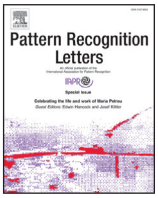
Pattern Recognition Letters 85 (2017) 79–83

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec



Multi-channel multi-model feature learning for face recognition

Melih S. Aslan[1],*, Zeyad Hailat[1], Tarik K. Alafif, Xue-Wen Chen*

Computer Science Dept., Wayne State University, 5057 Woodward Ave. Rm 3010, Detroit, MI 48202, United States

ARTICLE INFO

Article history:
Received 15 June 2016
Available online 30 November 2016

Keywords:
Unsupervised learning
Face recognition
Autoencoder
Sparse estimation
ADMM

ABSTRACT

Different modalities have been proved to carry various information. This paper aims to study how the multiple face regions/channels and multiple models (e.g., hand-crafted and unsupervised learning methods) answer to the face recognition problem. Hand crafted and deep feature learning techniques have been proposed and applied to estimate discriminative features in object recognition problems. In our Multi-Channel Multi-Model feature learning (McMmFL) system, we propose a new autoencoder (AE) optimization that integrates the alternating direction method of multipliers (ADMM). One of the advantages of our AE is dividing the energy formulation into several sub-units that can be used to paralyze/distribute the optimization tasks. Furthermore, the proposed method uses the advantage of K-means clustering and histogram of gradients (HOG) to boost the recognition rates. McMmFL outperforms the best results reported on the literature on three benchmark facial data sets that include AR, Yale, and PubFig83 with 95.04%, 98.97%, 95.85% rates, respectively.

1. Introduction

Ideally, object and face identification has four procedures - feature learning, feature extraction using labeled data, supervised training, and testing. Representative and discriminative features are desired to be learned and extracted from the object of interests. To boost the identification rate and to accelerate the learning process, many hand-crafted and unsupervised learning techniques have been developed that we will review a few of them below.

Since global representation methods, such as Eigenface [1] and Fisherface [2], fail to capture high-order statistics, local feature extraction techniques have been proposed such as local binary pattern (LBP) [3], scale-invariant feature transform (SIFT) [4], histograms of oriented gradients (HOG) [5], rotation-and scale-invariant, line-based color-aware descriptor (RSILC) [6], and correlation based features [7]. Although those techniques have proved that they are capable of obtaining good classification accuracy in limited scenarios, they are incapable of extracting the non-linear features.

Deep learning methods are designed to learn hierarchical representations in deep architectures for classification [8]. Traditional unsupervised models such as sparse Restricted Boltzmann Machine (RBM) [9], and sparse auto-encoder [10] have shown improved results in many classification tasks. Hierarchical model for sparse representation learning was proposed to build high level features [11]. Greedy layer wise pre-training [12,13] approach in deep learning [8] became very popular for deep hierarchical frameworks. Multi-layer of stacked sparse auto-encoder (SAE) [11,13,14], sparse deep belief net (DBN), and convolutional deep belief net (CDBN) [15] are few frameworks for learning sparse representation.

Several methods have been proposed in the literature that combines multiple modalities to enhance the face recognition performance. Ngiam et al. [16] proposed a multimodel learning technique that combines the features of the visual and audio information. Srivastava et al. [17] proposed a generative model of data that consists of multiple and diverse input modalities. They used a Deep Boltzmann Machines (DBM) to handle multimedia data feature learning such as image database with tags. Their model generates a fused representation from multiple data modalities. Shekhar et al. [18] proposed a multimedia or multi-biometric identification method that combines the information from different biometric modalities. Nilsback et al. [19] made a representative analysis on combining hand-crafted features (e.g., HOG, SIFT, and Hue-saturation-value) on flower classification. Huang et al. [20] proposed an idea that combines features from their deep learning system and hand-crafted techniques. The combination of multiple modalities slightly increased the face verification accuracy.

In this paper, we combine features extracted from multiple regions that are processed with multiple models such as hand-crafted and unsupervised feature learning methods. The main contributions are summarized as follows:

* Corresponding authors.
E-mail address: melih.aslan@wayne.edu (M.S. Aslan).

[1] These authors have equal contribution.

http://dx.doi.org/10.1016/j.patrec.2016.11.021
0167-8655/© 2016 Published by Elsevier B.V.

