Reconstruction-free Inference from Compressive Measurements

by

Suhas Anand Lohit

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved March 2015 by the
Graduate Supervisory Committee:

Pavan Turaga, Chair
Andreas Spanias
Baoxin Li

ARIZONA STATE UNIVERSITY

May 2015

ABSTRACT

As a promising solution to the problem of acquiring and storing large amounts of image and video data, spatial-multiplexing camera architectures have received lot of attention in the recent past. Such architectures have the attractive feature of combining a two-step process of acquisition and compression of pixel measurements in a conventional camera, into a single step. A popular variant is the single-pixel camera that obtains measurements of the scene using a pseudo-random measurement matrix. Advances in compressive sensing (CS) theory in the past decade have supplied the tools that, in theory, allow near-perfect reconstruction of an image from these measurements even for sub-Nyquist sampling rates. However, current state-of-the-art reconstruction algorithms suffer from two drawbacks – They are (1) computationally very expensive and (2) incapable of yielding high fidelity reconstructions for high compression ratios. In computer vision, the final goal is usually to perform an inference task using the images acquired and not signal recovery. With this motivation, this thesis considers the possibility of inference directly from compressed measurements, thereby obviating the need to use expensive reconstruction algorithms. It is often the case that non-linear features are used for inference tasks in computer vision. However, currently, it is unclear how to extract such features from compressed measurements. Instead, using the theoretical basis provided by the Johnson-Lindenstrauss lemma, discriminative features using smashed correlation filters are derived and it is shown that it is indeed possible to perform reconstruction-free inference at high compression ratios with only a marginal loss in accuracy. As a specific inference problem in computer vision, face recognition is considered, mainly beyond the visible spectrum such as in the short wave infra-red region (SWIR), where sensors are expensive.

To my grandparents

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Dr. Pavan Turaga for his infinite patience and constant support throughout the duration of this thesis. I am thankful to him for giving me the freedom to explore and work on various topics of interest. I would also like to thank my thesis committee members Prof. Andreas Spanias and Dr. Baoxin Li for taking time off their schedule and serving on my committee. Many thanks to my collaborators Dr. Aswin Sankaranarayanan and Jian Wang at Carnegie Mellon University, Pittsburgh for providing single pixel camera data for experiments, and Kuldeep Kulkarni at ASU for the discussions we had and his inputs, which greatly enhanced this thesis. I also need to thank my graduate advisor Toni Mengert for helping me make my stay in the ECEE department very smooth. I am grateful to Evie Selberg who helped me schedule my thesis defense. Lastly, I thank my parents and my sister for their love and support, without which none of this would have been possible.

TABLE OF CONTENTS

CHAPTER Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 Motivation and Background

"Measure what can be measured" is a quote attributed to Galileo Galilei and has been an important principle for scientific discovery for a long time. However, the world has witnessed an explosion in terms of amount of data generated, stored and analyzed over the last decade which are likely to increase at even faster rates. In this scenario, it is more apt to, as Thomas Strohmer [43] put it, *"measure what should be measured"*. This is a principle espoused by the theory of compressive sensing (CS).

The past decade has witnessed advances in theory and algorithms in the field of CS as well as developments in camera architecture – e.g. single pixel camera (SPC) that allow leveraging the tools supplied by CS theory. The most important aspect of compressive sensing is that it allows near-perfect reconstruction even when a signal is sampled at a rate much lower than the Nyquist rate, provided that the signal to be sampled is sparse in some known basis and the sampling mechanism satisfies certain conditions (explained in Chapter 2). This means that it is possible to sense signals directly in a compressed form, that can be perfectly reconstructed instead of the traditional two-step process of sensing all the measurements followed by compression. This feature of compressive sensing can be exploited in resource-constrained settings as well as in applications where sensing hardware would be otherwise expensive using conventional sensing technologies. Infrared imaging is the perfect example for the latter scenario and is the focus of this thesis. The sensors employed for imaging in the short wave IR region are very expensive. Thus, instead of using a large number of costly sensors, it is more sensible to use an SPC as described later.

However, there are some drawbacks with respect to the recovery of these compressed signals:

- Current state-of-the-art reconstruction algorithms are prohibitively expensive in terms of computation time.
- Near-perfect reconstruction is possible only when the number of measurements is higher than a certain threshold (explained in Chapter 2). Good quality reconstruction results are not possible at high compression ratios.
- Various parameters such as sparsity level of the signal and sparsifying basis need to be input to the algorithm and this is often done in a rather ad-hoc manner.

Consider the case when the images sensed through compressive sampling are used in an inference task such as face recognition. It is reasonable to assume that the recognition needs to be fast. Hence, the focus of research for the past few years has been to design better reconstruction algorithms (overcoming the issues listed above).

Instead of following this line of development, in this thesis, we propose a method to bypass reconstruction entirely and perform inference directly in the compressed domain. That is, we aim to extract features from the compressed measurements that can provide robust high-level inference capabilities. We develop a framework to extract discriminative correlational features from the compressed measurements. Correlational features have been used widely in computer vision for various applications [22] to devise inference algorithms such as face recognition.

1.2 Face Recognition

Recognizing faces - both identification and verification - has gained huge importance over the years, particularly in the area of security and law enforcement. As a specific example, we focus on the problem of face recognition in the NIR spectrum from compressively sensed measurements of the face. Infrared imaging has become attractive sensing modality for face recognition. The reason for this is visible imaging relies on reflected light off the skin which is a function of the illumination and hence, the accuracy of the system may decrease even with small changes in lighting conditions. By using infrared imaging, the problem of illumination variation can be minimized [28]. However, infrared cameras are very expensive, and this has prevented them from them being employed widely for tasks like face recognition. The single-pixel camera (SPC) architecture [46] provides a cost-effective solution for the acquisition problem. The SPC employs a single photodiode and a micro-mirror array to acquire images. This greatly reduces the cost of the camera as a single photodetector, sensitive to wavelengths of interest, is used for data acquisition.

An established method of performing face recognition is by first extracting features from face images and then using pattern recognition techniques for recognition. These features include linear features such as Gabor features [29], PCA [42], LDA [16] etc. and non-linear features such as HOG [12] and LBP [1] and combinations of these features with machine learning algorithms. Recently, deep learning and convolutional neural networks have been employed [44] on very large datasets to achieve very good recognition rates. *At present, it is unclear how to derive non-linear features from compressed measurements.*

1.3 Related Work in Compressed Inference

Calderbank et al. [7] showed theoretically that classifiers can be designed directly in the compressed domain. However, the classifiers are learnt from the compressed data, and do not consider the role of feature-extraction. Extracting features from compressed measurements is the central idea of this thesis. In [19], Haupt et al. investigate the use of CS measurements for signal classification rather than reconstruction i.e., using CS theory as a means of universally applicable non-adaptive dimensionality reduction technique. They derive theoretical results and error bounds that show that it is indeed possible to do so. The misclassification probability is shown to decrease exponentially with the number of measurements. Davenport et al. [10] propose the idea of the smashed filter to perform classification directly on the compressed measurements. A smashed filter is a dimensionality-reduced matched filter. They show how the general maximum likelihood classifier can deal with transformations such as translations and rotations. In this thesis, we design a correlation filter per class that can model many variations in each class and at the same time, maximize inter-class variations. In [39], Sankaranarayanan et al. develop a framework for acquiring CS videos and their reconstruction. This framework is limited to videos that can be modeled as linear dynamical systems such as dynamic textures and human activities. Neifeld and Premachandra [35] propose 'feature-specific imaging' where images are directly measured in the required task-specific basis such as Karhunen-Loeve or wavelet basis. In [31], a compressed sensing architecture is developed where, instead of perfect reconstruction of the CS images, only relevant parts of the scene i.e., the objects are reconstructed efficiently. This is halfway between doing reconstruction and bypassing reconstruction altogether. In [47] and [33], ideas from compressive sensing are used in face recognition with very good results. They rely on finding a sparse code for the test image in terms of the training set vectors which is analogous to reconstruction in compressed sensing.

1.4 Contributions

Following are the main contributions of this thesis:

1. A framework is proposed to extract linear features directly from compressive measurements without recovery using *smashed correlation filters*.

2. It is shown through extensive experiments that inference is indeed possible using these features with only a marginal loss in accuracy, compared to oracle sensing.

3. It is also demonstrated that the performance of this system is barely affected even at high compression ratios, where reconstruction would otherwise fail.

4. A framework, using a convolutional neural network, is proposed as a possible solution to the problem of extracting non-linear features directly from compressive measurements.

1.5 Organization

Chapter 2 describes the basic framework of compressive sensing. Chapter 3 presents details of correlation filters and their applications in pattern recognition. Construction of smashed correlation filters and using them to extract features for reconstruction-free inference are described in Chapter 4. For the specific problem of face recognition, Chapter 5 discusses the experiments performed and the results obtained on various databases. Chapter 6 discusses a possible solution for extracting non-linear features directly from compressive measurements using a convolutional neural network. The last chapter presents conclusions and scope for future research.

Chapter 2

COMPRESSIVE SENSING

The invention and wide-scale deployment of the digital computer has caused a revolution in every aspect of our lives. It has become possible to manufacture more robust and cheaper devices as a consequence of digitization. An important first step that allows moving from the analog to the digital domain is the Nyquist-Shannon sampling theorem. It states that, for band-limited signals, a sampling rate at least twice the maximum frequency component in the input signal, called the Nyquist rate, is sufficient for perfect reconstruction of the signal of interest. That is, the theorem provides a minimum sampling rate that ensures no information is lost during the sampling process.

The result of digital systems being employed everywhere is a huge amount of information being generated. For example, even storing a single image of size 2 million pixels with 8 bits per pixel would require about 4 MB of storage space. Storing and communicating such large signals, usually arising in applications such as imaging and video acquisition for remote surveillance, MRI etc. poses a difficult challenge. It may also be the case that traditional sensing methods are very expensive for emerging sensing modalities such as infrared imaging, which is discussed later in the chapter.

To address the demands of such high volume of data, we often rely on data compression. A popular method of compression is using *transform coding*. Here, a basis is used in which the signal of interest can be expressed very accurately with small number of coefficients. This is called a sparse or compressible representation since the original signal of dimension n can be represented to a high degree of accuracy with only k non-zero coefficients, with $k << n$. An example is the JPEG2000 compression standard for images that relies on the fact that a typical image has a compressible representations in wavelet basis. An image

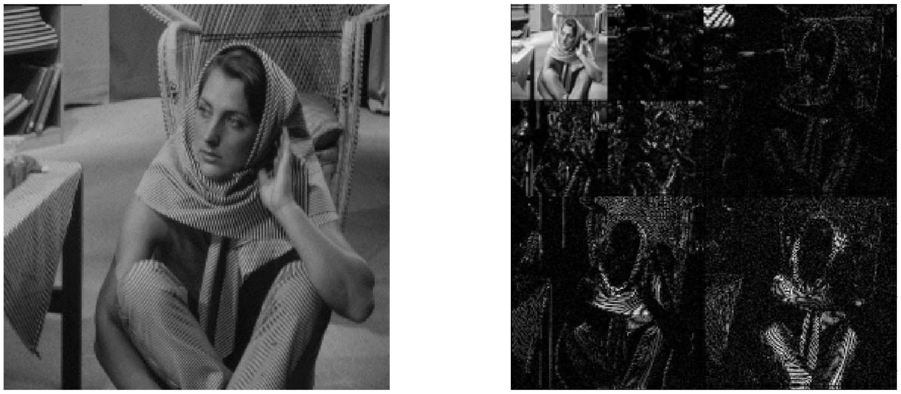

Figure 2.1: Left – Sample image of size 256×256. Right – Wavelet decomposition of the sample image containing lot of dark areas which shows that the image is compressible in wavelet domain.

and its wavelet decomposition (2.1) are shown to illustrate this fact. Clearly the wavelet decomposition has a lot of dark areas showing that the image is indeed compressible in the wavelet basis. For our example of a 2 megapixel image, only about 100,000 wavelet coefficients may be useful. However, the camera has to sense all the 2 million measurements since the knowledge of which coefficients matter is not known a priori.

