Robust human face recognition based on locality preserving sparse overcomplete block approximation

Dimche Kostadinov, Sviatoslav Voloshynovskiy and Sohrab Ferdowsi

University of Geneva, Computer Science Department,
7 Route de Drize, Geneva, Switzerland

ABSTRACT

Compressive Sensing (CS) has become one of the standard methods in face recognition due to the success of the family of Sparse Representation based Classification (SRC) algorithms. However it has been shown that in some cases, the locality of the dictionary codewords is more essential than the sparsity. Also sparse coding does not guarantee to be local which could lead to an unstable solution. We therefore consider the statistically optimal aspects of encoding that guarantee the best approximation of the query image to a dictionary that incorporates varying acquisition conditions. We focus on the investigation, analysis and experimental validation of the best robust classifier/predictor and consider frontal face image variability induced by noise, lighting, expression, pose, etc.. We compare two image representations using a pixel-wise approximation and an overcomplete block-wise approximation with two types of sparsity priors. In the first type we consider all samples from a single subject and in the second type we consider all samples from all subjects. The experiments on a publicly available dataset using low resolution images showed that several per subject sample sparsity prior approximations are as good as the results presented from SCR and that our simple overcomplete block-wise approximation provides superior performance in comparison to the SRC and WSRC algorithm.

Keywords: Compressive sensing, encoding, sparse approximations, face recognition

1. INTRODUCTION

Automatic human face recognition systems are used in a wide range of real world practical applications related to identification, verification, posture/gesture recognition, social network linking and multimodal interaction. In the last ten years, the problem of face recognition was intensively studied in different domains including biometrics, computer vision and machine learning with the main emphasis on recognition accuracy under various acquisition conditions and more recently on security and privacy. In different fields this problem has been addressed using different approaches mainly dependent on the type of representation and the used recognition method.

A standard preprocessing step in every face recognition system is face alignment and cropping where the face images are typically aligned according to the positions of the eyes. Afterwords the two common stages of a typical face recognition system are [1]: (i) *feature extraction*: where numerous methods have been proposed to project data to a low dimensional feature subspace, e.g. Principal Component Analysis (PCA) [2], Linear Discriminant Analysis (LDA) [3] and Laplasefaces (LPP) [4] and (ii) *classifier construction* and label prediction. In the past Nearest Neighbor (NN) [5] and Nearest Feature Subspace (NFS) [6] were used for classification. NN classifies the query image by only using its Nearest Neighbor. NN utilizes the local structure of the training data and is therefore easily affected by noise. NFS approximates the query image by using all the images belonging to an identical class, using the linear structure of the data. Class prediction is achieved by selecting that class of images that minimize the reconstruction error. NFS might fail in the case that classes are highly correlated to each other.

Certain aspects of these problems can be overcome by Sparse Representation based Classification (SRC) [7]. According to SRC, a dictionary is first learned from training images, which can be acquired from the same subject under different viewing conditions or from various subjects. Moreover, for memory-complexly reasons,

Further author information: (Send correspondence to S. Voloshynovskiy)
S. Voloshynovskiy: E-mail: svolos@unige.ch, Telephone: +41 (22) 379 01 58

the training images can be further clustered using vector quantization based on hierarchical k-means or random forests algorithms. In addition, to better cope with non-stationary distortions, the training images can be represented by a set of blocks. In this case, the enrollment stage includes the representation of each subject in the space of the earlier learned dictionary. At the recognition stage, a query image is first sparsely encoded using the codewords of the learned dictionary after which the classification is performed by verifying which class yields the smallest coding errors. In a direction related to sparsity, Elhamifar and Vidal [8] proposed a more robust classification method using a structured sparse representation, while Gao at al. [9] introduced a kernelized version of SRC. In a direction related to dimensionality reduction, Qiao at al. [10] proposed an unsupervised sparsity preserving projection method while Lu et al. [11] provided a supervised dimensionality reduction method for SRC. Qinfeng S. at al. [12] argue that the lack of sparsity in the data means that the compressive sensing approach cannot be guaranteed to recover the exact signal, and therefore that sparse approximations may not deliver the desired robustness and performance. It has also been shown [13, 14] that in some cases, the locality of the dictionary code words is more essential than the sparsity. The original sparse coding sheme does not guarantee to be local which leads to an unstable solution. Another extension of SRC [1], called Weighted Sparse Representation based Classification (WSRC) integrates the locality structure of the data into a sparse representation in a unified formulation. Finally, several recent studies [12, 13, 14, 15] indicate that dictionary learning is of secondary importance in comparison to proper encoding.

Most of the above approaches mainly rely on the discriminative nature of the sparse representation to perform classification. However, one can still argue on the most optimal way to exploit this discriminative nature and achieve a statistically optimal solution. We formulate and consider an optimal solution of the recognition problem as a linear approximation of the query image to all of the available images and we compare two image representations using a pixel-wise approximation and an overcomplete block-wise approximation. Considering the sparsity as assumption, we analyze several types of sparsity priors. The first is Sparsity Prior Per Sample Data from All Subjects and the second one is Sparsity Prior Per Subject Sample Data. Two types of approximations follow naturally along this line: global and local. In both of them our goal is to investigate, analyze and experimentally validate the design of the most accurate and robust classifiers considering face image variability induced by factors as noise, lightning, expression or pose.

This paper is organized as follows. In Section 2 we give the basic problem formulation, in Section 3 we shortly revisit two sparse approximation methods. We thoroughly present and explain our three proposed methods in Section 4. The results of the computer simulations for all of the methods are presented in Section 5. Finally Section 6 concludes the paper.

Notations: We use capital bold letters to denote real valued matrices (e.g, $\mathbf{W} \in \Re^{M \mathrm{x} N}$), small bold letters to denote real valued vectors (e.g. $\mathbf{x} \in \Re^{M}$). We use sub and upper indexed vector as sample data (vector) single realization from a given distribution (e.g. $\mathbf{x}_i(m) \in \Re^{M}$, where m denotes the sample from distribution). We denote an element of a vector as x. The estimate of $\mathbf{x}$ is denoted as $\hat{\mathbf{x}}$. All vectors have finite length, explicitly defined where appropriate. We denote an optimization problem that considers norm approximation without a prior with A, if that problem considers L_1-norm approximation we denote it with A_{L1}, if it considers L_2-norm approximation we denote it with A_{L2} and if it considers any other fidelity function, such as for example, the Huber robust function, we denote it with A_H. If the optimization problem that includes a fidelity function (e.g. L_2-norm approximation) and a prior (e.g. L_2-norm prior) we denote the problem with $A_{L2}P_{L2}$. If the considered prior includes weighting coefficients then we denote the problem as $A_{L2}P_{L2}W$. To denote any global optimization problem in our notation we add G and to denote any local optimization problem we add L. For example if we have a local L_2 norm optimization problem with an L_1 prior we denote it as $L - A_{L2}P_{L1}$. We denote a classifier operating on the L_2-norm by $C2$ and on the L_1-norm by $C1$.

2. PROBLEM FORMULATION

The face recognition system consists of two stages *enrolment* and *identification*.