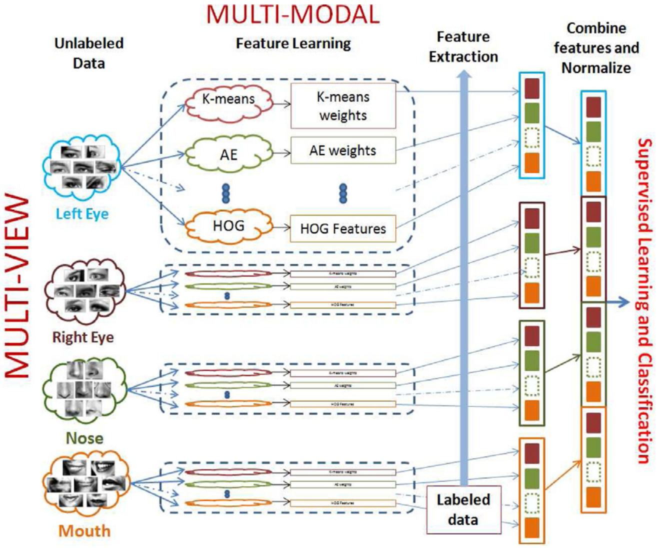


Fig. 1. Architecture of the proposed Multi-Channel Multi-Model feature learning (McMmFL) system.

- We propose a new AE optimization and draw upon the idea from the alternating direction method of multipliers (ADMM) formulation [21]. Our proposed encoder-decoder module efficiently extracts sparse representation of facial regions. One of the most important advantages of the ADMM-based optimization is the ability to divide the energy formulation into several units that can be used to paralyze/distribute the optimization tasks.
- The multi-channel learning procedure extracts representations that capture intra-region changes more precisely. Additionally, the unsupervised learning methods obtain specialized bases for corresponding regions. Instead of estimating a single centroid of a face region, feature learning for multi-region increases the detailed representation that learns more representative information as we assess this point in our experiments.
- Finally, fusing various features from multiple techniques enables us to achieve promising results.

The paper is organized as follows: Section 2 introduces the proposed method in details. The experimental setup and results are explained and discussed in Section 3. Finally, we conclude in Section 4.

2. Methods

Our system, as shown in Fig. 1, first extracts essential subregions from images, and applies preprocessing and normalization steps, followed by running the hand-crafted and unsupervised feature learning methods. After the system learns the bases, the features are extracted from the testing data. In this section, we will describe feature learning methods that we propose and employ.

2.1. The proposed autoencoder (AE)

We introduce a new encoder-decoder system for unsupervised feature learning. While learning, for given n data samples in R^m represented by matrix $\mathbf{X}=[\mathbf{x}_1,\ldots,\mathbf{x}_n]\in R^{m\times n}$, we want to learn a dictionary $\mathbf{W_d}=[\mathbf{w}_{d_1},\ldots,\mathbf{w}_{d_k}]\in R^{m\times k}$, sparse representation code vectors $\mathbf{Z}=[\mathbf{z}_1,\ldots,\mathbf{z}_n]\in R^{k\times n}$, and latent weight matrix $\mathbf{W_c}$, so that each input sample $\mathbf{x}_j$ can be approximated by $\mathbf{W_d}\mathbf{z_j}$. A nonlinear encoding function $f(\mathbf{x};\ \mathbf{W_c})$ has been used to map $\mathbf{X}\rightarrow\mathbf{Z}$, where $\mathbf{W_c}=[\mathbf{w_{c_1}},\ldots,\mathbf{w_{c_k}}]^T\in R^{k\times m}$. The decoder module reconstructs the input sample approximately by $\mathbf{X}\approx\mathbf{W_d}\mathbf{Z}$. This leads to the following optimization problem over $\mathbf{W_d}$, $\mathbf{Z}$ and $\mathbf{W_c}$:

$$\arg\min_{\mathbf{W_d},\mathbf{Z},\mathbf{W_d}}\frac{1}{2}\|\mathbf{X}-\mathbf{W_d}\mathbf{Z}\|_F^2+\lambda\|\mathbf{Z}\|_1+\frac{\alpha}{2}\|\mathbf{Z}-f(\mathbf{X};\mathbf{W_c})\|_F^2, \tag{1}$$

$$\text{subject to}:\ \|\mathbf{w_{d_i}}\|_2^2\le 1 \text{ for } i=1,\ldots,k,$$

where $\lambda>0$ is a parameter that controls the sparsity of the code vectors (features) and α is a penalty parameter. We consider $\|.\|_F$ and $\|.\|_1$ to represent Frobenius norm and element-wise L_1-norm respectively. In our experiment, we use sigmoid activation function, $f(\mathbf{X};\mathbf{W_c})=(1+exp^{-(\mathbf{W_c}\mathbf{X})})^{-1}$, and set α equals to 1. One can use different nonlinear activation functions, such as, hyperbolic tangent function and rectifier linear unit.