In the traditional sensing paradigm, there is no way around this problem. However, compressive sensing (CS) provides a way to integrate sampling and compression into a single step. CS is different from classical sampling as follows: (1) Instead of sampling at different points in time or space, CS systems sample by obtaining inner products of the entire signal with pseudo-random basis functions. (2) In the case of classical sampling, signal recovery is achieved through interpolation using sinc functions which is a very simple linear process. In the CS framework, computationally expensive non-linear reconstruction algorithms such as matching pursuit and basis pursuit need to be used.

2.1 Compressive Sampling Mechanism

Exploiting the sparsity of the signal and incoherent sampling, CS theory allows perfect reconstruction of signals sampled at sub-Nyquist rates. Let x, of dimension N and k non-zero values ($k << n$), be the signal to be sampled. Measurements are projections of the signal onto basis functions as shown below:

$$y_k = \langle x, \phi_k \rangle \quad k = 1, 2, \ldots, M, \tag{2.1}$$

In the case of an image, if the basis functions are Dirac delta functions in space, each spike corresponding to a pixel location, then the measurements would be the measurements obtained by a conventional camera. In the case of an arbitrary basis function, the measurement would be a linear combination of the pixel values. In CS framework, we restrict the number of measurements M such that $M << N$. This is shown in Figure 2.2.

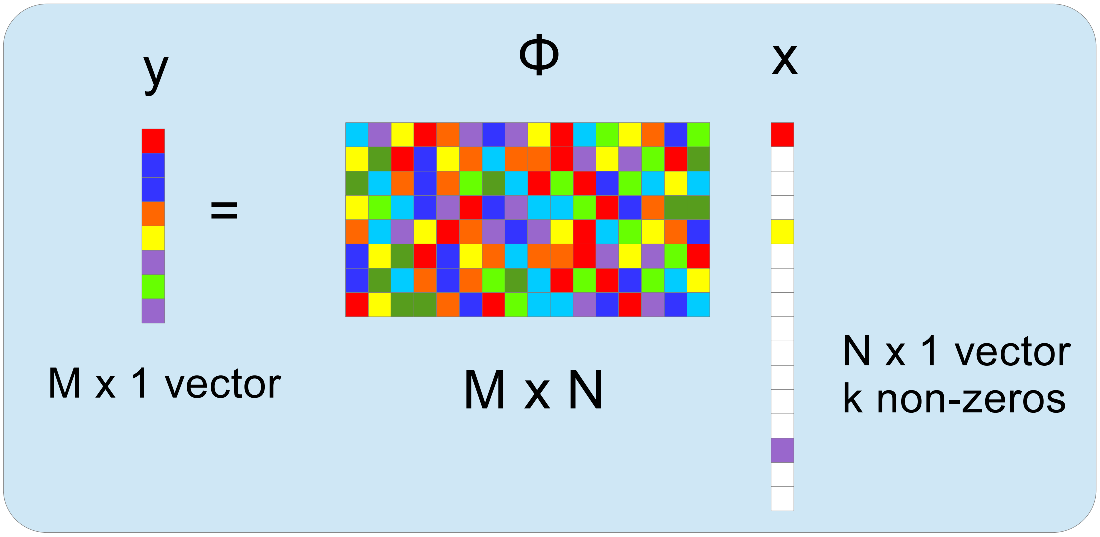

Figure 2.2: Compressive sampling. Each row of Φ is a basis vector. $\mathbf{x}$ is the input signal of dimension N and sparsity level k. $\mathbf{y}$ is the measured vector of dimension M. Note that $k < M << N$.

Representing the sensing matrix as Φ, where the size of Φ is $M \times N$ and each of its rows is a basis vector, we have $\mathbf{y} = \Phi\mathbf{x}$. Clearly, since $M < N$ this forms an underdetermined linear system and in general, computing $\mathbf{x}$ from $\mathbf{y}$ is ill-posed. This is because there are infinite possible solutions $\tilde{\mathbf{x}}$ such that $\mathbf{y} = \Phi\tilde{\mathbf{x}}$. But, as we shall see, with the sparsity prior imposed on $\mathbf{x}$, it becomes possible to recover $\mathbf{x}$ almost perfectly.

2.2 Requirements for Perfect Reconstruction of Compressive Measurements

In order to be able to recover the original signal at a sub-Nyquist rate, two conditions need to be met – (1) the signal needs to be sparse and (2) the sampling must be incoherent with the sparsifying basis. These are explained in more detail below:

2.2.1 Signal Sparsity

A signal is usually modeled as a vector living in a particular vector space or subspace. It is assumed that all vectors in this space are valid signals. However, this does not capture the structure of the signal space. Although the ambient dimensionality of the signal may be high, the number of degrees of freedom may be much lower and the signal can be represented in a lower dimensional model.

For example, natural images can be expressed to a high degree of accuracy with very few coefficients when transformed from the spatial domain to the wavelet domain as shown in Figure 2.2. Mathematically, a vector $v \in \mathbb{R}^N$ can be expressed in an orthonormal basis Ψ of dimension N as

$$v = \sum_{i=1}^{N} x_i \psi_i, \tag{2.2}$$

where $x_i = \langle v, \psi_i \rangle$. In the case of images, when Ψ is the wavelet basis, it is often the case that many x_i's are zero or close to zero. If the signal can be exactly represented with the few non-zero coefficients, it is called a sparse signal. If it can be approximated well by throwing away a large number of coefficients close to zero, it is called a compressible signal. The sparsity level of the signal is defined as the number of non-zero coefficients in its representation. It is denoted as the ℓ_0- quasinorm or $\|\mathbf{x}\|_0$. A signal with k non-zero entries is called a k-sparse signal.

2.2.2 Incoherent Sampling

The intuition for this can be derived by considering the discrete uncertainty principle [15] which says that a non-zero signal with a sparse representation in the time domain has a non-sparse representation in the frequency domain. For a given signal in time domain, with knowledge of only its sparsity level but not the location of the non-zero samples, it should be sampled in the Fourier basis as a small number of measurements is sufficient to reconstruct the original signal.

More formally, let $f \in \mathbb{C}^N$ be a discrete non-zero signal and let $\hat{f} \in \mathbb{C}^N$ be its discrete Fourier transform ($\hat{f} = Ff$). Let T and Ω be the support of f and $\hat{f}$ respectively. Then the following relationships hold:

$$|T|.|\Omega| \geq N \tag{2.3}$$

Since the geometric mean is dominated by the arithmetic mean, we have

$$|T| + |\Omega| \geq 2\sqrt{N} \tag{2.4}$$

Generally, for any pair of orthobases, their relationship, analogous to the above, is captured using the notion of mutual coherence. In the case of CS theory, if Ψ and Φ are the orthobases in which the signal $\mathbf{x}$, of dimension N, is represented and measured respectively, then the coherence [8], $\mu(\Psi, \Phi)$ between Ψ and Φ is given by

$$\mu(\Psi, \Phi) = \sqrt{N}. \max_{1 \leq k,j \leq N} |\langle \psi_k, \phi_j \rangle|. \tag{2.5}$$

It is to be noted that $\mu(\Psi, \Phi) \in [1, \sqrt{N}]$ [14]. For example, with the standard or canonical basis as Φ and the Fourier basis as Ψ, we can achieve maximal incoherence since $\mu(\Psi, \Phi) = 1$. A more important example is that of random matrices, as they are highly incoherent with any fixed basis Ψ. Thus a Gaussian matrix, with i.i.d entries or a Bernoulli matrix with ± 1 entries can be safely used as a sensing matrix without the knowledge of the sparsifying basis. In this sense, these sensing matrices are universal which is one of the many attractive properties of CS theory. With these matrices, the number of CS measurements required for perfect reconstruction with high probability is $O(k \log(\frac{N}{k}))$.

2.3 Restricted Isometry Property (RIP)

This is an important notion that tries to quantify robustness of measurement matrices used in compressive sensing. For each integer $k = 1, 2, \ldots$, the isometry constant δ_k of a matrix A is defined as the smallest number such that

$$(1 - \delta_k)\|\mathbf{x}\|_2^2 \leq \|A\mathbf{x}\|_2^2 \leq (1 - \delta_k)\|\mathbf{x}\|_2^2 \tag{2.6}$$

is true for all k-sparse vectors $\mathbf{x}$. When δ_k is not too close to 1, A is said to obey RIP of order k which means that, A approximately preserves the Euclidean length of k-sparse signals. Equivalently, all subsets of S columns from A need to be nearly orthogonal. This also implies that pairwise Euclidean distances of vectors are preserved in the measurement domain. This guarantees the existence of algorithms for discriminating these sparse vectors in the compressed domain.

We can construct A by sampling N column vectors uniformly at random on a unit sphere in $\mathbb{R}^M$ or by sampling i.i.d entries from a normal distribution with mean 0 and variance $1/M$ or by sampling i.i.d entries from a Bernoulli distribution with equiprobable symbols. It can be shown that, with overwhelming probability, these matrices obey the RIP if the number of rows $M \geq C.k \log(\frac{N}{k})$.

2.4 Signal Recovery

Given CS measurements $\mathbf{y}$, the goal is to find the original $\mathbf{x}$ by finding the sparsest solution that satisfies $\mathbf{y} = \Phi\mathbf{s}$, where $s = \Psi\mathbf{x}$. But, this amounts to minimizing the ℓ_0-norm which is NP hard. In [13], it has been shown that for most large underdetermined linear systems, ℓ_1-norm is equivalent to the ℓ_0-norm. Using the ℓ_1-norm makes the problem convex and still yields the required sparse solution. Thus, the reconstuction problem is posed as an optimization problem as follows:

$$\mathbf{x}^* = \min_{\tilde{\mathbf{x}} \in \mathbb{R}^N} \|\tilde{\mathbf{x}}\|_1 \quad s.t. \quad \mathbf{y} = \Phi\mathbf{s} \quad k = 1, 2, \ldots, M. \tag{2.7}$$

2.5 Reconstruction Algorithms

One of the main areas of CS research has been to devise faster and more accurate algorithms to reconstruct the sparse signal from the CS measurements. We briefly review two fundamental algorithms:

- Basis pursuit denoising (BPDN): Here [9], we solve a quadratic convex optimization problem of the form:

$$\min_x \frac{1}{2}\|y - Ax\| + \lambda\|x\|_1, \tag{2.8}$$

  where λ is the parameter that performs the trade-off between sparsity and quality of reconstruction.

- Matching pursuit: This is a greedy algorithm that finds the best projections of the given vector onto an over-complete dictionary $\mathcal{D}$. Given $\mathcal{D}$, the algorithm finds the atom in $\mathcal{D}$ that has the highest inner product with the vector, subtracts it from the vector, finds the next best atom and so on. Thus, at each step, it iteratively refines the representation. Due to the fact that $\mathcal{D}$ is over-complete, the output is a sparse vector. Extensions of this algorithm exist, such as Orthogonal Matching Pursuit [36, 45] and Compressive Sampling Matching Pursuit (CoSaMP) [34].

Basis pursuit is more robust to noise, but matching pursuit tends to be faster. However, all these algorithms are still very expensive in terms of time complexity.

