Transport Transforms and Its Application to Demultiplexing Orbital Angular Momentum Beams

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Electrical Computer Engineering

Se Rim Park
B.S., Electrical Computer Engineering, Yonsei University

Carnegie Mellon University
Pittsburgh, PA

August, 2017

Acknowledgements

First of all, Dr. Rohde, thank you for your guidance and support throughout the years. Without your advice that gravitated me toward the right direction, I would have been lost amid swirls of researches. I would also like to present my sincere gratitude to my exemplary committee members, Dr. Vijayakumar Bhagavatula, Dr. Dejan Slepcev, Dr. Aswin Sankaranarayanan, for their commitment and passion. I am grateful for the opportunities that I had to work with the smartest and the most fabulous collaborators and colleagues over the past, including Dr.Jonathan Nichols, Dr.Timothy Doster, and Dr.Abbie Watnik, Dr.Jinwon Lee, Dr.Soheil Kolouri, Dr.Liam Cartell, Dr.Mattew Thorpe, Chi Liu, and Dr.Shinjini Kundu. I acknowledge that this dissertation would not have been possible without funding from the Korea Foundation for Advanced Studies and Benjamin Garver Lamme/Westinghouse Graduate Fellowship. Thanks to all those whose life once crossed with mine, but I failed to acknowledge. I sincerely wish you all the best, and hope our life crosses once more. Lastly, Mom, Dad, and Seeun, your unconditional love and belief in me, are the foundation of my life. I love you.

Abstract

Discriminating data classes emanating from sensors is an important problem with many applications in science and technology. This study describes a new transform for pattern identification that interprets patterns as probability density functions, and has special properties with regards to classification. The transform, built upon the optimal transport theory, is invertible, with well defined forward and inverse operations. This study shows that the transform can be useful in 'parsing out' variations that are 'Lagrangian' (displacement and intensity variations) by converting these to 'Eulerian' (intensity variations) in transform space. This conversion is the basis for the main result that describes when the transforms can allow for linear classification to be possible in transform space. Demonstrated with computational experiments that used both real and simulated data, the transforms can help render a variety of real world problems simpler to solve.

Moreover, making use of a newly developed theory suggesting a link between image turbulence and photon transport through the continuity equation, the transform is utilized to perform a decoding task for orbital angular momentum carrying beam patterns. Free space optical communications utilizing orbital angular momentum beams have recently emerged as a new technique for communications with potential for increased channel capacity. Turbulence due to changes in the index of refraction emanating from temperature, humidity, and air flow patterns, however, add nonlinear effects to the received patterns, thus making the demultiplexing task more difficult. The decoding technique is tested and compared against previous approaches using deep convolutional neural networks. Results show that the new method can obtain comparable classification accuracies (bit error rate) at a fraction of the computational cost, thus enabling higher bit rates.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Mathematical Transformations in the Domain of Pattern Recognition

Mathematical transforms are useful tools in engineering, physics, and mathematics given that they can often render certain problems easier to solve in transform space. Fourier transforms [3] for example, are well-known for providing simple answers related to the analysis of linear time-invariant systems. Wavelet transforms, on the other hand, are well suited for detecting and analyzing signal transients (fast changes) [4]. These and other transforms have been instrumental in the design of sampling and reconstruction algorithms for analog-to-digital conversion, modulation and demodulation, compression, communications, etc, and have found numerous applications in science and technology.

On the other hand, the past few decades have brought about the emergence of ubiquitous, accurate, user friendly, and low cost digital sensing devices. These devices produce a wealth of data about the world we live in, ranging from digital microscopy images of sub-cellular patterns to satellite imagery and detailed telescope images of our universe. The relative ease with which vast amounts of data can be accessed and queried for information have brought about challenges related to 'telling signals apart',

or sensor data classification. Examples include being able to distinguish between benign and malignant tumors from medical images [5], between 'normal' and 'abnormal' physiological sensor data (e.g. flow cytometry) [6], identifying people from images of faces or fingerprints [7], identifying biological/chemical threats from resonant optical spectra [8] and others. For such problems, mathematical transforms have been used as low level representation models to facilitate pattern recognition by simplifying feature extraction from data.

Some interesting applications of transforms in pattern recognition are presented in [9, 10, 11, 12, 13]. In [9], discrete Fourier transform (DFT) was used for palm print identification. Monro et al. [10] used discrete cosine transform (DCT) for iris recognition. Wavelet coefficients were used as texture features in [11] for image retrieval. Mandal et al. [12] used curvelet-based features for face recognition. The Radon transform was used for Gait recognition in [13]. The list above is obviously not exhaustive. They represent just but a few examples of many applications of transforms in pattern recognition.

A common property among aforementioned transforms is that they are all invertible linear transforms that seek to represent a given image as a linear combination of a set of functions (or discrete vectors for digital signals). What we mean by an invertible linear transform, $\mathscr{F}$, is that for images I and J, $\mathscr{F}$ satisfies $\mathscr{F}(I) + \mathscr{F}(J) = \mathscr{F}(I + J)$, $\mathscr{F}(\alpha I) = \alpha\mathscr{F}(I)$, and $\mathscr{F}^{-1}$ exists. Linear transforms are unable to alter the 'shape' of image classes (i.e. distribution of the point cloud data) so as to fundamentally simplify the actual classification task. For example, linear operations are unable to render classification problems that are not linearly separable into linearly separable ones. When considering many important classification tasks, it is not hard to understand the problem at an intuitive level. One can often visually observe that in many categories (e.g. human faces, cell nuclei, galaxies, etc.) a common way in which one data differ from one another is not only in their intensities, but also in where the intensities are positioned. By definition, however, linear transforms must operate at fixed pixel coordinates. As such, they are unable to move or dislocate pixel intensities in any way. Hence, for pattern recognition purposes, linear transforms are usually followed by a nonlinear operator to demonstrate an overall nonlinear effect (e.g. thresholding in curvelet and

wavelet transforms, magnitude of Fourier coefficients, blob detection/analysis in Radon transform, etc.).

Many feature extraction methods have been developed for images [14, 15, 16] along side with the end to end deep neural network approaches such as convolutional neural networks [17, 18] and scattering networks (ScatNets) [19, 20, 16]. These recent methods have proven to be very successful in image classification and they have improved the state of the art classification for a wide range of image datasets. Such methods, however, are often not well suited for image modeling applications, including imaging and image reconstruction, as they provide a noninvertible nonlinear mapping from the image space to the feature space. Meaning that while the nonlinearity of the image classes are captured through the extracted features, any statistical analysis in the feature space does not have a direct interpretation in the image space as the mapping is noninvertible.

1.2 Motivations for a New Transform

Important practical questions often arise in the process of designing solution to many data classification problems. Examples would be: "Which features should be extracted?", "What classifier should be used?", "How can one model, visualize and understand any discriminating variations in the dataset?", etc. For many applications where optimal feature sets are yet to be discovered, researchers are faced with the task of utilizing a *trial and error* approach that involves testing for different combinations of features [21, 22], classifiers [23], kernels [24] in the effort to arriving at a useful solution of the problem. We note that many of the available signal transforms (Wavelet, Fourier, Hilbert, etc.) are linear transforms, and thus offer limited capabilities related to enhancing or facilitating separation in feature (transform) space unless some non-linear operations are performed.

The new signal data transformation framework described in this study renders certain classification problems linearly separable in the transform space. Linear separability in the transform space gains practical importance with datasets that contain a small number of high dimensional signals. When the number of available signals for training

are far less than their dimension, the nonlinear classifiers become prone to overfitting. This is a well known effect, and is addressed as the problem of high dimensional and low sample size (HDLSS) [25] in the literature. In addition, the overall variance of a classifier increases as the classifier becomes more complex [26], and often times simpler classifiers (e.g. linear) can yield higher accuracies than more sophisticated ones [27]. Transforming the data and rendering it to be linearly separable will help maintain small classification error, balance the bias/variance tradeoff, streamline the implementation of classification systems in many real world problems, and could bypass the often time consuming process of devising large sets of specially tailored numerical signal descriptors and testing each descriptor with various classifiers.

Signal Discrimination Problems

Let $\mathbb{P}$ and $\mathbb{Q}$ denote two disjoint classes of functions (signals) within a normed vector space V. The goal in classification is to deduce a functional to 'regress' a given label for each signal [28]. For a binary classification problem, the label of each signal can be considered 0/1 or -1/+1, and the problem of classifying a signal f can be solved by finding a linear functional $T : V \to \mathbb{R}$ and $b \in \mathbb{R}$ such that

$$\begin{aligned} T(f) &< b \qquad \forall f \in \mathbb{P}, \\ T(f) &> b \qquad \forall f \in \mathbb{Q}. \end{aligned} \tag{1.1}$$

Below we specifically consider the case when T is a linear classifier in V. For example, for real functions in L^2, one may find w such that $T(f) = \int_V w(x)f(x)dx$. For discrete signal data in countable domain $\mathbb{Z}$ one may find w such that $T(f) = \sum_{k\in\mathbb{Z}} w[k]f[k]$. Thus the goal is to obtain the linear function w and the scalar b from labeled data. In practice, linear classifiers are important given their efficient implementation, and favorable bias-variance trade off, especially in classification of high dimensional data [29].

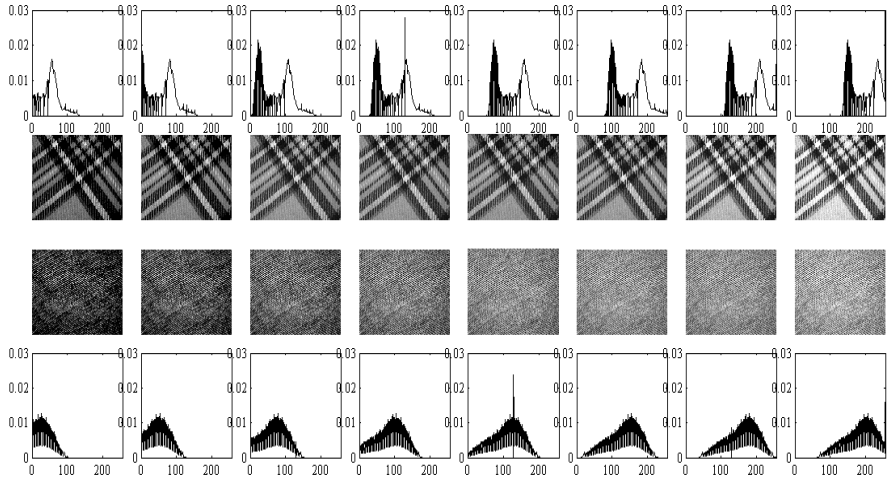


Figure 1.1: Two types of textures under illumination variation and their corresponding intensity histograms.

1.2.1 An illustrative example

Consider the problem of discriminating images of two different image patterns. The first column of Figure 1.1, contains two sample images from the UIUC Texture dataset [30], with their intensity histograms of the corresponding textures appearing directly above or beneath each texture. Now consider the same texture images, but under different brightness (which causes a translation of the histograms) and linear contrast (which causes a scaling of the histograms). Such variations in brightness and contrast are displayed in the different columns of Figure 1.1. A generative model for the histogram data corresponding to each texture class under brightness and contrast variations can be built by translation and scaling of the histograms. In other words, we

Table 1.1: Average Classification Error of the texture dataset

Classifier type	Dataset	L^2 space	CDT space
Fisher LDA	Training set	0 %	0%
	Testing set	**56.36** %	**0.84**%
PLDA	Training set	41.81 %	0%
	Testing set	**44.39** %	**0**%
Linear SVM	Training set	57.02 %	0.20%
	Testing set	**50.06** %	**1.60**%

generate a set of histograms $\{p_i\}_{i=1}^N$ and $\{q_j\}_{j=1}^N$, each belonging to class $\mathbb{P}$ and $\mathbb{Q}$, by appropriately scaling (a) and translating (μ) 'prototype' signals p_0 and q_0, such that $p_i(x) = p_0(a_i(x - \mu_i))$ and $q_j(x) = q_0(a_j(x - \mu_j))$. Finally, we note that stationary additive noise in these images can be modeled as a convolution of each signal p_i or q_j with the appropriate probability density of the noise model.

In order to illustrate the main difficulty with utilizing linear classification methods under these sources of signal variation, we attempted to train a linear classifier to a set of histograms under random brightness (μ) and contrast (a). We used a well-known Fisher Linear Discriminant Analysis (Fisher LDA) method [31] that seeks to maximize the differences in the projected mean of each class, while at the same time minimizing their intra class variances. We also generated a testing set by again applying the same brightness and contrast random model to the image data to create a testing data. Table 1.1 contains both the average training and testing error of 5-fold cross validation when using this simulated data model. It is clear that while the training error is very low, the resulting linear classifier fails to generalize to test data not used in training. We note that there is nothing special related to the use of the Fisher LDA criterion in solving for w in this example. That is, similar results are obtained utilizing linear Support Vector Machines instead (see Table 1.1).

Simple consideration of the structure of the problem can reveal the reason why it is hard to fit linear classifier to the testing dataset. This is because a single w, a linear classifier, is unable to 'cope' with the translation and scaling variations encountered in the test data $p_0(a_{ts}(x-\mu_{ts}))$. In other words, the operation $\int_V w(x)p_0(a_{ts}(x-\mu_{ts}))dx$ fails to satisfy equation (1.1) for randomly selected a_{ts} and μ_{ts} used to generate the test set. To be clear, it is well-known that, for a training set of fixed size, and for data of large enough dimension, a linear classifier w can always be found that will near perfectly separate the training data [32]. However, as this simple simulation is meant to clarify, such classifier may fail to generalize to testing data if such w fails to capture anything meaningful about the mathematical generative model of the problem. This is the phenomenon exemplified here.

Now, the histograms in this problem could be rendered linearly separable if, for any input histogram, one could simply 'mod out' the translation and scaling parameters,

thus removing the confounding variations rendering the problem not linearly separable. This is the intuition behind the Cumulative Distribution transform (CDT). It is able to handle variations such as translation, scaling, and others by computing rearrangements in the locations of the signal intensities with respect to a chosen reference, which does not require the estimation of the prototype histograms p_0 and q_0. Results in Table 1.1 show that the same Fisher LDA and SVM technique, when applied to data that have been transformed with the CDT, is perfectly able to separate the data.

1.3 Optimal Transport Primer

Let μ, ν be *probability* measures on probability spaces $(\mathcal{X}, \Sigma(\mathcal{X}))$, $(\mathcal{Y}, \Sigma(\mathcal{Y}))$. $\Sigma(A)$ refers to a σ-algebra of a measurable set A.

Marginals

Let π be a probability measure on $\mathcal{X} \times \mathcal{Y}$. Its marginal (or projection) on $\mathcal{X}$ (or $\mathcal{Y}$) is the measure $(proj_{\mathcal{X}})_{\#}\pi$ (or $(proj_{\mathcal{Y}})_{\#}\pi$), where $proj_{\mathcal{X}}$ and $proj_{\mathcal{Y}}$ stand for the projection maps $(x, y) \to x$ and $(x, y) \to y$.

Transport Plan

Let $\Pi(\mu, \nu)$ be the set of all joint probability measures on $\mathcal{X} \times \mathcal{Y}$ whose marginals are μ and ν. Then Π is transport plan. Note that there is always a trivial transport plan in which the variables X and Y are independent, i.e. $\pi(x, y) = \mu(x)\nu(y)$. If there exists a measurable function $f : \mathcal{X} \to \mathcal{Y}$ such that $Y = f(x)$, then f is called a transport map. Transport map does not always exist (for example, when μ is a Dirac mass and ν is not).

Transport Map

If a measurable map f pushes μ onto ν such that

$$\int_{f^{-1}(A)} d\mu = \int_A d\nu \qquad \text{for any measurable } A$$

then f is a mass preserving map, or a transport map. Informally, one can say that f transports the mass represented by the measure μ to the mass represented by the measure ν.

Optimal Transport

The optimal transport, introduces a cost function $c(x, y)$ on $\mathcal{X} \times \mathcal{Y}$, that can be interpreted as the work needed to move on unit mass from location x to location y. The Monge-Kantorovich minimization problem considers finding the solution for:

$$\inf_{\mathcal{X}\times\mathcal{Y}} c(x, y) d\pi(x, y), \tag{1.2}$$

where the infimum runs over all joint probability measures π on $\mathcal{X} \times \mathcal{Y}$ with marginals μ and ν. Such joint measures are called **transport plans**, those achieving the infimum are called **optimal transport plans**.

Under certain assumptions, there exists a **transport map** f. The search of the transport map is called the Monge problem:

$$\inf_{\mathcal{X}\times\mathcal{Y}} c(x, f(x)) d\mu(x).$$

When $c(x, y) = |x - y|^2$ in the Euclidean space, μ is absolutely continuous with respect to Lebesgue measure, and μ, ν have finite moments of order 2, then there is a unique optimal transport map between μ and ν.

Wasserstein distance

The p-Wasserstein metric, W_p, for $p \geq 1$ on $P_p(\Omega)$, set of Borel probability measures on Ω, can be defined as using the optimal transportation problem (1.2) with the cost function $c(x, y) = |x - y|^p$. For μ and ν in $P_p(\Omega)$,

$$W_p(\mu, \nu) = \left(\inf_{\pi\in\Pi(\mu,\nu)} \int_{\Omega\times\Omega} |x - y|^p d\pi(x, y) \right)^{\frac{1}{p}}. \tag{1.3}$$

For any $p \geq 1$, W_p is a metric on $P_p(\Omega)$.

Unique optimal transport map in $\mathbb{R}$

Let μ, ν be two probability measures on $\mathbb{R}$, and define their cumulative distribution function by

$$F(x) = \int_{-\infty}^{x} d\mu, \qquad G(y) = \int_{-\infty}^{y} d\nu.$$

Further, define their right-continuous inverse by

$$F^{-1}(t) = \inf\{x \in \mathbb{R}; F(x) > t\}$$
$$G^{-1}(t) = \inf\{y \in \mathbb{R}; G(x) > t\}$$

and set

$$f = G^{-1} \circ F.$$

If μ does not have atoms, then f is the optimal transport map that pushes μ onto ν.

Wasserstein distance

The p-Wasserstein metric, W_p, for $p \geq 1$ on $P_p(\Omega)$, set of Borel probability measures on Ω, can be defined as using the optimal transportation problem (1.2) with the cost function $c(x, y) = |x - y|^p$. For μ and ν in $P_p(\Omega)$,

$$W_p(\mu, \nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega} |x - y|^p d\pi(x, y) \right)^{\frac{1}{p}}. \tag{1.4}$$

For any $p \geq 1$, W_p is a metric on $P_p(\Omega)$.

1.4 Transport-based Approaches in Pattern Recognition

In contrast to commonly used linear signal transformation frameworks (e.g. Fourier and Wavelet transforms) which only employ signal intensities at fixed coordinate points, thus adopting an 'Eulerian' point of view, the idea behind the transport-based approaches is to consider the intensity variations together with the locations of the in-

tensity variations in the signal. Therefore, such transport-based solutions adopt a 'Lagrangian' point of view for analyzing signals.

Previous works have shown that the optimal transport theory can be utilized to solve pattern discrimination tasks. [33] utilized the optimal transport framework for nuclei image classification, and showed that the classification accuracy obtained using the L_2 Wasserstein metric are as good or better than those obtained utilizing a set of feature-based method. The work was further developed into [34], named the linear optimal transform (LOT) framework, which measures similarities between pairs of images by Kantorovich-Wasserstein distance. Under this framework, LOT Euclidean embedding is computed for each image, which can be viewed as a nonlinear image transformation method.

Given their suitability for comparing mass distributions, transport-based approaches for performing pattern recognition of morphometry encoded in image intensity values have also recently emerged. Recently described approaches for transport-based morphometry (TBM) [34, 35, 36] work by computing transport maps or plans between a set of images and a reference or template image. The transport plans/maps are then utilized as an invertible feature/transform onto which pattern recognition algorithms such as principal component analysis (PCA) or linear discriminant analysis (LDA) can be applied. These techniques has been recently employed to decode differences in cell and nuclear morphology for drug screening [35], and cancer detection histopathology [37, 38] and cytology [39], amongst other applications including the analysis of galaxy morphologies [36], for example.

Specifically, Fig. 1.2 shows how transport based learning can be applied to model the variation in a nuclei dataset (in this case malignant versus benign). The optimal transport maps between input images and a template image I0 are calculated. Next, linear statistical modeling such as principal component analysis (PCA), linear discriminant analysis (LDA), and canonical correlation analysis (CCA) is performed on the optimal transport maps. The resulting transport maps obtained from PCA, LDA, and CCA can then be inverted back to image space.

We note the strong similarity between deformation-based methods which have long been used to analyzed radiological images [40, 41], for example. The difference be-

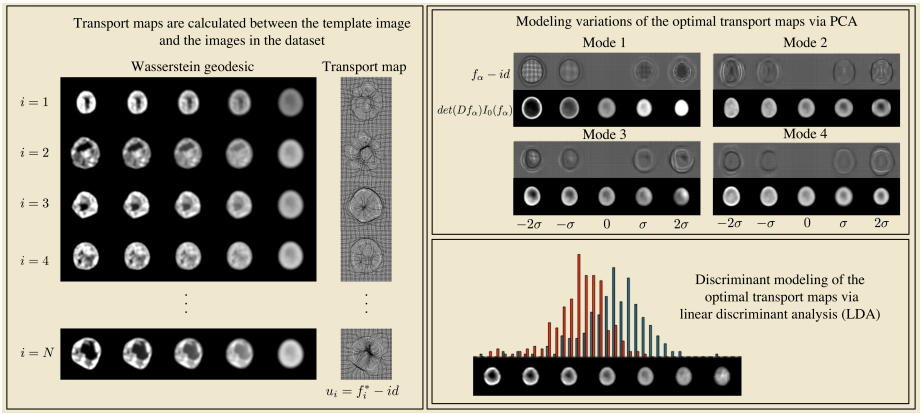


Figure 1.2: An Example of Transport-based Learning

ing that the transport based approach allows for numerically exact, uniquely defined solutions for the transport plans or maps used. That is, images can be matched with little perceptible error. The same is not true in methods that rely on registration via the computation of deformations, given the significant topology differences commonly found in medical images. Moreover, transport based approach allows for comparison of the entire intensity information present in the images (shapes and textures), while deformation-based methods are usually employed to deal with shape differences.

1.5 Notations and Symbols

Through out the thesis, we will consider two probability spaces $(X, \Sigma(X), \mathcal{I}_0)$ and $(Y, \Sigma(Y), \mathcal{I}_1)$ where X and Y are connected sets in $\mathbb{R}^n$. $\Sigma(A)$ refers to a σ-algebra of measurable set A, and $\mathcal{I}_0$ and $\mathcal{I}_1$ are probability measures, i.e. $\mathcal{I}_0(X) = 1, \mathcal{I}_1(Y) = 1$. Furthermore, let $\mathcal{I}_0(A) > 0, \mathcal{I}_1(A) > 0$ for Lebesgue measurable set A whose $\lambda(A) > 0$, and let I_0 and I_1 denote density functions associated with $\mathcal{I}_0$ and $\mathcal{I}_1$, respectively: $d\mathcal{I}_0(x) = I_0(x)dx$, $d\mathcal{I}_1(x) = I_1(x)dx$. Let $f_1 : X \to Y$ define a measurable map that pushes $\mathcal{I}_0$ onto $\mathcal{I}_1$ such that

$$\int_{f_1^{-1}(A)} d\mathcal{I}_0 = \int_A d\mathcal{I}_1 \quad \text{for any Lebesgue measurable } A \subset Y. \tag{1.5}$$

In our case, we will consider $d = 1$ and $\mathcal{I}_0$ and $\mathcal{I}_1$ that have densities as defined above. In this case, the relation above can be expressed, through Lebesgue integration, as

$$\int_{\inf(X)}^{x} I_0(\tau)d\tau = \int_{\inf(Y)}^{f_1(x)} I_1(\tau)d\tau. \tag{1.6}$$

for In addition, certain results shown below will require us to interpret measurable densities I_0, I_1 and maps f_1, f_2 as elements of L^2 function spaces. That is, given a measurable map $f_1 : X \to Y$ defined as above, for example, we can view it as an element of the space of functions whose absolute square value is Lebesgue integrable. In this case, the space is denoted as $L^2(X)$ and is defined as the set of functions that satisfy:

$$\|f\|_2 = \left(\int_X |f|^2 d\lambda\right)^{\frac{1}{2}} < \infty,$$

with λ referring to the Lebesgue measure in X.

1.6 Contributions and Outline of the Thesis

In this thesis, we propose a signal transformation framework, for signals and images, designed to facilitate the pattern recognition problem. This thesis consists of five main chapters, in which Chapter 2-5 are the main contributions of the author.

The specific contributions of this thesis are:

- **Contribution 1**: Developing a new mathematical transform and a theory that can render data linearly separable in the transform space,
- **Contribution 2:** Providing experimental validation that the transform indeed can facilitate pattern recognition using various examples,
- **Contribution 3**: Developing a classification pipeline utilizing the transform to enhance the performance of optical communications system.

The second chapter of this thesis introduces the Cumulative Distribution transform. We show that CDT is a nonlinear signal transformation, which takes an input a signal (treated as probability distribution function), and outputs an invertible function that is

related to morphing that signal to a chosen reference signal. We show that under some general conditions, the CDT transform can turn not-linearly separable classes of signals into linearly separable classes in the CDT space.

In the third chapter, we extend the CDT, which is developed for signal (a function with one independent variable), to Radon-Cumulative Distribution transform, which is developed for images (a function with two independent variables). We show that the Radon-CDT also shares the linear separability theorem of the CDT.

The fourth chapter of this thesis aims to provide the experimental justification of the method as well as developing pattern recognition pipeline utilizing the transforms. Five datasets are tested for each transform, 10 in total. The experiments demonstrate that under the condition where data are generated according to our model, the classes of different signals/images would belong to distinct convex sets in the transform space, and therefore be linearly separable.

The fifth chapter of this thesis utilizes the transform and the classification pipeline developed above to aid solving the demultiplexing problem in free space optical communications. Based on the recent finding in [42] that the light traveling in atmospheric turbulence would approximately follow the 'optimal transport path', we claim that the laser beams undergo deformation which can be decoded easily in the transform space. We demonstrate that the demultiplexing in transform space yields comparable BER as in the image space with a fraction of the computational cost.

Finally, Chapter six concludes the thesis and lists future work and directions.

Chapter 2

The Cumulative Distribution Transform

2.1 Introduction

In this chapter, we describe a new one-dimensional signal transformation framework, with well-defined analysis (forward transform) and synthesis (inverse transform) operations that, for signals that can be interpreted as probability density functions, can help facilitate the problem of recognition. Denoted as the Cumulative Distribution transform (CDT), the CDT can be viewed as a one to one mapping between the space of smooth probability densities and the space of differentiable functions, and therefore by definition retains all of the signal information. We show that the CDT can be computed efficiently, and can turn certain types of classification problems linearly separable in the transform space. In contrast to linear data transformation frameworks (e.g. Fourier and Wavelet transforms) which simply consider signal intensities at fixed coordinate points, thus adopting an 'Eulerian' point of view, the idea behind the CDT is to also consider the location of the intensities in a signal, with respect to a chosen reference, in the effort to 'simplify' pattern recognition problems. Thus, the CDT adopts a 'Lagrangian' point of view for analyzing signals. The idea is similar to our work on linear

optimal transport [34], and the links will be explicitly elucidated below. The chapter is organized as follows. We present the definition of the CDT in Section 2.2 then its properties in Section 2.3. The linear separability property in CDT space is presented in Section 2.4, and a numerical method for approximating the forward CDT for discrete signals is described in Section 2.5.

2.2 The 1D Cumulative Distribution Transform

Consider two probability density functions I_0 and I_1 defined as in Sec. 1.5. Considering I_0 to be a pre-determined 'reference' density, one can use relation (1.6) to uniquely associate f_1 with a given density I_1. We use this relationship to define the *Cumulative Distribution Transform (CDT)* of I_1 (denoted as $\widehat{I}_1 : X \to \mathbb{R}$), with respect to the reference I_0:

$$\widehat{I}_1(x) = (f_1(x) - x)\sqrt{I_0(x)}. \tag{2.1}$$

with $f_1 : X \to Y$ satisfying (1.6) for $x \in X$.

Now let $J_0 : X \to [0,1]$ and $J_1 : Y \to [0,1]$ be the corresponding cumulative distribution functions for I_0 and I_1, that is: $J_0(x) = \int_{\inf(X)}^{x} I_0(\tau)d\tau$, $J_1(x) = \int_{\inf(Y)}^{x} I_1(\tau)d\tau$. With f_1 defined in (1.6) one can re-write $J_0 : X \to [0,1]$ as

$$J_0(x) = J_1(f_1(x)). \tag{2.2}$$

For continuous cumulative distribution functions J_0 and J_1 (functions whose first derivative exists throughout their respective domains), f_1 is a continuous and monotonic function. If f_1 is differentiable, (2.2) can be rewritten as

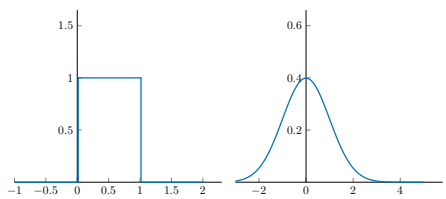
$$I_0(x) = f_1'(x)I_1(f_1(x)). \tag{2.3}$$

For measurable but discontinuous functions the relationship above does not hold for points at discontinuities.

The *inverse Cumulative Distribution Transform* of $\widehat{I}_1$ is defined as:

$$I_1(y) = \frac{d}{dy} J_0(f_1^{-1}(y)) = (f_1^{-1})' I_0(f_1^{-1}(y)) \tag{2.4}$$

where $f_1^{-1} : Y \to X$ refers to the inverse of f_1 (i.e. $f_1^{-1}(f_1(x)) = x$), $f_1(x) = \widehat{I}_1(x)/\sqrt{I_0(x)} + x$. Naturally, formula (2.4) holds for points where J_0 and f_1 are differentiable. By the construction above, f_1 will be differentiable except for points where I_0 and I_1 are discontinuous. Note that in practice, we have control over the definition of I_0, and in our numerical implementation described in section 6, we take it to be the uniform density. The example presented below shows the CDT of normal distribution density.



(a) $I_0(x)$ **(b)** $I_1(x)$

Figure 2.1: Example 2.2.1

Example 2.2.1. *Consider a probability density of uniform distribution $I_0 : [0, 1] \to \mathbb{R}$:*

$$I_0(x) = 1,$$

and a normal distribution density $I_1 : \mathbb{R} \to \mathbb{R}$ with zero-mean and unit-variance (see Figure 2.1):

$$I_1(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}.$$

$\int_{-\infty}^{\infty} I_1(\tau)d\tau = \int_0^1 I_0(\tau)d\tau = 1$ *holds by definition. To find the CDT for I_1 with respect to the reference I_0, we first solve for $f_1 : [0, 1] \to \mathbb{R}$:*

$$\int_{-\infty}^{f_1(x)} I_1(\tau)d\tau = \int_{-\infty}^{f_1(x)} \frac{1}{\sqrt{2\pi}} e^{-\tau^2/2} d\tau = \int_0^x 1 d\tau = x. \tag{2.5}$$

By setting $\Phi(x) = 1/\sqrt{2\pi}\int_{-\infty}^{x} e^{-\tau^2/2}d\tau$, (2.5) *can be rewritten as*

$$\Phi\left(f_1(x)\right) = x.$$

$\Phi(x)$ *is a monotonically increasing function, and the inverse exists. Hence, we get*

$$f_1(x) = \Phi^{-1}(x). \tag{2.6}$$

By substituting (2.6) *into* (2.1)*, we have found the CDT,* $\widehat{I}_1(x) : [0,1] \to \mathbb{R}$

$$\widehat{I}_1(x) = \Phi^{-1}(x) - x. \tag{2.7}$$

Figure 2.2 shows the plot (black dotted line) for the CDT of a normal distribution density function with zero mean and unit variance.

2.3 CDT Properties

Here we describe a few basic properties of the CDT, with the main purpose of elucidating certain of its qualities necessary for understanding its ability to linearly separate particular types of densities.

Property 2.3.1. ***Nonlinearity*** *The CDT is a non-linear transformation.*

For transformation A to be linear, we must have that $A(\alpha I_1 + \beta I_2) = \alpha A(I_1) + \beta A(I_2)$. It is easy to check by example 2.2.1 that this relation does not hold. Suppose $\alpha = 1/2$, $\beta = 1/2$, I_1 be a normal density and I_2 be a uniform density. Then $A(\alpha I_1 + \beta I_2) \neq \alpha A(I_1) + \beta A(I_2)$.

Before going on to state further properties of the CDT, it is worth expanding upon the geometric meaning of the CDT. We first note that, using the standard definition of the L^2 norm, i.e. $\|\widehat{I}_i\|_{L^2} = \left(\int_X |\widehat{I}_i(x)|^2 dx\right)^{1/2}$, we have:

$$\|\widehat{I}_1\|_{L^2}^2 = \int_X (f_1(x) - x)^2 I_0(x)dx. \tag{2.8}$$

As such, the quantity $\|\widehat{I}_1\|^2_{L^2}$ computes the 'amount' of intensity from I_0 at coordinate x that will be *displaced to* coordinate $f_1(x)$. Because f_1 is uniquely defined for nonzero probability densities, the quantity $\|\widehat{I}_1\|^2_{L^2}$ can be viewed as the minimum amount of 'effort' (quantified as density intensity $\times$ displacement) that must be applied to 'morph' I_1 onto I_0. This quantity can be interpreted as the optimal transport (Kantorovich-Wasserstein) distance between I_0 and I_1 [43]. Moreover, the set of continuous density functions is formally a *Riemannian manifold* [43] meaning that at any point in probability density space, there is a tangent space endowed with an inner product corresponding to the incremental intensity flow (see [34] for more details). Therefore the distance between I_1 and I_0 expressed in (2.8) can be interpreted as a geodesic distance over the associated manifold.