To solve Eq. (1), we propose to use the ADMM form [21], which is used for the convex optimization, to solve the general L_1 regularized loss optimization, and the stochastic gradient descent. $\mathbf{Z}$ is estimated using the ADMM optimization, and $\mathbf{W_d}$ and $\mathbf{W_c}$ are estimated using the stochastic gradient descent. In the ADMM form, the problem can be written as:

$$\text{minimize}:\ f(\mathbf{Z})+g(\mathbf{Y}), \tag{2}$$

$$\text{subject to}:\ \mathbf{Z}-\mathbf{Y}=0, \tag{3}$$

where

$$f(\mathbf{Z})=\frac{1}{2}\|\mathbf{X}-\mathbf{W_d}\mathbf{Z}\|_F^2+\frac{\alpha}{2}\|\mathbf{Z}-f(\mathbf{X};\mathbf{W_c})\|_F^2, \tag{4}$$

$$g(\mathbf{Y})=\lambda\|\mathbf{Z}\|_1. \tag{5}$$

The augmented Lagrangian will be

$$L(\mathbf{X},\mathbf{W_d},\mathbf{W_c},\mathbf{Z},\mathbf{Y})=f(\mathbf{Z})+g(\mathbf{Y})+\frac{\rho}{2}\|\mathbf{Z}-\mathbf{Y}^k+\mathbf{U}^k\|_F^2. \tag{6}$$

Then, the ADMM solution becomes

$$\begin{aligned}\mathbf{Z}^{k+1}&=\frac{1}{2}\|\mathbf{X}-\mathbf{W_d}\mathbf{Z}\|_F^2+(0.5)\|\mathbf{Z}-f(\mathbf{X};\mathbf{W_c})\|_F^2\\&\quad+\frac{\rho}{2}\|\mathbf{Z}-\mathbf{Y}^k+\mathbf{U}^k\|_F^2,\end{aligned} \tag{7}$$

$$\mathbf{Y}^{k+1}=\lambda\|\mathbf{Y}\|_1+\frac{\rho}{2}\|\mathbf{Z}-\mathbf{Y}^k+\mathbf{U}^k\|_F^2, \tag{8}$$

$$\mathbf{U}^{k+1}=\mathbf{U}^k+\mathbf{Z}^{k+1}-\mathbf{Y}^{k+1}. \tag{9}$$

From here, $\mathbf{Z}^{k+1}$ and $\mathbf{Y}^{k+1}$ are estimated using the gradient descent and soft-thresholding [21], respectively. In the same iteration loop, we, then, estimate and update $\mathbf{W_d}$, $\mathbf{W_c}$ using stochastic gradient descent method.

$$\mathbf{W_d}\leftarrow\mathbf{W_d}-\eta_1\nabla_{\mathbf{W_d}}J(\theta), \tag{10}$$

$$\mathbf{W_c}\leftarrow\mathbf{W_c}-\eta_2\nabla_{\mathbf{W_c}}J(\theta), \tag{11}$$

where gradient calculations are given by $\nabla_{\mathbf{D}}J(\boldsymbol{\theta})$ and $\nabla_{\mathbf{W}}J(\boldsymbol{\theta})$ with respect to $\mathbf{D}$ and $\mathbf{W}$ correspondingly.

2.2. K-Means and hand-crafted features

The K-means clustering method obtains specialized bases for the corresponding region of data. Coates et al. [22] proved that the K-means method can achieve comparative or better results than other possible unsupervised learning methods. The algorithm takes the dataset $\mathbf{X}$ and outputs a function f: $R^n\rightarrow R^k$ that maps an input vector $\mathbf{x}$ to a new feature vector of k features. We follow to minimize the following equation:

$$f_a(x)=\max\{0,\mu(q)-q_a\}, \tag{12}$$

where $q_a = \|\mathbf{x} - \mathbf{C}^{(a)}\|_2$ and $\mu(q)$ is the mean of the elements of q. Refer to [22] for more description of this method.

