Feature Extraction From Compressive Cameras

With Application to Activity Recognition

by

Kuldeep Sharad Kulkarni

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

Approved June 2012 by the
Graduate Supervisory Committee:

Pavan Turaga, Chair
Andreas Spanias
David Frakes

ARIZONA STATE UNIVERSITY

August 2012

ABSTRACT

Recent advances in camera architectures and associated mathematical representations now enable compressive acquisition of images and videos at low data-rates. While most computer vision applications of today are composed of conventional cameras, which collect a large amount redundant data and power hungry embedded systems, which compress the collected data for further processing, compressive cameras offer the advantage of direct acquisition of data in compressed domain and hence readily promise to find applicability in computer vision, particularly in environments hampered by limited communication bandwidths. However, despite the significant progress in theory and methods of compressive sensing, little headway has been made in developing systems for such applications by exploiting the merits of compressive sensing. In such a setting, we consider the problem of activity recognition, which is an important inference problem in many security and surveillance applications. Since all successful activity recognition systems involve detection of human, followed by recognition, a potential fully functioning system motivated by compressive camera would involve the tracking of human, which requires the reconstruction of atleast the initial few frames to detect the human. Once the human is tracked, the recognition part of the system requires only the features to be extracted from the tracked sequences, which can be the reconstructed images or the compressed measurements of such sequences. However, it is desirable in resource constrained environments that these features be extracted from the compressive measurements without reconstruction. Motivated by this, in this thesis, we propose a framework for understanding activities as a non-linear dynamical system, and propose a robust, generalizable feature that can be extracted directly from the compressed measurements without reconstructing

the original video frames. The proposed feature is termed *recurrence texture* and is motivated from recurrence analysis of non-linear dynamical systems. We show that it is possible to obtain discriminative features directly from the compressed stream and show its utility in recognition of activities at very low data rates.

DEDICATION

To my grandfather,

ACKNOWLEDGEMENTS

I am indebted to several people who have contributed directly or indirectly to this thesis.

Words cannot express my gratitude to my thesis advisor, Dr. Pavan Turaga, for his constant guidance, his inspiring thoughts and moral support during my MS study. His enthusiastic attitude towards research and his dedication to work have inspired me. His approach to problem formulation and presentation of the same is a quality, which I try to emulate and has totally transformed my attitude towards research. My interactions with me have positively benefited me to mature as a researcher and will hold me in good stead as I look forward to my Phd under his guidance.

I would like to thank my committee members, Professor Andreas Spanias and Dr. David Frakes for serving as members of my thesis committee and examining my thesis report.

I am grateful to Darleen Mandt for helping me out whenever I was lost in administrative maze.

This thesis would not have been possible without the support of my entire family. I wish to thank my grandparents, my parents and my brother for their endless love and unconditional support. No words are enough to express my gratitude for all that they had to sacrifice and endure, so that I may have this opportunity. This thesis owes its genesis to my grandfather, to whom I dedicate this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

INTRODUCTION

1.1 MOTIVATION

Recent years has seen the generation of huge volume of visual information due to advances in camera capabilities. This has led to a growing need for information processing systems to efficiently compress, analyze and store highly redundant information data captured by imaging devices. Most visual data is stored in some compressed form or the other. Therefore, it is desirable that low-level features are directly extracted in compressed domain. Low-level features are compact, mathematical representations of the physical properties of the image data. Although the compressed-domain approach imposes many difficulties, it opens up an opportunity to reduce the computational complexity because it greatly reduces the amount of data to be analyzed for indexing, high-level understanding and classification. Feature extraction in compressed domain is not a new concept and significant research has been done over the last 3 decades. Earliest work in this area can be referred to [20] who used Mandala transform as a way of automatic target recognition in compressed images. A simple model of feature extraction in compressed domain is shown in figure 1.1.

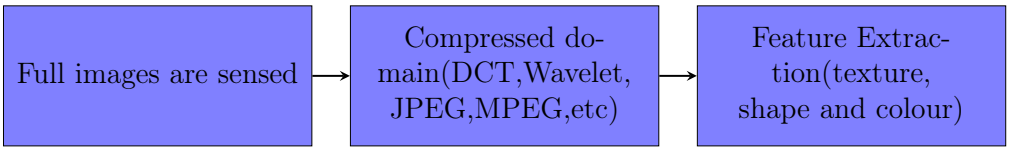


Figure 1.1: A simple existing model of feature extraction in compressed domain is shown. First full-blown images are acquired, then data is compressed by exploiting its structure and then features are extracted from the compressed data.

However, all these methods employ the 'worst case approach' of 'sam-

ple first, ask questions later', where large amounts of data are sampled first and the structure of the data is exploited to compress the data for storage and transmission. Breakthrough research in recent years has seen the emergence of a new sensing method called 'Compressive Sensing or Sampling', which allows us to integrate the process of sensing and compression of data. According to this theory, instead of using a enormous number of sensors to acquire whole data, we can sense a compressed version directly in the form of significantly less number of measurements using very few sensors. A significant progress in the field of compressive sensing allows signal reconstruction at sub-Nyquist sampling rates by exploiting the additional structure on the signal being sensed. This is most often in the form of sparsity in an appropriately chosen basis [5]. A large body of work now exists that deals with algorithms for recovery of the original signal from such compressed measurements. There is a tremendous breadth of such techniques, and the readers are referred to recent compilations for a comprehensive survey [16, 18]. However, much less attention has been devoted to the question of whether higher-level inference tasks such as detection and recognition can be performed without reconstructing the original signal/images. Recent work shows that simpler tasks like background subtraction [8] and optical flow estimation [34] are possible using compressive sensing without reconstruction.

Before we move onto discuss the problem of higher-level inference tasks, we wish to drive home the motivation behind the usage of compressive cameras for such tasks. To this end, lets consider the application of unmanned aerial vehicles(UAVs) which provide realtime video and high resolution aerial images on demand. These UAVs encounter very high video handling requirements such as collection of data, followed by transmission of the same to a ground

station using a low-bandwidth communication link. This results in expensive methods being employed for video capture, compression, and transmission implemented on the aircraft. Other similar applications like real-time monitoring of patients, children and elderly persons, sports play analysis are also resource constrained and have to be achieved in real-time, thus demanding low communication overheads. These applications have a activity recognition system, as their primary component. Activity recognition systems of today involve acquisition of the enormous amount of redundant video data, followed by extraction of rich set of features, which involves expensive computations, thus rendering real-time applications impossible. It is important that we exploit the inherent structure in the acquired imagery to transmit the small number of measurements in order to address communication requirements.

The general problem of activity recognition is difficult to address, since many features that are useful for object and activity recognition tasks require non-linear feature extraction techniques. Typical features useful for activity analysis include histogram of gradients (HOG) [15], optical flow [10], 3D SIFT [22], contours [36] etc. Activity recognition has a rich and long history in computer vision, and the readers are referred to recent surveys on this topic [1]. [12] explored the utility of CS as a compression tool for features that have already been extracted from the original video, but did not address direct feature extraction from CS measurements of images. It is quite difficult to obtain such complex features directly from the compressive measurements without an intermediate step of signal reconstruction. Thus, there is a growing need to explore novel features that retain robustness and accuracy, yet are amenable to extraction directly from compressed measurements. Recently a linear dynamical system (LDS) was used to recover videos from CS cameras in [32].