At the enrolment stage, the photos from each subject are acquired and organized in the form of a codebook. We will assume that the recognition system should recognize K subjects. The photo of each subject i, $1 \leq i \leq K$, are acquired under different imaging conditions such as lighting, expression, pose, etc., which will represent the

variability of face features and serve as intraclass statistics. We will also assume that the frontal face images are aligned to the same scale, rotation and translation (as in [7]). Therefore each subject i is defined by $\mathbf{x}_i(m) \in \Re^N$ vectors representing a concatenation of aligned image columns with $1 \leq m \leq M$. The samples from all subjects are arranged into a codebook represented by a matrix:

$$\mathbf{W} = [\mathbf{x}_1(1), ..., \mathbf{x}_1(m), ..., \mathbf{x}_1(M), ..., \mathbf{x}_i(1), ..., \mathbf{x}_i(m), ..., \mathbf{x}_i(M), ..., \mathbf{x}_K(1), ..., \mathbf{x}_K(m), ..., \mathbf{x}_K(M)] \in \Re^{N \mathrm{x}(K*M)} . \tag{1}$$

At the recognition stage, a probe or query $\mathbf{y} \in \Re^N$ is presented to the system. The system should identify the subject i as accurate as possible based on $\mathbf{y}$ and $\mathbf{W}$. It is also assumed that $\mathbf{y}$ always corresponds to one of the subjects represented in the database. The codebook construction is shown in Fig. 1.

In the scope of this paper, face recognition is considered as a classification problem where the classifier should produce the decision in favour of some class i whose codebook codewords produce the most accurate approximation of probe $\mathbf{y}$. One important class of approximations is represented by a *sparse linear approximation* [7], when the probe $\mathbf{y}$ is approximated by $\hat{\mathbf{y}}$ in the form of:

$$\hat{\mathbf{y}} = \mathbf{W}\boldsymbol{\alpha}\,, \tag{2}$$

where $\boldsymbol{\alpha} \in \Re^{M \mathrm{x} K}$ is a sparse coding vector shown in Fig. 1. The coding vector $\boldsymbol{\alpha}$ weights the codebook codewords gathered for all classes to favour the contribution of codewords corresponding to the correct class $\hat{i}$. The model of approximations can be represented as:

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{r}\,, \tag{3}$$

where $\mathbf{r} \in \Re^N$ is the residual error vector of the approximation.

For each class i, let $\boldsymbol{\delta}_i : \Re^{K*M} \rightarrow \Re^{K*M}$ be a function that selects the coefficients associated with the ith class. For $\boldsymbol{\alpha} \in R^{K*N}$, $\boldsymbol{\delta}_i(\boldsymbol{\alpha})$ is a new vector whose only nonzero entries are the entries in $\boldsymbol{\alpha}$ that are associated with class i. Using the coefficients that are only associated with the ith class, one can find the class approximation $\hat{\mathbf{y}}_i \in R^M$ to the given test sample $\mathbf{y}$:

$$\hat{\mathbf{y}}_i = \mathbf{W}\boldsymbol{\delta}_i(\boldsymbol{\alpha})\,, \tag{4}$$

and

$$\mathbf{y} = \hat{\mathbf{y}}_i + \mathbf{r}_i\,, \tag{5}$$

where $\mathbf{r}_i \in \Re^N$ is the residual error vector of the approximation to class i.

Then the probe $\mathbf{y}$ is classified based on the approximation $\hat{i}$ that minimizes the L_2-norm of the residual error vector between $\mathbf{y}$ and $\hat{\mathbf{y}}_i$. This classifier is denoted as $C2$:

$$C2: \ \hat{i} = \arg\min_{1 \leq i \leq K} (\|\mathbf{r}_i\|_2) = \arg\min_{1 \leq i \leq K} \|\mathbf{W}\boldsymbol{\delta}_i(\boldsymbol{\alpha}) - \mathbf{y}\|_2. \tag{6}$$

Because we use the $\boldsymbol{\delta}_i$ function in this approximation, which is a hard assignment non-linear function, it might change the optimality of the found solution in the L_2 norm, which is known to be unstable. In order to be able to more robustly tackle this problem, we propose a classifier based on the approximation $\hat{i}$ that minimizes the L_1-norm of the residual error vector between $\mathbf{y}$ and $\hat{\mathbf{y}}_i$ and we denote this classifier as $C1$:

$$C1: \ \hat{i} = \arg\min_{1 \leq i \leq K} (\|\mathbf{r}_i\|_1) = \arg\min_{1 \leq i \leq K} \|\mathbf{W}\boldsymbol{\delta}_i(\boldsymbol{\alpha}) - \mathbf{y}\|_1\,. \tag{7}$$

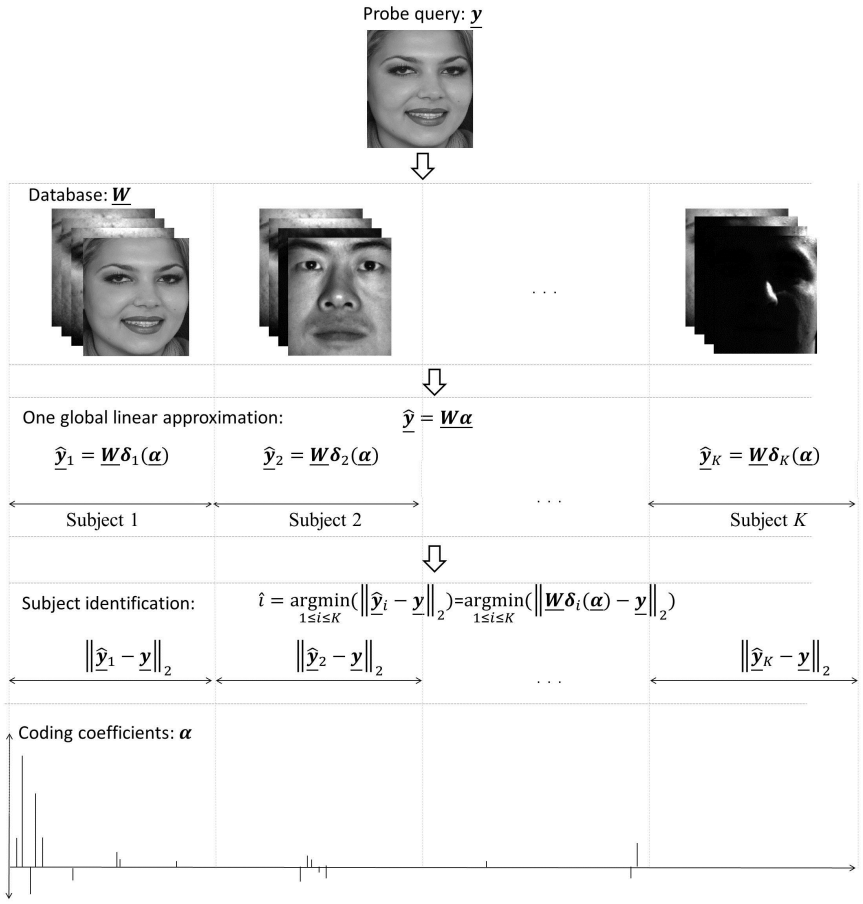

Figure 1. General problem formulation, sparse linear approximation $\hat{\mathbf{y}}$ of the test sample $\mathbf{y}$ in the codebook $\mathbf{W}$.

In more general case, the equations (6) and (7) correspond to the minimum L_p distance classification, where if $p = 2$ one has the Euclidean distance and if $p = 1$ one has the Manhattan distance. A natural extension to (6) and (7) that might be considered is a bounded distance decoding (BBD) rule when:

$$\hat{i} = \{i \in \{1, \cdots, M\} : \|\mathbf{W}\boldsymbol{\delta}_i(\boldsymbol{\alpha}) - \mathbf{y}\|_p \leq \eta\}, \tag{8}$$

where $\eta \geq 0$. The BDD rule is useful when the classifier should reject the database unrelated probes. In the general case, the BDD will produce a list of candidates that satisfy the above condition. To have only one unique $\hat{i}$ on the list, the parameter η should be chosen accordingly. Geometrically in the L_p space, it means that the L_p spheres with radius η around each approximate for each class should not overlap thus producing a unique classification.