In our system, one of the most powerful hand-crafted feature descriptors is employed to boost the rates. We believe that in addition to original gray-level information, image gradient will also contribute to the multi-model object feature learning and classification. The traditional HOG features are estimated on gradient information of images. We refer to this method hereafter as *HOGgrad*. In our experiments, the HOG and HOGgrad features were obtained every 8 pixels on each image view; and the dimension of each HOG descriptor for an image view is 128.

2.3. Feature extraction

In the unsupervised learning process, we calculate new bases for each method (i.e., K-means and our AE). In the testing stage, the new projected data is calculated using the correlation information between the labeled data and estimated bases.

Let $\mathbf{X}_i$ be any image region and $\mathbf{C}_i$ and $\mathbf{W}_{\mathbf{d_i}}$ are the corresponding bases using the K-means and our AE methods, respectively. The features of labeled data corresponding to image regions are calculated as $\mathbf{Y}_i = \mathbf{X}_i\mathbf{C}_i^t$ (for K-means features) and $\mathbf{A}_i = \mathbf{X}_i{\mathbf{W}_{\mathbf{d_i}}}^t$ (for AE features). Then, the extracted features are combined together one by one to get the multi-model representation as $\mathbf{Y} = [\mathbf{Y}_1; \mathbf{Y}_2; \ldots; \mathbf{Y}_M]$ and $\mathbf{A} = [\mathbf{A}_1; \mathbf{A}_2; \ldots; \mathbf{A}_M]$, where M equals to the number of image region (and sometimes multimedia data such as speech). HOG and HOGgrad features can be represented as $\mathbf{H} = [\mathbf{H}_1; \mathbf{H}_2; \ldots; \mathbf{H}_M]$ and $\mathbf{G} = [\mathbf{G}_1; \mathbf{G}_2; \ldots; \mathbf{G}_M]$. Finally, the feature vector that represents a whole image is represented as $\mathbf{V} = [\mathbf{Y}; \mathbf{A}; \mathbf{H}; \mathbf{G}]$. In our experiment, each method estimates 128 feature units for each image region.

3. Experiments and results

In our experiments, we assess the performance of our proposed method on three data sets: AR [23], Yale [2], and wild PubFig83 [24] data. All images in our experiments are locally normalized to have the Gaussian distribution and whitened as in [22]. In the unsupervised learning part, we train the entire labeled training set of images before the classification step. One of the most important detail is the feature normalization procedure. To be more specific, while each channel-feature and each model-feature are normalized using L_2-norm individually, we observe improved results. We use the linear support vector machine (SVM) for the classification.

3.1. Evaluation on AR face database

The aligned AR database [23] contains 100 subjects (50 men and 50 women), with 14 different images per subject which totals to 1,400 images (excluding the occluded images) taken in two sessions. There are facial expression (neural, smile, anger, scream) and illumination challenges. We segment four essential facial regions with sizes of 39 × 51 (left eye and right eye), 30 × 60 (mouth), and 45 × 42 (nose). We conduct 10 runs for train-test procedure to get the average recognition rate for each partition.

Table 1 presents the detailed experimental results and comparison between our system and some of representative methods. We follow the same framework [22,26] for each method to obtain a fair comparison. We achieved 81.35% and 94.42% recognition accuracy using 2 and 5 training images per subject, respectively. The best results were obtained using the features of K-means, HOG, HOG (Gradient), and the proposed AE. The closest rates were achieved by Wang et al. [28] that are 75.5% (using 2 training and 180 feature units), 94.71% (using 7 training images per subject and 540 feature units). We also assess the feature dimensions of the K-means and AE. Using the 256 units for each method increased the recognition accuracy more than 0.6%.

We assess the response of our method to missing facial region/information as shown in Table 2. Results show that eye regions contains the most effective, important, discriminative information. Missing nose and mouth features decreases the rates around 1.5%, whereas missing both eyes decreases the original rates more than 20%. However, achieving 81.80% should not be underestimated using just nose and mouth regions in one hundred subjects. Shekhar et al. [18] obtained 75.0% recognition accuracy on sun-glass occluded database. Naseem et al. [29] achieved only 26% correct classification rates on subjects that were wearing scarf that closes only mouth region. Table 3 shows the results using various dimensions.

