Image Analysis Applications of the Maximum Mean Discrepancy Distance Measure

by

Michael Diu

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Electrical and Computer Engineering

Waterloo, Ontario, Canada, 2013

I hereby declare that I am the sole author of this thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Abstract

The need to quantify distance between two groups of objects is prevalent throughout the signal processing world. The difference of group means computed using the Euclidean, or ℓ_2 distance, is one of the predominant distance measures used to compare feature vectors and groups of vectors, but many problems arise with it when high data dimensionality is present. Maximum mean discrepancy (MMD) is a recent unsupervised kernel-based pattern recognition method which may improve differentiation between two distinct populations over many commonly used methods such as the difference of means, when paired with the proper feature representations and kernels. MMD-based distance computation combines many powerful concepts from the machine learning literature, such as data distribution-leveraging similarity measures and kernel methods for machine learning.

Due to this heritage, we posit that dissimilarity-based classification and changepoint detection using MMD can lead to enhanced separation between different populations. To test this hypothesis, we conduct studies comparing MMD and the difference of means in two subareas of image analysis and understanding: first, to detect scene changes in video in an unsupervised manner, and secondly, in the biomedical imaging field, using clinical ultrasound to assess tumor response to treatment. We leverage effective computer vision data descriptors, such as the bag-of-visual-words and sparse combinations of SIFT descriptors, and choose from an assessment of several similarity kernels (e.g. Histogram Intersection, Radial Basis Function) in order to engineer useful systems using MMD. Promising improvements over the difference of means, measured primarily using precision/recall for scene change detection, and k-nearest neighbour classification accuracy for tumor response assessment, are obtained in both applications.

Acknowledgements

I would like to thank all the people who made this possible. First, my adviser, Prof. Mohamed Kamel, for your guidance and support, and for taking the initial leap of faith. To Mehrdad, soon to be Dr. Gangeh, a debt of gratitude for mentoring me throughout my studies; your papers and presentations are great footsteps to follow in. Special acknowledgments are due to CPAMI colleagues Céline and Meena; grad school would have been tougher without your social and intellectual support. Prof. Oleg Michailovich for inspiring teaching, scholarship, and always useful advice. I'm fortunate to have secured the substantial image processing and biomedical engineering expertise of committee members Professors Zhou Wang and Dan Stashuk, who provided many useful review comments across all areas of the thesis. Any remaining errors are my own, of course.

The assistance of Sunnybrook Health Sciences Centre in supplying a dataset used for our core experiments is gratefully acknowledged.

Last, but certainly not least, my incredibly tolerant wife, Maggie.

Dedication

To Mom and Dad.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AUC	Area Under Curve
BOW	Bag of Words
CDF	Cumulative Density Function
EMD	Earth Mover's Distance
FLD	Fisher Linear Discriminant
GLCM	Grey Level Co-occurrence Matrix
HIK	Histogram Intersection Kernel
HSIC	Hilbert-Schmidt Independence Criterion
KCD	Kernel Change Detection
KDE	Kernel Density Estimation
KFDR	Kernel Fisher Discriminant Ratio
kNN	k-Nearest-Neighbour
LDC	Linear Discriminant Classifier
LLC	Locally constrained Linear Coding
MBF	Mid-Band Fit
MMD	Maximum Mean Discrepancy
MRI	Magnetic Resonance Imaging

PCA	Principal Component Analysis
PDF	Probability Density Function
QUS	Quantitative Ultrasound
RBF	Radial Basis Function
RKHS	Reproducing Kernel Hilbert Space
RMMD	Robust Maximum Mean Discrepancy
ROC	Receiver-Operator Curve
ROI	Region of Interest
SAD	Sum of Absolute Differences
SIFT	Scale Invariant Feature Transform
SPM	Spatial Pyramid Matching
SVC	Support Vector Classifier
SVM	Support Vector Machine

List of Symbols

Symbol	**Meaning**
§1.2	reference to a section or chapter in the text
$\hat{\cdot}$	an empirical estimate of the dotted quantity
A	an uppercase letter denotes a matrix
$K \succcurlyeq 0$	matrix K is positive semidefinite
$K \succ 0$	matrix K is positive definite
D	histogram size
d	the dimensionality of a vector; i.e. X is a $m \times d$ matrix
$d(\mathbf{x}, \mathbf{y})$	dissimilarity (distances are a subset of dissimilarities)
$\mathbf{E}_x[\mathbf{x}']$	the expectation of $\mathbf{x}'$ with respect to distribution X
$F^d(x; X)$	cumulative distribution function of the d-th dimension of x on X
$\mathcal{F}_t[f]$	a *functional* operating on function f
$\mathcal{H}$	(reproducing kernel) Hilbert space
h_i	the value of the i-th bin of a histogram
I_n	the $n \times n$ identity matrix
K	kernel matrix (Gram matrix)
$k(\mathbf{x}, \mathbf{y})$	kernel function
L	number of pyramid levels
$\lambda^{(k)}$	the k-th eigenvalue
λ	regularization parameter
ℓ_p	Minkowski distance of order p
$M\hat{M}D_b$	Empirical, biased estimate of MMD
MMD_u	Unbiased estimate of MMD
m	the number of instances in a dataset
m_1, m_2	the number of instances in sets X, Y
$\mu_{\mathbf{x}}, \mu_{\mathbf{y}}$	sample means of X, Y
$\mathbf{P}_{\mathbf{x}}, \mathbf{P}_{\mathbf{y}}$	probability distributions X, Y in the RKHS
$p_X(\mathbf{x})$	probability distribution functions of X (when disambiguation is needed)
$p(\mathbf{x})$	pdf of X (when context is obvious which distribution is meant)

Symbol	**Meaning**
$\phi(\mathbf{x})$	mapping function of x from original feature space to RKHS
$\mathbb{R}^n$	the n-dimensional space of real numbers
r	Pearson correlation coefficient
$\mathbf{\Sigma}$	covariance matrix
σ_x, σ_y	sample standard deviations of X, Y
σ	hyperparameter for a kernel
t	time index
$V^{(k)}$	the k-th eigenvector
w	window size
X, Y	two feature sets that we wish to compare
$\mathbf{x}, \mathbf{y}$	individual $1 \times d$-dimensional instances $\{x_1, x_2, \cdots, x_d\}$ from X, Y.
$\mathbf{z}$	bolded lowercase variables indicate a vector quantity

Chapter 1

Introduction

Dissimilarity measures between feature vectors, or groups of vectors, are embedded in an overwhelming majority of machine learning algorithms such as classification, regression, clustering and dimensionality reduction. It is therefore not difficult to see that due to their ubiquity, even small improvements to dissimilarity measures may have a wide impact in signal processing, image analysis and machine learning applications.

The Euclidean, or ℓ_2, distance measure has been traditionally used to compare both low- and high-dimensional data vectors, but performs poorly compared to alternatives when high data dimensionality is present. Divergence measures which compare probability distributions, such as the Kullback-Leibler divergence, are often used in the place of the Euclidean distance, but require expensive computations of integrals and empirical density estimates.

Maximum mean discrepancy (MMD) [1] is a modern unsupervised kernel-based pattern recognition method that, paired with the proper feature representations and kernels, may improve differentiation between two distinct populations over many commonly used methods such as the difference of means computed using the ℓ_2 distance. MMD-based discrimination of data sources and distance computation combines several powerful concepts from the machine learning literature: data distribution-leveraging similarity measures, and kernel methods for machine learning.

Due to this heritage, we posit that dissimilarity-based classification and visual change-point detection using MMD can lead to enhanced group discrimination and accuracy. This is far from a foregone conclusion, however; recent works such as [2] have found that in a comparative study of similarity measures for diffusion magnetic resonance imaging (MRI) analysis applications, the ℓ_2 distance, surprisingly, had superior noise robustness and sensitivity to application-specific criteria compared to over 11 other distance measures and similarity criteria.

To test our hypothesis, we research desirable properties for distance measures and conduct studies comparing MMD and the difference of means computed using the ℓ_2 distance in two subareas of image analysis and understanding: detecting scene changes in video in an unsupervised manner, and in biomedical imaging, assessing tumor response to treatment using quantitative ultrasound. Suitable features, transformations, classifiers, and kernels are selected in order to engineer a useful system.

One comment on the scope of the thesis is in order. We focus on image analysis, not image processing; our applications have the aim of extracting higher-level patterns and measurements, as opposed to outputting processed images for downstream consumption. Accordingly, we restrict ourselves to dissimilarity measures that can compare two *sets* of objects, and do not dwell on measures designed to compare just two objects, such as two images.

We bring together concepts from statistics (changepoint detection, significance tests), computer vision (colour descriptors, object descriptors), machine learning/data mining (classifiers; kernel methods), and statistical signal processing (sparse linear combinations, spectrum analysis methods) in this work.

1.1 Maximum Mean Discrepancy

The concept of maximum mean discrepancy is based on Müller's definition of an *integral probability metric* [3]. This metric was designed as a measure to compare the dissimilarity of *probability measures*[1] P, Q, and depends on finding a function f from amongst the space of functions $\mathcal{F}$ that can maximize the distance

$$d(P,Q) := \sup_{f \in \mathcal{F}} \left| \int f dP - \int f dQ \right| \tag{1.1}$$

Using the properties of a Reproducing Kernel Hilbert Space (RKHS), it is shown in §3.2 that this concept may be represented as

$$MMD := ||\mu[\mathbf{P}_P] - \mu[\mathbf{P}_Q]||_{\mathcal{H}} \tag{1.2}$$

[1] A probability measure has unit area.

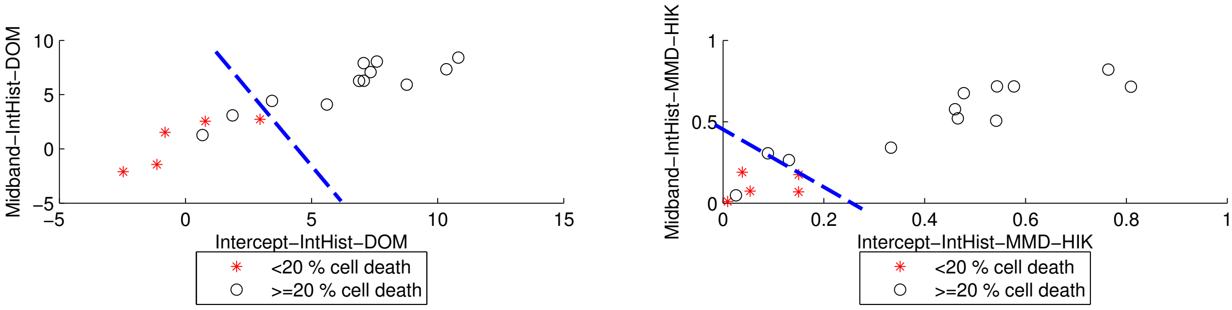


Figure 1.1: Feature plots illustrating the increased separation achieved on actual cancer treatment data. 'Midband-IntHist' and 'Intercept-IntHist' are terms specific to the dataset detailed in §5.

where $\mu[\mathbf{P}_P]$ denotes the mean of probability measure P, and similarly for Q. The symbol $||\cdot||_{\mathcal{H}}$ denotes that the norm is taken in the Hilbert space.

Thus, MMD is related to the unnormalized Fisher ratio, defined as the squared difference of group means, but MMD is computed in a higher-dimensional (possibly infinite-dimensional) RKHS. The data is mapped from the original feature space to the RKHS using a *kernel* $k(x_i, x_j)$, a positive semidefinite function which can perform nonlinear transformations on the data, thereby potentially enhancing the accuracy of linear discriminants in this alternate feature space.

This concept is illustrated in Figure 1.1. On the left, each data point represents the data for a different mouse, used as a test subject. The graph plots the value of the ℓ_2 difference of means (DOM) between features computed on pre- and post-cancer treatment ultrasound scans, $\mu_{post} - \mu_{pre}$, taken in the original feature space. Each axis represents one of the features, called 'Midband' and 'Intercept'. Some of the subjects have been treated with placebos, and we therefore expect very little distance between pre- and post-treatment images. These groundtruth labels are indicated with asterisks and circles. On the right, each point represents the MMD distance between the same two sets of imagery. Both dissimilarity measures used the same underlying feature data. Drawing a potential linear discriminant decision boundary (the dotted line) shows that fewer misclassifications may be achieved with MMD in this scenario. Furthermore, it may be seen that a larger distance between class geometric centers is present with MMD.

MMD can be computed efficiently in $O(m^2)$ time for m data samples, or *instances*, and so compared to other dissimilarity measures for distributions such as Parzen estimation or divergences, MMD is far more suited to real-time image analysis. In practice, a biased version $M\hat{M}D_b^2$ of MMD can be empirically estimated as

$$M\hat{M}D_b^2(X,Y) = \frac{1}{m^2}\sum_{i,j=1}^{m} k(\mathbf{x_i},\mathbf{x_j}) - 2k(\mathbf{x_i},\mathbf{y_j}) + k(\mathbf{y_i},\mathbf{y_j}) \tag{1.3}$$

where the b subscript denotes ‘biased’.

