing a fast prototyping of efficient image classification applications. However, these solutions take advantage of filters that may not extract high discriminative features in more complicated computer vision problems (*e.g.* containing high degree of overlap between the distributions of the data). To generate more discriminative information, we introduce a discriminative canonical correlation network (DCCNet), that employs filters constructed from discriminative canonical correlations analysis (DCC). Learning filters from DCC ensures that the network will produce discriminative features, generating more representative and discriminative information. We demonstrate the applicability of DCCNet through experiments on four datasets.

Index Terms— discriminative canonical correlations analysis, PCA-based network, image classification

1. INTRODUCTION

Image classification is a crucial task for the success of several applications including human-computer interaction, image and video retrieval, video surveillance, biometrics and social media networks [1, 2]. Accordingly, image classification still plays a decisive role in pattern recognition, computer vision and image analysis. One of the central issues on applying image classification techniques in real world problems is the intra-class variation, which is due to many factors, like misalignment of the target objects, illumination conditions, occlusions, low image contrast and incorrect camera position.

In the last decade, many efforts have been made to alleviate these drawbacks, such as low-level features. These methods, including local binary patterns (LBPs) [3], histogram of oriented gradients (HOG) [4] and scale-invariant feature transforms (SIFT) [5], have achieved high recognition accuracy when employed by a traditional machine learning technique such as a support vector machines (SVM) classifier. However, even by extracting color, textures and gradients of an image in an efficient approach, these strategies lack flexibility when dealing with new datasets, leading the adoption of such hand-crafted features to be restricted or sometimes impractical in more advanced applications [6].

The problem of image classification has received a considerable amount of attention, particularly in the area of deep learning [7, 8]. The recent improvement over the hand-crafted features motivate the use of deep architectures. A main concept of deep learning is that all relevant information required to recognize image patterns is contained in hierarchical neural network models. 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 important neural network is convolutional neural network (CNN) that reached the state-of-the-art performance in various applications [9, 10]. In despite of its favorable results, in the CNN method, parameters fine-tuning is a time consuming task [11], even when using machines equipped with GPU. In order to alleviate the high computational cost of training a CNN, a few networks have been proposed based on PCA [12], LDA [13], Gabor and ICA [14] filter banks. For example, PCANet employs a CNN architecture with no pooling layers, nor active functions and without using back-propagation to learn the weights of the layers. Accordingly, these filter banks may be generated by employing PCA or LDA. Moreover, these approaches exhibit competitive results compared to the state-of-the-art results for image classification tasks.

Although the aforementioned learning networks have shown to be effective, these methods can still be improved in order to achieve more robust performance.For instance, PCANet does not make any distinction between different class features, creating a common subspace that reflects the training set, without considering discriminative information between the different image classes. Therefore, instead of employing PCA or LDA to learn the filter banks, we use discriminative canonical correlations (DCC) [15]. More

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

precisely, we employ DCC to develop a transformation matrix T that maximizes the canonical correlations of within-class image sets and minimizes the canonical correlations of between-class image sets. Our assumption is that these filters can reveal more discriminative information of the same input patterns compared to PCANet or LDANet.

It worth noting that DCC was originally proposed for image-set classification as an alternative to subspace-based methods [16, 17, 18]. We understand that the subspace-based methods, including PCA and LDA, are suitable for single image classification. Figure 1 shows the conceptual diagram of DCCNet. We summarize our contributions as follows:

(1) We investigate the use of a novel filter bank based on the transformation matrix of DCC. This transformation has show a powerful representation and may achieve more discriminative information when employed as filter banks than the filter banks provided by PCA.

(2) The proposed method is designed to handle supervised multiple class features, therefore it is able to cope with the high variability of the data, which is not feasible by PCA. The capabilities of DCC filter banks are therefore strengthening, since making use of supervised features leads to a solution that efficiently increases the robustness of the produced features, improving its performance. Hence, the proposed method may be applied to many computer vision tasks, as will be shown in comprehensive experiments.