2.6 Single Pixel Camera

The single pixel camera (SPC) was developed at Rice University [46] as a proof of concept of compressive acquisition of images. The SPC framework is shown in Figure 2.3. The illumination from the scene is projected onto the digital micromirror device (DMD). A DMD consists of millions of micromirrors each representing one of its pixels. Each micromirror can be in one of two states depending on its rotation - on or off. The configuration of the mirrors is encoded as a pseudorandom binary pattern and stored in memory. The light reflected off the DMD is focused onto a single sensor - a photodiode that is sensitive to the required wavelengths. This process optically computes the inner product between the image and the pseudorandom pattern and forms one CS measurement. Different CS measurements are obtained by changing the mirror configuration as many number of times.

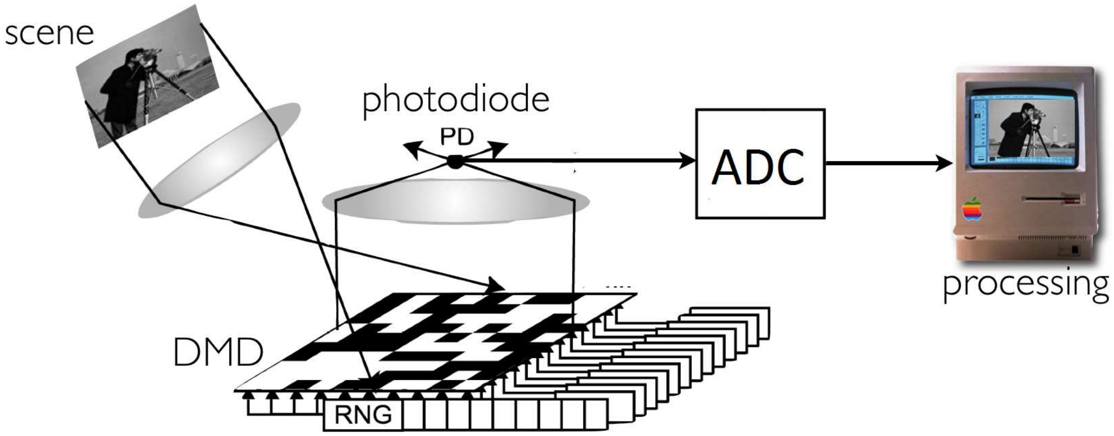

Figure 2.3: Single pixel camera architecture and image acquisition mechanism from [46].

In addition to reducing the dimensionality, there is another advantage to using SPCs. In infrared imaging – e.g. in the short wave infrared (SWIR) range, the pixels are very expensive. The SPC, on the other hand, requires just one photodiode that can sense the appropriate wavelengths. This feature of the SPC, combined with CS theory, is exploited in this thesis to perform infrared face recognition, as explaining in the following chapters.

Chapter 3

CORRELATION FILTERS FOR VISUAL RECOGNITION

By visual recognition, we mean the task of assigning an input image to one of the many predefined classes accurately. An example of this task is face recognition which is the automatic identification or verification of a person from their facial image.

The standard method for visual recognition using labeled data is a three-step process illustrated in Figure 3.1. This procedure is also referred to as supervised learning in the parlance of machine learning. Pre-processing an image may include contrast enhancement, image registration, noise reduction etc. Feature extraction in a method of reducing the amount of data by extracting only the relevant information that might aid in classification. The final step in the pipeline is classification of the features extracted with the help of a classification algorithm – called the classifier – that has been trained using the labeled database.

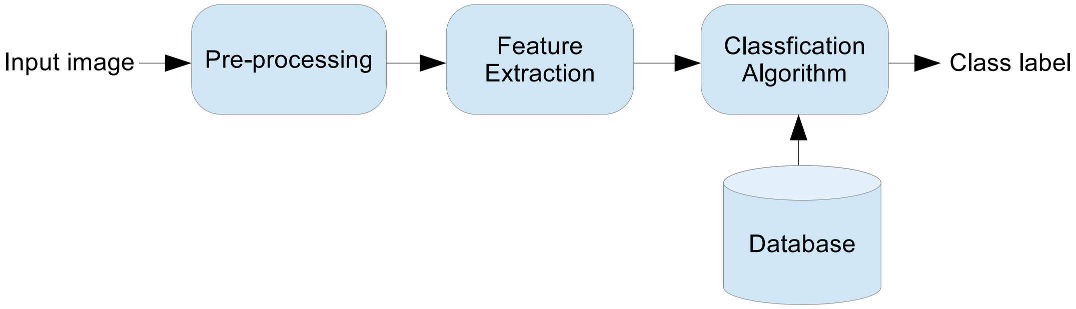

Figure 3.1: Main steps in visual recognition.

Instead of using hand-crafted, application-specific features, the training dataset can itself be used for extracting discriminative features for classification. Correlation filters (CFs) are filters that are generated directly from the training images. One CF is computed per class in the case of a multi-class classification problem. For each class, the CF is designed so as to be able to give a high correlation output only when a test image belonging to that particular class is presented to it. Thus, using the correlation outputs from all the filters for each of the classes, it becomes possible to classify the input image. However, before designing CF, it is necessary to understand the concept of matched filters, which is presented next.

3.1 Matched Filters

Historically, the matched filter (MF) was developed for target detection in the received signal of a radar system using cross correlation. Cross-correlation is a particularly attractive method for pattern recognition since it implicitly provides shift invariance. That is, if the target shifts, so does its correlation, by the same amount. But, the magnitude of correlation at the target location is preserved.

Consider a binary detection problem which requires classifying a received signal $r(t)$ that is a corrupted version of the signal $s(t)$ corrupted by additive white noise $n(t)$. Under these conditions, matched filter is the linear filter $H(f)$ that provides optimal performance by maximizing the output SNR. The impulse response of the matched filter $h(t)$ is proportional to the time reversed version of the transmitted signal i.e., $s(-t)$ [25]. From an optimization point of view, for a signal $\mathbf{x}$ and desired output $\mathbf{g}$, the MF $\mathbf{h}$ is obtained by minimizing

$$r(\mathbf{x}, \mathbf{f}, \mathbf{g}) = \|\mathbf{h} \otimes \mathbf{x} - \mathbf{g}\|, \tag{3.1}$$

where $\otimes$ is the cross-correlation operation. Unfortunately, MFs are not applicable to practical pattern recognition problems. This is because their performance degrades significantly when the test patterns deviate from the template. This motivates the need of more advanced correlation filters (CFs) whose output is more stable when presented with variability at the input for a particular class.

3.2 Correlation Filters

CFs have found various applications such as target recognition [41], object detection [37], face detection [6], face recognition [40] etc. The block diagram of a correlation filter in shown in Figure 3.2. The correlation output usually has a characteristic peak if the input image is a match i.e., it belongs to the same class as that of the correlation filter. Otherwise, no peak is observed.

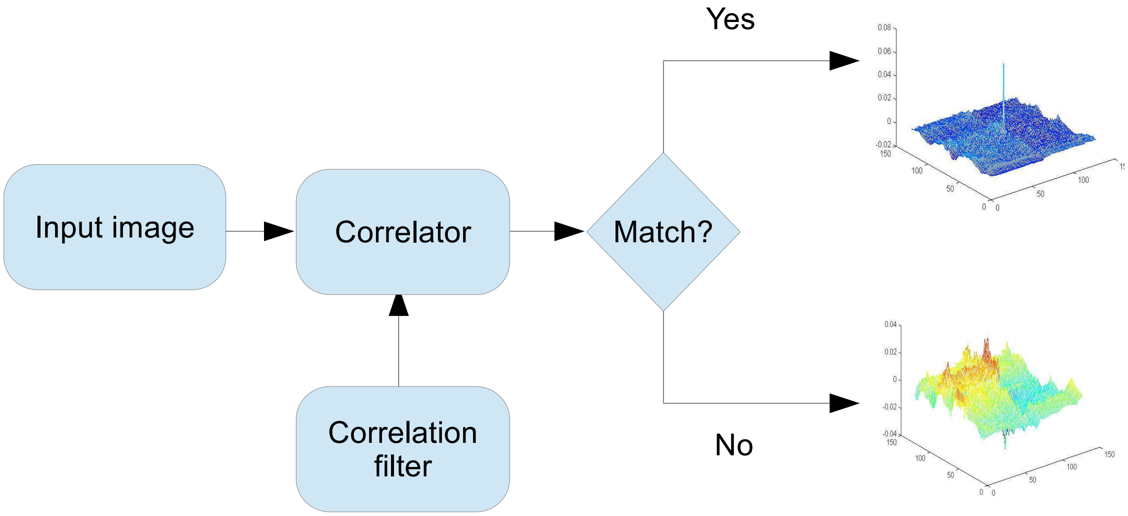

Figure 3.2: Block diagram of correlation based pattern recognition.

Using the correlation outputs called “correlation planes”, a performance measure called the Peak-to-sidelobe ratio (PSR) is defined, that characterizes the sharpness of the peak. PSR is calculated using the formula PSR $= \frac{peak-\mu}{\sigma}$, where μ is the mean and σ is the standard deviation of the correlation values in a bigger region around a mask centered at the peak as explained in [40].

These advanced CFs are designed by using multiple training images per class as well as incorporating regularization in order and improve noise tolerance and generalization. We have two kinds of correlation filters:

3.2.1 Unconstrained Correlation Filters

Here, the correlation filter is derived to be the solution of an optimization problem over the entire training set with n images $\mathbf{x}_i$, $i = 1, 2, \ldots, n$, taking into account both the localization loss as well as regularization:

$$\mathbf{h}^* = \arg\min_{\mathbf{h}} \quad \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{h} \otimes \mathbf{x}_i - \mathbf{g}_i\|_2^2 + \lambda \|\mathbf{h}\|_2^2, \tag{3.2}$$

where λ is the parameter used to control the trade-off between localization and regularization and $\mathbf{g}_i$ is the ideal correlation plane for the i^{th} image. As explained in [5], the optimization problem in 3.2 can be solved efficiently by expressing it in the frequency domain:

$$\hat{\mathbf{h}}^* = \arg\min_{\hat{\mathbf{h}}} \quad \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{h}}^\dagger \hat{\mathbf{X}}_i \hat{\mathbf{X}}_i^\dagger \hat{\mathbf{h}} - \frac{2}{n}\sum_{i=1}^{n} \hat{\mathbf{g}}_i^\dagger \hat{\mathbf{X}}_i \hat{\mathbf{h}} + \lambda \hat{\mathbf{h}}^\dagger \hat{\mathbf{h}}, \tag{3.3}$$

where $\hat{\mathbf{x}}$ is the DFT of $\mathbf{x}$ and $\hat{\mathbf{X}}$ is a diagonal matrix with the elements of $\hat{\mathbf{x}}$ on its diagonal. By solving this optimization problem, we get a closed form expression for the optimal CF:

$$\hat{\mathbf{h}}^* = \left[\lambda I + \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{X}}_i \hat{\mathbf{X}}_i^\dagger\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{X}}_i \hat{\mathbf{g}}_i\right]. \tag{3.4}$$

Depending on the value of $\mathbf{g}$, the CF obtained from 3.4 can be an Unconstrained Minimum Average Correlation Energy (UMACE) filter [30], Maximum Average Correlation Height (MACH) filter [32] etc.