Does applying this data transformation to the RKHS result in a measurable and significant performance improvement on real-life pattern recognition problems, as opposed to toy examples on simulated data? This is a key question of this dissertation. We will elaborate further on RKHS, kernel methods, and MMD in the chapters to follow.

1.2 Application 1: Video Changepoint Detection

Our first application[2] of MMD is to video changepoint detection, a scenario with a time series of objects (image frames), i.e. a series of ordered objects, where we do not know the group membership of an object. A vast literature has established that scene change detection algorithms have broad application in video indexing, analytics, summarization, and compression. We apply MMD and leverage several powerful data representations from the supervised image classification world, such as bag-of-visual-words and sparse combinations of SIFT descriptors, to locate scene change points in videos with promising results.

We introduce a novel method for detecting scene changes in videos, with several desirable properties — it is unsupervised, can work in an online or offline fashion, is not sensitive to thresholds or the genre of the video, allows for decimation of framerates and resolutions for high speed processing, and enables detection of different scenes, not just shot boundaries. It is tolerant, in theory, to rotations, fast movement, and other non-semantic changes.

Our system differs from others in two main ways. First, we adopt a more modern and powerful feature descriptor, the visual bag of words [5] using densely sampled scale-invariant feature transform (SIFT) keys [6] as the base words, which ensures robustness to noise, rapid motion, rotations, colour shifts, and global brightness/contrast changes. This

[2] Copyright acknowledgement: This work is based on an accepted manuscript scheduled to appear in the image analysis conference, ICIAR [4]. The final publication will be available at http://link.springer.com.

approach has been shown to perform strongly in still-image scene recognition applications [7].

The second difference is in our use of the MMD. This kernelized distance measure allows us to efficiently use very high dimensional feature descriptors, by enabling computation of the MMD to occur in dissimilarity space and not using the original feature descriptors. The MMD is computed over the frames of a video sequence in an overlapping sliding window fashion, successively forming ‘current’ and ‘next’ groups of frames. A standard peak finding routine is used on the MMD sequence to find local maxima, which are interpreted as scene change points.

1.3 Application 2: Cancer Treatment Prognosis

The second application demonstrates MMD applied to unordered objects, where the size and membership of each group is known *a priori*. We develop a computer assisted cancer treatment prognosis system using quantitative ultrasound. Quantitative ultrasound (QUS) methods provide a promising alternative framework to non-invasively, inexpensively and quickly assess tumor response to cancer treatments using standard ultrasound equipment. We review features, feature transformations and other statistical techniques presently used in the QUS literature to differentiate between subjects responding vs non-responding to treatment. Next, the concept of using the MMD distance measure as an indicator of cell death level, and as a feature for classification is introduced. Three alternative, commonly used feature representation and distance schemes are implemented for comparison purposes.

While all tested feature representation and distance combinations showed statistically significant differences between pre- and post-treatment groups, significant improvements in both correlation to histologically determined cell death ratios and classification accuracy were observed using MMD and intensity histograms of QUS backscatter parameters in a study of mice bearing human breast cancer xenograft tumours, treated with chemotherapy, and imaged with QUS. Our system achieves classification accuracy of 84.7% when given the target of predicting if cell death in a subject is greater than 20%.

1.4 Summary of Contributions

1. *Using MMD as a feature.* Previous works have applied MMD as a two-sample statistical test, cost function, or feature selection method. We propose using MMD as a

feature in a dissimilarity-based pattern recognition framework for classification. To the best of our knowledge, this has not been previously reported in the literature.

2. *Contribution to computer vision.* We propose a scene change detection system with two elements novel to the field: the visual bag of words [5] using densely sampled scale-invariant feature transform (SIFT) keys, coupled with MMD as the similarity measure. Our system is unsupervised, unlike many alternatives. To our knowledge, only one previous work [8] has published results on time-series change detection using MMD, but this was with univariate data, not the 2-D images used here.

3. *Biomedical engineering contribution.* We present a cancer treatment prognosis system utilizing MMD and clinical ultrasound. It is able to give an early indication of the fraction of cells undergoing apoptosis (cell death) within 24 h after treatment, using the quantitative parametric maps obtained from coarse human-selected, ultrasound guided, region of interest (ROI) windows. It is one of the first works to apply supervised classification techniques on quantitative ultrasound (QUS) data. We also introduce the use of nonparametric density estimates of extracted spectrum features, as a feature for QUS analysis.

1.5 Organization of the Thesis

The thesis attempts to follow a logical order in developing the foundations and alternatives to MMD, the methods common to both proposed applications, followed by experimental results arising from our application of MMD.

After some basic definitions in §2.1, we review dissimilarity measures suited for comparing sets of instances in image analysis, their desirable properties and evaluation criteria. Next, in §2.2, we examine Reproducing Hilbert Kernel Spaces and kernel methods, which form the theoretical framework underlying MMD, and review past applications of MMD. Our main tools, dissimilarity-based pattern recognition, MMD, and nonparametric density estimation, are detailed in §3. We present the results of our work on video scene change detection in §4, and our computer assisted cancer treatment prognosis system in §5. Conclusions and future work are addressed in the final chapter, §6.

Chapter 2

Background

Before introducing the formulation of maximum mean discrepancy, we first review dissimilarity measures on distributions, kernel methods, and prior applications of MMD and its links to other statistical tools. We defer the review of application-specific background material, such as image descriptors and past approaches, to their respective sections in §4 and §5.

The motivating problem in this review can be stated thusly: given two sets of 2-D image data, of size m_1 images and m_2 images, respectively, how can we we quantify the distance between their underlying feature distributions?

We note that in image processing, many situations involve the possibility of matching portions of one image to another, i.e. a *registration* is desired. However, this is not the case in our motivating applications, either in the ultrasound parametric maps, which are 2-D maps of features derived from frequency spectra, nor in the scene change detection scenario, in which we have to detect changes in the semantic 'gist' ([9], Ch. 14) or general concept of each scene. In short, we do not have the expectation that one set of imagery should have exact visual similarity to the other. Therefore, we do not further discuss dissimilarity measures specialized for this case, such as the normalized cross correlation, difference image entropy [10], and structural similarity index (SSIM, [11]).

2.1 Dissimilarity measures

(Dis)similarity measures are at the core of most pattern recognition, machine learning, data mining, and information retrieval algorithms. They may be defined between individual variables $\mathbf{x}, \mathbf{y}$, or between distributions $\mathbf{P}_\mathbf{x}, \mathbf{P}_\mathbf{y}$. We review the definitions of these terms, comment on performance criteria for choosing dissimilarity measures, and compare the significant measures used in the literature.

Definitions

A *metric* d, also known as a *distance*, is a function $d : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ that maps its two operands, members of a set $\mathcal{F}$, onto the real number space if and only if it meets these conditions:

$$d(\mathbf{x}, \mathbf{y}) = 0 \qquad \text{iff } \mathbf{x} = \mathbf{y}\text{; coincidence axiom} \tag{2.1}$$
$$d(\mathbf{x}, \mathbf{y}) \geq 0 \qquad \text{Non-negativity} \tag{2.2}$$
$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}) \qquad \text{Symmetry / reflectivity} \tag{2.3}$$
$$d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}) \qquad \text{Triangle inequality} \tag{2.4}$$

A *dissimilarity measure* satisfies the non-negativity and coincidence axioms above, but may not be symmetrical ([12], Appendix A); distances are subsets of dissimilarities. If it additionally satisfies the triangle inequality, then it may have a geometrical interpretation, e.g. the underlying objects are embeddable in an Euclidean space, and the dissimilarity measure is a *metric*. There is a useful relation between distance metrics and norms of vector spaces. Given a norm, we can always define the following metric:

$$d(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} - \mathbf{y}|| \tag{2.5}$$

Conversely, $||\mathbf{x}|| = d(\mathbf{x}, 0)$ if the metric additionally satisfies translation invariance and homogeneity properties, that is

$$d(\mathbf{x}, \mathbf{y}) = d(\mathbf{x} + a, \mathbf{y} + a) \tag{2.6}$$
$$d(\alpha\mathbf{x}, \alpha\mathbf{y}) = |\alpha| d(\mathbf{x}, \mathbf{y}) \tag{2.7}$$

Similarity measures $s : \mathcal{F} \times \mathcal{F} \to \mathbb{R}$ obey the non-negativity axiom, but may not obey symmetry and often do not obey the triangle inequality. While *distances* may possess values in the interval $[0, \infty]$, similarities vary from $[0, 1]$, as they cannot have a similarity greater than one, and so have a natural probabilistic interpretation.

Using our definitions, similarities may be converted to distances using $d = 1 - s$ if they also obey the conditions above; another possible transformation is $s = \dfrac{1}{1 + d}$. In other words, there is no canonical transformation.

2.1.1 Selection criteria for dissimilarity measures

A review of surveys and comparisons of dissimilarity measures for image processing applications ([2] for diffusion tensor magnetic resonance imaging (MRI) data, [10] for registering computed tomography (CT) data, and [13] for image and texture retrieval) suggests that while there is no single ‘best’ dissimilarity measure, we can define several properties of an ‘ideal’ measure:

It is metric. When the measure has the metric property, we may accurately compute the geometric mean of a set of distances. This is essential for many dissimilarity-based pattern recognition algorithms that depend on class centroids (§3.1).

Appropriate computational complexity. The timing deadlines of the task may not permit the most complicated measures to be used. Does the dissimilarity measure involve parallelizable operations; does it need explicit estimates of the probability distribution function, logarithms, square roots, or exponents? Is the measure iterative or require solving optimization problems? We may further divide complexity into online and offline components; an example of the latter is precomputing sample averages or variances.

Selective invariance. Recognizing that certain tasks need to be sensitive to features that other tasks regard as noise, this includes measures such as robustness to scaling, rotations and other affine transforms, normalizing by intra-class variance, and ignoring ‘noisy’ dimensions. The relative positions of the histogram bins may be ignored, or exploited to encourage correlations between bins. We can see that the ‘dissimilarity selection’ problem is similar to the ‘feature selection’ problem.

For very high-dimensional descriptors, Aggarwal [14] argues that (dis)similarity can no longer be viewed as a higher-dimensional extension of proximity in the Euclidean sense. Instead of loss functions that penalize non-matching bins, he argues for functions that count the number of statistically significant features that both have non-zero values. Statistically significant features are defined as those which exceed a per-feature threshold. While his work was developed in the context of text classification and information retrieval, we also see these concepts in computer vision contexts, such as max pooling [15] and the histogram intersection kernel (Eqn. 2.37).

Finally, the ideal measure has multivariate support and gives intuitive results for the application at hand.

Next, we review the mainstream dissimilarity measures used in image analysis, in order to look for similarities compared to MMD. We draw on the taxonomy and coverage of Rubner [13] and Webb [12] in categorizing distance functions.

2.1.2 Distance measures between individual objects/instances

The Euclidean, or ℓ_2 distance

The Minkowski family of distances are likely the most common distances in use across science and engineering. Formally, the Minkowski distance of order p is defined as:

$$d(X, Y) = \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p} \tag{2.8}$$

The case $p = 1$ is known as the Manhattan, or city-block distance; the $p = 0$ is the pseudo-distance of counting the number of dimensions with differences, and the case $p = \infty$, also known as the Chebyshev distance, essentially compares only the largest-magnitude dimension between X, Y. The case when $p = 2$ is the Euclidean, or ℓ_2 distance, which is of special interest to us, as it is the most commonly used Minkowski metric, the natural distance of Euclidean geometry, and the basis behind the most common loss function in statistics, the mean squared error (MSE). Specifically, the squared Euclidean distance is equal to the distance between a true value $\mathbf{y}$ and its estimate $\hat{\mathbf{y}}$, by Pythagoras' theorem.

Many drawbacks are known with the ℓ_2 distance and MSE [16, 17]; it weights all dimensions equally and does not take into account the variance of particular dimensions. One attempt to improve upon it is the Mahalanobis distance d_M, a generalization of the ℓ_2 distance, and a specialized case of the K-L divergence (§2.1.3, [18]). It utilizes the covariance matrix Σ between dimensions to decorrelate and 'whiten' the input matrix.

$$d_M(X, Y) = \big((\mathbf{x} - \mathbf{y})^T \Sigma^{-1} (\mathbf{x} - \mathbf{y}) \big)^{1/2} \tag{2.9}$$

An object may not necessarily be described with a vector of numbers from a Euclidean space. Its *descriptor* may be composed of heterogeneous variable types. *Nominal variables*, also known as categorical variables, are unordered, discrete items, such as {plane, train, car}. *Ordinal variables* are also discrete, but they are ordered; the interval between them may not be uniform; an example set is {kindergarten, high school, undergraduate}. Entirely different similarity measures are employed for binary variables, such as the simple matching coefficient (SMC) and Jaccard coefficient; they all depend on the construction of a 2-D histogram-like contingency matrix.

