1-1-2010

Dimensionality reduction using compressed sensing and its application to a large-scale visual recognition task

Son Lam Phung
University of Wollongong, phung@uow.edu.au

Fok Hing Chi Tivive
University of Wollongong, tivive@uow.edu.au

Abdesselam Bouzerdoun
University of Wollongong, bouzer@uow.edu.au

Jie Yang
University of Wollongong, jy962@uow.edu.au

Follow this and additional works at: <https://ro.uow.edu.au/infopapers>

Part of the [Physical Sciences and Mathematics Commons](#)

Recommended Citation

Phung, Son Lam; Tivive, Fok Hing Chi; Bouzerdoun, Abdesselam; and Yang, Jie: Dimensionality reduction using compressed sensing and its application to a large-scale visual recognition task 2010, 1607-1614. <https://ro.uow.edu.au/infopapers/807>

Dimensionality reduction using compressed sensing and its application to a large-scale visual recognition task

Abstract

This paper presents a novel algorithm for the dimensionality reduction which employs compressed sensing (CS) to improve the generalization capability of a classifier, especially for large-scale data. Compared to traditional dimensionality reduction methods, the proposed algorithm makes no use of the problem-dependent parameters, nor does it require additional computation for the eigenvalue decomposition like PCA or LDA. Mathematically, the derived algorithm regards the input features as the dictionary in CS, and selects the features that minimize the residual output error iteratively, thus the resulting features have a direct correspondence to the performance requirements of the given problem. Furthermore, the proposed algorithm can be regarded as a sparse classifier, which selects discriminative features and classifies the training data simultaneously. Experimentally, the CS-based algorithm is tested with a hierarchical visual pattern recognition architecture. The simulation results show that not only does the proposed method utilize only 25% of full features while achieving the test accuracy of the original full architecture, but also its performance is competitive when compared to existing dimensionality reduction methods.

Keywords

application, large, scale, visual, dimensionality, recognition, reduction, task, compressed, sensing, its

Disciplines

Physical Sciences and Mathematics

Publication Details

Yang, J., Bouzerdoun, A., Tivive, F. & Phung, S. (2010). Dimensionality reduction using compressed sensing and its application to a large-scale visual recognition task. WCCI 2010 IEEE World Congress on Computational Intelligence (pp. 1607-1614). USA: IEEE.

Dimensionality Reduction using Compressed Sensing and its Application to a Large-Scale Visual Recognition Task

Jie Yang, Abdesselam Bouzerdoum *Senior Member, IEEE*,
Fok Hing Chi Tivive *Member, IEEE*, and Son Lam Phung *Member, IEEE*

Abstract—This paper presents a novel algorithm for the dimensionality reduction which employs compressed sensing (CS) to improve the generalization capability of a classifier, especially for large-scale data. Compared to traditional dimensionality reduction methods, the proposed algorithm makes no use of the problem-dependent parameters, nor does it require additional computation for the eigenvalue decomposition like PCA or LDA. Mathematically, the derived algorithm regards the input features as the dictionary in CS, and selects the features that minimize the residual output error iteratively, thus the resulting features have a direct correspondence to the performance requirements of the given problem. Furthermore, the proposed algorithm can be regarded as a sparse classifier, which selects discriminative features and classifies the training data simultaneously.

Experimentally, the CS-based algorithm is tested with a hierarchical visual pattern recognition architecture. The simulation results show that not only does the proposed method utilize only 25% of full features while achieving the test accuracy of the original full architecture, but also its performance is competitive when compared to existing dimensionality reduction methods.

I. INTRODUCTION

The automated analysis of images in visual recognition has historical interest and becomes popular immediately after the path finding work of [1] in which they capture salient patterns/features within prespecified regions of the visual field. Existing techniques for capturing the primitive features include filter banks [2], statistical analysis [3] or convolutional neural networks (CNNs) [4]. Over the past decade, Bouzerdoum and his colleagues [5], [18], [22] have presented a new class of CNNs inspired by the biophysical mechanism of shunting inhibition, with various systematic connection schemes and training algorithms. The proposed architecture has a generic structure in which shunting inhibitory neurons are trained to capture visual features from the images. These convolutional networks have been employed in face detection [6] and gender recognition [7].

How to make use of the features properly plays a fundamental role in the visual recognition process. Usually, determining the valid features is problem-dependent and requires specialized prior knowledge. The problem, nevertheless, seems to be in an awkward predicament in that it is difficult to find the balance between the number of required features and the model performance. A too large number of features would not only result in overfitting and

poor generalization but also increasing the computational overhead. On the other hand, few features may not be sufficient for the generalization ability due to some lack of information.

Therefore, dimensionality reduction approaches are required, and the reason is obvious in that fewer features, with respect to (w.r.t.) the number of original ones, have more advantages in addition to better generalization. For example, it costs usually less computational complexity to build the classifier. A variety of methods have been developed to deal with the dimensionality reduction problem, which can be broadly categorized as feature extraction [8] and feature selection [9] algorithms. Traditionally, feature extraction is more effective than feature selection due to the algebraic transformation adapted in feature extraction algorithms; However, feature extraction algorithms may break down in high dimensional problems since they typically assign nonzero loading to all the features; By contrast, the computational requirement for feature selection algorithms is always much lower than that of extraction algorithms since they usually depend on a binary transformation [10]. More importantly, most feature extraction approaches, such as principal component analysis (PCA) and linear discriminant analysis (LDA), involve eigenvalue decomposition, which is extremely time-consuming. In addition, if the dimension exceeds the number of training data (smaller samples), LDA may encounter a singularity problem due to the matrix inversion.