LDS models are useful for video reconstruction, but being generative models they are sensitive to spatial/view transforms, thus require further processing to obtain robust recognition performance. In this thesis, we propose a framework to the understanding of activitiy recognition as a non-linear dynamical system which involves feature extraction without the reconstruction of original data.

1.2 CONTRIBUTIONS AND ORGANIZATION:

The main contributions of the thesis are the following:

1. We study the problem of activity recognition from compressive cameras using the geometric properties of high-dimensional video data,
2. We present a conceptually simple yet robust method for quantifying this geometric information in terms of recurrence textures,
3. We show the utility of this method for performing robust activity recognition at very low data rates.

This thesis is organized into 5 chapters. Chapter 2 presents the basics of compressive sensing. In Chapter 3, a theoretical framework to consider the problem of human activity analysis in compressive cameras is described and the proposed geometric analysis of video via recurrence analysis, and associated feature extraction are discussed. Chapter 4 provides a discussion of results with the proposed method. The last chapter concludes the work and explores the scope for future research in this direction.

Chapter 2

COMPRESSIVE SENSING

2.1 COMPRESSIVE SENSING: A BACKGROUND

The Shannon-Nyquist sampling theorem states that the sampling frequency of a signal should be at least twice the highest frequency contained in the signal, in order to avoid loss of information. In applications like digital image and video cameras, the rate specified by this theorem is so high that it makes compression essential before transmission. In imaging systems and high speed analog to digital converters, high sampling rate is very expensive. This section provides the necessary background about the theory of compressive sensing, a new paradigm in signal aquisition of compressible signals. Simply put, CS theory asserts that it is possible to recover signals fully from fewer number of measurements than what is required by Nyquist rate, provided certain conditions are met. CS is based on two principles: sparsity of the signals and incoherence, which deals with the manner in which the signal is sensed. Sparsity communicates the idea that the 'information rate' of a continuous time signal may be much smaller than suggested by its bandwidth, or that a discrete-time signal depends on a number of degrees of freedom which is very small when compared to the length of the signal itself. To be more specific, CS exploits the fact that many natural signals are sparse or compressible in the sense that they have compact representations when transformed to appropriate basis ψ. Incoherence extends the notion of the classical uncertainty principle, 'A time-limited signal cannot be band-limited signal' and conveys the idea that signals having a sparse representation in ψ must be spread out in the domain in which they are acquired, just as a Dirac or a spike in the time domain is spread out in the frequency domain. In other words, incoherence

says that unlike the signal of interest, the sampling/sensing waveforms have an extremely dense representation in ψ. The crucial observation is that it is possible to design efficient sensing or sampling systems that capture the useful information content in a sparse signal and condense it into a small amount of data. These systems implement correlations of the signal with a small number of fixed waveforms that are incoherent with the sparsifying basis. Thus the systems will have sensors to very efficiently capture the information in a sparse signal without trying to understand that signal. Finally, there are numerical optimization methods to reconstruct the whole signal from a small number of measurements. Thus CS is a very efficient protocol by which data can be sensed at very low rates in the form of incomplete set of measurements and later uses the computational power to reconstruct the original signal from such acquired data.

2.2 MEASUREMENT PRINCIPLE

Unlike in the Shannon/Nyquist sampling, we do not measure the point samples for representing a signal. However, we obtain linear measurements of the signal which are projections of the signals onto a measurement space. Thus for a image, information is not in terms of actual pixels but in terms of a set of linear measurements. Let $g(t)$ be a signal obtained by projections

$$y_k =< g, \phi_k > \quad k = 1, ..., M \tag{2.1}$$

We obtain M correlations of the signal with M different sensing waveforms, ϕ_k, which can be Dirac delta functions (spikes) or sinusoids as shown in Figure 2.1. Here we restrict our attention to discrete signals $g \in \mathbb{R}^N$. This results in a undersampled situation in which, the number M of available measurements is much smaller than the dimension N of the signal g. Now we are confronted

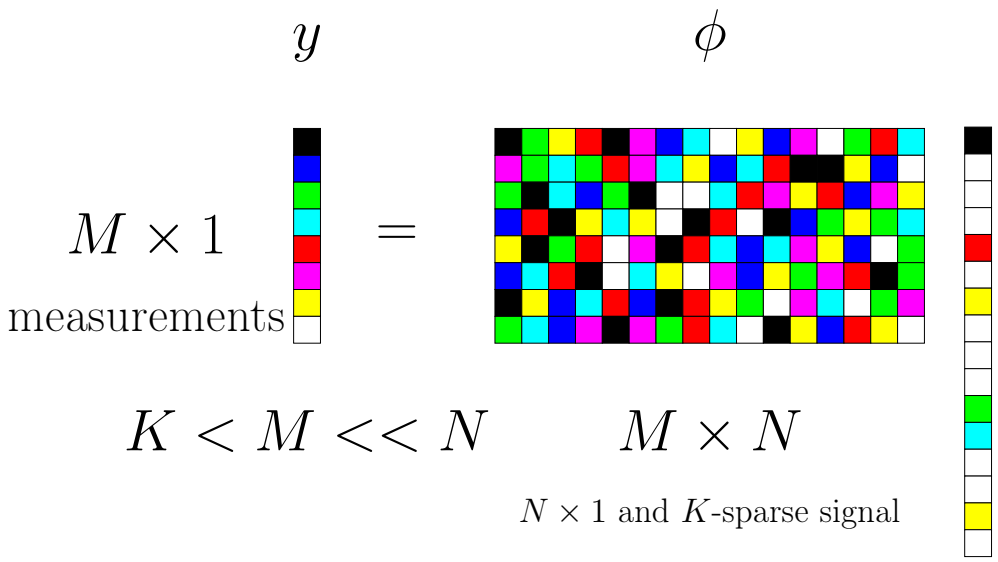


Figure 2.1: Illustration of coded acquistion by compressive sensing. The signal to be sensed, x is correlated with M sensing waveforms, $\phi_1^*, ..., \phi_M^*$ which form the rows of the sensing matrix ϕ, yielding M linear measurements.

with an important question about accurate reconstruction from $M << N$ measurements only. Letting ϕ denote the $M \times N$ sensing matrix with the vectors $\phi_1^*, ..., \phi_M^*$ as rows (a^* is the complex transpose of a), the process of recovering $g \in \mathbb{R}^N$ from $y = \phi g \in \mathbb{R}^N$ is ill-posed in general when $M < N$: there are infinitely number of signals, $\hat{g}$ for which $\phi\hat{g} = y$.

2.3 INCOHERENCE AND SPARSITY OF SIGNALS

This section presents the two fundamental principles underlying CS: sparsity and incoherence.