Dissimilarity measures between individual objects are not the focus of this work, but provide a useful context and contrast to the dissimilarity measures defined for distributions, covered in the next section. We refer interested readers to [19] for more details and examples of dissimilarity measures for non-numeric data.

2.1.3 Distances between groups of objects

In our image analysis applications, we wish to work at a higher level of abstraction than comparing individual instances, in order that we may utilize the spatial and temporal information contained in *bags* or *sets* of image-like structures. One compelling way to achieve this is to compare the feature-space means of the two groups; another is to compare the probability distribution functions (pdf) of the two sets. We organize the most commonly used dissimilarity measures in Table 2.1, according to four categories: Fisher Criterion-style distances, which measure the difference between means, measures used as test statistics, information-theoretic divergence measures, and measures that take the proximity between two histogram bins into account (which Rubner *et al.* call the *ground distance* [13]).

Fisher Criterion-style distances

The Fisher criterion, or Fisher ratio (Eqn. 2.11, where $\mu_{\mathbf{x}} = \frac{1}{m}\sum_{i=1}^{m} \mathbf{x}_i$, and $\mu_{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{y}_i$, is a popular measure that requires no probability estimation, relying on the empirical means and variance only. The Euclidean distance is often used for the same purpose (Eqn. 2.10), and may be seen as a non-normalized version of the Fisher criterion.

Dissimilarity based on Test Statistics

Test statistics are the measures used in statistical hypothesis testing to determine, within a user-specified probability $1 - \alpha$, whether the observed sample values in two groups of data samples could have occurred due to random variations, a case known as the *null hypothesis*, or if it is because the two groups have truly different underlying populations, that is, *statistically significant differences* [29]. Although the original χ^2-statistic (Eqn. 2.12) is not symmetric, this can be replaced using $d_{\chi^2 sym}(X, Y) = d_{\chi^2}(X, Y) + d_{\chi^2}(Y, X)$, an idea which holds for the other non-symmetric dissimilarities.

The Kolmogorov-Smirnov (Eqn. 2.13) and Cramér–von Mises (Eqn. 2.14) measures utilize the cumulative distribution function (CDF) $F^c(x) = \int_{x'=-\infty}^{x} p(x')dx'$, where c denotes the channel in the multivariate histogram case (e.g. red, green, blue channels). CDF-based distances are able to bypass problems with mismatched histogram bin alignment and widths, as they do not compare dimension-by-dimension. These problems will be especially apparent when sample sizes or dimensionality is small. However, X, Y must be one-dimensional.

Table 2.1: Dissimilarities used to compare groups of objects

Measure name	Formula	
Fisher Criterion-style distances		
Euclidean distance between means	$d_{\ell_2}(X,Y) = (\mu_{\mathbf{x}} - \mu_{\mathbf{y}})^2$	(2.10)
Fisher Criterion/Ratio	$d_{fish}(X,Y) = \frac{(\mu_{\mathbf{y}} - \mu_{\mathbf{x}})^2}{\sigma_y + \sigma_x}$	(2.11)
Test Statistics		
χ^2-statistic [20]	$d_{\chi^2}(X,Y) = \sum_i \frac{p(\mathbf{x}) - \hat{f}(i)^2}{\hat{f}(i)}$	(2.12)
	$\hat{f}(i) = \frac{1}{2}\left(p_X(\mathbf{x}) + p_Y(\mathbf{y})\right)$	
Kolmogorov-Smirnov [21]	$d_{KS}(X,Y) = \max_i \lvert F^c(i;X) - F^c(i;Y) \rvert$	(2.13)
Cramér–von Mises criterion [22]	$d_{CvM}(X,Y) = \sum_i \left(F^c(i;X) - F^c(i;Y)\right)^2$	(2.14)
Information-theoretic divergences		
Bhattarcharyya [23]	$d_{Bhat}(X,Y) = -\ln \sum_{\mathbf{m} \in Dom(X,Y)} \sqrt{p_X(\mathbf{m}) p_Y(\mathbf{m})}$	(2.15)
Kullback-Leibler [24, 25]	$d_{KL}(\mathbf{x} \Vert \mathbf{y}) = \sum_{m \in Dom(X,Y)} \ln\left(\frac{p_X(\mathbf{m})}{p_Y(\mathbf{m})}\right) p_X(\mathbf{m})$	(2.16)
Jeffrey / Jensen-Shannon [26]	$d_{JS}(X,Y) = \sum_i \left(p_X(\mathbf{m}) \ln \frac{p_X(\mathbf{m})}{\hat{f}(i)} + p_Y(\mathbf{m}) \ln \frac{p_Y(\mathbf{m})}{\hat{f}(i)}\right)$	(2.17)
	$\hat{f}(i) = \frac{1}{2}\left(p_X(\mathbf{x}) + p_Y(\mathbf{y})\right)$	
Distances considering ground distance		
Quadratic form [27]	$d_{QF}(X,Y) = \sqrt{(X-Y)^\top \mathbf{A}(X-Y)}$	(2.18)
Earth Mover's Distance [28]	$d_{EMD}(X,Y) = \frac{\sum_{i,j} g_{ij} d_{ij}}{g_{ij}}$	(2.19)

Dissimilarities based on information-theoretic divergences

Divergence functions $D(\cdot||\cdot)$, where $\mathbf{x}||\mathbf{y}$ denotes the non-symmetric divergence from $\mathbf{x}$ to $\mathbf{y}$, are dissimilarity measures used specifically for probability functions; they are not necessarily symmetric and may not satisfy the triangle inequality. They are distinguished from test statistics and other dissimilarities by having the property [30] that $g^{(D)}$ is positive semidefinite everywhere on the statistical manifold S (the space of all probability spaces $(\Omega, \mathcal{F}, P)$), where matrix $g^{(D)}$ is the inner product, or *Riemannian metric*.

The Jensen-Shannon divergence (Eqn. 2.17) is a symmetric version of the Kullback-Leibler divergence (Eqn. 2.16) $d_{KL}(\mathbf{x}||\mathbf{y})$. In turn, the Kullback-Leibler divergence is closely related to the *mutual information* $I(X, Y)$ between X, Y: $I(X; Y) = D_{KL}(p_{X,Y}||p_X p_X)$. They are commonly employed in object trackers, e.g. the mean shift tracker [31].

Dissimilarities considering ground distance

The Earth Mover's Distance (Eqn. 2.19) is conceptually simple, yet statistically powerful. The idea is to view the problem of comparing two distributions X, Y as determining the optimal amount of 'earth' g_{ij} to be moved from a histogram bin X_i to another bin Y_j (the inter-bin distance is d_{ij}), which is solved using linear optimization in order to minimize the overall cost $g_{ij}d_{ij}$. This elegantly deals with the resolution vs. misaligned bins tradeoff involved with higher numbers of bins, works with multivariate distributions, and allows each X, Y image to have different numbers of (possibly) differently sized bins.

EMD-based dissimilarities and their applications remain a very active area of research over ten years after publication of the original papers [28], with over 15,600 search results for "`earth movers distance`" on Google Scholar. One theoretical difficulty of using EMD is that while EMD is a metric for normalized histograms, the kernel matrix composed of EMD similarities is not guaranteed to be positive semidefinite. This means that in theory, we cannot use it as a valid kernel for machine learning applications. Nonetheless, in practice, it is still done anyway and often gives good results [32].

The computational complexity of EMD, which is $O(n^3)$, is its biggest drawback. Although enormous computational speedups have been made (e.g. [33]), at the time of writing, it was not feasible to employ EMD in most real-time applications such as object recognition.

The quadratic form distance (Eqn. 2.18) is another dissimilarity in the same vein as EMD. It uses a matrix $\mathbf{A}$ of bin-to-bin similarities, but requires no optimization to solve.

2.1.4 Discussion

An examination of the formulae in Table 2.1 reveals several commonalities. Dissimilarities rely either on the difference of means, or on the additive, multiplicative, or ratio combination of distributions. Attempts are then made to normalize them, using the variance of individual populations, or the average PDF between the two groups. Log-compression, or square roots may be employed to dampen the impact of very large magnitude differences.

The algorithm designer can decide amongst them depending on the need for high sensitivity, false positive tolerance, computational complexity, data dimensionality, and knowledge of the underlying data distributions. Some choices may be ruled out right away using these criteria, leaving a few that will need to be empirically compared.

The concept of measuring the dissimilarity between two groups of objects by comparing their distributions appeals for several reasons. It handles outliers and noisy data elegantly; as $m \to \infty, d \to \infty$, noisy data will be averaged away. Maximizing the divergence has a strong theoretical link to information theory concepts of minimizing mutual information. However, practical considerations impede obtaining the necessary probability estimates $\mathbf{P_x}$; when the training set is small, the available data is too sparse, especially for multivariate data, and we run into the *curse of dimensionality* [17]. The computed dissimilarity measures will be computationally unstable. Even when greater quantities of data are available, evaluating the divergence integrals is, practically speaking, extremely slow and unsuited for real-time operation. We discuss histograms, kernel density estimation and other nonparametric density estimation methods further in §3.3.

The MMD framework is a compromise between the 'difference of means' concept and the 'integrated difference of probabilities' concept. It follows the basic unnormalized Fisher Ratio idea of taking the squared difference of means, while we show in §2.3.4 that MMD is also equivalent to the integrated difference of the two distribution functions. The probabilities are estimated using a Parzen window, and are never explicitly calculated. Furthermore, the designer can plug in kernel functions that possess the desired traits for the task, with the aim that MMD may allow improved discrimination over and above that provided by the base kernel, or similarity value, alone.

2.2 Kernel Methods for Machine Learning

The MMD is a kernel-based distance measure, implying that its computations are reliant on inner products taken in a Reproducing Kernel Hilbert Space (RKHS). In this section, we clarify what a RKHS is, provide an introductory background overview of kernel methods, and introduce the kernel trick.

We will attempt to answer practical questions such as:

- Why do we want to work in the RKHS? Why not simply work in the original feature space?
- Which operations should be performed in the RKHS?
- Which RKHS should we use?
- How do we convert our data from the original feature space to this RKHS ?

2.2.1 What is a kernel?

A Mercer kernel is a symmetric, continuous function that maps $K : [a, b] \times [a, b] \rightarrow \mathbb{R}$; that is, a two-operand function that outputs a real number, and satisfies the following condition: Mercer's Theorem states [34] that a symmetric function $k(x, y)$ is a *Mercer kernel* or simply a *kernel*, if and only if the kernel matrix

$$K(X, Y) = [\begin{smallmatrix} k(\mathbf{x}_1, \mathbf{y}_1) & \dots & k(\mathbf{x}_1, \mathbf{y}_n) \\ & \dots & \\ k(\mathbf{x}_n, \mathbf{y}_1) & \dots & k(\mathbf{x}_n, \mathbf{y}_n) \end{smallmatrix}] \tag{2.20}$$

is positive semidefinite.

The kernel matrix K is positive semidefinite if any only if, for any choices of vectors $\mathbf{c}$,

$$\mathbf{c}^\top K \mathbf{c} \geq 0, \tag{2.21}$$

which we denote as $K \succcurlyeq 0$.

The kernel must have a closed form to be of practical significance. Any positive definite function is a reproducing kernel for a specific RKHS ([35]).

2.2.2 The Kernel Trick

The main mechanism for incorporating kernel methods into statistical pattern recognition algorithms is the *kernel trick*. The key concept is to map both operands of an inner product $\langle \mathbf{x}, \mathbf{y} \rangle$ into the RKHS, and then compute the dot product there: $\langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$. We can do this because the mapping is an *isomorphism*, between X and a Hilbert space

$\mathcal{H}$. An isomorphism has a key property: they are *injective*, meaning they are one-to-one mappings. The 'trick' aspect arises because we do not actually need to calculate, or even have a formula for, the individual mappings $\phi(x)$; we just need the kernel function $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}\rangle$, as we shall detail in §2.2.3.

In summary, the kernel trick seeks to replace matrix inner products $XX^\top$ with the kernel matrix K_x, which is composed of individual kernel entries (Eqn. 2.20). This $m \times m$ matrix of all pairwise kernel values is also known as the Gram matrix. No claim is made that the dot products performed in the RKHS output identical numerical results as the original space X, which would be pointless.