To overcome these difficulties, we propose a novel algorithm, termed compressed sensing-based dimensionality reduction (CDR), for feature selection. Its advantage benefits from the compressed sensing (CS) [11], [12], [13] theory, which addresses the sparse signal representation. The CS method can help in recovering signals that have a sparse representation from a number of measurements/projections of dimensionality lower than the number of samples required by the Shannon/Nyquist sampling theory. In [21], the author proposed a pruning-like algorithm for neural networks via CS in which the most significant element (weights and hidden neurons) has been selected iteratively. Similarly, if we consider the discriminative features as a sparse subset of the full features, then CS can be employed to find out those required features. In this context, the proposed algorithm is the extensive enhancement of [21] toward the feature selection.

The validity of our approach is supported by a visual recognition system which is based on a combination of fixed and adaptive filters. The fixed filters are used to detect

Jie Yang, Abdesselam Bouzerdoum, Fok Hing Chi Tivive, and Son Lam Phung are with the School of Electrical, Computer and Telecommunications Engineering, and ICT Research Institute, University of Wollongong, Wollongong, Northfields Avenue, NSW 2522, Australia (email: {jy962, bouzer, tivive, phung}@uow.edu.au).

primitive features that are common to most visual recognition tasks. The adaptive filters are tuned to find out features specific to a type of visual objects. At the dimensionality reduction stage, the CS-based algorithm is employed to detect the most discriminative features. In this context, the process of selecting features can be thought of as finding out the one with the maximal correlation between the residual output error and the candidate features.

Fig. 1. An illustration of the proposed system for visual recognition tasks. From (a), the original images have been collected which are processed by an adaptive hierarchical network in (b) using a variety of fixed and adaptive filters to detect the primitive features. In (c) the novel dimensionality reduction algorithm (CDR) is employed to select those discriminative features that can be processed by a classifier in (d).

The remainder of the paper is organized as follows. The next two sections present a brief review of compressed sensing theory and existing dimensionality reduction algorithms. Then Section IV details the development of the CDR algorithm by describing the link between selecting discriminative features and the sparse representation in CS. The implementation issues and experimental results are discussed in Section VI, followed by concluding remarks.

II. SPARSE REPRESENTATION USING COMPRESSED SENSING

This section discusses the theory of compressed sensing, a novel sensing/sampling paradigm that is different from the Shannon/Nyquist sampling theory. This theory supposes that if we allow for a degree of residual error ϵ , CS guarantees the successful recovery of the given signal from a number of projections (measurement vectors/samples). We refer the

reader to [11], [12], [13] for a more detailed discussion on CS and its wide applications.

According to the number of measurement vectors, the CS problem can be categorized into Single-Measurement Vector (SMV) [14] or Multiple-Measurement Vectors (MMV) [15]. Mathematically, the SMV problem is expressed as follows. Given a measurement vector $\mathbf{y} \in \mathbb{R}^m$ and a dictionary $\mathcal{D} \in \mathbb{R}^{m \times n}$ (the columns of $\mathcal{D}$ are referred to as the atoms), we seek a vector solution satisfying:

$$(P): \min \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = \mathcal{D}\mathbf{x} \quad (1)$$

where $\|\mathbf{x}\|_0$ (known as l_0 -norm) is the cardinality or number of nonzero elements in $\mathbf{x}$. Several algorithms have been developed including greedy and non-convex local optimization algorithms.

- Greedy algorithms, such as orthogonal matching pursuit (OMP) [14], select the atom that minimizes the residual error at each iteration;
- Non-convex local optimization methods, namely FOCUSS algorithm [16], on the other hand, focus on a re-weighted minimum norm rather than l_0 -norm.

In this paper, we denote all optimization algorithms for the SMV problem as SMVALG. In particular, the convergence property of SMVALG is demonstrated by the following theorem:

Theorem 1: [13] For any vector $\mathbf{y}$, there exists a time function $\beta(t) \in (0, 1)$, which depends only on the dictionary $\mathcal{D}$, such that the residual error calculated by SMVALG decays as: $\|\mathbf{r}_t\|^2 \leq \beta(t) \|\mathbf{r}_{t-1}\|^2$, where $\mathbf{r}_t = \mathbf{y} - \mathcal{D}\mathbf{x}_t$ is the reconstruct error after t iterations. The upper limit for the residual error is given by $\|\mathbf{r}_t\|^2 \leq \beta(t) \|\mathbf{r}_{t-1}\|^2 \leq \dots \leq \prod_{i=1}^t \beta(i) \|\mathbf{r}_0\|^2$.