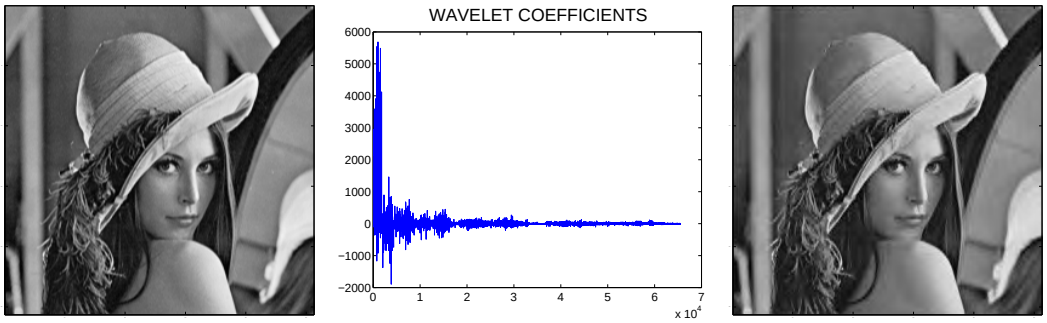
2.3.1 SPARSITY

Most natural signals have compact representations when transfomed to a appropriate basis. For the image in Figure 2.2(a), wavelet coefficients provide a very compact representation. In mathematical terms, we wish to express a

vector $g \in \mathbb{R}^N$ (such as the N-pixel image in Figure 2.2) in an orthonormal basis (such as a wavelet basis) $\psi = [\psi_1, \psi_2, ...\psi_N]$ as follows:

$$g(t) = \sum_{i=1}^{N} x_i \psi_i(t) \tag{2.2}$$

where x is the coefficient sequence of g, $x_i =< g, \psi_i >$. Compactly we can write g as ψx where (ψ is $N \times N$ matrix with $\psi_1, \psi_2, ..., \psi_N$ as columns). Thus when a signal has a sparse expansion, we can reject the small coefficients without any significant loss. Now, consider $g_K(t)$ obtained by keeping only the terms corresponding to the K largest values of (x_i) in the expansion (2). By definition, $g_K := \psi x_K$, where x_K is the vector of coefficients x_i with all but the largest K set to zero. This vector is sparse, since barring few, all of its entries are zero. Such objects with at most K nonzero entries are called K-sparse. Since ψ is an orthonormal basis (or orthobasis), we have $|| \ g - g_K \ ||_2 = || \ x - x_K \ ||_2$, and if x is sparse or compressible in the sense that the sorted magnitudes of the x_i decay quickly, then x is well approximated by x_K and therefore, the error $|| \ g - g_K \ ||_2$ is small. Thus one can throw away a significant portion of the coeffiecients without much loss. In figure, 2.2(c) we show an example where the perceptual loss is barely observable from the image of 65,536 pixels to its approximation obtained by throwing away 90% of the coefficients. Sparsity is a fundamental modeling tool which allows efficient fundamental signal processing; e.g., accurate statistical estimation and classification, efficient data compression, and so on. By determining how efficiently signals can be acquired nonadaptively, sparsity has a significant impact on the process of aquisiton process itself.



(a) (b) (c)

Figure 2.2: a) Original image with 65,536 pixels with pixel values in the range [0,255], b) A very large number of wavelet coefficients are nearly zero, indicating the image's compressible nature which is true for most natural images, c) No perceptual loss in reconstructed image after rejecting 90% of the wavelet coefficients

2.3.2 INCOHERENT SAMPLING

Suppose we have a pair (ϕ, ψ) of orthobases of $\mathbb{R}^N$. The first basis ϕ is used for sensing the signal g and the second is used to represent g. The coherence between the sensing basis ϕ and the representation basis ψ is

$$\mu(\phi, \psi) = \sqrt{n}.max \mid< \phi_k, \psi_j >\mid, \quad \forall 1 \leq k, j \leq n \tag{2.3}$$

Thus correlation measures the largest correlation between any two elements of ϕ and ψ. If ϕ and ψ contain highly correlated elements, the coherence is large. Otherwise, it is small. From basic linear algebra, the coherence, $\mu(\phi, \psi) \in [1, \sqrt{n}]$. Since in compressive sensing, the pairs of bases of interest are required to have low coherence, we will now give examples of such cases. Firstly, ϕ is the canonical or spike basis $\phi_k(t) = \delta(t - k)$ and ψ is the Fourier basis, $\psi_j(t) = n^{-1/2} e^{i2\pi jt/n}$. Since ϕ is the sensing matrix, this corresponds to the traditional sampling scheme in time. The coherence of time-frequency pair follows the relation, $\mu(\phi, \psi) = 1$ and therefore, we have maximal incoherence. In the second example, we have wavelets bases for ψ and noiselets [11] for ϕ. The coherence between noiselets and Haar wavelets is $\sqrt{2}$ and that between

noiselets and Daubechies D4 and D8 wavelets is respectively, about 2.2 and 2.9. Noiselets are also maximally incoherent with spikes and incoherent with the Fourier basis. The noiselets are very important for efficient CS implementations since they are incoherent with bases providing sparse representations of image data [6]. As a third example, we have random matrices which are highly incoherent with any fixed basis ψ. We select an orthobasis ϕ uniformly at random, which can be done by orthonormalizing N vectors sampled independently and uniformly on the unit sphere. Then with high probability, the coherence between ϕ and ψ is about $\sqrt{2log(N)}$.

2.4 MEASUREMENT SYSTEMS AND SPARSE SIGNAL RECOVERY CONDITIONS

We wish to recover all the N coefficients of g, but we get to observe only a subset of the samples $M \subset 1, 2, ..N$. These samples are encoded in the following manner.

$$y_k =< g, \phi_k > \quad k = 1, ..., M \tag{2.4}$$

The reconstruction equation $\hat{g} = \psi\hat{x}$, where $\hat{x}$ is the solution obtained through l_1-norm minimization through the convex optimization program given by

$$\max_{\hat{x}\in\mathbb{R}^N} || \hat{x} ||_1 \; s.t \quad y_k =< g, \phi_k > \quad k = 1, ..., M \tag{2.5}$$

Suppose the signal $g \in \mathbb{R}^N$ in terms of the coefficient x is K-sparse, then selecting M measurements in the ϕ domain uniformly at random gives the following.

If $M \geq C.\mu^2(\phi, \psi).K.log(N)$, for some positive constant C, the solution to equation 2.5 is exact with overwhelming probability. It follows that the role of coherence is very simple; the smaller the coherence, the fewer samples are needed, and hence we look for systems with low coherence. Also the signal g can be exactly recovered from smaller data set through minimizing a

convex function which need not have any knowledge about number of nonzero coefficients and their locations or values.

2.5 RECONSTRUCTION ALGORITHMS

The objective of the CS decoder is to reconstruct the K-sparse signal $g \in \mathbb{R}^N$ from its compressive measurements $y \in \mathbb{R}^M$. One method of solving this l_1 optimization problem is through Basis Pursuit (BP). Yet another method of reconstruction through Basis Pursuit Denosing (BPDN) is well suited in cases where measurements are noisy. The measuring process with noise can be given by

$$y = \phi x + z, \quad y \in \mathbb{R}^M, x \in \mathbb{R}^N, z \in \mathbb{R}^M \tag{2.6}$$

where z is a stochastic noise or a deterministic unknown error term. The solution to the BPDN optimization problem is given by

$$\hat{x} = argmin \mid\mid \hat{x} \mid\mid_1 s.t. \mid\mid y - \phi\psi x \mid\mid < \epsilon \tag{2.7}$$

where ϵ is a constant which takes into account the variance of the noise z.

