PostDoc Journal
Vol. 1, No. 10 October 2013

Journal of Postdoctoral Research
www.postdocjournal.com

Recent progress in face recognition based on sparse coding

Heyan Zhu[1], Shengping Zhang[2*]
(* indicates the corresponding author)

[1]School of Opto-electronic Information, Yantai University, China
[2]Department of Cognitive, Linguistic & Psychological Sciences, Brown University, USA

Abstract: Sparse coding has been attracting increasing interests in computer vision filed, due to its adaptive learning ability and biological inspiration from human vision system. Since sparse representation based classification method for face recognition got great success in 2009, many subsequent improved methods were proposed. In this paper, we aim at providing a comprehensive review of the recent state-of-the-art face recognition methods based on sparse coding. By analyzing their advantages and disadvantages, we summarize the roles of sparse coding in face recognition and discuss the potential improvements in the future.

1. Introduction

Sparse coding was first proposed by [Olshausen and Field, 1996] to model natural image statistics. Given a collection of image patches, sparse coding learns a set of basis functions that are capable of sparsely representing any image patch. The learned basis functions have similar properties with the receptive fields of simple cells in visual cortex. Although sparse coding was proposed in 1996, it has just been attracting increasing interests in signal processing and computer vision since compressive sensing became popular [Donoho, 2006; Candès et al., 2006]. Sparse coding and compressive sensing are two different concepts but have similar formulization. Due to the success of compressive sensing in signal compress, sparse representation also became popular and got success in computer vision. The representative work in computer vision is sparse representation based face recognition, which linearly represents a test face image using training face images from all classes. With assumption that the test face image is from one of the training classes, the representation coefficients should be sparse. Therefore, the representation coefficients can be solved by L1 norm minimization. Then the test face image is classified as the class that reconstructs the test face image using its training face images and the corresponding coefficients with the smallest error.

Although sparse coding based face recognition methods achieve state-of-the-art performance, the role of sparse coding in these successes is still not very clear. Some researchers also validate through extensive experiments that sparse coding may not be as effective as most researchers expected. They found non-sparse representation even outperforms sparse representation in some popular datasets. Therefore, it is quite necessary to provide perspectives from all angles (both positive and negative comments) on the roles of using sparse coding in face recognition. The purpose of this paper is to firstly review related work that achieves success in face recognition and then introduce some negative comments on these work. Finally, we conclude how sparse representation is really useful for face recognition.

The paper is organized as follows: Section 2 introduces the basic principles of sparse coding. Section 3 reviews representative work on face recognition based on sparse representation as well as its variants. Conclusion is presented in Section 4.

2. Basics of sparse coding

To model how visual cortex encodes visual signals received from the external world, a usual assumption is that visual signals are represented by generative models. [Olshausen and Field, 1996] further proposed to factorize visual signals into a linear combination of a set of basis functions, which can be considered as feature vocabulary used to describe the image content. To get feature vocabulary to represent structural features underlying in the image, unsupervised learning can be used to learn a set of basis functions that produce a sparse and independent representation. This method is called sparse coding in natural image statistics field. The basis functions learned from natural images have similar properties with simple cells in visual cortex.

Mathematically, sparse coding finds a set of basis $\mathbf{D} = [\mathbf{d}_1, \mathbf{d}_2, \ldots, \mathbf{d}_k] \in \mathbb{R}^{n\times k}$ to sparsely represent an image

$$\begin{aligned} \mathrm{y} &= \sum_{\boldsymbol{i=1}}^{\boldsymbol{k}} a_i \mathbf{d}_i + \mathbf{v} \\ &= \mathbf{Da} + \mathbf{v} \end{aligned} \tag{1}$$

where $\mathrm{y} \in \mathbb{R}^n$ is the vector with elements being intensities of pixels in the image, a_i is the coefficient of the i-th basis function, $\mathbf{a} = [a_1, a_2, \ldots, a_k]^T \in \mathbb{R}^k$ is the coefficient vector, $\mathbf{v} \in \mathbb{R}^n$ is the noise term.