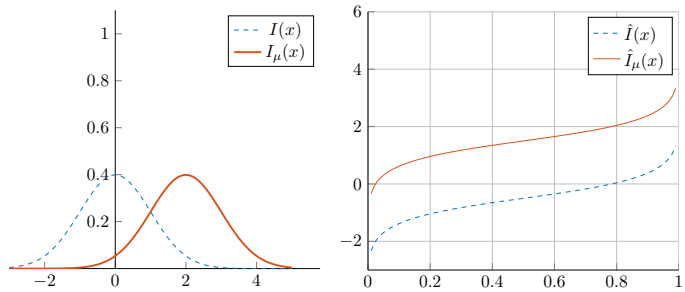
Now consider the distance between the CDT of two different densities I_1 and I_2, computed with respect to the same reference I_0:

$$\|\widehat{I}_1 - \widehat{I}_2\|^2_{L^2} = \int_X \left((f_1(x) - x) - (f_2(x) - x)\right)^2 I_0(x)dx \tag{2.9}$$

where f_1 and f_2 correspond to the mappings between I_1 and I_0, and I_2 and I_0 respectively. In two or more dimensions, as described in [34], this distance can be thought of as the 'linearized' optimal transport (generalized geodesic) metric between density functions I_1 and I_2. It can be interpreted as a azimuthal equidistant projection of I_1 and I_2 onto the plane associated with the incremental intensity flows about the point I_0. For one dimensional density functions, however, f is uniquely determined. Hence the optimal transport distance computed between densities I_1 and I_2 can also be expressed through (2.9) above. In short, the CDT of a given probability density function I_i can be viewed as an invertible embedding of the function onto a linear space that is isometric with respect to the standard optimal transport (also known as Earth Mover's) distance.

We now describe important properties of the CDT operation relative to density coordinate changes such as translation, scaling, and more generally diffeomorphisms applied to a given density.

Property 2.3.2. ***Translation.*** *Let I_μ represent a translation of the probability density I_1 by μ, $I_\mu(x) = I_1(x - \mu)$. The CDT of I_μ with respect to the reference probability*



(a) $I_\mu(x)$ **(b)** $\widehat{I}_\mu(x)$

Figure 2.2: Example 2.3.3

density $I_0 : X \to \mathbb{R}$ *is given by* $\widehat{I}_\mu : X \to \mathbb{R}$*:*

$$\widehat{I}_\mu(x) = \widehat{I}_1(x) + \mu\sqrt{I_0(x)}. \tag{2.10}$$

For a proof, see A.1.

Example 2.3.3. *Consider a translation of the density function* $I_1(x)$ *in Example 2.2.1 by* μ

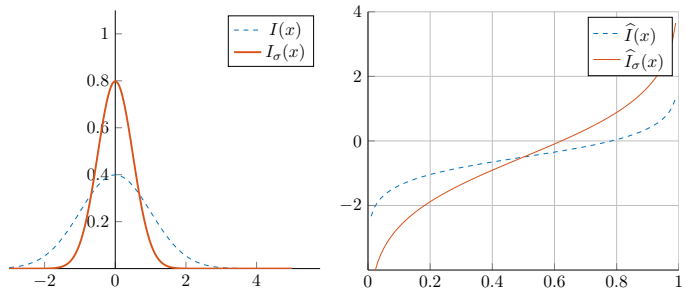
$$I_\mu(x) = I_1(x-\mu) = \frac{1}{\sqrt{2\pi}} e^{-(x-\mu)^2/2}.$$

This is a normal distribution with mean μ *and unit variance. The corresponding CDT,* $\widehat{I}_\mu : [0,1] \to \mathbb{R}$*, for* I_μ *with respect to the uniform reference density* $I_0 : [0,1] \to \mathbb{R}$ *can be found by the translation property* (2.10) *and by the CDT found in* (2.7)

$$\widehat{I}_\mu(x) = \widehat{I}_1(x) + \mu = \Phi^{-1}(x) - x + \mu,$$

which is translation constant μ *plus the CDT of zero-mean normal distribution. Figure 2.2 is plotted for case when* $\mu = 2$*.*

Property 2.3.4. ***Scaling.*** *Let* I_a *represent a scaling of the probability density* I_1 *by* a*,* $I_a(x) = aI_1(ax)$*. The CDT of* I_a *respect to the reference probability density* $I_0 : X \to$



(a) $I_\sigma(x)$ **(b)** $\widehat{I}_\sigma(x)$

Figure 2.3: Example 2.3.5

$\mathbb{R}$ *is given by* $\widehat{I}_a : X \to \mathbb{R}$*:*

$$\widehat{I}_a(x) = \frac{\widehat{I}_1(x) - x(a-1)\sqrt{I_0(x)}}{a}. \tag{2.11}$$

For a proof, see A.2.

Example 2.3.5. *Consider the density function* $I_1(x)$ *in Example 2.2.1 scaled with a factor* a*, such as*

$$I_a(x) = aI_1(ax) = \frac{a}{\sqrt{2\pi}} e^{-\frac{(ax)^2}{2}}.$$

This is identical to a normal distribution with zero-mean and a standard deviation $\frac{1}{a}$*. The corresponding CDT,* $\widehat{I}_a : [0,1] \to \mathbb{R}$*, for* I_a *with respect to the uniform reference density* $I_0 : [0,1] \to \mathbb{R}$ *can be found by the scaling property* (2.11) *and by the CDT found in* (2.7)*:*

$$\begin{aligned}\widehat{I}_a(x) &= \frac{\widehat{I}_1(x) - x(a-1)}{a} \\ &= \frac{\Phi^{-1}(x) - x - ax + x}{a} \\ &= \frac{\Phi^{-1}(x)}{a} - x.\end{aligned}$$

Figure 2.3 plots this function for the case when $a = 2$*.*

Property 2.3.6. ***Composition.*** *Let* $I_g : Z \to \mathbb{R}$ *represent a probability density that has*

the following relation with the probability density $I_1 : Y \to \mathbb{R}$

$$J_g(x) = J_1(g(x)).$$

$J_1 : Y \to \mathbb{R}$ and $J_g : Z \to \mathbb{R}$ represent the corresponding cumulative distribution for I_1 and I_g respectively. $g : Z \to Y$ is an invertible, differentiable function. The CDT of the corresponding density I_g with respect to the reference probability density $I_0 : X \to \mathbb{R}$ is given by

$$\widehat{I}_g(x) = \left(g^{-1} \left(\frac{\widehat{I}_1(x)}{\sqrt{I_0(x)}} + x \right) - x \right) \sqrt{I_0(x)}.$$

See A.3 for a proof. Property 2.3.6 summarizes one of the main characteristics of the CDT transform so far, as rendering diffeomorphic transport changes 'Eulerian' in the CDT space. In detail, in CDT space, the changes in $\widehat{I}_g$ at coordinate x_0 is only affected by the change of the same coordinate x_0, i.e. $\widehat{I}_1(x_0)$. On the other hand, in L^2 space, the changes in I_g at coordinate x_0 is affected by the changes in both coordinates x_0 and $g(x_0)$, i.e. $I_g(x_0) = g'(x_0)I_1(g(x_0))$.

2.4 Linear Separability in the CDT space

One of the main contributions of this paper is to describe how the CDT transformation can enhance linear separability of signal classes. Before stating the main result regarding linear separation, a few preliminary results are necessary. As is well-known, the linear separability of two sets in $\mathbb{R}^n$ is determined by the existence of a separating hyperplane. If two sets are convex and disjoint, a separating hyperplane always exists, and hence the sets are linearly separable. Furthermore, the converse holds when at least one set is an open set [44]. The Hahn-Banach Separation Theorem is a generalization of the separating hyperplane theorem for infinite dimensional spaces.

Theorem 2.4.1 (Hahn-Banach Separation Theorem for Normed Vector Spaces)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be nonempty, convex subsets of a real normed vector space V. Furthermore, assume $\mathbb{P}$ and $\mathbb{Q}$ are disjoint and that one is closed and the other is compact. Then,*

there exists a continuous linear functional T on V and $b \in \mathbb{R}$ that strictly separates set $\mathbb{P}$ and $\mathbb{Q}$ such that

$$T(p) < b < T(q), \qquad \forall p \in \mathbb{P}, \forall q \in \mathbb{Q}. \tag{2.12}$$

For a non-zero linear functional T and a real number b, a hyperplane $\mathcal{H}(T, b) = \{v \in V | T(v) = b\}$ can be defined, and a hyperplane that satisfies (2.12) is called a separating hyperplane. For a proof and more details on the Hahn-Banach separation theorem, please refer to [45, 46]. For L^2 spaces, the Hahn-Banach Separation Theorem implies that there exists a unique *linear* classifier w that *linearly* separates two convex sets. To derive this, we need the following theorem, which states that every linear functional T on L^2 is of the form (2.13) for some $w \in L^2$.

Theorem 2.4.2. *For every continuous linear functional T on L^2 there is a unique $w \in L_2$ so that*

$$T(f) = \int_X f(x)w(x)dx, \qquad \forall f \in L^2. \tag{2.13}$$

In other words, there exists a separating hyperplane in L^2 space, $\mathcal{H}(w, b) = \{x \in X | w(x) = b\}$. For a proof and more details, please refer to [47]. Therefore, for a continuous linear functional T on L^2, a unique w can always be found. The following Lemma is a consequence of Theorem 2.4.1 and Theorem 2.4.2 that state there exists a *linear* classifier w that can separate two disjoint, convex sets in L^2 space.

Lemma 2.4.3 (Linear Classifier for Convex Sets in L^2 Space)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be nonempty, convex subsets of L^2 space, where $\mathbb{P}$ and $\mathbb{Q}$ are disjoint and that one is closed and the other is compact. Then, there exists a continuous hyperplane $\mathcal{H}(w, b) = \{x \in X | w(x) = b\}$ that separates set $\mathbb{P}$ and $\mathbb{Q}$ such that*

$$\begin{aligned} \int_X w(x)p_i(x)dx &< b, \qquad \forall p_i \in \mathbb{P} \\ \int_X w(x)q_j(x)dx &> b, \qquad \forall q_j \in \mathbb{Q}, \end{aligned} \tag{2.14}$$

and $\mathcal{H}(w, b)$ is called a linear classifier.

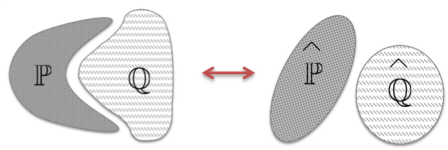


Figure 2.4: Depiction for linear separability properties of the CDT.

So far, we have seen that a linear classifier always exists for two disjoint, convex sets in L^2 with one being compact and the other closed. Moreover, the linear classifier would also linearly separate any subset pair from each convex hull of each convex set. In other words, two linearly separable convex sets imply that any subset pair from each convex hull is linearly separable, and vice versa. Therefore, in order to determine whether or not two sets are linearly separable, it suffices to show whether any subset pair from each convex hull is linearly separable. The following Lemma states this argument and will be used to show the main result of the paper.

Lemma 2.4.4. *[Linear Separation of Compact Convex Hulls of Convex Sets in L^2 Space] Two nonempty, compact subsets $\mathbb{P}$ and $\mathbb{Q}$ in L^2 space are linearly separable if and only if both their convex hulls are disjoint, i.e. when the following equation holds:*

$$\sum_{i=1}^{N_p} \alpha_i p_i \neq \sum_{j=1}^{N_q} \beta_j q_j, \tag{2.15}$$

for any subset $\{p_i\}_{i=1}^{N_p} \subset \mathbb{P}$ and $\{q_j\}_{j=1}^{N_q} \subset \mathbb{Q}$, and for any $\alpha_i, \beta_j > 0$ that satisfies $\sum_i \alpha_i = \sum_j \beta_j = 1$.

For proof, see A.4.

We now discuss the conditions under which the CDT can render classes of 1-dimensional probability densities linearly separable. We begin by defining a generative model for classes $\mathbb{P}$ and $\mathbb{Q}$.

Definition 2.4.5. *$\mathbb{H}$ is a set of monotonic and differentiable functions. $\mathbb{P}$ and $\mathbb{Q}$ are two disjoint sets satisfying*

i) $h'(p_0 \circ h) \in \mathbb{P}, \;\; h'(q_0 \circ h) \in \mathbb{Q}, \qquad \forall h \in \mathbb{H}, \;\; p_0 \in \mathbb{P}, \;\; q_0 \in \mathbb{Q}$

ii) $\forall p \in \mathbb{P}, \forall q \in \mathbb{Q}, \;\; p \neq q$ *(disjoint).*

Note that in the definition above we have used the notation $p \circ h(x) = p(h(x))$. The definition provides a framework which one can use to construct (or interpret) signal classes. In more practical language, we envision signal classes as being generated from fundamental patterns, but with distortions or confounds applied to them. For example, let p_0 and q_0 be two distinct probability densities, which we denote as 'mother' densities. Furthermore, let $\mathbb{H}$ be composed of all translations: $h_\tau(x) = x - \tau$, with τ a random variable. Elements of the sets $\mathbb{P}$ and $\mathbb{Q}$ are thus $p_0 \circ h_\tau$, and $q_0 \circ h_\tau$, respectively, and can be viewed as translations of the original mother densities. In this case, the translation makes up the 'nuisance' (confound) parameter a classifier must decode to enable accurate separation of the classes. Note that we have used the translation case as an example here, and the model specified above allows for more complex classes to be created. We note that since $h \in \mathbb{H}$ is monotonic and differentiable, its inverse h^{-1} exists and is also differentiable.

We now describe the main Theorem of this paper clarifying the linear separation properties of the newly proposed CDT.

Theorem 2.4.6. ***Linear Separability Theorem in CDT Space*** *Let $\mathbb{P}, \mathbb{Q}, \mathbb{H}$ follow be defined according to Definition 2.4.5. In addition, let $h \in \mathbb{H}$ satisfy the following conditions:*

i) $\forall h \in \mathbb{H}, \quad h^{-1} \in \mathbb{H}$.

ii) $\forall h \in \mathbb{H}$ *and* $\alpha_i > 0$ *that satisfies* $\sum_i \alpha_i = 1$, $h_\alpha^{-1} = \sum_i \alpha_i h_i^{-1} \in \mathbb{H}$.

iii) $\forall h_1, h_2 \in \mathbb{H}, \quad h_1 \circ h_2 \in \mathbb{H}$.

Then the corresponding sets in the CDT space $\widehat{\mathbb{P}}, \widehat{\mathbb{Q}}$ are linearly separable.

We note that the linear separability theorem is independent of the choice of the reference I_0. For a proof, see A.5.

2.5 Numerical implementation

We now describe a numerical method for approximating the CDT given discrete data. Recall that the CDT is defined for continuous-time functions in contiguous, finite do-

main. In order to compute the CDT for a discrete-time signal, we need a way of estimating its cumulative function at any arbitrary coordinate. We do so via interpolation. Given a discrete signal of N points and an interpolating model, the forward CDT can be estimated numerically at all N points. Our numerical method is designed when the reference function is $I_0(x) = 1$ for $x \in [0, 1]$ (recall the linear separation properties of the CDT are independent of the choice of reference). The computation is formulated with the aid of B-splines [48]. We use the B-spline of degree zero which guarantees that the reconstructed signals are always positive, which yields a low complexity algorithm ($O(N)$). We note that under the specific construction below the approximated density functions will be discontinuous at the half way point between sampled nodes, and, as stated above, reconstruction at these points is not possible.

Let $\pi(x)$ be the B-spline of degree zero of width r

$$\pi(x) = \begin{cases} 1 & x \in [-\frac{1}{2}r, \frac{1}{2}r] \\ 0 & \text{elsewhere} \end{cases}$$

and define $\Pi(x) = \int_{-\infty}^{x} \pi(\tau) d\tau$ as

$$\Pi(x) = \begin{cases} 0 & x < -\frac{1}{2}r \\ x + \frac{1}{2}r & x \in [-\frac{1}{2}r, \frac{1}{2}r] \\ r & x > \frac{1}{2}r. \end{cases} \tag{2.16}$$

Let's denote a N-point discrete-time signal as $\mathbf{c} = [c_1, \cdots, c_N]$ and x_i as the ith sample location of $\mathbf{c}$, i.e. $\mathbf{c}(x_i) = c_i$, $\forall i = 1, \cdots, N$. We interpolate the discrete-time signal $\mathbf{c}$ with the B-spline of degree zero to be a continuous-time signal such as $I_1(x) = \sum_{i=1}^{N} c_i \pi(x - x_i)$ for $x \in [x_1 - \frac{1}{2}r, x_N + \frac{1}{2}r]$. Rewriting (1.6), we have

$$\int_{x_1 - \frac{1}{2}r}^{f_1(x)} I_1(\tau) d\tau = \int_{x_1 - \frac{1}{2}r}^{f_1(x)} \sum_{i=1}^{N} c_i \pi(\tau - x_i) d\tau = x \tag{2.17}$$

which can be simplified further by interchanging the sum and the integral, and then

using Π to denote the cumulative integral function of π as

$$\sum_{i=1}^{N} c_i \Pi(f_1(x) - x_i) = x. \tag{2.18}$$

By substituting (2.16) into (2.18) and taking the inverse of $\Pi(x)$ which is piecewise linear, $f_1(x)$ is computed according to the following algorithm:

1. When $0 < x < rc_1$, we have $c_1 \Pi(f_1(x) - x_1) = x$. Thus,

$$f_1(x) = \frac{x}{c_1} + x_1 - \frac{1}{2}r$$

2. When $-rc_n + \sum_{i=1}^{n-1} c_i < x < rc_n + \sum_{i=1}^{n-1} c_i$, we have $\sum_{i=1}^{n} c_i \Pi(f_1(x) - x_i) = x$. Thus,

$$f_1(x) = \frac{x - \sum_{i=1}^{n-1} c_i}{c_n} + x_n - \frac{1}{2}r. \tag{2.19}$$

3. Proceed until $n = N$.

Chapter 3

The Radon-Cumulative Distribution Transform

3.1 Introduction

Intensity vector flows represent an interesting alternative for encoding the pixel intensity movements which help simplify certain pattern recognition tasks. Past works have been done [34, 36] to describe a framework that makes use of the L^2 optimal transport metric (Earth Mover's distance) to define a new invertible image transform. The transport-based approach, however, depends on obtaining a unique transport map that encodes an image via minimization of a transport metric, which is relatively cumbersome and slow for large images.

In this chapter, we describe a new 2D image transform by combining the standard 2D Radon transform of an image with the 1D Cumulative Distribution Transform (CDT) proposed earlier. As the CDT, the transform for 2D, namely Radon-CDT, utilizes a reference (or template), and it can be computed with a (nonlinear) closed form formula without the need for a numerical minimization method. Also, The Radon-CDT shares similar properties as the CDT including enhancements in linear separation, and therefore can be used to improve the linear separability of image classes.

The Radon-CDT is a nonlinear and invertible image transform that enables any statistical analysis in the transform space to be directly inverted to the image space. In other words, given that the transform is invertible, the approach enables visualization of any regression applied in transform space. It thus enables one to visualize variations in texture and shapes, as well as to visualize discriminant information by inverting classifiers.

In what follows, we start by introducing the Radon cumulative transform. Its properties are enumerated in Section 3.3, and the linear separation theorem is presented in Section 3.4. The details of the numerical implementation of the method is presented in Section 3.5.

3.2 The Radon-Cumulative Distribution Transform

3.2.1 The Radon transform

The Radon transform of a function $I : \mathbb{R}^2 \to \mathbb{R}^+$, which we denote by $R = \mathscr{R}(I)$, is defined as:

$$R(t, \theta) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x, y)\delta(t - x\cos\theta - y\sin\theta)dxdy$$

where $\delta(t)$ is a Dirac delta function, $x\cos\theta + y\sin\theta = t$ defines a line, t is the perpendicular distance from the line to the origin, and θ is the angle between the line and the y-axis as shown in Fig. 3.1.

The inverse Radon transform, $I = \mathscr{R}^{-1}(R)$, is defined with the aid of the Fourier Slice Theorem as [49, 50]:

$$I(x, y) = \int_0^{\pi} R^*(x\cos\theta + y\sin\theta, \theta)d\theta,$$

where $R^*(t, \theta) = w(t) * R(t, \theta)$ is the one-dimensional convolution with respect to variable t, $w(t) = \mathscr{F}^{-1}(|\omega|)$ is the ramp filter, $\mathscr{F}^{-1}$ is the inverse Fourier transform.

The total integral of the function $I(x, y)$ along x and y is equivalent to the line

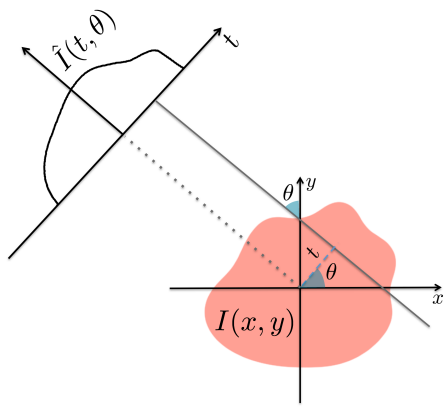


Figure 3.1: Geometry of the line integral associated with the Radon transform

integral of $R(t,\theta)$ along t:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} I(x,y)dxdy = \int_{-\infty}^{\infty} R(t,\theta)dt, \qquad \forall\theta \in [0,\pi]. \tag{3.1}$$

Here we combine the CDT [51] and the Radon transform to describe the Radon Cumulative Distribution Transform (Radon-CDT). We then derive a few properties of the Radon-CDT, and extend the CDT results [51] on linear separability of classes of one-dimensional signals [51] to classes of images.

3.2.2 The Radon-CDT transform

Consider an image $I : \mathbb{R}^2 \to \mathbb{R}^+$ and a reference image $I_0 : \mathbb{R}^2 \to \mathbb{R}^+$, where both images are normalized such that

$$\int_{\mathbb{R}}\int_{\mathbb{R}} I(x,y)dxdy = \int_{\mathbb{R}}\int_{\mathbb{R}} I_0(x)dxdy = 1.$$

The forward Radon-CDT for I is defined as,

$$\widehat{I}(t,\theta) = (f(t,\theta) - \mathbb{I}(t))\sqrt{R_0(t,\theta)}, \tag{3.2}$$

where

- R_0 and R are corresponding Radon transforms for I and I_0,
- For each θ, the monotonic differentiable map $f(t,\theta)$ is found that tranports the one-dimensional signal $R_0(t,\theta)$ to $R(t,\theta)$, i.e.

$$\int_0^{f_\theta(t)} R_\theta(\tau)d\tau = \int_0^t R_{0,\theta}(\tau)d\tau.$$

  Here, we used $f(t,\theta) = f_\theta(t)$, $R(t,\theta) = R_\theta(t)$, and $R_0(t,\theta) = R_{0,\theta}(t)$ in abuse of notation to explicitly state that the θ is fixed.
- $\mathbb{I}$ represents the identity map, i.e. $f^{-1}(f(t,\theta),\theta) = \mathbb{I}$, or equivalently $\mathbb{I}(t) = t$.

The inverse Radon-CDT is defined as,

$$I = \mathscr{R}^{-1}(det(D\mathbf{g})(R_0 \circ \mathbf{g})) \tag{3.3}$$

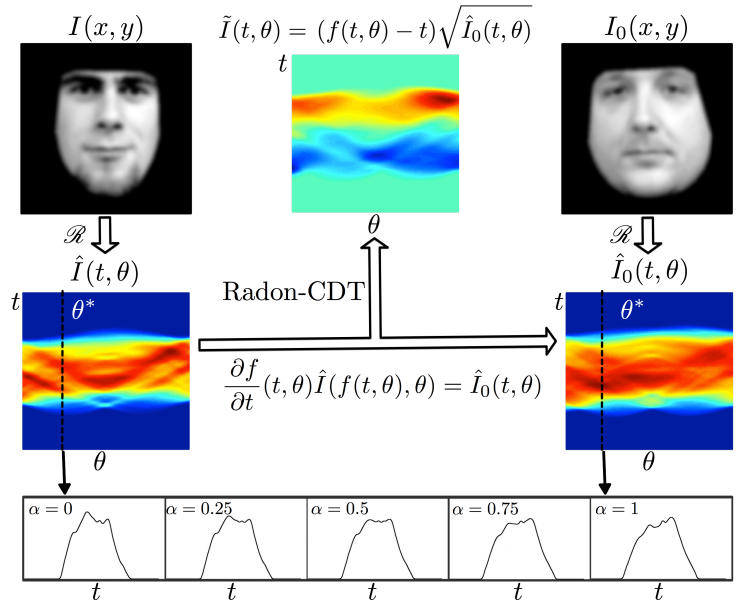
where

- $\mathbf{g}(t,\theta) = [f^{-1}(t,\theta),\theta]^T$,
- $D\mathbf{g}$ is the Jacobian of $\mathbf{g}$,
- $det(D\mathbf{g}(t,\theta)) = \frac{\partial f^{-1}(t,\theta)}{\partial t}$.

Fig. 3.2 visualizes the steps for computing Radon-CDT of a sample image I with respect to a reference image I_0. The radon transforms for images are first computed, denoted as R and R_0. Then for each $\theta = \theta^*$ where θ^* is an arbitrary projection angle, the one-dimensional transport map from $R_0(t,\theta^*)$ to $R(t,\theta^*)$ is computed. Finally, the Radon-CDT is obtained from f and R_0.

3.3 Radon-CDT properties

Here we describe a few basic properties of the Radon-CDT, with the main purpose of elucidating certain of its qualities necessary for understanding its ability to linearly separate certain types of two-dimensional densities.



$$\hat{I}_\alpha(t,\theta^*) = \frac{\partial f_\alpha}{\partial t}(t,\theta^*)\hat{I}(f_\alpha(t,\theta^*),\theta^*)$$
$$f_\alpha(t,\theta^*) = (1-\alpha)t + \alpha f(t,\theta^*), \quad \alpha \in [0,1]$$
$$\hat{I}_{\alpha=0}(t,\theta^*) = \hat{I}(t,\theta^*), \quad \hat{I}_{\alpha=1}(t,\theta^*) = \hat{I}_0(t,\theta^*)$$

Figure 3.2: The process of calculating the Radon-CDT transform of image I with respect to the reference image I_0.

Property 3.3.1. Translation. *Let $J(x,y) = I(x-x_0, y-y_0)$ and let $\widehat{I}$ be the Radon-CDT of I. The Radon-CDT of J with respect to a reference image I_0 is given by,*

$$\widehat{J}(t,\theta) = \widehat{I}(t,\theta) + (x_0\cos(\theta) + y_0\sin(\theta))\sqrt{R_0(t,\theta)},$$
$$t \in \mathbb{R} \text{ and } \theta \in [0,\pi]. \tag{3.4}$$

See A.6 for a proof. Similar to the CDT example in Eq. 2.10, it can be seen that while $I(x-x_0, y-y_0)$ is nonlinear with respect to $[x_0, y_0]$ the presentation of the image in the Radon-CDT, $\widehat{I}(t,\theta) + (x_0\cos(\theta) + y_0\sin(\theta))\sqrt{R_0(t,\theta)}$ is linear.

Property 3.3.2. Scaling. *Let $J(x,y) = \alpha^2 I(\alpha x, \alpha y)$ with $\alpha > 0$ and let $\widehat{I}$ be the Radon-CDT of I. The Radon-CDT of J with respect to a reference image I_0 is given*

by,

$$\widehat{J}(t,\theta) = \frac{\widehat{I}(t,\theta)}{\alpha} + \left(\frac{1-\alpha}{\alpha}\right)\sqrt{R_0(t,\theta)},$$
$$t \in \mathbb{R} \text{ and } \theta \in [0,\pi]. \tag{3.5}$$

See A.7 for a proof. Similar to the translation property, it can be seen that while $\alpha^2 I(\alpha x, \alpha y)$ is nonlinear with respect to α the corresponding presentation in the Radon-CDT space, $\frac{\widehat{I}(t,\theta)}{\alpha} + \left(\frac{1-\alpha}{\alpha}\right)\sqrt{R_0(t,\theta)}$, is linear in $\frac{1}{\alpha}$.

Property 3.3.3. Rotation. *Let $J(x,y) = I(x\cos(\phi)+y\sin(\phi), -x\sin(\phi)+y\cos(\theta))$ and let $\widehat{I}$ be the Radon-CDT of I. For a circularly symmetric reference image I_0, the Radon-CDT of J is given by,*

$$\widehat{J}(t,\theta) = \widehat{I}(t,\theta-\phi), \;\; t \in \mathbb{R} \text{ and } \theta \in [0,\pi] \tag{3.6}$$

See A.8 for a proof. Note that unlike translation and scaling, for rotation the transformed image remains nonlinear with respect to ϕ.

3.3.1 The Radon-CDT Representation

Here we show some implications of these properties in image modeling. Let I_0 be an arbitrary image and let $I(x,y) = I_0(x-x_0, y-y_0)$ be a translated version of I_0. A natural interpolation between these images follows from $I_\alpha(x,y) = I_0(x-\alpha x_0, y-\alpha y_0)$ where $\alpha \in [0,1]$. The linear interpolation between these images in the image space, however, is equal to,

$$\begin{aligned} I_\alpha(x,y) &= \alpha I_1(x,y) + (1-\alpha) I_0(x,y) \\ &\neq I_0(x-\alpha x_0, y-\alpha y_0) \end{aligned} \tag{3.7}$$

In fact, above equation is also true for any linear image transform. Take the Radon transform for example, where the linear interpolation in the transform space is equal

to,

$$\begin{aligned}
\hat{I}_\alpha(t,\theta) &= \alpha\mathscr{R}(I_1(x,y)) + (1-\alpha)\mathscr{R}(I_0(x,y)) \\
&= \mathscr{R}(\alpha I_1(x,y)) + (1-\alpha)I_0(x,y)) \Rightarrow \\
I_\alpha(x,y) &= \mathscr{R}^{-1}(\hat{I}_\alpha(t,\theta)) = \alpha I_1(x,y) + (1-\alpha)I_0(x,y) \\
&\neq I_0(x-\alpha x_0, y-\alpha y_0). \qquad (3.8)
\end{aligned}$$

On the other hand, due to its nonlinear nature, this scenario is completely different in the Radon-CDT space. Let I_0 be the template image for the Radon-CDT (the following argument holds even if the template is chosen to be different from I_0), then from Equation (3.4) we have $\tilde{I}_0(t,\theta) = 0$ and $\tilde{I}_1(t,\theta) = (x_0 cos(\theta) + y_0 sin(\theta))\sqrt{\hat{I}_0(t,\theta)}$. The linear interpolation in the Radon-CDT space is then equal to,

$$\begin{aligned}
\tilde{I}_\alpha(t,\theta) &= \alpha\tilde{I}_1(t,\theta) + (1-\alpha)\tilde{I}_0(t,\theta) \\
&= \alpha(x_0 cos(\theta) + y_0 sin(\theta))\sqrt{\hat{I}_0(t,\theta)} \Rightarrow \\
I_\alpha(t,\theta) &= I_0(x-\alpha x_0, y-\alpha y_0). \qquad (3.9)
\end{aligned}$$

which is the natural interpolation between these images and captures the underlaying translation. Figure 3.3 summarizes the equations presented above and provides a visualization of this effect.

3.4 Linear separability in the Radon-CDT space

Here, as in the CDT case, we describe that the linear separability theorem holds for the Radon-CDT. With the same framework of constructing image classes based on the 'mother images', we show that the sets of images corrupted by confounds $\mathcal{H}$ can be effectively linearly separated in the Radon-CDT space.

Let $\mathbb{P}$ and $\mathbb{Q}$ be sets of normalized images born from two mother images p_0 and q_0

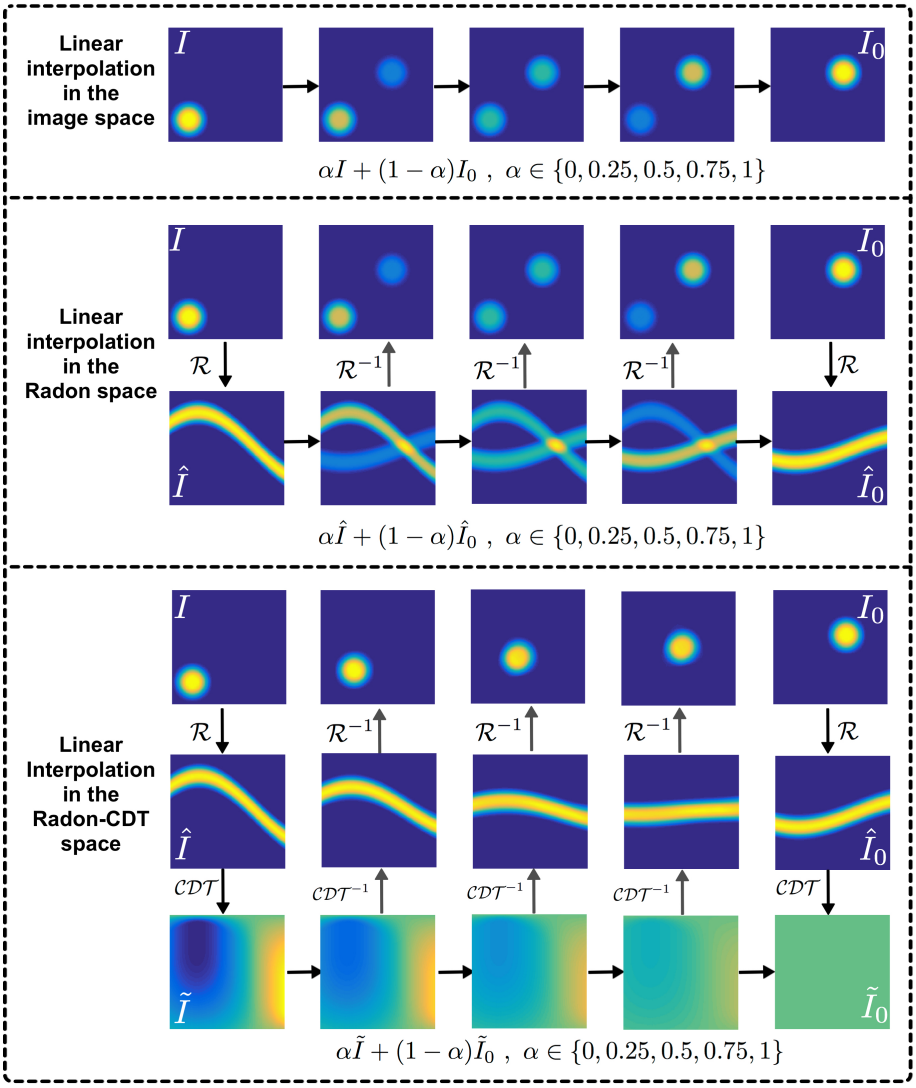


Figure 3.3: A simple linear interpolation between two images in the image space, the Radon transform space (which is a linear transform), and the Radon-CDT space.

as follows,

$$\mathbb{P} = \{p | p = \mathscr{R}^{-1}(det(D\mathbf{h})(\tilde{p}_0 \circ \mathbf{h})) \tag{3.10}$$

$$\mathbb{Q} = \{q | q = \mathscr{R}^{-1}(det(D\mathbf{h})(\tilde{q}_0 \circ \mathbf{h})) \tag{3.11}$$

where

- $\tilde{p}_0$, $\tilde{q}_0$ corresponds to the Radon transform of p_0, q_0,
- $\mathcal{H}$ is a set of measurable maps, with $\mathbf{h} \in \mathcal{H}$,
- $\mathbf{h}(t, \theta) = [h(t, \theta), \theta]^T, \forall h \in \mathcal{H}, \forall \theta \in [0, \pi]$.