2.6 COMPRESSIVE IMAGING

In this section we describe the application of compressive sensing to a imaging device, developed in Rice University [37]. The compressive imaging system developed by them embodies a microcontrolled mirror array propelled by pseudorandom and other measurement bases and a single or multiple photodiode optical sensor. CS camera by employing a single photon detector, provides a significant advantage over conventional cameras in that it can be adapted to image at wavelengths which are not possible with the latter. The compressive measurements of the image are computed optically, as per the CS theory. And finally, CS reconstruction algorithms are employed to recover the actual im-

ages. In addition to this, it provides the provison of acquiring measurements of a video signal which can be reconstructed by either 2-D reconstruction of one frame at a time or joint 3-D reconstruction. The measurement bases used in the camera are incoherent with any sparse bases, and hence the camera can be used to capture all kinds of images. The compressive imaging block diagram from [37] is shown in the Figure 2.3. The hardware implementation

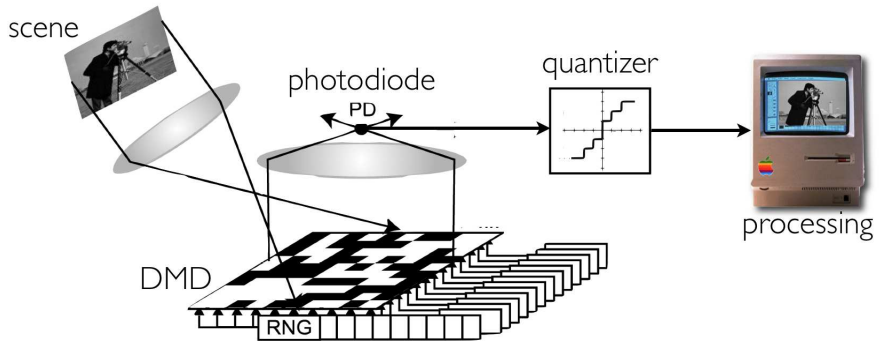


Figure 2.3: Figure from [37]: Compressive Imaging (CI) camera block diagram. Incident lightfield (corresponding to the desired image x) is reflected off a digital micromirror device (DMD) array whose mirror orientations are modulated in the pseudorandom pattern m supplied by the random number generators (RNG). Each different mirror pattern produces a voltage at the single photodiode that corresponds to one measurement y_m.

of the above imaging system is a single-pixel camera. It incorporates a micro-controlled mirror array displaying a time sequence of M pseudorandom basis images, ϕ_m which is combined with a single optical sensor to compute incoherent image measurements y. The camera provides the luxury of adaptively selecting the number of measurements to be computed, by trading off the extent of compression versus acquisition time, while the conventional cameras trade off the resolution against the the number of pixel sensors.

Chapter 3

PROBLEM FORMULATION

3.1 DYNAMICAL MODELS OF ACTIVITY RECOGNITION

In this chapter we formalize the problem of activity recognition from fundamentals, develop the theory and cast it as a problem of texture recognition. The problem of action recognition can be studied at two levels of complexity, one being the simple movements performed by a single human, termed as 'actions' and other being 'activities', the coordinated combination of several simple movements performed by a small group of humans. Examples of actions can be running, swimming, walking, bending etc. and examples of activities are two persons shaking hands, a group of people dancing in a certain coordinated manner. Here, we give a overview of approaches of action recognition. Since our method (explained later in this chapter) relies on quantifing fine variations in non-linear dynamical system, we restrict to only dynamical models used for action recognition. [29] lists down 3 major dynamical models used for action recognition, namely Hidden Markov Models, Linear Dynamical Systems and Non-linear Dynamical System.

3.1.1 HIDDEN MARKOV MODELS

The Hidden Markov Model(HMM) is a statistical tool for modeling generative sequences governed by an underlying process which generates an observable sequence. The system being modeled is assumed to be a Markov process with unobserved (hidden) states. Each state has a probability distribution over the possible output tokens. Therefore the sequence of tokens generated by an HMM gives information about the sequence of states. In the case of action recognition, the temporal evolution of a activity is modeled by HMM. HMMs

first gained popularity in speech recognition [30]. [25, 2] applied HMMs to model the temporal evolution of human gait patterns for action recognition. Since these methods assume that the feature sequences on which HMM is enforced, are obtained from actions perfomed by a single person, they are not useful in modeling activities perfomed by more than one person. HMMs are limited in their applicability to stationary actions due to the assumption of Markovian dynamics and the time-invariant nature of the model.

3.1.2 LINEAR DYNAMICAL MODELS

Linear Dynamical models are an extension of HMMs to continuous space. Thus the state-space is allowed to assume values in $\mathbb{R}^k$ where k is the dimensionality of the state-space. A first-order time-invariant Gauss-Markov Processes as described in [29] is given by the following.

$$x(t) = Ax(t-1) + w(t), \quad w \sim N(0, Q) \tag{3.1}$$

$$y(t) = Cx(t) + v(t), \quad v \sim N(0, R) \tag{3.2}$$

where $x \in \mathbb{R}^d$ is the d-dimensional state vector and $y \in \mathbb{R}^n$ is a n-dimensional feature vector with $d << n$, A, the transition matrix and C, the measurement matrix. w and v are the process and observation noise respectively which are Gaussian distributed with zero-means and covariance matrices Q and R respectively. There is a rich literature to obtain the closed form solutions for learning the model parameter (A, C) from the feature sequence y. This model has been successfully used in applications of recognition of actions and actions based on gait, most notably in [27, 13]. Recently a linear dynamical system (LDS) [31] was used to recover videos from CS cameras and recognize actions. Here the compressive measurements, instead of feature vectors were used to form the temporal sequence y. However, as with HMMs, since LDSs

are developed using Markovian dynamics, the drawback of this model is its time-invariant nature which renders it useless for non-stationary actions.

3.1.3 NONLINEAR DYNAMICAL MODELS

A activity is composed of sequences of actions of short durations. Hence it is not possible to model the whole activity by a single LDS. In such a case, each action can be modeled by a different LDS. This gives rise to the notion of switching LDSs. Thus the model parameters (A, C) will now vary with time and are replaced by $(A(t), C(t))$ [29]. Approaches to model activities using LDSs are restricted to use time series data, which lies on Euclidean space. However most successful features used in computer vision are non-linear features like SIFT, HOG. [10] describes a activity recognition method in which the temporal evolution of histogram of oriented optical flow(HOOF) features is modeled using Nonlinear Dynamical Systems(NLDSs).

3.2 PROBLEM FORMULATION

Now, we formulate the problem of action recogntion from compressive measurements into one of identifying discriminative features in a non-linear dynamical system. To start with, when a sequence of images is acquired by a compressive camera, the measurements are generated by a sensing strategy which maps the image space $\mathcal{I} \in \mathbb{R}^N$ to an observation space $Z \in \mathbb{R}^M$. The overall mapping consists of a transformation F from the 3D scene-space $\mathcal{S}$ to image-space, with the addition of noise n in the sensor, followed by the measurement matrix ϕ, which gives measurements Z,

$$I(t) = F \circ S(t) + n(t) \tag{3.3}$$

$$Z(t) = \phi I(t) \tag{3.4}$$

Here $S(t)$ refers to a model of the scene (such as a CAD model) with a human performing an action. Compressive sensing represents a succession of data-reduction operations, going from the full-blown space of 3D scenes to image-space, and then to measurement-space. Assuming that the changes in the scene are due to a human performing some activity, we seek features that can be extracted directly from the sequence of measurements $\{Z(t)\}$. Since we do not intend to reconstruct the image-sequence, we are restricted in our ability to extract meaningful features. However, the JL-lemma suggests that the general geometric relations of a set of points in a high-dimensional space can be preserved by certain embeddings into a low-dimensional space. In the case of compressive sensing, this embedding is achieved by the random measurement matrix ϕ, in other words orthogonally projecting to $\mathbb{R}^M$. Formally stated, the Johnson-Lindenstrauss lemma is given as follows:

Given $0 < \epsilon < 1$, a set X of Q points in $\mathbb{R}^N$, and a number $M > N_0 = \mathcal{O}(\frac{log(Q)}{\epsilon^2})$, there is a Lipschitz function $f : \mathbb{R}^N \to \mathbb{R}^M$ such that

$$(1-\epsilon)\|v-u\|^2 \leq \|f(v)-f(u)\|^2 \leq (1+\epsilon)\|v-u\|^2 \tag{3.5}$$

At this point in time, we recall the fact that many human activities lie intrinsically on low-dimensional manifolds. For example, the deformations seen in the shape of a tracked human silhouette performing a activity like walking are governed by physical body constraints, more specifically the joint-angles in the body. Thus number of degrees of freedom required to determine a human activity is significantly small when compared to high-dimensional image data. It is shown in [4] that these images of tracked humans when considered as points in a high-dimensional visual input space, lie on a low dimensional manifold and a lot of work has been done to explicitly model the manifold and extract its structure from full-blown images for tracking and activity recogni-

tion [24, 23]. Using JL-lemma, it is shown in [3] that if sufficient number M of random projections of a manifold-modeled signal are taken, then with high probability, all pairwise Euclidean and geodesic distances between points on the manifold are well preserved. Thus manifold structure obtained from CS measurements will be about the same as that obtained from the images. Last few years has seen considerable amount of research to exploit the above notion to reveal the structure of the underlying manifold from CS measurements, most notably in [19], where a greedy algorithm is developed to estimate the smallest dimension to which the high-dimensional data can be projected and perform manifold learning. While we do not attempt to explicitly determine manifold structure, our work is related to the above mentioned works in that we wish to utilize the inherent geometric structure for activity recognition which is what activity recognition methods using manifolds are based on and that we wish to do it from CS measurements directly. Motivated by this, we wish to explore the notion of recurrence plots which encode the geometric structures of the data [17]. Further, considering that the system defined in equation (3.4) is a non-linear dynamical system, we attempt to understand the system properties via its recurrence properties [26, 38].

3.3 RECURRENCE TEXTURES AND CLASSIFICATION OF ACTIVITIES

Recurrence plots (RPs) are a visualization tool for dynamical systems. These plots often reveal correlations in the data that are not easily detected in the original time series. A recurrence matrix defined as

$$R(i,j) = \theta(\epsilon - \|x_i - x_j\|_2) \tag{3.6}$$

where x_t is the observed time series and $\theta(.)$ is the Heaviside step function. RPs which are thus binary images displaying black dots where the values

are within the threshold ϵ, are shown to capture the system's behavior and be distinctive for different dynamical systems. Recurrence plots are intricate and visually appealing. They are also useful for finding hidden correlations in highly complicated data. Moreover, because they make no demands on the stationarity of a data set, RPs are particularly useful in the analysis of systems whose dynamics may be changing. For example, Casdagli[7] used RPs to characterize time series generated by dynamical systems driven by slowly varying external forces.

At the time instant t, the compressive measurement of the image observation (the t^{th} frame of the video sequence) is $Z(t) \in \mathbb{R}^M$. Thus, if a sufficient number of measurements are taken, then with high probability the RPs for the compressed $\{Z(t)\}$ and uncompressed signals $\{I(t)\}$ will be the same. Though these seems like a straightforward consequence of the JL-lemma, we formally quantify the exact error between these RPs in section 3.4. Thus, we propose to use the recurrence relations of $\{Z(t)\}$ as a means to acquire discriminative features from activities. In order to quantify the structures in RPs, a set of measures known as Recurrence Quantitative Analysis have been proposed by [17, 26, 38]. Recurrence Quantification Analysis, is particularly useful in finding locations in the data where the underlying dynamics change. RQA is the best available approach to analyze the dynamics of a system from recurrence plots. In order to perform RQA on a data set, we first construct a RP, choosing a threshold and then use that RP to compute statistical values namely % recurrence(REC), % determinism (DET) and entropy. The first of these statistics, termed % recurrence(REC), is simply the percentage of points on the RP that are darkened. The second RQA statistic is called % determinism(DET); it measures the percentage of recurrent points in a RP that are contained in

lines parallel to the main diagonal. Diagonal lines are included in the analysis, if and only if they meet or exceed some prescribed minimum length threshold. Intuitively, DET measures how organized a RP is. The third RQA statistic called entropy, is closely related to % determinism. Entropy(ENT) is calculated by binning the diagonal lines according to their lengths and using the following formula:

$$ENT = -\sum_{k=1}^{N} P_k log(P_k) \tag{3.7}$$

where N is the number of bins and P_k is the percentage of all lines that fall into bin k. However, the lumped nature of RQA measures do not capture the dynamics of different system unambiguously, sometimes yielding similar RQA measures for structurally dissimilar RPs. For example in figure 3.1, it is shown

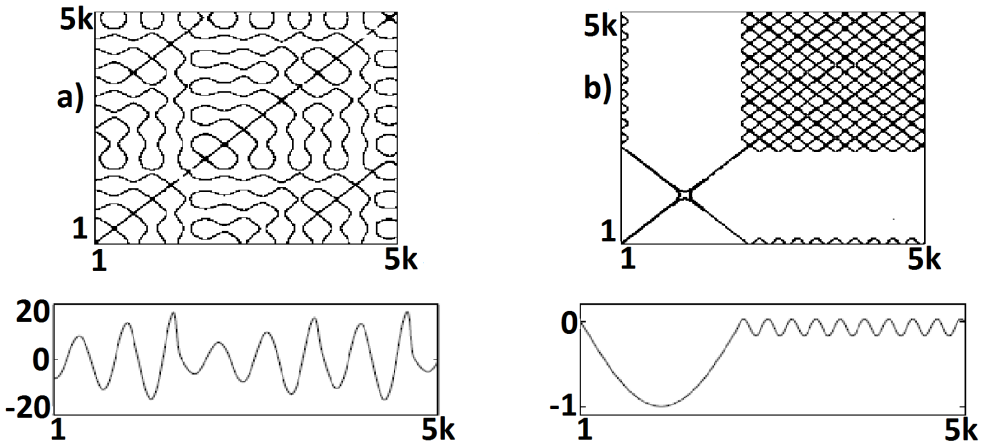


Figure 3.1: Figure from [21]: RQA results on structurally dissimilar RPs can be almost identical. These two very different RPs, one (left) from the Rossler system and other,(right) a sine-wave signal of varying period, have equal or near-equal values of REC (2.1%) and DET (42.9% for the Rossler data and 45.8% for the varying-period sine wave).

that two structurally different RPs that are almost identical from the standpoint of RQA. Moreover, the RPs themselves are very sensitive to the threshold, leading to different structures for different thresholds for the same system.