Table 1
Comparison of face recognition rates on *AR* database with some of the representative methods and individual feature learning methods that we use/propose in this paper. In the table *T* represents 'Train'.

Methods	Acc. (%)	
	2 Train	*5 Train*
PCA [25]	34.94	56.13
NPE [25]	40.45	61.12
LPP [25]	55.07	71.58
ONPP [25]	62.20	81.76
EPP [25]	72.45	86.23
Sparse filtering [26]+SVM	63.14	84.56
Coates et al. [22]	65.24	85.56
McDFR [27]	70.92	91.54
Wang et al. [28]	75.50	(94.71) 7*T*
	180 d.	*540 d.*
K-means	75.56	89.40
HOG	71.96	89.67
HOGgrad	67.32	86.60
AE (*128*)	74.60	90.13
AE (*256*)	78.07	91.33
AE (*128*) + K-means	75.23	90.73
AE (*256*) + K-means	78.40	91.87
HOG + HOGgrad	77.20	91.33
K-means + HOG	82.61	93.91
K-means + HOGgrad	83.06	93.42
AE + HOG	82.38	93.40
AE + HOGgrad	80.76	92.26
McMmFL(*128*)	**81.35**	**94.42**
McMmFL(*256*)	**82.12**	**95.04**

Table 2
Classification in *AR* database on missing information.

Missing Region	Acc. (%) with 5 Train
Mouth	93.69
Nose	94.06
Right eye	91.78
Left eye	91.91
Mouth and nose	93.00
Right and left eye	81.80

Table 3
Comparison of face recognition rates on *AR* database with respect to the dimension.

Methods	Acc. (%) with 5 Train			
	32 d.	*64 d.*	*128 d.*	*256 d.*
K-means	86.93	88.18	89.40	89.75
Our AE	85.67	88.87	90.13	91.33
McMmFL	93.71	93.89	94.42	95.04

Table 4
Noise test on *AR*.

dim.	SNR		
	Original	*20db*	*10db*
128	94.42	93.72	86.86
256	95.04	94.83	90.15

Table 5
Learning on multi-region versus whole facial region on *AR*.

Region	Training			
	Input dim.	*Feat. dim.*	*2 T*	*5 T*
Whole Face	8800	1028	68.12	86.26
Multi-region	5679	512	75.23	90.73

Table 6
Comparison of face recognition rates on *Yale* database (See [31,32] for the abbreviations).

Methods	Recognition rate(%)		
	2 T	*4 T*	*8 T*
LocLDA [31]	55.30	73.80	-
PCA [32]	42.63	52.86	64.33
LPP [32]	57.19	75.14	84.11
LPDP [32]	56.74	78.90	90.67
DLPP/MMC [32]	58.19	78.14	89.56
LDA [32]	45.19	68.95	83.22
SNPE1 [32]	66.77	73.61	79.33
DSNPE1 [32]	72.33	86.85	96.00
McDFR [27]	76.58	89.90	97.78
K-means	66.2	82.7	90.22
HOG	73.18	84.38	93.78
HOGgrad	69.94	80.13	89.34
AE	68.87	83.13	91.12
McMmFL	**85.34**	**93.87**	**98.97**

We also test our system on various signal-to-noise ratio (SNR). Table 4 shows the results using images with 20db and 10db SNR versus features with 128 and 256 dimensions. We use the same data trough all steps, i.e., in unsupervised and supervised learning and hand-crafted feature extraction stages. It is observed that the more dimensions the features, the more robust the recognition to the noise.

To explore how the multi-region unsupervised learning extracts more representative features rather than the learning features from the whole facial region. The whole faces are in the sizes of 110 × 80. Since the dimension is bigger than each facial region, we choose to learn 1028 dimensional features. Although the learned feature dimension of whole facial region is doubled, the multi-region technique using AE and K-means achieves much better recognition rates as shown in Table 5.

In terms of the execution time, the proposed AE method learns 128 dimensional features in 334 s whereas the sparse coding method [30] extracts the same dimensional features in 2565 s for one eye region that is in 39 × 51 size.

