J. Vis. Commun. Image R. 25 (2014) 313–321

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

A robust elastic net approach for feature learning ☆

Ling Wang [a], Hong Cheng [a,*], Zicheng Liu [b], Ce Zhu [a]

[a] University of Electronic Science and Technology of China, 2006 Xiyuan Avenue, Chengdu 611731, China
[b] Microsoft Research Redmond, One Microsoft Way, Redmond, WA 98052, USA

ARTICLE INFO

Article history:
Received 21 August 2013
Accepted 4 November 2013
Available online 16 November 2013

Keywords:
Feature learning
Principal component analysis
Elastic net
Spars representation
Robust statistics
Object recognition
Background reconstruction
Maximum likelihood estimation

ABSTRACT

Unsupervised feature learning has drawn more and more attention especially in visual representation in past years. Traditional feature learning approaches assume that there are few noises in training data set, and the number of samples is enough compared with the dimensions of samples. Unfortunately, these assumptions are violated in most of visual representation scenarios. In these cases, many feature learning approaches are failed to extract the important features. Toward this end, we propose a Robust Elastic Net (REN) approach to handle these problems. Our contributions are twofold. First of all, a novel feature learning approach is proposed to extract features by weighting elastic net. A distribution induced weight function is used to leverage the importance of different samples thus reducing the effects of outliers. Moreover, the REN feature learning approach can handle High Dimension, Low Sample Size (HDLSS) issues. Second, a REN classifier is proposed for object recognition, and can be used for generic visual representation including that from the REN feature extraction. By doing so, we can reduce the effect of outliers in samples. We validate the proposed REN feature learning and classifier on face recognition and background reconstruction. The experimental results showed the robustness of this proposed approach for both corrupted/occluded samples and HDLSS issues.

1. Introduction

Unsupervised feature learning approaches directly from images and videos have drawn more and more attentions in past years [1–6]. Hand-designed features, such as Scale Invariant Feature Transform (SIFT) [7], Histogram of Gradient (HOG) [8], Histogram of Flow (HOF), and Spatial–Temporal Interesting Point (STIP) [9], are successfully applied to visual representa tion especially for object recognition. However, it is really challenging for designing high quality features for visual representation. Recent advancements in feature learning have shown promising results in unsupervised manner from images and videos [10,11].

As the oldest single-layer feature extraction algorithm, Principal Component Analysis (PCA) [12,13] combines the probabilistic, auto-encoder and manifold views of feature learning. Its main idea is to find a projection matrix (a.k.a. loading matrix) that maximizes the variance of a sample set. The PCA can be formulated as a ℓ_2-norm regression type optimization problem. Thus the general PCA can work well under Gaussian assumptions. However, the real samples could be corrupted or occluded thus resulting in high noises, which may violate the Gaussian assumptions. In order to alleviate this problem, ℓ_1-norm PCA approaches are formulated by using Maximum Likelihood Estimation (MLE) to the given samples, where the error between projection space and the original sample is assumed to follow a Laplacian distribution instead of a Gaussian distribution [14–16]. But in practice, this may not work very well especially there are corruptions, occlusions and noises [17]. In this case, the outliers have a very large influence on the ℓ_1-norm residual. Hence, robust statistics have been developed for handling those issues, such as M-estimators [18,19] were proposed to replace the ℓ_2/ℓ_1-norm. In [19], the authors perform a gradient descent approach with local quadratic approximation to solve the M-estimators problem. It is an efficient and robust approach for feature learning, and we will compare it with our approach in the following, so we named it as "FRPCA".