These limitations motivate us to make use of the full geometric information encoded in the non-thresholded recurrence matrices or the non-thresholded recurrence plots(NTRPs). We term the non-thresholded recurrence matrices simply as 'Distance' matrices. But instead of calculating the distance matrix for the time series obtained from the sequence of measurements, we calculate it for the time series obtained by taking the first derivative measurements (successive difference operation). Thus, for each sequence of compressive measurements $\{Z(t)\}$ the distance matrix is a square-symmetric matrix, D of size $(T-1) \times (T-1)$, given by

$$D(i,j) = \|\dot{Z}(i) - \dot{Z}(j)\|_2 \tag{3.8}$$

where $\dot{Z}(i) = Z(i+1) - Z(i)$. We perform this successive difference operation as a way to remove the effects of a static background, so that features are more sensitive to movement in the scene. On visualizing the distance matrices as intensity images as shown in figure 3.2, it is clear that different activities give rise to widely different *recurrence textures*. Motivated by this, we pose the problem of classification of the dynamical system as a texture recognition problem. To this end, we utilize a computationally simple yet powerful texture classification method based on local binary patterns (LBPs) [28]. Certain LBPs termed as 'uniform' are fundamental properties of image texture and and their occurrence histogram is proven to be a powerful texture feature.

3.4 QUANTIFICATION OF ERROR IN NTRPS

Before we move to explain how texture recognition is performed using LBPs, we wish to know exactly, by how much the non-thresholded recurrence plots(NTRPs) obtained from CS measurements differ from those obtained from original images. Formally, we will quantify the error between two 'Distance' matrices

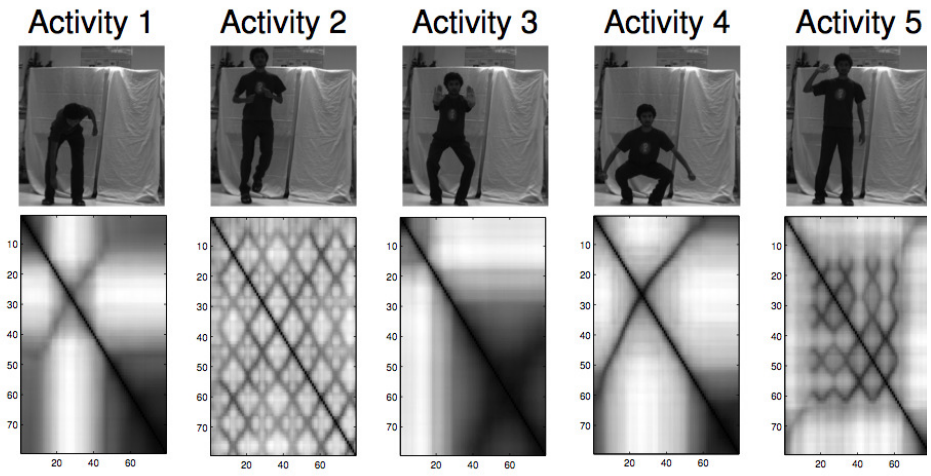


Figure 3.2: Row1: Examples of different activities from UMD dataset; Row2: Corresponding recurrence texture representations of the actions.

in terms of Q, the number of frames used to obtain those matrices, M, the number of CS measurements, N, the dimension of the original image. From [14], we gather that a linear mapping represented by a $M \times N$ matrix ϕ, whose entries are randomly drawn from certain probability distributions, can be con-

sidered as Lipschitz function f. Hence from JL-lemma it follows that for every two points x_i and x_j in image space $\mathcal{I} \in \mathbb{R}^N$ and $M > \mathcal{O}(\frac{log(Q)}{\epsilon^2})$, we have the relation in (3.9).

$$(1-\epsilon)\|x_i - x_j\|^2 \leq \|\phi x_i - \phi x_j\|^2 \leq (1+\epsilon)\|x_i - x_j\|^2 \tag{3.9}$$

By defining similar inequalities for every pair of points in a sequence of Q images and then adding them, we arrive at the relation in (3.10).

$$(1-\epsilon)\sum_{i=1}^{Q}\sum_{j=1}^{Q}\|x_i - x_j\|^2 \leq \sum_{i=1}^{Q}\sum_{j=1}^{Q}\|\phi x_i - \phi x_j\|^2 \leq (1+\epsilon)\sum_{i=1}^{Q}\sum_{j=1}^{Q}\|x_i - x_j\|^2 \tag{3.10}$$

The summation terms in this inequality are nothing but the squares of forbenius norms of the respective NTRPs, which in effect yields us the following inequality.

$$\sqrt{(1-\epsilon)}\|I\|_F \leq \|Z\|_F \leq \sqrt{(1+\epsilon)}\|I\|_F \tag{3.11}$$

where $\|I\|_F$ and $\|Z\|_F$ denote the forbenius norms of the 'Distance matrices' in image and compressed domain respectively. We denote the ratio of forbenius norm in the compressed domain to that in the image domain as R_F and it follows the bounds in equation (3.12).

$$\sqrt{(1-\epsilon)} \leq R_F \leq \sqrt{(1+\epsilon)} \tag{3.12}$$

This ratio R_F requires to be as close to unity as possible to ensure minimum deviation in compressed domain. We define this deviation from unity as E_F, and hence the bounds for it are given by equation (3.13)

$$\sqrt{(1-\epsilon)} - 1 \leq E_F \leq \sqrt{(1+\epsilon)} - 1 \tag{3.13}$$

where $E_F = R_F - 1$. Since ϵ is directly related to M and Q, we can now say that we have quantified the error between the NTRPs obtained from original

image domain and those from compressed domain in terms of the number of measurements and the number of points used to obtain those NTRPs. Since M is directly proportional to Q and inversely proportional to the square of ϵ, to be able to force E_F close to zero, by taking very few measurements, it is necessary that we construct the NTRPs from sequences of small number of images.

3.5 LOCAL BINARY PATTERNS

Local binary patterns (LBP) is a type of feature used for classification in computer vision. LBP was first described in 1994 [33]. It has since been found to be a powerful feature for texture classification.

The LBP feature vector in its simplest form is created in the following manner.

- Divide the examined window to cells (e.g. 16x16 pixels for each cell)
- For each pixel in a cell, compare the pixel to each of its 8 neighbors (on its left-top, left-middle, left-bottom, right-top, etc.). Follow the pixels along a circle, i.e. clockwise or counter-clockwise.
- Where the center pixel's value is greater than the neighbor, write '1'. Otherwise, write '0'. This gives an 8-digit binary number (which is usually converted to decimal for convenience)
- Compute the histogram, over the cell, of the frequency of each 'number' occurring (i.e., each combination of which pixels are smaller and which are greater than the center)
- Normalize the histogram

- Concatenate normalized histograms of all cells. This gives the feature vector for the window

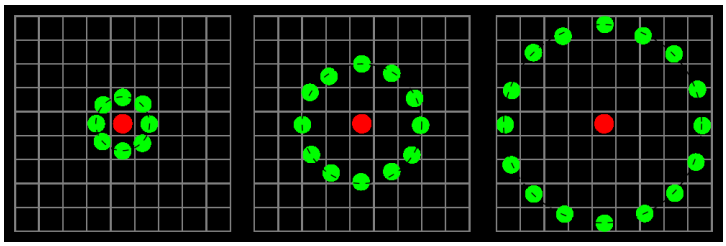


Figure 3.3: Local Binary Patterns

Chapter 4

EXPERIMENTS AND RESULTS

4.1 ACTIVITY RECOGNITION

For experiments, we choose the UMD Human Activity Dataset [35] and the UCSD Traffic Dataset[9]. The UMD database consists of 10 different activities: Bend, Jog, Push, Squat, Wave, Kick, Batting, Throw, Turn Sideways and Pick Phone. Each activity was repeated 10 times, so there were a total of 100 sequences in the dataset. Each sequence consists of 80 images and were cropped to a resolution of 331×301. Some samples of the various activites

Figure 4.1: Samples images from various activities from UMD dataset

are shown in Figure 4.1. Each image is sensed compressively at measurement factors of 100, 400, 800, 1000 and 1200 by taking the corresponding number of random measurements. To achieve this, we multiplied the full images with

a sensing matrix ϕ, which contained Gaussian i.i.d entries with expectation 0 and variance $\frac{1}{M}$, where M is the number of the measurements, corresponding to the compression factor. Since the background is relatively static, in each sequence, differences of compressive measurements of successive images are taken to remove the effect of the static background. These difference measurements are used to generate a distance matrix of size 79×79 for each sequence. As explained before these distance matrices are viewed as textures.