(3) We provide extensive experiment results with a number of benchmark image datasets for the proposed DCC Network. The datasets considered are: LFW dataset [19], which are images of faces collected from the web. For object recognition, we use CIFAR-10 [20] dataset. NYU Depth V1 dataset [21], which includes depth information which contains both geometric information and distance of objects. We also employ Street View House Numbers dataset (SVHN) [22], which are images of house numbers collected by Google Street View.

The organization of this work is as follows: Section 2 presents the details of the introduced CNN based architecture. In Section 3, we provide the experimental results and discussions. Finally, the paper is summarized and concluded with future research directions in 4.

2. PROPOSED METHOD

2.1. Discriminative Canonical Correlation Filtering

Let us assume that C sets of training images are given by $\{A_1, A_2, \dots, A_C\}$, where A_i is a $n \times M$ matrix containing M $d-$dimensional images. Let us assume that each A_i set belongs to one of the C object classes. Then, we assume that there is a linear transformation that represents each A_i set in terms of its variance by a subspace spanned by P_i.This new representation provides a more compact manner to represent each A_i set. Each P_i basis vectors spans a reference class subspace, that can be efficiently computed by:

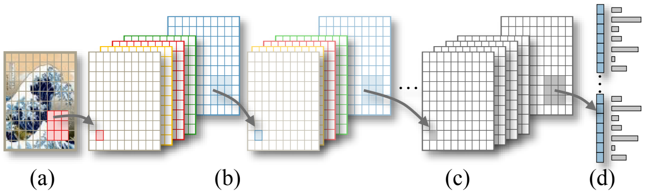


Fig. 1. The Convolutional Neural Network architecture introduced in this work: (a) An input image is processed by a convolutional feature extraction layer based on DCC filters (b). Then, Binary Hashing is applied on the features (c). Finally, a block-wise histogramming creates the final feature (d).

$$A_i A_i^T \cong P_i \Lambda_i P_i^T, \tag{1}$$

where Λ_i is the top$-k$ largest eigenvalue and P_i is the corresponding eigenvectors forming unitary orthogonal bases for a linear subspace. After computing all the P_i basis vectors, we are able to compute the T transformation matrix of DCC. Scatter measures of within-class and between-class can be related to pairwise correlation by the scatter matrices:

$$B = \sum_i (x_i)^T (x_i)^B, W = \sum_i (x_i)^T (x_i)^W, \tag{2}$$

and x_i^B, x_i^W indicate the closest between-class and within-class vectors of a given vector x_i. The aim is to find a transformation T that transform all P_i subspaces to another space containing maximal correlation within the same class and minimal correlation between different classes. In order to compute T, we employ the discriminant analysis of canonical correlations by solving:

$$T = \arg\max_T \frac{tr(T^T S_b T)}{tr(T^T S_w T)}, \tag{3}$$

where S_b and S_w are the between-class and within-class scatter matrices. The transformation matrix T is solved as the eigen-decomposition of $S_w^{-1} S_b$. By adopting the T matrix as filter banks $\{t_i\}_{i=1}^n$ on a CNN architecture, we achieve more robustness than PCANet and LDANet, as we can efficiently represent the different image classes, which is not achievable by PCANet based filter banks. Also, T is able to discriminant the input patters, increasing the network robustness.

2.2. Discriminative Canonical Correlation Network

The architecture of DCCNet is shown in Figure 1. DCCNet consists of DCC multistage convolutional layers, hashing and histogram. Assume that we have N training images of size $m \times n$. In each image, we take a patch of size

$k_1 \times k_2$ around each pixel. After collecting all the patches, we vectorize the patches and combine them into a matrix of $k_1 \times k_2$ rows and $(mk_1 + 1)(nk_2 + 1)$ columns. For the $i-$th image I_i of the $K-$th class, we obtain a matrix X_i: $\mathrm{X} = [X_1, X_2, \dots, X_N] \in R^{k_1 k_2 \times Nc}$, where c stands for the number of rows of X_i. Then, we move on to obtain the eigenvectors of $X_i X_i^T$, as shown in Equation 1, and save the ones corresponding to the L_1 largest eigenvalues as the subspace of the X_i image class.