Example 1. *Regularized kernel regression.*

We provide a simple example of the kernel trick. It is well known [36] that the regularized least-squares fit of regression targets $\mathbf{y}$ and feature matrix $\mathbf{X}$ is given by

$$\hat{y}(x) = \mathbf{w}^\top \cdot \mathbf{x}, \tag{2.22}$$

where $\mathbf{w} = (\lambda \mathbf{I_n} + \mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{y}$, and $\mathbf{I_n}$ is the identity matrix. The next step is to move to the RKHS domain using the mapping $\mathbf{\Phi}$, such that $x \to \phi(\mathbf{x})$, and $\mathbf{X} \to \mathbf{\Phi}$. This yields

$$\begin{aligned} \mathbf{w} &= (\lambda \mathbf{I_n} + \mathbf{\Phi}\mathbf{\Phi}^\top)^{-1}\mathbf{\Phi}\mathbf{y} && (2.23)\\ &= \mathbf{\Phi}(\mathbf{\Phi}\mathbf{\Phi}^\top + \lambda \mathbf{I_n})^{-1}\mathbf{y} && (2.24) \end{aligned}$$

where Eqn. 2.24 is obtained [37] by applying the matrix identity for positive definite, invertible matrices P, R [38]

$$(P^{-1} + B^\top R^{-1} B)^{-1} = PB^\top(BPB^\top + R)^{-1}, \tag{2.25}$$

and then dividing by $(\lambda \mathbf{I_n})$. We next represent $\mathbf{w}$ as a weighted sum of the training instances,

$$\mathbf{w} = \sum_{i=1}^{m} \phi(\mathbf{x})(\mathbf{\Phi}\mathbf{\Phi}^\top + \lambda \mathbf{I_n})^{-1}\mathbf{y} = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}) \tag{2.26}$$

where $\alpha_i = (\mathbf{\Phi}\mathbf{\Phi}^\top + \lambda \mathbf{I_n})^{-1}\mathbf{y}$. This leads to our final result,

$$\hat{y}(x) = \mathbf{w}^\top \cdot \Phi(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i \phi(\mathbf{x}) \cdot \phi(\mathbf{x}) = \sum_{i=1}^{m} \alpha_i k(\mathbf{x}, \mathbf{x}_i), \tag{2.27}$$

where $k(x_i, x_j) = \phi(\mathbf{x_i})\phi(\mathbf{x_j})$, and kernel matrix $K := \Phi\Phi^\top$. We note several benefits: a $d \times d$ multiplication and matrix inverse has been replaced by a $m \times m$ operation; the linear kernel $k(x, y) = \mathbf{x} \cdot \mathbf{y}$ implicit in the non-kernel version (Eqn. 2.22) may be swapped out for more discriminative kernels, such as the Radial Basis Function kernel (§2.2.4); and we have reduced storage requirements, as α can be of substantially smaller size than $\mathbf{w}$. These benefits generalize to other kernelized learning algorithms that we mention in the next section.

Benefits of the Kernel Trick and Kernel Methods

Up to now, we have discussed the mechanics of mapping our data into a RKHS, and discussed which operations should be performed in the RKHS. The kernel trick allows the algorithm designer to customize, or 'plug in' a variety of application specific kernels, while leaving the underlying algorithm unchanged. It allows re-invention of many classic algorithms, such as principal component analysis (PCA), penalized regression, and dissimilarity based classification. We provide a representative list of kernelized algorithms in Table 2.2, and note that they are all well-regarded, high performing algorithms in their respective sub-fields.

Kernels and the kernel trick allow vector-space operations over data types which do not have an intuitive representation in $\mathbb{R}^n$, or even similar dimensionality to be used in these pattern recognition algorithms. Examples range from graphs, or text documents, to structured objects such as trees ([39]).

Kernels allow nonlinear transformations to be applied to the data, which can enhance separation between classes, in part by leveraging a greater number of basis functions (dimensions) in the RKHS (we see in the next section that in this function space, there may be an infinite number of eigenfunctions and eigenvalues).

2.2.3 What is a Reproducing Kernel Hilbert Space?

A Hilbert space is a *complete*, optionally infinite-dimensional vector space with a norm and inner product (Fig. 2.1). By complete vector space, we mean that for every infinite sequence of decreasing elements (formally, every Cauchy sequence), and any arbitrarily small $\epsilon \in \mathbb{R}$, we can find an index N in the sequence such that $||x_m - x_n|| < \epsilon$, and $m, n > N$.

It is not a particular space we are interested in so much as a family of hypothesis spaces, called Reproducing Kernel Hilbert Spaces (RKHS). As we will see, each pair of operands in

Table 2.2: Examples of kernelized learning algorithms

Learning problem	**Algorithm**
Dimensionality Reduction	Kernelized Principal Component Analysis (KPCA) [40] Kernel Fisher discriminant (≈ Kernel LDA) [41]
Classification	Support Vector Machines (SVM) [39] Gaussian Process Classification [42] Kernel perceptron [43] (a type of neural network)
Clustering	Kernel k-means [44]
Regression	Penalized kernel regression [45] Gaussian Process Regression [46]
Significance test	Maximum Mean Discrepancy

a kernel function is represented by a unique RKHS. Much of our treatment follows [35, 47], but we also recommend [48] and the online resources of Gretton [49] as helpful references.

The elements of this space are functions, and so the 'meta-functions' defined on this space are *functionals* $\mathcal{F}_t[f]$ that work on functions. Such a function can for example be a probability distribution. It is useful because it lets us use the vector-space concepts of min, max, sum, norm, inner product, etc. on spaces of functions.

An *evaluation functional* is a functional that evaluates all the possible functions in its Hilbert space at a point t, i.e. $\mathcal{F}_t[f] = f(t)$. It maps an input function in Hilbert space to a 1-D scalar, $\mathcal{F}_t : \mathcal{H} \to \mathbb{R}$. Note that f refers to the function, and $f(t)$ refers to it parametrized at a specific point.

A *reproducing kernel* Hilbert space is a Hilbert space where for every possible evaluation point t, all functions f are bounded by some $M > 0$, i.e. no functions have infinite values.

$$|\mathcal{F}_t[f]| = |f(t)| < M||f||_{\mathcal{H}} \tag{2.28}$$

However, it is hard to work with this definition, and we instead turn to the Riesz representer theorem.

Theorem 1. *The Riesz representer theorem. For all functions f in the RKHS, and all $\mathbf{x}_1 \in X$, where X is the original feature space (formally, some set or field), the evaluation functional of f is equal to the inner product between f and a* representer *of $\mathbf{x}_1$, a function $K_{\mathbf{x}_1}$.*

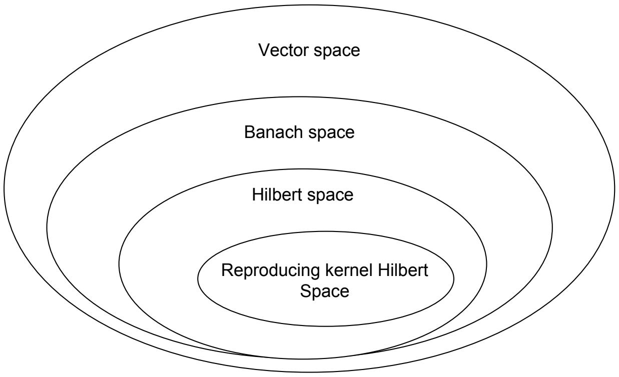


Figure 2.1: Venn diagram showing the relation between Reproducing Kernel Hilbert Spaces and other spaces.

In turn, the *reproducing property* says that this is equal to the function evaluated at that point $\mathbf{x}_1$ in the original feature space. Mathematically,

$$\mathcal{F}_t[f] = \langle K_{x_1}, f \rangle_{\mathcal{H}} = f(\mathbf{x}_1) \quad \forall f \in \mathcal{H}, \forall \mathbf{x}_1 \in X \tag{2.29}$$

Now, let us evaluate the case where the function f is equal to the representer function K_{x_1}. Using the reproducing property, and evaluating f at point $\mathbf{x}_2$ instead, we arrive at

$$K_{x_1}(\mathbf{x}_2) = \langle K_{x_2}, K_{x_1} \rangle_{\mathcal{H}}. \tag{2.30}$$

This is our first example of a *reproducing kernel*, which is any kernel that uses the reproducing property. Each reproducing kernel defines a unique RKHS, and *vice versa.* In the literature, we often see this representer of x_1, K_{x_1} denoted as $\phi(\mathbf{x_1})$, i.e. it is a mapping from X to $\mathbb{H}$, leading to

$$K(\mathbf{x}_1, \mathbf{x}_2) = \langle \phi(\mathbf{x_1}), \phi(\mathbf{x_2}) \rangle_{\mathcal{H}}. \tag{2.31}$$

Formally, the properties for a reproducing kernel are:

1. It is symmetric, $K(\mathbf{x}_1, \mathbf{x}_2) = K(\mathbf{x}_2, \mathbf{x}_1)$
2. Positive definite, that is

$$\sum_{i,j=1}^{n} c_i c_j K(t_i, t_j) \geq 0 \tag{2.32}$$

for any dimension $n \in \mathbb{N}$, and any choice of $t_1, \ldots, t_n \in X$, and $c_1, \ldots, c_n \in \mathbb{R}$.

We conclude by noting that in a function space defined by a RKHS, the familiar eigenvector equation $Av = \lambda v$ has an equivalent,

$$\int K(x, x')\phi(x')dx' = \lambda\phi(x') \tag{2.33}$$

or, equivalently, $\langle K(\mathbf{x}, \cdot), \phi\rangle_X = \lambda\phi$. Given $\mathbf{y} \in \ell_2$, an inner product in a RKHS is defined as

$$\langle \mathbf{y}, \mathbf{y}' \rangle_{\mathcal{H}_K} = \sum_{i=0}^{\infty} \frac{y_i y_i'}{\lambda_i} \tag{2.34}$$

Note that there are an infinite number of eigenvalues and eigenvectors! This is why function f evaluated at a point x can be thought of as a possibly infinite-dimension 'weighting vector' $\mathbf{w}$ for that input x, mapped into a RKHS: $f(\mathbf{x}) = \langle \mathbf{w}, \phi(\mathbf{x})\rangle$. An interesting implication of this is that higher-order moments of features may be computed via this expansion, compared to the regular inner product. When the representer represents a probability distribution in a RKHS space, as in the case with MMD, this amounts to computing higher order moments of the probability distribution.

2.2.4 Selecting kernels for image analysis

The final question from the objectives we laid out at the beginning of this section was how to select an appropriate RKHS to work in, or equivalently, how to select an appropriate kernel. Once a reproducing kernel is chosen, [35] shows how a RKHS can be formally constructed from it. We review several common kernels used in image analysis and the rationale for selecting them.

The simplest kernel is the identity mapping, which may work reasonably for high-dimensional problems [50]:

$$K_{\text{linear}}(X, Y) = \langle X \cdot Y \rangle = \sum_{j=1}^{d} x_j y_j \tag{2.35}$$

The radial basis function (RBF) kernel has good analytical properties [51] and is the default kernel for generic data [50], but performs very poorly on histogram descriptors [52] used frequently in image analysis and computer vision, our area of interest.

$$K_{RBF}(X,Y) = \exp\left\{-\frac{1}{2\sigma^2}||X-Y||_{\ell 2}^2\right\} \tag{2.36}$$

One important consideration with the RBF kernel is the selection of the hyperparameter σ. A substantial literature has built up around the optimal choice of this parameter (e.g. [53]); an often-used heuristic is to use the median distance between all entries of the kernel matrix, or some quantile of this value.

One principled method is to learn the kernel from training data, using convex optimization techniques (semidefinite programming) [54]; while very computationally intensive, it is an interesting avenue for future work in our application. In practice, the selection of the kernel depends on the nature of the features, as with the selection of dissimilarity measures (§2.1.1). We utilize several experimental surveys [32, 55] that compare kernels for different domains and descriptor types, particularly for the histogram-like descriptors we use in our work.

One kernel consistently recommended in several works ([32, 7]) for computer vision descriptors is the Histogram Intersection Kernel, which is parameter free and easy to compute.

$$K_{HIK}(X,Y) = \sum_{j=1}^{d} \min\left(x_j, y_j\right) \tag{2.37}$$

We can see that it does not penalize dissimilar dimensions, one of the high-dimensional criteria discussed in §2.1.1. We report on experiments using the RBF and HIK kernels in our tests in Chapters 4 and 5.

2.3 Previous Work on Maximum Mean Discrepancy

So far, we have reviewed the most common distance measures, and pointed out that comparing probability distributions is more discriminative than simply comparing scalar indices/measures derived from high dimensional data.

We also introduced the kernel trick, the advantages it brings, and reviewed the mathematics behind the kernel trick and RKHSs. One might assume the next logical step is to utilize kernel-based dissimilarity measures. However, kernel-based distance measures that rely on probabilities, such as the unnormalized Fisher kernel [56]

$$k(\mathbf{x},\mathbf{x}') := U_\theta^\top(\mathbf{x})U_\theta(\mathbf{x}) \tag{2.38}$$

where

$$U_{theta}(x) := -\partial_\theta \log p(\mathbf{x}|\theta) \tag{2.39}$$

still need an estimate of a conditional or marginal probability density in order to compute most probability measures for distributions, as seen in Table 2.1. As [56] points out, this requires sophisticated bias correction or feature space partitioning schemes.