While the SMV problem is finding a sparse signal representation, the MMV problem looks for a joint-sparse representation of several signals, or a sparse matrix representation [15]. The MMV can be stated as follows (to avoid confusion, we use the upper-case letters $\mathbf{Y}$ and $\mathbf{X}$ to denote the matrices):

$$(Q): \min \|\mathbf{m}(\mathbf{X})\| \quad \text{s.t.} \quad \mathbf{Y} = \mathcal{D}\mathbf{X} \quad (2)$$

Where $\mathbf{Y} \in \mathbb{R}^{m \times k}$, $\mathbf{X} \in \mathbb{R}^{n \times k}$ and $\mathbf{m}(\mathbf{X})$ denotes the matrix norm. Methods used in SMV (such as OMP) have been extended to solve the above problem such as the OMPMMV Algorithm [15]. Without loss of generality, we denote them as MMVALG in this paper. Theoretically, the convergence results for MMVALG algorithms have been analyzed which shows that the norm of the residual error decays with the increase of the number of the selected atoms:

Theorem 2: For any MMVALG algorithms, the objective function $\|\mathbf{Y} - \mathcal{D}\mathbf{X}_t\|_F^2$ decreases, i.e. $\|\mathbf{Y} - \mathcal{D}\mathbf{X}_t\|_F^2 = k \|\mathbf{Y} - \mathcal{D}\mathbf{X}_{t-1}\|_F^2$, where $k \in (0, 1)$ is a constant.

III. DIMENSIONALITY REDUCTION

The dimensionality reduction approach, trying to maintain the performance of the classifier while reducing the number

of required features, is a fundamental technique for the large-scale data processing. We consider N training samples with d features $X \in \mathbb{R}^{d \times N} = \{\mathbf{x}_i | i \in [1, N]\}$, the dimensionality reduction problem can be mathematically stated as the problem of finding a transformation matrix $W \in \mathbb{R}^{p \times d}$, where p is the dimension of data after dimensionality reduction ($p \ll d$), by taking into account some optimization criteria $J(W)$. Thus, according to different objective functions $J(W)$, many state-of-the-art dimensionality reduction approaches fall into two categories: feature extraction and feature selection:

- Feature extraction (FE) aims to extract features by projecting the original high-dimensional data to a lower-dimensional space through an algebraic transformation;
- Feature selection (FS) finds out a subset of the most representative features using a binary transformation matrix.

A. Feature extraction

Here we consider two popular algorithms, PCA and LDA, for FE. PCA aims to determine a linear combination of the input variables that maximizes the variance. Let $C = \frac{1}{N} \sum_{i=1}^N (\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})^T$ be the covariance matrix of X , where $\bar{\mathbf{x}}$ is the mean vector of the data. Then, PCA aims to maximize the objective function:

$$J(W) = \text{trace}(W^T C W) \quad (3)$$

and return p first eigenvectors of C to form W .

On the other hand, LDA is used to find the feature space in which

$$J(W) = \text{trace} \left[(W^T S_W W)^{-1} (W^T S_B W) \right] \quad (4)$$

is maximized. The within-class scatter matrix $S_W = \sum_{i=1}^c P_i C_i$ while the between-class scatter matrix S_B is defined as $S_B = \sum_{i=1}^c P_i (m_i - \bar{\mathbf{x}})(m_i - \bar{\mathbf{x}})^T$, where c is the number of classes, P_i denotes the prior probability of class i , C_i is the covariance matrix of class i , and m_i is the mean vector of class i .

A limitation of PCA and LDA is that both of them require the eigenvalue decomposition, which is extremely time-consuming to compute. In addition, if the dimension exceeds the number of training data (smaller samples), LDA may encounter a singularity problem.

B. Feature selection

In contrast to the FE algorithms, FS algorithms are much more efficient since their aim is to detect a subset of the most representative features from the original dataset instead of generating new features. Normally, all the features are ranked according to some criteria; then a greedy strategy is employed to select those features with higher rank. Mutual information (MI) [17] is one of most popular approaches. According to Shannon's probability and information theory, MI is used as a similarity measurement between two random variables. In terms of feature selection, the relevance of the desired output w.r.t the feature matrix X can be calculated

using MI. Mathematically, the uncertainty of a given output class Y can be measured by the entropy:

$$H(Y) = - \sum_{c=1}^{N_c} p(c) \log(p(c)) \quad (5)$$

where $p(c)$ denotes the prior probability of class c . Given a feature vector $\mathbf{x}$ from X , the remaining uncertainty of Y becomes the conditional entropy:

$$H(Y|\mathbf{x}) = - \int p(\mathbf{x}) \left(\sum_{c=1}^{N_c} p(c|\mathbf{x}) \log(p(c|\mathbf{x})) \right) d\mathbf{x} \quad (6)$$

where $p(c|\mathbf{x})$ is the conditional probability for class c given the feature $\mathbf{x}$. In general, the Mutual Information between Y and $\mathbf{x}$ is defined as:

$$I(Y; \mathbf{x}) = H(Y) - H(Y|\mathbf{x}) \quad (7)$$

Based on the above process, while selecting the discriminative feature at t th iteration, we calculate MI for all the remaining features w.r.t the desired output Y and choose the one with the maximal MI value. That is to say, the X^{i^*} feature is selected subject to $i^* = \arg \max I(Y; \mathcal{S}_{t-1} \cup X^{i^*})$, where the subset $\mathcal{S}_{t-1}$ denotes the selected features until $(t-1)$ th iteration, and $\mathcal{S}_0 = \emptyset$ is the initial empty set.