From the point of view of probabilistic inference, the purpose of sparse coding is to learn a generative model, $p(\mathbf{y}|\mathbf{D})$, to make the image generated from this model follows the true distribution $p^*(\mathbf{y})$. Probability $p(\mathbf{y}|\mathbf{D})$ can be factorized as

$$\begin{aligned} p(\mathbf{y}|\mathbf{D}) &= \int p(\mathbf{y}\,, \mathbf{a}|\mathbf{D})d\mathbf{a} \\ &= \int p(\mathbf{y}|\mathbf{a}, \mathbf{D})p(\mathbf{a})d\mathbf{a} \end{aligned} \tag{2}$$

Given the dictionary $\mathbf{D}$ and coefficient vector $\mathbf{a}$, $p(\mathbf{y}|\mathbf{a}, \mathbf{D})$ is determined by the noise term v. Assuming noise follows Gaussian distribution, $p(\mathrm{y}|\mathrm{a}, \mathrm{D})$ can be computed as

$$\begin{aligned} & p(\mathbf{y}|\mathbf{a}, \mathbf{D}) \\ & = \frac{1}{Z} e^{-\frac{\|\mathbf{y}-\mathbf{Da}\|_2^2}{2\sigma^2}} \end{aligned} \tag{3}$$

where σ^2 is the variance of the noise, Z is the normalization constant, $\|\mathbf{y} - \mathbf{Da}\|_2^2$ is the reconstruction error. The prior $p(\mathbf{a})$ contains the constrains on the representation coefficients. Assuming coefficients are independent statistically, prior $p(\mathbf{a})$ can be factorized as

$$p(\mathbf{a}) = \prod_{i=1}^{K} p(a_i) \tag{4}$$

When constraining the coefficients to be sparse, prior on each coefficient can be computed as

$$p(a_i) = \frac{1}{Z_\beta} e^{-\beta \mathcal{S}(a_i)} \qquad (5)$$

where $\mathcal{S}(a_i)$ determines the shape of the distribution, β controls the shape of the distribution, Z_β is the normalization constant. The dictionary can be computed via maximum a posteriori estimation

$$\mathbf{D}^* = \arg max_{\,\mathbf{D}} \langle \log p(\mathbf{y}|\mathbf{D}) \rangle \qquad (6)$$

Defining energy function $E(\mathbf{y}, \mathbf{a}|\mathbf{D}) = -\log p(\mathbf{y}|\mathbf{a}, \mathbf{D}) p(\mathbf{a})$, Eq.(6) is then equivalent to

$$\mathbf{D}^* = \arg min_{\mathbf{D}} \langle min_{\mathbf{a}} E(\mathbf{y}, \mathbf{a}|\mathbf{D}) \rangle \qquad (7)$$

where the energy function can be further expanded as

$$E(\mathbf{y}, \mathbf{a}|\mathbf{D}) = \|\mathbf{y} - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \sum_{i=1}^{K} \mathcal{S}(a_i) \qquad (8)$$

where $\lambda = 2\sigma^2\beta$. The energy function consists of two parts. The first part is the reconstruction error, which constrains dictionary to reconstruct the input image with minimal error. The second part constrains the coefficients to be sparse. Fig. 1 shows an illustration of the coding process.

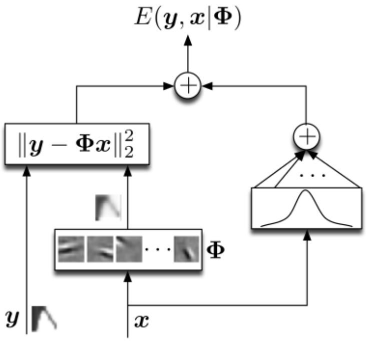
$E(\boldsymbol{y}, \boldsymbol{x}|\boldsymbol{\Phi})$
$\|\boldsymbol{y} - \boldsymbol{\Phi x}\|_2^2$
$\cdots$
$\boldsymbol{\Phi}$
$\boldsymbol{y}$
$\boldsymbol{x}$

Fig. 1. An illustration of sparse coding process.

Given training samples $[\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_m] \in \mathbb{R}^{n \times m}$ and an initial dictionary, the dictionary can be learned by alternating the following two steps:

1) Computing sparse representation coefficients for each training sample using the fixed dictionary

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{y}_i - \mathbf{D}\mathbf{a}\|_2^2 + \lambda \sum_{i=1}^{K} \mathcal{S}(a_i) \qquad (9)$$

2) Updating dictionary

$$\mathbf{D} = \mathbf{D} - \eta \sum_{i=1}^{m} (\mathbf{D}\mathbf{a}_i - \mathbf{y}_i)\mathbf{a}_i^T \qquad (10)$$

3. Face recognition via sparse coding

The most success application of sparse coding in computer vision is face recognition [John et al., 2009]. The idea is very intuitive. The motivation behind this idea is that the face image is inside the subspace spanned by training samples from the same class. An example of illustrating the subspace, coefficients and the reconstructed image is shown in Fig.2.