It is important to note that h must be absolutely continuous in t and θ, so that $det(D\mathbf{h})(\tilde{p}_0 \circ \mathbf{h})$ and $det(D\mathbf{h})(\tilde{q}_0 \circ \mathbf{h})$ remain in the range of the Radon transform [52]. Now, we can state the linear separation theorem.

Theorem 3.4.1. *Linear Separation Theorem in $\mathbb{R}^2$ Under the signal generative model described in* (3.11)*, the sets $\mathbb{P}$ and $\mathbb{Q}$ become linearly separable in the transform space if $\mathcal{H}$ satisfies the following conditions,*

i) $h \in \mathcal{H} \Rightarrow h^{-1} \in \mathcal{H}$

ii) $h_1, h_2 \in \mathcal{H} \Rightarrow \alpha h_1 + (1 - \alpha) h_2 \in \mathcal{H},\ \forall \alpha \in [0, 1]$

iii) $h_1, h_2 \in \mathcal{H} \Rightarrow h_1(h_2), h_2(h_1) \in \mathcal{H}$

iv) $det(D\mathbf{h})(\tilde{p}_0 \circ \mathbf{h}) \neq \tilde{q}_0,\ \mathbf{h}(t, \theta) = [h(t, \theta), \theta]^T,\ \forall h_\theta \in \mathcal{H}$

The theorem holds regardless of the choice of the reference image I_0. See A.9 for a proof.

3.5 Numerical Implementation

3.5.1 The Radon transform

A large body of work on numerical implementation of the Radon transform exists in the literature [53]. Here, we use a simple numerical integration approach that utilizes

nearest neighbor interpolation of the given images, and summation.

3.5.2 Measure preserving map

Here, we follow the similar computation algorithm as in Sec. 2.5. The measure preserving map that warps $\widehat{I}(.,\theta)$ into $\widehat{I}_0(.,\theta)$ is found with the aid of the B-spline π of degree zero of width r,

$$\pi(x) = \begin{cases} 1/r & |x| \leq \frac{1}{2}r \\ 0 & |x| > \frac{1}{2}r \end{cases}. \tag{3.12}$$

and define Π as,

$$\Pi(x) = \begin{cases} 0 & & x < -\frac{1}{2}r \\ \frac{x}{r} + \frac{1}{2} & -\frac{1}{2}r \leq & x \leq \frac{1}{2}r \\ 1 & & x > \frac{1}{2}r \end{cases} \tag{3.13}$$

Using the B-spline of degree zero, we approximate the continuous sinograms, $\widehat{I}(.,\theta)$ and $\widehat{I}_0(.,\theta)$, with their corresponding discrete counterparts c and c_0 as follows,

$$\widehat{I}(t,\theta) \approx \sum_{k=1}^{K} c[k]\pi(t - t_k) \tag{3.14}$$

$$\widehat{I}_0(t,\theta) \approx \sum_{k=1}^{K} c_0[k]\pi(t - t_k). \tag{3.15}$$

Now the goal is to find $f(.,\theta)$ such that,

$$\int_{-\infty}^{f(t,\theta)} \widehat{I}(\tau,\theta)d\tau = \int_{-\infty}^{t} \widehat{I}_0(\tau,\theta)d\tau$$

which is equivalent to,

$$\sum_{k=1}^{K} c[k]\Pi(f(t,\theta) - t_k) = \sum_{k=1}^{K} c_0[k]\Pi(t - t_k)$$

let $\boldsymbol{\rho} = [0, \frac{1}{L}, \frac{2}{L}, ..., \frac{L-1}{L}, 1]^T$ for $L > 1$, and define $\boldsymbol{\tau}_0$ and $\boldsymbol{\tau}$ such that,

$$\sum_{k=1}^{K} c[k]\Pi(\boldsymbol{\tau}[l] - t_k) = \boldsymbol{\rho}[l]$$
$$\sum_{k=1}^{K} c_0[k]\Pi(\boldsymbol{\tau}_0[l] - t_k) = \boldsymbol{\rho}[l]$$

for $l = 1, ..., L + 1$, where $\boldsymbol{\tau}$ and $\boldsymbol{\tau}_0$ are found using the algorithm defined in Sec. 2.5. From the equation above we have that $f(\boldsymbol{\tau}_0[l], \theta) = \boldsymbol{\tau}[l]$. Finally we interpolate f to obtain its values on the regular grid, t_k for $k = 1, ..., K$.

3.5.3 Computational complexity

The computational complexity of the Radon transform of $N \times N$ images at M projection angles is $O(N^2M)$, and the computational cost for finding the mass preserving map, $f(t, \theta)$, from a pair of sinograms is $O(MN \log(N))$, hence, the overall computational cost of the Radon-CDT is dominated by the computational complexity of the Radon transform, $O(N^2M)$.

Chapter 4

Applications to the Linear Classification

In the second part of the thesis, we experimentally evaluate the properties of the CDT and Radon-CDT by comparing linear classification performed in the transform space with that in original signal space (L^2). We present computational examples that show the CDT and Radon-CDT can significantly increase linear classification accuracy compared to simply treating signals in ℓ^2 space.

The chpater demonstrates linear separability in the CDT space for signals with 1 independent variables. We investigate five cases of signal classification: classification of texture images from histograms, classification of accelerometer signals, classification of flow Cytometry data, classification of histograms from hand gesture image data, and classification of cell images from orientation histograms.

The second chapter demonstrates linear separability in the Radon-CDT space, we investigate five cases of image classifications on real images – classification of face images, classification of liver nuclei images, classification of animal face images, classification of handwriting digit images, and classification or shape images – as well as a synthetic example.

Note that the goal is not to propose the ultimate, or optimal, classification method

for each application, but rather to experimentally validate Theorem 2.4.6 and Theorem 3.4.1 using both simulated (manufactured) data and diverse, real datasets. We note that, with the exception of the simulated cases in Section 1.2 and in Section 4.2.1, we have no precise knowledge of whether conditions $i)$, $ii)$, and $iii)$ for $\mathbb{H}$ specified in the Theorem hold. Results seem to confirm, however, that the generative model specified in these conditions has a least some bearing on each problem investigated here.

4.1 Linear Classification in the CDT space

In this chapter, we demonstrate the application of the CDT in pattern recognition by demonstrating its capability to linearly separate the data. We quantify the degree of linear separability of the data by computing classification error using linear classifiers. As the CDT does not actually prescribe an optimal classifier, we compute three different linear classifiers using a standard cross validation procedure (or leave-one-out cross validation when data size is small) that separates training and testing data. In addition, we provide qualitative (visual) evidence, by computing a low dimensional projection of the data using training data only, that the CDT indeed tends to make data more linearly separable.

4.1.1 Experimental procedure

Average classification error is compared using three different linear classifiers: Fisher's linear discriminant analysis (Fisher LDA) method [31], the penalized LDA (pLDA) method of Wang et al [54], and the linear Support Vector Machine (SVM) method [55]. All experiments were performed using the MATLAB [56] programming language, while the SVM method was implemented using the LIBSVM package [57]. While the Fisher LDA method does not require parameter tuning, the linear SVM and pLDA methods require parameter tuning steps which were performed using 2nd depth cross validation utilizing the training set only. In the SVM method, the parameter is set to reflect how much error the separating hyperplane is to tolerate, while the parameter in the pLDA method determine the regularization to be applied when computing the

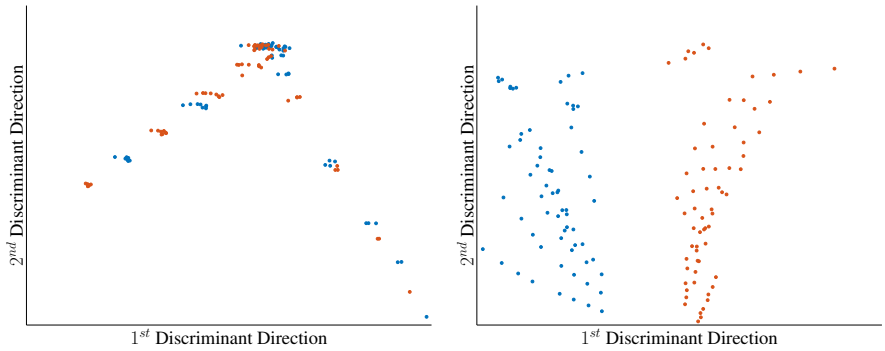
covariance matrix (refer to references [55, 54] for more details).

The low dimensional visualization plots were computed using the pLDA method, which in contrast to the standard LDA method can yield multi dimensional embeddings for the given data. The dimensions of each embedding are weighted according to a optimization metric, which combines a data separation term (given by LDA) and a 'data fitting' term (given by the standard Principal Component Analysis cost function). For each experiment reported below, we utilize the pLDA method to visualize a 2-dimensional embedding of the *testing* data. In each case, a subset of the data was used to estimate the lower dimensional embedding. Remaining (testing) data was used to obtain the visualizations.

The computational experiments shown in subsections 4.1.2, 4.1.4, 4.1.5, 4.1.6 were computed using a five-fold cross validation strategy, with 80% of the data used for training, and 20% for testing. For experiment in subsection 4.1.3, due to small sample size, a leave-one-out cross validation is used instead. The experimental procedure is summarized in Algorithm 1. For more details on cross validation experimental procedures, refer to references [29, 28].

Algorithm 1: 5-fold cross validation

```
Partition the dataset into 5 groups. Leave one fold out for testing and use the
remaining fold for training.
foreach training set do
    1. For SVM and PLDA, partition the training set into 5 groups again. Leave
    one fold out for validation and use remaining fold for training. For LDA,
    skip to step 2.
    foreach training set (parameter sweep) do
        1. Learn the classifier for different parameter values.
        2. Compute the validation error.
    Return the best parameter of average validation error.
    2. Learn the classifier with the optimal parameter.
    3. Compute the testing error.
2. Compute the average classification error.
```



(a) PLDA projection in L^2 space **(b)** PLDA projection in CDT space

Figure 4.1: PLDA projection for texture dataset

4.1.2 Texture classification from intensity histograms

In this application, already discussed in the introduction as a motivating example, our goal is to utilize the CDT to distinguish between two types of texture images, under brightness and contrast variations, from their intensity histograms. Consider the textures displayed in the middle rows of Figure 1.1. Their corresponding histograms are shown directly under and above each image, with variations in brightness and contrast. Variations in brightness correspond to translations in the histograms, while variations in contrast correspond to scalings (dilations) of the histograms. We note that such variations (translation and scaling) satisfy the necessary properties described in Theorem 2.4.6. Our theory thus predicts that the histogram data would be perfectly separable in CDT space. For testing this hypothesis, we generated a set of 128 images (2 sets of 64 images) by applying 8 random variations in brightness, with the translations in the range of $[0, 0.5]$, and 8 random variations in contrast, with scalings in the range of $[0.6, 1.67]$. Results are shown in Table 1.1, and confirm that 1) the data is not linearly separable in histogram space and 2) becomes linearly separable in CDT space. The lower dimensional representation of the original data using Penalized LDA also confirms this (see Figure 4.1).

Table 4.1: Average classification error of the accelerometer dataset

Classifier type	Dataset	L^2 space	CDT space
Fisher LDA	Training set	0 %	0%
	Testing set	**50** %	**10**%
PLDA	Training set	0%	0 %
	Testing set	**60** %	**5** %
Linear SVM	Training set	8.75%	7.5 %
	Testing set	**55** %	**10** %

4.1.3 Activity Recognition with Accelerometer Data

An accelerometer is a device that records the acceleration of a moving object. Modern 'smartphones' are commonly equipped with a 3-axis accelerometer that keeps track of the acceleration in 3 different directions x, y, and z, and accelerometers have been widely adapted to various wearable devices (e.g. watches) for human activity recognition. In this example, we aim to detect (classify) two different activities given accelerometer data obtained from an iphone 5. Class 1 consists of a person swinging arms while holding the phone. Class 2 consists of a phone being dropped to the ground. Figure 4.2a shows the raw data recorded from the accelerometer for both cases. We note that in this case, the signals varied in length given the different duration of the episodes. Signals were zero-padded so that they match the length of the largest signal. 10 sample signals were acquired for each class. For each instance, the Energy = $x^2 + y^2 + z^2$ is computed from the tri-axis measurements (see Figure 4.2b). Here we compare the ability of the linear classification in original (energy) signal space versus in CDT space.

Results are shown in Table 4.1, and clearly indicates that the data becomes linearly separable in CDT space. The lower dimensional representation of the original data using Penalized LDA (PLDA) [54] indicates (see Figure 4.3) that each class forms convex hulls that are linearly separable in CDT space but not in energy signal space. For this example, both training and testing data are represented in the lower dimensional embedding in Figure 4.3. By seeing Figure 4.3, we can verify that the linear classifier computed using only training set correctly separates both training and testing set in CDT space, but not in energy signal space.

In this experiment, it is apparent that the signals varied in terms of intensity and

the location where the maximum peak has occurred, and this explains the inability of linear classifiers to perform well in original (energy) signal space. As explained above, the CDT is able to overcome such variations.

4.1.4 Flow Cytometry

Flow Cytometry is a technique used to analyze light emission properties of grouped cells using fluorescence markers. In this example, we utilize an existing database (the FlowRepository database [6]) to distinguish DNA histograms between normal subjects and donors diagnosed with acute myeloid leukemia (AML), obtained from peripheral blood or bone marrow aspirates. The data included 8 measurements per each subject, where fluorochrome signals were detected at the 620nm wavelength specifically. Sample data is shown in Figure 4.4a, where the x-axis represents each cell that passed through the flow cytometry sensor, and the y-axis correspond to the DNA intensity measurement of the cell at wavelength 620nm. The intensity histogram with 1024 intensity levels are computed and their corresponding CDTs (see Figure 4.4b and Figure 4.4c).

The average classification error is reported in Table 4.2. We note that the classification in the signal space using LDA (test accuracy of 84.99%) is worse than the line of chance (87.5%), given the uneven distribution of patient data. Comparison with the line of chance and the classification accuracy in histogram space using PLDA or SVM also suggests that linear classifiers trained are more or less equivalant to random classification. However, classifying data in CDT space suggests that linear separation is possible, and the Cohen's Kappa for this computation (0.3) confirms fair agreement

Table 4.2: Average Classification Error of the flow cytometry dataset

Classifier type	Dataset	L^2 space	CDT space
Fisher LDA	Training set	6.81 %	5.82%
	Testing set	**15.01** %	**11.31** %
PLDA	Training set	11.55 %	7.75%
	Testing set	**12.03** %	**9.15** %
Linear SVM	Training set	10.39 %	8.65%
	Testing set	**11.46** %	**8.88**%

[58].

4.1.5 Cambridge Hand Dataset

The Cambridge hand gesture dataset consists of 900 image sequences of 3 primitive hand shapes (see Figure 4.6a) where each image sequence consists of around 60 frames of 3 different motions [59]. In this example, we try to distinguish 3 different hand shapes; flat, spread, and v-shape. There are 2678 images for flat hands, 2992 images for spread hands, and 2764 images for v-shape hands, and each image was taken under arbitrary positioning and illumination. A preprocessing step computes the edge of each image (240 x 320 pixels large) and the corresponding indices of the edge pixels. Two histograms are created counting x coordinates and y coordinates of the edge pixels per image Figure 4.6b. Corresponding CDTs are computed for each x and y histogram Figure 4.6c. The classification is done with concatenation of two x and y histograms and concatenation of two x and y CDTs.

Table 4.3: Average Classification Error of the hand gestures dataset

Classifier type	Dataset	L^2 space	CDT space
Fisher LDA	Training set	13.92 %	4.58%
	Testing set	**16.11** %	**5.76**%
PLDA	Training set	38.02 %	6.73%
	Testing set	**38.21** %	**6.97**%
Linear SVM	Training set	13.77%	1.27%
	Testing set	**15.73** %	**1.65**%

Results are shown in Table 4.3, which clearly indicate that the data becomes more linearly separable in CDT space. As in previous examples, the two dimensional representation of the original testing data using Penalized LDA (PLDA) [54] indicates (see Figure 4.7) that classes form convex hulls that are linearly separable in CDT space and not in histogram space. Moreover, this example shows that the CDT can be applied to multi-class problems which would enhance the simplicity of the classification problem.

Table 4.4: Average Classification Error of the HeLa dataset

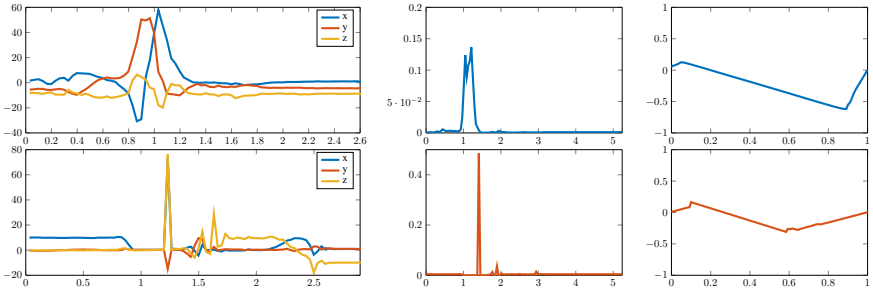
Classifier type	Dataset	L^2 space	CDT space
Fisher LDA	Training set	0.53 %	0.40%
	Testing set	**2.66** %	**2.59**%
PLDA	Training set	0.14 %	0.92%
	Testing set	**1.59** %	**1.07**%
Linear SVM	Training set	0 %	0.26%
	Testing set	**0.53** %	**1.05**%

4.1.6 Actin and Microtubules Classification

Our goal in this experiment is to quantify how well actin and microtubule filaments in HeLa cells [60] differ from one another in terms of their orientation distributions. Fluorescence microscope images of HeLa cells were grouped into two classes according to their protein structure: rhodamine-conjugated phalloidin, which labels F-actin and a monoclonal antibody against beta-tubulin (microtubules). Each image was preprocessed such that outside the cropped region was set to 0 and contrast-stretched to have full scale (see Figure 4.8a). In order to compute the orientation of each pixel, the images were filtered with 32 Gabor filters of size 9×9, and for each pixel, the filter with the maximum response is selected and labeled from 1 to 32 (see Figure 4.8b). A histogram of orientation filter responses are computed for each image (see Figure 4.8c) and then the CDT is computed for each histogram (see Figure 4.8d). In this example, both histogram and CDT show excellent classification accuracy, given that the difference between two protein structures are hard to be recognized by visual inspection. It is an instance where data is already well (linearly) separated in Euclidean space, and is also linearly separable in CDT space (i.e. the CDT did not destroy linear separation in this example).

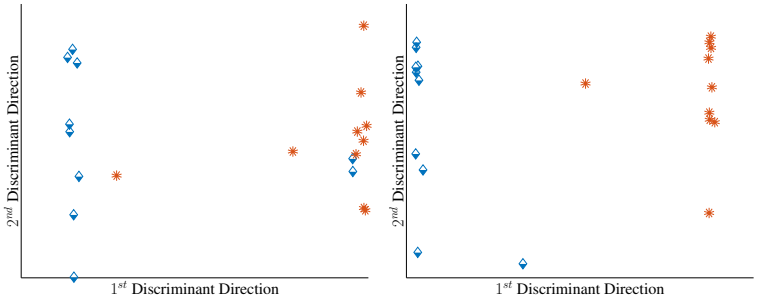
4.2 Linear Classification in the Radon CDT Space

In this chapter, we describe the application of Radon-CDT in pattern recognition by demonstrating its capability to simplify data structure on real image datasets. Our goal is to demonstrate that the data classes in the Radon-CDT space become more linearly



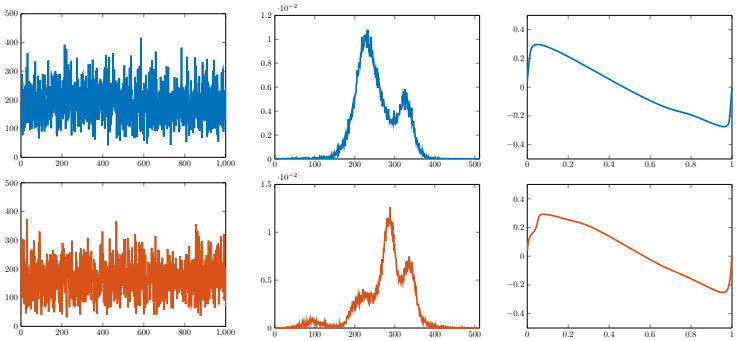
(a) Raw data **(b)** Energy **(c)** CDT

Figure 4.2: Two classes of accelerometer dataset, swinging (top row) vs free falling (bottom row)



(a) PLDA projection in L^2 space **(b)** PLDA projection in CDT space

Figure 4.3: PLDA projection for accelerometer dataset



(a) Raw data **(b)** Histogram **(c)** CDT

Figure 4.4: Two classes of flow cytometry data, AML (top row) vs. Normal (bottom row)

separable. This is done by utilizing linear classifiers (e.g. linear discriminant analysis and linear support vector machine (SVM) classifiers) in the image and the Radon-CDT space. Although the image classes do not exactly follow the class structures stated in subsection 3.4, linear classifiers in the Radon-CDT space consistently lead to higher classification accuracy compared to that in the image space.

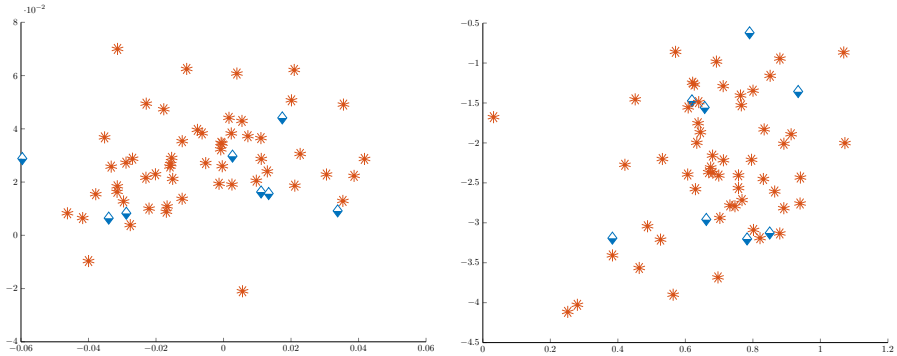
4.2.1 Synthetic example

Before delving into real examples, let us demonstrate the linear separability property of our proposed image transform with the simulated examples which follow the model provided in Theorem 3.4.1:

Two classes of images $\mathbb{P}$ and $\mathbb{Q}$ are generated as follows,

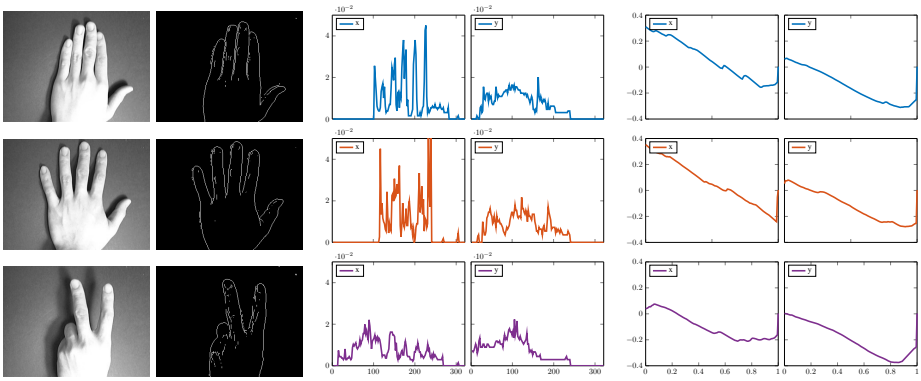
$$\mathbb{P} = \left\{ p | p(\mathbf{x}) = \frac{1}{2\pi\sigma^2} e^{-\frac{\|\mathbf{x}-\boldsymbol{\mu}\|^2}{2\sigma^2}}, \boldsymbol{\mu} \sim \text{unif}([0,1]^2) \right\}$$
$$\mathbb{Q} = \left\{ q | q(\mathbf{x}) = \frac{1}{4\pi\sigma^2} (e^{-\frac{\|\mathbf{x}-\boldsymbol{\mu}_1\|^2}{2\sigma^2}} + e^{-\frac{\|\mathbf{x}-\boldsymbol{\mu}_2\|^2}{2\sigma^2}}), \boldsymbol{\mu}_1 \sim \text{unif}([0,1]), \boldsymbol{\mu}_2 = \boldsymbol{\mu}_1 + [0, 0.2]^T \right\}$$

Figure 4.10 illustrates these classes of images (left: images, right: Radon-CDT). 2000 images were generated for the training set, 1000 images per class. Also, 2000 images were generated for the testing set, 1000 images per class. Each image is of size 40×40.



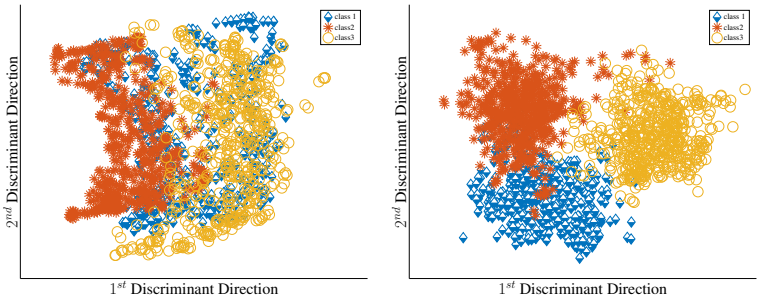
(a) PLDA projection in L^2 space **(b)** PLDA projection in CDT space

Figure 4.5: PLDA projection for flow cytometry dataset



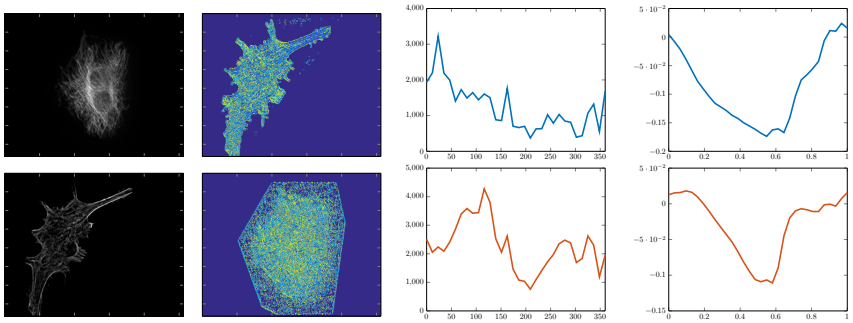
(a) Raw data **(b)** Histogram **(c)** CDT

Figure 4.6: Three different classes of hand gestures dataset



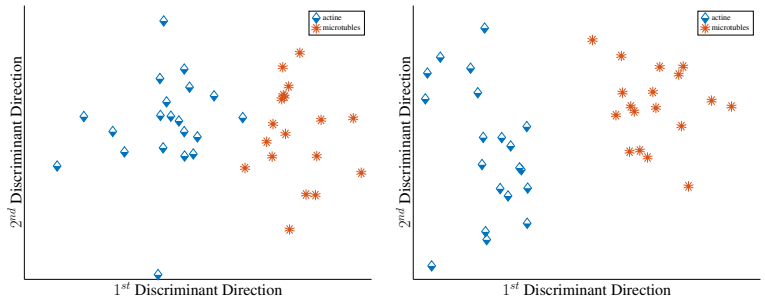
(a) PLDA projection in L^2 space **(b)** PLDA projection in CDT space

Figure 4.7: PLDA projection for hand gesture dataset



(a) Raw data **(b)** Filter Response **(c)** Histogram **(d)** CDT

Figure 4.8: Two classes of HeLa dataset, Actin (top row) vs. Microtubules (bottom row)



(a) PLDA projection in L^2 space **(b)** PLDA projection in CDT space

Figure 4.9: PLDA projection for HeLa dataset

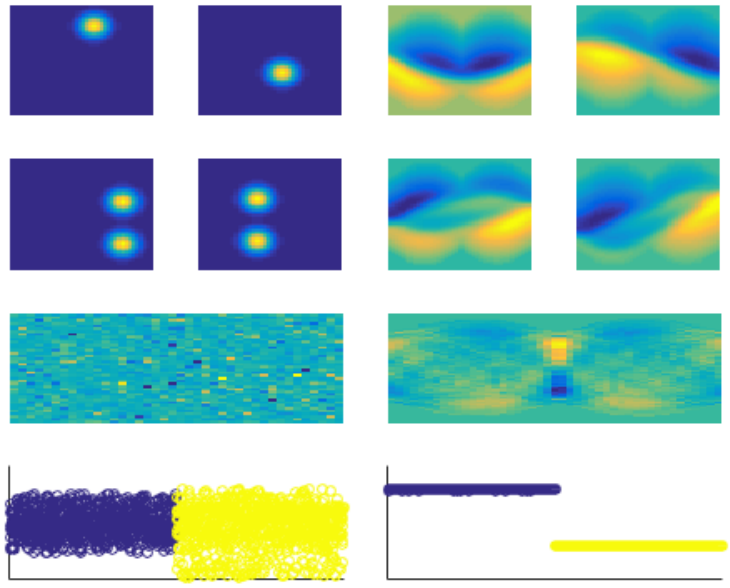


Figure 4.10: Two example image classes $\mathbb{P}$ and $\mathbb{Q}$ and their corresponding Radon-CDT, and the corresponding linear classifiers in each space

Classes $\mathbb{P}$ and $\mathbb{Q}$ are disjoint, however, they are not linearly separable in the image space. This is demonstrated by computing a linear classifier and seeing the ability to separate the training data and testing data. Specifically, the linear discriminant analysis was performed in each space. The 5-fold average classification accuracy in each space

is reported in Table. 4.5. The discriminant hyperplane computed in the training dataset doesn't generalize to the testing dataset, which suggests that the data is not linearly separable in the image space.

Perfect testing classification accuracy in the Radon-CDT space suggests that the data is linearly separable in the Radon-CDT space. Next, we demonstrate the linear separability property of our proposed image transform by calculating the Radon-CDT of classes $\mathbb{P}$ and $\mathbb{Q}$ with respect to an arbitrary image I_0 (here we used uniform density for a reference). Perfect classification accuracy in both training and testing data set implies that the image classes have become linearly separable in the Radon-CDT space.

The third row of 4.10 plots the linear hyperplane, and the fourth row shows the inner product between the *test* data samples and the linear hyperplane. In Radon-CDT space, the clear line can be drawn to distinguish two classes, suggesting that the data is linearly separable.

	TR	TS
L2	71.75 %	66.95 %
Radon CDT	100 %	100 %

Table 4.5: Linear classification accuracy for the synthetic dataset

4.2.2 Datasets

1. **MNIST Digit Classification**: MNIST [61] dataset consist of 70000 images of handwritten digits from 0-9 (see Fig. 4.11 (a)) of size 28×28.

2. **Shape Classification**: MPEG-7 [62] dataset consist of 1400 images of shapes of various size, of 70 different kinds (see Fig. 4.11 (b)). For this dataset, we resized each image into 32×32, provided that there is no loss of quality in the image.

3. **Face Image Classification**: The Carnegie Mellon University Face Images database [63] includes frontal images of 40 subjects, and contains two classes of expressions, namely 'neutral' and 'smiling' (see Fig. 4.11 (c)).