IV. PROBLEM FORMULATION AND DERIVATION OF CS-BASED ALGORITHM

In this section, we formulate the problem of dimensionality reduction as a sparse representation in compressed sensing. Before explaining the main idea, it is necessary to introduce some basic definitions.

Definition 1: Given a vector $\mathbf{x}$, its *sparse rank* $VS(\mathbf{x})$ is the number of non-zero entries in $\mathbf{x}$, i.e., $VS(\mathbf{x}) = \|\mathbf{x}\|_0$.

Definition 2: Given a matrix X , its *sparse rank* $MS(X)$ is the number of non-empty rows in X . A non-empty row must have at least one non-zero entry.

Without loss of generality, consider the traditional classification problem where the training data consists of N pairs $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^N$, and d and m are the dimensions of the input $\mathbf{x}_i$ and output $\mathbf{y}_i$ respectively. The aim then is to extract some decision rules, say

$$\mathbf{y}_i = f(\mathbf{x}_i) + \mathbf{e}_i \quad (8)$$

where $f: \mathbb{R}^d \rightarrow \mathbb{R}^m$ is an unknown smooth vector valued function and $\mathbf{e}_i$ is the corresponding error. In general $f(\mathbf{x}_i)$ may be expressed as

$$f(\mathbf{x}_i) = W\varphi(\mathbf{x}_i) \quad (9)$$

where W is an $m \times n$ weight matrix and $\varphi(\mathbf{x}_i)$ is an n -dimensional feature vector. The mapping $\varphi(\cdot): \mathbb{R}^d \rightarrow \mathbb{R}^n$ is usually learned or approximated from the training data. In the case of support vector machines (SVMs), it represents the user-defined kernel function, whereas in feedforward neural networks, the mapping $\varphi(\cdot)$ is generated by the hidden layers of the network. Now consider the training data set consisting of the input data matrix X , where each column represents an

input vector $\mathbf{x}_i$, and the corresponding desired output matrix Y , then Eq. (8) can be written in matrix form as follows:

$$Y = W\varphi(X) + E \quad (10)$$

The main problem with a classifier that uses Eq. (10) is that all n features in $\varphi(\mathbf{x}_i)$ contribute to the final output even though some of them may be redundant or irrelevant. When the dimensionality reduction approaches are considered, a sparse classifier is built with lower dimension (only part of $\varphi(\mathbf{x}_i)$ is considered). Normally, the dimensionality reduction approaches are applied to the input $\mathbf{x}_i$. In this paper, the proposed method aims to reduce the dimension of $\varphi(\mathbf{x}_i)$, that is, find a representative subset from $\varphi(\mathbf{x}_i)$ instead of $\mathbf{x}_i$ while minimizes the training error. Because we believe that this has a more direct influence on the output data. Obviously, the method can also be employed to the input data by simply replacing $\varphi(X)$ with X due to the simple mapping relationship between $\varphi(X)$ and X .

Note that each column in W corresponds to a certain row vector in $\varphi(X)$ which now is a feature vector. Consequently, if only few columns in W consist of nonzero elements, then their corresponding row vectors in $\varphi(X)$ are regarded as the required features while the remaining ones are less informative to the classifier.

Based on the above discussion, the problem of dimensionality reduction can be formulated as minimizing the following equation:

$$\min_W S(W) \quad s.t. \quad \|Y - W\varphi(X)\|_F^2 \leq \xi \quad (11)$$

where $S(W)$ is a sparsity measure of W , ξ is the given minimal error, and $\|\cdot\|_F^2$ denotes the frobenius norm. Note that $\|Y - W\varphi(X)\|_F^2 = \|Y^T - [\varphi(X)]^T W^T\|_F^2$, then Eq. (11) can be rewritten as follows:

$$\min_W MS(W^T) \quad s.t. \quad \|Y^T - [\varphi(X)]^T W^T\|_F^2 \leq \xi \quad (12)$$

Comparing the modified optimization problem with Eq. (2), we can see that the original feature matrix $[\varphi(X)]^T$ serves as the dictionary $\mathcal{D}$ and the desired output Y^T plays the role of the signal matrix in CS. By doing so, the process of feature selection can be regarded as finding sparse solutions for the weight matrix W . Therefore, we propose a CS-based dimensionality reduction method (Algorithm 1), which combines the ideas of sparse representation from CS and feature selection; it is capable of fully exploiting all information available for locating the most representative features.