We investigate and analyze the pixel-wise approximation and the overcomplete block-wise approximation. Accordingly, we statistically analyze two types of priors for sparsity:

-*global sparsity*: a sparsity prior per all the images in the available data set such that the vector $\boldsymbol{\alpha}$ has a few non zero values that might be related to images from different subject class;

-*local sparsity*: explicit sparsity prior per subject class such that the vector $\boldsymbol{\alpha}$ might have non zero values at just one object class and zero at all the rest.

In the next sections, we will present two different types of approximations and the corresponding classifiers based on the local and global sparsity models. Two samples of two different coding coefficients for the two types of sparsity assumptions are exemplified in Fig. 2.

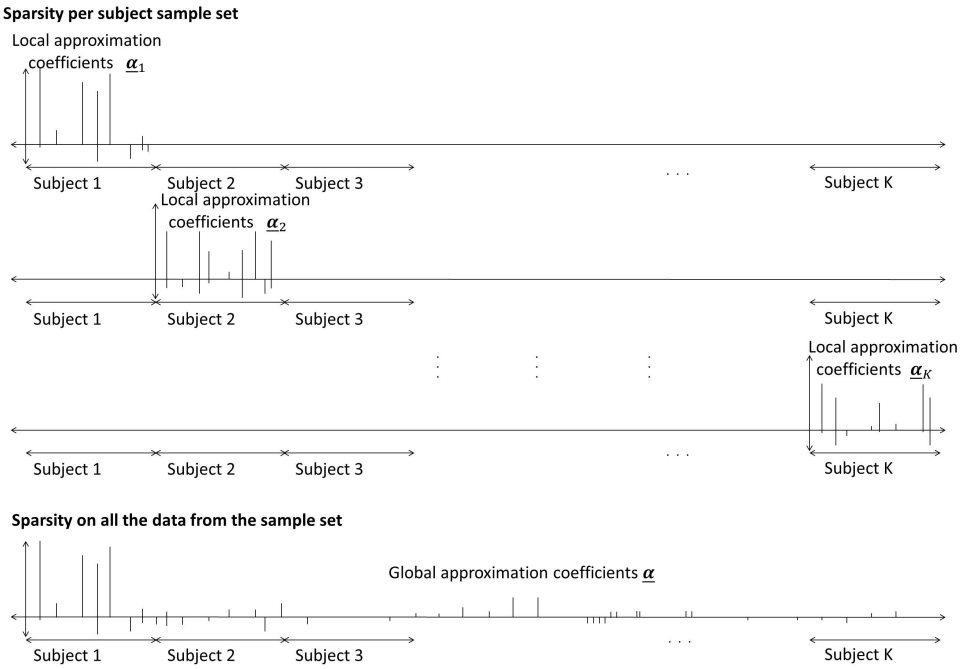

Figure 2. Two samples of the two different coding coefficients for the two types of sparsity assumptions. On the top sparsity prior per all the images in the available data set is assigned such that the coefficients $\boldsymbol{\alpha}$ have few non zero values that might be related to images from a different subject class. On the bottom an explicit sparsity prior per subject class is shown such that the coefficients $\boldsymbol{\alpha}$ might have non zero values at just one object class and zero at all the rest.

3. OVERVIEW OF SPARSE APROXIMATION MODELS

In the first part of this section, we briefly overview the SRC algorithm and in the second part, one of its more robust extensions known as the WSRC algorithm.

3.1 Classification Based on Sparse Representation

In the spirit of SRC, assuming that the training samples $\mathbf{x}_i(1), \mathbf{x}_i(2), \mathbf{x}_i(3), ..., \mathbf{x}_i(M)$ from a single subject i lie on a subspace, then any new (test) sample $\mathbf{y}$ from the same subject i should approximately lie in the linear span of the training samples $\mathbf{W}_i = [\mathbf{x}_i(1), \mathbf{x}_i(2), \mathbf{x}_i(3), ..., \mathbf{x}_i(M)] \in \Re^{N \mathrm{x} M}$ associated with subject i [7]:

$$\mathbf{y}_i = \mathbf{W}_i \boldsymbol{\alpha}_i \ , \tag{9}$$

where $\boldsymbol{\alpha}_i = [\alpha_i(1), \alpha_i(2), \alpha_i(3), ..., \alpha_i(M)] \in \Re^M$.

Since i is unknown, the linear representation of $\mathbf{y}$ can be rewritten as approximation $\hat{\mathbf{y}}$ of all training samples:

$$\hat{\mathbf{y}} = \mathbf{W}\boldsymbol{\alpha} \ . \tag{10}$$

Motivated by the sparse coefficient approximation, SRC aims at solving the L_0-minimization problem:

$$\hat{\boldsymbol{\alpha}} = \ \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_0 \ \ \textit{subject to } \mathbf{y} = \mathbf{W}\boldsymbol{\alpha} \ , \tag{11}$$

where the L_0-norm counts the number of nonzero entries in a vector. However, the problem of finding the most sparse solution of an under-determined system of linear equations is NP-hard and difficult even to approximate [19]. The theory of compressive sensing (some of the practicable aspects exploited in [7]) reveals

that if the solution $\boldsymbol{\alpha}$ is sparse enough, the solution of the L_0-minimization problem is equal to the following L_1-minimization problem:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_1 \ \textit{subject to}\ \mathbf{y} = \mathbf{W}\boldsymbol{\alpha} \ . \tag{12}$$

The L_1-minimization problem can be extended to the following stable L_1-minimization problem:

$$G\text{-}A_{L2}P_{L1}\text{:}\ \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\boldsymbol{\alpha}\|_1 \ \textit{subject to}\ \|\mathbf{W}\boldsymbol{\alpha} - \mathbf{y}\|_2 < \epsilon \ , \tag{13}$$

where $\epsilon > 0$ is a given tolerance. We note that this problem formulation can be expressed in another form known as the Least Absolute Shrinkage and Selection Operator LASSO [16].

After obtaining the most sparse solution $\hat{\boldsymbol{\alpha}} \in \Re^{MxK}$, in SRC the classification is performed using equation (6) or (7).

In the SRC algorithm, if the solution $\boldsymbol{\alpha}$ is sparse enough, then with overwhelming probability [7], it can be correctly recovered via L_1-minimization from any sufficiently large number of linear measurements. However, if the number of linear measurements is not large enough, SRC might perform poorly. This indicates that the discriminative information from the linearity structure of the data especially in lower dimensional feature subspaces, is not enough for SRC. In the next subsection, we describe an extension of SRC by imposing the locality constraint on the sparsity regularized reconstruction problem.

We also extend our analysis to the L_2-norm approximation defined as:

$$G\text{-}A_{L2}\text{:}\ \hat{\boldsymbol{\alpha}} = \|\mathbf{W}\boldsymbol{\alpha} - \mathbf{y}\|_2 \ , \tag{14}$$

and also to L_2-norm approximation with L_2-norm prior:

$$G\text{-}A_{L2}P_{L2}\text{:}\ \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{W}\boldsymbol{\alpha} - \mathbf{y}\|_2 + \lambda\|\boldsymbol{\alpha}\|_2 \ , \tag{15}$$

where λ corresponds to the Lagrangian multiplier.