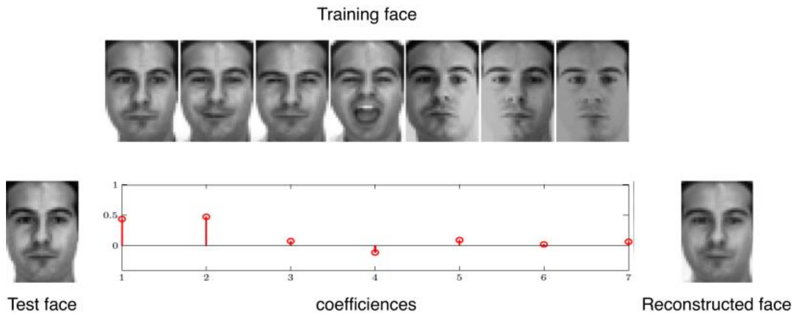

Fig. 2. An illustration of how a face image (left) is in the space spanned by face images (top) from the same class. The coefficients of representing the test face image using all training face images is shown in the bottom. The reconstructed face image is shown in the right.

Let $\mathbf{A}_i = [\mathbf{y}_{i,1}, y_{i,2}, \dots, \mathbf{y}_{i,m_i}] \in \mathbb{R}^{n\times m_i}$ be the matrix consisting of training samples from i-th class and m_i is the number of training samples of class i . Let $\mathbf{A} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m] \in \mathbb{R}^{n\times m}$ be the matrix consisting of training samples from all k classes and $m = m_1 + m_2 +, \dots, + m_k$. For a test sample **y**, to compute which class it is from, we can linearly represent it using training samples from all classes

$$\mathbf{y} = [\mathbf{A}_1, \mathbf{A}_2, \dots, \mathbf{A}_m] \begin{bmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \\ \vdots \\ \mathbf{a}_m \end{bmatrix} = \mathbf{Aa} \qquad (11)$$

where $\mathbf{a}^T = [\mathbf{a}_1^T, \mathbf{a}_2^T, \dots, \mathbf{a}_m^T]$ is the coefficient vector. A reasonable assumption is that the test sample comes from one of the training classes; therefore, the coefficient vector should be sparse. So the coefficient vector can be solved by L1-norm minimization

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{y} - \mathbf{Aa}\|_2^2 + \lambda\|\mathbf{a}\|_1 \qquad (12)$$

Then the reconstruction error when representing y using training samples from class i can be computed as $\|\mathbf{y} - \mathbf{A}_i\mathbf{a}_i\|_2^2$. The test sample $\mathbf{y}$ can finally be classified as the class that corresponds to the minimal reconstruction error. Occlusion is a big challenge for face recognition. The sparse representation based face recognition framework can be easily extended to handle occlusion. If the dictionary $\mathbf{A}$ is replaced by $\widehat{\mathbf{A}} = [\mathbf{A}, \mathbf{I}] \in \mathbb{R}^{n\times(n+m)}$ where $\mathbf{I} \in \mathbb{R}^{n\times n}$ is the identity matrix, the framework described above is capable of handling partial face occlusion. The rational behind is that when the test face image is partially occluded, the occluded part will activate the identity basis in $\mathbf{I}$ to represent it, which will result in the unoccluded part be represented by $\mathbf{A}$. Therefore, the framework described above can still be used for face recognition when only $\mathbf{A}$ and the corresponding coefficients are considered.

The method above uses pixel intensity values as features to describe each face image. From the point of view of discriminability, pixel intensity values are not good features for a classification problem. In addition, pixel intensity is very sensitive to noise. To increase discriminability, [Yang and Zhang, 2010] proposed to use Gabor filters to extract features from each face image and then used the similar sparse representation framework with [John et al., 2009] to perform classification. In particular, the identity basis functions used in [John et al., 2009] are also transformed by Gabor filters and form a compact occlusion dictionary.

The success of the above method relies on the alignment of the face images. In other words, the face needs to be normalized to the same size and positioned at the same position in the image. However, in practical applications, especially when the face image is captured by a hand-held camera, it is difficult to get well-aligned images. Therefore, the above method cannot be directly used in practice. To overcome this drawback, [Wagner et al., 2012] proposed a practical face recognition system based on sparse representation. They used a transformation operator to map the un-aligned face images into well-aligned ones. They then inferred both the transformation and the coefficients in the sparse representation framework. After the transformation and the coefficients are solved, the same classification strategy with [John et al., 2009] can be used to performance classification.