Activity	1	2	3	4	5	6	7	8	9	10
1	10	0	0	0	0	0	0	0	0	0
2	0	10	0	0	0	0	0	0	0	0
3	0	0	9	1	0	0	0	0	0	0
4	0	0	0	10	0	0	0	0	0	0
5	0	0	0	0	10	0	0	0	0	0
6	3	0	0	0	0	6	0	1	0	0
7	0	0	0	0	0	0	10	0	0	0
8	1	0	0	0	0	0	0	7	1	1
9	0	0	0	0	0	0	0	0	10	0
10	0	0	0	0	0	0	0	2	0	8

Table 4.1: Confusion table for activity recognition experiment using compressive measurements at a compression ratio = 100. The confusion matrix exhibits a strong diagonal structure, which implies that most activities are recognized correctly.

We used local binary pattern features [28] to classify the textures. Thus, each sequence is represented by LBP feature descriptor of length 38 which gives the normalized histograms of 38 binary patterns. For this experiment, we performed a leave-one-execution-out test, in which we trained on 9 executions and tested on the remaining execution for all activities using a simple nearest-neighbor classifier. In table 4.1, we show the confusion matrix obtained for the activity recognition experiment using the proposed method for a compression factor of 100. The classification accuracy is obtained to be 90%.

Compression factor	Recognition Rate
Uncompressed	90%
100	90%
400	86%
800	84%
1000	81%
1200	80%

Table 4.2: Activity recognition rate for different compression factors. The recognition rates are quite stable even at very high compression rates.

In table 4.2, we present average recognition results when the compression ratio was varied across a broad range of values. We observe that the proposed framework works very well across a wide variety of compression factors. These are encouraging and positive results, which suggest that significant performance improvements are possible by a careful choice of features and classifiers. The UCSD Traffic Dataset[9] consists of 254 videos capturing the highway traffic in Seattle. These videos are acquired from a single stationary camera over two days. The database contains different kinds of traffic patterns and weather conditions like overcast, sunny, rain drops on the camera lens. Each video is of length 50 frames at a resolution of 320×240 pixels. The database was labeled according to the amount of the traffic congestion in each video. Out of the 254 sequences, 44 are of heavy traffic, 45 of medium traffic, and 165 of light traffic. In figure 4.2, sample images from the three types of traffic are shown. We perform a classification experiment of the videos into these three categories. There are four different train-test scenarios provided with the dataset. For comparison, firstly at fixed compression ratio of $25\times$, we perform the same experiments with CS-LDS [31] as well as our method. The results show that our method performs significantly better than CS-LDS method for compression ratio equal to 25. Secondly, we perform the 4 experiments using our method for different compression ratios.

(a) Light

(b) Medium

(c) Heavy

Figure 4.2: Samples images from 3 different types of traffic from UCSD dataset: Light, Medium and Heavy

	Expt.1	Expt.2	Expt.3	Expt.4
Our method	92.06	92.19	85.94	92.06
CS-LDS(d=10)	84.12	87.5	89.06	85.71

Table 4.3: Classification results (in %) on the UCSD Traffic Dataset.

Compression ratio	Expt.1	Expt.2	Expt.3	Expt.4
25×	92.06	92.19	85.94	92.06
150×	88.89	78.13	78.13	82.54
300×	87.30	82.81	76.56	82.54

Table 4.4: Classification results at different compression ratios (in %) on the UCSD Traffic Dataset.

4.2 ERROR IN NTRPS

In section 3.4, we derived a relation to quantify the error between the NTRPs in the image domain and compressed domain. Here we empirically verify the inequalities derived for this deviation by calculating NTRPs for sequences of images for activities from the UMD dataset. It is shown in figure 4.3 that empirical results are consistent with the bounds given by equation (3.13). We

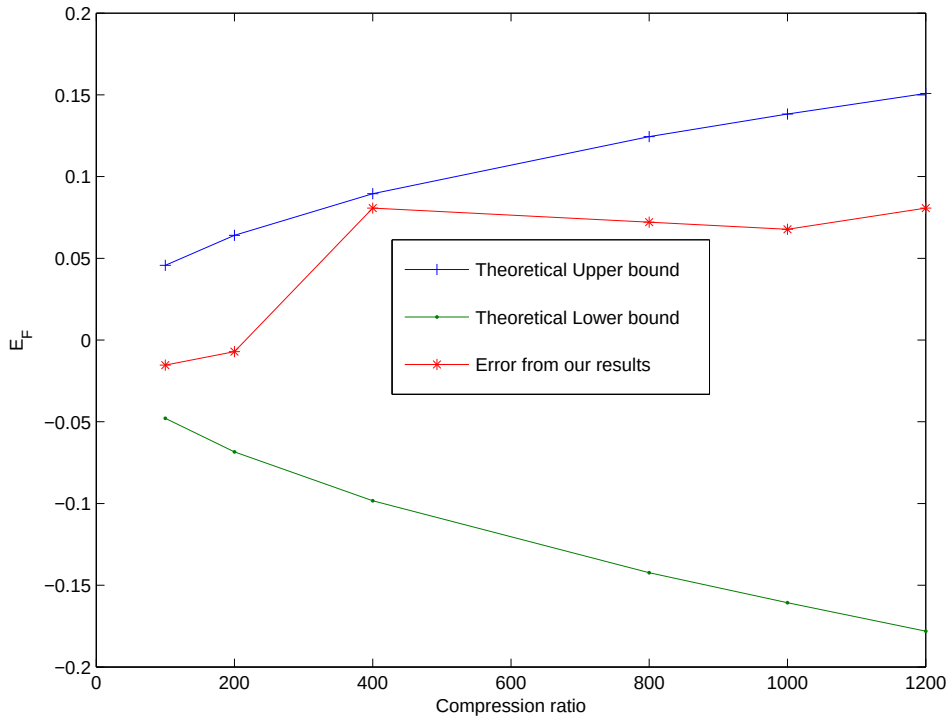


Figure 4.3: Comparison of the upper and lower bounds for E_F with empirical results. The errors in NTRPs obtained for sequences of images for activities from UMD dataset, for different compression ratios are consistent with the theoretical bounds calculated. The errors are very much within tolerable limits even for high compression ratios, thus promising the applicability of our approach for similar inference problems.

notice that the deviation in NTRPs is very much within tolerable limits even for a very high compression ratio of 1200, thus promising the applicability of our approach for similar inference problems. The deviation, E_F naturally increases with ϵ which in turn is proportional to logarithm of number of points needed to obtain NTRPs and inversely proportional to the number of measurements M. Thus for a fixed deviation, it is possible to decrease the number of measurements, if we use less number of points to obtain NTRPs. From this we can conclude that if we know in advance that a small number of frames gives discriminative information about a activity, we can reduce the number of measurements to capture those fewer number of frames.