Considering the robustness of the approximation we also analyze L_1-norm approximation:

$$G\text{-}A_{L1}\text{:}\ \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{W}\boldsymbol{\alpha} - \mathbf{y}\|_1 \ , \tag{16}$$

the L_1-norm approximation with L_2-norm prior:

$$G\text{-}A_{L1}P_{L2}\text{:}\ \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{W}\boldsymbol{\alpha} - \mathbf{y}\|_1 + \lambda\|\boldsymbol{\alpha}\|_2 \ , \tag{17}$$

and L_1-norm approximation with L_1-norm prior:

$$G\text{-}A_{L1}P_{L1}\text{:}\ \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{W}\boldsymbol{\alpha} - \mathbf{y}\|_1 + \lambda\|\boldsymbol{\alpha}\|_1 \ . \tag{18}$$

3.2 Classification Based on Weighted Sparse Representation

To overcome the shortcomings of the SRC algorithm, the authors of [1] proposed a more robust weighted sparse representation method which integrates both sparsity and data locality into a unified formulation. WSRC preserves the similarity between the test sample and its neighboring training data while seeking a sparse linear representation. Similar to SRC, the WSRC algorithm uses all the training data as a dictionary and the locality is imposed by the L_1 regularization. WSRC is formulated as the following weighted L_1 minimization problem:

$$\hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{C}\boldsymbol{\alpha}\|_1 \ \textit{subject to}\ \mathbf{y} = \mathbf{W}\boldsymbol{\alpha}\,, \tag{19}$$

where $\mathbf{C} \in \Re^{(K*M)\mathrm{x}(K*M)}$ is a block-diagonal matrix:

$$diag(\mathbf{C}) = [\|\mathbf{y} - \mathbf{x}_1(1)\|_{1.5}, \|\mathbf{y} - \mathbf{x}_1(2)\|_{1.5}, ..., \|\mathbf{y} - \mathbf{x}_i(m)\|_{1.5}, ..., \|\mathbf{y} - \mathbf{x}_K(N-1)\|_{1.5}, \|\mathbf{y} - \mathbf{x}_K(N)\|_{1.5}]\,, \tag{20}$$

which is the locality adapter that penalizes the distance between $\mathbf{y}$ and each training data $\mathbf{x}_i(m)$ and $\|.\|_{1.5}$ is L_p-norm with $p = 1.5$ defined as:

$$\|\mathbf{x}\|_{1.5} = (x_1^{1.5} + x_2^{1.5} + x_3^{1.5} + ... + x_N^{1.5})^{\frac{2}{3}}\ . \tag{21}$$

This problem can be extended to the following more stable L_1-minimization problem:

$$G\text{-}A_{L2}P_{L1}W_{L1.5}\text{:}\ \hat{\boldsymbol{\alpha}} = \underset{\boldsymbol{\alpha}}{\operatorname{argmin}} \|\mathbf{C}\boldsymbol{\alpha}\|_1 \ \textit{subject to}\ \|\mathbf{W}\boldsymbol{\alpha} - \mathbf{y}\|_2 < \epsilon \tag{22}$$

where $\epsilon > 0$ is a given tolerance.

After computing the residual approximation the same decision rule is used as in the SRC algorithm.

4. PROPOSED METHODS FOR SPARSE APPROXIMATION

Here we present the formulations of the two proposed sparse approximations. In the first part, we propose several approximations with a prior sparsity per class and in the second one we present one simple overcomplete block-wise approximation with a prior sparsity per all data samples from all classes.

4.1 Sparsity Prior Per Subject Sample Data

Since the object class i is unknown in advance, one can assume that as in Section 3.2, the training samples $\mathbf{x}_i(1), \mathbf{x}_i(2), \mathbf{x}_i(3), ..., \mathbf{x}_i(M)$ from a single subject i lie on a subspace. However, instead of solving an approximation problem for the whole set of samples we assume a priori sparsity per all samples from single subject class and consider the approximation using just all the samples per subject. We refer to this as a local prior sparse approximation per class.

By the corollary of the a priori assumptions the coding coefficients must have a priori zero values at all the corresponding samples from multiple subjects except the samples from one subject. These values (coding coefficients) might have non zero values. Again we want to find an approximation to $\mathbf{y}$. However, instead of using all the samples from all the subjects, we make K independent local approximations $\hat{\mathbf{y}}_i \in \Re^N$, $1 \leq i \leq K$ to $\mathbf{y}$, equivalent to the number of subjects K by using all the samples from a given subject i:

$$\hat{\mathbf{y}}_i = \mathbf{W}_i\boldsymbol{\alpha}_i\,, \tag{23}$$

where $\mathbf{W}_i = [\mathbf{x}_i(1), \mathbf{x}_i(2), \mathbf{x}_i(3), ..., \mathbf{x}_i(M)] \in \Re^{N\mathrm{x}M}$, $i \in (1, 2, 3...K)$ and $\boldsymbol{\alpha}_i \in \Re^M$.

It is interesting to investigate the impact of sparsity imposed by priors on the vector $\boldsymbol{\alpha}_i$ and the accuracy versus the robustness of the fidelity constraint imposed on the residual $(\mathbf{y}_i - \mathbf{W}_i\boldsymbol{\alpha}_i)$. To cover the general form we will consider the approximation in the general form:

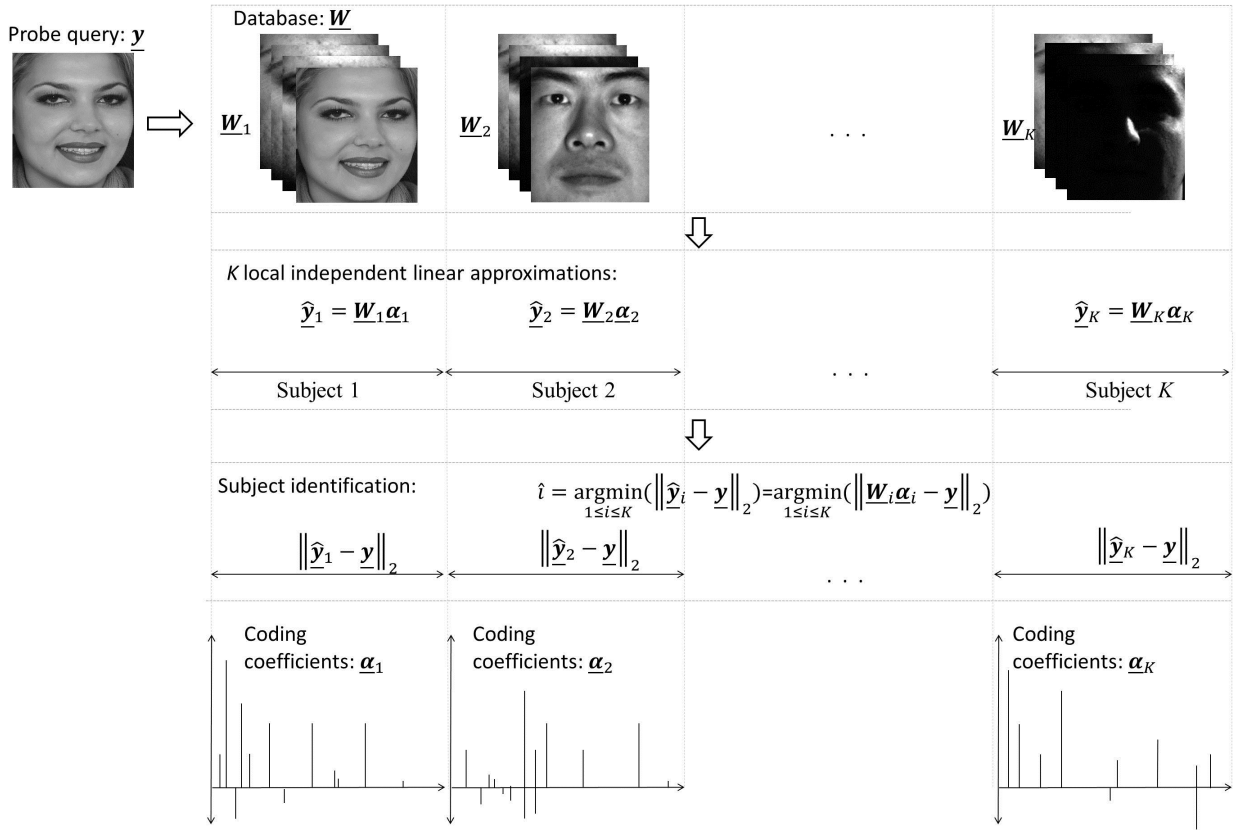

Figure 3. Two samples of the two different coding coefficients for the two types of sparsity assumptions. On the left side sparsity prior per all the images in the available data set such that the coefficients $\boldsymbol{\alpha}$ have few non zero values that might be related to images from different subject class, on the right explicit sparsity prior per subject class such that the coefficients $\boldsymbol{\alpha}$ might have non zero values at just one object class and zero at all the rest.

$$\hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \boldsymbol{\rho}(\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}) + \lambda\|\boldsymbol{\alpha}_i\|_p \ , \tag{24}$$

where $\boldsymbol{\rho}(.)$ is the penalty function, $\|.\|_p$ stands for L_p-norm and λ corresponds to the Lagrangian multiplier. In the following we will consider several approximation algorithms for different penalty functions and priors.

L_2-norm approximation: To investigate the impact of priors and the accuracy of the penalty function, we start with the least square solution:

$$L\text{-}A_{L2}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \|\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}\|_2 \ , \tag{25}$$

which yields a close form solution:

$$\hat{\boldsymbol{\alpha}}_i = (\mathbf{W}_i^T\mathbf{W}_i)^{-1}\mathbf{W}_i^T\mathbf{y} \ . \tag{26}$$

Next, we investigate the impact of two prior models with L_2 and L_1 regularizations.

L_2-norm approximation with L_2-norm regularization: The L_2-norm regularization corresponds to the Tikhonov regularization with the regularization term $\lambda\|\mathbf{G}\boldsymbol{\alpha}_i\|_2$:

$$L\text{-}A_{L2}P_{L2}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \|\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}\|_2 + \lambda\|\mathbf{G}\boldsymbol{\alpha}_i\|_2 \ , \tag{27}$$

where G is the regularization matrix in the from of smoothness that leads to the closed form solution:

$$\hat{\boldsymbol{\alpha}}_i = (\mathbf{W}_i^T\mathbf{W}_i + \lambda\mathbf{G}^T\mathbf{G})^{-1}\mathbf{W}_i^T\mathbf{y}\,, \tag{28}$$

where a simple selection of the regularizer $\mathbf{G}$ corresponds to the identity matrix $\mathbf{G} = \mathbf{I}$.

L_2-norm approximation with L_1-norm regularization: The L_1 regularization resembles the SRC formulation with the subject related priors:

$$L\text{-}A_{L2}P_{L1}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} \|\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}\|_2 + \lambda\|\boldsymbol{\alpha}_i\|_1 \ . \tag{29}$$

The solution corresponds to LASSO.

***Huber* penalty function as approximation without priors**: This formulation aims at capturing the impact on accuracy introduced by the approximation term and the related error deviations by introducing the robust M-estimator in the form of the Huber penalty function:

$$L\text{-}A_H\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} (\boldsymbol{\psi}_H(\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}))\,, \tag{30}$$

where:

$$\boldsymbol{\psi}_H(a) = \begin{cases} (\frac{1}{2})a^2 \ \textit{if}\ |a| \le \beta\,, \\ \beta(|a| - \frac{1}{2})\beta) \ \textit{otherwise.} \end{cases} \tag{31}$$

***Huber* penalty function as approximation with L_2-norm prior** :

$$L\text{-}A_HP_{L2}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} (\boldsymbol{\psi}_H(\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}) + \lambda\|\mathbf{G}\boldsymbol{\alpha}_i\|_2)\,. \tag{32}$$

***Huber* penalty function as approximation with L_1-norm prior**:

$$L\text{-}A_HP_{L1}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} (\boldsymbol{\psi}_H(\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}) + \lambda\|\boldsymbol{\alpha}_i\|_1). \tag{33}$$

***Tukey* penalty function as approximation without priors**: Similar to the Hubert estimator this formulation also intends to capture the impact in accuracy introduced by the approximation term and the related error deviations:

$$L\text{-}A_T\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} (\boldsymbol{\psi}_T(\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}))\,, \tag{34}$$

where:

$$\boldsymbol{\psi}_T(a) = \begin{cases} \frac{k^2}{6}\{1 - [1 - (\frac{a}{k})^2]^3\} \ \textit{if}\ |a| \le k \\ \frac{k^2}{6} \ \textit{otherwise.} \end{cases} \ . \tag{35}$$

***Tukey* penalty function as approximation with L_2-norm prior**:

$$L\text{-}A_TP_{L2}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}} (\boldsymbol{\psi}_T(\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}) + \lambda\|\mathbf{G}\boldsymbol{\alpha}_i\|_2)\,. \tag{36}$$

***Tukey* penalty function as approximation with L_1-norm prior**:

$$L\text{-}A_T P_{L1}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}}(\boldsymbol{\psi}_T(\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}) + \lambda\|\boldsymbol{\alpha}_i\|_1)\,, \tag{37}$$

The solutions to all the unregularized and regularized robust M-estimators are computed using the iterative reweighted least squares (IRLS) algorithm [17].

L_1-norm approximation: Similarly, we consider the L_1-norm approximation without prior:

$$L\text{-}A_{L1}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}}\|\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}\|_1\ . \tag{38}$$

L_1-norm approximation with L_2-norm regularization:

$$L\text{-}A_{L1}P_{L2}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}}\|\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}\|_1 + \lambda\|\mathbf{G}\boldsymbol{\alpha}_i\|_2\ . \tag{39}$$

L_1-norm approximation with L_1-norm regularization:

$$L\text{-}A_{L1}P_{L1}\text{:}\ \hat{\boldsymbol{\alpha}}_i = \underset{\boldsymbol{\alpha}_i}{\operatorname{argmin}}\|\mathbf{W}_i\boldsymbol{\alpha}_i - \mathbf{y}\|_1 + \lambda\|\boldsymbol{\alpha}_i\|_1\ . \tag{40}$$

The solutions to the all unregulated and regularized L_1-norm approximation problems are solved numerically [18].

4.2 Sparse Representation based on Overlapping Image Blocks with Sparsity Prior Per Sample Data from All Subjects

In the general case, the distribution of the data sample variability and how this variability is spatially distributed in the image is unknown in advance. Here we assume non-stationary and non-regular spatially distributed variability. By using a sparse representation on overlapping image blocks we expect to reduce the information uncertainty, thus to increase the identification accuracy.