Remark 1: Consider the one-output case $\mathbf{y}$, the problem in Eq. (12) is converted to solving a sparse solution for the weight vector $\mathbf{w}$, i.e.

$$\min_{\mathbf{w}} VS(\mathbf{w}^T) \quad s.t. \quad \|\mathbf{y}^T - [\varphi(X)]^T \mathbf{w}^T\|_F^2 \leq \xi \quad (13)$$

Similarly, we can employ SMVALG algorithms for the above optimization, i.e., $(\mathbf{w}^T) \leftarrow \text{SMVALG}(\mathbf{y}^T, [\varphi(X)]^T)$. Correspondingly, we have $p = VS(\mathbf{w}^T)$.

input : The desired matrix $Y \in \mathbb{R}^{m \times N}$, the feature matrix $[\varphi(X)] \in \mathbb{R}^{n \times N}$ and the maximal iteration M

output: A transformation matrix W

for $i \leftarrow 1$ **to** M **do**
 $(W)^T \leftarrow \text{MMVALG}(Y^T, [\varphi(X)]^T)$
Check the termination criterion
end
 In total $p = MS(W^T)$ features have been selected.

Algorithm 1: CS-based algorithm for dimensionality reduction (CDR).

Remark 2: We employ the OMPMMV [15] algorithm for solving Eq. (12) while the OMP [14] algorithm is applied to (13) because of their simplicity and flexibility. Most importantly, these two methods are iterative-like, which means we can control the sparsity of the solution, i.e. how many features are selected. Because both these two OMP-based algorithms only select one atom at each iteration, and they do not select the same atom (feature) twice. Consequently the number of features can be determined as t at the t th iteration.

Remark 3: The proposed CDR algorithm can be regarded as a sparse linear classifier. Note that a linear classifier (LC) is normally achieved based on a linear combination of the input features. That is to say, given the training data $\{X, Y\}$, the linear classifier, with $\varphi(X) = X$, aims to solve the following problem;

$$V = \arg \min_V \|Y - VX\|_F^2 \quad (14)$$

where V is the weight matrix. If we compare LC with CDR, it is easy to see that the CDR-based feature selection methods share the same procedure for calculating the corresponding weight matrix. The only difference lies in how many features are involved since the LC algorithm takes all the existing features into account while the CDR method depends on the selected discriminative features. In other words, CDR can be regarded as a sparse case of linear classifier with the constraint on the number of selected features. One of its advantages is that the presented algorithm can select important features and classify the training data simultaneously. On the other hand, existing algorithms require the calculation of output weight matrix for these new features. In this context, CDR saves in computational cost since the output weight matrix can be obtained directly from the inner product between the features and the desired output.

Remark 4: In our paper, the model generalization capability has been considered as the termination criterion. One common approach is to estimate the validation errors during selection of features until some stopping criterion is satisfied. Here the CDR algorithm is terminated when the validation error keeps on increasing for T successive times, where T is a parameter given by the user. The decision is made based on

the assumption that an increase in error from the validation set indicates the beginning of overfitting.

V. AN ADAPTIVE HIERARCHICAL ARCHITECTURE FOR FEATURE DETECTION

The proposed feature detection method has two stages comprising fixed, directional filters and adaptive filters, which are arranged in a hierarchical structure to detect features from input images, as shown in Fig. 2. These filters have nonlinear characteristics as they are based on the mechanism of shunting inhibition [18], [20], [22]. The first stage is designed to detect primitive features, such as orientations or edges that are present in a wide range of objects, whereas the second stage is emphasized on detecting more specific, salient features for a given problem by taking the kernels of the adaptive filters into account.

Fig. 2. An overview of the two-layer hierarchical structure for detecting the possible features from the original images.

A. Architecture for feature detection

To capture elementary features from the input, the filter in the first stage is designed as a directional filter, whose output response is given by

$$Z_{1,i} = \frac{D_i * I}{G * I}, \quad (15)$$

where I is a 2-D input pattern, $Z_{1,i}$ is the output of the i -th filter, D_i and G are the filter coefficients, and “*” denotes 2-D convolution. The convolution mask D_i is defined as a first derivative Gaussian kernel. For a given direction θ_i , the first-order derivative Gaussian kernel is defined as

$$D_i(x, y) = \cos(\theta_i) G'_x(x, y) + \sin(\theta_i) G'_y(x, y), \quad (16)$$

where

$$G'_x(x, y) \partial G(x, y) / \partial x = \frac{-x}{2\pi\sigma^4} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (17)$$

and

$$G'_y(x, y) \partial G(x, y) / \partial y = \frac{-y}{2\pi\sigma^4} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right). \quad (18)$$

The other convolution mask G is simply chosen as an isotropic Gaussian, which is given by

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right). \quad (19)$$

Each filter is associated with an angle θ_i , where $\theta_i = \frac{(i-1)\pi}{N_1}$ for $i = 1, 2, \dots, N_1$, and the number of filters, N_1 , for Stage 1 is chosen according to the complexity of the given problem.

The next processing step is motivated by the center-surround receptive fields that take place in the early stages of the mammalian visual systems. There are two major configurations: on-center and off-center. Accordingly, we separate each sub-sampled map $Z_{1,i}$, where $i = 1, 2, \dots, N_1$, into an on-response map and an off-response map, using zero as the threshold:

$$Z_{1,i} \rightarrow \begin{cases} \text{on response map : } Z_{2,2i-1} = \max(Z_{1,i}, 0) \\ \text{off response map : } Z_{2,2i} = -\min(Z_{1,i}, 0) \end{cases}. \quad (20)$$

Essentially, for the on-response map, all negative entries are set to 0, whereas for the off-response map, positive entries are set to 0 and the entire map is then negated. At the end of Stage 1, the features in each map are normalized, using the following transformation

$$Z_{3,i} = \frac{Z_{2,i}}{Z_{2,i} + \mu}. \quad (21)$$

where μ is the mean value of the absolute response of the output map of the directional filter.