4. **Liver Nuclei Classification** The Liver Nuclei dataset contains 500 images of segmented liver nuclei extracted from histology images obtained from the archives

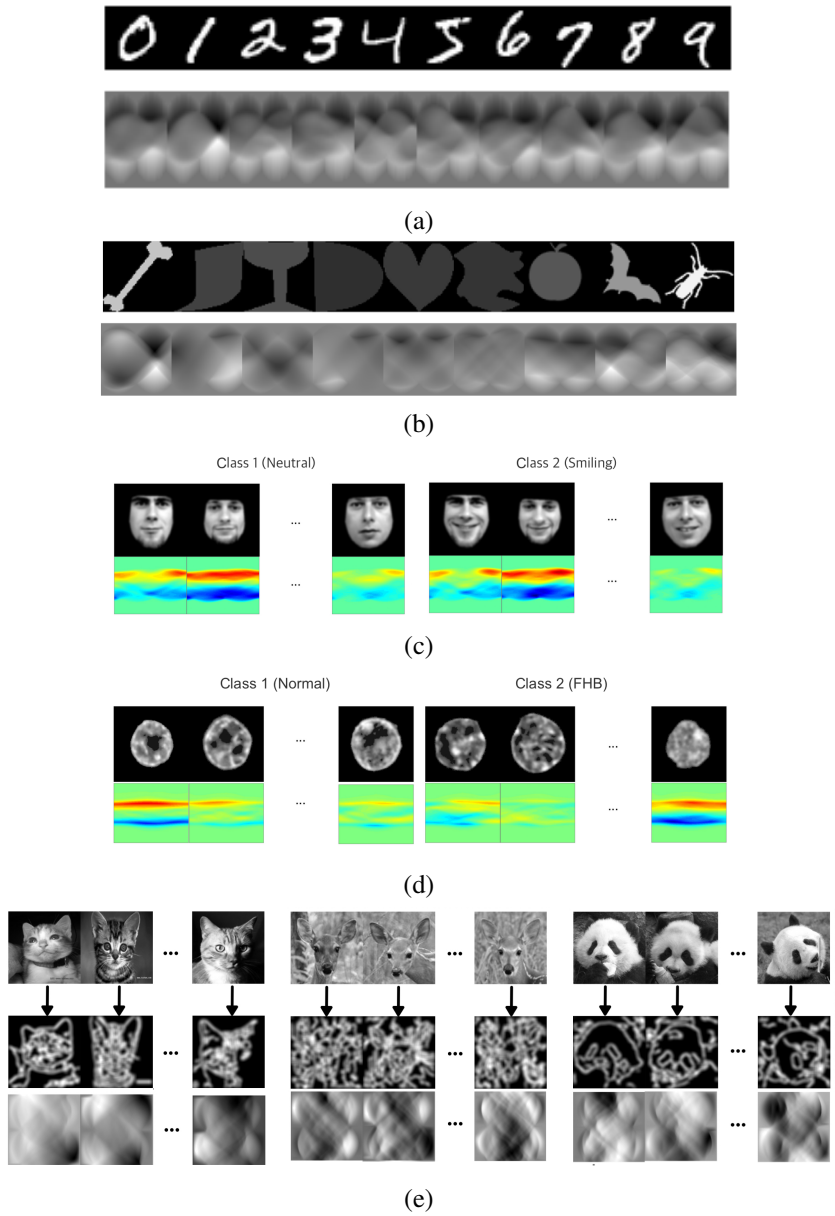


Figure 4.11: The image classes and their Radon-CDT for the MNIST(a), MPEG-7 (b), face (c), liver nuclei (d), animal face(e) data set

of the University of Pittsburgh Medical Center (UPMC). The nuclei belong to 10 different subjects including five cancer patients suffering from fetal-type hep-

Algorithm 2: Training Linear Discriminant Classifier with 5-fold Cross Validation

```
Partition the dataset into 5 groups. Leave one fold out for testing and use the
  remaining fold for training.
foreach training set do
    1. Fit the gaussian distribution for each class samples. 2. Compute the
       testing error.
3. Repeat the steps for all folds, and return the average testing error.
```

atoblastoma (FHB), with the remaining images from the liver of five healthy individuals [33] (in average 50 nuclei are extracted per subject). The classes in the nuclei dataset are 'fetal-type hepatoblastoma' (type of a liver cancer) and 'benign' for liver nuclei (see Fig. 4.11 (d)). .

5. **Animal Faces Classification**: The last dataset contains facial images of three different animals [64], , namely cat, deer, and panda under a variety of variations including translation, pose, scale, texture, etc. The dataset includes 159 images of cat faces, 101 images of deer faces, and 116 images of panda faces. The animal face dataset is preprocessed by first calculating the image edges using the Canny operator and then filtering the edge-maps of the images with a Gaussian low pass filter. (see Fig. 4.11 (e)).

4.2.3 Experimental procedure

We describe below the experimental procedures that was performed on the real images to demonstrate the linear separability in the Radon-CDT space.

For the MNIST and MPEG dataset, we fit gaussian distribution with equal covariance directly to the images of same class (i.e. linear discriminant analysis). MNIST dataset consists of 70,000 images of size 28×28, which is a reliable amount of samples to learn a distribution in $\mathbb{R}^{28\times28}$ with lesser chance of overfitting. MPEG dataset consists of 1,400 images of size 32×32, which can guide us how Radon-CDT can perform against mid-size classification problem. The data was cross-validated with 5 folds, and the detailed procedure is described in 2.

Algorithm 3: Training Linear SVM with 10-fold Cross Validation with 2nd-depth Parameter Sweep

```
Partition the dataset into 5 groups. Leave one fold out for testing and use the
  remaining fold for training.
foreach training+validation set do
    1. Compute the PCA subspace. 2. Partition the training set into 5 groups
    again. Leave one fold out for validation and use remaining fold for training.
    foreach training set do
        1. Learn the classifier for different hyper-parameter values.
        2. Compute the validation error.
        3. Return the validation error.
    4. Choose the hyper-parameter (and corresponding classifier) with the
    smallest validation error.
    5. Compute the testing error.
3. Repeat the steps for all folds, and return the average testing error.
```

For rest of the datasets, the linear SVM was learned and cross-validated on the data. A linear SVM was specifically chosen to mitigate the effect of overfitting due to small sample size. The data was divided into 10 folds, where 1 fold (20%) was used for testing and the remaining was used for training+validation.

The principal components of the datasets were calculated using the entire dataset and all data points were projected to these principal components (i.e. the dimensions which are not populated by data points were discarded).

For each 1-st depth cross validation, the linear-SVM was found using the training set, and tested on the validation set and the testing set. During training, a five-fold 2-nd depth cross validation scheme was used, to train the linear SVM (using training set) while finding the correct hyper-parameter (for margin/error trade off) for the classifier (using validation set). This was repeated for each fold, and then average accuracy is reported. The procedure is summarize in 3.

4.2.4 Discussion

The classification accuracy for training and testing set for MNIST and MPEG-7 dataset is reported in Table. 4.6 using LDA classifier. It can be clearly seen that the image classes become more linearly separable in the Radon-CDT space.

MNIST data	LDA		
	Training accuracy	Testing Accuracy	
Image space	87.36 ± 0.05	86.56 ± 0.27	(a)
Radon-CDT space	92.97 ± 0.07	92.37 ± 0.25	

MPEG-7 data	LDA		
	Training accuracy	Testing Accuracy	
Image space	100 ± 0	62.07 ± 1.83	(b)
Radon-CDT space	100 ± 0	73.93 ± 1.73	

Table 4.6: Average classification accuracy for the (a) MNIST dataset and (b) MPEG-7 dataset, calculated from five-fold cross validation using linear discriminant analysis in the image space and the Radon-CDT space

The classification accuracy for remaining dataset is reported in in Table 4.7 using linear SVM. Also, here we compared our Radon-CDT with well-known image transforms such as the Radon transform and the Ridgelet transform [65]. Radon-CDT space consistently provided higher classification accuracy than that of Radon transform or Ridglet transform. We can confirm that the Radon transform block in Radon-CDT played no role in boosting linear separability in Radon-CDT space, and rather the linear separability is derived from the Radon-CDT's original property. Throughout the dataset and regardless of the linear classification methods, It can be seen that the linear classification accuracy is not only higher in the Radon-CDT space but also computed with smaller standard deviation.

Moreover, we can check that the Radon-CDT captures the nonlinearity of the data and simplifies the data structure significantly by looking at the cumulative percent variance (CPV) captured by the principal components. Figure 4.12 shows the CPV calculated from the image space, the Radon transform space, the Ridgelet transform space, and the Radon-CDT space as a function of the number of principal components for all the datasets. It can be seen that the variations in the datasets are captured more efficiently and with fewer principal components in the Radon-CDT space as compared to the other transformation spaces. This indicates that the data structure becomes simpler in the Radon-CDT space, and the variations in the datasets can be explained with fewer parameters. We can also visually verify that the representation in Radon-CDT space appears more simpler than that in image space from Figures 4.11 to 4.11.

Face data	Linear SVM		
	Training accuracy	Testing Accuracy	
Image space	100 ± 0	76.0 $\pm$ 11.94	(a)
Radon space	100 ± 0	79.12 ± 12.25	
Ridgelet space	100 ± 0	68.75 ± 78.12	
Radon-CDT space	100 ± 0	82.62 ± 11.5	

Nuclei data	Linear SVM		
	Training accuracy	Testing Accuracy	
Image space	100 ± 0	65.2 ± 6.6	(b)
Radon space	100 ± 0	62.56 ± 6.7	
Ridgelet space	100 ± 0	63.8 ± 66.7	
Radon-CDT space	100 ± 0	75.56 ± 6.21	

Animal Face data	Linear SVM		
	Training accuracy	Testing Accuracy	
Image space	59.77 $\pm$ 2.68	58.24 $\pm$ 2.07	(c)
Radon space	100 ± 0	77.38$\pm$ 5.18	
Ridgelet space	94.21 ± 0.03	85.11 ± 5.59	
Radon-CDT space	92.48 ± 0.53	86.43 ± 5.29	

Table 4.7: Average classification accuracy for the face dataset (a), the nuclei dataset (b), and the animal face dataset (c), calculated from ten-fold cross validation using linear SVM in the image space, the Radon transform space, and the Radon-CDT spaces. The improvements are statistically significant for all datasets.

In summary, provided that the image dataset is generated according to the model specified in Theorem. 3.4.1, the image classification in the Radon-CDT space can be performed using linear classifiers with less overfitting. This can be a useful solution when only small number of samples available. Also, simpler representation (using a fewer number of bases) in Radon-CDT space provides additional benefit.

We emphasize here that the use of linear classifier over any other nonlinear classifiers is intentional. The classification experiments in this subsection serve as a measure of linear separability of image classes in the corresponding transform spaces and are designed to test our theorem on the linear separability of image classes in the Radon-CDT space.

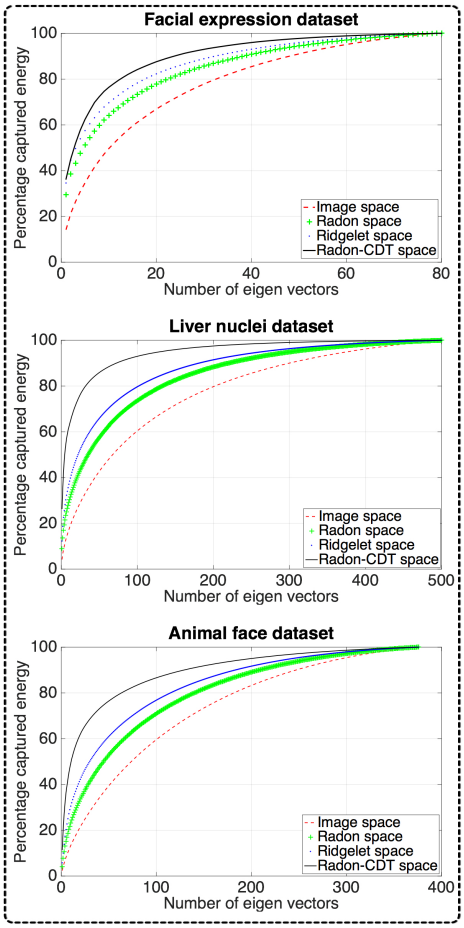


Figure 4.12: Cumulative percent variance (CPV) captured by the principal components

Chapter 5

Applications to Demultiplexing Optical Spatial Patterns

5.1 Introduction to Free Space Optical (FSO) Communications

Free-space optics (FSO) communication systems use optical or laser beams with optical wavelengths such as ultraviolet, visible, and infrared for communication. In comparison to fiber optics, FSO communications systems have fewer channel limitation and can be used in air, water or terrain where optical fiber might be too expensive to install. Also, FSO systems are capable of providing larger signal intensity at the receiver – due to smaller beam divergence– than the radio frequency (RF) communications systems, which promoted the recent advances in deep space communications between Mars and Earth [66].

In FSO communications, to achieve higher channel capacity, modulated optical (or laser) beams have been multiplexed either in time, frequency, polarization, or separate locations. A decade ago, [67] demonstrated that the light could also be multiplexed in orbital angular momentum (OAM). Paraxial beams possessing different orbital angular momentum are orthogonal, and therefore can be multiplexed and demultiplexed

without information interference. There exist some controversies over using OAM multiplexing that there is no gain in data rate when compared to traditional spatial multiplexing (MIMO) [68] and that OAM does not increase the channel capacity compared to a mode set void of OAM [69]. But the diverse signal path for spatial MIMO cannot always be guaranteed, and successive researches [70, 71, 72, 73] have demonstrated the prospects of utilizing OAM carrying beams in FSO communication.

OAM beams are orthogonal in nature. However, demultiplexing OAM beams becomes nontrivial when the orthogonality breaks down in data channels. In fibers, OAM modes become unstable when fibers are bent and stressed. In free space communications, atmospheric effects such as scintillation, turbulence, beam wandering, will challenge demultiplexing OAM beams. Instead of direct detection of OAM beams using its orthogonality principles [74], indirect methods have been investigated, where [71] demultiplexed the OAM beams using offline DSP (digital signal processor) based on coherent detection and MIMO, while [72, 75] took an approach utilizing machine learning methods.

Here we sought to demultiplex OAM carrying beams as in [72, 75], by adopting a pattern classification methods. Specifically, the OAM carrying beams are captured by a CCD camera, and will be represented as a unique pattern, and can be identified with a classification system. Moreover, based on the recent finding [42] that the light traveling in atmospheric turbulence would approximately follow the 'optimal transport path', we investigate the hypothesis that the OAM beams undergo deformation which can be decoded easily in the transport based transforms described earlier. We demonstrate that the OAM carrying beam patterns are (nearly) linearly separated in the Radon-CDT space, and are near perfectly separated in the Radon-CDT space with the aid of the shallow convolutional network.

The last chapter is organized as follows. In Section 5.2, we overview the basics of FSO communications utilizing OAM carrying beams. In Section 5.3, we introduce how we can apply classification system to demultiplex OAM carrying beams. In Section 5.4, we explain the experimental procedures, including the procedures for the data collection and the outline of the computational methods. In Section 5.5, we demonstrate that the demultiplexing OAM carrying beams is feasible with a simple classifier,

which is robust to turbulence, spatial down-sampling, and beam-wandering.

5.2 FSO Communication with OAM Carrying Beams

5.2.1 Orbital Angular Momentum (OAM)

Mathematically, an electromagnetic wave can be described as a field $u(x, y, z, ; t)$ with spatial coordinates (x, y, z) and time t, which follows the hyperbolic partial differential equation [76]:

$$\frac{\partial^2 u}{\partial t^2} = c^2 \nabla^2 u,$$

where ∇^2 is Laplacian operator defined by

$$\nabla^2 u = \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2}.$$

For electromagnetic waves, $c = 3x10^8 m/s$ is the speed of flight. If the field variation is sinusoidal (i.e. a monochromatic wave), $u(x, y, z; t) = U(x, y, z)e^{-i\omega t}$, where ω is the angular frequency and $U(x, y, z)$ is the complex amplitude of the wave, then we get the time-independent reduced wave equation (or Helmholtz equation):

$$\nabla^2 U + k^2 U = 0, \tag{5.1}$$

where $k = \omega/c = 2\pi/\lambda$ is the optical wavenumber and λ is the wavelength.

If we now change to cylindrical coordinates, Eq. (5.1) becomes

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial \tilde{U}}{\partial r}\right) + \frac{\partial^2 \tilde{U}}{\partial z^2} + k^2 \partial U = 0.$$

Using a simplification, $V(r, z) = \tilde{U}(r, z)e^{-ikz}$, and the paraxial assumption, $\partial^2 V/\partial z^2 = 0$, we get the paraxial wave equation:

$$\frac{1}{r}\frac{\partial}{\partial r}\left(r\frac{\partial V}{\partial r}\right) + 2ik\frac{\partial V}{\partial z} = 0. \tag{5.2}$$

By solving Eq. (5.2) in different coordinate system with different symmetry assumptions, several beams that carry OAM can be defined. For example, Gaussian-tapered Bessel beams (BGB) arise from assuming the problem is cylindrically symmetric and solving circular cylindrical coordinates and Bessel function. If the problem is cylindrically smmetric and cylindrical coordinates (r, ϕ, z) is used, the Laguerre-Gauss beam (LGB) arise.

5.2.2 Laguerre-Gauss Beam (LGB)

Laguerre-Gauss beams, an OAM carrying beam with spiral phase distribution $\exp(im\theta)$ are described as

$$u_{LG(p,m)}(r,\theta,z) = \frac{C_{LG(p,m)}}{w(z)} \left(\frac{r\sqrt{2}}{w(z)}\right)^{|m|} L_p^{|m|}\left(\frac{2r^2}{w^2(z)}\right) \times \tag{5.3}$$

$$\exp\left(\frac{-r^2}{w^2(z)} - \frac{ikr^2 z}{2(z^2+z_R^2)}\right) \times \exp(i(2p+|m|+1)\psi(z))\exp(im\theta) \tag{5.4}$$

or when $z = 0$

$$u_{LG(p,m)}(r,\theta,z=0) = C_{LG(p,m)} \left(\frac{r\sqrt{2}}{w_0}\right)^{|m|} L_p^{|m|}(2r^2) \times \exp(im\theta).$$

L_p^m is the generalized Laguerre-Gauss polynomial of p (radial mode) and m (angular mode), $C_{LG(p,m)}$ is a normalization constant, $\psi(z) = \tan^{-1}(z/z_R)$ is the Gouy phase, $w(z) = w_0\sqrt{1+(z/z_R)}$ is the beam radius, w_0 is the beam waist, and $z_R = \pi w_0^2/\lambda$ is the Rayleigh range. An example of LGB with OAM mode $m = 1, \cdots, 5$ (left to right) propagating in z direction is shown in Fig. 5.1. The top row shows the magnitude of the transverse plane at $z = 0$, and the bottom row shows the phase front of the same transverse plane. LGB beams in each column are realizations of different orbital angular momentum, and hence orthogonal to one another.

Fig. 5.2 and Fig. 5.3 shows the LGB with OAM mode $m = 1, \cdots, 5$ (left to right) propagating in z at $z = 10$ and $z = 100$ respectively. As the beam propagates forward, we can see that the phase in the field circulates around the the points of zero intensity

(i.e. vortex).

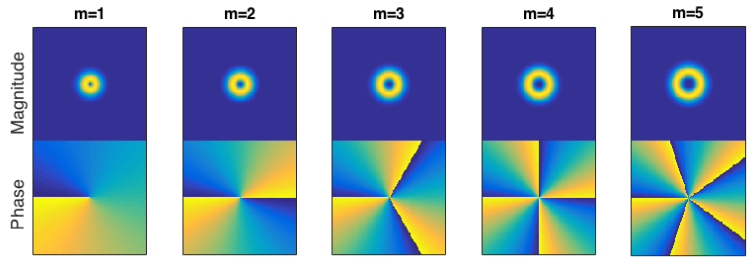


Figure 5.1: Laguerre-Gauss Beam, z = 0.

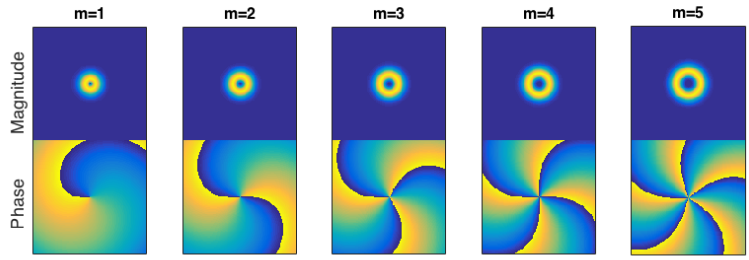


Figure 5.2: Laguerre-Gauss Beam, z = 10.

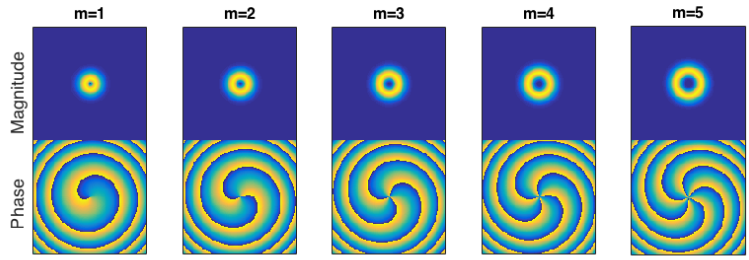


Figure 5.3: Laguerre-Gauss Beam, z = 100.

5.2.3 Gaussian-tapered Bessel Beam (BGB)

Ideal Bessel beams, an OAM carrying beam with spiral phase distribution $\exp(im\theta)$, are described as

$$u_{B(m)}(r,\theta,z) = C_b J_m(\beta r)\exp(-ik_z z)\exp(im\theta)$$

where J_m is the order of m Bessel function and β is the radial frequency, $k = \sqrt{k_z^2 + \beta^2} = 2\pi/\lambda$. m defines the OAM mode number of photons that can take any integer value. Example Ideal BG beams with OAM mode $m = 1, \cdots, 5$ (left to right) are shown in Fig. 5.4. The top row shows the magnitude of the transverse plane at $z = 0$, and the bottom row shows the phase front of the same transverse plane. Fig. 5.5 and Fig. 5.6 show the ideal BG beam propagating in the direction of z at $z = 10(m)$, and $z = 100(m)$ respectively. Because the ideal BG beam is diffraction free, the magnitude of the BG beam stays the same, and only the phase rotates as it propagates in the medium.

And ideal Bessel beam cannot be realized, however, because it would require an infinite amount of energy. Gaussian-tapered Bessel beam (BGB) can approximate ideal Bessel beam for a finite distance (pseudo diffraction-free beam) instead. For the circular symmetric case, the pseudo BGB can be realized by [77]:

$$u_{BG(m)}(r, \theta, z) = \frac{C_{BG} w_0}{w(z)} J_m \left(\frac{\beta r}{1 + iz/z_r} \right) \tag{5.5}$$

$$\times \exp \left(i(\left(k - \frac{\beta^2}{2k} \right) z - \psi(z) - \frac{1}{w^2(z)} \right) \tag{5.6}$$

$$\times \exp \left(\frac{ik}{2R(z)} \right) \left(r^2 + \beta^2 \frac{z_r}{k^2} \right) \exp(im\theta), \tag{5.7}$$

where $R(z) = z \left(1 + \left(\frac{z_R}{z} \right)^2 \right)$ is the radius of curvature of the beam, $\psi(z) = \tan^{-1}(z/z_R)$ is the Gouy phase, $z_R = \pi w_0^2 \lambda$ is the Rayleigh range, $w(z) = w_0 \sqrt{1 + (z/z_R)}$ is the beam radius, w_0 is the beam waist. When $z = 0$, the above equation simplifies to:

$$u_{BG(m)}(r, \theta, z = 0) = C_{BG} J_m(\beta r) \exp \left(-(r/w_0)^2 \right) \exp(im\theta).$$

Fig. 5.7 shows the magnitude (top) and phase front (bottom) of the Pseudo Bessel-Gauss Beam at $z = 0$. Fig. 5.8 and Fig. 5.9 show the Pseudo BG beam at $z = 100(m)$ and $z = 10000(m)$ respectively. The pseudo BG beam retains its diffraction-less property up-to several kilometers, compared to the LG beam. Also, the Gaussian phase profile, $\exp(im\theta)$, allows these beams to exhibit OAM. The BG beams in each column

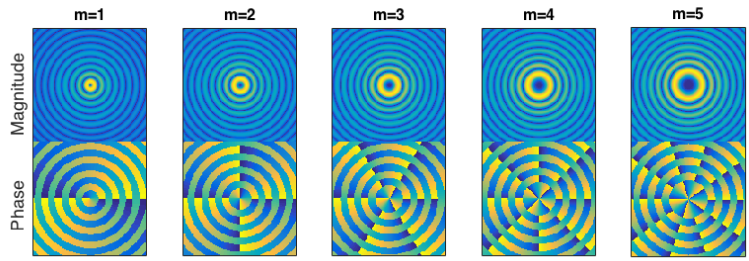


Figure 5.4: Ideal Bessel-Gauss Beam, z = 0.

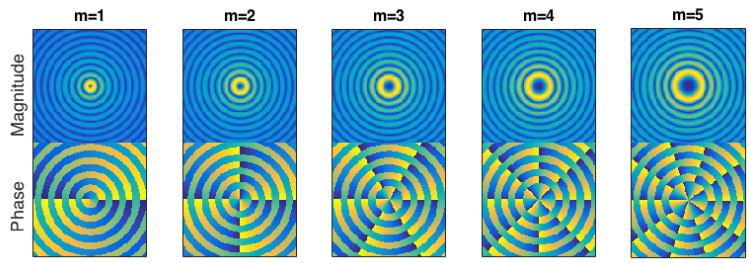


Figure 5.5: Ideal Bessel-Gauss Beam, z = 10.

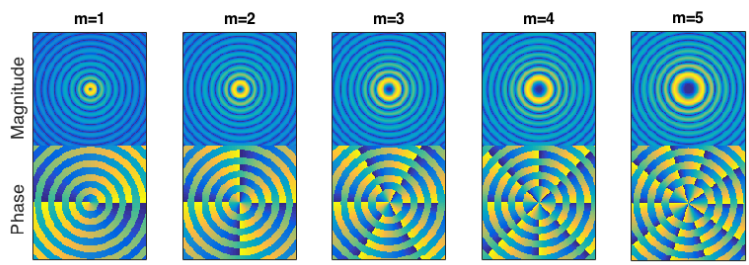


Figure 5.6: Ideal Bessel-Gauss Beam, z = 100.

(of different mode m) are orthogonal to one another, i. e.

$$\int_r \int_\theta u_{BG(m_1)} u^*_{BG(m_2)} dr d\theta = 0.$$

A BGB is produced by the superposition of Gaussian beams whose axes are uniformly distributed on a cone of angle θ_c. The radial frequency of the BGB is given by: $\beta = k \sin(\theta_c)$. For a fixed θ_c, after a certain propagation distance, the BGB beam

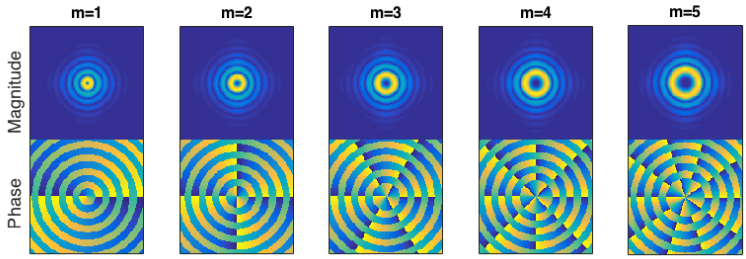


Figure 5.7: Pseudo Bessel-Gauss Beam, z = 0 (m).

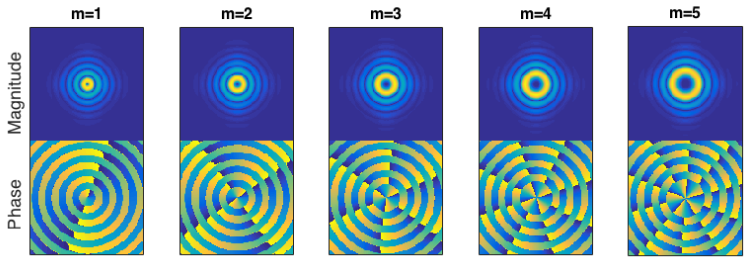


Figure 5.8: Pseudo Bessel-Gauss Beam, z = 100 (m).

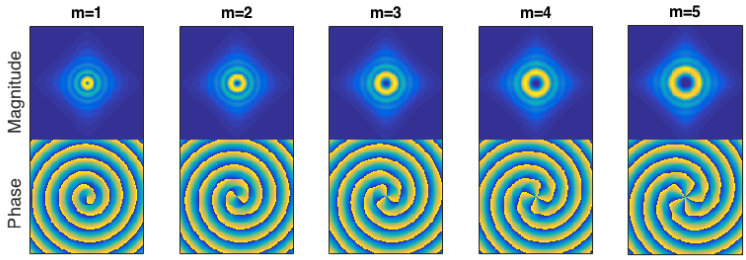


Figure 5.9: Pseudo Bessel-Gauss Beam, z = 10000 (m).

will result in normal Gaussian beams. The radial frequency (or the angle of the cone) supporting the propagation distance Z can be found by

$$\beta = k \sin(w_0/Z).$$

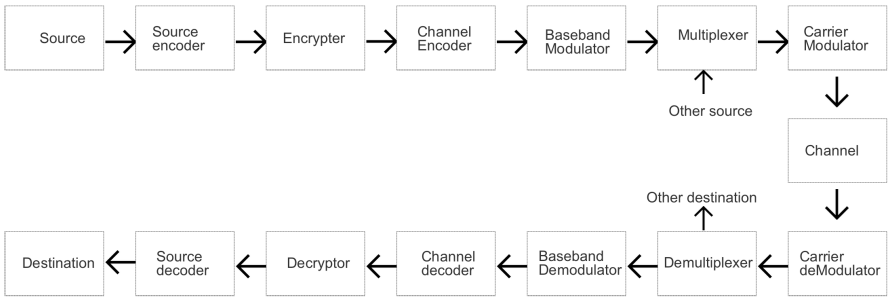


Figure 5.10: RF Communications Diagram

5.2.4 OAM Communications System

Free space communications using OAM carrying beams are mostly adapted from the RF communications systems. Fig. 5.10 illustrates the standard RF communications pipeline: (a) analog signals are digitized to digital bits in the source encoder, (b) error correcting codes are added in the channel encoder so that the bits can be more robust to the errors in noisy channels, (c) source symbols are represented with analog signals in the baseband modulator, (d) single data source is multiplexed with other sources in the multiplexer, and (e) baseband signals are modulated into carrier frequency so that it can be transmitted over the channel in the carrier modulator. OAM carrying beams have been utilized to substitute conventional baseband modulations (denoted as (c)) or subcarrier multiplexing (denoted as (d)). [78, 73].

Baseband Modulation

In baseband modulation, single OAM mode can be used to represent digital symbols of single data stream. So far, most of the work has been carried out using On-off keying (OOK) modulation scheme [79, 80, 81, 82] or pulse position modulation (PPM) [83, 84] because of its simple implementation. In OOK, the transmission of binary data is represented by the presence or absence of light pulse, i.e., if the information bit is 1, the laser is turned on for the duration T_b, and if it is 0, nothing is transmitted. Fig. 5.11 shows the OOK modulation scheme for the transmission of message 110010. In M-PPM scheme, each symbol period is divided into M time slots each of duration T_s

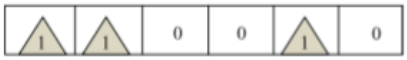


Figure 5.11: OOK modulation scheme for the transmission of message 110010

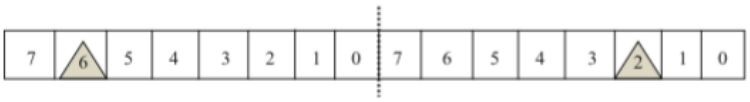


Figure 5.12: 8-PPM scheme with eight slots for the transmission of message 110010

seconds, and the information is placed in one of the M time slots to represent a data word. Here, $M = 2^n$ where n is the number of information bits. Therefore, each PPM symbol is mapped directly to n bit sequence and thus allows $\log_2 M$ bits within each PPM symbol. Fig. 5.12 describes how PPM modulations scheme can be used to transmit message 110010 [85].

On the other hand, we can expect OAM carrying beams be utilized in a similar manner as Frequency shift keying (FSK) modulation scheme. In FSK modulation, different frequencies can be used to represent different data symbols. As such, we can use OAM carrying beams with different OAM modes to represent data symbols. Table. 5.1 illustrates how we can encode 2-bit symbol $[0,0], [0,1], [1,0], [1,1]$ onto OAM beams with 4 modes $\{m_1, m_2, m_2, m_4\}$.

OAM Modes	m_1	m_2	m_3	m_4
Symbol	[1,1]	[0,1]	[1,0]	[0,0]

Table 5.1: Modulation Example

Multiplexing

To utilize OAM modes for multiplexing [73], separate data streams are first modulated via any modulation schemes of choice. Any modulation scheme like binary phase-shift keying (BPSK), quadrature phase-shift keying (QPSK), quadrature amplitude modulation (QAM), amplitude modulation (AM), or frequency modulation (FM) are a valid choice. Then each data stream can be multiplexed using OAM beams in a similar manner as the frequency division multiplexing (where each frequency channel serves as the

optical carrier). As FDM, each OAM mode channel can serve as the optical carrier due to their orthogonality. Moreover, OAM carrying beams can share the same frequency channel without interfering with one another, which provides much higher channel capacity. Also, OAM can multiplex single stream data as well - dividing high data rate to multiple low data rates and transmitting them in parallel form.

Various spatial patterns can be created by multiplexing OAM carrying beams. For example, Fig. 5.13 shows the analytical plot for the multiplexed patterns of LG beam using mode set $= [25, 20, 15, 10, 5]$. The bit string indicates the presence of the OAM beam; for example, 10000 implies that only mode 25 is multiplexed, and 11000 implies that mode 25, 20 are multiplexed.

Similarly, Fig. 5.14 shows analytical plots for the multiplexed patterns of BG beam of modes $= [-3, -1, 2, 4, 6]$, Fig. 5.15 shows the patterns of BG beam of modes $= [-4, -1, 2, 5, 8]$, and Fig. 5.16 shows patterns of BG beam of modes $= [-7, -2, 3, 8, 13]$.

By using different types of beams (e.g. Lauguerre Guass, Hermit Gauss, etc.) and by the combination of different mode sets, distinctive spatial patterns can be designed. Each distinctive spatial pattern can carry information, either by encoding symbols (i.e. modulation) or carrying independent information streams (i.e. multiplexing).