Let $\mathbf{b}_{ij}(m) \in \Re^Z$ be a block (patch) j from image m coming from subject i and let $\mathbf{y}_j \in \Re^Z$ be the same numbered block (patch) j from the query (test) image $\mathbf{y}$, $1 \leq i \leq K, 1 \leq m \leq N, 1 \leq j \leq B$ where K is the number of subjects, N is the number of images per subject, Z is the number of pixels per block, B is the number of overlapping blocks per image. This method has the following formulation:

$$G\text{-}A_{L2}P_{L1}W_{L2}\text{:}\ \hat{\boldsymbol{\alpha}}_j = \underset{\boldsymbol{\alpha}_j}{\operatorname{argmin}}\|\mathbf{W}_j\boldsymbol{\alpha}_j - \mathbf{y}_j\|_2 + \lambda\|\mathbf{C}_j \odot \boldsymbol{\alpha}_j\|_1\ \ \textit{such that}\ \sum_{k=1}^{K*M}\alpha_j(k) = 1\ , \tag{41}$$

where the coding coefficients $\hat{\boldsymbol{\alpha}}_j \in \Re^{K*M}$ and $\odot$ denotes the element-wise multiplication and the data sample set consists of the image blocks j from all the images:

$$\mathbf{W}_j = [\mathbf{b}_{1j}(1), \mathbf{b}_{1j}(2), ...\mathbf{b}_{ij}(m-1), \mathbf{b}_{ij}(m), ..., \mathbf{b}_{Kj}(N-1), \mathbf{b}_{Kj}(N)]\,, \tag{42}$$

$\mathbf{C}_j \in \Re^{K*M}$ is a vector that penalizes the distance between test image block $\mathbf{y}_j$ and each training data block $\mathbf{b}_{ij}(m)$, defined as follows:

$$\mathbf{C}_j = [dist(\mathbf{y}_j, \mathbf{b}_{1j}(1)), dist(\mathbf{y}_j, \mathbf{b}_{1j}(2)), ..., dist(\mathbf{y}_j, \mathbf{b}_{ij}(m)), ..., dist(\mathbf{y}_j, \mathbf{b}_{Kj}(N-1)), dist(\mathbf{y}_j, \mathbf{b}_{Kj}(N))\,, \tag{43}$$

where $dist(\mathbf{y}_j, \mathbf{b}_{ij}(m)) = \exp(\|\mathbf{y}_j - \mathbf{b}_{ij}(m)\|_2/\sigma)$, with σ a spread parameter estimated experimentally; in our experiments we used $\sigma = 100$ which is roughly equal to the data dimensionality. We call this method modified-WSRC (MWSRC).

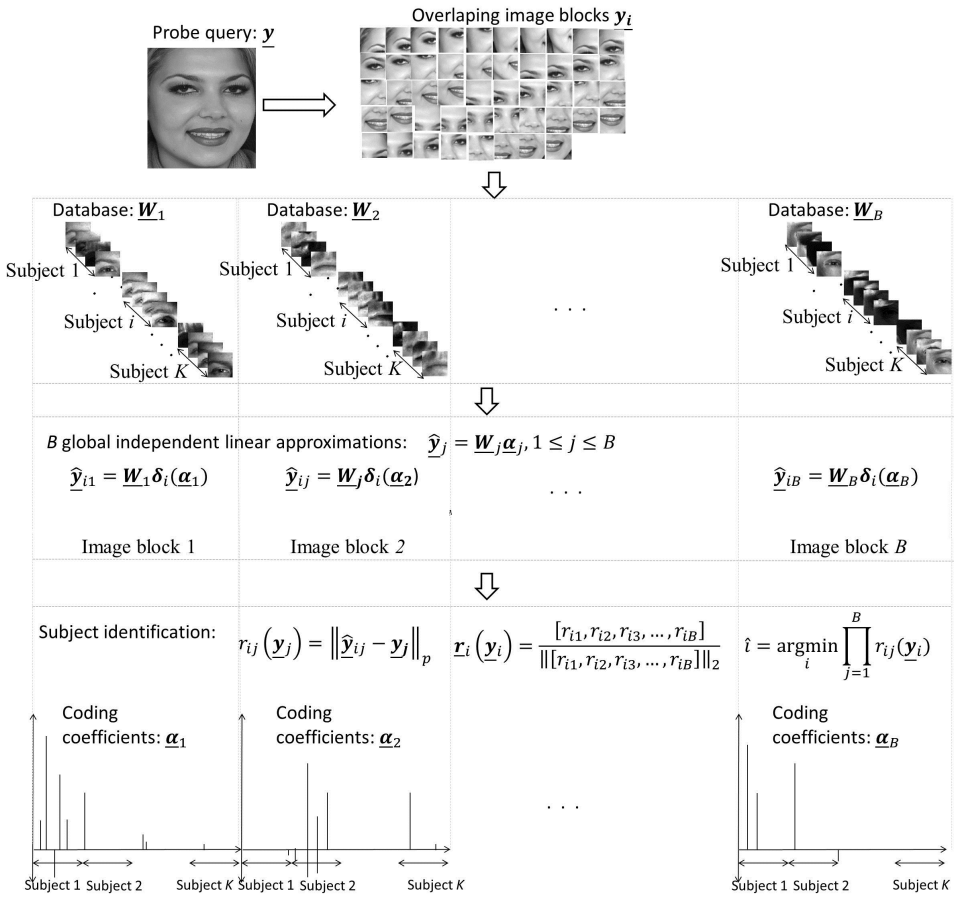

Figure 4. Weighted Sparse Representation on Overlapping Image Blocks with Sparsity Prior Per Sample Data from All Subjects.

Our analysis also extends to the locality constrained linear coding (LCLC) [19] problem formulated as follows:

$$G\text{-}A_{L2}P_{L2}W_{L2}\text{:}\ \hat{\boldsymbol{\alpha}}_j = \underset{\boldsymbol{\alpha}_j}{\operatorname{argmax}} \|\mathbf{W}_j\boldsymbol{\alpha}_j - \mathbf{y}_j\|_2 + \lambda\|\mathbf{C}_j \odot \boldsymbol{\alpha}_j\|_2\,. \tag{44}$$

This is a similar problem to the previous one, the only difference is the induced norm on the regularization term.

Operating on the overcomplete dictionary that consists of overlaping image blocks, the extended analysis further includes the SRC block-wise approximation, the $L2$ norm block-wise approximation and the $L1$ norm block-wise approximation with no prior, the L_2 norm and the $L1$ norm prior.

The final decision function essentially represents a simple ensemble classifier. The fusion rule consists of normalizing the assigned decisions per block and taking the product from those decisions that have the smallest values, followed by a minimum search per subject product. More formally:

$$\hat{\mathbf{y}}_{ij} = \mathbf{W}_j\boldsymbol{\delta}_i(\boldsymbol{\alpha}_j)\ \ , \tag{45}$$

$$r_{ij}(\mathbf{y}_j) = \|(\hat{\mathbf{y}}_{ij} - \mathbf{y}_j)\|_p\ \ , \tag{46}$$

where $p \in \{1, 2\}$, when $p = 1$ this is related to L_1 norm error and when $p = 2$ this is related to L_2 norm error.

$$\mathbf{r}_i(\mathbf{y}_i) = \frac{[r_{i1}, r_{i2}, r_{i3}, ..., r_{iB}]}{\|[r_{i1}, r_{i2}, r_{i3}, ..., r_{iB}]\|_2} \ , \tag{47}$$

$$\hat{i} = \arg \min_{1 \leq i \leq K} \prod_{j=1}^{B} r_{ij}(\mathbf{y}_i) \ . \tag{48}$$

5. COMPUTER SIMULATION

In this section we present the results of the computer simulation that are organized in three parts. In the first part, we present the results using a pixel-wise approximation with per all data samples sparsity priors, the results using per subject data sample sparsity prior are presented in the second part and the results using an overcomplete block-wise approximation with per all data samples sparsity priors are given in the last part. In all three parts of the computer simulation, we present the results using two classifiers: $C1$ (equation (6)) and $C2$ (equation (7)). The goal of the computer simulation is to underline the importance and the significance of the right assumptions including the type of the approximation, prior and classifier related to the most optimal sparse representation that achieves highest accuracy.