After collecting all the subspaces P_i of each image set class, we now compute the $T = \{t_i\}_{i=1}^{n}$ transformation matrix in order to create the DCC filters. We perform this operation by employing the Equation 3.

The leading principal eigenvectors capture the main variation of all the training patches. Thus, we finish the first stage. At the second stage, we share similar process with stage 1. The input images T_i^l of stage 2 should be: $I_i^l = I_i * W_l^1, i = 1, 2, \dots, N$ the boundary of I_i is zero-padded so that T_i^l have the same size of I_i. We collect all the patches of T_i^l and, we combine the Y_l together as a matrix: $Y = \left[Y^1, Y^2, \dots, Y^{L_1}\right] \in R^{k_1 k_2 \times L_1 Nc}$. After that, we obtain the eigenvectors of YY^T, maintaining the ones corresponding to the L_2 largest eigenvalues in order to create the subspaces and second stage filters, according to Equation 1: $W_\ell^2 = q_\ell\left(YY^T\right) \in R^{k_1 k_2}, \ell = 1, 2, \dots, L_2$. At the final stage, for each input image of stage 2, we get:

$$T_i^l = \sum_{\ell=1}^{L_2} 2^{\ell-1} H\left(I_i^l * W_\ell^2\right), l = 1, 2, \dots L_1$$

The function $H(.)$ binaries these outputs, i.e. the value of the function is one for positive inputs and zero otherwise. For each of the L_1 images $T_i^l, l = 1, 2, 3, \dots, L_1$ we partition it into B blocks, whose size is $k_1 k_2 \times B$, and we compute the $2^{L_2} \times B$ histogram matrix in each block ranging from $\left[0, 2^{L_2} - 1\right]$, followed by vectorizing the matrix into a row vector $Bhist(T_i^l)$. Finally, we concatenate the $Bhist(T_i^l)$ of $T_i^l, l = 1, 2, 3 \dots, L_1$ as the feature

$$f_i = \left[Bhist\left(T_i^1\right), \dots, Bhist\left(T_i^{L_1}\right)\right]^T \in R^{\left(2^{L_2}\right) L_1 B}$$

According to the literature [23, 24], the local blocks can be either overlapping or non-overlapping, and in our work, we keep it as non-overlapping.

3. EXPERIMENTAL RESULTS

To evaluate the effectiveness of the proposed DCCNet, we conduct experiments using several public databases, including LFW dataset [19], which are images of faces collected from the web. For object recognition, we use CIFAR-10 [20] dataset. NYU Depth V1 dataset [21], which includes depth information which contains both geometric information and distance of objects. We also employ Street View House Numbers dataset (SVHN) [22], which are images of house numbers collected by Google Street View. These datasets cover various unconstrained scenarios including different point of view and illumination conditions. All the images have been converted to grayscale before processing and, for a fair comparison, we employ the Coiflets and Daubechies orthogonal wavelet transform to extract the low frequency sub-images of the original images to generate two view features for the CCANet [25]. In addition, we do not use the TR-Normalization introduced by [26]. For the classification, we use a linear SVM classifier.

LFW dataset consists of more than 13000 images of faces. 1680 of the people pictured have two or more distinct photos in the data set. In addition, the images were aligned with deep funneling [27]. We center crop the images to size 150×150 to exclude most of the background to focus on recognizing the face. Then we employ image patches of size 32×32 and we set $L_1 = L_2 = 8$. For the 3 stages layer experiments, we employ $L_3 = 4$. For this dataset, these parameter settings are the same for all the evaluated methods. CIFAR-10 dataset consists of 60000 32×32 color images in 10 classes, with 6000 images per class. There are 50000 training images and 10000 test images. In this dataset, we employ filter size $k_1 = k_2 = 5$, the number of filters $L_1 = 40$, $L_2 = 8$, $L_3 = 4$ and block size equal to 8×8.