Detection of OAM Carrying Beams

Receiver design of the OAM detector is based on the principle of orthogonality. Although the superposition and simultaneous propagation of a set of OAM beams with different OAM mode can produce an unrecognizable intensity pattern, the inner product of any two of the orthogonal fields is zero after propagating in a vacuum. Thus, co-propagating modes can be perfectly split and recombined.

Given the received field $u_n(r, \phi, z)$ with OAM state n and the analyzing field $u_m^*(r, \phi, z)$ (where the $*$ indicates the complex conjugate) with OAM state m, the mth

channel output signal (observed at the detector of channel m) is

$$u_n(r,\phi,z)\cdot u_m(r,\phi,z) = \int rdrd\phi u_n(r,\phi,z)u_m^*(r,\phi,z)$$
$$= \begin{cases} 0 & \forall n \neq m \\ \int rdrd\phi |u_m(r,\phi,z)|^2 & n = m \end{cases}$$

It is clear that for $n = m$, the outcome is equal to the total power of the observed field. With this approach one can use on-off keying (OOK), pulse-position modulation (PPM), or any other modulation scheme appropriate for direct-detection.

Conjugate mode sorting

The conjugate mode sorting [86] is a method to determine the OAM mode of a detected beam based on its orthogonality properties. Given a transmitted OAM beam, $u_n(r,\phi,z)$, OAM Beams of support of the mode set $u_m^*(r,\phi,z)$ are cycled through. In optical implementation, the beam from the transmitter is reflected from the spatial light modulator (SLM), programmed with the analysing-hologram pattern of negative mode. If the m value of the beam is added with the $-m$ value of the hologram, then the resulting beam has planar phase fronts and therefore can be focussed through a pin hole.

Fig. 5.17 illustrates the concept of conjugate mode sorting with Laguerre-Gauss (LG) beam. The LG beam of mode $m = 5$ (left, plotting magnitude only) is multiplied with its complex conjugate, resulting in real-valued planar phase wave (middle, real-valued). When the multiplication of the two, $u_{m=5}(r,\theta)u_{m=5}^*(r,\theta)$, is Fourier transformed, we get a transform of high-intensity at the origin. However, when the transmitted mode n and analyzing mode m are mismatched (as in Fig. 5.18, which illustrates when the mode $m = 10$ is used for detection instead), the Fourier transform results in a doughnut-shaped transform indicating that the transmitted signal does not contain the OAM mode n.

This sorting method is dependent on having good alignment between the transmitter and the receiver; mis- alignment is shown to have comparable effects to turbulence

in the correct determination of the OAM mode. Due to the effect of turbulence, the normalized energy will not be concentrated exactly at the origin of the correct conjugate mode, thus we have to look at the relative energy across all the modes. For the non-multiplexing case, one can simply take the maximum value across the support of the mode set; for the multiplexing case, a threshold must be chosen so as to decide whether a mode is present or not in the signal.

Other classical methods of detecting OAM beams include optical transformation sorting [87], counting spiral fringes [88], using dove prism interferometer [89], or measuring the doppler effect [90], which are all built upon the orthogonality principle of OAM beams.

However, OAM beams arriving at the transmitter would undergo undesired deformations due to the various atmospheric effects, such as changes in air pressure and temperatures, the presence of cloud and particles, and OAM beams would be no longer orthogonal when arriving at the receiver. As a consequence, the performance of the classical detectors degrades substantially in the presence of turbulence. Therefore, to detect the multiplexed OAM patterns deformed by strong atmospheric effects, we instead turn to utilizing a machine learning approach.

5.3 Detection of OAM Carrying Beams via Classification

5.3.1 Problem Formulation

Consider a OAM mode set $\{m_i\}_{i=1}^{\log_2(K)}$ and suppose $\varphi_{m_i}(\vec{x})$ is the OAM beam of mode m_i. For example, $\varphi_{m_i}(\vec{x})$ may correspond to the Laguerre-Gauss beam presented in (5.4) or the Bessel-Gauss beam presented in (5.7). Let $I(\vec{x})$ be generated by linearly combining the beams with OAM:

$$I(\vec{x}) = \sum_{i=1}^{\log_2(K)} c_i \varphi_{m_i}(\vec{x}).$$

c_i, either 0 or 1, indicates whether the mode m_i is present in $I(\vec{x})$. Modes can be linearly combined in K $(= 2^{\log_2(K)})$ different ways, resulting in K distinct beams. Each combination patten is associated with an integer label $[1, \cdots, K]$.

Given a set of arbitrary OAM beams $[I_1, \cdots, I_N]$, now detecting which modes are present translates into a problem of assigning one of K integers to the beam, i.e. identifying its corresponding label $[\ell_1, \cdots, \ell_N]$. In order to do so, a classification model $\mathcal{W}$ that identifies the label, $\hat{\ell}_i = \mathcal{W}(I_i)$, shall be derived, and we want that identification to match its original label, i.e. $\hat{\ell}_i = \ell_i$.

Formally speaking, given a training set $[I_1, \cdots, I_N]$ and its corresponding labels $[\ell_1, \cdots, \ell_N]$, a classifier $\mathcal{W}$ can be learned such that it minimizes mis-classification measure $\mathcal{L}$:

$$\mathcal{W}^* = \arg\min_{\mathcal{W}} \mathcal{L}(\mathcal{W}(I_i), \ell_i).$$

Once the model is learned, I_{test} can be identified by

$$\hat{\ell}_{test} = \arg\min_{\ell=\{1,\cdots,K\}} \mathcal{L}(\mathcal{W}(I_{test}), \ell). \tag{5.8}$$

Especially, when $\mathcal{L}$ is the expected value of the 0-1 loss, $E[1(\ell_i, \hat{\ell}_i)]$, minimizing $\mathcal{L}$ leads to a bayes classifier:

$$\hat{\ell}_{test} = \mathcal{W}(I_{ts}) = \arg\max_{\ell} P(\ell | I = I_{ts}). \tag{5.9}$$

5.3.2 Classification Methods

5.3.3 Linear Discriminant Analysis

Linear discriminant analysis (LDA) [91] is a bayes classifier that arise when conditional probability density functions for each class are modeled as a multivariate normal distribution:

$$P(I|\ell = k) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)},$$

and when each class shares the same covariance matrix Σ.

The training step involves finding the model parameters $\{\mu_\ell\}_{\ell=1}^K$ and Σ. Given the training set $\mathcal{I} = \{I_n\}_{n=1}^N$, the estimates of the class means are given as

$$\hat{\mu}_k = \frac{\sum_{n=1}^N p_{nk} I_n}{\sum_{n=1}^N p_{nk}}.$$

$p_{nk} = 1$ if observation n is from class k and $p_{nk} = 0$ otherwise. The unbiased estimate of the covariance matrix is given as

$$\hat{\Sigma} = \frac{\sum_{n=1}^N \sum_{k=1}^K p_{nk}(I_n - \hat{\mu}_k)(I_n - \hat{\mu}_k)^T}{N - K}.$$

The testing phase involves finding the label that maximizes the Bayes classification rule in Eq. (5.9):

$$\hat{\ell}_{ts} = \arg\max_\ell p(\ell | I = I_{ts}) = \frac{p(I_{ts}|\ell)p(\ell)}{\sum_{\ell=1}^K p(I_{ts}|\ell)p(\ell)},$$

among all possible $\ell = [1, \cdots, K]$. Specifically, determining whether the data I_{ts} belongs to label k_1 or k_2 can be done by looking at the log-ratio the posteriori probabilities:

$$\begin{aligned}
&\log \frac{P(\ell = k_1 | I = I_{ts})}{P(\ell = k_2 | I = I_{ts})} = \log \frac{P(I|k_1)}{P(I|k_2)} + \log \frac{p(k_1)}{p(k_1)} \\
&= \frac{p(k_1)}{p(k_1)} - \frac{1}{2}(\mu_{k_1} + \mu_{k_2})^T \Sigma^{-1} (\mu_{k_1} - \mu_{k_2}) + x^T \Sigma^{-1} (\mu_{k_1} - \mu_{k_2}).
\end{aligned}$$

The set yielding $P(\ell = k_1 | I = I_{ts}) = P(\ell = k_2 | I = I_{ts})$, whose log ratio is 0, is the decision boundary between the two classes k_1 and k_2. As the name LDA suggests, the decision boundary is represented as a linear hyperplane:

$$\begin{aligned}
p(\ell = k_1 | I = I_{ts}) &= p(\ell = k_2 | I = I_{ts}) \\
&\Leftrightarrow I_{ts}^T w = b
\end{aligned}$$

where $w = \Sigma^{-1}(\mu_{k_1} - \mu_{k_2})$ and $b = -\log \frac{p(k=k_1)}{p(k=k_2)} + \frac{1}{2}(\mu_{k_1} + \mu_{k_2})^T \Sigma^{-1} (\mu_{k_1} - \mu_{k_2})$.

5.3.4 Convolutional Neural Networks

Convolutional Neural Networks (CNN) is a supervised learning method for finding a non-linear mapping between input and output pairs. The mapping is made up of composition of functions:

$$y^{(L)} = a_L \circ a_{(L-1)} \circ \cdots \circ a_{(1)}(I).$$

$a_{(i)}$ denotes the ith layer's mapping function and $y^{(i)}$ denotes its output. Functions range from linear operations such as convolution and downsampling to non-linear functions such as max-pooling and rectified unit activations. Network architectures can vary greatly depending on the combinations of functions, and because of this versatility, CNN can produce much flexible and complex decision rules.

For multi-class classification, we specifically choose the final layer a_L to be a *soft-max* function, i.e.

$$y_\ell^{(L)} = a_{(L)}\left(y_\ell^{(L-1)}\right) = \frac{e^{y_\ell^{(L-1)}}}{\sum_{k=1}^{K} e^{y_k^{(L-1)}}}.$$

In probabilistical view point, the output $y_\ell^{(L)}$ can be interpreted as posterior probability $p(\ell|I = I_{ts})$. By Bayes classification rule, the negative log of the posterior is minimized (equivalent to maximizing the posterior probability):

$$\begin{aligned}
-\log p(\ell|I = I_{ts}) &= -\log\left(\frac{e^{y_\ell^{L-1}}}{\sum_{j=1}^{K} e^{y_j^{L-1}}}\right) \\
&= -\sum_{k=1}^{K} 1(\ell = k)\log\left(\frac{e^{y_k^{L-1}}}{\sum_{j=1}^{K} e^{y_j^{L-1}}}\right) \\
&= -\sum_{k=1}^{K} p_k \log \hat{p}_k,
\end{aligned}$$

which is coined as cross-entropy loss. $p_k = 1$ when $k = \ell$ and otherwise $p_k = 0$. The training step involves finding the parameters for subsequent layers $a_1, \cdots, a_L$ that minimizes the cross-entropy loss.

At the testing step, the label is assigned based on the bayes classification rule:

$$\ell_{ts}^* = \arg \max_{k=1\cdots,K} \frac{e^{y_k}}{\sum_{j=1}^{K} e^{y_j}}.$$

5.4 Experimental Setup

5.4.1 Laboratory Setup and Data Collection

The OAM carrying BGB beams were simulated in the Naval Research Laboratory, with 633-nm Gaussian laser, binary phase ferroelectric spatial light modulators (SLM), Dalsa GigE camera, and several optical tools such as mirrors and pinhole filters. The diagram of the experiment is depicted in Fig. 5.19. The collimated Gaussian plane wave is emitted, and as it interferes with the spatial light modulator, the OAM modes are encoded into the beam. Specifically, the SLM is programmed with a binary phase hologram which represents the desired OAM mode multiplexing and the effect of turbulence. The beam propagates through free space until the beam is recorded by the camera. High spatial frequency components are filtered by the pinhole located before the camera. The SLM and camera are driven by a MATLAB program.

Simulating the Effect of Turbulence

The light propagating through the turbulence is simulated through the phase screen method [92]. In this method, light passes through a layer of changing diffractive index, and is assumed to diffract and propagate to the next layer. In order to produce this behavior of light, statistically generated phase screens are placed at a layer. These phase screens represent the phase that would have occurred in propagation from the previous layer. Specifically, phase screens can be generated as follows. Complex Gaussian white noise (two 2-D array of i.i.d. Gaussian random numbers for the real and imaginary parts) C is multiplied by the square root of the spectrum $\Psi(\kappa)$, and then inverse Fourier transformed:

$$P = \mathcal{F}^{-1}\{\sqrt{\Psi(\kappa)}C\}. \tag{5.10}$$

where $\mathcal{F}^{-1}$ is the inverse 2-D Fourier transform, and $\Psi(\kappa)$ corresponds to modified Kolmogorov turbulence model [93]:

$$\Psi(\kappa) = 0.033C_n^2(\kappa^2+1/L_0^2)^{-11/6}\exp(-\kappa^2/\kappa_\ell^2)\times(1+1.082(\kappa/\kappa_\ell)-0.254(\kappa/\kappa_\ell)^{7/6}),$$

where κ is the spatial frequency (rad/m), L_0 is the outer scale of turbulence, ℓ_0 is the inner scale of turbulence, and $\kappa_\ell = \kappa/(3.3\ell_0)$. C_n^2 is the structure constant of the index refraction, which measures the strength of the turbulence (more details on the turbulence theory and the phase screen method can be found in Section C.2.1).

The strength of the simulated turbulence are varied controlling D/r_0, where D is the dimension of the SLM and r_0 is the Fried parameter in Eq. (C.2). In the experiments shown below, 3 levels of turbulence $D/r_0 = [5, 10, 15]$, corresponding to medium, strong, very strong turbulence were simulated via the phase screen method.

Encoding the Hologram

The ideal OAM beam multiplexed with modes $\{m_1, m_2, \cdots, m_M\}$ under the effect of turbulence – which is encoded into the phase screen $P[x, y]$ – can be represented as

$$S[x, y] = \exp(iP[x, y])\sum_{j=1}^{M} u_{BG(m_j)}[x, y].$$

The hologram is written to the SLM, which, when illuminated by a Gaussian plane wave, will create a desired multiplexed OAM beam with simulated turbulence. The phase of the hologram imprinted on SLM is created as follows. Uniform amplitude beams with a tilt phase and the multiplexed beam phase are added to create an off-axis hologram:

$$H[x, y] = |\exp(ix) + \exp(i\angle(S[x, u]))|^2.$$

H is then binarized

$$\hat{H}[x, y] = \begin{cases} \pi & H[x, y] > (|H_{\max}| + |H_{\min}|)/2 \\ 0 & \text{otherwise.} \end{cases}$$

The SLM had pixel dimensions of 1280 x 1024 (but only centered 1024x1024 region was used to keep things square). The pixel size was 13.62 micrometers (μ m), and the entire SLM dimension D was about 1.3cm assuming that the laser fully illuminated the SLM. The binary phase-only SLM was used for efficient projection, however, but can create the light that is an approximation of the true light field.

5.4.2 Description of the Dataset

We remark that the dataset was generated and shared by Dr.Nichols, Dr.Watnik, and Dr.Doster from optical science division in Naval research laboratory with Dr.Rohde's laboratory at UVA.

Three mode sets, which consist of 5 mode numbers, were configured to see whether a particular set is more detectable than the other. Table. 5.3 describes the mode numbers for each set. The sets are all approximately centered around $m = 0$, but contain different spacing between adjacent modes. Increasing the spacing between modes in the encoding set diminishes the effects of mode coupling. However, because including higher mode numbers is inevitable, the OAM beams interfere with turbulence field more due to wider beam diameter.

Five mode numbers in each set, when multiplexed, generated 32 (2^5) distinctive mode patterns. The multiplexed patterns of set 1, 2, and 3 are visualized in Fig. 5.20, Fig.5.21, and Fig. 5.22 respectively. Each mode set generates different multiplexing patterns, where patterns for set 3 appears to be coarser than that for set 1.

Each pattern was transmitted and recorded 1000 times while being interfered with one of 1000 phase screens that simulated the effect of random turbulence. The phase screens were kept same across all patterns. Pattens transmitted from phase screens indexed by 1 to 850 were used for the training set, and the remainder was used for the testing set.

Turbulence Level	Very Strong	Strong	Medium
D/r_0	15	10	5

Table 5.2: Simulated Turbulence Levels

The demultiplexing performance was tested under 3 turbulence conditions – medium , strong, very strong – to test how robust the classification methods are under different level of atmospheric turbulence. The OAM beams propagating through different level of turbulence are shown in Fig. 5.26, Fig. 5.27, and Fig. 5.28 for set 1, set 2, and set 3 respectively.

Set #	Modes (m)
1	{ -7, -2, 3, 8, 13}
2	{ -4, -1, 2, 5, 8}
3	{-3, -1, 2, 4, 6 }

Table 5.3: Mode Sets Used in Experiment

5.4.3 Experimental Procedures

The experiment was designed to evaluate how accurately different classifications systems can decode demultiplexed patterns distorted by the turbulence. In detail, we were interested whether the classification system could be applicable in the presence of severe turbulence. Three levels of turbulence, $D/r_0 = [5, 10, 15]$, corresponding to medium, strong, very strong turbulence, were tested. Also, we investigated whether multiplexing with certain mode set is more favorable performance-wise, i.e. whether it is more easily decoded by the classifier, or less susceptible to the turbulent atmosphere. Three mode sets were compared. In addition, we examined the advantage of the Radon-CDT transform in designing simple classifications system, bearing in mind that the demultiplexing system should be computationally cheap to achieve high data rate.

We trained a 32-class classifier to demultiplex 5 OAM carrying beams for fixed mode set and the turbulence level (unless stated otherwise). The training set consisted of 850 images per pattern, and the testing set consisted of 150 images per pattern, which have a dimension of 151×151. The Radon transform[1] and the Radon-CDT were applied to normalized images (i.e. $\int\int I(x, y)dxdy = 1$), using equidistant, N projections, i.e. θ, the angle between the projection line and the y-axis, is configured to

[1]Radon space was tested to support the claim that classification performance gained in Radon-CDT space was not due to the Radon transform.

be ($\theta = \left[0, \frac{180}{N}, \frac{180}{N}2, \cdots, \frac{180}{N}(N-1)\right]$). Example images for Radon-CDT data are shown in Fig. 5.23, Fig. 5.24, and Fig. 5.25 for set 1,2, and 3 respectively.

Performance Measures

We evaluated the performance of classifiers on their ability to decode symbols and the computation complexity. The demultiplexing power is measured by the classification accuracy and the bit error rate (BER). The computation complexity is determined through the complexity of the classifier and the number of floating point operations required at the testing phase. (The symbols represent; I_n: multiplexed pattern, ℓ_n: label, N: number of samples)

1. Classification accuracy (%):

$$\text{Acc} = \frac{\sum_{n=1}^{N} 1\left(\hat{\ell}_n = \ell_n\right)}{N} \times 100,$$

   where ℓ_n is the true label of pattern I_n, $\hat{\ell}_n$ is the predicted label, $1(\hat{\ell}_n = \ell_n)$ is an indicator function – 1 if $\hat{\ell}_n = \ell_n$ and otherwise 0, and N is the number of samples considered.

2. Bit error rate (BER):

$$\text{BER} = \frac{\sum_{n=1}^{N}\sum_{i=1}^{5} 1\left(\hat{p}_{ni} = p_{ni}\right)}{\sum_{n=1}^{N}\sum_{i=1}^{5} \hat{p}_{ni}} \times 100,$$

   where p_{ni} is the true mode indicator of pattern I_n – 1 if the pattern I_n consists of mode m_i, and otherwise 0– and $\hat{p}_{ni}$ is the predicted mode indicator. In this BER formulation, each OAM carrying beam represents '1', and therefore the OOK modulation is inherently assumed.

3. the number of trainable parameters in classifiers,

4. the number of total floating point operations (FLOPs).

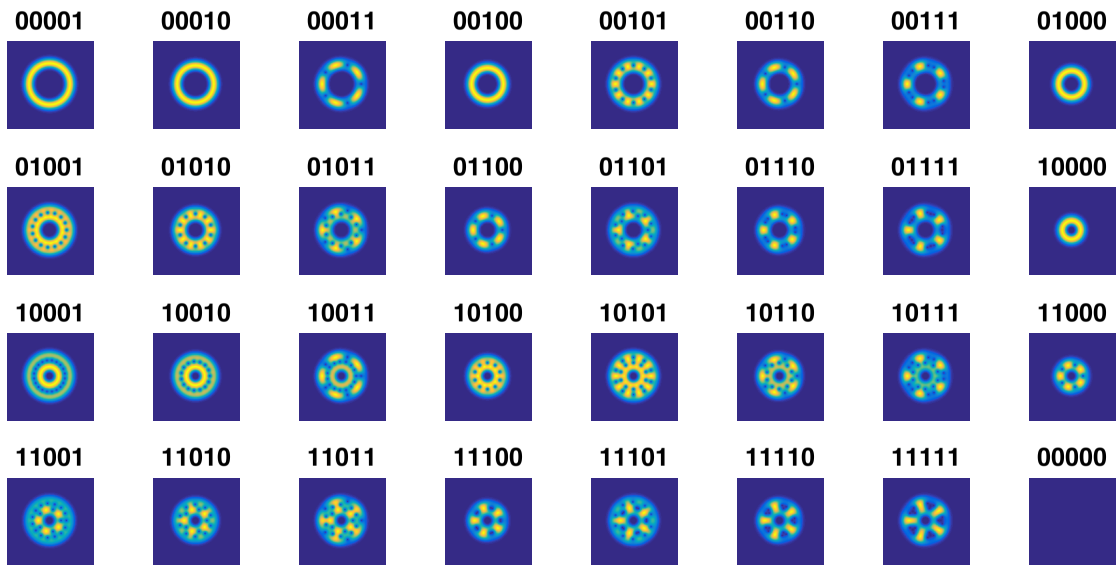


Figure 5.13: Multiplexed Patterns of Laguerre-Gauss Beam

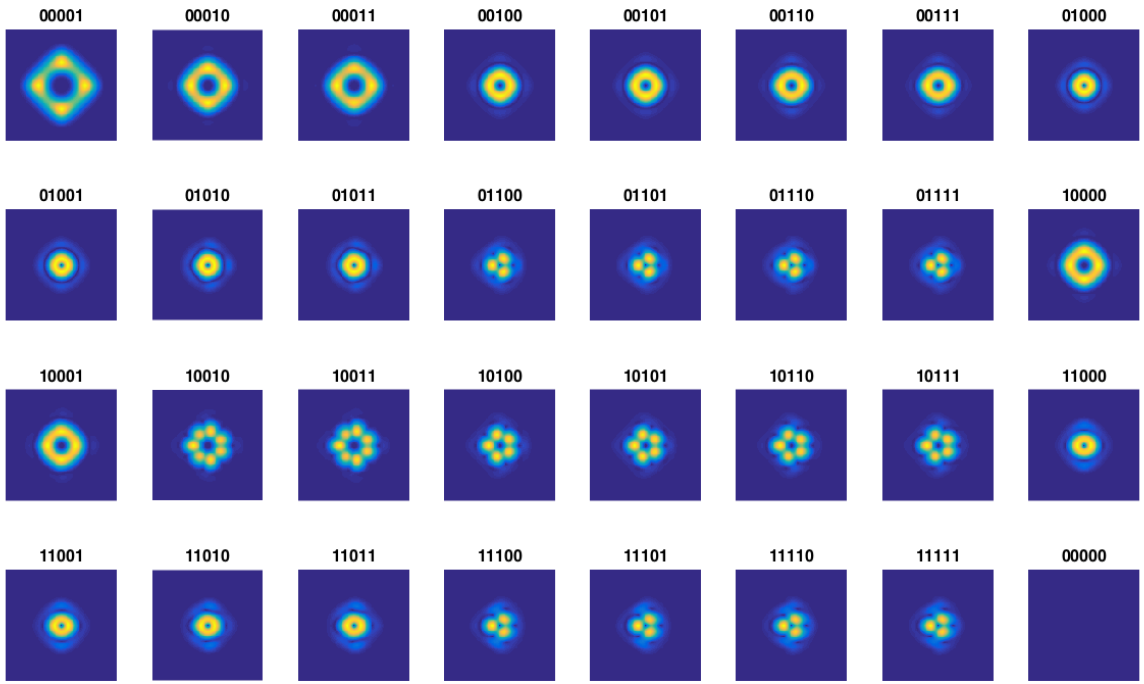


Figure 5.14: Multiplexed Patterns of Bessel-Gauss Beam, set 1

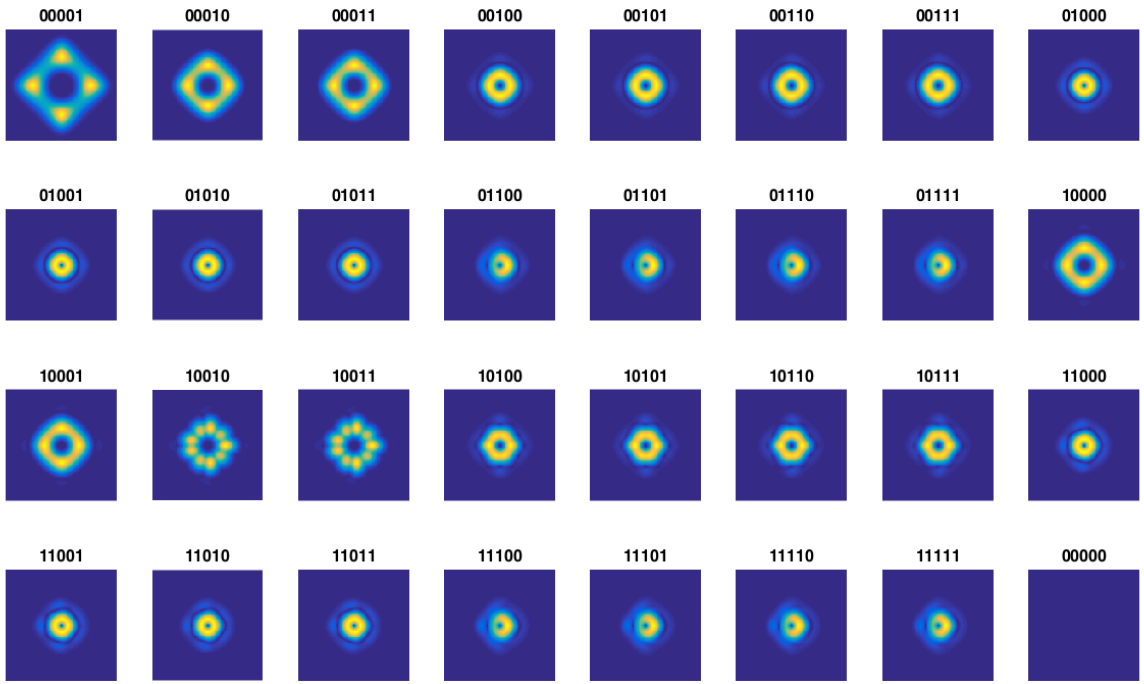


Figure 5.15: Multiplexed Patterns of Bessel-Gauss Beam, set 2

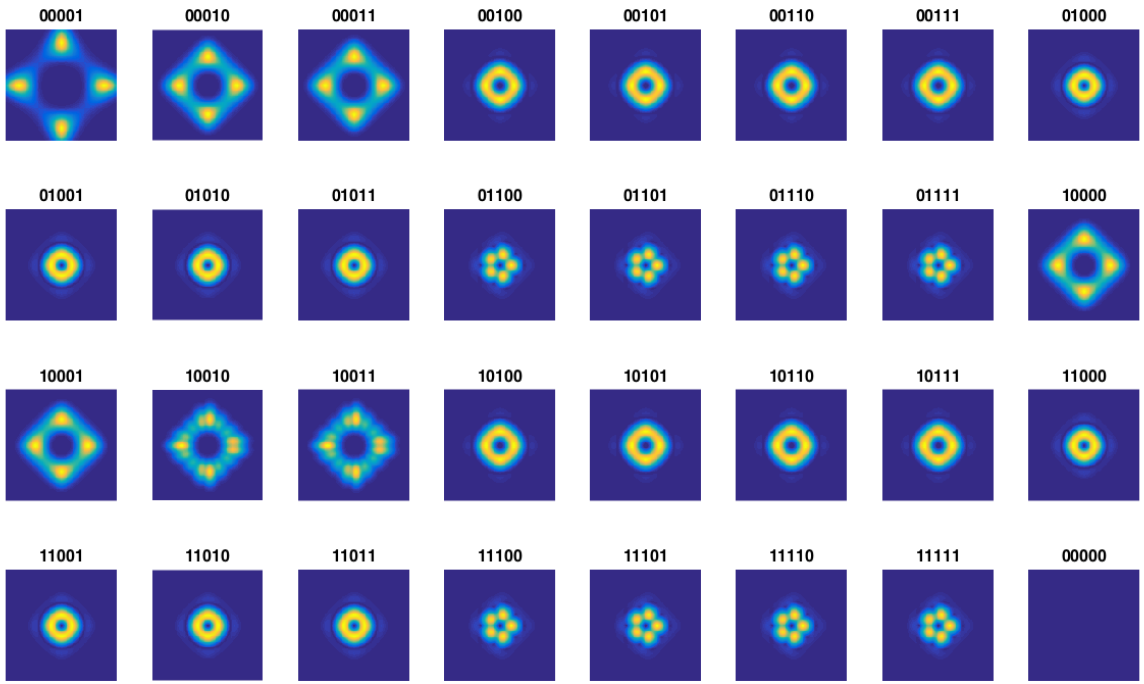


Figure 5.16: Multiplexed Patterns of Bessel-Gauss Beam, set 3

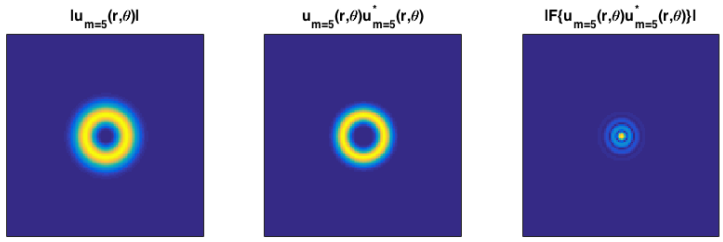


Figure 5.17: Conjugate Mode Sorting With LG Beam, n=5 transmitted, m=5 used for detection

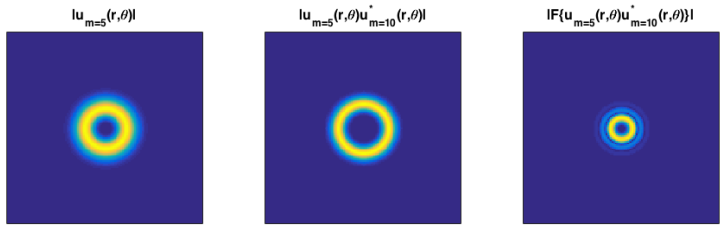


Figure 5.18: Conjugate Mode Sorting With LG Beam, n=5 transmitted, m=10 used for detection