The aim of Stage 2 is to detect more specific features from the preceding stage that will simplify the classification task. Hence, each directional filter in Stage 1 is linked to two filters in stage 2: one filter is used to process the on-response and the other filter for the off-response. Therefore, Stage 2 has twice the number of filters in Stage 1. Similarly, the filters in Stage 2 are also based on the shunting inhibition mechanism. Consider an input map $Z_{3,i}$ to Stage 2. The filter output is calculated as

$$Z_{4,i} = \frac{g(P_k * Z_{3,i} + \mathbf{b}_k) + \mathbf{c}_k}{\mathbf{a}_k + f(Q_k * Z_{3,i} + \mathbf{d}_k)}, \quad (22)$$

where P_k and Q_k are the adaptive kernels for the filter, $\mathbf{a}_k$, $\mathbf{b}_k$, $\mathbf{c}_k$ and $\mathbf{d}_k$ are adjustable bias terms, and f and g are activation functions.

B. Training method

A variety of training methods can be employed to tune the adaptive filters in Stage 2, such as the error-backpropagation technique. Without loss of generality, the levenberg-marquardt (LM) algorithm has been implemented and the feature vectors from Stage 2 are inputted to a linear classifier, whose output Y is given as

$$Y = W Z_{4,i} + \mathbf{b}, \quad (23)$$

where W is the adjustable weight matrix, $\mathbf{b}$ is an adjustable bias term, $Z_{4,i}$ is the feature matrix by concatenating all the

features produced from Stage 2. Given that all parameters of the adaptive filters and the linear classifier are arranged as a column vector, $W = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_N]^T$. Then the LM algorithm is applied here to calculate the weight update at iteration t as follows:

$$\Delta \vec{W}(t) = [J^T(t)J(t) + \mu(t)\Psi]^{-1} J(t)^T E(t),$$

where $J(t)$ is the Jacobian of the error function $E(t)$, Ψ is the identity matrix and $\mu(t)$ is a regularization term to avoid the singularity problem.

VI. PEDESTRIAN DETECTION

To investigate the efficiency of the proposed CDR algorithm for feature selection, we employ the architecture described in the previous section to detect visual features for the pedestrian detection problem. We show the comparison between the CDR-based system and the original one with full features. Furthermore, the proposed method is compared with other existing FE and FS methods.

A. Classification problem

Pedestrian detection has many applications in traffic safety, law enforcement and car industry. Its objective is to determine the presence and the location of people or pedestrians in images and video. In recent years, detecting pedestrian has attracted significant research interest as the pedestrian appearance can change drastically due to clothing, lighting conditions, and walking poses. Munder and Gavrilu [19] analyzed several classifiers for pedestrian detection, tested on the Daimler-Chrysler database. These classifiers include neural networks (NNs), support vector machines (SVM), and adaptive boosting (AdaBoost) classifier. The Daimler-Chrysler database consists of three training sets, labeled as 1, 2, and 3, and two test sets, labeled as T1 and T2. Each set has 4800 pedestrian patterns and 5000 non-pedestrian patterns; each image pattern has the size of 36×18 pixels. Some exemplars of pedestrian and non-pedestrian are shown in Fig. 3.

Fig. 3. Image patterns from the Daimler-Chrysler pedestrian detection database.

The configuration of the feature detection model is obtained using 1 and 2 as training set and 3 as validation set. The Stage 1 has 9 directional filters with a kernel size of 7×7 , whereas Stage 2 has 18 filters with a kernel size of 5×5 . The activation functions f and g for Stage 2 are chosen as hyperbolic tangent and exponential functions, respectively. The feature vector generated at Stage 2 has 2016 dimensions.

B. CDR vs. the original architecture

Before digging into too much detail, an overview of our proposed framework may be helpful. At first, we investigate how the proposed algorithm influences the generalization ability by utilizing the selected significant features rather than all the available ones. Methodologically, the CDR algorithm is directly applied after the original structure has been trained. In [22], the original structure was compared with existing pedestrian detection methods as reported in [19], such as the SVM classifiers using PCA (84.2%), Haar (86.2%) and LRF (89.8%) features. The architecture from [22] achieves better performance. Here the comparison is emphasized between the CDR-based system and the full architecture.

TABLE I

THE COMPARISON BETWEEN THE ORIGINAL ARCHITECTURE WITH THE CDR-BASED ALGORITHM FOR THE PEDESTRIAN DETECTION.

$p = 100$		$p = 300$	
Acc (Train)	Acc (Test)	Acc (Train)	Acc (Test)
93.27%	87.52%	96.65%	90.94%
$p = 500$		$p = 1000$	
Acc (Train)	Acc (Test)	Acc (Train)	Acc (Test)
97.57%	91.76%	98.26%	91.86%

Note **Acc(Train)** and **Acc(Test)** represent the accuracy on Training and Test Datasets respectively. p is the number of the selected features from the CDR algorithm. The original architecture achieves 99.70% and 91.80% for training and test data, respectively.