By mapping probability distributions into a RKHS, this gives us tools to compare distributions without having to estimate its density, either parametrically or nonparametrically. In this subsection, we review the relevant works leading up to MMD, key applications of MMD, and its relationships to other machine learning concepts such as kernel density estimation. The formulation of MMD is deferred to §3.2. We shall see in this chapter that these tools are a means to quantify dependence between distributions (using the Hilbert Schmidt Independence Criterion, HSIC), compute metric distances between distributions (MMD), or compute the PDF $p(x)$ in a nonparametric way (using a Parzen estimator).

2.3.1 MMD as a statistical test of significance

MMD was originally proposed [57] as a family of test statistics used to determine if two sets of data vectors are statistically distinguishable. These types of tests may be generically called as *two-sample* or *homogeneity* tests, and this remains the most common use of MMD [58]. Instead of computing the probability of the null hypothesis being true, known as the *p value*, as is done in the t-test, a test threshold τ is computed for a given significance level α. If $\mathrm{MMD}(X, Y, \alpha) > \tau$, then the two samples are deemed to be statistically different. The significance level α sets an upper bound on the Type I error (the false positive rate) of the test.

Gretton *et al.* compared [1] these MMD-based tests against several alternatives, including the multivariate t-test, on the *attribute matching* problem for databases. Suppose we have have two tables, which we believe to represent the same information, but have differently named or ordered features. For example, one table has a column ‘Gender’, and the other ‘Sex’. This problem involves identifying the likeliest matches between pairs of candidate features.

They did this in two ways. The first, was to simply count the percentage of positive binary null hypothesis test results, when the matched features were truly the same. An ideal test would report 100% for cases where the features are different, and 0% when the two features involved are the same.

The second method is more relevant to our use of MMD. First, they define a separate kernel and compute MMD separately for each feature/dimension. Then, they find the optimal permutation π of features which minimizes the total sum of $\sum_{i=1}^{m} ||\mu_i(X_i) - \mu_i(Y_{\pi(i)})||^2$. Setting this expression as the cost function, they then treat it as a linear programming problem, which may be solved in $O(n^3)$ time using the Hungarian algorithm [59]. In other words, they treated MMD as a distance measure to be minimized as part of an objective function.

2.3.2 MMD Variants and Extensions

Kernel Change Detection

One related algorithm that predated MMD by several years was Kernel Change Detection (KCD), by Desobry *et al.* [60]. In it, they propose computing a quantity similar to the Fisher ratio (Eqn. 2.11) in a RKHS. Their work is based on training two soft margin single-class Support Vector Machines (a ν-SVM); one for the immediate past set of descriptors, another for the immediate future set. It is particularly relevant to our scene change detection application, which also seeks to compute a dissimilarity measure in an online, sliding window manner. Each SVM outputs a binary indicator matrix of chosen support vectors $\alpha_{\mathbf{i}}$ which is used in the proposed dissimilarity:

$$d_{\mathcal{KCD}}(X,Y) \propto \arccos\left(\frac{\alpha_1^T K_{12}\alpha_2}{\sqrt{\alpha_1^T K_{11}\alpha_1}\sqrt{\alpha_2^T K_{22}\alpha_2}}\right) \tag{2.40}$$

Eqn. 2.40 is further normalized by a measure representing the intra-class variance, which we omit here. In the case of a RBF kernel, and in the limit as $m \to \infty$, Desobry *et al.* show that

$$d_{\mathcal{H}}(X,Y)\ \overrightarrow{m \rightarrow \infty}\ \frac{g(||\mu_1 - \mu_2||_{\chi}^2)}{g(\beta_1\sigma_1^2) + g(\beta_2\sigma_2^2)} \tag{2.41}$$

where $g(u) = \arccos(\exp(-\frac{1}{2\sigma^2}u))$, and β_1, β_2 are constants.

The technique was demonstrated by segmenting the different musical notes from a recording of a church pipe organ. This is a difficult task, due to the lengthy room reverberation present in a church, as well as the rapidly changing musical score that presents abrupt frequency changes intermixed with lengthy notes.

Several criticisms of the Parzen-style approach to distance computation, which includes MMD, are offered in [60]: they point out that the empirical mean map $\mu[\mathbf{P}_x] := \mathbf{E}_x[\phi(\mathbf{x})]$

used is not robust to outliers, whereas they estimate the empirical probability distribution with a sparse weighted mean, using only the support vectors of each class. Secondly, the claim is made that the empirical covariance estimate, $\hat{\sigma}$, used to optionally normalize Eqn. 2.11, is a poor one, due to the limited window sizes needed to detect abrupt changes. Despite these valid points, KCD was shown to have similar experimental precision and recall audio segmentation performance to MMD by [61]; we also note that training two SVMs for every advance of a sliding window will be more computationally intensive than MMD. A third difference is that they advocate the use of a fixed global threshold on KCD chosen heuristically to identify the changepoints, whereas our approach in Chapter 4 relies on local peak-finding.

Kernel Fisher Discriminant Ratio

Harchaoui *et al.* [61] proposed a modification of MMD, the Kernel Fisher Discriminant Ratio (KFDR), in which they normalize the measure by the within-class covariance matrix $\hat{\Sigma}_W$. Furthermore, they perform spectral truncation on the kernel matrix, which is reminiscent of Kernel PCA [40].

$$\begin{aligned} \mathrm{KFDR}(X,Y) = && \frac{m_1 m_2}{m_1+m_2}\left\langle \hat{\mu}_2 - \hat{\mu}_1, (\hat{\Sigma}_W + \gamma I)^{-1}(\hat{\mu}_2 - \hat{\mu}_1)\right\rangle_{\mathcal{H}} && (2.42) \\ = && \mathbf{m}\mathbf{K}\mathbf{V}\boldsymbol{\Lambda}^{\mathbf{-1}}\mathbf{V}^\top\mathbf{K}^\top\mathbf{m}^\top && (2.43) \end{aligned}$$

where $\mathbf{V}, \boldsymbol{\Lambda}$ are the first p eigenvectors and eigenvalues, respectively, of the bi-centered kernel matrix, and $\mathbf{m}$ is a $m \times m$ matrix normalizing for the number of elements.

KFDR was demonstrated by segmenting audio tracks from television shows, both semantically (applause vs movie vs music), and by identifying individual speakers, which is called the *speaker diarization* problem. The authors claim substantially improved precision and recall over MMD and KCD, and comparable performance to a supervised Hidden Markov Model. While the computational cost of the eigendecompositions may be manageable for real-time operation if the window sizes are kept reasonably small, the issue we discovered in our informal experiments is that this technique may be less effective for very small window sizes. In the paper, they recommend the heuristic of keeping only the first p eigenvalues, where $p = \max\{d, \lambda_d > \epsilon\}$, and $\epsilon = 10^{-10}$. In practice, for window sizes $w \in \{10 - 30\}$, we found that only one or two eigenvectors were being removed, and thus the difference with MMD was negligible.

Robust MMD

Another attempt to make the empirical mean map more robust to outliers is from Arif and Vela [62], who proposed Robust MMD (RMMD). Like KFDR, eigenvector-based dimensionality reduction is utilized to do eigenvector decomposition. The novelty is that the PDF is approximated using an orthogonal series density estimator – that is, a weighted sum of M orthonormal functions,

$$p(x) = \sum_{k=1}^{M} \omega^{(k)} \Psi^{(k)}(\mathbf{x}). \tag{2.44}$$

Given the eigendecomposition of the kernel matrix K_x into eigenvectors and eigenvalues (V_i, λ_i) respectively, the k-th orthonormal functions $\Psi^{(k)}(\mathbf{x})$ are found by using the eigenvector-eigenvalue ratio $v_i^{(k)} = V_i^{(k)} / \sqrt{\lambda_i^{(k)}}$ as a weight on a Parzen estimate:

$$\Psi^{(k)}(x) = \langle V_k, \phi(\mathbf{x}) \rangle = \sum_{i=1}^{d} v_i^{(k)} k(\mathbf{x}, \mathbf{x}_i) \tag{2.45}$$

If we think of each orthonormal function as a weighted Parzen estimate, the weight vector $\omega^{(k)}$ is obtained by averaging all such estimates:

$$\omega^{(k)} = \mathbf{E}\{\Psi^{(k)}(\mathbf{x})\} = \frac{1}{m} \sum_{i=1}^{m} \Psi^{(k)}(\mathbf{x_i}) \tag{2.46}$$

The RMMD distance between two samples is given by the distance between weights ω. One major approximation made is to skip the eigendecomposition for the second sample group Y, and instead reuse the eigenvectors derived from X. Due to this trick, it requires the two samples being compared to be of the same size, i.e. $m_1 = m_2$. This permits the approximation

$$RMMD(X, Y) := || \sum_{k=1}^{M} \omega_x^{(k)} V_x^{(k)} - \sum_{k=1}^{M} \omega_y^{(k)} V_y^{(k)} || \approx ||\omega_{\mathbf{x}} - \omega_{\mathbf{y}}||. \tag{2.47}$$

where M is the user-defined number of eigenvectors to retain; a heuristic for finding M is given in [62].

RMMD is applied to the problem of visual tracking, which is, loosely speaking, finding the visual subregion at frame $t + 1$ whose feature density has the minimum distance to a predefined subregion's target density in frame t (the tracker must be initialized in some fashion, e.g. by having the user click on an object to be tracked). The algorithm ran fairly slowly at $0.5 - 1$ fps for 320×240 video, and tracked object sizes of around 30×30 pixels. Favorable tracking results are obtained against two well-known tracking algorithms, but no quantitative comparison was conducted against MMD.

As with KFDR, the same observation about having limited effectiveness for small sample sizes applies; when no eigenvalues are removed, RMMD is equivalent to MMD.

2.3.3 Link to the Hilbert-Schmidt Independence Criterion

The Hilbert-Schmidt Independence Criterion is a measure of *similarity*, using RKHS concepts, from the same research team [63] that developed MMD. The key idea of the HSIC is to quantify the degree of dependence between two samples X, Y using $\Delta := ||\mu[\mathbf{P}_{xy}] - \mu[\mathbf{P}_x \times \mathbf{P}_y]||$ in the RKHS. Recalling that from the definition of covariance of random variables X, Y, we have

$$\sigma^2_{X,Y} = \mathbf{E}[XY] - \mathbf{E}[X]\mathbf{E}[Y] \tag{2.48}$$

and if the two random variables are independent, then $\mathbf{E}[XY] = \mathbf{E}[X]\mathbf{E}[Y]$, we see that Δ will be very small for two statistically independent sets.

Gretton *et al.* [63] show that Δ may be empirically estimated (estimates are denoted with a hat, $\hat{}$), albeit with a bias that varies with $O(m^{-1})$, and a deviation from the true expected distance bounded by $O(m^{-0.5})$, as

$$H\hat{S}IC = \hat{\Delta}^2 = \frac{1}{m^2} \text{ tr } KHLH \tag{2.49}$$

where $\text{tr}(\cdot)$ is the matrix *trace* operator, K is the kernel matrix with elements $K_{ij} := k(\mathbf{x_i}, \mathbf{x_j})$, L is a second kernel matrix with elements $L_{ij} := l(\mathbf{y_i}, \mathbf{y_j})$, $H = I - \mathbf{1}/m$, and $\mathbf{1}$ is the $m \times m$ matrix of ones; it is a centering matrix which makes the average of each kernel matrix row and column ≈ 0, i.e. the average similarity between a feature point to all other training points is zero.

Smola *et al.* [56] show that many feature extraction criteria, such as the Pearson correlation r, t-statistic, and the signal to noise ratio, may be expressed as special cases of the HSIC using specific kernels. Most pertinently, it is shown that tr $KHLH$ can lead to the distance (in the original feature space) between the two class means, i.e. the Fisher Ratio (Eqn. 2.11). This is the same idea as the MMD, but in a different feature space.

Both MMD and the HSIC arise from the key concept of embedding probabilities of a given feature space in a RKHS. The HSIC is a dependency measure, and is the *Hilbert-Schmidt norm* of the covariance operator function (the Hilbert-Schmidt norm is akin to the Frobenius matrix norm of the covariance matrix in the original feature space. The Frobenius norm, in turn, is the matrix equivalent of the ℓ_2 norm); higher values indicate more dependence. MMD is a metric distance and test statistic based on population means only; higher values indicate more independence. MMD is a drop-in replacement for other information-theoretic distance measures like the Bhattacharyya distance; the HSIC clearly is not.

2.3.4 Link to the Parzen estimator

The Parzen kernel window estimator [53] is used to obtain a PDF estimate $p(x)$ in a nonparametric fashion, i.e. without assuming a specific probability distribution for the data. It takes a one-operand window function, or kernel, $K(x)$, whose area sums to one. This kernel should not be confused with the two-operand Mercer kernel we use in the kernel methods of §2.2.

$$\hat{p}(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} h(n)^{-m} K\left(\frac{||\mathbf{x} - \mathbf{x}_i||}{h(n)}\right) \tag{2.50}$$

$h(n)$ is a sequence of positive numbers which acts as a per-instance weight. It may also be set to a constant, in which case it is the familiar smoothing parameter σ as seen in the RBF kernel.