In [Shi et al., 2011], the authors argued that face recognition is not a sparse representation problem. They removed the sparse constraint on the coefficients and get the following least square formulation

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{y} - \mathbf{A}\mathbf{a}\|_2^2 \quad (13)$$

More importantly, the resulting least square problem can be more efficiently solved by

$$\mathbf{a} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y} \quad (14)$$

Similar to [John et al., 2009], after the coefficient vector $\mathbf{a}$ is solved, reconstruction error of using training face images from each class and the corresponding coefficients is computed. The test face image is classified as the class with the smallest reconstruction error.

[Zhang et al., 2011] also questioned that role of sparse constraint in face recognition. They replace the L1 norm constraint in Eq. (12) with the L2 norm constraint. The object function used in their method is

$$\mathbf{a} = \arg\min_{\mathbf{a}} \|\mathbf{y} - \mathbf{A}\mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_2 \quad (15)$$

They called the representation problem defined by Eq. (15) as collaborative representation. Similar to the least square problem, collaborative representation has also an efficient solution

$$\mathbf{a} = \left(\mathbf{A}^{\mathrm{T}}\mathbf{A} + \boldsymbol{\lambda}\mathbf{I}\right)^{-1}\mathbf{A}^{\mathrm{T}}\mathbf{y} \quad (16)$$

They compared the performance between sparse representation and collaborative representation and found that collaborative representation has very competitive classification performance with sparse representation.

It should be noted that the reconstruction error in Eq. (12) is measured using L2 norm, which is obtained by assuming the noise term follows Gaussian distribution. However, in practical applications, this assumption may not be true. For example, in the presence of occlusion and illumination change, the noise term cannot be exactly modeled as Gaussian distribution. To release this unreasonable assumption, [Yang et al., 2011] proposed to model sparse coding as a sparsity-constrained robust regression problem, which seeks for the MLE (maximum likelihood estimation) solution of the sparse coding problem. Their experimental results indicate their model is much more robust to outliers than traditional sparse coding models.

4. Conclusions

Existing face recognition methods based on sparse coding exploit the global classification ability of sparse representation. To reduce computation complexity, the images used as basis functions are down-sampled into a low-dimensional space. Although low-dimensional space makes the computation more efficient, the classification performance is also affected. Little work tries to extract more discriminative while low-dimensional features from images. In addition, sparse coding can also be used to extract features from images. This ability has been widely used in object recognition where the image is firstly densely sampled to get a collection of image patches. Each image patch is sparsely represented by a set of basis functions that are learned from a set of training patches. The sparse coefficients are used as features to describe the appearance of the image patch. Although this idea achieves success in object recognition, it is rarely used in face recognition. Therefore, the research in the future can be along this way.

Reference

[Olshausen and Field, 1996] B. Olshausen and D. Field, Emergence of simple-cell receptive field properties by learning a sparse code for natural images, Nature, 381:607-609, 1996

[Donoho, 2006] D. Donoho, Compressed sensing, IEEE Transactions on Information Theory, 52(4): 1289-1306, 2006

[Candès et al., 2006] E. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements, Communications on Pure and Applied Mathematics 59 (8): 1207-1223, 2006

[John et al., 2009] J. Wright, A. Yang, A. Ganesh, S. Sastry and Y. Ma, Robust face recognition via sparse representation, IEEE Transactions on Pattern Recognition Analysis and Machine Intelligence, 31(2):210-227, 2009

[Yang and Zhang, 2010] M. Yang, L. Zhang, Gabor Feature Based Sparse Representation for Face Recognition with Gabor Occlusion Dictionary, Proc. European Conference on Computer Vision, pp. 448-461, 2010

[Wagner et al., 2012] A. Wagner, J. Wright, A. Ganesh, Z. Zhou, H. Mobahi and Y. Ma, Robust alignment and illumination by sparse representation, IEEE Transactions on Pattern Recognition Analysis and Machine Intelligence, 34(2):372-386, 2012

[Shi et al., 2011] Q. Shi, A. Eriksson, A. van den Hengel and C. Shen, Is face recognition really a compressive sensing problem?, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 553-560, 2011

[Zhang et al., 2011] L. Zhang, M. Yang and X. Feng, Sparse representation or collaborative representation: which helps face recognition?, Proc. IEEE Conference on Computer Vision, pp. 471-478, 2011

[Yang et al., 2011] M. Yang, L. Zhang, J. Yang and D. Zhang, Robust sparse coding for face recognition, Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp. 625-632, 2011