The computer simulation is carried out on publicly available data. The used database is Extended Yale B for face recognition. This database consists of 2414 frontal face images of 38 subjects captured under various laboratory-controlled lighting conditions [20]. All the images from this database are cropped and normalized to 192x168 pixels.

In our set up, the images from the dataset are rescaled to 10x12 pixels using nearest neighbor interpolation. In all of the computer simulations we use raw, basic, elementary image pixel values (block of image pixel values) as features. To be unbiased in our validation of the results we use 5-fold cross validation, where for single validation for each subject, half of the images are selected at random for training and the remainder for testing.

All of the formulated optimization problems presented in the previous chapters are solved using CVX [21]. In all of the regularized optimization problems, the regularization parameters were chosen to maximize the classification accuracy.

5.1 Pixel-wise approximation with per all data samples sparsity priors

In this series of computer simulations, we analyse the identification accuracy related to the assumption of *sparsity prior per all the subject samples*. We investigate the recognition accuracy under 4 models of approximation: $G\text{-}A_{L2}$, $G\text{-}A_{L1}$, $G\text{-}A_{LCLC}$ and $G\text{-}A_{WSRC}$. To study the impact of priors on the recognition performance we consider 3 models: no prior, P_{L2} and P_{L1} that correspond to the L_2-norm and the L_1 norm, respectively. Finally, the resulting estimates are tested with the $C1$ and $C2$ classifiers. In the L_2 norm and L_1 norm approximations with L_2 norm prior the Lagrangian multiplier λ is set to 330, in the L_2 norm approximation with L_1 norm prior the Lagrangian multiplier λ is set to 1000, in the L_1 norm approximation with L_1 norm prior the Lagrangian multiplier λ is set to 500. In the LCLC algorithm the Lagrangian multiplier λ is set to 300 and in the WSRC algorithm the tolerance ϵ is set to 10^{-4}. Table 1 shows the results for the above methods. The results presented in bold signifies the best achieved results.

Table 1. Identification precision under pixel-wise approximation with per all the data samples sparsity prior

	$G\text{-}A_{L2}$		$G\text{-}A_{L1}$		$G\text{-}A_{LCLC}$		$G\text{-}A_{WSRC}$	
	C2	C1	C2	C1	C2	C1	C2	C1
No prior	90.8717%	91.0855%	87.8125%	82.9441%	-	-	-	-
P_{L2}	90.6415%	91.0197%	87.5493%	89.3092%	**91.58%**	89.8729%	-	-
P_{L1}	93.7381 %	**94.5066**%	91.6941%	**94.1941**%	-	-	93.24%	**94.0790**%

In most of the obtained results, the $C1$ classifier demonstrates superior performance in comparison to the $C2$ classifier. It is explained by the impact of the $\delta_i(\alpha)$ operator in the classification rules (6) and (7), which

changes the optimality of the sparse solution. The application of the $\delta_i(\alpha)$ operator creates a number of outliers, and the L_1 norm is known to be more robust to outliers in comparison to the L_2 norm. In addition, L_1 norm prior P_{L1} demonstrates a significant gain in performance in almost all approximation models and classifiers in comparison to the L_2 norm prior P_{L2} and no prior cases. It is also interesting to point out that the P_{L2} in $G - A_{L2}$ and $G - A_{L1}$ give no improvement over the no prior cases with the exception of the $G - A_{L1}P_{L2}C1$ case. This explains the non-informative impact of the L_2 norm prior. Finally, the impact of the approximation model is negligible under the P_{L1} prior and $C1$ classifier. The best result is achieved for the $G - A_{L2}P_{L1}C1$ setup.

5.2 Pixel-wise approximation with per subject data samples sparsity priors

In this series of computer simulations we present the identification accuracy when we assume sparsity priors per subject sample data. We compare the performance of several linear, quadratic and robust approximations, including the L_2 norm, L_1 norm, Huber and Tukey penalty functions. The performance of all approximations are evaluated using no prior, the L_2 norm prior and the L_1 norm prior. In the L_2 norm, L_1 norm approximation with L_2 norm and L_2 norm prior the Lagrangian multiplier λ is set to 10, in the Huber and Tukey approximations with L_1 norm L_2 norm prior the Lagrangian multiplier λ is set to 15. Table 2 shows the results for the above methods.

Table 2. Identification precision assuming sparsity prior per subject sample data

	L-A_{L2}		L-A_{L1}		L-A_H		L-A_T	
	C2	C1	C2	C1	C2	C1	C2	C1
No prior	92.054%	92.0724%	91.86%	**92.98**%	91.58%	92.75%	92.75%	92.34%
P_{L2}	92.0888%	92.1053%	91.1390%	92.8442%	91.1225%	92.8429%	91.0306%	92.8659%
P_{L1}	**92.81**%	92.2039	91.3294%	92.9621%	91.1569%	**92.9087**%	90.1298%	**92.9009**%

First, the results from the $C2$ classifier shown in Table 2 indicate that several of the proposed approximations using sparsity prior per subject samples without priors give good identification rates that are comparable to the identification rates of the SRC algorithm. The local L_1-norm approximation has the best identification accuracy. The performance relies on the type of the local approximations and all the available subject samples (or the variability in all the available subject samples). At first this may seem to be restrictive because we do not use the data samples from all the subjects, however it turns out that the per sample data are sufficient in cases where: (1) we have equally likely per class data variability and (2) when we are able to capture the impact of the deviations to some extent in the approximation by introducing robust estimators.

Secondly, the results from the $C1$ classifier shown in Table 2 indicate that with the L_1-norm and the L_2-norm priors we have little improvement. In this approach, the type of sparsity is less important than the type of approximation due to the fact that only local per subject data is used. This is in a good agreement with the assumptions about the local and global sparsity we wanted to validated in this paper.

5.3 Overcomplete block-wise approximation with per all data samples sparsity priors

This series of computer simulations shows the results on the identification accuracy using the overcomplete block-wise approximation related to the assumption: *sparsity prior per all the data samples*. The block size is 4x4, and in total we have 63 overlaping image blocks. All of the image blocks that form the overcomplete dictionary were normalized to have unit one norm. We compare the performance of 2 block-wise approximations with 3 different priors, 2 block-wise approximations with no prior and ensemble classifier that aggregates the independent, individual decisions made by the $C1$ or the $C2$ classifier. The block-wise approximations are the following: L_2 norm, L_1 norm, LCLC algorithm and MWSRC algorithm. The performance of the L_2 norm, and the L_1 norm approximations are evaluated using no priors, the L_2 norm and the L_1 norm priors (note that the L_2 norm approximations with the L_1 norm prior corresponds to SRC algorithm). In the block-wise L_2 norm approximation with L_2 norm prior the Lagrangian multiplier λ is set to 500, in the block-wise L_2 norm approximation with the L_1 norm prior the Lagrangian multiplier λ is set to 3, in the block-wise L_1

norm approximation with the L_2 norm prior the Lagrangian multiplier λ is set to 3, in the block-wise L_1 norm approximation with the L_1 norm prior the Lagrangian multiplier λ is set to 3, in the L_{LCLC} block-wise approximations the Lagrangian multiplier λ is set to 0.01 and in the L_{MWSRC} block-wise approximations the Lagrangian multiplier λ is set to 7. Table 3 shows the results for the proposed strategy using overlapping blocks.