Chapter 5

FUTURE WORK

In this thesis, we presented a framework to address activity recognition from compressive cameras. This has potential applications in a wide variety of resource constrained contexts such as in remote air-borne surveillance, or home-based security and health-care systems. We proposed a solution based on dynamical analysis via recurrence relations, which has an interpretation in terms of geometric structures of high-dimensional data. We showed that these geometric structures are preserved even in the compressed domain, and do contain significant discriminative information to recognize activities at very low data-rates. We further quantified the deviation in geometric structures in compressed domain from those in original image domain and showed that the deviations are within tolerable limits even for a very high compression ratio. Having explored feature extraction at the most basic level for action recognition, future research can be pursued in following two ways. Firstly, we plan to the explore the possiblity of extraction of more sophisticated and traditionally successful features like motion vectors for activity recognition. Secondly, we will be looking into the problem of extraction of more general features like shapes, integral images directly from compressed images, the kind of features that are useful in general computer vision problems and not necessarily only activity recognition.

[1] J.K. Aggarwal and M.S. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43:16:1–16:43, April 2011.

[2] A. N. Rajagopalan Naresh P. Cuntoor Amit K. Roy-chowdhury Volker Krüger Amit Kale, Aravind Sundaresan. Identification of humans using gait. 2004.

[3] Richard G. Baraniuk and Michael B. Wakin. Random projections of smooth manifolds. *Found. Comput. Math.*, 9(1):51–77, 2009.

[4] R. Bowden. Learning statistical models of human motion. *In IEEE Workshop on Human Modelling, Analysis and Synthesis*, 2000.

[5] E. J. Candès, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.

[6] Emmanuel J. Candès and Michael B. Wakin. An introduction to compressive sampling. *IEEE Signal Processing Magazine*, pages 21 – 30, 2008.

[7] M.C. Casdagli. Recurrence plots revisited. *Physica D: Nonlinear Phenomena*, 108:12 – 44, 1997.

[8] V. Cevher, A. C. Sankaranarayanan, M. F. Duarte, D. Reddy, R. G. Baraniuk, and R. Chellappa. Compressive sensing for background subtraction. In *Euro. Conf. Comp. Vision*, Oct. 2008.

[9] A. B. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. In *IEEE Conf. Comp. Vision and Pattern Recog*, June 2005.

[10] R. Chaudhry, A. Ravichandran, G. D. Hager, and R. Vidal. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In *IEEE Conf. Comp. Vision and Pattern Recog*, pages 1932–1939, 2009.

[11] R. Coifman, F. Geshwind, and Y. Meyer. Noiselets. *Applied and Computational Harmonic Analysis*, 10(1):27 – 44, 2001.

[12] O Concha, R. Xu, and M Piccardi. Compressive sensing of time series for human action recognition. In *Proceedings of the 2010 International*

Conference on Digital Image Computing: Techniques and Applications, DICTA '10.

[13] N. P. Cuntoor and R. Chellappa. Epitomic representation of human activities. 2007.

[14] Achlioptas. D. Database-friendly random projections. pages 274–281, 2001.

[15] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *IEEE Conf. Comp. Vision and Pattern Recog*, pages 886–893, 2005.

[16] M. F. Duarte, M. A. Davenport, D. Takhar, J. N. Laska, T. Sun, K. F. Kelly, and R. G. Baraniuk. Single-pixel imaging via compressive sampling. *IEEE Signal Process. Mag.*, 25(2):83–91, 2008.

[17] J. P. Eckmann, S. O. Kamphorst, and D. Ruelle. Recurrence plots of dynamical systems. *Europhysics Letters*, 5(9):973–977, 1987.

[18] Michael Elad. *Sparse and Redundant Representations - From Theory to Applications in Signal and Image Processing.* Springer, 2010.

[19] Chinmay Hegde, Michael B. Wakin, and Richard G. Baraniuk. Random projections for manifold learning. In *Neural Information Processing Systems (NIPS)*, Dec. 2007.

[20] Y. S. Hsu, S. Prum, J. H. Kagel, and H. C. Andrews. Pattern recognition experiments in the mandala/cosine domain. *IEEE Trans. Pattern Anal. Mach. Intell.*, 5(5):512–520, May 1983.

[21] J. S. Iwanski and E.. Bradley. Recurrence plots of experimental data: To embed or not to embed? *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 8(4):861–871, 1998.

[22] I. Laptev and T. Lindeberg. Space-time interest points. *IEEE Intl. Conf. Comp. Vision.*, 2003.

[23] C-S. Lee and A. Elgammal. Modeling view and posture manifolds for tracking. pages 1 –8, oct. 2007.

[24] C-S. Lee and A. Elgammal. Coupled visual and kinematics manifold models for human motion analysis. *International Journal on Computer Vision*, 87, 2010.

[25] Z. Liu and S. Sarkar. Improved gait recognition by gait dynamics normalization. 28(6):863–876, 2006.

[26] N. Marwan, M. C. Romano, M. Thiel, and J. Kurths. Recurrence plots for the analysis of complex systems. *Physics Reports*, 438(5-6):237, 2007.

[27] A. K. Roy-Chowdhury N. Vaswani and R. Chellappa. "shape activity": a continuous-state hmm for moving/deforming shapes with application to abnormal activity detection. 14(10):1603–1616, 2005.

[28] Timo Ojala, Matti Pietikäinen, and Topi Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(7):971–987, 2002.

[29] V. S. Subrahmanian O. Udrea P. Turaga, R. Chellappa. Machine recognition of human activities: A survey. *Circuits and Systems for Video Technology, IEEE Transactions on In Circuits and Systems for Video Technology*, 18(11):1473–1488, 2008.

[30] Lawrence R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Readings in speech recognition*, pages 267–296.

[31] A. C. Sankaranarayanan, P. Turaga, R. Baraniuk, and R. Chellappa. Compressive acquisition of dynamic scenes. In *under review at SIAM J. Imaging Sciences*.

[32] A. C. Sankaranarayanan, P. Turaga, R. Baraniuk, and R. Chellappa. Compressive acquisition of dynamic scenes. In *Euro. Conf. Comp. Vision*, Sep. 2010.

[33] M. Pietikäinen T. Ojala and D. Harwood. Performance evaluation of texture measures with classification based on kullback discrimination of distributions. *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, 1994.

[34] V. Thirumalai and P. Frossard. Correlation estimation from compressed images. *accepted in Journal of Visual Communication and Image Representation*, 2012.

[35] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury. The function space of an activity. *IEEE Conf. Comp. Vision and Pattern Recog*, pages 959–968, 2006.

[36] A. Veeraraghavan, A. Roy-Chowdhury, and R. Chellappa. Matching shape sequences in video with an application to human movement analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1896–1909, 2005.

[37] 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 *Image Processing, 2006 IEEE International Conference on*, pages 1273 –1276, oct. 2006.

[38] Webber Jr. C.L. Zbilut, J.P. Embeddings and delays as derived from quantification of recurrence plots. *Physics Letters A*, 171(3-4):199–203, 1992.