3.2.2 Constrained Correlation Filters

In equality constrained correlation filters, in addition to the minimizing the objective in (3.2), we also constrain the correlation value obtained at the target location for each training image to a particular value c_i. This results in the modified optimization problem:

$$
\begin{aligned}
\mathbf{h}^* = \arg\min_{\mathbf{h}} \quad & \frac{1}{n}\sum_{i=1}^{n} \|\mathbf{h}\otimes\mathbf{x}_i - \mathbf{g}_i\|_2^2 + \lambda\|\mathbf{h}\|_2^2 \\
s.t. \quad & \mathbf{h}^T\mathbf{x}_i = c_i, \quad i = 1, 2, \ldots, n.
\end{aligned} \tag{3.5}
$$

As before, by transforming this to the frequency domain, we get

$$
\begin{aligned}
\hat{\mathbf{h}}^* = \arg\min_{\hat{\mathbf{h}}} \quad & \frac{1}{n}\sum_{i=1}^{n} \hat{\mathbf{h}}^\dagger\hat{\mathbf{X}}_i\hat{\mathbf{X}}_i^\dagger\hat{\mathbf{h}} - \frac{2}{n}\sum_{i=1}^{n}\hat{\mathbf{g}}_i^\dagger\hat{\mathbf{X}}_i\hat{\mathbf{h}} + \lambda\hat{\mathbf{h}}^\dagger\hat{\mathbf{h}} \\
s.t. \quad & \hat{\mathbf{h}}^T\hat{\mathbf{x}}_i = c_i, \quad i = 1, 2, \ldots, n.
\end{aligned} \tag{3.6}
$$

By choosing appropriate values for $\mathbf{g}$ and λ, we can derive the Minimum Average Correlation Energy (MACE) filter [30] and Optimal Trade-off Synthetic Discriminant Function (OTSDF) filter [24]. The general solution to the above has a closed form expression given by

$$
\hat{\mathbf{h}}^* = \hat{\mathbf{S}}^{-1}\hat{\mathbf{X}}(\hat{\mathbf{X}}^\dagger\hat{\mathbf{S}}^{-1}\hat{\mathbf{X}})^{-1}\mathbf{c}, \tag{3.7}
$$

where $\hat{\mathbf{X}}$ is the data matrix with each column corresponding to the vectorized version of DFT of the training images. $\mathbf{S} = \gamma\mathbf{I} + \frac{1}{n}\sum_{i=1}^{n}\hat{\mathbf{X}}_i^\dagger\hat{\mathbf{X}}_i$. With $\gamma = 0$, we get the MACE filter.

3.3 Maximum Margin Correlation Filters

Maximum margin correlation filters (MMCFs), used later in this thesis, are a kind of constrained correlation filters with non-equality constraints. Traditional CFs, although they provide very good localization, do not perform as well as SVMs in terms of generalization. MMCFs combine the strengths of traditional CFs and the SVM. Specifically, they are designed to provide:

(i) High PSR: The correlation output is high only at the target location; it is much lower at all other locations. This criterion comes from the earlier work in MACE filters.

(ii) Max-margin: A large margin between the positive and negative training examples is achieved. This is the objective of the SVM and it reduces to correlations at the center of the image being well-separated.

The above goals are achieved by solving an optimization problem of the form:

$$\begin{aligned} \min_{\mathbf{h},b} \quad & (\|\mathbf{h}\|_2^2 + C\sum_{i=1}^{N}\xi_i, \sum_{i=1}^{N}\|\mathbf{h}\otimes\mathbf{x}_i - \mathbf{g}_i\|_2^2) \\ \text{s.t.} \quad & t_i(\mathbf{h}^T\mathbf{x_i} + b) \geq c_i - \xi_i, \quad i = 1, 2, \ldots, N, \end{aligned} \tag{3.8}$$

where the minimizer $\mathbf{h}^*$ is the required MMCF, $\mathbf{g}_i$ is the desired value of the correlation output, $c_i = 1$ for images in the true class and $c_i = 0$ for those in the false class. ξ_i are the positive slack variables that take care of outliers and $t_i \in \{-1, 1\}$ are the labels. As shown in [37], with appropriate transformations, (3.8) can be reduced to a single optimization problem that can be solved on any standard SVM solver.

In this thesis, we employ MMCFs in our compressive framework in order to perform face recognition. This is explained in detail in the next chapter.

Chapter 4

FACE RECOGNITION USING SMASHED CORRELATION FILTERS

In this chapter we describe a method to perform face recognition using measurements obtained from a single pixel camera (SPC), *without reconstruction*. We employ MMCFs (described in section 3.3) trained on the full-blown face images and use them to construct smashed filters [10] as described later in this chapter. The theoretical basis for compressed inference using correlation filters is provided by the Johnson-Lindenstrauss (JL) lemma.

Consider a training set of P non-compressed face images belonging to Q classes. Let each such image be denoted by $X_p, p = 1, 2, \dots, P$. Q MMCFs, one for each of the Q classes, are trained by solving an optimization of the form shown in Equation 3.8. Let us denote each MMCF thus obtained by $H_q, q = 1, 2, \dots, M$. Let each image be of size $N = N_1 \times N_2$ pixels. Each H_q is also of the same size. Face recognition in the traditional framework is done using full-blown images as shown in Figure 4.1. We will call this the oracle method and use this to compare the accuracy of face recognition with compressed measurements.

Each test image is correlated with each H_q in order to obtain Q correlation planes, $c_q, q = 1, 2, \dots, Q$. The size of c_q is $(2N_1 - 1) \times (2N_2 - 1)$. The correlation plane is given by the equation:

$$c_q(i, j) = \sum_{k=0}^{N_1-1} \sum_{l=0}^{N_2-1} X(k, l) H_q(k - i, l - j). \tag{4.1}$$

This can be written in the form of an inner product as

$$c_q(i, j) = \langle X, H_q^{i,j} \rangle, \tag{4.2}$$

where $H_q^{i,j}$ is the shifted version of H_q by i and j units in the x and y directions respectively. In the case of compressed measurements obtained from the SPC, the high-dimensional

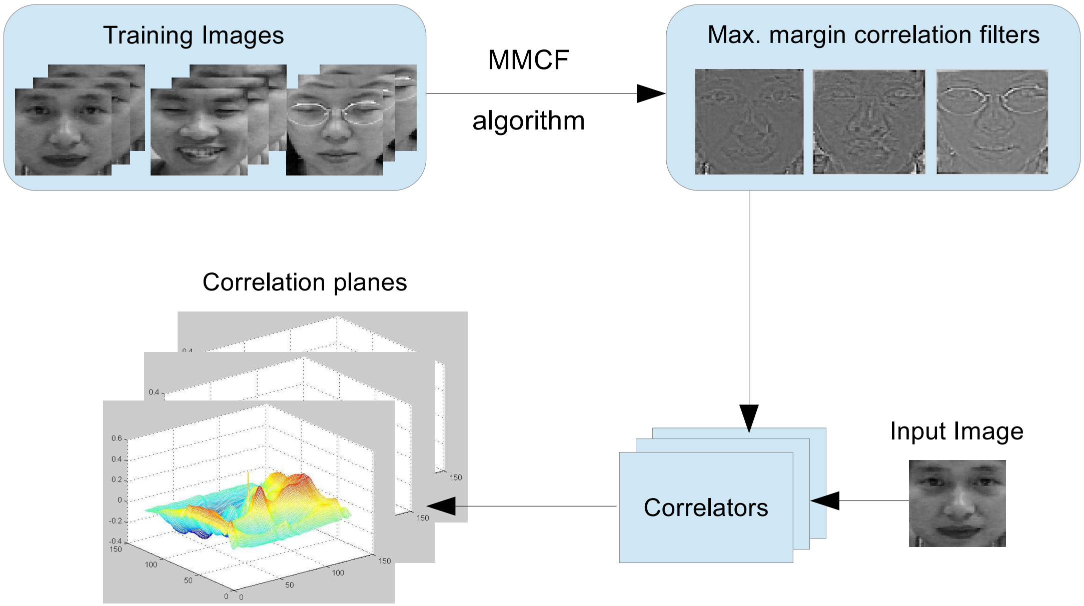

Figure 4.1: Face recognition in a traditional correlation pattern recognition framework - oracle

image space, say $\mathbb{R}^N$ is mapped to a lower dimensional $\mathbb{R}^M$. We have access only to the compressed measurements ΦX. Since our aim is to bypass reconstruction, we need to be able to use the same correlational framework for the compressed sensing case. This is where we resort to the JL lemma, the details of which are explained in the next section.

4.1 JL Lemma and Correlation-preserving Maps

According to the Johnson-Lindenstrauss lemma, certain embeddings exist that preserve the general geometric relations of a set of points in a high-dimensional space, when mapped to a lower dimensional space. In the case of compressive sensing, the mapping is provided by a pseudo-random sensing matrix, Φ. It is stated more formally as follows:

Given $0 < \epsilon < 1$, a set X of P points in $\mathbb{R}^N$, and a number $M > N_0 = \mathcal{O}(\frac{log(P)}{\epsilon^2})$, there exists a linear map $f : \mathbb{R}^N \rightarrow \mathbb{R}^M$ such that

$$(1-\epsilon)\|v-u\|^2 \leq \|f(v)-f(u)\|^2 \leq (1+\epsilon)\|v-u\|^2 \tag{4.3}$$

It is required that the mapping be at least Lipschitz continuous, which limits the rate at which the function changes.

Recently, the manifold structure of random projections has been studied in detail. In [3], using the JL lemma, it is shown that, by using random projections of data points on a smooth manifold, all pairwise geodesic distances are preserved with high probabiity. In [21] and [17], manifold learning algorithms are developed using random projections. This is related to the thesis in that the distance preserving property of these random projections is used to perform inference in the compressed domain. The JL lemma is also intimately connected with the Restricted Isometry Property (RIP) of sensing matrices in CS theory [2].

Using the above results, as proven in [11], we get an expression for the correlation values as

$$c_m(i,j) - \epsilon \leq \langle f(X), f(H_m^{i,j}) \rangle \leq c_m(i,j) + \epsilon. \tag{4.4}$$

Even though the JL lemma does not tell us how to find f, as long as it satisfies RIP, we can construct f as a matrix Φ of size $M \times N$. Given that $M \geq C.k \log(\frac{N}{k})$, Φ satisfies RIP with overwhelming probability [8] if:

- the entries are i.i.d realizations of a standard Gaussian or
- the entries are i.i.d realizations of a Bernoulli random variable

Using one of the above matrices as the sensing matrix, we can rewrite Equation 4.4 as

$$c_m(i,j) - \epsilon \leq \langle \phi X, \phi H_m^{i,j} \rangle \leq c_m(i,j) + \epsilon. \tag{4.5}$$

Thus, correlation values are preserved in the lower dimensional mapping, to a certain degree of accuracy determined by ϵ. This provides the theoretical foundation for compressive classification described in the next section.

4.2 Face Recognition with Smashed MMCFs

The $H_m^{i,j}$ from (4.5) is called the smashed filter [10]. By employing these filters, we can modify the traditional method of face recognition with correlation filters shown in Figure 4.1. The block diagram for face recognition with smashed correlation filters is shown in Figure 4.2.