Table I presents the average results from the proposed algorithms over 10 independent runs using the linear classifier. Several observations can be made from these results: firstly, with the increase in the number of the selected features from the CDR algorithm, the performance is improved in terms of the training and test accuracy. On the other hand, the average test errors obtained from the CDR-based algorithm is comparable to that of the full architecture, especially when 500 or 1000 features are selected using the CDR approach. Although the average training errors are larger than that of the initial structure, the proposed algorithm is built on fewer features. For example, the compact architecture can classify the test data with 91.76% accuracy while considers only 500 features. In other words, when CDR is employed, the same performance can be obtained with 75% of the features are ingored.

To investigate how the proposed algorithm can be applied to other classifiers, Fig. 4 represents the performance when NNs are considered as the classifier with different numbers of hidden neurons. The training parameters in NNs are manually tuned as follows: the maximum number of epochs for training is 100; the minimum performance gradient is $1e - 6$, and The activation function for the hidden neurons was set to the standard sigmoid function and a linear transfer function for the output neuron.

We find from Fig. 4 that when fewer features have been considered, some useful information may be lost and this

Fig. 4. The comparison in terms of classification accuracy between the original architecture with the CDR-based algorithm using the NNs;

results in worse performance. For example, when only 100 features were used, CDR achieves 87.52% on average. Obviously, more accurate performance requires more features. Thus, as we can see from the simulations, the CDR algorithm obtains on average 90.30%, 91.21%, and 90.60% when 300, 500 and 1000 features were used, respectively. Note that using the same NNs classifier, the original architecture with 2016 features only achieves 90.37% on average.

On the other hand, it is quite interesting to notice that all the performances drop when NNs are employed in the final classification stage. One possible reason is that a more powerful classifier (such as NNs) may overfit the training data and lead to a decrease in the testing accuracy. However, the proposed CDR algorithm is relatively robust compared to the original system. For example, in terms of the CDR-based system, only 0.55% accuracy has been lost using 500 features while the one with the entire features degrades 1.43%. Thus, we illustrate that the CDR-based feature selection method is capable of improving the generalization ability of the visual recognition system using a more compact feature representation. Fig. 5 gives the corresponding illustration obtained by the CDR-based pedestrian detection system.

Fig. 5. The pedestrian detection outputs for some test images. The detection score has been highlighted.

C. Comparison with state-of-the-art methods

In this section, we compare the proposed approach with some popular algorithms such as PCA, LDA for FE and MI for FS. The extracted features from these algorithms are then trained using a linear classifier. The different algorithms are evaluated in terms of the generalization ability on the test data and memory requirement. Fig. 6 represents the average performance with respect to the number of features for 10 independent runs. To achieve better generalization ability, we have to employ more features.

Fig. 6. Performance of four different dimensionality reduction algorithms: PCA, LDA (from FE) and MI, CDR (from FS) via the pedestrian detection problems.

Normally, we cannot compare the FS methods with the FE since the latter methods take advantages of all the available features to form the new features. However, the FS methods, depending on a binary transformation matrix, only select a subset of the original features. It is expected that FS may achieve worse performance since less information has been considered. It is one of the reasons why MI performs poorly compared to other three algorithms.

An interesting observation is that the proposed CDR algorithm outperforms the LDA and MI. In fact, as we can see from Fig. 6, from the very beginning until the number of selected features is up to 400, CDR achieves the best generalization ability (91.59%). Although PCA achieves a better result (91.81%) when using 500 features, CDR is still competitive (91.76%) with the PCA. However, in terms of memory requirements, the proposed CDR approach is much better since CDR only needs to record the indexes of the discriminative features ($O(p)$) while PCA or LDA needs additional memory storage for the transformation matrix ($O(p \times d)$) where p is the number of the new discriminative features and d is the total number of the available features.

The following discussion briefly explains the possible reasons behind the good performance of CDR compared with the MI-based algorithm. Note that CDR selects the significant features from the available features iteratively. The methodology has been implemented by searching the features which minimize the residual output error in each iteration. On the other hand, the selection criterion of MI-based approach is based on a similarity measure between the features and classes. However, based on the estimation of the

entropies from the data, the result from MI is sensitive to the computation of this measure. Compared to MI, the proposed strategy in this paper is more efficient as the results show in this section. Figure 7 presents different features selected by the CDR and MI-based algorithm from six images chosen from the training data. In this experiment, 500 features are considered and the resulting features are highlighted in the original image.

Fig. 7. An example of different features (regions) selected by CDR and MI-based algorithm when the same number of features (500) is employed. The first row displays six input images from the training data. The second and third rows return the corresponding highlighted regions from CDR, MI-based algorithm respectively.

We note that although the same number of features (500) is applied, the feature region from CDR seems bigger than that of the MI-based algorithm. It indicates that CDR may consider more information from the original input image. However, the features from MI-based method may cluster in certain parts which may not be as informative for the classification, compared to CDR. For example, as we can see from the input image *B*, CDR almost covers the entire body while MI-based method ignores the person's head which may decrease the generalization ability.