3.2. Face recognition on Yale database

The Yale database contains 165 images with 15 subjects and 11 frontal images per subject. Each image has one type of facial expressions and configurations. Four essential facial regions are segmented as 40 × 60 (left eye and right eye), 32 × 46 (mouth), and 60 × 48 (nose). The analysis of the experimental results on Yale database is shown in Table 6. We compare the recognition accuracy with various number of training images per subject. For example, K-means obtains 90.22% classification rate, whereas our AE achieves 91.12% when using 8 training images. In the same situation, HOG and HOGgrad get 93.78% and 89.34%, respectively. Perhaps, the less number of training samples for the unsupervised learning methods (i.e., K-means and our AE) should be the reason of the lower classification rates than HOG features. When we combine the features from all techniques, the rate is increased to 98.97%. The closest rate to our results was achieved by Chen et al. [27] that is 97.78%.

Fig. 2. Some example images from the aligned PubFig83 database with various real-world changes on facial expression, pose, illumination, occlusion, resolution, etc.

Table 7
Analysis of face recognition rates on the *PubFig83* database. The unit number is 96 for the unsupervised learning methods.

Methods	Acc. (%)
	90 Train
McDFR [27]	90.14
Chiachia et al. [33]	92.28
McMmFL	**95.87**

3.3. Face recognition on selected pubfig database

Unlike the traditional controlled databases, unconstrained databases contain unrestricted varieties of expression, pose, lighting, occlusion, resolution, etc. We use the PubFig83 database [33] with 83 subjects and at least 100 images per subject. Fig. 2 shows some random images from this data. We randomly select 90 images per subject as the training set, and the rest of the images are used as the testing set in the supervised learning step. The facial regions are in the sizes of 32 × 52 (eyes), 48 × 76 (mouth), and 60 × 48 (nose).

We present our recognition results on Table 7. Chiachia et al. [33] and Chen et al. [27] achieved 92.28% and 90.14% recognition rates, respectively. Chen et al. used the discriminative features learned from the supervised deep neural network. Our system outperforms and achieves 95.87% rate. This comparison also shows that each region and each model contribute unique and discriminative features.

4. Conclusion

We have presented the analysis on multi-channel multi-model feature learning for face recognition. Our experiments verify again that learning features from various techniques and regions boost the classification rates. Although recent convolutional neural network (CNN) techniques that have more than 6 convolutional layers may be the best candidate to achieve the state-of-the-art results on many large scale databases, they have some drawbacks to be used in all applications. One is their time consuming training process that can end up days. Our new AE system can be applied to solve energy formulations with a time and cost efficient parallelized system that will be our one of future search. The other drawback of recent CNNs is that they need high number of samples to avoid over-fitting whereas it can be difficult to find many labeled samples as in this paper. A remedy for this problem can be the transferring the weights of a pre-trained CNN structure.

Acknowledgement

This project is partially Supported by NSF award IIS-1554264.

References

[1] M. Turk, A. Pentland, Eigenfaces for recognition, in: *Journal of Cognitive Neuroscience*, 3:1, 1991, pp. 71–86.

[2] P.N. Belhumeur, J.P. Hespanha, D. Kriegman, Fisherfaces: recognition using class specific linear projection, in: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:7, 1997, pp. 711–720.

[3] T. Ahonen, A. Hadid, M. Pietikainen, Face description with local binary patterns: application to face recognition, in: *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28, 2006, pp. 2037–2041.

[4] D.G. Lowe, Distinctive image features from scale-invariant keypoints, in: *International Journal of Computer Vision*, 60, 2004, pp. 91–110.

[5] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: *Proceedings IEEE Conference on Computer Vision and Pattern Recognition*, 2005, pp. 886–893.

[6] S. Candemir, E. Borovikov, K. Santosh, S. Antani, G. Thoma, Rsilc: rotation-and scale-invariant, line-based color-aware descriptor, in: *Image and Vision Computing*, 42:1, 2015.

[7] A. Bal, Image feature extraction by dynamic neural filtering and phase-only joint transform correlation, in: *Optics and Laser Technology*, 39:1, 2007, pp. 2–7.

[8] X. Chen, X. Lin, Big data deep learning: challenges and perspectives, in: *IEEE Access*, 2:8, 2014, pp. 514–525.

[9] H. Lee, C. Ekanadham, A.Y. Ng., Sparse deep belief net model for visual area v2, in: *Advances in neural information processing systems*, 2008, pp. 873–880.

[10] M. Ranzato, F.J. Huang, Y.-L. Boureau, Y. LeCun, Unsupervised learning of invariant feature hierarchies with applications to object recognition, in: *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, 2007, pp. 1–8.

[11] Q.V. Le, Building high-level features using large scale unsupervised learning, in: *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 8595–8598.