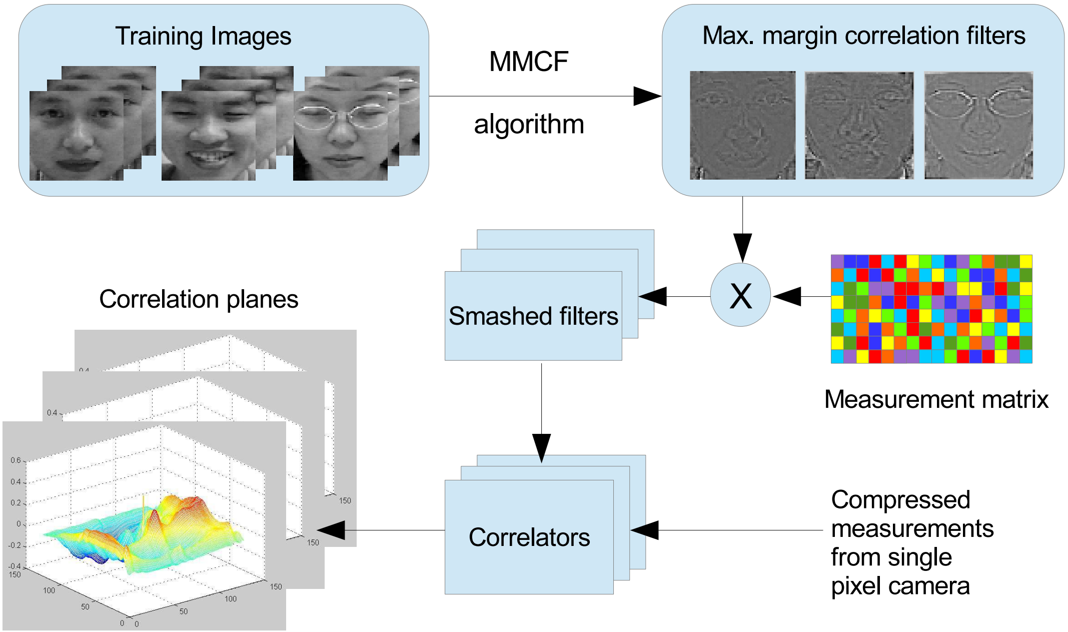

Figure 4.2: Face Recognition in a compressed sensing framework with smashed MMCFs

Equation 4.5 the correlation outputs of the compressed measurements can be obtained to a certain degree of accuracy (determined by the number of measurements) without reconstruction. Clearly, if there are Q subjects, Q correlation filters are trained and Q correlation planes are obtained for each test ”image”, as shown in the block diagram.

Each correlation plane is divided into non-overlapping blocks and for each block, the peak and peak to side-lobe ratio (PSR) are determined. PSR is calculated using the formula PSR $= \frac{peak-\mu}{\sigma}$, where μ is the mean and σ is the standard deviation of the correlation values in a bigger region around a mask centered at the peak as explained in [40]. The peaks and

PSRs of the different blocks are concatenated. Similar vectors are obtained for the each of the Q correlation planes. All these vectors are concatenated to form a single feature vector for the particular test image. This feature vector is input into Q linear SVMs for a one vs all classification. It is to be noted that the SVMs are trained on feature vectors obtained in the same fashion from the training set.

Chapter 5

EXPERIMENTS AND RESULTS

In this chapter, we test the framework described in the previous to perform face recognition in the compressed domain without reconstruction. We carry out two sets of experiments:

(i) Controlled experiments: Here, we use publicly available databases - NIR and AMP and proceed to train the max–margin correlation filters. For testing, we use the testing set of the database and simulate the process of getting the compressed measurements in software. We carry out the face recognition experiment at different compression ratios and different noise levels. We study the effect of using different measurement matrices - (a) Gaussian random matrix, Φ^G (b) Low rank column permuted Hadamard matrix, Φ^H and (c) Simple downsampling, Φ^D.

(ii) Experiments on the single pixel camera (SPC): We perform a similar face recognition experiment on actual compressed measurements of face images. The compressed data were obtained through a collaboration with researchers at Carnegie Mellon University, Pittsburgh, who have built an SPC.

5.1 Controlled Experiments

In this section, we use two publicly available face datasets in order to perform compressive face recognition by simulating the process of compressive sensing in software. The two datasets are described below:

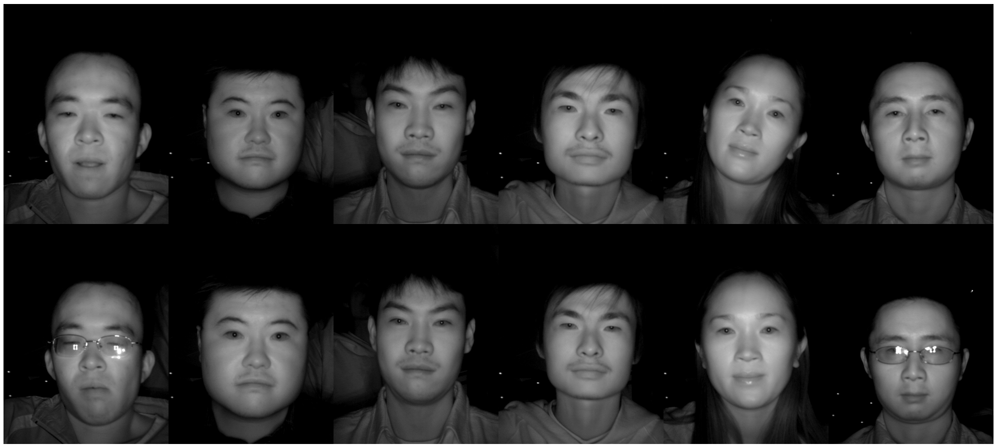

Figure 5.1: Sample images from the NIR database.

5.1.1 NIR Database

The NIR database [28] consists of near infrared images of 197 subjects with 20 images per subject in grayscale. The images have been captured using an active NIR imaging system that is shown to be able to produce high quality images irrespective of the surrounding lighting conditions. Each image was resized to 256 × 256 from the original size of 640 × 480. For each subject, 10 images are used for training and the remaining 10 images are used for testing. Sample images from the dataset are shown in Figure 5.1.

5.1.2 AMP Database

The AMP database[1] is a facial expression database compiled by the Advanced Multimedia Processing lab at CMU. The dataset consists of 975 grayscale facial images belonging to 13 people, each of size 64 × 64 pixels. Sample images from the database are shown in Figure 5.2. For each subject, 25 images are used for training and the remaining 50 images are used for testing.

[1] http://chenlab.ece.cornell.edu/projects/FaceAuthentication/

Figure 5.2: Sample images from the AMP database

5.2 Training and Testing Protocol

Here, we describe the protocol used for both datasets. We describe in detail the parameters chosen for the NIR database only where images are of resolution 256×256 (obtained after rescaling). The AMP database uses corresponding scaled parameters suitable to the image-size 64×64. For the NIR database, the MMCFs, one for each of the 197 classes, are trained using the 256×256 images from the training set. Then, each image in the dataset is vectorized to get a vector $\mathbf{x}_i$. The process of getting the measurements $\mathbf{y}_i$ from a single pixel camera is simulated using the equation $\mathbf{y}_i = \Phi\mathbf{x}_i$, where Φ is the measurement matrix.

First, the measurement matrix is chosen to be a Gaussian matrix (Φ^G) such that the entries of the matrix are i.i.d. standard Gaussian. The number of rows, M of the matrix corresponding to the number of compressed measurements is varied. Three values of M are chosen – 65536, 625 and 121 corresponding to compression ratios CR $= 1, 105, 542$ respectively.

Then, according to equation 4.5, the trained correlation filters $\{H_i\}$ are also compressed to obtain the smashed filters $\tilde{H}_i = \Phi H_i$. As explained in Chapter 4, each compressed image in the training set is correlated with all the smashed filters to obtain 197 correlation planes. Using Equation 4.5 directly to get the correlation estimates $\hat{c}(i, j)$ is not efficient since it requires that the smashed filter be computed separately for each shift (i, j). Instead, equivalently, we first project the compressed measurements $\Phi\mathbf{x}$ back into the pixel space by premultiplying with Φ^T. That is, we compute

$$\hat{c}(i, j) = \langle \Phi^T \Phi \mathbf{x}, H^{i,j} \rangle. \tag{5.1}$$

This can be computed efficiently using the FFT. Then, each correlation plane is divided into $B = 16$ square non-overlapping blocks (of size 128×128) and the PSR and peak values of each block is extracted. These values, in addition to the PSR and peak value for the entire correlation plane, are concatenated to form a feature vector of size 1×6698. These features are used to train 197 linear SVMs, one for each class. In the testing phase, feature vectors of the compressed images are obtained in the same fashion and input to the trained SVMs for a one vs all classification. The accuracy of the face recognition system is determined as the ratio of number of correctly recognized faces to the total number of faces.

The above experiment is then repeated with Φ^H , the matrix containing a random subset of rows of a permuted Hadamard matrix. As before, the accuracy is determined for different numbers of measurements, M (the number of randomly chosen rows of Φ^H. Finally, the images in the dataset are downsampled by the same factors and the same experiment is carried out.

Next, the effect of adding noise on face recognition accuracy is considered. Each of the above experiments is repeated after adding measurement noise – Gaussian noise – of standard deviation σ calculated using $\sigma = \eta \frac{\|\Phi\mathbf{x}\|}{\sqrt{M}}$, where $\eta = 0, 0.1, 0.2, 0.3$ is the noise level. The recognition accuracies are determined at each noise level at each compression factor for each of the measurement matrices. Figure 5.3 shows the variation of accuracy with respect to noise, CR held constant.

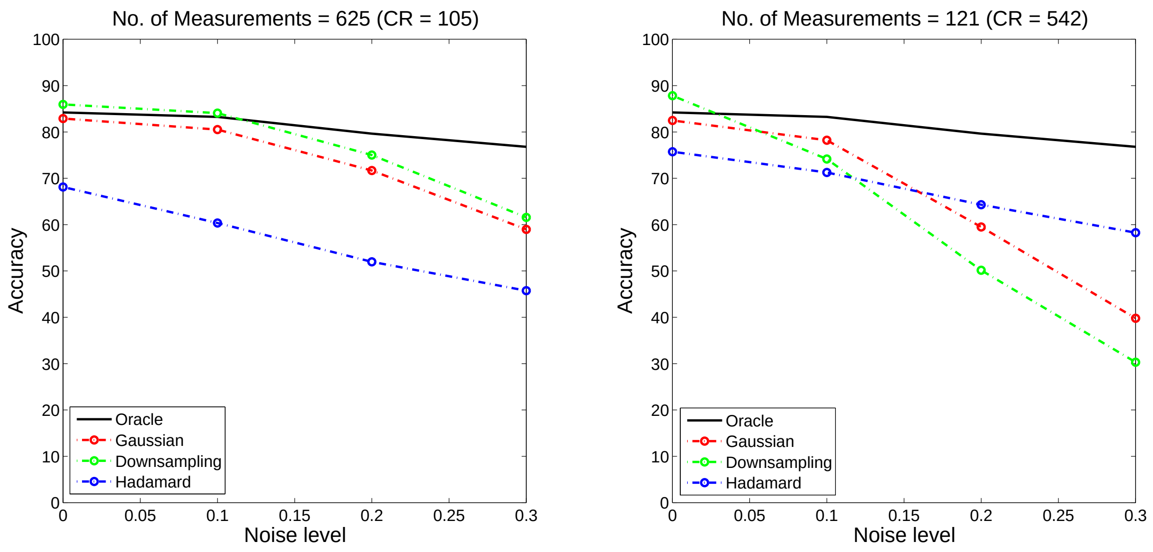

Figure 5.3: The figures show the variation of recognition accuracy for the NIR database for Oracle (no compression), Gaussian measurements, low-rank permuted Hadamard measurements, downsampling, for varying amounts of measurement noise. Note that results indicate that performance is close to Oracle for low-noise levels, and Hadamard is more stable in performance than Gaussian and downsampling operators

Similar experiments are conducted on the AMP database at two CRs of 28 and 114. Since there are 13 subjects, 13 correlation filters are trained, each filter corresponding to one of the subjects. Each correlation plane is divided into $B = 4$ blocks and features are computed similar to above. Figure 5.4 shows the variation of accuracy with respect to noise, CR held constant.

