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/

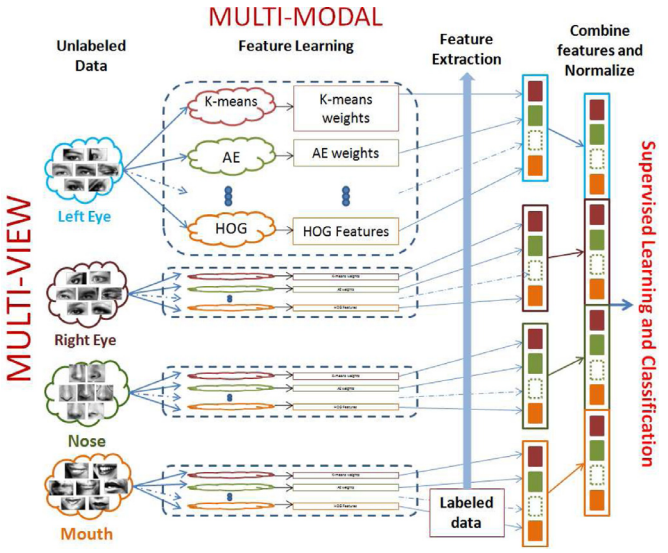

Fig. 1. Architecture of the proposed Multi-Channel Multi-Model feature learning (McMmFL) system.

- 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 $\mathbf{W_d z_j}$. 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 $\mathbf{X} \approx \mathbf{W_d Z}$. This leads to the following optimization problem over $\mathbf{W_d}$, $\mathbf{Z}$ and $\mathbf{W_c}$:

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

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

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, $f(\mathbf{X}; \mathbf{W_c}) = (1 + exp^{-(\mathbf{W_c X})})^{-1}$, 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:

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

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

where

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

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

The augmented Lagrangian will be

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

Then, the ADMM solution becomes

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

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

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

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.

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

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

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:

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