Even though the ℓ_1-norm based PCA and other robust PCA can handle the outliers to some degree, but most of those approaches need to assume that the observation samples are large enough, namely the sample set is non-singular matrix. However, such as in scene monitoring, we only have limited observation numbers which is much less than the number of features. We named it as High Dimension, Low Sample Size (HDLSS) problem. In this case, the traditional eigen-decomposition based PCA methods

☆ The preliminary version of this paper appeared in CCPR2012.
* Corresponding author.
E-mail addresses: eewangling@uestc.edu.cn (L. Wang), hcheng@uestc.edu.cn (H. Cheng), zliu@microsoft.com (Z. Liu), eczhu@uestc.edu.cn (C. Zhu).
URL: http://www.uestcrobot.net (H. Cheng).

1047-3203/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.jvcir.2013.11.002

can not give the proper solutions. Feature learning from HDLSS data is challenging for visual representation [20–26].

The Elastic Net (EN) approach [20] is proposed to solve the this problems. Afterwards, the Sparse Principal Component Analysis (SPCA) [21] is proposed by formulating PCA as the EN regression optimization problem, which can generate modified Principal Components (PCs) with sparse loadings and has the HDLSS asymptotic property [22]. However, since SPCA and EN approaches are based on ℓ_2-norm, they are sensitive to outliers. Fig. 1 illustrates the performance of PCA and its variants on a toy data with outliers. The toy data are some points of a line, but there are several points far away from their correct coordinates. The PCA, ℓ_1-PCA [27], SPCA, FRPCA [19] and the proposed Robust Elastic Net (REN) feature learning approaches are used to find the loading vector of this line. It is shown that the ℓ_1-norm estimator performs better than ℓ_2-norm based PCA and SPCA. The M-estimator based FRPCA and the proposed REN get the best performance. However, the residual error of FRPCA is much bigger than that of REN, which will be shown in our experiments in Section 5. It is also noted that, for the reasons of that the HDLSS problem can not be expressed in 2D space clearly, we just illustrates the robustness of our proposed approach in Fig. 1.

In this paper, we propose a robust elastic net approach for feature learning and classification. First of all, the proposed feature learning approach is formulated as a weighted elastic net problem by using MLE approach. A distribution induced weight function is used to leverage the importance of different samples thus reducing the effects of outliers. Moreover, the REN approach can handle High Dimension, Low Sample Size (HDLSS) issues. Thus, we can extract better visual representation from training samples even in case of occlusions, corruptions and noises. Second, a REN classifier is proposed for object recognition, and can be used for generic visual representation including that from the REN feature extraction. By doing so, we can reduce the effect of outliers in a test sample. Combining the REN feature extraction and classifier can reduce different outliers not only from training samples but also from testing samples. We validate the proposed REN feature learning and classifier on face recognition and background reconstruction. The experiment results illustrated the robustness of this proposed approach especially for both corrupted/occluded samples and HDLSS samples.

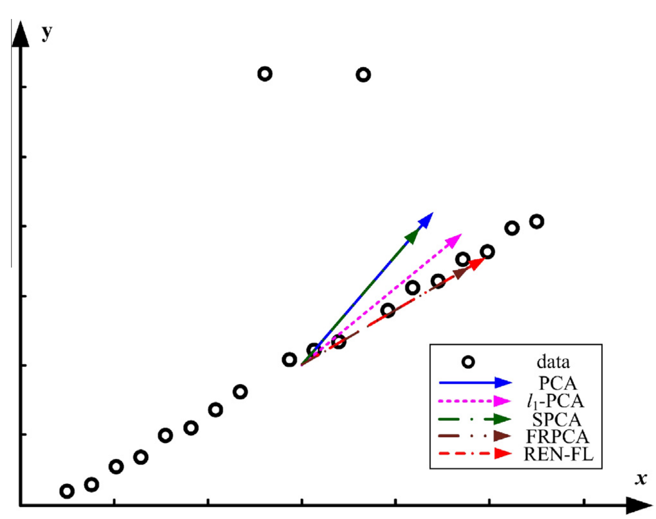

Fig. 1. A toy data with outliers. The 'o' denotes the original data; the lines with arrow denote the first loading vector yielded by PCA, ℓ_1-PCA, SPCA, FRPCA and the proposed REN, respectively.