From the plots above, we make the following important observations:

(i) At low noise levels, reconstruction-free inference results at high compression ratios are very close to the results obtained with Oracle sensing (no compression).

(ii) Hadamard measurement matrices are much more robust to noise, especially at high compression ratios, compared to Gaussian and downsampling measurements.

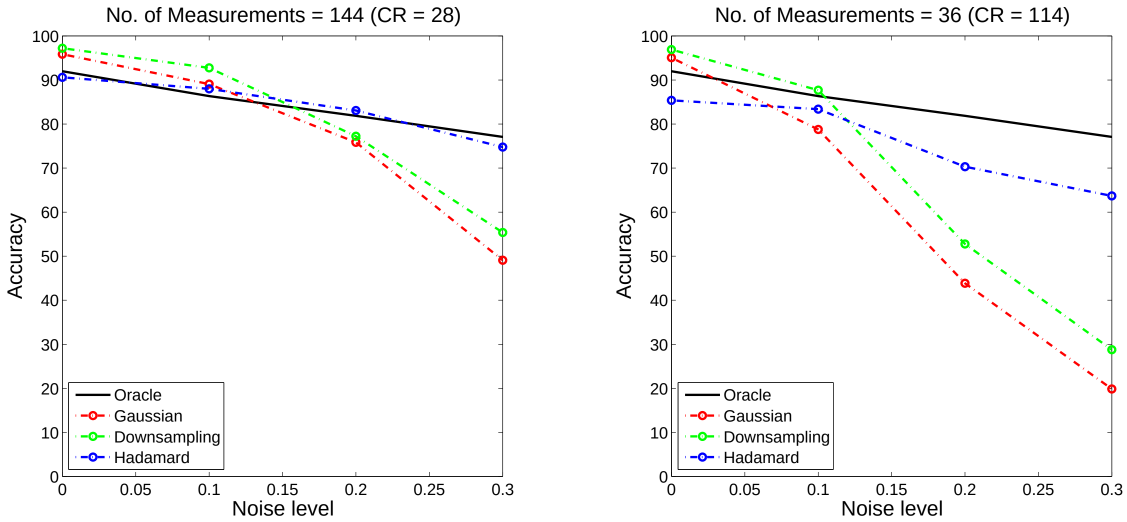

Figure 5.4: The figures show the variation of recognition accuracy for the AMP database for Oracle (no compression), Gaussian measurements, low-rank permuted Hadamard measurements, downsampling, for varying amounts of measurement noise. Note that results indicate that performance is close to Oracle for low-noise levels, and Hadamard is more stable in performance than Gaussian and downsampling operators

5.3 Experiments on Single Pixel Camera

The single pixel camera we used to obtain data was built by researchers at Carnegie Mellon University. It uses a digital micro-mirror device (DMD) with a resolution of 1024×768 and changes the micro-mirror configurations at a frame-rate of 22.7 kHz. The measurement rate of the SPC is determined primarily by the operating speed of the DMD; hence, we obtain 22.7k measurements per second. For example, to capture an image of resolution 128×128 without CS recovery, we would need 0.72 seconds at the operating rate of 22.7kHz. With CS, this can be reduced to as little as 0.1 seconds without significant loss in quality.

Based on the observations from the controlled experiments reported in section 5.1 and because it is easy to implement in hardware, we use a permuted Hadamard matrix for sensing. More specifically, for an $N \times N$ image, we first generate a $N^2 \times N^2$ column-permuted Hadamard matrix. Each row of this matrix is shaped into an $N \times N$ image that is upsampled and mapped to the 1024×768 mircomirror array. Given that the DMD can only direct light towards or away from the photodetector, this implements a $0/1$ measurement matrix. To obtain measurements corresponding to the ± 1 Hadamard matrix, we subtract half the average light level from the observed measurements in post-processing.

The new dataset consists of 120 face images, belonging to 30 subjects with 4 images per subject. The images are captured using the SPC at a resolution of 128×128. The dataset is divided into four train-test splits. For each split, the train set consisted of three images per subject, and the test set contained one image per subject. The recognition experiment was conducted at various compression ratios. The results are shown in Table 5.1.

Compression ratio	No. of Measurements	Recognition Accuracy
1 (Oracle)	16384	60%
10	1638	62.5%
50	328	58.33%
100	164	53.33%
200	82	49.17%

Table 5.1: Face Recognition results obtained on compressed measurements from a single pixel camera.

Reconstruction failure Here, we demonstrate that, for high compression ratios (CR), inference is not possible even after reconstruction using state-of-the-art algorithms. From the SPC measurements of a face image, we reconstruct a face image using the CoSaMP algorithm [34] at compression ratios of 5, 10 and 100 as shown in Figure 5.5. Clearly, reconstructed images at high CRs retain no valuable information that can be exploited for inference. Hence, we need to employ a framework – such as the one described in this thesis – for direct inference on compressed measurements.

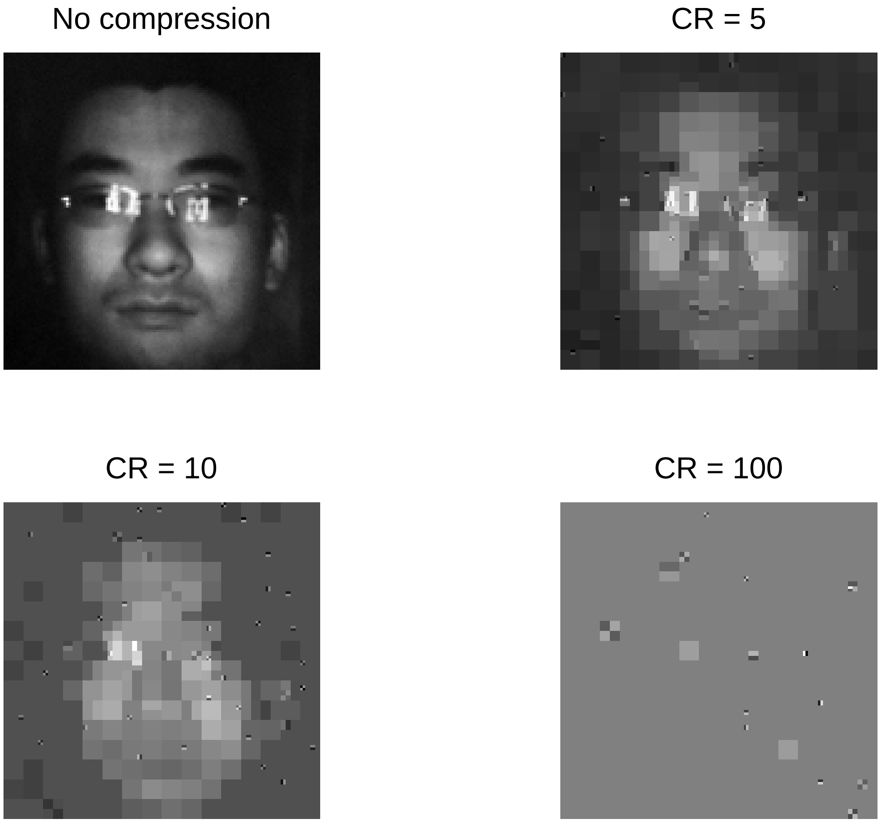

Figure 5.5: The figures show the reconstruction of images of a face at different compression ratios (CR) using the CoSaMP [34] algorithm. Note how reconstruction quality degrades very rapidly across compression rates which makes 'reconstruction-then-inference' a losing proposition.

5.4 Error Analysis

In this section, we try to visualize the error in correlation estimation produced due to random projection. As shown in Equation 4.5, this error is quantified in terms of ϵ, which is the absolute difference between correlation in the original domain, c and correlation in compressed domain, c^{comp}. That is, at each location (i, j), ϵ can be defined as

$$\begin{aligned} \epsilon &= |c(i,j) - c^{comp}(i,j)| \qquad (5.2) \\ &= |\langle H^{i,j}, X\rangle - \langle \Phi H^{i,j}, \Phi X\rangle| \qquad (5.3) \end{aligned}$$

In order to visualize this error, we considered 2 face images belonging to different subjects, X_1 and X_2 and the 2 corresponding filters H_1 and H_2 from the AMP dataset. Each image is correlated with both the filters to obtain the correlation planes c_{11}, c_{12} for X_1 with H_1 and H_2 respectively and c_{21}, c_{22} for X_2 with H_1 and H_2 respectively. The compressed measurements of the images are computed at CR = 28 and the corresponding correlation estimates in the compressed domain are obtained – $c_{11}^{comp}, c_{12}^{comp}, c_{21}^{comp}, c_{22}^{comp}$. Using Equation 5.2, ϵ_{11}, ϵ_{12}, ϵ_{21} and ϵ_{22} are computed respectively. The results are shown in Figure 5.6. Next, each ϵ is cross-correlated with all the other ϵ's. The plots obtained from this are shown in Figure 5.7. It can be observed that there is significant correlation between the ϵ's. This means that the error in the correlation estimates are correlated with each other irrespective of the input image or the class that the image belongs to. Thus, in the case of compressed measurements, even though the correlation plane itself does not seem to show the peak that is required to classify the images, it nevertheless contains the information required for classification. This also motivates the use of a classifier – SVM in our case – that uses a feature extracted from these correlation planes as described in Section 5.2.

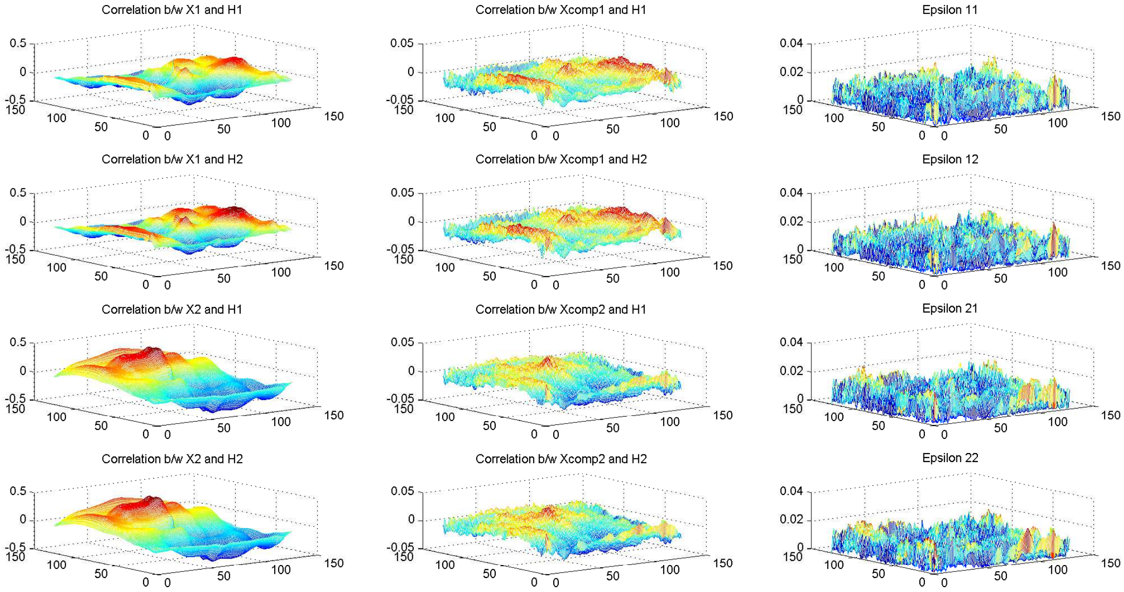

Figure 5.6: The left column shows the correlation planes of two images X_1 and X_2 (oracle sensing) belonging to different classes with corresponding filters H_1 and H_2. The middle column shows the correlation plane estimates obtained with compressed measurements. The right column shows the difference, ϵ, between the correlation planes for oracle sensing and the correlation plane estimates obtained using compressed measurements.