VII. CONCLUSIONS

We have presented a novel method for feature selection, termed compressed sensing-based dimensionality reduction (CDR). The algorithm is inspired by the fact that there is a strong similarity between the original features and the dictionary in compressed sensing. CDR selects the features that minimize the residual output error at each iteration and the stopping criterion is determined automatically using the feedback on the validation set. Also, it has been shown that the CDR algorithm can be regarded as a sparse linear classification method, which means that we can select discriminative features and classify the training data simultaneously. Experimental results on detecting pedestrians show clearly that the proposed algorithm outperforms the original architecture and some traditional dimensionality reduction methods in terms of memory complexity and accuracy.

REFERENCES

- [1] A. Yarbus, "Eye Movements and Vision," *Plenum*, 1967.
- [2] T. R. Lim and A. T. Guntoro, "Car recognition using Gabor filter feature extraction," *Asia-Pacific Conference on Circuits and Systems, APCCAS*, 2002.
- [3] K. Wang and L. Xin etc., "Application of Kernel Method on Face Feature Extraction," *International Conference on Mechatronics and Automation, ICMA*, pp. 3560-3564, 2007.
- [4] Y. LeCun, B. Boser, J. S. Denker, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, pp. 541-551, 1989.
- [5] A. Bouzerdoum and R. B. Pinter, "Shunting inhibitory cellular neural networks: derivation and stability analysis," *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, vol. 40, pp. 215-221, 1993.
- [6] F. H. C. Tivive and A. Bouzerdoum, "A face detection system using shunting inhibitory convolutional neural networks," *IEEE International Joint Conference on Neural Networks*, vol. 4, pp. 2571-2575, 2004.
- [7] F. H. C. Tivive and A. Bouzerdoum, "A Gender Recognition System using Shunting Inhibitory Convolutional Neural Networks," *IEEE International Joint Conference on Neural Networks*, pp. 5336-5341, 2006.
- [8] J. Yang, D. Zhang and A. F. Frangi, "Two-dimensional PCA: a new approach to appearance-based face representation and recognitions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 26, pp. 131-137, 2004.
- [9] W. Zuo, D. Zhang, "BDPCA plus LDA: a novel fast feature extraction technique for face recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 36, pp. 946-953, 2006.
- [10] B. ZMojaradi, H. Abrisami-Moghaddam, M. J. V. Zoej, and R. P. W. Duin, "Dimensionality Reduction of Hyperspectral Data via Spectral Feature Extraction," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, pp. 2091-2105, 2009.
- [11] E. J. Candes, J. Romberg, T. Tao, "Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information," *IEEE Transactions on Information Theory*, vol. 52, pp. 489-509, 2006.
- [12] H. Rauhut, K. Schnass, P. Vandergheynst, "Compressed Sensing and Redundant Dictionaries," *IEEE Transactions on Information Theory*, vol. 54, pp. 2210 - 2219, May 2008.
- [13] R. Gribonval, P. Vandergheynst, "On the Exponential Convergence of Matching Pursuits in Quasi-Incoherent Dictionaries," *IEEE Transactions on Information Theory*, vol. 52, pp. 255-261, 2006.
- [14] J. A. Tropp and C. Gilbert Anna, "Signal Recovery From Random Measurements Via Orthogonal Matching Pursuit," *IEEE Transactions on Information Theory*, vol. 53, pp. 4655-4666, 2007.
- [15] S. F. Cotter et al., "Sparse solutions to linear inverse problems with multiple measurement vectors," *IEEE Transactions on Signal Processing*, vol. 53, pp. 2477-2488, 2005.
- [16] B. D. Rao, K. Engan, S. F. Cotter, J. Palmer, K. Kreutz-Delgado, "Subset selection in noise based on diversity measure minimization," *IEEE Transactions on Signal Processing*, vol. 51, pp. 760-770, 2003.
- [17] J. X. Hong and G. Hu, "Two-Phase Construction of Multilayer Perceptrons Using Information Theory," *IEEE Transactions on Neural Networks*, vol. 20, pp. 715-721, 2009.
- [18] F. H. C. Tivive, A. Bouzerdoum, "Efficient training algorithms for a class of shunting inhibitory convolutional neural networks," *IEEE Transactions on Neural Networks*, vol. 16, pp. 541-556, 2005.
- [19] S. Munder, D. M. Gavrilu, "An experimental study on pedestrian classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1863-1868, 2006.
- [20] G. Arulampalam, A. Bouzerdoum, "A generalized feedforward neural network architecture for classification and regression," *Neural Networks*, vol. 16, no. 5-6, pp. 561-568, 2003.
- [21] J. Yang, A. Bouzerdoum, S. L. Phung, "A Neural Network pruning approach based on Compressive Sampling," *IEEE International Joint Conference on Neural Networks*, pp. 3428-3435, 2009.
- [22] F. H. C. Tivive, A. Bouzerdoum, S. L. Phung, and K. M. Iftekharuddin, "Adaptive hierarchical architecture for visual recognition," *Applied Optics*, vol. 49, no. 10, pp. B1-B8, 2010.