Table 3. Identification precision using block-wise overcomplete dictionary assuming sparsity prior per all sample data

	G-A_{L2}		G-A_{L1}		G-A_{LCLC}		G-A_{MWSRC}	
	C2	C1	C2	C1	C2	C1	C2	C1
No prior	72.25%	72.323%	72.6480%	72.7467%	-	-	-	-
P_{L2}	72.25%	72.42%	83.3059%	76.9737	**86.27**%	81.9572%	-	-
P_{L1}	**94.2599**%	**94.2928** %	**96.5296**%	**97.0230**%	-	-	**94.398**%	**94.7502**%

Because the data sample variability is unknown in advance, nor is it's spatial distribution known, the independent per block decisions alone can not be reliable. Some blocks decisions might be very noisy while others are not. However, in the general case, on average most of them are not noisy. Therefore the resulting increase in performance shown in Table 3 is due to the fact that the we have an ensemble classifier (48), where the final decision is made by fusing all individual results originating from the overlapping image blocks. This shows us that even a simple overcomplete dictionary (using overlapping image blocks) can increase performance.

The results also highlight one important aspect. That is the impact of the results originating either from the L_2 norm or the L_1 norm classifier ($C2$ or $C1$ respectively) operating on the individual block approximations is negligible in the ensemble classifier as long as on average the majority of decisions about the individual block approximations is correct.

In the end one can conclude that a key for high accuracy is the encoding, which in relationship to identification breaks down in three parts: (1) the used data priors in the encoding; what is used and how it is used as prior in the approximation and coefficients estimation, (2) the method of approximation and (3) the used classifier.

6. CONCLUSION

In this paper we considered some aspects of statistical image modeling in sparse image classification. Our focus was on the encoding that guarantees the best approximation of the query image when considering a dictionary that incorporates varying acquisition conditions. Considering the sparsity as assumption, we analyzed different types of sparsity priors. The first is Sparsity Prior Per Sample Data from All Subjects and the second is Sparsity Prior Per Subject Sample Data. We analyzed two different types of approximations that followed naturally along this line: global and local and two types of image representations using a pixel-wise approximation and an overcomplete block-wise approximation. We presented a comparative analysis to outlier robustness of two classifiers, one related to the residual error measured under L_2 norm and other related to the residual error measured under the L_1 norm. In relation to robustness we also presented and analysed one simple ensemble classifier. This classifier aggregates the results originating from the L_2 norm or the L_1 norm classifier operating on the individual block-wise approximations from the overcomplete dictionary. Our experiments on a publicly available dataset using low resolution images showed that several per subject sample sparsity prior approximations are as good as the results presented from SCR. On the other hand, our very simple over complete block-wise approximation using Sparsity Prior Per Sample Data from All Subjects provides superior performance in comparison to several state-of the-art methods.

ACKNOWLEDGMENTS

The research has been partially supported by the research project PSPB-125/2010. The authors are thankful to Maurits Diephuis for the discussions on the links to the deep learning encoding algorithms and his comments in the early versions of the paper.

7. REFERENCES

[1] Can-Yi Lua,Hai Mina, Jie Gui, Lin Zhua, Ying-Ke Lei "Face recognition via Weighted Sparse Representation", JVCIR, Volume 24, Issue 2, February 2013, Pages 111116.

[2] M.A. Turk, A.P. Pentland, "Face recognition using eigenfaces", IEEE Computer Society Conference on Computer Vision and, Pattern Recognition, 1991, pp. 586591.

[3] P.N. Belhumeur, J.P. Hespanha, D.J. Kriegman, "Eigenfaces vs. sherfaces: recognition using class specic linear projection", IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (1997) 711720.

[4] X.F. He, S.C. Yan, Y.X. Hu, P. Niyogi, H.J. Zhang, "Face recognition using Laplacianfaces", IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (2005) 328340.

[5] T.M. Cover, P.E. Hart, "Nearest neighbor pattern classication", IEEE Transactions on Information Theory 13 (1967) 2127.

[6] S. Shan, W. Gao, D. Zhao, "Face identication from a single example image based on face-specic subspace (FSS)", IEEE International Conference on Acoustic, Speech and, Signal Processing, 2002, pp. II/2125II/2128.

[7] J. Wright, A.Y. Yang, A. Ganesh, S.S. Sastry, Y. Ma, "Robust face recognition via sparse representation", IEEE Transactions on Pattern Analysis and Machine Intelligence 31 (2009) 210227.

[8] E. Elhamifar, R. Vidal, "Robust classication using structured sparse representation", IEEE Conference on Computer Vision and Pattern Recognition, 2011, pp. 18731879.

[9] S. Gao, I. Tsang, L.-T. Chia, "Kernel sparse representation for image classication and face recognition", K. Daniilidis, P. Maragos, N. Paragios (Eds.), Computer Vision ECCV 2010, Springer, Berlin/Heidelberg, 2010, pp. 114.

[10] L.S. Qiao, S.C. Chen, X.Y. Tan, "Sparsity preserving projections with applications to face recognition", Pattern Recognition 43 (2010) 331341.

[11] C.-Y. Lu, "Optimized projection for sparse representation based classication", Advanced Intelligent Computing (2012) 8390.

[12] Qinfeng Shi, Anders Eriksson, Anton van den Hengel, Chunhua Shen, "Is face recognition really a Compressive Sensing problem?", IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 11), Colorado Springs, USA, June 21-23, 2011.

[13] Adam Coates and Andrew Y. Ng., "The Importance of Encoding Versus Training with Sparse Coding and Vector Quantization", ICML 28, 2011.

[14] Adam Coates, "Demystifying Unsupervised Feature Learning", PhD thesis. Stanford University (2012).

[15] Adam Coates and Andrew Y. Ng., "Learning Feature Representations with K-means", In Neural Networks: Tricks of the Trade, Reloaded, Springer LNCS, 2012.

[16] Tibshirani, R. "Regression shrinkage and selection via the lasso", Journal of the Royal Statistical Society, Series B, Vol 58, No. 1, pp. 267288, 1996.

[17] Holland, P. W., and R. E. Welsch. "Robust Regression Using Iteratively Reweighted Least-Squares." Communications in Statistics: Theory and Methods, A6, 1977, pp. 813827.

[18] Dantzig, G.B., A. Orden, and P. Wolfe, "Generalized Simplex Method for Minimizing a Linear Form Under Linear Inequality Restraints," Pacific Journal Math., Vol. 5, pp. 183195, 1955.

[19] Jinjun Wang, Akiira Media Syst., Palo Alto, CA, USA Jianchao Yang; Kai Yu ; Fengjun Lv ; Huang, T. ; Yihong Gong "Locality-constrained Linear Coding for image classification", IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3360 - 3367, 13-18 June 2010.

[20] A.S. Georghiades, P.N. Belhumeur, D.J. Kriegman, "From few to many: illumination cone models for face recognition under variable lighting and pose", IEEE Transactions on Pattern Analysis and Machine Intelligence 23 (2001) 643660.

[21] Michael Grant and Stephen Boyd. CVX: Matlab software for disciplined convex programming, version 2.0 beta. http://cvxr.com/cvx, September 2013.