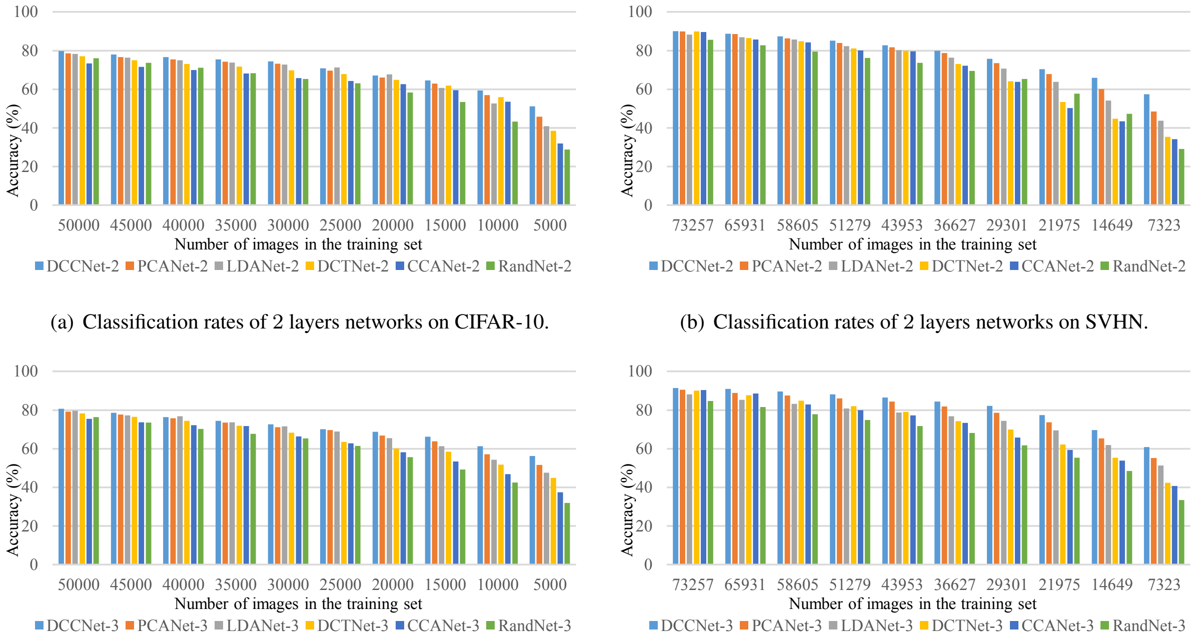
NYU Depth V1 dataset is comprised of video sequences from a variety of indoor scenes as recorded by both the RGB and Depth cameras from the Microsoft Kinect. The dataset has RGB and depth information and has 7 scene classes including bathroom, bedroom, bookstore and kitchen. In this dataset, we employ $L_1 = L2 = 8$, $L_3 = 4$ and blocks of 8×8. SVHN dataset consists of 10 classes, 1 for each digit. We used 73257 digits for training, 26032 digits for testing. The input image digits have size of 32×32 and we convert the images to grayscale. We use image patches of size 8×8 and we set $L_1 = L_2 = 4$. For the 3 stages layer experiments, we employ $L_3 = 2$.

In the first experiment, we evaluate the performance of the networks using 2 and 3 stages. This experiment is important to show the improvement of the accuracy when employing more network layers. Table 1 shows the mean accuracy and the standard deviation of different methods on the LFW, CIFAR-10, NYU Depth V1 and SVHN datasets, when 2 and 3 layers are employed. We can see that most of the methods have improvements when using 3 layers. DCCNet demonstrated the best accuracy network among the evaluated networks, confirming the robustness of the method by employing the discriminative transformation matrix. Not surprisingly, RandNet is much worse than the other methods. Hence, PCANet and LDANet shows high performance even compared to CCANet.