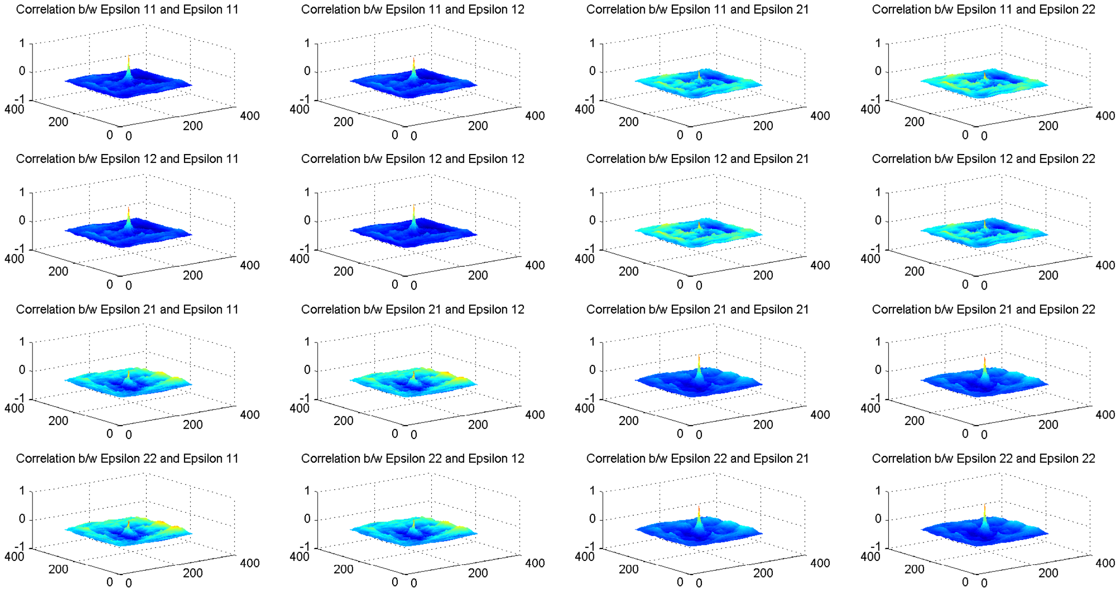

Figure 5.7: Each ϵ from Figure 5.6 is cross-correlated with the all the other ϵ's. Note the prominent peak at the center of each correlation plane. This shows that the errors in correlation estimates obtained using compressed measurements for different images and filters are all correlated with each other.

Chapter 6

EXTRACTING NON-LINEAR FEATURES USING CONVOLUTIONAL NEURAL NETWORKS

The previous chapters have shown how to extract simple linear features from compressive measurements using correlation filters for visual inference. The next step would be to design a framework capable of extracting non-linear features from these measurements. Hand-crafted and usually task-specific non-linear features – e.g. Histograms of Oriented Gradients (HOG) and Local Binary Patterns (LBPs) – are extracted from pixel values and have been shown to be very useful for computer vision applications such as image recognition. However, it is not straightforward as to what kind of non-linear features can be extracted from compressive measurements that can be used reliably for high-level inference. In this chapter, we show that feature learning using a multi-layered convolutional neural network (CNN) is a possible solution to this problem.

The CNN architecture has been shown to be very good at tasks like hand written digit recognition [26]. More recently, using deep CNNs (with many layers), huge leaps have been possible in a variety of computer vision tasks such as image recognition [23], object detection [18], face recognition [44] etc. The success of such architectures has been attributed to the network's ability to learn an optimal set of rich, discriminative features directly from training set rather than relying on hand-crafted features.

CNNs can be very briefly described as follows. They are multi-layered neural network architectures [4] that contain many layers of small groups of neurons that are active to small regions in an input image. The output from groups of neurons of one layer are combined in the next layer and so on. This results in some amount of translational invariance in the input. The main difference between conventional neural networks and CNNs is that the not all neurons in a layer are connected to every neuron from the previous layers. Thus the weight matrix, that contains the weights between all pairs of neurons of consecutive layers, is sparse. This results in a reduction in the number of parameters that need to be optimized in the network and better convergence.

6.1 MNIST Database

The MNIST database[1] contains hand written digits of size 28×28 in grayscale. The dataset is already divided into three files - training set containing 50000 images, validation set containing 10000 images and the testing set containing 10000 images. Each image belongs to one of 10 classes – digits 0 through 9. MNIST was chosen since training a CNN for this database is simple and does not need special hardware. Sample images are shown in Figure 6.1.

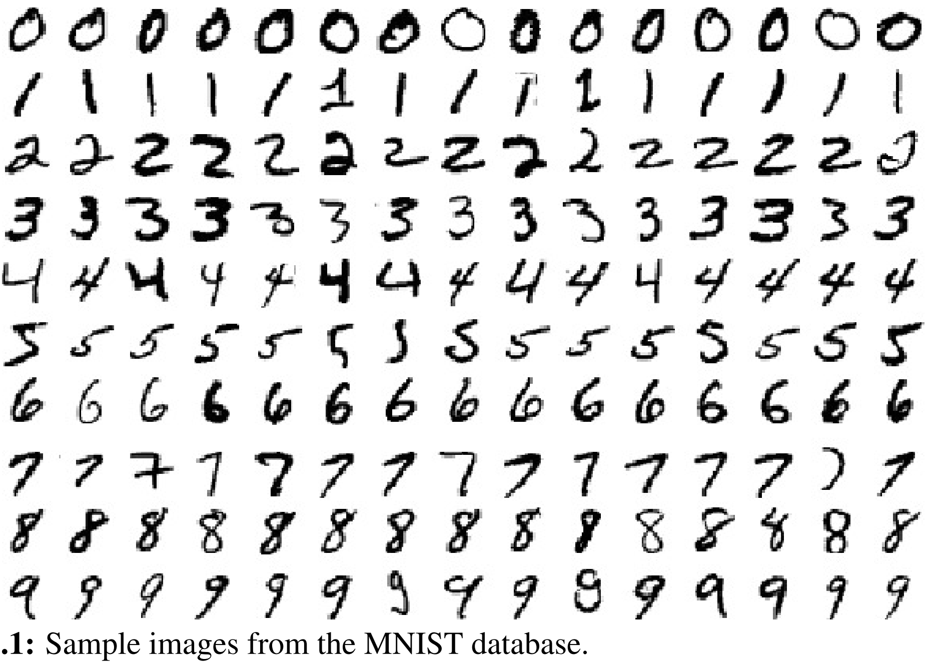

Figure 6.1: Sample images from the MNIST database.

[1] http://yann.lecun.com/exdb/mnist/

6.2 CNN Architecture

The architecture we use in this thesis is based on the LeNet-5 model [27] with two convolutional (convolution followed by tanh non-linearity) and max-pooling layers followed by a single fully connected layer and a 10-way softmax classifier. This is shown in Figure 6.2. The input image is fed into the first convolutional layer that consists of 20 filters of size 5×5 and produces 20 feature maps of size 24×24. This is followed by a maxpooling layer that reduces the size of feature maps to 12×12. The second convolutional layer contains 50 filters of size 5×5 and produces 50 feature maps of size 8×8, followed by maxpooling which further reduces the size to 4×4. These feature maps are flattened into a single vector of size 1×800 and fed into the fully connected layer, which yields the final feature vector of size 1×500. This feature vector is used as input for a softmax classfier that outputs the probabilities for each of the 10 classes. The final prediction is simply the class that has the highest probability for the given input image.

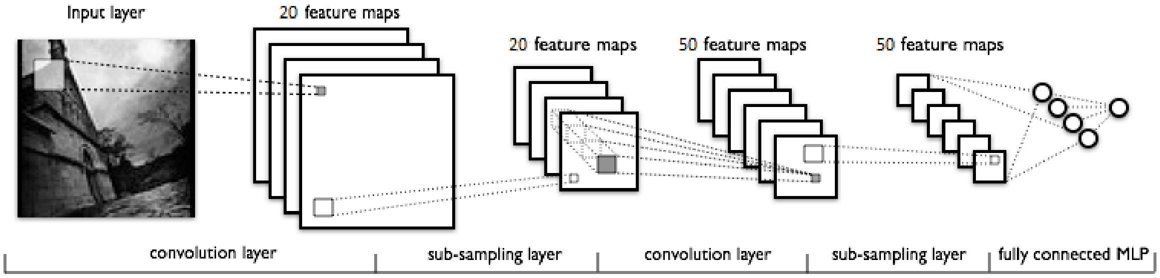

Figure 6.2: LeNet-5 architecture[2]

6.3 CS-MNIST Recognition

Here, we demonstrate that CS measurements can be used directly for classification, without reconstruction, using the CNN described above. The entire database was compressed with a random Gaussian short, fat matrix, Φ. That is, for each vectorized image $\mathbf{x}$, compressed sensing was simulated using $\mathbf{y} = \Phi\mathbf{x}$. Each $\mathbf{y}$ was then projected back

[2] http://deeplearning.net/tutorial/lenet.html#lenet

into the pixel space using $\hat{\mathbf{x}} = \Phi^T \mathbf{y}$. This new database was used to train the CNN using mini-batch gradient descent. The validation set was used for *early stopping* [48], which is a method used to prevent overfitting. The testing set was used to measure its performance. The results obtained at different compression ratios (CR) are shown in Table 6.1. Figure 6.3 shows how the validation error varies as training progresses.

Compression Ratio	No. of Measurements	Test Error
1 (Oracle)	784	0.93%
5	156	1.86%
10	78	3.02%
20	39	6.40%
100	8	39.13%

Table 6.1: MNIST recognition results obtained on compressed measurements using LeNet-5 CNN trained separately at each CR.

It can be seen from Table 6.1 that CNNs provide a possible way of extracting non-linear features from compressive measurements. For example, using just 78 of the 784 measurements yields an impressive test-error rate of 3.02%. From Figure 6.3, we observe that the validation error decays quickly and approaches the final value within a small number of epochs at all compression ratios. The results are encouraging on the MNIST database and it remains to be seen whether similar trends will be observed for natural images.

In the experiment described above, a different CNN is trained for each compression ratio. Next, we investigate the possibility of training a single CNN using the original images from the MNIST database(oracle) and testing the network using compressed images at different CRs. The results obtained are shown in Table 6.2.