Arif and Vela [64] point out that the link between $\mu[\mathbf{P}_u]$ (Eqn. 3.7), the mean mapping to RKHS used in MMD, and the Parzen estimator is

$$p(u) = \langle \mu[\mathbf{P}_u], \phi(\mathbf{u}) \rangle \approx \frac{1}{m} \sum_{i=1}^{m} k(\mathbf{u}, \mathbf{u}_i) \tag{2.51}$$

which clearly has the same form as Eqn. 2.50 above. An even closer linkage to MMD was reported by Anderson *et al.* in 1994 [65], who proposed the two-sample test statistic based on two Parzen estimates

$$T_{h_1 h_2} = \int \left(\hat{p}_1(\mathbf{u}) - \hat{p}_2(\mathbf{u})\right)^2 du, \tag{2.52}$$

where $\hat{p}$ is defined as in Eqn. 2.50. This has the same form as MMD^2 (Eqn. 3.12), but was arrived at without any link to RKHS methods, some twelve years before the work of Gretton [1]. The linkage to RKHS methods contributed by making the connections to other reproducing kernel based measures, such as the HSIC (Sec. 2.3.3), clearer.

Chapter 3

Methods

3.1 Dissimilarity-Based Pattern Recognition

Statistical pattern recognition tasks, such as predicting categorical labels (classification) of a test set T, can be performed using properties that characterize an object, which we call *features*, as inputs. Alternatively, the inputs to the algorithm may use dissimilarity values from T to a set of representative objects R (we will call them *prototypes*). Examples of algorithms designed for the former category include the Fisher Linear Discriminant and logistic regression [17], while algorithms designed from the outset for (dis)similarity inputs include Support Vector Machines, clustering algorithms, and the kernelized algorithms listed in Table 2.2.

Presumably, there are some benefits to working in dissimilarity space instead of feature space, otherwise there would be little profit in discussing dissimilarity measures, kernels, and MMD. We describe three such benefits here, and preview specific techniques that we use with MMD.

First, using dissimilarities, we can compare objects that have varying dimensionality; that is, we can calculate inter-object distances even when they compare from two different feature spaces. These may be two images with different numbers of colour channels or colourspaces; different shapes with varying numbers of polygons; or two video frames with varying numbers and locations of interest points (an *interest point* is a ‘landmark’ in an image, such as the intersection of two edges, used in computer vision).

Second, objects with missing features or non-continuous valued variables can be compared using dissimilarities. Some features might consist of categorical or ordinal data

(§2.1.2), which are easier to compare in pairs, than to quantify in isolation. Certain features might be unknown or unreliable in some instances, either during model training (*induction*) or testing time. Instead of requiring each algorithm to specially handle missing data, this can be abstracted away using dissimilarity measures.

Lastly and most importantly, dissimilarity based pattern recognition can, at least in certain instances, improve discrimination performance over feature-based representation. This viewpoint is perhaps most strongly espoused today by Duin and Pękalska [66, 67, 68]. One recent work on schizophrenia detection using MRI [69] reported, "classification onto the dissimilarity space shows improvements of the standard NN (nearest-neighbour) rule and the support vector classifier on the original space." The precise reasons and theory behind the improvement are still an active area of research, but we observe that dissimilarity representation offers dimensionality reduction (thus, reducing the effect of the *curse of dimensionality*), while inheriting the invariance and robustness properties of its underlying dissimilarity measure (§2.1).

Dissimilarity space representations may be further subdivided into two methods: *embedding* the dissimilarities into a Euclidean space, if the dissimilarity measure is a metric, or by treating dissimilarities directly as features for classification. Embedding of dissimilarities is related to multi-dimensional scaling (MDS) [70], in which eigenvector decomposition and truncation on the kernel matrix of similarities K is used to compute a Euclidean projection X. Classification and other operations are then performed on X as usual. Several classifiers are suggested in the literature for this pseudo-Euclidean space [67]: the generalized nearest-mean-classifier (NMC), which is akin to k-nearest-neighbours (Alg. 1), but using the nearest class centroid instead of nearest neighbour; Fisher's Linear Discriminant (FLD), and the Support Vector Classifier (SVC).

Algorithm 1 The k-nearest-neighbour classifier

```
procedure KNNC($x, k$)
    for $i = 1$ to $k$ do
        $l_i \leftarrow$ the class of the $i$-th closest prototype of $x$ in dissimilarity space
    end for
    return $mode([l_1, l_2, \cdots, l_k])$                ▷ break ties randomly
end procedure
```

The second approach, performing classification on dissimilarities directly, requires a symmetric dissimilarity measure, ideally obeying the triangle inequality. Tests [67] comparing the two approaches, involving the well-known MNIST digits dataset and another dataset of binary shapes, showed that classifying with the dissimilarity kernel directly had lowest error using a variety of classifiers (k-NN, FLD, SVC). The only times that the embedding approach was superior in classification accuracy was when the size, m, of the training set, consisting of representative objects, was sufficiently large and the embedded dimensionality, d, small.

Given that in our applications, the retrieved (target) dimensionality is small while the training set is also small, we adopt the second approach of classifying on dissimilarity values directly. Dissimilarity representation is also a natural fit to our second application, changepoint detection, as it seeks the maxima of a time-series feature. We work with the k-NN classifier, one of the simplest possible classifiers, as it is the most commonly used choice in dissimilarity space [67] and performs fairly well. In this way, attention can be focused on the relative performance differences between MMD and alternatives. We leave to future work efforts to test more advanced classifiers such as Support Vector Machines with MMD.

3.2 Maximum Mean Discrepancy

3.2.1 Formulation

After several preliminaries, we are now ready to present the formulation of the maximum mean discrepancy, for which we have used [57] and [56] as the primary reference sources.

One of the inspirations for the development of MMD was the work of Müller ([3]), who defined an *integral probability metric* on probability measures P, Q as

$$d(P,Q) := \sup_{f \in \mathcal{F}} \left| \int f dP - \int f dQ \right| \tag{3.1}$$

where $\mathcal{F}$ is the space of all functions, and where f is subject to

$$||f||_{\mathcal{H}} \leq 1 \tag{3.2}$$

That is, it was the maximum difference between the mapped distributions. We seek the feature mapping f which maximizes this; ideally, it is large on points in P, and small-valued on points in Q, and with a RKHS norm less than or equal to unity. Such a function is constrained to be “in the unit ball”, or hypersphere, of the RKHS.

The empirical MMD_b (the b subscript means biased) was therefore defined on data samples (not probability distributions) X, Y following Eqn. 3.1,

$$MMD_b(X,Y,\mathcal{F}) := \sup_{f \in \mathcal{F}} \left(\frac{1}{m_1} \sum_{i=1}^{m_1} f(\mathbf{x_i}) - \frac{1}{m_2} \sum_{i=1}^{m_2} f(\mathbf{y_i}) \right) \tag{3.3}$$

and its theoretical counterpart is

$$\begin{aligned} MMD &:= \sup_{f \in \mathcal{F}} \left(\mathbf{E}_x[f(\mathbf{x})] - \mathbf{E}_y[f(\mathbf{y})] \right) && (3.4) \\ &= \sup_{f \in \mathcal{F}} \langle \mu[\mathbf{P}_x] - \mu[\mathbf{P}_y], f \rangle_{\mathcal{H}} && (3.5) \\ &= ||\mu[\mathbf{P}_x] - \mu[\mathbf{P}_y]||_{\mathcal{H}} && (3.6) \end{aligned}$$

We used the reproducing property, Eqn. 2.29, to proceed from Eqn. 3.4 to Eqn. 3.5, as well as the property that the RKHS mapping is injective (§2.2.2). To move from Eqn. 3.5 to Eqn. 3.6, we use Eqn. 3.13 together with the geometric interpretation of the inner product: the maximum value of the inner product will be reached when there is zero angle between the vectors $\langle a, b \rangle$, i.e. when a, b are identical. The phrase 'maximum mean discrepancy' in MMD is named for this concept.

We stop for a moment, and define the *mean mapping* functional $\mu[\mathbf{P}_x]$. Its input is a function (a probability distribution), and its output is a scalar, as expected. The functional is defined as the expectation of the evaluation functional – empirically, the sample mean of the mapped data $\phi(x)$.

$$\mu[\mathbf{P}_x] := \mathbf{E}_x[k(\mathbf{x}, \cdot)] = \mathbf{E}_x[\phi(\mathbf{x})] \qquad (3.7)$$

Its empirical counterpart is

$$\mu[X] := \frac{1}{m} \sum_{i=1}^{m} k(\mathbf{x_i}, \cdot) \qquad (3.8)$$

We still require an empirical formula to evaluate Eqn. 3.6, and so proceed as follows:

$$\begin{aligned} MMD^2(\mathbf{P}_x, \mathbf{P}_y) &= ||\ \mu[\mathbf{P}_x] - \mu[\mathbf{P}_y]\ ||^2_{\mathcal{H}} && (3.9) \\ &= \langle \mu[\mathbf{P}_x] - \mu[\mathbf{P}_y], \mu[\mathbf{P}_x] - \mu[\mathbf{P}_y] \rangle && (3.10) \\ &= ||\mathbf{E}[\phi(\mathbf{x})] - \mathbf{E}[\phi(\mathbf{y})]||^2 && (3.11) \\ &= \mathbf{E}\{(\phi(\mathbf{x}) - \phi(\mathbf{y}))(\phi(\mathbf{x}) - \phi(\mathbf{y}))\} \\ &= \mathbf{E}\{\phi(\mathbf{x})\phi(\mathbf{x}) - \phi(\mathbf{y})\phi(\mathbf{x}) - \phi(\mathbf{x})\phi(\mathbf{y}) + \phi(\mathbf{y})\phi(\mathbf{y})\} \\ &= \mathbf{E}_{x,x'}[k(\mathbf{x}, \mathbf{x}')] - 2\mathbf{E}_{x,y}[k(\mathbf{x}, \mathbf{y})] + \mathbf{E}_{\mathbf{y},\mathbf{y}'}[k(\mathbf{y}, \mathbf{y}')] && (3.12) \end{aligned}$$

where Eqn. 3.10 is obtained by recalling that all inner product spaces have a naturally defined norm,

$$||\mathbf{x}|| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}. \qquad (3.13)$$

An unbiased version of Eqn. 3.12, $M\hat{M}D_u^2$, where u denotes unbiased, can be empirically estimated, assuming $m_1 = m_2 = m$, using

$$M\hat{M}D_u^2(X,Y) = \frac{1}{m(m-1)} \sum_{i \neq j} k(\mathbf{x_i}, \mathbf{x_j}) - k(\mathbf{x_i}, \mathbf{y_j}) - k(\mathbf{y_i}, \mathbf{x_j}) + k(\mathbf{y_i}, \mathbf{y_j}) \tag{3.14}$$

This is an unbiased estimate (a *U-statistic* in the statistics literature). However, for our proposed applications, we require a biased estimate (known as a *V-statistic*) of MMD in order that it fulfills the second condition of a metric, non-negativity (Eqn. 2.2). Using the fact that the kernel function is symmetric, we finally obtain **the main formula used in this thesis**,

$$M\hat{M}D_b^2(X,Y) = \frac{1}{m^2} \sum_{i,j=1}^{m} k(\mathbf{x_i}, \mathbf{x_j}) - 2k(\mathbf{x_i}, \mathbf{y_j}) + k(\mathbf{y_i}, \mathbf{y_j}). \tag{3.15}$$

Unless otherwise specified, we adopt the simplifying convention that empirical results referring generically to ‘MMD’ utilize $M\hat{M}D_b^2$.

Metric property

Eqn. 3.15 has the metric property. Recalling the definition of a *distance* or *metric* in §2.1, the coincidence axiom follows by inspection; if $X = Y$, MMD^2 is obviously zero. That the converse also holds requires a more technical proof, and is shown in ([57], theorem 5). Symmetry is obvious by inspection, since the reproducing kernel is also symmetric. The triangle inequality is slightly more complicated to prove, and is shown in ([57], theorem 18). Finally, non-negativity is is implied by the other three axioms. □

Convergence bounds

The following convergence bound, proved in [57], allows us to approximate the distance between populations $D(\mathbf{P}_x, \mathbf{P}_y)$ using just the empirical means, and means that MMD is a consistent estimator [1]. It holds that under certain conditions, the error between finite samples will be bounded by $O(m^{-0.5})$. The conditions are that the Rademacher average $R_m(\mathcal{F}, \mathbf{P})$, a measure used in defining statistical learning bounds, is itself $O(m^{-0.5})$ bounded.