In order to evaluate the robustness of the proposed network, we also perform experiments changing the number of

Dataset	Layers	PCANet [12]	LDANet [12]	RandNet [12]	DCTNet [26]	CCANet [25]	DCCNet
CIFAR-10 [20]	2	78.67 ± 2.11	78.33 ± 2.19	76.11 ± 2.2	77.13 ± 2.33	73.44 ± 1.88	**79.87 ± 1.67**
	3	79.13 ± 2.17	79.66 ± 2.03	75.36 ± 2.23	78.27 ± 2.38	74.49 ± 1.79	**80.68 ± 1.59**
LFW dataset [19]	2	85.20 ± 1.46	**85.67 ± 1.87**	81.77 ± 2.11	84.20 ± 1.93	83.33 ± 1.97	**85.67 ± 1.84**
	3	85.67 ± 1.39	85.67 ± 1.73	82.69 ± 2.03	84.66 ± 2.17	83.69 ± 1.81	**85.93 ± 1.71**
NYU Depth V1 [21]	2	**81.59 ± 1.55**	80.20 ± 1.67	78.81 ± 1.53	79.33 ± 1.71	80.67 ± 1.55	80.89 ± 1.77
	3	**82.25 ± 1.51**	80.20 ± 1.55	79.63 ± 1.84	80.13 ± 1.59	80.81 ± 1.47	81.11 ± 1.67
SVHN dataset) [22]	2	89.97 ± 1.33	88.33 ± 1.51	85.67 ± 1.33	89.88 ± 1.45	89.67 ± 1.49	**90.11 ± 1.67**
	3	90.55 ± 1.25	88.13 ± 1.48	84.72 ± 1.29	90.13 ± 1.67	90.43 ± 1.51	**91.37 ± 1.61**

Table 1. The best average classification rates and standard deviation results are shown in bold.



(a) Classification rates of 2 layers networks on CIFAR-10.

(b) Classification rates of 2 layers networks on SVHN.

(c) Classification rates of 3 layers networks on CIFAR-10.

(d) Classification rates of 3 layers networks on SVHN.

Fig. 2. Classification rates of different image classification networks when decreasing the number of images in the training set.

images in the training set. For this experiments, we only employ CIFAR-10 and SVHN datasets, as the number of samples are huge comparing to LFW and NYU Depth V1 datasets. Therefore, with less samples per training set, it is possible to analyze the accuracy of the proposed network in relation to the number of images in the training set. Figure 2 show the results for this experiment. Most of the methods have shown high robustness level when decreasing the number of images in the training set. Indeed, PCANet-2 and PCANet-3 have shown strong feature descriptor, as an unsupervised feature. LDANet-2 and LDANet-3 seems to produce weak features when more than half of the training set is absent. DCCNet-2 and DCCNet-3 shows high ability to handle small amount of images for training due to its generative model to describe the different image classes as subspaces. Hence, the DCC subspace T enforces the feature invariance.

4. CONCLUSIONS AND FUTURE DIRECTIONS

We presented a new image classification framework for face recognition, numbers recognition and scene understanding, namely Discriminative Canonical Correlation Network. In order to show the flexibility of the proposed method, we perform experiment on LFW, CIFAR-10, SVHN and NYU Depth V1 datasets. We showed that by employing DCC filters we could create strong discriminant features, when compared to PCANet and LDANet. These features demonstrated high discriminative power, even when we decrease the number of available training images. For future work, we will investigate how to automatically select the number of basis vectors employed by DCC filter banks. Another important avenue is to develop a tensor version of the DCCNet in order to hanfle gesture recognition and action recognition problems.

5. REFERENCES

[1] Wenhao Zhang, Melvyn L Smith, Lyndon N Smith, and Abdul Farooq, "Gender and gaze gesture recognition for human-computer interaction," *Computer Vision and Image Understanding*, vol. 149, pp. 32–50, 2016.