The rest of this paper is organized as follows. Section 2 reviews the related work. We introduce the proposed REN feature learning and its iterative algorithm in Section 3. Section 4 proposes the REN classifier. Section 5 presents experimental results and analysis. Section 6 concludes this paper and the future work.

Notations

In this paper, the bold letter $\mathbf{X}/\mathbf{e}$ denotes a matrix/vector, the italic lower letter x_{ij}/e_i denotes the element of a matrix/vector, $\mathbf{x}_j$ denotes the jth column of matrix $\mathbf{X}$, $\mathbf{x}^{(i)}$ denotes the ith row of matrix $\mathbf{X}$, $\mathbf{X}_k$ denotes the first k columns of $\mathbf{X}$, the Greek letter α/β denotes a vector. $()^{\mathrm{T}}$ denotes the transpose operations, $\|\cdot\|_p$ denotes p-norm of a matrix/vector.

2. Related work

Unsupervised feature learning approaches, such as random weights hierarchy feature learning approach [2], unsupervised hierarchy feature learning approach [3], and the SIFT [7], HOG [8], STIP [9], provide better performance on feature learning for the properties of translation invariant, scale invariant and spatio-temporal feature invariant. But most of these approaches are hand-designed, need more human intervene. On the other hand, in those multi-level representations, the number of features and the locations will be computed in each layers. To these complexity and expense, Coates et al. analyzed the algorithm performance with the relationship of the number of features [4]. It pointed out that, large numbers of features are critical to achieving high performance using only a single-layer of features. Thus, we analyze the single-layer robust feature extraction approach in this paper. PCA as the oldest single-layer features extraction approach, we first review its main idea and its variants.

Suppose the data matrix is $\mathbf{X} \in \mathbb{R}^{n\times p}$, where n is the number of observations or samples, p is the number of feature or variable dimensions. The Singular Value Decomposition (SVD) of $\mathbf{X}$ is $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^T$, or the eigenvalue decomposition of $\mathbf{X}^T\mathbf{X} = \mathbf{V}\Sigma^2\mathbf{V}^T$. The PCs of $\mathbf{X}$ are defined as $\mathbf{P} = \mathbf{X}\mathbf{V}_k = \mathbf{U}_k\Sigma_k$, which capture the maximum variability of $\mathbf{X}$ and guarantees minimal information loss, where $\mathbf{V}_k \in \mathbb{R}^{p\times k}$ is called as principal loading matrix or projecting matrix. Since the entries of $\mathbf{V}_k$ usually are dense, then the PCs of PCA are the linear combination of all the observations. Thus, the PCA is not the real feature extraction approach. Then, Tibshirani proposed a sparse regression algorithm: LASSO. Following this expression, the $\mathbf{V}$ is sparse and its nonzero entries corresponding to the main features of the $\mathbf{X}$.

Obviously, when $n > p$ and $\mathbf{X}$ is a column full rank matrix, the PCA and LASSO have unique solutions. However, in many scenarios, the observations are limited and the feature dimensions are much more than observation numbers. For example, in background modeling, it usually has $p > n$. In this case, both standard PCA and LASSO do not have unique solutions. They select at most n features that seems to be a limiting feature selection. On the other hand, if a group of features correlated to each other, the PCA and LASSO only randomly select one of them. That will loss some performance on feature extracting. To solve the $p > n$ problem and give the group selection, Zou et al. proposed the EN algorithm [20] and a new sparse PCA (SPCA) approach [21]. Moreover, Zhou et al. extend the EN on the sparse manifold learning [28].

When the observations either be corrupted or be occluded by noise, some robust feature learning and sparse coding approaches are used to solve the outliers. Fernando et al. proposed a fast M-estimation based approach by using gradient iteration algorithm [19]. Wang et al. proposed a kernel-based feature extraction algorithm to handle nonlinear classification problem [29]. But neither