Theorem 2. $||\mu[\mathbf{P}_x] - \mu[X]|| \leq 2R_m(\mathcal{H}, \mathbf{P}_x) + R\sqrt{-m^{-1}\log(\delta)}$, *with probability at least* $1 - \delta$, *where* m = *number of instances, and* R *is an arbitrary bound.*

Smola *et al.* [56] pointed out that this concept of bounding the maximum difference between empirical and expected means is also seen in the Kolmogorov-Smirnov statistic (Table 2.1, Eqn. 2.13).

3.2.2 MMD, a distance

To date, MMD has been utilized as a two-sample test statistic (§2.3.1), as a clustering criterion ([71]), as a feature selection method ([72]), and as a cost function for *domain transfer learning* [56, 73]. (Real-life machine learning algorithms are usually tested on data with different underlying feature distributions than they were trained on. Domain transfer learning is the sub-field of machine learning that tries to reweight the available training data to minimize the expected loss on the test set.)

MMD as a distance

In our first application, video scene change detection (§4), we utilize MMD as a distance, for the problem of *changepoint detection*. Changepoint detection can be summarized as the task of detecting when the underlying data distribution has changed 'significantly'. The peaks of MMD in a local window are used to identify the first frame of a new scene. While a recent work [8] also used MMD for changepoint detection in electroencephalogram (EEG) data, our method has two key differences. First, we use a different method of determining window sizes, rather than exhaustively trying all combinations and second, we search for visual changepoints in a time series of 2-D data, not of 1-D data.

MMD as a feature for supervised learning

One of our main contributions is to instead propose using MMD as a feature in a dissimilarity-based, supervised pattern recognition framework (§3.1), which is applied in §5. To the best of our knowledge, this has not been previously reported in the literature. We use the k-nearest-neighbour classifier (Alg. 1), in the dissimilarity framework, without embedding the MMD distances in a Euclidean space, but rather using them directly. As MMD is a metric, we conjecture that it should perform well.

One may wonder why we don't simply use the MMD hypothesis tests in our visual changepoint and prognosis assessment applications, in place of the approaches to be introduced in Chapters 4 and 5. We present three justifications.

Firstly, by using the test statistic itself as a feature, which contains distance information not present in the binary significance decision, we may combine it with supervised learning approaches to enhance group discrimination. Supervised learners can add context, such as the distance levels needed to achieve utility in a given application (for example, the distance needed to indicate clinical significance). Secondly, the temporal or spatial behavior of the test statistic, such as its local minima and maxima, contains information not captured in the two-sample test. Lastly, we can combine the MMD distance with other features to again increase context and discrimination.

We provide more implementation details on using MMD as a distance and as a feature, in their respective application chapters, §4 and §5.

3.3 Nonparametric Density Estimation

Probability density estimates, used as features, can be an effective data descriptor for a large, multidimensional object, such as an image. We next address the question of how to obtain these estimates.

Remark. It may appear a bit unusual that we have advocated MMD as a distance that can avoid explicit density computation, but then return to the topic of nonparametric density estimation. In fact, there is no contradiction, because the density estimation described here is performed *within* an object, whereas we advocate for the use of MMD *between* objects. We do not use parametric density estimation to represent the entire set of objects (e.g. images) as a single histogram; rather, we collect an array of histograms.

Parametric density estimation refers to the practice of assuming a data sample follows a certain data distribution, allowing us to use the *sufficient statistics* of that distribution to estimate the probability density function (PDF). In the example of a multivariate Gaussian, the sufficient statistic is comprised of the sample mean $\hat{\mu}$ and variance $\hat{\Sigma}$, giving:

$$p(x) = \frac{1}{\sqrt{(2\pi)^m|\hat{\mathbf{\Sigma}}|}} \exp\left(-\frac{1}{2}(\mathbf{x}-\hat{\mu})^\top \hat{\Sigma}^{-1}(\mathbf{x}-\hat{\mu})\right). \tag{3.16}$$

If the data truly is sampled from this assumed distribution, then we have successfully managed to condense our representation enormously – we have performed data compression. This rarely holds in complex domains such as image and audio processing; the data may have multiple modes, each following a different distribution. In this case, *mixture models* of distributions may be estimated using techniques such as expectation maximization [74], but this is iterative and computationally expensive.

Due to the very high intrinsic dimensionality and unconstrained input environments of these domains, we believe it inappropriate to assume a distribution on the data. As an alternative, we use nonparametric density estimates of the features to represent the data in our experiments with video and ultrasound data. No assumptions are made about an underlying distribution of the data (whether it be Gaussian, Poisson, etc.). Nonparametric models support multimodal data, without the complexities of estimating mixture models, although one disadvantage of such representations is that they are frequently higher-dimensional than a parametric model.

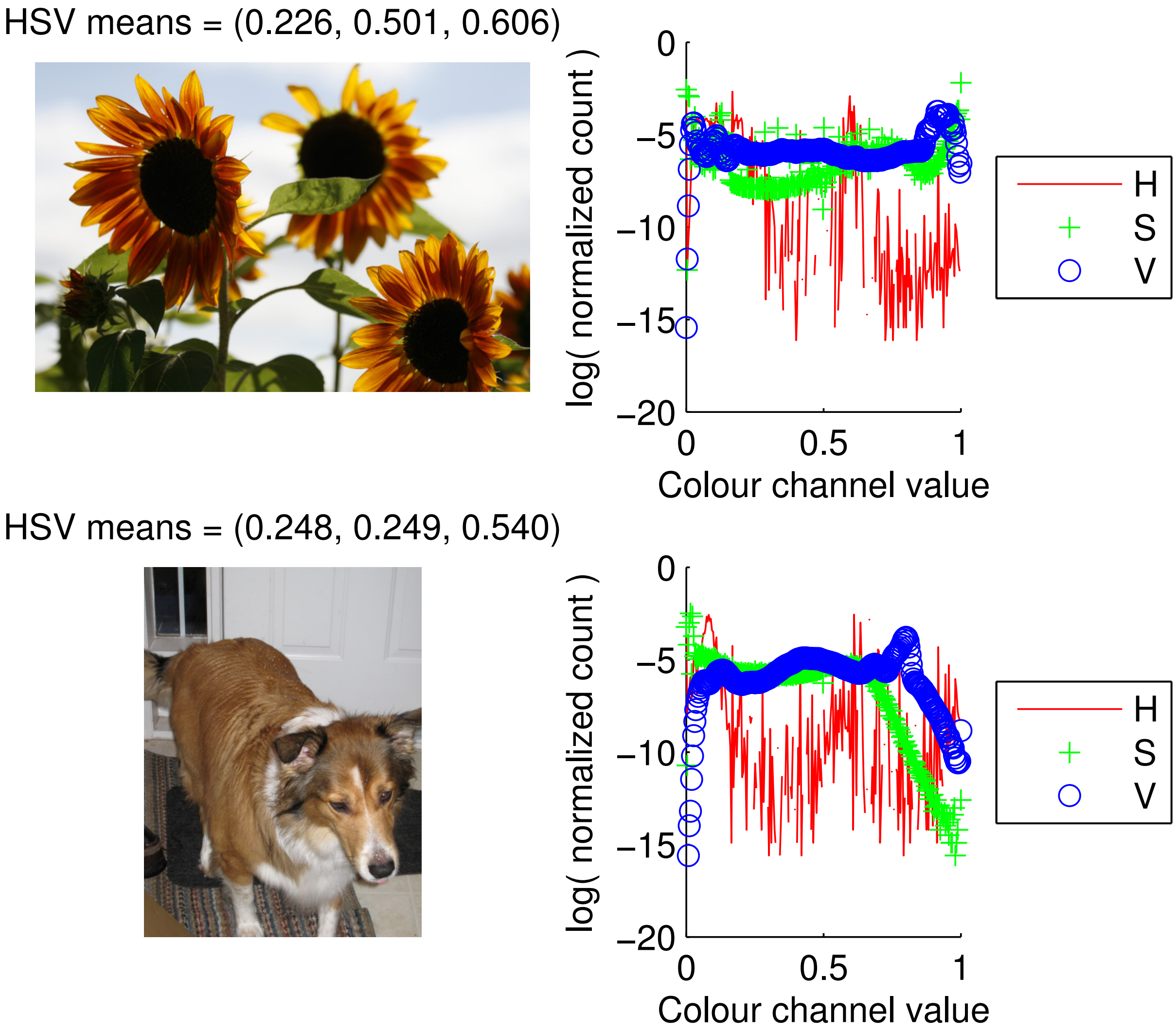
This concept is illustrated in Figure 3.1, using two images and their Hue, Saturation, Value (HSV) representations. Using the scalar mean of each colour channel as the feature (e.g. $\mu_{dog} = [\mu_H \; \mu_S \; \mu_V]$), the ℓ_2 distance between the two image means is 0.0219, which, in the context of colour channels each normalized to a range of [0–1.0], provides insufficient discrimination between the very different images. Taking the greyscale difference of means instead does not help, with a distance of 7.14 (in a range of [0–255]). A dissimilarity measure that compares the estimated densities of each image on a local subregion-by-subregion basis should be more discriminative.

Several leading methods for nonparametric density estimation include histograms, regression fits based on histograms (e.g. smoothing splines [17]) and kernel density estimators.

Kernel density estimators (KDEs) such as the Parzen estimator (Eqn. 2.3.4) generate smooth probability estimates that are averages of nearby points, with the closest values weighted most strongly. Two decisions must be made, the choice of kernel and the value of the kernel hyperparameter σ. While satisfactory algorithms exist to generate KDEs in the 1-D case, the multivariate case is still a matter of active research [75]. This reason, along with the extra computational load of KDEs (there are at least $O(m^2)$ more calculations), influenced our decision to use histograms for density estimation.

3.3.1 Histograms

One of the most popular, if not the most popular, density estimation method is the histogram, which was first described in the published scientific literature under this name by Pearson in 1895 [76]. They are rotation invariant, scale invariant, simple to compute and have many distance metrics suited for comparing distributions using them.


HSV means = (0.226, 0.501, 0.606)

HSV means = (0.248, 0.249, 0.540)

Figure 3.1: The ℓ_2 distance between group means often performs poorly on high dimensional data when trying to distinguish two distinct populations (here, the flower vs. dog ℓ_2 distance is only 0.0219). Using nonparametric density estimates (right), and dissimilarity measures that operate on them, may be more effective.

More formally, given a collection of m samples $X = \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m$, we define the histogram with D bins as a vector of natural numbers $\mathbf{h}$ that meets the constraint $m = \sum_{i=1}^{D} h_i$, where h_i is the value of bin i. The samples $\mathbf{x}_i$ may be scalars or multivariate.

As there can be varying number of instances (pixels or samples), the histograms are normalized to sum to one, using

$$h_i^{'} = \frac{h_i}{\sum_{i=1}^{D} h_i} \tag{3.17}$$

Additional measures used to reduce data dimensionality include clamping all the values above x_{max} and below x_{min} to the last and first bin, respectively. This is useful when the variable is a real number and is not inherently quantized.

In a one-dimensional histogram, it is straightforward to find the correct bin to increment, given the underlying variable's value. However, when the underlying data is multivariate, then vector quantization is frequently used for $d \geq 2$, as it is clearly a case where we run into the *curse of dimensionality* [17]. Vector quantization methods, such as k-means clustering [19], identify k representative *prototypes* in the original feature space (akin to finding the Voronoi tessellation); the ℓ_2 distance is then used to find the nearest prototype for each test instance. Each prototype is assigned one of the histogram's bin indices, while the bin ordering loses the natural interpretation it had in the 1-D case.

The question arises of how many bins are appropriate, a tradeoff between resolvability of peaks, and the risk that noise will randomly cause an instance to fall into the neighboring bin. Cross-validation measures are generally used to find an empirically acceptable result. Equal depth bins, where the observed population is allocated in equal quantities to variable-width bins, are also possible and have been found to give good results ([14]), but require sophisticated dissimilarity measures such as the EMD (Eqn. 2.19).

There are several well-known drawbacks with histograms. The spatial connectivity contained within neighbouring pixels in a certain configuration is lost, for example, when we build histograms of pixel intensities. Part of the solution was the development of a higher layer of abstraction, the *bag of words*, such that in many computer vision problems, we build histograms of *atoms*, which may be 2-D image intensity patches, SIFT keys, or textons (we introduce SIFT keys and textons in §4 and §5, respectively). Spatial Pyramid Matching [7] is an effective method to capture spatial connectivity at an even larger spatial scale using histograms.

3.4 Summary

Dissimilarity-based pattern recognition enables us to compare groups of potentially varying dimensionality, and can improve discrimination of the groups compared to operations performed in the original feature space. Simple classifiers, such as k-NN, are often used when treating dissimilarities as features for classification.

Maximum mean discrepancy is a kernelized distance measure used to quantify the dissimilarity between two groups of data. MMD may be treated as a distance, as a statistical two-sample test, or as a feature for dissimilarity-based classification. When used as a feature, we gain the benefits of supervised learning, and we are able to combine the MMD feature with other features for classification. Furthermore, by examining the time-domain behavior of MMD, we may gain additional context for classification purposes.