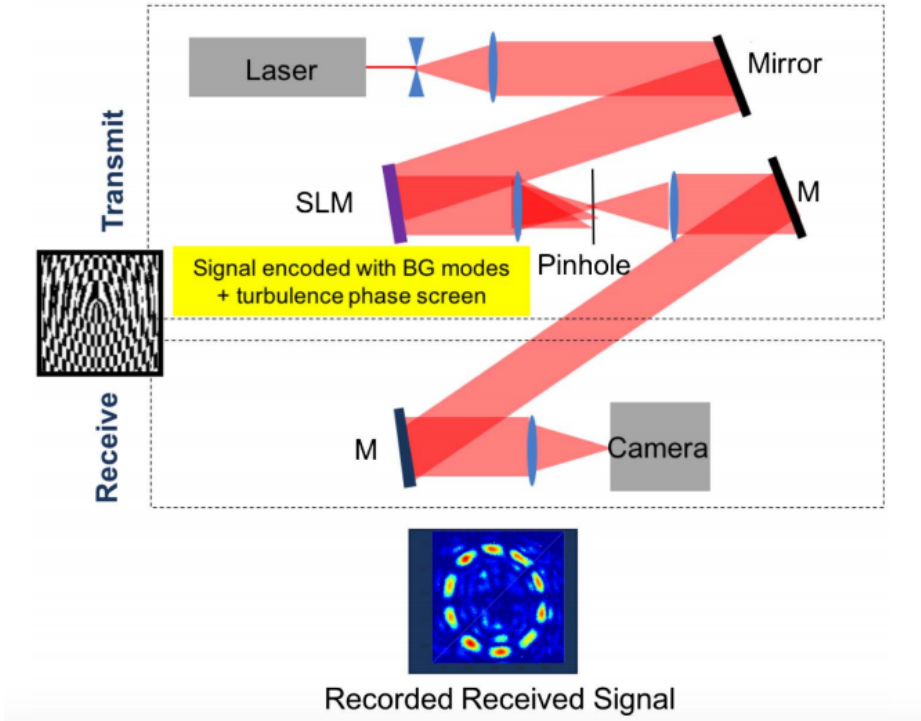


Figure 5.19: Experiment Diagram

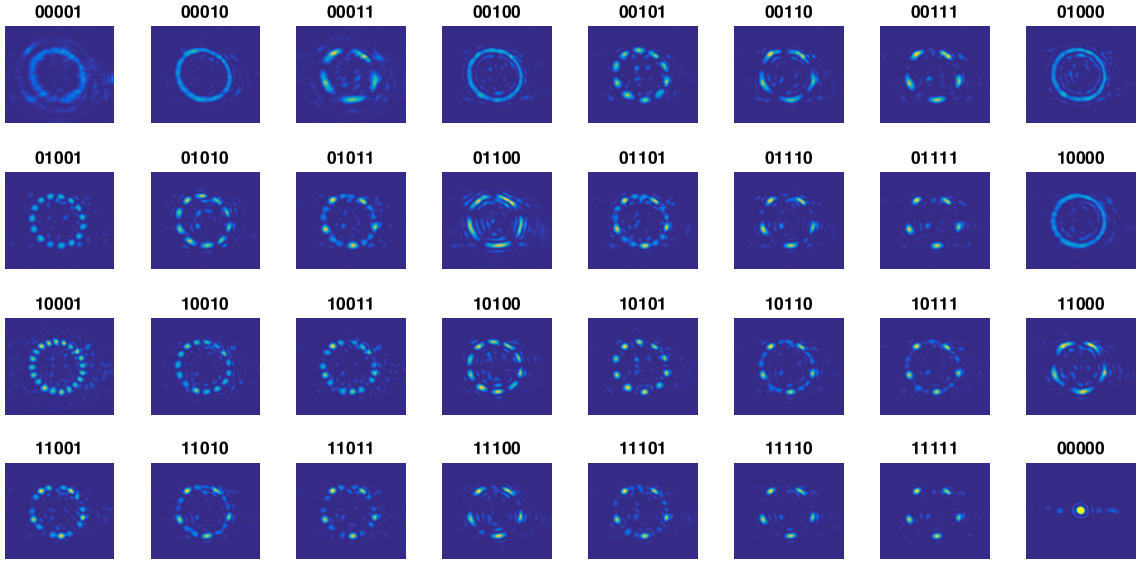
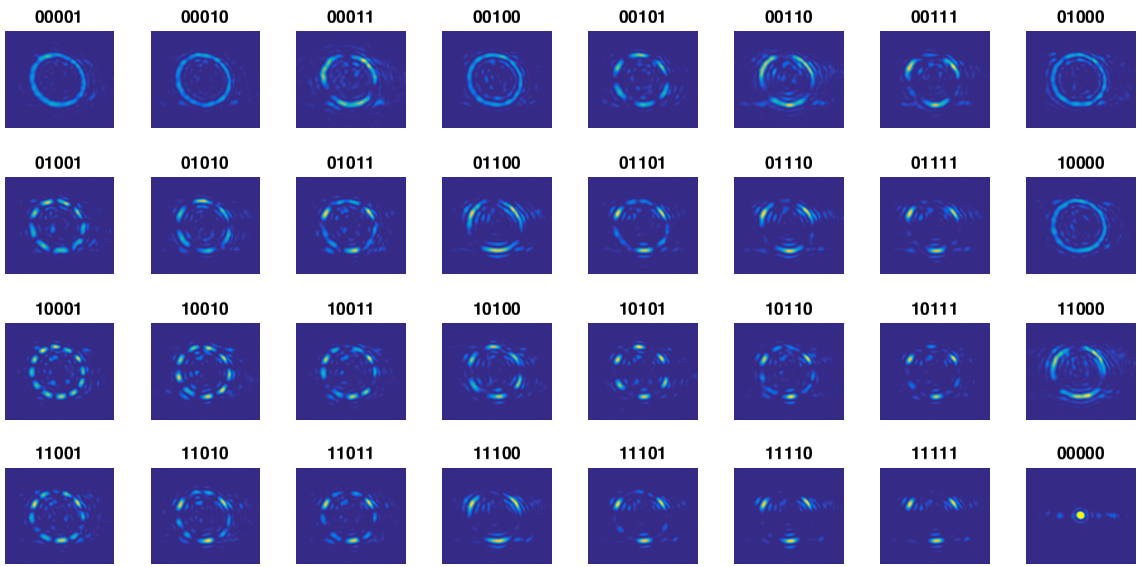


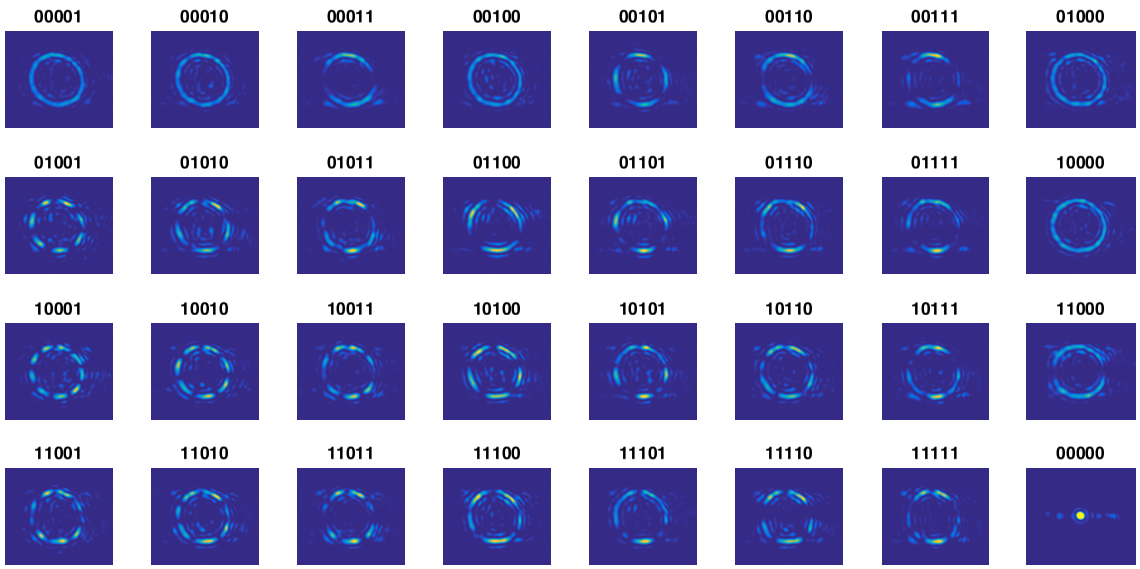
Figure 5.20: Mode Set 1

Bit string	00001	00010	00100	01000	10000
Mode #	13	8	3	-2	-7



Bit string	00001	00010	00100	01000	10000
Mode #	8	5	2	-1	-4

Figure 5.21: Mode Set 2



Bit string	00001	00010	00100	01000	10000
Mode #	6	4	2	-1	-3

Figure 5.22: Mode Set 3

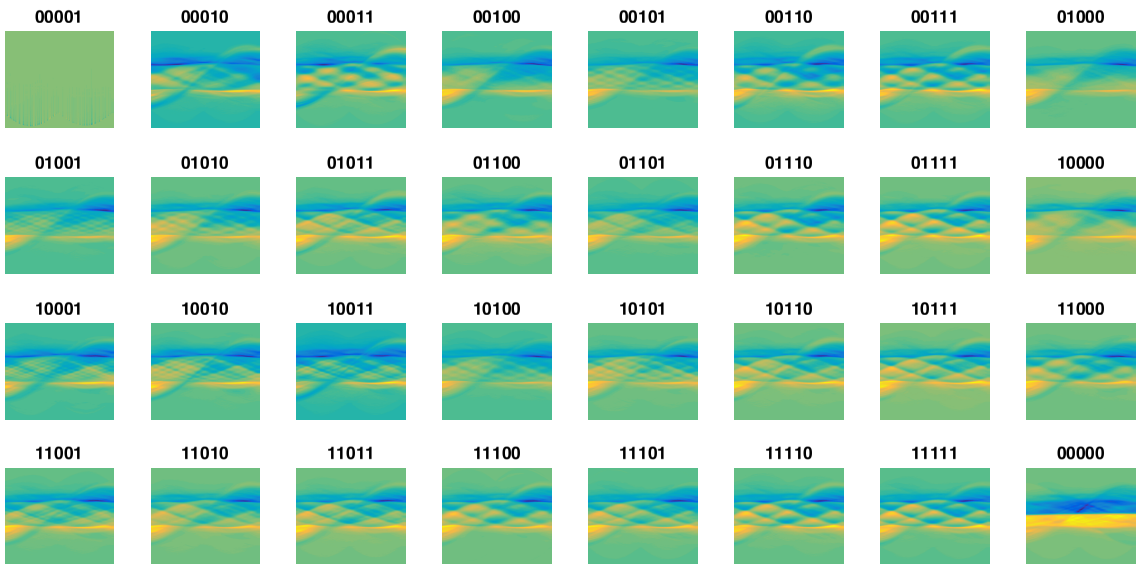
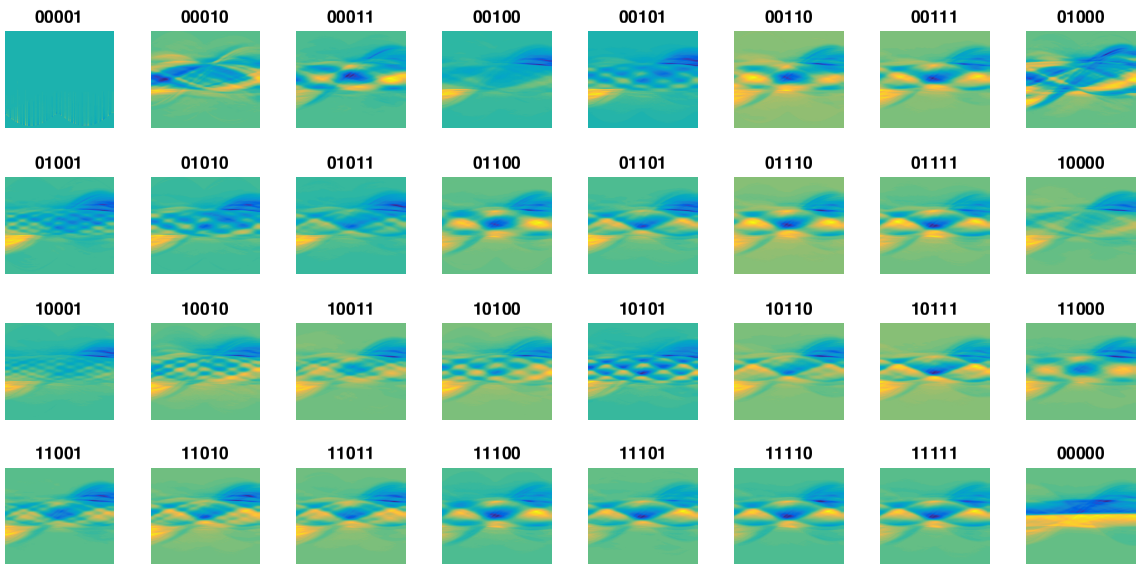


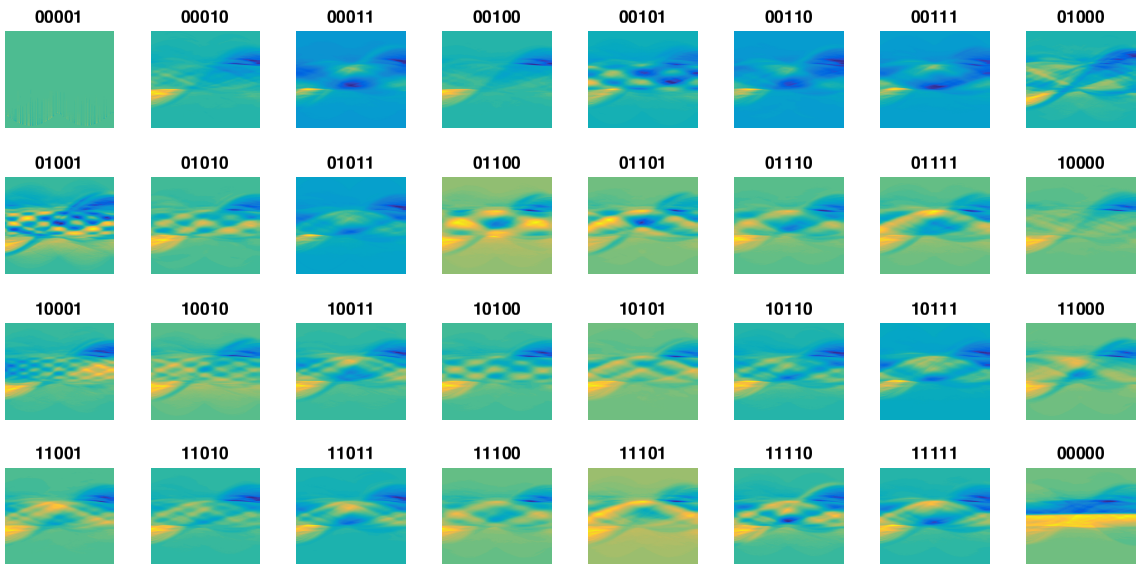
Figure 5.23: Mode Set 1 Radon CDT

Bit string	00001	00010	00100	01000	10000
Mode #	13	8	3	-2	-7



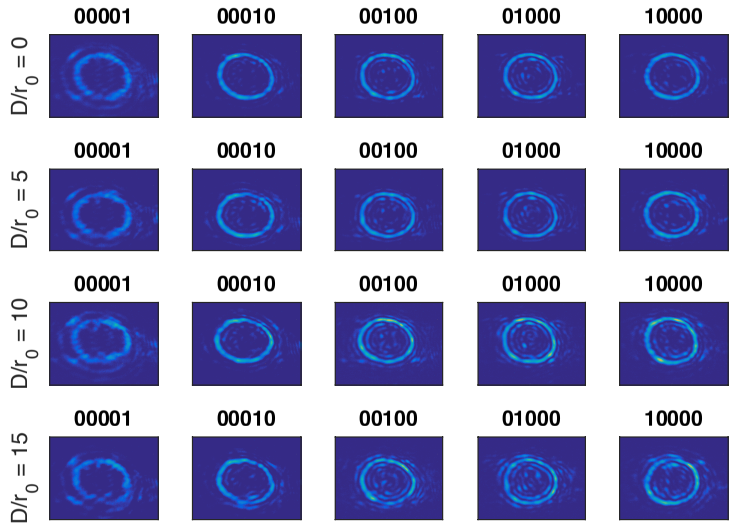
Bit string	00001	00010	00100	01000	10000
Mode #	8	5	2	-1	-4

Figure 5.24: Mode Set 2 Radon CDT



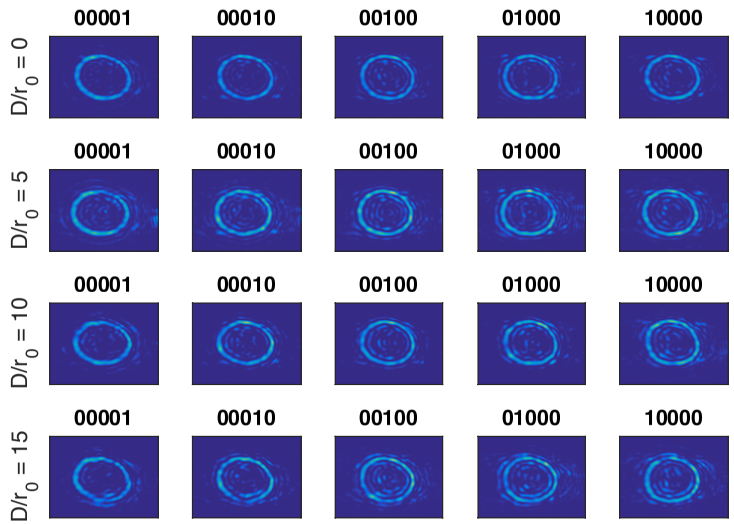
Bit string	00001	00010	00100	01000	10000
Mode #	6	4	2	-1	-3

Figure 5.25: Mode Set 3 Radon CDT



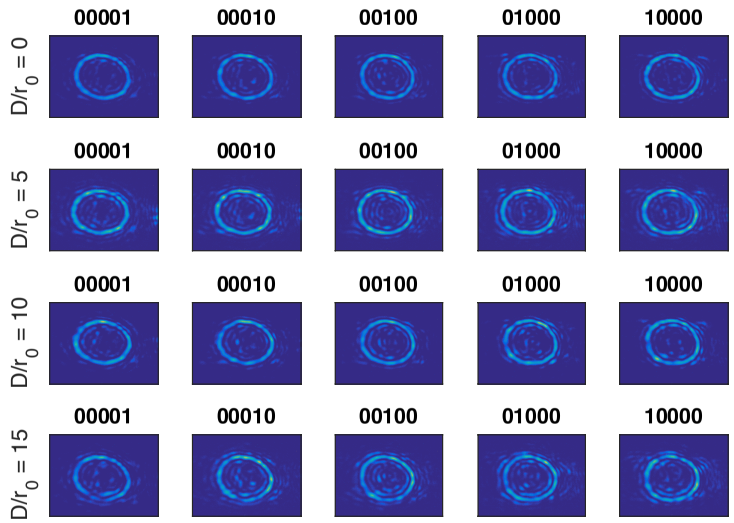
Bit string	00001	00010	00100	01000	10000
Mode #	13	8	3	-2	-7

Figure 5.26: Fundamental Mode Set 1 under Different Turbulence Levels



Bit string	00001	00010	00100	01000	10000
Mode #	8	5	2	-1	-4

Figure 5.27: Fundamental Mode Set 2 under Different Turbulence Levels



Bit string	00001	00010	00100	01000	10000
Mode #	6	4	2	-1	-3

Figure 5.28: Fundamental Mode Set 3 under Different Turbulence Levels

5.5 Experimental Results

Our study was conducted with several aims. The foremost goal was to design a classification system that can accurately demultiplex OAM beam patterns deformed by atmospheric turbulence. For reliable communication, we aimed for the BER (bit error rate) $\sim 1.0e^{-3}$, which can be combined with the current out-of-shelf error correcting codes such as LDPC (low-density parity-check code) to achieve coding gain about $5dB$ [94]. [95] also showed that in the presence of turbulence, LDPC coding is sufficient to decrease BERs within the threshold of threshold of $3e^{-2}$. Therefore, we set the maximum tolerable BER to be $3e^{-2}$ for our system.

Also, bearing in mind that the demultiplexing system should achieve high data rate, we aimed to design computationally fast and cheap system, which can operate with the lesser number of FLOPs (floating point operations).

In addition, we investigated whether multiplexing with certain mode set is more favorable performance-wise. Three mode sets were compared; the set #1 had the most gap between the mode numbers and hence included the highest mode number, and set #3 had the least separation between the mode numbers. We also examined the potential possibility of including more mode numbers to a set, i.e. to multiplex more than 5 OAM carrying beams. Lastly, we examined the robustness of the classification system under several adverse effects, such as spatial downsampling – due to low-resolution CCD camera – and beam-wandering – due to turbulence and misalignment between the laser and receiver.

Our experiment revealed that demultiplexing 5 OAM carrying beams could be performed reliably under severe turbulence. In conjunction with the Radon-CDT, 1-layer CNN system achieved BER near $1e^{-3}$ regardless of the turbulence level. The experiment with enlarged mode set revealed that potentially 15 or more OAM carrying beams could be multiplexed without compromise in BER. Also, we validated that the Radon-CDT was also robust to spatial down-sampling and beam wandering.

Data Space	dim	$\#\theta$	$D/r_0 = 5$			$D/r_0 = 10$			$D/r_0 = 15$		
			Set 1	Set 2	Set 3	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
Image	4000	-	98.50	99.48	99.71	83.37	87.78	90.15	58.71	79.08	84.27
RadonCDT	4000	180	99.29	99.96	100	98.19	95.98	97.58	97.1	95.31	98.85
Radon	3906	18	98.5	99.48	99.75	90.05	91.64	93.54	73.88	82.94	87.88
RadonCDT	3906	18	98.98	99.95	99.95	96.81	98.23	99.21	94.73	94.58	98.27

Table 5.4: Linear Classification Accuracy for Testing Sets

5.5.1 Linear Classification

In a vacuum, an EM wave will propagate through a straight path, or the optimal transport path (more details can be found in Section. B.3). If the mixing of air is present causing turbulence, EM wave will wander locally, but still be approximated by the straight path defined by the optimal transport map. Each instance of EM wave received under the random effect of turbulence can be associated with its optimal transport map. The maps affect the transmitted beam regardless of the type of the multiplexed pattern and will break down orthogonality at the receiver. We hypothesize that the collection of optimal transport maps belongs to the map of diffeomorphisms that satisfy the conditions stated in Section 3.4.

The first experiment was conducted to validate the claim that the received patterns can be linearly decoded in the Radon-CDT space. The linear discriminant analysis (LDA) was chosen as the linear classifier. We compute the linear classifier and compare the linear classification accuracy between image space and Radon-CDT space.

Computing LDA involves inverting a data matrix, and for 32000 images of size 151×151, this is practically infeasible with a standard computer given the storage requirement and the cost of diagonalization. Therefore, the data size was reduced to 4000 using PCA and then to 31 using Fisher-LDA. The entire training set was used to compute the PCA subspace since PCA doesn't require label information. However, the Fisher-LDA finds the discriminant subspace based on the label. 80% of the training set was used to compute the Fisher-LDA subspace, and the 20% of the training set was allocated as a validation set. The LDA classifier was trained using only this 80% of the training set.

Table. 5.4 shows the classification accuracy of LDA classifiers on different input

space (TR: training, VAL: validation, and TS: testing) of various levels of turbulence. Among all input space considered, Radon-CDT space consistently showed the highest demultiplexing accuracy, with the least 'subspace' overfitting and the 'classifier' overfitting. High linear classification accuracy in the image space for $D/r_0 = 5$ suggests that the patterns are linearly separable. The advantage of utilizing Radon-CDT becomes more prominent under severe turbulence, indicating that the Radon-CDT transform decoded out the deformation from the OAM beam patterns caused by turbulent atmospheric effects.

We demonstrated here that Radon-CDT could enhance linear classification. To achieve lesser BER, however, we adopted a shallow convolutional neural network to add non-linear feature extraction in prior to the linear classification.

5.5.2 Non-linear Classification: 1-layer Convolution Network

To improve upon the linear classification of the Radon-CDT data, a 1-layer convolutional neural network was chosen to exploit an additional non-linearity. With the softmax function at the final output layer and the cross entropy loss, CNN can be described as a concatenation of non-linear mapping plus a linear classifier. 1-layer CNN is specifically chosen for two folds: i) to learn a dimension reduction mapping that wouldn't overfit as much as PCA and ii) to boost the classification accuracy by adding a slight non-linearity to the classifier in addition to ones accounted by the Radon-CDT.

Our 1-layer CNN consists of a convolution layer, a batch normalization layer, a max-pooling layer, and a reLu activation layer, followed by a softmax layer. The cross-entropy loss was used, and the CNN was minimized via stochastic gradient descent with Adaptive Moment Estimation [96]. The exact configurations are shown in Table.5.8.

Table. 5.5 shows demultiplexing accuracy for the TS set for Radon-CDT space (TR set accuracy is omitted because they were all 100%), for the highest turbulence level ($D/r_0 = 15$), utilizing 90 projections in Radon-CDT. Compared to the linear classification accuracy, providing a non-linear feature extraction using convolutional layer helped in improving the performance. BER is also measured, which lie in the

D/r_0	Acc (%)			BER		
	set 1	set 2	set 3	set1	set 2	set 3
5	99.50	100	99.75	0.002625	0	0.001917
10	99.91	99.67	99.75	0.000917	0001583	0.0015
15	99.23	99.17	99.60	0.004167	0.003958	0.002250

Table 5.5: 1-Layer CNN Performance in the Radon-CDT space for Testing Sets, $\theta = 90$

D/r_0	Acc (%)			BER		
	set 1	set 2	set 3	set1	set 2	set 3
5	99.90	99.94	99.92	0.00042	0.00025	0.00004
10	99.63	99.69	99.56	0.00150	0.00125	0.00200
15	99.30	99.43	99.48	0.00183	0.00489	0.00258

Table 5.6: Alexnet Performance in the Image Space

	Conv 1, # θ : 90	Alexnet
# Params	550.4 K	69M
# Flops	9.11 M	832 M

Table 5.7: Computational Complexity

conv1	# filter : 96	Size: 11$\times$ 5	Strides:3$\times$3
Max pooling		Size: 3$\times$3	Strides: 2$\times$2
Batch Normalization			
Relu			
Softmax			

Table 5.8: 1 Layer CNN Architectures

margin of the safe zone ($\sim 1e^{-3}$) for reliable communication when additional error correcting code schemes are taken into account.

It is notable that decoding in Radon-CDT space provides comparable results compared to the accuracy and BER reported previously in [75], shown in Table. 5.6. The authors utilized Alexnet for decoding, which is about 100 the size of the network we implemented here and requires about 100 times more Flops to decode a single image.

In summary, the non-linearity introduced into the data by atmospheric turbulence effect, that can be not decoded linearly in Radon-CDT space, can be reliably decoded utilizing a simple non-linear feature extractor (a convolutional layer).

5.5.3 Generalization to Demultiplexing with Low Resolution Images

In this experiment, we test whether we can share the same classifier on the channel that is affected by different levels of turbulence. We combined the non-turbulent and turbulent datasets for each mode set, and trained the classifier.

Table. 5.9 shows the corresponding classification accuracy and BER. The BER of each set is almost similar or even better to that of the BER of worst turbulence type. We assume that during the optimization process, the CNN classifier converged to the better solution due to increased number of samples.

Through this experiment, we could confirm that that the CNN classifier can be generalized to decode multiplexed OAM patterns regardless of the strength of the turbulence.

-	set 1	set 2	set 3
BER	0.0029	0.0029	0.0033
Acc	99.32	99.30	99.48

Table 5.9: Results for mixed turbulence sets

5.5.4 Demultiplexing with Low Resolution Images

Downsample Factor	Data Space	Size	Data Space	Size
16		9×9		15×9
8	Image	19×19	R-CDT	29×19
4		38×38		57×38

Table 5.10: Size for down-sampled data

Here we investigated whether we can utilize low-resolution image for demultiplexing. The images were down-sampled by a factor of $\alpha = [1/4, 1/8, 1/16]$ followed by the Radon-CDT transform. Fig. 5.29 shows the down-sampled images (top) and Radon-CDT data (bottom). Table. 5.10 summarizes the size of the down-sampled data. Two classifiers were tested: a linear classifier (LDA) and a 1-layer regular neural network (NN). We note that the reduced image size enabled to train both classifiers without

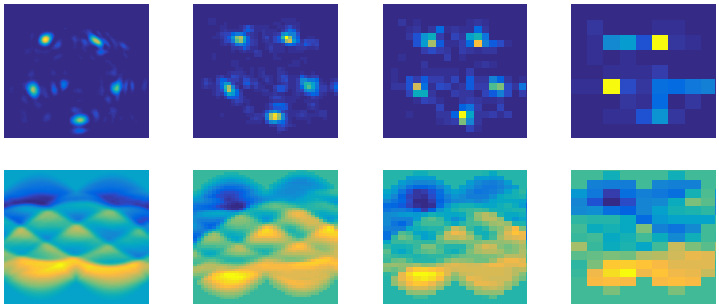


Figure 5.29: Downsampled image (top) and R-CDT (bottom) by factor of [1, 4, 8, 16] from left to right

Downsampl e Factor	Data Space	set 1		set 2		set 3	
		TR	TS	TR	TS	TR	TS
16	Image	36.27	44.19	29.14	37.54	23.47	29.65
8		6.68	13.29	6.36	12.00	4.42	7.46
4		3.44	13.19	3.43	11.81	2.24	7.71
16	R-CDT	34.54	42.62	28.14	36.40	20.62	26.50
8		2.21	4.69	2.96	1.55	1.2	2.1
4		0.28	2.40	0.19	1.31	0.22	0.96

Table 5.11: LDA Classification Results for down-sampled testing set, $D/r_0 = 15$

Downsample Factor	Data Space	# Dense Nodes	# Params	set 1	set 2	set 3
16	Image	5000	580k	13.09	8.36	5.19
8		1500	594k	3.29	5.57	4.05
4		400	591k	3.42	6.07	3.17
16	R-CDT	3000	510k	18.51	13.23	8.82
8		1000	588k	2.56	1.63	0.83
4		250	550k	2.94	4.46	0.94

Table 5.12: NN Classification Results for down-sampled testing set, $D/r_0 = 15$

any computational issues. The number of dense nodes in NN was devised to render the network size consisting of approximately 500k parameters.

Table. 5.11 shows the classification performance for LDA. For a down-sample factor of 4 and 8, classification performance was not significantly compromised by reducing the size of images. In fact, in the image space, training in reduced dimension facilitated higher linear classification accuracy than using full dimension images. Yet, again, Radon-CDT yielded consistently higher accuracy. Table. 5.12 shows the clas-

-	Radon-CDT set 1+3	Radon-CDT set 1+2+3
BER	0.002962	0.003941
Acc	99.35	99.15

Table 5.13: Results for larger mode set

$Data$	/w Beam Wandering			/wo Beam Wandering		
	set 1	set 2	set 3	set1	set 2	set 3
Image	48.33	38.44	47.81	65.43	64.14	68.06
Radon-CDT	98.64	98.50	98.71	98.58	98.62	98.79

Table 5.14: Classification Accuracy for Testing Set for 1-Layer CNN, $D/r_0 = 15, \alpha = 0.2$

sification performance for NN. As similar to the previous result with 1-layer CNN, adding non-linear feature extraction step enhanced classification performance. For a down-sample factor of 4 and 8, the demultiplexing accuracy is nearly similar to full resolution images. We conclude that demultiplexing using low-resolution images can greatly reduce the computation cost without damaging the performance.

5.5.5 Demultiplexing with Larger Mode Sets

In this experiment, we test whether we can share the same detector on the channel when mode sets are all used to communicate data. We train a detector for two scenarios: i) when mode set 1 and set 3 are used to transmit data over the same channel, and ii) when all three mode sets are used to transmit data. When mode set 1 and three are used, there are 63 (32+32-1, minus 1 for '00000' case) distinctive patterns, and when mode set 1, 2, three are used, there are 92 (discarding overlapping multiplexing cases) distinctive patterns.

Table. 5.13 shows the classification accuracy and BER. There exists a performance drop compared to the single mode set performance, but considering that the number of classes is doubled or tripled, the drop is marginal. Therefore, we conclude that a single detector can be utilized when different mode sets are used for multiplexing/modulating data over the same communication channel.

5.5.6 Robustness to Beam Wandering

The OAM carrying beam, generated in the laboratory, propagated through a pre-defined path. The beam patterns were collected by a fixed camera that heads towards the mirror, whose center was matched to the center of the beam. However, if we were to transmit the visible lights via air, beam wandering – due to atmospheric effect but also by inherent jitter present in a laser – is unavoidable, and the patterns captured by the camera would not be aligned properly. In this experiment, we test whether our classification system can robustly demultiplex off balance OAM patterns, using 1-layer CNN in the radon-CDT space.

To mimic the effect of beam wandering, the images were randomly translated by $[\Delta x, \Delta y]$, where Δx and Δy was independently drawn from a uniform distribution in a range $[-151\alpha, 151\alpha]$, where $\alpha \in [0, 1]$ controls the severity of the beam wandering. The mentioned preprocessing was performed on-line, while training the Convolutional Neural Network, which resulted in augmenting the dataset. Note that the translations in the image space convert to the vertical shifts in the Radon-CDT space by the Translation Property Stated in (3.4). The Radon-CDT under the effect of beam wandering can, therefore, be computed on-line without the additional cost of computation given that the Radon-CDT before the translation is already computed.

Table. 5.14 shows the classification accuracy for the testing set in the Radon-CDT space and the image space. Two types of testing sets were tested: i) the original testing set without the effect of beam-wandering and ii) the augmented testing set with the effect of beam-wandering.For both types of testing sets, in the Radon-CDT space, the CNN successfully decoded the multiplexed patterns, whereas, in the image space, it failed to. The Radon-CDT removed the wandering effect from the image and therefore reduced the task of 'shallow' CNN substantially, rendering CNN to classify the patterns correctly. However, in the image space, the shallow CNN could not decode the translation and the turbulent confounds at the same time. We hypothesize that this is due to the linear separability property of the CDT which accounts for the translation confounds and mode them out in the transform space.

In summary, in the Radon-CDT space, the classifier is robust to beam-wandering

effects, and also provides an additional advantage of training better classifier.

Chapter 6

Conclusion

In this study, we have described a new nonlinear operation, termed the Cumulative Distribution Transform (CDT), that takes as input signals that can be understood as probability density functions, and outputs a continuous function that is related to morphing that signal to a chosen reference signal. Also, we extended the CDT by combining the 2D Radon transform, to a new, non-linear, low-level image transform, termed the Radon-CDT.The transforms are invertible as it contains well-defined forward (analysis) and inverse (synthesis) operations.

In addition to describing a few of its properties, we have extensively studied the ability of the transforms to improve the linear separability in comparison to the linear separability in original signal space. We experimentally validated with signal and image applications involving both simulated and real data. In all examples shown, the results of the Theorem are confirmed, and the theory and experimental results here add to our understanding in explaining why transport-based approaches have been able to improve the state of the art in certain cancer detection from microscopy images problems [38, 39].

The CDT is cheap to compute. We described a numerical approximation for discrete signals that is $O(n\log(n))$, with n the length of the signal. The Radon-CDT, which is built upon the CDT, also has a closed form, and hence does not require numerical optimization for computation. Its computational efficiency combined with the-

oretical and experimental results presented above suggest that the transforms could be a useful tool for building more complex signal pattern recognition systems.

Moreover, guided by newly developed theory suggesting a link between image turbulence and photon transport through the continuity equation, we utilized the transform to perform a decoding task for orbital angular momentum carrying beam patterns. The fact that the transforms are a mathematically invertible transform ensures that no information will be lost in this step, and we hypothesized that the transforms to be a useful pre-processing step in the search for solutions. The decoding technique was tested in the Radon-CDT space and was compared against previous approaches using deep convolutional neural networks. Results showed that the new method could obtain comparable classification accuracies (bit error rate) at a fraction of the computational cost, thus enabling higher bit rates.

Our study was conducted with several aims. The foremost goal was to design a classification system that can accurately demultiplex OAM beam patterns deformed by atmospheric turbulence. Our experiment revealed that the demultiplexing could be performed reliably under severe turbulence achieving BER $\sim< 1.0e3$ for medium or strong turbulence, and $\sim\ 1.0e-3$ for very severe turbulence. Also, we aimed to design fast and cheap method, and the advantage of using Radon-CDT space was clear, it enabled simpler representations in the Radon-CDT space, that facilitated better classification using simple 1-layer CNN. Also, we investigated whether multiplexing with certain mode set is more favorable, and concluded that although the difference is minimal, the set #3 is more robust to classification. The mode set #3 has the least gap between the modes, and therefore is more susceptible to mode coupling, but less perturbed by turbulence.