Clearly, the accuracy for a particular CR in Table 6.2 is considerably lower than the corresponding accuracy in Table 6.1. This is expected since in the first case the network

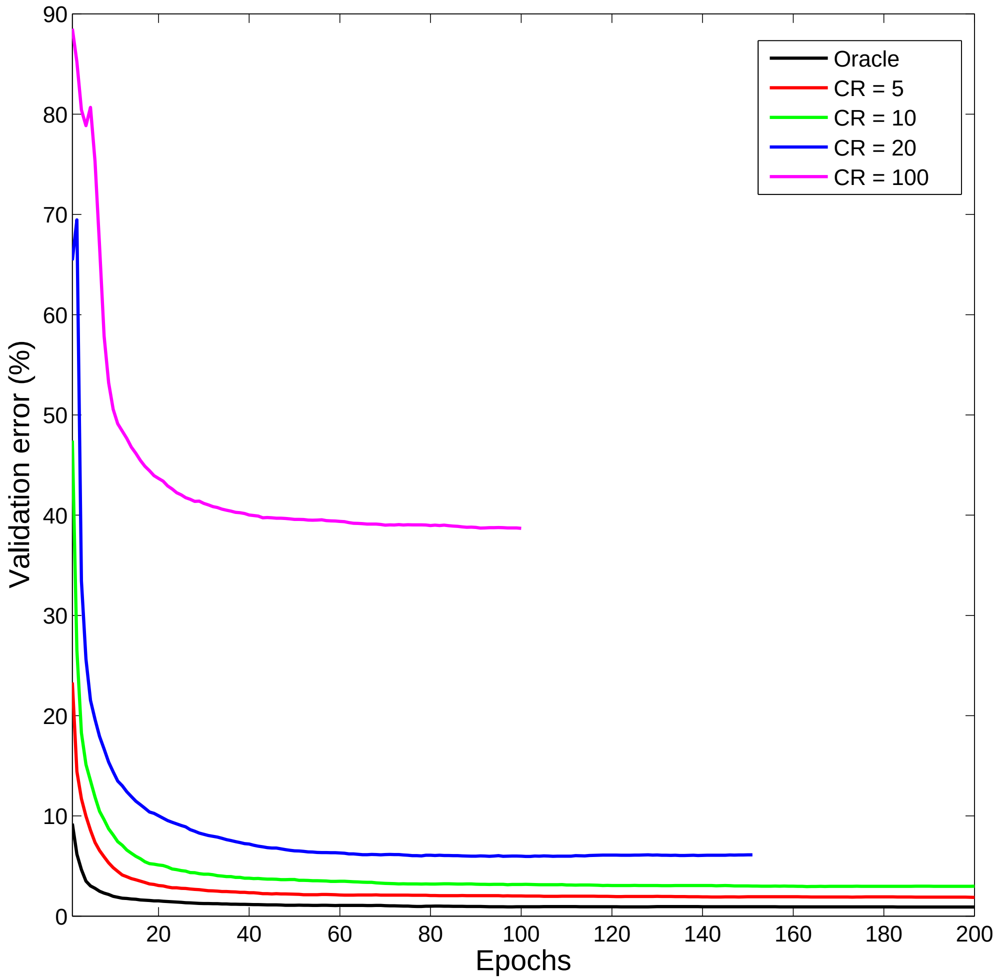

Figure 6.3: Results from LeNet-5 CNN on CS-MNIST. The plot shows how the validation error converges at various compression ratios (CR). Maximum number of epochs was set to 200.

trained on the compressed measurements and thus, the weights learned are optimized for the compressed measurements for a particular CR. In the second case, the network learned on the original MNIST database is used for testing on compressed measurements. Clearly, this does not perform as well as the first case since the inputs at high CRs are too different from the full-blown images.

Compression Ratio	No. of Measurements	Test Error
1 (Oracle)	784	0.92%
5	156	31.15%
10	78	65.58%
20	39	84.1%
100	8	88.65%

Table 6.2: MNIST recognition results obtained on compressed measurements using LeNet-5 CNN trained once on original images.

Chapter 7

CONCLUSIONS AND FUTURE WORK

Compressive sensing is a revolutionary method in signal acquisition and processing that is well suited for many resource constrained environments. It combines sampling and compression into one step, while allowing near-perfect reconstruction at sub-Nyquist sampling rates. However, current reconstruction algorithms are suffer from some drawbacks. In this thesis, we have presented a framework to address the problem of high level inference from compressive measurements without reconstruction. To this end, we have constructed smashed correlation filters and have shown that it is indeed possible to do so. We have demonstrated that, as a consequence of the JL lemma, correlations are preserved in the compressed domain and the correlational features thus extracted contain discriminative information for robust classification, even at high compression ratios. As a specific example, we have shown how this framework can be readily applied to the problem of face recognition, with very good results. This is especially important in the infrared domain, where compressive sensing architecture – e.g., the single pixel camera – provides a cost effective solution. There is a lot of scope for future research in this emerging field of compressive inference. Developing methods to extract non-linear features tailored for computer vision from compressive measurements is of prime concern. We have discussed one possible way of extracting non-linear features using convolutional neural networks. Considering the astronomical amounts of data being generated around the world, storage and more importantly smarter processing are issues that deserve attention. Compressive sensing provides a smart technique of sensing, storing and communicating data. Designing better algorithms that can exploit this directly for other tasks such as inference, in a computationally efficient fashion, is the challenge that lies ahead and this thesis is an important step in this direction.

REFERENCES

[1] T. Ahonen, A. Hadid, and M. Pietikäinen. Face recognition with local binary patterns. In *ECCV 2004*, pages 469–481. Springer, 2004.

[2] R. Baraniuk, M. Davenport, R. Devore, and M. Wakin. A simple proof of the restricted isometry property for random matrices. *Constr. Approx*, 2008, 2007.

[3] R. G. Baraniuk and M. B. Wakin. Random projections of smooth manifolds. *Foundations of computational mathematics*, 9(1):51–77, 2009.

[4] C. M. Bishop et al. Neural networks for pattern recognition.

[5] V. N. Boddeti. *Advances in Correlation Filters: Vector Features, Structured Prediction and Shape Alignment*. PhD thesis, Carnegie Mellon University, 2012.

[6] D. S. Bolme, B. A. Draper, and J. R. Beveridge. Average of synthetic exact filters. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2105–2112. IEEE, 2009.

[7] R. Calderbank, S. Jafarpour, and R. Schapire. Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. *Preprint 2009*.

[8] E. J. Candès and M. B. Wakin. An introduction to compressive sampling. *Signal Processing Magazine, IEEE*, 25(2):21–30, 2008.

[9] S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM journal on scientific computing*, 20(1):33–61, 1998.

[10] M. A. Davenport, M. F. Duarte, M. B. Wakin, J. N. Laska, D. Takhar, K. F. Kelly, and R. G. Baraniuk. The smashed filter for compressive classification and target recognition. In *Electronic Imaging*, pages 64980H–64980H. International Society for Optics and Photonics, 2007.

[11] M. A. Davenport and M. B. Wakin. Analysis of orthogonal matching pursuit using the restricted isometry property. *Information Theory, IEEE Transactions on*, 56(9):4395–4401, 2010.

[12] O. Déniz, G. Bueno, J. Salido, and F. De la Torre. Face recognition using histograms of oriented gradients. *Pattern Recognition Letters*, 32(12):1598–1603, 2011.

[13] D. L. Donoho. For most large underdetermined systems of linear equations the minimal 1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

[14] D. L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *Information Theory, IEEE Transactions on*, 47(7):2845–2862, 2001.

[15] D. L. Donoho and P. B. Stark. Uncertainty principles and signal recovery. *SIAM Journal on Applied Mathematics*, 49(3):906–931, 1989.

[16] K. Etemad and R. Chellappa. Discriminant analysis for recognition of human face images. *JOSA A*, 14(8):1724–1733, 1997.

[17] Y. Freund, S. Dasgupta, M. Kabra, and N. Verma. Learning the structure of manifolds using random projections. In *Advances in Neural Information Processing Systems*, pages 473–480, 2007.

[18] R. Girshick, J. Donahue, T. Darrell, and J. Malik. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 580–587. IEEE, 2014.

[19] J. Haupt, R. Castro, R. Nowak, G. Fudge, and A. Yeh. Compressive sampling for signal classification. In *Signals, Systems and Computers, 2006. ACSSC'06. Fortieth Asilomar Conference on*, pages 1430–1434. IEEE, 2006.

[20] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(3):328–340, 2005.

[21] C. Hegde, M. Wakin, and R. Baraniuk. Random projections for manifold learning. In *Advances in Neural Information Processing Systems 20*, pages 641–648. Curran Associates, Inc., 2008.

[22] R. Juday. Application of correlation filters. In *Correlation Pattern Recognition*, pages 357–382. Cambridge University Press, 2005. Cambridge Books Online.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[24] B. V. Kumar, A. Mahalanobis, and D. W. Carlson. Optimal trade-off synthetic discriminant function filters for arbitrary devices. *Optics Letters*, 19(19):1556–1558, 1994.

[25] B. V. Kumar, A. Mahalanobis, and R. D. Juday. *Correlation pattern recognition*, volume 27. Cambridge University Press Cambridge, 2005.

[26] Y. LeCun, B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[27] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[28] S. Z. Li, R. Chu, S. Liao, and L. Zhang. Illumination invariant face recognition using near-infrared images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):627–639, 2007.

[29] Y. Liang, W. Gong, Y. Pan, W. Li, and Z. Hu. Gabor features-based classification using svm for face recognition. In *Advances in Neural Networks–ISNN 2005*, pages 118–123. Springer, 2005.

[30] A. Mahalanobis, B. Kumar, and D. Casasent. Minimum average correlation energy filters. *Applied Optics*, 26(17):3633–3640, 1987.

[31] A. Mahalanobis and R. Muise. Object specific image reconstruction using a compressive sensing architecture for application in surveillance systems. *Aerospace and Electronic Systems, IEEE Transactions on*, 45(3):1167–1180, 2009.

[32] A. Mahalanobis, B. Vijaya Kumar, S. Song, S. Sims, and J. Epperson. Unconstrained correlation filters. *Applied Optics*, 33(17):3751–3759, 1994.

[33] P. Nagesh and B. Li. A compressive sensing approach for expression-invariant face recognition. In *IEEE CVPR*, pages 1518–1525. IEEE, 2009.

[34] D. Needell and J. A. Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Commun. ACM*, 53(12):93–100, Dec. 2010.

[35] M. A. Neifeld and P. Shankar. Feature-specific imaging. *Applied optics*, 42(17):3379–3389, 2003.

[36] Y. C. Pati, R. Rezaiifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Asilomar Conf. Signals Sys. Comp.*, Nov. 1993.

[37] A. Rodriguez, V. N. Boddeti, B. V. Kumar, and A. Mahalanobis. Maximum margin correlation filter: A new approach for localization and classification. *IEEE Transactions on Image Processing*, 22(2):631–643, 2013.

[38] S. T. Roweis, L. K. Saul, and G. E. Hinton. Global coordination of local linear models. In *Advances in Neural Information Processing Systems 14*, pages 889–896. MIT Press, 2002.

[39] A. C. Sankaranarayanan, P. K. Turaga, R. G. Baraniuk, and R. Chellappa. Compressive acquisition of dynamic scenes. In *ECCV 2010*, pages 129–142. Springer, 2010.

[40] M. Savvides, B. V. Kumar, and P. Khosla. ation using correlation filters. *3rd IEEE Automatic Identification Advanced Technologies*, pages 56–61, 2002.

[41] S. R. F. Sims and A. Mahalanobis. Performance evaluation of quadratic correlation filters for target detection and discrimination in infrared imagery. *Optical Engineering*, 43(8):1705–1711, 2004.

[42] L. Sirovich and M. Kirby. Low-dimensional procedure for the characterization of human faces. *JOSA A*, 4(3):519–524, 1987.

[43] T. Strohmer. Measure what should be measured: progress and challenges in compressive sensing. *Signal Processing Letters, IEEE*, 19(12):887–893, 2012.

[44] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *IEEE CVPR*, pages 1701–1708. IEEE, 2014.

[45] J. A. Tropp and A. C. Gilbert. Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Transactions on Information Theory*, 53(12):4655–4666, 2007.

[46] M. B. Wakin, J. N. Laska, M. F. Duarte, D. Baron, S. Sarvotham, D. Takhar, K. F. Kelly, and R. G. Baraniuk. An architecture for compressive imaging. In *IEEE International Conference on Image Processing*, pages 1273–1276. IEEE, 2006.

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

[48] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.