[12] G. Hinton, S. Osindero, Y.-W. Teh, A fast learning algorithm for deep belief nets, in: *Neural Computation*, 2006, pp. 1527–1554.

[13] Y. Bengio, P. Lamblin, D. Popovici, e.a. H. Larochelle, Greedy layer-wise training of deep networks, in: *Advances in neural information processing systems*, 19:153, 2007, pp. 1527–1554.

[14] M. Ranzato, G.E. Hinton, Modeling pixel means and covariances using factorized third-order Boltzmann machines, in: *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 2551–2558.

[15] H. Lee, R. Grosse, R. Ranganath, A.Y. Ng, Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations, in: *Proceedings of the 26th Annual International Conference on Machine Learning*, 2009, pp. 609–616.

[16] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A.Y. Ng, Multimodal deep learning, in: *Proceedings of the International Conference on Machine Learning*, 2011, pp. 689–696.

[17] N. Srivastava, R. Salakhutdinov, Multimodal learning with deep Boltzmann machines, in: *Proceedings of the Neural Information Processing Systems*, 2012, pp. 2231–2239.

[18] S. Shekhar, V.M. Patel, N.M. Nasrabadi, R. Chellappa, Joint sparse representation for robust multimodal biometrics recognition, in: *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 36:1, 2014, pp. 113–126.

[19] M. Nilsback, A. Zisserman, Automated flower classification over a large number of classes, in: *Indian Conference on Computer Vision, Graphics and Image Processing*, 2008, pp. 722–729.

[20] G.B. Huang, H. Lee, E. Learned-Miller, Learning hierarchical representations for face verification with convolutional deep belief networks, in: *Computer Vision and Pattern Recognition, IEEE Conference on*, 2012, pp. 2518–2525.

[21] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, in: *Foundations and Trends in Machine Learning*, 3:1, 2010, pp. 1–122.

[22] A. Coates, A.Y. Ng, H. Lee, An analysis of single-layer networks in unsupervised feature learning, in: *International Conference on Artificial Intelligence and Statistics*, 2011, pp. 215–223.

[23] A. Martınez, R. Benavente, The AR face database, *Computer Vision Center, Technical Report*, 1998.

[24] B.C. Becker, E.G. Ortiz, Evaluating open-universe face identification on the web, in: *Proceedings of IEEE Conf. Computer Vision and Pattern Recognition Workshops*, 2013, pp. 904–911.

[25] F. Zang, J. Zhang, J. Pan, Face recognition using elastic faces, in: *Pattern Recognition*, 4:11, 2012, pp. 3866–3876.

[26] J. Ngiam, Z. Chen, S.A. Bhaskar, P.W. Koh, A.Y. Ng, Sparse filtering, in: *Advances in Neural Information Processing Systems*, 2011, pp. 1125–1133.

[27] X. Chen, M. Aslan, K. Zhang, T. Huang, Learning multi-channel deep feature representations for face recognition, in: *NIPS*, 2015, pp. 60–71.

[28] J. Wang, C. Lu, M. Wang, P. Li, S. Yan, X. Hu, Robust face recognition via adaptive sparse representation, in: *Cybernetics, IEEE Transactions on*, 44:12, 2014, pp. 2368–2378.

[29] I. Naseem, R. Togneri, M. Bennamoun, Linear regression for face recognition, in: *IEEE Trans. Pattern Anal. Mach. Intell.*, 32:11, 2010, pp. 2106–2112.

[30] H. Lee, A. Battle, R. Raina, A.Y. Ng, Efficient sparse coding algorithms, in: Advances in Neural Information Processing Systems, 2006, pp. 801–808.

[31] X. Shu, Y. Gao, H. Lu, Efficient linear discriminant analysis with locality preserving for face recognition, in: *Pattern Recognition*, 45:15, 2012, pp. 1892–1898.

[32] J. Gui, Z. Sun, W. Jia, R. Hu, Y. Lei, S. Ji, Discriminant sparse neighborhood preserving embedding for face recognition, in: *Pattern Recognition*, 45:8, 2012, pp. 2884–2893.

[33] G. Chiachia, A.X. Falcao, A.R. N. Pinto, D. Cox, Learning person-specific representations from faces in the wild, in: *Information Forensics and Security, IEEE Transactions on*, 9:12, 2014, pp. 2089–2099.