Provided that the light propagation path can be explained by the transport phenomena, we hypothesized that certain deformations could be (nearly) linearly decoded in the Radon-CDT space, and showed empirical evidence. In the presence of turbulence, the light would only approximately follow the optimal transport path. With the aid of non-linear feature extractor (1-layer CNN), which decoded out the confounds introduced by the turbulence, the OAM patterns are decoded near perfectly in the Radon-CDT space. We omitted the analysis of the light propagation and its relation to the

optimal transport for brevity in previous chapters, but for interested readers, we included the details in Appendix B.

The main limitation of the CDT and Radon-CDT model is that the linear separability properties depend on the signals being generated from mother signals through the application of a differential, one to one diffeomorphic mapping with additional restrictions. In certain cases, a physical model for the data can help determine whether the conditions for linear separability in the transform space are applicable. In some cases, the transform can indeed be a poor match for the problem. In such cases, the transform can still be applied with no loss of information, though we currently offer no information regarding whether the CDT would enhance (or help destroy) linear separability. The variety of examples shown above, however, have helped us confirm the model is applicable, at least to some extent, to not an insignificant number of applications. We envision that the transforms could be used as a step in pattern recognition pipeline that could simplify (and enhance the performance) subsequent feature extraction and classification.

Appendix A

Proofs

A.1 Proof for translation property of CDT

Consider a probability density $I_1 : [y_1, y_2] \to \mathbb{R}$, and let $I_\mu : [y_1 + \mu, y_2 + \mu] \to \mathbb{R}$ represent a translation of the probability density I_1 by μ, i.e. $I_\mu(x) = I_1(x - \mu)$. To find the CDT for I_μ with respect to the reference probability density $I_0 : X \to \mathbb{R}$, we solve for $f_\mu : X \to [y_1 + \mu, y_2 + \mu]$:

$$\int_{y_1+\mu}^{f_\mu(x)} I_\mu(\tau)d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau = x. \tag{A.1}$$

And similarly, to find the CDT for I_1 with respect to the reference I_0, we solve for $f_1 : X \to [y_1, y_2]$:

$$\int_{y_1}^{f_1(x)} I_1(\tau)d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau = x. \tag{A.2}$$

(A.1) and (A.2) can be set equal,

$$\int_{y_1+\mu}^{f_\mu(x)} I_\mu(\tau)d\tau = \int_{y_1}^{f_1(x)} I_1(\tau)d\tau. \tag{A.3}$$

By substituting I_1 for I_μ in (A.3), we have

$$\int_{y_1+\mu}^{f_\mu(x)} I_1(\tau-\mu)d\tau = \int_{y_1}^{f_1(x)} I_1(\tau)d\tau. \tag{A.4}$$

By the change of variables theorem, we can substitute $u = \tau - \mu$ in (A.4)

$$\int_{y_1}^{f_\mu(x)-\mu} I_1(u)du = \int_{y_1}^{f_1(x)} I_1(\tau)d\tau.$$

Since upper limit on left and right side of the integrals are equal, we have $f_\mu(x) = f_1(x) + \mu$. Substituting this into expression for $\widehat{I}_\mu(x) = (f_\mu(x) - x)\sqrt{I_0(x)}$, we have

$$\widehat{I}_\mu(x) = (f_1(x) + \mu - x)\sqrt{I_0(x)}.$$

By substituting $\widehat{I}_1(x) = (f_1(x) - x)\sqrt{I_0(x)}$, we have proved the translation property

$$\widehat{I}_\mu(x) = \widehat{I}_1(x) + \mu\sqrt{I_0(x)}.$$

A.2 Proof for scaling property of CDT

Consider a probability density $I_1 : [y_1, y_2] \to \mathbb{R}$, and let $I_a : [y_1/a, y_2/a] \to \mathbb{R}$ represent a scaling of the probability density I_1 by a, i.e. $I_a(x) = aI_1(ax)$. To find the CDT for I_a with respect to the reference $I_0 : X \to \mathbb{R}$, we solve for $f_a : X \to$: $[y_1/a, y_2/a]$:

$$\int_{y_1/a}^{f_a(x)} I_a(\tau)d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau. \tag{A.5}$$

And similarly, to find the CDT for I_1 with respect to the reference I_0, we solve for $f_1 : X \to [y_1, y_2]$:

$$\int_{y_1}^{f_1(x)} I_1(\tau)d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau \tag{A.6}$$

(A.5) and (A.6) can be set equal,

$$\int_{y_1/a}^{f_a(x)} I_a(\tau)d\tau = \int_{y_1}^{f_1(x)} I_1(\tau)d\tau. \tag{A.7}$$

By substituting $I_a = aI_1(ax)$ in (A.7), we have

$$\int_{y_1/a}^{f_a(x)} aI_1(a\tau)d\tau = \int_{y_1}^{f_1(x)} I_1(\tau)d\tau. \tag{A.8}$$

By the change of variables theorem we can substitute $a\tau = u$, $ad\tau = du$ in (A.8),

$$\int_{y_1}^{af_a(x)} I_1(u)du = \int_{y_1}^{f_1(x)} I_1(\tau)d\tau.$$

Since the upper limit on left and right side of the integrals are equal, we have $f_a(x) = \frac{f_1(x)}{a}$. Substituting this expression for $\widehat{I}_a(x) = (f_a(x) - x)\sqrt{I_0(x)}$, and cleaning up some algebras, we get $\widehat{I}_a : X \to \mathbb{R}$:

$$\widehat{I}_a(x) = \frac{\widehat{I}_1(x) - x(a-1)\sqrt{I_0(x)}}{a}.$$

A.3 Proof for composition property of CDT

Let $I_1 : Y \to \mathbb{R}$ represent a probability density, and $J_1 : Y \to \mathbb{R}$ its cumulative distribution function. Let $I_g : Z \to \mathbb{R}$ represent a probability density that has the following relation with I_1:

$$J_g(x) = J_1(g(x)). \tag{A.9}$$

$J_g : Z \to \mathbb{R}$ represent the corresponding cumulative distribution for I_g, and $g : Z \to Y$ is an invertible, differentiable. By differentiating each side of (A.9), we have

$$I_g(x) = g'(x)I_1(g(x)).$$

To find the CDT for I_g with respect to the reference probability density $I_0 : X \to \mathbb{R}$, we solve for $f_g : X \to Z$:

$$\int_{\inf(Z)}^{f_g(x)} I_g(\tau)d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau \tag{A.10}$$

And similarly, to find the CDT for I_1, we solve for $f_1 : X \to Y$:

$$\int_{\inf(Y)}^{f_1(x)} I_1(\tau)d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau \tag{A.11}$$

(A.10) and (A.11) can be set equal,

$$\int_{\inf(Z)}^{f_g(x)} I_g(\tau)d\tau = \int_{\inf(X)}^{f_1(x)} I_1(\tau)d\tau. \tag{A.12}$$

By substituting $I_g(x) = g'(x)I_1(g(x))$ in (A.12), we have

$$\int_{\inf(Z)}^{f_g(x)} g'(\tau)I_1(g(\tau))d\tau = \int_{\inf(Y)}^{f_1(x)} I_1(\tau)d\tau. \tag{A.13}$$

By the change of variables theorem we can substitute $g(\tau) = u, g'(\tau)d\tau = du$ in (A.13),

$$\int_{\inf(Y)}^{g(f_g(x))} I_1(u)du = \int_{\inf(Y)}^{f_1(x)} I_1(\tau)d\tau.$$

Since the upper limit on left and right side of the integrals are equal, we have

$$g(f_g(x)) = f_1(x).$$

Since g is an invertible function, $f_g(x) = g^{-1}(f_1(x))$ holds. By substituting this expression for $\widehat{I}_g(x) = (f_g(x) - x)\sqrt{I_0(x)}$, and cleaning up some algebra, we get $\widehat{I}_g : Z \to \mathbb{R}$:

$$\begin{aligned}\widehat{I}_g(x) &= \left(g^{-1}\left(f_1(x)\right) - x\right)\sqrt{I_0(x)} \\ &= \left(g^{-1}\left(\frac{\widehat{I}_1(x)}{\sqrt{I_0(x)}} + x\right) - x\right)\sqrt{I_0(x)}.\end{aligned}$$

A.4 Proof for Lemma 2.4.4

Proof. (if) The convex hulls of compact convex sets are compact in L^2 space. Therefore, the convex hulls are compact. For disjoint, compact convex sets sets, we know from Lemma 2.4.3 that there exists a hyperplane that linear separates the two. Therefore, if convex hulls are disjoint (i.e. (2.15) holds), then $\mathbb{P}$ and $\mathbb{Q}$ are linearly separable.

(only if) Suppose $\mathbb{P}$ and $\mathbb{Q}$ are linearly separable but there exists convex hulls of $\mathbb{P}$ and $\mathbb{Q}$ that are not disjoint, i.e. there exist $\{p_i\}_{i=1}^{N_p} \subset \mathbb{P}$, $\{q_j\}_{j=1}^{N_q} \subset \mathbb{Q}$, and $\alpha_i, \beta_j > 0$ that satisfies $\sum_{i=1}^{N_p} \alpha_i = 1$, $\sum_{j=1}^{N_q} \beta_j = 1$ s.t.

$$\sum_{i=1}^{N_p} \alpha_i p_i = \sum_{j=1}^{N_q} \beta_j q_j, \tag{A.14}$$

for finite N_p, N_q. We can easily see that this contradicts linear separability. Suppose there exists a linear classifier (i.e. $w(x) = b$ exists that satisfies (2.14)). By multiplying each side of (A.14) with $w(x)$ and integrating over X, we have

$$\int_X w(x) \left(\sum_i \alpha_i p_i(x) \right) dx = \int_X w(x) \left(\sum_j \beta_j q_j(x) \right) dx. \tag{A.15}$$

The left side of (A.15) is always smaller than b because

$$\int_X w(x) \left(\sum_i \alpha_i p_i(x) \right) dx = \sum_i \alpha_i \int_X w(x) p_i(x) dx < \sum_i \alpha_i b = b. \tag{A.16}$$

On the other hand, the right side of (A.15) is always larger than b because

$$\int_X w(x) \left(\sum_j \beta_j q_j(x) \right) dx = \sum_j \beta_j \int_X w(x) q_j(x) dx > \sum_j \beta_j b = b \tag{A.17}$$

However, (A.16) and (A.17) contradict to the equivalence in (A.15), which implies that the linear classifier w cannot exist. Therefore, the convex hulls must be disjoint if linear classifier exists. □

A.5 Proof for Theorem 2.4.6: Linear Separability in the CDT Space

Proof. We show that $\widehat{\mathbb{P}}$, $\widehat{\mathbb{Q}}$ must be linearly separable. If not, it would contradict Definition 2.4.5 that they are disjoint. Suppose $\widehat{\mathbb{P}}$, $\widehat{\mathbb{Q}}$ are not linearly separable. Then by Lemma 2.4.4, there exist $\{p_i\}_{i=1}^{N_p} \subset \mathbb{P}$, $\{q_j\}_{j=1}^{N_q} \subset \mathbb{Q}$, and $\alpha_i, \beta_j > 0$ that satisfies $\sum_{i=1}^{N_p} \alpha_i = \sum_{j=1}^{N_q} \beta_j = 1$ such that the convex combination of $\{p_i\}_{i=1}^{N_p}$ and $\{q_j\}_{j=1}^{N_q}$ are equivalent, i.e.

$$\sum_{i=1}^{N_p} \alpha_i \widehat{p}_i = \sum_{j=1}^{N_q} \beta_j \widehat{q}_j.$$

By substituting $\widehat{p}_i = (f_i - \mathbb{1})\sqrt{I_0}$ and $\widehat{q}_j = (g_j - \mathbb{1})\sqrt{I_0}$, where $\mathbb{1}$ refers to an identity map, we have

$$\sum_{i=1}^{N_p} \alpha_i (f_i - \mathbb{1})\sqrt{I_0} = \sum_{j=1}^{N_q} \beta_j (g_j - \mathbb{1})\sqrt{I_0}.$$

By using $\sum_{i=1}^{N_p} \alpha_i = \sum_{j=1}^{N_q} \beta_j = 1$, and dividing each side of the equation by I_0, we have

$$\sum_{i=1}^{N_p} \alpha_i f_i = \sum_{j=1}^{N_q} \beta_j g_j.$$

By substituting $f_i = h_i^{-1} \circ f_0$ and $g_j = h_j^{-1} \circ g_0$ (see Lemma E.1 presented below), we have

$$\sum_{i=1}^{N_p} \alpha_i (h_i^{-1} \circ f_0) = \sum_{j=1}^{N_q} \beta_j (h_j^{-1} \circ g_0).$$

By substituting $h_\alpha^{-1} = \sum_{i=1}^{N_p} \alpha_i h_i^{-1}$ and $h_\beta^{-1} = \sum_{j=1}^{N_q} \beta_j h_j^{-1}$, we have

$$h_\alpha^{-1} \circ f_0 = h_\beta^{-1} \circ g_0.$$

By composing each side of the equation with h_α, we have

$$f_0 = h_\alpha \circ h_\beta^{-1} \circ g_0. \tag{A.18}$$

Note that $h_\alpha^{-1}, h_\alpha, h_\alpha \circ h_\beta^{-1} \in \mathbb{H}$ by conditions *i), ii), iii).* From the definition of the CDT in (2.3) with respect to reference I_0, we have

$$f_0'(p_0 \circ f_0) = g_0'(q_0 \circ g_0) = I_0.$$

By substituting f_0 with the right side of (A.18), we have

$$\begin{aligned} & (h_\alpha \circ h_\beta^{-1} \circ g_0)'(p_0 \circ (h_\alpha \circ h_\beta^{-1} \circ g_0)) = g_0'(q_0 \circ g_0) \\ \Leftrightarrow \quad & (h_\alpha \circ h_\beta^{-1})'(p_0 \circ (h_\alpha \circ h_\beta^{-1})) = q_0 \\ \Leftrightarrow \quad & h'_{\alpha\beta^{-1}} p_0(h_{\alpha\beta^{-1}}) = q_0. \end{aligned}$$

The last step of the equation is derive by setting $h_{\alpha\beta^{-1}} = h_\alpha \circ h_\beta^{-1}$, where $h_{\alpha\beta^{-1}} \in \mathbb{H}$. However, the last statement contradicts the Definition 2.4.5 that $h'p_0(h)$ and $h'q_0(h)$ each belong to disjoint set $\widehat{\mathbb{P}}$ and $\widehat{\mathbb{Q}}$. Therefore, $\widehat{\mathbb{P}}$, $\widehat{\mathbb{Q}}$ must be linearly separable.

□

Lemma E.1. *Let f_0, f_i be monotonic functions from $X \to Y$ for probability densities $p_0 : Y \to \mathbb{R}$, $p_i : Y \to \mathbb{R}$ with respect to reference $I_0 : X \to \mathbb{R}$, such that*

$$\int_{\inf(Y)}^{f_0(x)} p_0(\tau)d\tau = \int_{\inf(Y)}^{f_i(x)} p_i(\tau)d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau. \tag{A.19}$$

Then $p_i = h_i'(p_0 \circ h_i)$ implies $h_i \circ f_i = f_0$.

Proof. Substituting (A.19) with $p_i = h_i' p_0(h_i)$, we have

$$\int_{\inf(Y)}^{f_i(x)} h_i'(\tau)p_0(h_i(\tau))d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau.$$

By change of variables theorem, substituting $h_i(\tau) = u$ and $h_i'(\tau)d\tau = du$, we have

$$\int_{\inf(Y)}^{h_i(f_i(x))} p_0(u)du = \int_{\inf(X)}^{x} I_0(\tau)d\tau.$$

Since $\int_{\inf(Y)}^{f_0(x)} p_0(\tau)d\tau = \int_{\inf(X)}^{x} I_0(\tau)d\tau$ holds (see (A.19)), we have

$$\int_{\inf(Y)}^{h_i(f_i(x))} p_0(\tau)d\tau = \int_{\inf(Y)}^{f_0(x)} p_0(\tau)d\tau. \tag{A.20}$$

The upper limits on each side of the integrals in (A.20) can be set to be equal since both f_i and h_i are strictly increasing functions:

$$h_i(f_i(x)) = f_0(x). \tag{A.21}$$

Equivalently, we have $f_i(x) = h_i^{-1}(f_0(x))$ by inverting (A.21). □

A.6 Proof for Translation property of Radon-CDT

For $J(x, y) = I(x - x_0, y - y_0)$ and using the properties of Radon transform we have,

$$\widehat{J}(t, \theta) = \widehat{I}(t - x_0 \cos(\theta) - y_0 \sin(\theta), \theta)$$

Therefore the Radon-CDT of J can be written as,

$$\tilde{J}(t, \theta) = (g(t, \theta) - t)\sqrt{\widehat{I}_0(t, \theta)}$$

where $g(t, \theta)$ satisfies,

$$\int_{-\infty}^{g(t,\theta)} \widehat{J}(\tau, \theta)d\tau = \int_{-\infty}^{t} \widehat{I}_0(\tau, \theta)d\tau$$

The left hand side of above equation can be rewritten as,

$$\begin{aligned}
\int_{-\infty}^{g(t,\theta)} \widehat{J}(\tau,\theta)d\tau &= \int_{-\infty}^{g(t,\theta)} \widehat{I}(\tau - x_0\cos(\theta) - y_0\sin(\theta),\theta)d\tau \\
&= \int_{-\infty}^{g(t,\theta)-x_0\cos(\theta)-y_0\sin(\theta)} \widehat{I}(u,\theta)du \Rightarrow \\
g(t,\theta) - x_0\cos(\theta) - y_0\sin(\theta) &= f(t,\theta) \Rightarrow \\
g(t,\theta) &= f(t,\theta) + x_0\cos(\theta) + y_0\sin(\theta) \Rightarrow \\
(g(t,\theta)-t)\sqrt{\widehat{I}_0(t,\theta)} &= (f(t,\theta)-t)\sqrt{\widehat{I}_0(t,\theta)} + \\
&(x_0\cos(\theta) + y_0\sin(\theta))\sqrt{\widehat{I}_0(t,\theta)} \Rightarrow \\
\tilde{J}(t,\theta) &= \tilde{I}(t,\theta) + (x_0\cos(\theta) + y_0\sin(\theta))\sqrt{\widehat{I}_0(t,\theta)}
\end{aligned}$$

where $\frac{\partial f}{\partial t}(t,\theta)\widehat{I}(f(t,\theta),\theta) = \widehat{I}_0(t,\theta)$.

A.7 Proof for Scaling property of Radon-CDT

For $J(x,y) = \alpha^2 I(\alpha x, \alpha y)$ with $\alpha > 0$ and using the properties of Radon transform we have,

$$\widehat{J}(t,\theta) = \alpha\widehat{I}(\alpha t,\theta).$$

The Radon-CDT of J can be written as,

$$\tilde{J}(t,\theta) = (g(t,\theta)-t)\sqrt{\widehat{I}_0(t,\theta)}$$

where $g(t,\theta)$ satisfies,

$$\int_{-\infty}^{g(t,\theta)} \widehat{J}(\tau,\theta)d\tau = \int_{-\infty}^{t} \widehat{I}_0(\tau,\theta)d\tau$$

The left hand side of above equation can be rewritten as,

$$\begin{aligned}
\int_{-\infty}^{g(t,\theta)} \widehat{J}(\tau,\theta)d\tau &= \int_{-\infty}^{g(t,\theta)} \alpha\widehat{I}(\alpha\tau,\theta)d\tau \\
&= \int_{-\infty}^{\alpha g(t,\theta)} \widehat{I}(u,\theta)du \Rightarrow \\
g(t,\theta) &= \frac{f(t,\theta)}{\alpha} \Rightarrow \\
(g(t,\theta)-t)\sqrt{\widehat{I}_0(t,\theta)} &= (\frac{f(t,\theta)}{\alpha}-t)\sqrt{\widehat{I}_0(t,\theta)} \\
= \frac{(f(t,\theta)-t)\sqrt{\widehat{I}_0(t,\theta)}}{\alpha} &+ (\frac{1-\alpha}{\alpha})\sqrt{\widehat{I}_0(t,\theta)} \Rightarrow \\
\tilde{J}(t,\theta) &= \frac{\tilde{I}(t,\theta)}{\alpha} + (\frac{1-\alpha}{\alpha})\sqrt{\widehat{I}_0(t,\theta)}
\end{aligned}$$

A.8 Proof for Rotation property of Radon-CDT

For $J(x,y) = I(x\cos(\phi)+y\sin(\phi), -x\sin(\phi)+y\cos(\theta))$ and using the properties of Radon transform we have,

$$\widehat{J}(t,\theta) = \widehat{I}(t,\theta-\phi).$$

Given a circularly symmetric reference image, the Radon-CDT of J can be written as,

$$\tilde{J}(t,\theta) = (g(t,\theta)-t)\sqrt{\widehat{I}_0(t,\theta)}$$

where $g(t,\theta)$ satisfies,

$$\int_{-\infty}^{g(t,\theta)} \widehat{J}(\tau,\theta)d\tau = \int_{-\infty}^{t} \widehat{I}_0(\tau,\theta)d\tau$$

The left hand side of above equation can be rewritten as,

$$\begin{aligned}
\int_{-\infty}^{g(t,\theta)} \widehat{J}(\tau,\theta)d\tau = \int_{-\infty}^{g(t,\theta+\phi)} \widehat{I}(\tau,\theta)d\tau \Rightarrow \\
f(t,\theta) = g(t,\theta+\phi) \Rightarrow g(t,\theta) = f(t,\theta-\phi) \Rightarrow \\
(g(t,\theta)-t)\sqrt{\widehat{I}_0(t,\theta)} = (f(t,\theta-\phi)-t)\sqrt{\widehat{I}_0(t,\theta)} \\
= (f(t,\theta-\phi)-t)\sqrt{\widehat{I}_0(t,\theta-\phi)} \Rightarrow \\
\tilde{J}(t,\theta) = \tilde{I}(t,\theta-\phi)
\end{aligned}$$

A.9 Proof for Theorem 3.4.1: Linear separability in the Radon-CDT space

Let image classes $\mathbb{P}$ and $\mathbb{Q}$ be generated from Eq. (3.11). Here we show that the classes are linearly separable in the Radon-CDT space.

Proof. By contradiction we assume that the transformed image classes are not linearly separable,

$$\begin{aligned}
\sum_i \alpha_i \tilde{p}_i(t,\theta) &= \sum_j \beta_j \tilde{q}_j(t,\theta) \Rightarrow \\
\sum_i \alpha_i f_i(t,\theta) &= \sum_j \beta_j g_j(t,\theta)
\end{aligned}$$

where $\sum_i \alpha_i = \sum_j \alpha_j = 1$, $\frac{\partial f_i}{\partial t}(t,\theta)\widehat{p}_i(f_i(t,\theta),\theta) = \widehat{I}_0(t,\theta)$, and $\frac{\partial g_j}{\partial t}(t,\theta)\widehat{q}_j(g_j(t,\theta),\theta) = \widehat{I}_0(t,\theta)$. Figure A.1 shows a diagram which illustrates the interactions between $\widehat{q}_j$s, $\widehat{p}_i$s, and $\widehat{I}_0$. It is straightforward to show that $f_i(t,\theta) = h_i^{-1}(f_0(t,\theta),\theta)$ and $g_j(t,\theta) = h_j^{-1}(g_0(t,\theta),\theta)$ as can also be seen from the diagram in Figure A.1. Therefore we can write,

$$\begin{aligned}
\sum_i \alpha_i f_i(t,\theta) &= \sum_j \beta_j g_j(t,\theta) \Rightarrow \\
\sum_i \alpha_i h_i^{-1}(f_0(t,\theta),\theta) &= \sum_j \beta_j h_j^{-1}(g_0(t,\theta),\theta)
\end{aligned}$$

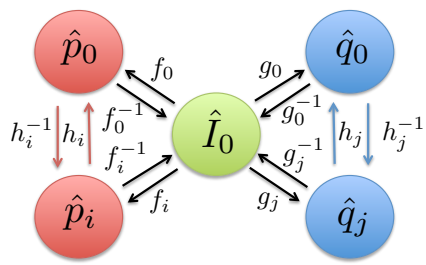


Figure A.1: The diagram of interactions of the images mass preserving maps.

Defining $h_\alpha(t,\theta) = \sum_i \alpha_i h_i^{-1}(t,\theta) \in \mathcal{H}$ and $h_\beta(t,\theta) = \sum_i \beta_j h_j^{-1}(t,\theta) \in \mathcal{H}$, we can rewrite above equation as,

$$\begin{aligned} h_\alpha(f_0(t,\theta),\theta) &= h_\beta(g_0(t,\theta),\theta) \Rightarrow \\ f_0(t,\theta) &= h_\alpha^{-1}(h_\beta(g_0(t,\theta),\theta),\theta). \end{aligned}$$

Defining $h(t,\theta) = h_\alpha^{-1}(h_\beta(t,\theta),\theta) \in \mathcal{H}$ we have,

$$f_0(t,\theta) = h(g_0(t,\theta),\theta)$$

which implies that $\exists h \in \mathcal{H} \rightarrow \frac{\partial h}{\partial t}(t,\theta)\widehat{p}_0(h(t,\theta),\theta) = \widehat{q}_0(t,\theta)$, which contradicts with the fourth condition of $\mathcal{H}$.

$\square$

Appendix B

Link from Wave Propagation to the Optimal Transport

B.1 Transport and The Conservation of Mass

The continuity equation (or conservation of mass formula) in continuum physics states:

$$\frac{\partial \rho}{\partial t} + \nabla \cdot (\rho v) = 0. \tag{B.1}$$

Here $\rho = \rho(t, \vec{x})$ stands for the density of a system of particles at time t and position $\vec{x}$; $v = v(t, \vec{x})$ for the velocity field at time t and position $\vec{x}$, and $\nabla\cdot$ stands for the divergence operator. The natural setting for this equation is a Riemannian manifold M. This Eulerian description (B.1), can be alternatively expressed in Lagrangian description with the dynamic coordinates [97]:

$$\vec{x}_z \equiv f(\vec{x}_0, z), \tag{B.2}$$

the coordinates is labeled according to their 'new' location defined by the Lagrangian map f_t, which evolves the starting coordinates $\vec{x}_0$ forward in space to location z. The

velocity can be expressed as

$$v(t, f_t(\vec{x})) = \frac{d}{dt} f_t(\vec{x}).$$

If $v(t, \vec{x})$ is (locally) Lipschitz continuous, this is equivalent to saying that there always exist a transport map f_t that pushes the density ρ_0 onto ρ_t:

$$\rho_t = (f_t)_{\#}\rho_0.$$

B.2 Parabolic wave equation to continuity equation

The parabolic wave equation can be written as

$$i2k_0 \frac{\partial \Psi(\vec{x}, z)}{\partial z} + \nabla_X^2 \Psi(\vec{x}, z) - 2k_0^2 \eta(\vec{x}, z)\Psi(\vec{x}, z) = 0. \tag{B.3}$$

where

- $\vec{x} = (x_1, x_2)$ defines the plane in the direction traverse to propagation,
- wave is propagating horizontally (in the z direction) with wavenumber k_0,
- ∇_X^2 denotes the Laplacian operating in the transverse coordinates (x_1, x_2),
- $\tilde{n}(\vec{x}, z) \equiv n(\vec{x}, z) + i\kappa(\vec{x}, z)$ is the complex index of refraction, where $n(\vec{x}, z)$ is the refractive index and $\kappa(\vec{x}, z)$ is extinction coefficient, and
- $\eta(\vec{x}, z) \equiv n(\vec{x}, z) - 1$ is the deviation in refractive index from unity.

Assuming $\rho(\vec{x}, z) \geq 0$ and using the Madelung transformations [98] $\Psi(\vec{x}, z) = \sqrt{\rho(\vec{x}, z)} \exp(i\phi(\vec{x}, z)/2)$, Eq. (B.3) becomes:

$$\frac{\partial \rho(\vec{x}, z)}{\partial z} + \nabla_X \cdot (\rho(\vec{x}, z) v(\vec{x}, z)) = 0, \tag{B.4}$$

$$\frac{\partial v(\vec{x}, z)}{\partial z} + (v(\vec{x}, z) \cdot \nabla_X) v(\vec{x}, z) = \nabla_X p(\vec{x}, z) + 2\nabla_X \eta(\vec{x}, z). \tag{B.5}$$

where $v(\vec{x}, z) \equiv \nabla_X \phi(\vec{x}, z)$ and $p(\vec{x}, z) = \frac{2\nabla_X^2 \left(\rho(\vec{x},z)^{1/2}\right)}{\rho(\vec{x},z)^{1/2}}$. Therefore, the parabolic wave equation given in Eq. (B.3) can be interpreted as the continuity and momentum equation in Eq. (B.4) and Eq. (B.5) in fluid mechanics. The 'density' $\rho(\vec{x}, z) = \Psi(\vec{x}, z)\Psi(\vec{x}, z)^*$ is the image intensity, and the phase gradient $v(\vec{x}, z) \equiv \nabla_X \phi(\vec{x}, z)$ plays the role of the velocity.

If we assume additional constraints that the the polarization in the direction of propagation z is minimal (and thus can be neglected) and that the extinction coefficient $\kappa = 0$ [jon], we can rewrite Eq. (B.5) as

$$\frac{\partial v(\vec{x}, z)}{\partial z} + (v(\vec{x}, z) \cdot \nabla_X) v(\vec{x}, z) = 2\nabla_X g(\eta(\vec{x}, z)). \tag{B.6}$$

where the right hand side is now written entirely in terms of the material index

$$g(\eta(\vec{x}, z)) = -2\nabla_X^2 \log(n(\vec{x}, z)) + 4(\nabla_X \log(n(\vec{x}, z)))^2 + \eta(\vec{x}, z),$$

In other words, the refractive index creates the potential function $g(\eta(\vec{x}, z))$ of the momentum equation in Eq. (B.5). In the absence of turbulence, or other index fluctuations, the right hand side disappears and the momentum equation becomes simply $Dv(\vec{x}, z)/Dz = 0$ which suggests a constant velocity solution. Because the right hand side is a function of the traverse index gradient, a constant velocity solution applies when in the case that the refractive index n is varying in z only. Therefore, each localized portion of the electric field moves in straight lines from source to destination, as illustrated in Fig. B.1.

B.3 Solution for the Parabolic Wave Equation via Optimal Transport

In the previous section, we derived continuity and momentum Eq. (B.4), Eq. (B.5) from the parabolic wave equation (B.3). On the other hand, we saw from Eq. (B.1) that when the velocity field is locally Lipschitz continuous, than there always exist a transport map which satisfies Eq. (B.1).

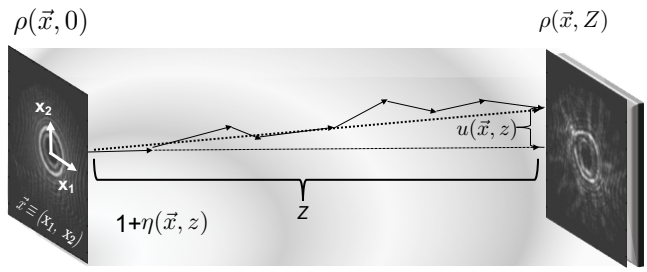


Figure B.1: Illustration of the transport problem. Intensity is transported in the transverse plane as the associated EM field moves through space from $z = 0$ to $z = Z$. Absent fluctuations in the refractive index the intensity is transported along constant velocity paths, i.e., straight lines.

We can furthermore relate the the solution of the *optimal* transport map in Eq. (B.4) to the solution the parabolic equation in (B.3) with two assumptions:

1. We are minimizing Kinetic energy, namely *action*:

$$\min \mathcal{A} \equiv Z \int_{\mathbb{R}^2} \int_0^Z \rho(\vec{x}, z) |v(\vec{x}, z)|^2 dz d\vec{x}, \tag{B.7}$$

   which is associated with moving 'image' intensity over a distance $z = [0, Z]$ in time interval $t = [0, T]$.

2. The potential function $V(\vec{x}, z) = 2g(\eta(vx, z))$ in Eq.(B.5) is neglected in forming the action (or $\eta(\mathbf{x}, z) \ll 1$).

We mention that it has been recently shown in [99] that minimization of the specific action (B.7) given the constraint (B.4) and the assumption that intensity is conserved yields precisely (B.5) along with the requirement that $v(\vec{x}, z) = \nabla_X \phi(\vec{x}, z)$.

In summary, if we accept the minimization of action principle, if the index fluctuations are small, if two intensity points ρ_0, ρ_T are given, and if the total mass between two points are conserved, then the solution to the parabolic wave equation can be found by seeking for the optimal transport map that pushes ρ_0 to ρ_T.

Specifically, the flow map, f, so that resulting intensity and velocity field minimizes

the action Eq. (B.7) can be obtained via minimizing the Kantorovich-formulation:

$$d_p(0,Z)^2 = \inf_f \int_{\mathbb{R}^2} \|f(\vec{x}_0,Z) - \vec{x}_0\|^2 \rho(\vec{x}_0,0) d\vec{x} = \min_v \mathcal{A} \tag{B.8}$$

subject to the constraints imposed by continuity equation (conservation of mass) which can be rewritten as,

$$\int \rho(\vec{x}_z,z) d\vec{x} = \int \rho(\vec{x}_0,0) d\vec{x} \tag{B.9}$$

$$\det(J_f(\vec{x}_0,z))\rho(\vec{x}_z,z) = \rho(\vec{x}_0,0). \tag{B.10}$$

where $J_f(\vec{x}_0,z)$ denotes the Jacobian of $f(\vec{x}_0,z)$ (see [97], [100] or [99]).