[2] Yan Li, Ruiping Wang, Zhiwu Huang, Shiguang Shan, and Xilin 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] Timo Ojala, Matti Pietikainen, and Topi Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

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

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

[6] Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus, "Deconvolutional networks," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 2528–2535.

[7] Peng Tang, Jin Zhang, Xinggang Wang, Bin Feng, Fabio Roli, and Wenyu Liu, "Learning extremely shared middle-level image representation for scene classification," *Knowledge and Information Systems*, pp. 1–22, 2016.

[8] Peng Tang, Xinggang Wang, Bin Feng, and Wenyu Liu, "Learning multi-instance deep discriminative patterns for image classification," *IEEE Transactions on Image Processing*, 2016.

[9] Clement Farabet, Camille Couprie, Laurent Najman, and Yann LeCun, "Learning hierarchical features for scene labeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1915–1929, 2013.

[10] Yi Sun, Yuheng Chen, Xiaogang Wang, and Xiaoou Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems*, 2014, pp. 1988–1996.

[11] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[12] Tsung-Han Chan, Kui Jia, Shenghua Gao, Jiwen Lu, Zinan Zeng, and Yi Ma, "Pcanet: A simple deep learning baseline for image classification?," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5017–5032, 2015.

[13] Matthias Dorfer, Rainer Kelz, and Gerhard Widmer, "Deep linear discriminant analysis," *arXiv preprint arXiv:1511.04707*, 2015.

[14] Cheng-Yaw Low, Andrew Beng-Jin Teoh, and Cong-Jie 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.

[15] Tae-Kyun Kim, Josef Kittler, and Roberto Cipolla, "Discriminative learning and recognition of image set classes using canonical correlations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, 2007.

[16] Kazuhiro Fukui and Osamu Yamaguchi, "The kernel orthogonal mutual subspace method and its application to 3d object recognition," in *Asian Conference on Computer Vision*. Springer, 2007, pp. 467–476.

[17] Bernardo B Gatto, SS Waldir, and Eulanda M dos Santos, "Kernel two dimensional subspace for image set classification," in *Tools with Artificial Intelligence (ICTAI), 2016 IEEE 28th International Conference on*. IEEE, 2016, pp. 1004–1011.

[18] Bernardo B Gatto and Eulanda M dos Santos, "Image-set matching by two dimensional generalized mutual subspace method," in *Intelligent Systems (BRACIS), 2016 5th Brazilian Conference on*. IEEE, 2016, pp. 133–138.

[19] Gary B Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Tech. Rep., Technical Report 07-49, University of Massachusetts, Amherst, 2007.

[20] Alex Krizhevsky and Geoffrey Hinton, "Learning multiple layers of features from tiny images," 2009.

[21] Nathan Silberman and Rob 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.

[22] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng, "Reading digits in natural images with unsupervised feature learning," in *NIPS workshop on deep learning and unsupervised feature learning*, 2011, vol. 2011, p. 5.

[23] Yufei Gan, Tong Zhuo, and Chu He, "Image classification with a deep network model based on compressive sensing," in *Signal Processing (ICSP), 2014 12th International Conference on*. IEEE, 2014, pp. 1272–1275.

[24] Lei Tian, Chunxiao Fan, Yue Ming, and Yi Jin, "Stacked pca network (spcanet): an effective deep learning for face recognition," in *Digital Signal Processing (DSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1039–1043.

[25] Xinghao Yang, Weifeng Liu, Dapeng Tao, and Jun Cheng, "Canonical correlation analysis networks for two-view image recognition," *Information Sciences*, 2017.

[26] Cong Jie Ng and Andrew Beng Jin 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.

[27] Gary B. Huang, Marwan Mattar, Honglak Lee, and Erik Learned-Miller, "Learning to align from scratch," in *NIPS*, 2012.