Also, the velocity (which is constant in z) can be expressed as a phase gradient [99],

$$v(\vec{x}_z,z) = (f(\vec{x}_0,Z) - \vec{x}_0)/Z = \nabla_X \phi(\vec{x}_z,z). \tag{B.11}$$

The solution is exact if the index perturbations are zero as the light will travel in straight lines, with constant velocity. In the event that the index is fluctuating, the constant velocity solutions are approximating a wandering path with a straight line. And once the map f is found, the displacement coordinates can be linearly interpolated via

$$\vec{x}_z = f(\vec{x}_0,z) = (1 - z/Z)\vec{x}_0 + \frac{z}{Z} f(\vec{x}_0,Z).$$

Appendix C

Atmospheric Turbulence

C.1 Atmospheric Turbulence as Random Fields

C.1.1 Spatial Covariance Function

A random field $u(\vec{r})$ is a collection of random numbers whose indices are identified with a spatial coordinates in $\vec{r} = [x, y, z]$.

We define the mean or expected value of the random field $u(\vec{r})$ by

$$m(\vec{r}) = \langle u(\vec{r}) \rangle$$

where the brackets $\langle\rangle$ denote an ensemble average.

The associated spatial autocovariance function, or simply the covariance function is

$$B_u(\vec{r}_1, \vec{r}_2) = \langle (u(\vec{r}_1) - m(\vec{r}_1))(u^*(\vec{r}_2) - m^*(\vec{r}_2)) \rangle,$$

where $u^*(\vec{r}_1)$ denotes the complex conjugate of $u(\vec{r}_1)$.

The random field $u(\vec{r})$ is *statistically homogeneous* if its moments are invariant under a spatial translation. In other words, it is equivalent to saying that:

1. its mean value $\langle u(\vec{r}) \rangle = m(\vec{r}) = m$ is independent of the spatial position $\vec{r}$, and

2. the covariance function depends only on the spatial vector difference $\vec{r} = \vec{r}_2 - \vec{r}_1$, and the covariance function can be represented as:

$$B_u(\vec{r}) = \langle u(\vec{r}_1) u^*(\vec{r}_1 + \vec{r}) \rangle - |m|^2.$$

The random field $u(\vec{r})$ is *statistically isotropic* if the moments are invariant under rotations; the covariance function depends only on the scalar distance R, i.e. $4B_u(\vec{r}_1, \vec{r}_2) = B_u(r)$.

C.1.2 Spatial power spectrum

If $u(\vec{r})$ is statistically *homogeneous and isotropic* complex random field with zero mean, its covariance function $B_u(r)$ can be expressed in the Fourier integral form

$$B_u(r) = \int \exp^{i\kappa r} V_u(\kappa) d\kappa,$$

where κ (in units of rad/m) denotes the wave number (spatial frequency) and $V_u(\kappa)$ is the one-dimensional spectrum of the random field $u(\vec{r})$.

If $u(\vec{r})$ is statistically *homogeneous* with zero mean, the covariance $B_u(\vec{r})$ can be represented as

$$B_u(\vec{r}) = \int \int \int \exp^{i\mathbf{K}\cdot\vec{r}} \Psi_u(\mathbf{K}) d^3\kappa$$

where the function $\Psi_u(\mathbf{K})$ is the three-dimensional *spatial power spectrum* of the random field $u(\vec{r})$. This function also can be obtained directly from the covariance function through the inverse Fourier transform relation.

C.1.3 Structure Function of Random Fields

Random fields are often times not strictly stationary, and in practice, the theoretical description of spatial fluctuations of a random field in terms of the covariance function and power spectral density are very limiting. For instance, velocity fields in turbulence are not strictly homogeneous because the average velocity field cannot be constant over widely separated portions of the random medium. Nonetheless, the velocity difference

at two distinct points almost always behaves like a statistically homogeneous field. Formally speaking, the random field $u(\vec{r})$ is decomposed into the sum

$$u(\vec{r}) = m(\vec{r}) + u_1(\vec{r})$$

where $m(\vec{r}) = \langle u(\vec{r}) \rangle$ is the mean and $u_1(\vec{r})$ is statistically homogeneous fluctuation with mean value 0. Random fields that permit a decomposition into a varying mean and a statistically homogeneous fluctuation are called locally homogeneous.

Locally homogeneous fields are characterized by the structure function. In general, the structure function for a locally homogeneous random field $u(\vec{r})$ can be expressed in the form

$$\begin{aligned} D_u(\vec{r}_1, \vec{r}_2) = D_u(r) &= \langle (u(\vec{r}_1) - u(\vec{r}_1 + \vec{r}))^2 \rangle \\ &= \langle (u_1(\vec{r}_1) - u_1(\vec{r}_1 + \vec{r}))^2 \rangle \end{aligned}$$

and the spectrum is related to the structure function by

$$D_u(\vec{r}) = 2 \int \int \int \Psi_u(\mathbf{K})(1 - \cos(\mathbf{K} \cdot \vec{r})) d^2\kappa$$

C.2 Turbulence Theory

Light propagating through the atmosphere is affected by random fluctuations caused by the atmospheric turbulence. The atmospheric turbulence can be described in temperature fluctuations and index of refraction fluctuations. Specifically, air of different temperature leads to inhomogeneities in the index of refractions. For example, when the plane wave front light propagates through the atmosphere through the regions of high refractive index, the light will be delayed with respect to other regions. And when the light is received at the transmitter, the plane wave front would no longer be flat but severely distorted. The phase interference caused by the delay across the beam would eventually result in fluctuations in intensity. Spatial spreading of the beam occurs as well, if the turbulence eddies of size greater than the beam diameter act like moving

lenses [2].

To describe atmospheric turbulence, statistical approaches - modeling turbulence as a random field - have been taken since Kolmogorov's theory of turbulence [101]. For mathematical simplicity, Kolmogorov assumed that the turbulence field is locally homogeneous and isotropic. Recall from Sec. C.1 that statistical local homogeneity of the random field implies that the field can be decomposed into a varying mean and statistically homogeneous fluctuations, and that isotropic field can be described by only on their vector separation (independent of the chosen observation points).

Here we describe the turbulence caused by the fluctuations in the index of refraction $n(\vec{r})$, which are primarily to random temperature fluctuations in the visible and near-IR region of the spectrum. Under the assumption that the random field of the index of refraction $n(\vec{r})$ can be considered locally homogeneous and isotropic, the spatial distribution of the field can be characterized by the *structure function*, which is the variance of index of refraction difference between two points separated by a vector [102].

$$
\begin{aligned}
D_n(\vec{r}) &= \langle |n(\vec{r},\cdot) - n(\vec{r}+\vec{r}_1,\cdot)|^2 \rangle \\
&= \begin{cases} C_n^2 l_0^{-4/3} r^2, & 0 \ll r \ll l_0 \\ C_n^2 r^{2/3}, & l_0 \ll r \ll L_0 \end{cases}
\end{aligned}
$$

where L_0 and l_0 are the outer and inner scale respectively, which bounds the subrange which turbulence properties are assumed to be statistically homogeneous and isotropic. C_n is the index-of-refraction structure constant (in units of $m^{-2/3}$), which is a measure of the strength of the fluctuations in the refractive index. Values of C_n^2 typically range from 10^{-17} or less for *weak* turbulence and up to 10^{-13} or more for *strong* turbulence. At constant height above the ground and for short time intervals at a fixed propagation distance, it may be reasonable to assume that C_n^2 is essentially constant. For example, Fig. C.1 shows the refractive structure parameter over a three-day period in August 2002 [1] in Washington D.C.

From the structure function, the well-known Kolmogorov power-law spectrum can

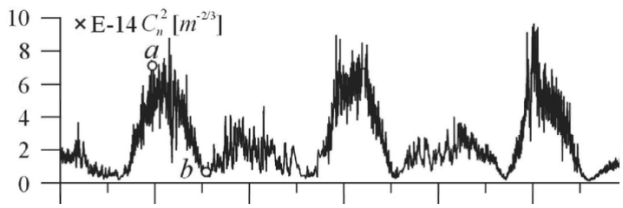


Figure C.1: Daytime C_n^2 Profile over a three-day period in August 2002 [1].

be derived:

$$\Psi_n(\kappa) = 0.033 C_n^2 \kappa^{-11/3}, 1/L_0 \ll \kappa \ll l_0.$$

C.2.1 Simulating Turbulence through Phase Screen

One of the technique to simulate the light propagating through the turbulence is the phase screen method [92]. In this method, the light propagation path is modeled as layers of medium of fluctuating diffractive index (see Fig. C.5). When light passes through layer of changing diffractive index, the light is assumed to diffract and propagate to the next layer. In order to produce this behavior of light, statistically generated phase screens are placed at each layer. These phase screens represent the phase that would have occurred in propagation from the previous layer.

Specifically, phase screens can be generate as follows. Here we used modified Kolmogorov spectrum, but other spectrum model can be used as well. Complex Gaussian white noise (two 2-D array of Gaussian random numbers for the real and imaginary parts) C is multiplied by the square root of the spectrum $\Psi(\kappa)$, and then inverse Fourier transformed:

$$P = \mathcal{F}^{-1}\{\sqrt{\Psi(\kappa)}C\}. \tag{C.1}$$

where $\mathcal{F}^{-1}$ is the inverse 2-D Fourier transform, and $\Psi(\kappa)$ corresponds to modified Kolmogorov turbulence model [93]:

$$\Psi(\kappa) = 0.033 C_n^2 (\kappa^2 + 1/L_0^2)^{-11/6} \exp(-\kappa^2/\kappa_\ell^2) \times (1 + 1.082(\kappa/\kappa_\ell) - 0.254(\kappa/\kappa_\ell)^{7/6}),$$

where κ is the spatial frequency (rad/m), L_0 is the outer scale of turbulence, ℓ_0 is the inner scale of turbulence, and $\kappa_\ell = \kappa/(3.3\ell_0)$. C_n^2 is the structure constant of the index refraction, which measures the strength of the turbulence.

Fig. C.2 shows the Kolmogorov spectrum (left), the gaussian random numbers (middle), and the resulting phase screens (right). Note that multiplication of Kolmogorov spectrum with the gaussian can be viewed as filtering the gaussian field with the low frequency spatial filter $\Psi(\kappa)$. Fig. C.3 shows the random phase screens which all conform to the same Kolmogorov spectral model. By averaging $|FT|^2$ of 100 random phase screens and plotting the spectrum in k_x direction, we can verify from Fig. C.4 that the averaging approximates the true Kolmogorov spectrum very well.

When these phase screens are placed in the propagation path, they introduce spatially stochastic variations to the wave phase, and will develop amplitude fluctuations during wave propagation through the free space from screen to screen. Fig. C.6 shows a simulation example of how phase screens interrupts and distorts the beam intensity during propagation. Fig. C.6 (top) shows LG beam propagating in vacuum at $z = 100m$ (top). When a random phase screen (one of Fig. C.3) is inserted, it will distort the phase as shown in Fig. C.6 (bottom, right). The beam intensity after propagating $z = 1m$ is shown in Fig. C.6 (bottom, left). In later chapters, these generate phase screens are used to simulate the effect of turbulence on the propagation of light.

The fried parameter, or coherent length, r_0, measures the quality of the optical transmission through the atmosphere along the defined path. r_0 is related to $C_n(z)^2$ by

$$r_0 = \left(0.432\left(\frac{2\pi}{\lambda}\right)^2 \sec(\alpha)\int_{path} zC_n(z)^2 dz\right)^{-3/5}, \tag{C.2}$$

where α is the zenith angle. When a constant turbulence strength is assumed over the propagation distance and $\alpha = 0$, (C.2) can be rewritten in relation to the structure constant of the index refraction C_n^2 as or

$$r_0 = \left(0.432\left(\frac{2\pi}{\lambda}\right)^2 \Delta z C_n^2\right)^{-3/5}.$$

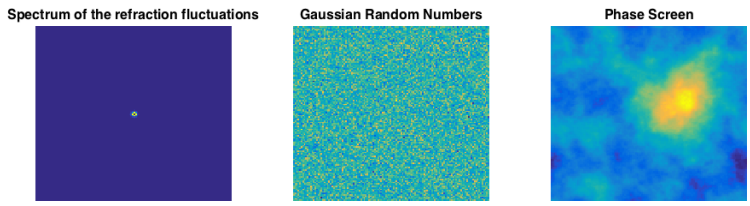


Figure C.2: Generation of Phase Screen

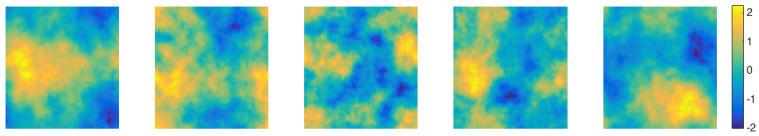


Figure C.3: Random Realization of Turbulence

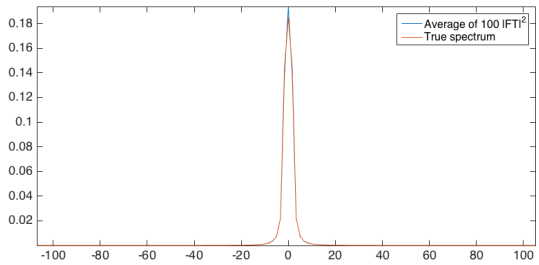


Figure C.4: Verification of Kolmogorov Spectrum, of 100 random realizations

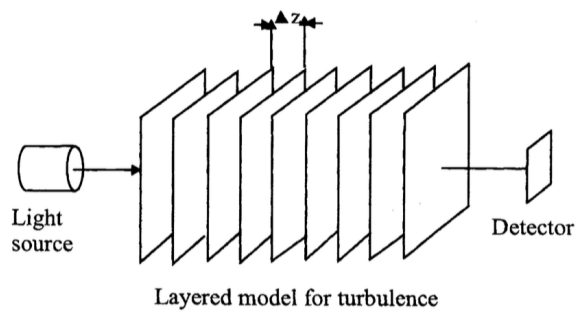


Figure C.5: Layered Propagation System for Modeling Propagation of light through Turbulent Atmosphere [2].

It has the dimension of length – at visible wavelengths, r_0 varies from 20 cm at the best locations to 5 cm at typical sea-level sites. It defines an important length scale of the theory of *seeing*: the scale length over which phase errors in a wave front are of the order of 1 radian, or put in another way, the phase variance σ^2 over an aperture of diameter D is approximately 1 rad^2 [103]:

$$\sigma^2 = 1.0299 \left(\frac{D}{r_0}\right)^{5/3}.$$

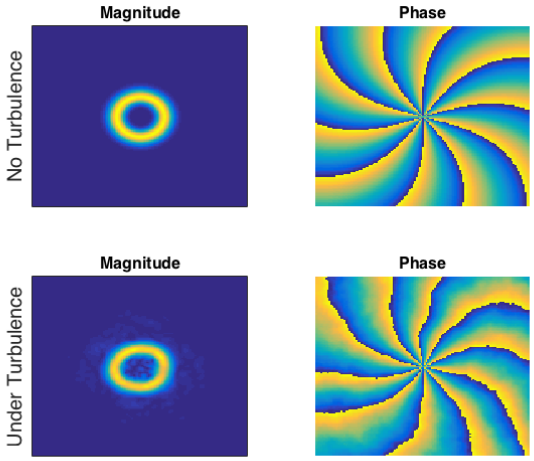


Figure C.6: Effect of a phase screen

Bibliography

[1] Jennifer C Ricklin, Stephane Bucaille, and Frederic M Davidson. Performance loss factors for optical communication through clear air turbulence. In *Optical Science and Technology, SPIE's 48th Annual Meeting*, pages 1–12. International Society for Optics and Photonics, 2004.

[2] Alastair D McAulay. Generating kolmogorov phase screens for modeling optical turbulence. In *AeroSense 2000*, pages 50–57. International Society for Optics and Photonics, 2000.

[3] David W Kammler. *A first course in Fourier analysis*. Cambridge University Press, 2007.

[4] Stéphane Mallat. *A wavelet tour of signal processing*. Academic press, 1999.

[5] Hu Huang, Akif Burak Tosun, Jia Guo, Cheng Chen, Wei Wang, John A Ozolek, and Gustavo K Rohde. Cancer diagnosis by nuclear morphometry using spatial information. *Pattern recognition letters*, 42:115–121, 2014.

[6] Josef Spidlen, Karin Breuer, Chad Rosenberg, Nikesh Kotecha, and Ryan R Brinkman. Flowrepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytometry Part A*, 81(9):727–731, 2012.

[7] Wen Gao Wenchao Zhang, Shiguang Shan and Xilin Chen. Local gabor binary pattern histogram sequence (lgbphs): a novel non-statistical model for face rep-

resentation and recognition. *Tenth IEEE International Conference on Computer Vision, 2005*, 1:786–791, 2005.

[8] Daniel C Harris. *Quantitative chemical analysis*. Macmillan, 2010.

[9] Wenxin Li, David Zhang, and Zhuoqun Xu. Palmprint identification by fourier transform. *International Journal of Pattern Recognition and Artificial Intelligence*, 16(04):417–432, 2002.

[10] Donald M Monro, Soumyadip Rakshit, and Dexin Zhang. DCT-based iris recognition. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(4):586–595, 2007.

[11] Minh N Do and Martin Vetterli. Wavelet-based texture retrieval using generalized gaussian density and kullback-leibler distance. *Image Processing, IEEE Transactions on*, 11(2):146–158, 2002.

[12] Tanaya Mandal, Angshul Majumdar, and QM Jonathan Wu. Face recognition by curvelet based feature extraction. In *Image Analysis and Recognition*, pages 806–817. Springer, 2007.

[13] Nikolaos V Boulgouris and Zhiwei X Chi. Gait recognition using radon transform and linear discriminant analysis. *Image Processing, IEEE Transactions on*, 16(3):731–740, 2007.

[14] Lei Zhang, Xiantong Zhen, and Ling Shao. Learning object-to-class kernels for scene classification. *Image Processing, IEEE Transactions on*, 23(8):3241–3253, 2014.

[15] Ling Shao, Li Liu, and Xuelong Li. Feature learning for image classification via multiobjective genetic programming. *Neural Networks and Learning Systems, IEEE Transactions on*, 25(7):1359–1371, 2014.

[16] Fan Zhu and Ling Shao. Weakly-supervised cross-domain dictionary learning for visual recognition. *International Journal of Computer Vision*, 109(1-2):42–59, 2014.

[17] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[19] Joan Bruna and Stéphane Mallat. Invariant scattering convolution networks. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 35(8):1872–1886, 2013.

[20] Laurent Sifre and Stéphane Mallat. Rotation, scaling and deformation invariant scattering for texture discrimination. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 1233–1240. IEEE, 2013.

[21] Isabelle Guyon. *Feature extraction: foundations and applications*, volume 207. Springer Science & Business Media, 2006.

[22] Huan Liu and Hiroshi Motoda. *Feature extraction, construction and selection: A data mining perspective*. Springer, 1998.

[23] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

[24] Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

[25] Shuiwang Ji and Jieping Ye. Generalized linear discriminant analysis: a unified framework and efficient model selection. *Neural Networks, IEEE Transactions on*, 19(10):1768–1782, 2008.

[26] David JC MacKay. *Information theory, inference, and learning algorithms*, volume 7. Citeseer, 2003.

[27] Jerome H Friedman. On bias, variance, 0/1loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, 1(1):55–77, 1997.

[28] Christopher M Bishop et al. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[29] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*, volume 2. Springer, 2009.

[30] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1265–1278, 2005.

[31] Peter N. Belhumeur, João P Hespanha, and David Kriegman. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 19(7):711–720, 1997.

[32] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2000.

[33] Wei Wang, John A Ozolek, and Gustavo K Rohde. Detection and classification of thyroid follicular lesions based on nuclear structure from histopathology images. *Cytometry Part A*, 77(5):485–494, 2010.

[34] Wei Wang, Dejan Slepčev, Saurav Basu, John A Ozolek, and Gustavo K Rohde. A linear optimal transportation framework for quantifying and visualizing variations in sets of images. *International journal of computer vision*, 101(2):254–269, 2013.

[35] Saurav Basu, Soheil Kolouri, and Gustavo K Rohde. Detecting and visualizing cell phenotype differences from microscopy images using transport-based morphometry. *Proceedings of the National Academy of Sciences*, 111(9):3448–3453, 2014.

[36] S. Kolouri, A. B. Tosun, J. A. Ozolek, and G. K. Rohde. A continuous linear optimal transport approach for pattern analysis. *Pattern Recognition*, 51:453–462, 2016.

[37] Wei Wang, John A Ozolek, Dejan Slepcev, Ann B Lee, Cheng Chen, and Gustavo K Rohde. An optimal transportation approach for nuclear structure-based pathology. *Medical Imaging, IEEE Transactions on*, 30(3):621–631, 2011.

[38] John A Ozolek, Akif Burak Tosun, Wei Wang, Cheng Chen, Soheil Kolouri, Saurav Basu, Hu Huang, and Gustavo K Rohde. Accurate diagnosis of thyroid follicular lesions from nuclear morphology using supervised learning. *Medical image analysis*, 18(5):772–780, 2014.

[39] Akif Burak Tosun, Oleksandr Yergiyev, Soheil Kolouri, Jan F Silverman, and Gustavo K Rohde. Detection of malignant mesothelioma using nuclear structure of mesothelial cells in effusion cytology specimens. *Cytometry Part A*, 2015.

[40] Ulf Grenander and Michael I Miller. Computational anatomy: An emerging discipline. *Quarterly of applied mathematics*, 56(4):617–694, 1998.

[41] Sarang C Joshi and Michael I Miller. Landmark matching via large deformation diffeomorphisms. *Image Processing, IEEE Transactions on*, 9(8):1357–1370, 2000.

[42] Jonathan M Nichols, Abbie T Watnik, Timothy Doster, Serim Park, Andrey Kanaev, Liam Cattell, and Gustavo K Rohde. An optimal transport model for imaging in atmospheric turbulence. *arXiv preprint arXiv:1705.01050*, 2017.

[43] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[44] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[45] Lawrence Baguette. *Functional Analysis*. Marcel-Dekker, 1991.

[46] Nikolaj Kapitonovič Nikol'skij. *Functional analysis I: linear functional analysis*, volume 19. Springer Science & Business Media, 1992.

[47] Elias M Stein and Rami Shakarchi. *Functional Analysis: Introduction to Further Topics in Analysis*, volume 4. Princeton University Press, 2011.

[48] A. Aldroubi M. Unser and M. Eden. B-spline signal processing: Part i - theory. *IEEE Transactions on Signal Processing*, 41(2), 1993.

[49] Eric Todd Quinto. An introduction to X-ray tomography and radon transforms. In *Proceedings of symposia in Applied Mathematics*, volume 63, page 1, 2006.

[50] Frank Natterer. *The mathematics of computerized tomography*, volume 32. Siam, 1986.

[51] Se Rim Park, Soheil Kolouri, Shinjini Kundu, and Gustavo K Rohde. The cumulative distribution transform and linear pattern classification. *Applied and Computational Harmonic Analysis*, 2017.

[52] Izrail Moiseevič Gelfand, Mark I Graev, and N Ya Vilenkin. *Generalized functions. Vol. 5, Integral geometry and representation theory*. Academic Press, 1966.

[53] A Averbuch, RR Coifman, DL Donoho, M Israeli, and J Walden. *Fast Slant Stack: A notion of Radon transform for data in a Cartesian grid which is rapidly computable, algebraically exact, geometrically faithful and invertible*. Department of Statistics, Stanford University, 2001.

[54] Wei Wang, Yilin Mo, John A Ozolek, and Gustavo K Rohde. Penalized fisher discriminant analysis and its application to image-based morphometry. *Pattern recognition letters*, 32(15):2128–2135, 2011.

[55] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[56] MATLAB. *version 8.4.0 (R2014b)*. The MathWorks Inc., Natick, Massachusetts, 2014.

[57] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[58] Douglas G Altman. *Practical statistics for medical research*. CRC Press, 1990.

[59] Tae-Kyun Kim, Kwan-Yee Kenneth Wong, and Roberto Cipolla. Tensor canonical correlation analysis for action classification. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[60] Michael V Boland and Robert F Murphy. A neural network classifier capable of recognizing the patterns of all major subcellular structures in fluorescence microscope images of hela cells. *Bioinformatics*, 17(12):1213–1223, 2001.

[61] Yann LeCun, Corinna Cortes, and Christopher JC Burges. Mnist handwritten digit database. *AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist*, 2, 2010.

[62] Xiang Bai, Xingwei Yang, Longin Jan Latecki, Wenyu Liu, and Zhuowen Tu. Learning context-sensitive shape similarity by graph transduction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(5):861–874, 2010.

[63] Mikkel B Stegmann, BK Ersbll, and Rasmus Larsen. FAME-a flexible appearance modeling environment. *Medical Imaging, IEEE Transactions on*, 22(10):1319–1331, 2003.

[64] Zhangzhang Si and Song-Chun Zhu. Learning hybrid image templates (hit) by information projection. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(7):1354–1367, 2012.

[65] Minh N Do and Martin Vetterli. The finite ridgelet transform for image representation. *Image Processing, IEEE Transactions on*, 12(1):16–28, 2003.

[66] Stephen A Townes, Bernard L Edwards, A Biswas, DR Bold, RS Bondurant, D Boroson, JW Burnside, DO Caplan, AE DeCew, R DePaula, et al. The mars laser communication demonstration. In *Aerospace Conference, 2004. Proceedings. 2004 IEEE*, volume 2, pages 1180–1195. IEEE, 2004.

[67] Graham Gibson, Johannes Courtial, Miles J. Padgett, Mikhail Vasnetsov, Valeriy Pas'ko, Stephen M. Barnett, and Sonja Franke-Arnold. Free-space information

transfer using light beams carrying orbital angular momentum free-space information transfer using light beams carrying orbital angular momentum. *Opt. Express*, 12(22):5448–5456, Nov 2004.

[68] Matteo Oldoni, Fabio Spinello, Elettra Mari, Giuseppe Parisi, Carlo Giacomo Someda, Fabrizio Tamburini, Filippo Romanato, Roberto Antonio Ravanelli, Piero Coassini, and Bo Thidé. Space-division demultiplexing in orbital-angular-momentum-based mimo radio systems. *IEEE Transactions on Antennas and Propagation*, 63(10):4582–4587, 2015.

[69] Mauritz Andersson, Eilert Berglind, and Gunnar Björk. Orbital angular momentum modes do not increase the channel capacity in communication links. *New Journal of Physics*, 17(4):043040, 2015.

[70] Fabrizio Tamburini, Elettra Mari, Anna Sponselli, Bo Thid, Antonio Bianchini, and Filippo Romanato. Encoding many channels on the same frequency through radio vorticity: first experimental test. *New Journal of Physics*, 14(3):033001, 2012.

[71] R. Ryf, S. Randel, A. H. Gnauck, C. Bolle, A. Sierra, S. Mumtaz, M. Esmaeelpour, E. C. Burrows, R. J. Essiambre, P. J. Winzer, D. W. Peckham, A. H. McCurdy, and R. Lingle. Mode-division multiplexing over 96 km of few-mode fiber using coherent 6 ,×, 6 mimo processing. *Journal of Lightwave Technology*, 30(4):521–531, Feb 2012.

[72] Mario Krenn, Robert Fickler, Matthias Fink, Johannes Handsteiner, Mehul Malik, Thomas Scheidl, Rupert Ursin, and Anton Zeilinger. Communication with spatially modulated light through turbulent air across vienna. *New Journal of Physics*, 16(11):113028, 2014.

[73] Jian Wang, Jeng-Yuan Yang, Irfan M Fazal, Nisar Ahmed, Yan Yan, Hao Huang, Yongxiong Ren, Yang Yue, Samuel Dolinar, Moshe Tur, et al. Terabit free-space data transmission employing orbital angular momentum multiplexing. *Nature Photonics*, 6(7):488–496, 2012.

[74] Mohammad Mirhosseini, Mehul Malik, Zhimin Shi, and Robert W Boyd. Efficient separation of the orbital angular momentum eigenstates of light. *Nature communications*, 4, 2013.

[75] Timothy Doster and Abbie T Watnik. Machine learning approach to oam beam demultiplexing via convolutional neural networks. *Applied Optics*, 56(12):3386–3396, 2017.

[76] Larry C Andrews and Ronald L Phillips. *Laser beam propagation through random media*, volume 1. SPIE press Bellingham, WA, 2005.

[77] Robert L Nowack. A tale of two beams: an elementary overview of gaussian beams and bessel beams. *Studia Geophysica et Geodaetica*, 56(2):355–372, 2012.

[78] Ivan B Djordjevic. Deep-space and near-earth optical communications by coded orbital angular momentum (oam) modulation. *Optics express*, 19(15):14277–14289, 2011.

[79] Yuan Fang, Jianjun Yu, Nan Chi, Junwen Zhang, and Jiangnan Xiao. A novel pon architecture based on oam multiplexing for efficient bandwidth utilization. *IEEE Photonics Journal*, 7(1):1–6, 2015.

[80] Yongxiong Ren, Long Li, Zhe Wang, Seyedeh Mahsa Kamali, Ehsan Arbabi, Amir Arbabi, Zhe Zhao, Guodong Xie, Yinwen Cao, Nisar Ahmed, et al. Orbital angular momentum-based space division multiplexing for high-capacity underwater optical communications. *Scientific Reports*, 6, 2016.

[81] Marc Sorel, Michael J Strain, Siyuan Yu, and Xinlun Cai. Photonic integrated devices for exploiting the orbital angular momentum (oam) of light in optical communications. In *Optical Communication (ECOC), 2015 European Conference on*, pages 1–3. IEEE, 2015.

[82] Jaime A Anguita, Camilo Quezada, and Joaquin Herreros. Demonstration of multi-user laser communication using orbital-angular-momentum channels. In

SPIE Optical Engineering+ Applications, pages 851708–851708. International Society for Optics and Photonics, 2012.

[83] Mohammad Mirhosseini, Omar S Magana-Loaiza, Changchen Chen, Brandon Rodenburg, Mehul Malik, and Robert W Boyd. Rapid generation of light beams carrying orbital angular momentum. *Optics express*, 21(25):30196–30203, 2013.

[84] Ivan B Djordjevic and Murat Arabaci. Ldpc-coded orbital angular momentum (oam) modulation for free-space optical communication. *Optics express*, 18(24):24722–24728, 2010.

[85] Hemani Kaushal, Subrat Kar Kar, and V. K. Jain. *Free Space Optics Communications*. Springer India, 2017.

[86] Graham Gibson, Johannes Courtial, Miles J. Padgett, Mikhail Vasnetsov, Valeriy Pas'ko, Stephen M. Barnett, and Sonja Franke-Arnold. Free-space information transfer using light beams carrying orbital angular momentum. *Opt. Express*, 12(22):5448–5456, Nov 2004.

[87] Martin PJ Lavery, Gregorius CG Berkhout, Johannes Courtial, and Miles J Padgett. Measurement of the light orbital angular momentum spectrum using an optical geometric transformation. *Journal of Optics*, 13(6):064006, 2011.

[88] MS Soskin, VN Gorshkov, MV Vasnetsov, JT Malos, and NR Heckenberg. Topological charge and angular momentum of light beams carrying optical vortices. *Physical Review A*, 56(5):4064, 1997.

[89] Martin PJ Lavery, Fiona C Speirits, Stephen M Barnett, and Miles J Padgett. Detection of a spinning object using light's orbital angular momentum. *Science*, 341(6145):537–540, 2013.

[90] Jonathan Leach, Miles J Padgett, Stephen M Barnett, Sonja Franke-Arnold, and Johannes Courtial. Measuring the orbital angular momentum of a single photon. *Physical review letters*, 88(25):257901, 2002.

[91] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

[92] RG Lane, A Glindemann, JC Dainty, et al. Simulation of a kolmogorov phase screen. *Waves in random media*, 2(3):209–224, 1992.

[93] L.C. Andrews. An analytical model for the refractive index power spectrum and its application to optical scintillations in the atmosphere. *Journal of Modern Optics*, 39(9):1849–1853, 1992.

[94] Zhen Qu and Ivan B Djordjevic. Ldpc-coded oam based fso transmission system in the presence of strong atmospheric turbulence. In *Signals, Systems and Computers, 2015 49th Asilomar Conference on*, pages 999–1002. IEEE, 2015.

[95] Ivan B Djordjevic and Zhen Qu. Coded orbital angular momentum modulation and multiplexing enabling ultra-high-speed free-space optical transmission. In *Optical Wireless Communications*, pages 363–385. Springer, 2016.

[96] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[97] Y. Brenier. The least action principle and the related concept of generalized flows for incompressible perfect fluids. *Journal of the American Mathematical Society*, 2(2):225–255, 1989.

[98] Madelung Erwin. Quantum theory in hydrodynamic form. *Zeit. f. Physics*, 40:322–326, 1927.

[99] J.-G. Liu, R. L. Pego, and D. Slepčev. Least action principles for incompressible flows and optimal transport between shapes. *https://www.math.cmu.edu/cna/Publications/publications2016/papers/16-CNA-004.pdf*, N/A:N/A, 2016.

[100] S. Kolouri, S. Park, M. Thorpe, D. Slepčev, and G. K. Rohde. Transport-based analysis, modeling, and learning from signal and data distributions. *N/A*, N/A:N/A, 2017.

[101] Andrei N Kolmogorov. The local structure of turbulence in incompressible viscous fluid for very large reynolds numbers. In *Dokl. Akad. Nauk SSSR*, volume 30, pages 301–305. JSTOR, 1941.

[102] Karel Johannes Gerardus Hinnen. *Data-driven optimal control for adaptive optics*. PhD thesis, TU Delft, Delft University of Technology, 2007.

[103] Robert J Noll. Zernike polynomials and atmospheric turbulence. *JOsA*, 66(3):207–211, 1976.