Comparing nonparametric density estimates can be much more discriminative vs. comparing the difference of means for two individual instances, and are very useful for complex, high-dimensional, multi-modal data such as images. Histograms are one of the most popular methods of obtaining this density estimate.

By combining different variants of the three powerful techniques presented in this chapter — dissimilarity-based pattern recognition, maximum mean discrepancy, and descriptors based on nonparametric density estimates — we are able to design statistical pattern recognition systems that may outperform the traditional difference of means approach on two very different applications – video scene change analysis and computer assisted cancer treatment prognosis.

In the next chapter, we utilize a more sophisticated version of the histogram, the bag of words, which is able to represent underlying data that is multivariate. We utilize MMD as a distance and perform dissimilarity-based peak finding to detect scene changes in videos.

Dissimilarity-based classification, using the maximum mean discrepancy distance as a feature, computed on histogram estimates of the data density, is the essence of our system in §5 to assist clinicians with determining the prognosis of cancer treatment using sets of ultrasound imagery.

Chapter 4

Application 1: Video Scene Change Detection

Utilizing MMD as a distance, on time series data

4.1 Introduction

A vast literature has established that scene change detection algorithms have broad application in video indexing, analytics, summarization, and compression (see [77] for one representative survey). We introduce a novel method for detecting scene changes in videos, with several desirable properties — it is unsupervised, can work in an online or offline fashion, is not sensitive to thresholds or the genre of the video, allows for decimation of framerates and resolutions for high speed processing, and enables detection of different scenes, not just shot boundaries. It is tolerant to rotations, fast movement, and other non-semantic changes.

Work in this field has been underway for many years, under names such as video summarization, keyframe extraction, shot boundary detection, and changepoint detection. Inspired by the nomenclature of [78], such systems may be broadly decomposed and distinguished by the features used, their spatial support, the feature similarity metric employed, the region of temporal support chosen, and finally the boundary detection method.

Our system differs from others in two main ways. First, we adopt a more modern and powerful feature descriptor, the visual bag of words [5] using densely sampled scale-invariant feature transform (SIFT) keys [6] as the base words, which ensures robustness to noise, rapid motion, rotations, colour shifts, and global brightness/contrast changes. This approach has been shown to perform strongly in still-image scene recognition applications [7]. Secondly, we use a kernelized distance metric, the maximum mean discrepancy (MMD) [1] that is computationally simple, involving the inner product between the difference in means of the two distributions, yet statistically powerful, because these distributions are mapped into a high-dimensional, nonlinear feature space using kernels, whereupon the means are estimated via the Parzen estimator. The kernel representation allows us to efficiently use very high dimensional feature descriptors, by enabling computation of the MMD to occur in dissimilarity space and not using the original feature descriptors.

The MMD is computed over the frames of a video sequence in an overlapping sliding window fashion, successively forming 'current' and 'next' groups of frames. A standard peak finding routine is used on the MMD sequence to find local maxima, which are interpreted as scene change points. Previous work has demonstrated the usefulness of MMD and closely related variants in tasks such as speaker discrimination [61]; for segmenting musical notes [60], and EEG/ECG data [8]; and in matching small image patches for visual tracking [62].

To the best of our knowledge, neither of these two elements (the feature descriptor or similarity measure) have been previously proposed for scene boundary detection, although high-level video feature extraction [79] uses similar descriptors to summarize an entire video, obtaining its 'gist', without localizing the endpoints of individual scenes.

4.1.1 Previous Work

A review of the recent literature and surveys from the last decade ([77, 78, 80]) suggests continued research interest and activity in shot boundary and keyframe extraction techniques; fifty-seven groups and approaches participated in the annual NIST TRECVID [81] shot boundary detection competition held between 2001 and 2007. We suspect this popularity is in part because this application provides a nice testbed for new theoretical approaches in feature representations, dimensionality reduction techniques, distance measures, classifiers and clustering algorithms. Commercial applications of video indexing/summarization techniques include the keyframe summary that is available in the Google Youtube video service, and the multimedia indexing and search features for large corporate video collections in Autonomy Virage [82].

Based on the final TRECVID SBD workshop report [81], we may remark on some commonly observed characteristics of the top-performing shot boundary detection systems. They are based on local or global changes, as measured by some similarity measure, of raw pixel feature vectors, colour histograms, or 2D-transformed versions of the frame (e.g. Fourier transform, wavelets), or of the frame's edges. One recent, representative work is in [83], which used wavelet features modeled using a generalized Gaussian distribution, and compared with the Kullback-Leibler divergence. The local maxima of the divergence in a fixed size window is marked as a shot boundary (*cluster* in their terminology) if it exceeds a threshold determined experimentally. However, a 'shot segment' or cluster derived using such an approach does not necessarily correspond to notable semantic changes in the video; rather, new shot segments may be formed due to camera or object motion.

A more general concern shared with other approaches, particularly parametric modeling approaches, is that there are a great number of parameters to be set. One way to handle this is to learn these parameters in a supervised manner; seven out of the top ten systems in the TRECVID 2005 competitions used SVMs [81], which require manually annotated groundtruth and may limit the system's ability to generalize to other video content genres.

Statistical models of scene change probability can also be built using supervised training data, and an *a posteriori* estimate of scene change probability computed, as in [84]. However as one might imagine, they are highly content-dependent (consider sports vs. nightly news) and also medium-dependent; that is, very different models may be seen in movies vs. surveillance video vs. mobile robotics.

Our approach differs from these works in major ways that have not been previously considered in the scene recognition field. We use a visual descriptor more robust to inter-frame motion, translation, and colour and lighting shifts, borrowing from the scene recognition world, in order that the similarity measures may discriminate between different scenes more precisely while being less sensitive to intra-scene changes.

4.2 Methodology

4.2.1 A Spatial Relationship-Preserving Scene Descriptor

We first turn our attention to the base feature vector used in the MMD computations when comparing one group of frames against another, that is, the scene descriptor that describes each frame. Given a greyscale input frame, SIFT keys are first computed on a dense grid spacing. The closest match to this key is found from a linear combination of visual 'words' in a dictionary, using a technique called locality-constrained linear coding (LLC) [15], and the coefficients of the linear combination used to increment each word's histogram bin by an amount proportional to their coefficients, thus forming the descriptor for the image. Rather than choosing a combination of words that minimizes ℓ_1 or ℓ_2 constraints, as in many sparse coding approaches, LLC regularizes the constraint equation using the distances between keys $\mathbf{x}_i$ and atoms $\mathbf{d}_i$, as shown in (4.1), where $\mathbf{B}$ is the dictionary of words, and $\mathbf{c}_i$ the coefficient vector ($\odot$ represents element-by-element multiplication).

$$\min_{\mathbf{C}} \sum_{i=1}^{m} ||\mathbf{x}_i - \mathbf{B}\mathbf{c}_i||^2 + \lambda ||\mathbf{d}_i \odot \mathbf{c}_i||^2 \qquad s.t.\ \mathbf{1}^T \mathbf{c}_i = 1, \forall i \tag{4.1}$$

The coefficients are constrained to be shift-invariant, which was found in [15] to have highest accuracy amongst the alternatives evaluated. The dictionary is built offline using the k-means unsupervised clustering technique with a dictionary size D. Larger dictionary sizes led to higher precision and recall, but are limited by increasing clustering time and memory requirements. Dictionaries from other videos may be used to enable scene change detection in online applications. Spatial relationships between words are utilized through the use of the spatial pyramid matching (SPM) technique, which, in a series of levels L, recursively subdivides the image by factors of four into regions, as illustrated in Fig. 4.1, and computes a LLC-based histogram within each region. These histograms are then concatenated together to form a final descriptor. We refer readers to [7] for further details.

4.2.2 Scene Detection System

Finally, we discuss the integration of the MMD and the scene descriptor into a scene detection system, which is summarized in a block diagram, Fig. 4.2.

Aside from hard cuts from one scene to another, other common editing techniques that may confuse scene detectors (causing false negatives) are dissolves, fades, and wipes. Whereas many previous works have dealt with the problems of these special effects by essentially developing an independent detector for each special case [78], our scene descriptor deals with them in a simplified, unified, and elegant manner. Dissolves and wipes to a new scene, or fades to white/black, appear as prominent maxima in the MMD due to the significant change in the bag of words histogram. For accurate localization, we only need to ensure that the window size is longer than the longest expected fade. One benefit of the LLC-based feature descriptor is that dissolves and wipes involving two scenes may be directly modeled as a linear combination of words from the past and future scenes. This should, in theory, give us a smooth MMD peak rather than creating a noisy MMD signal by incorrectly allowing unrelated words to be matched to the image patch. Camera flashes are another kind of video content that can cause false positives, but may be dealt with by

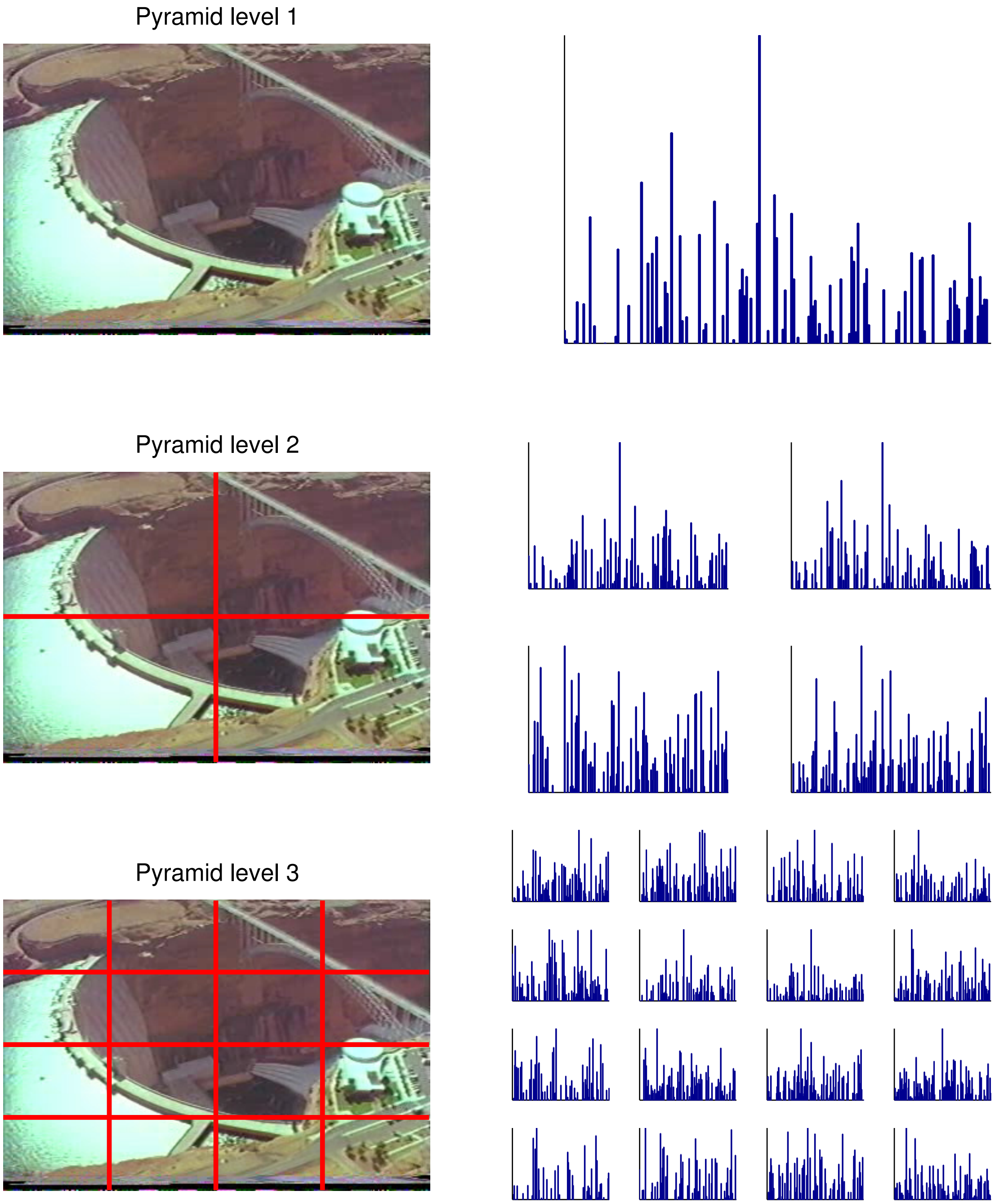


Figure 4.1: An illustration of the spatial pyramid matching technique, showing a three-level pyramid of histograms. Each histogram bin represents one 'visual word'.

ensuring that there are no descriptors immediately on both sides of the large peak that have low mutual dissimilarity.

The histogram intersection kernel (HIK), (4.2), popularized in [7], is used to compare the descriptors of two images X and Y, as it is parameterless and outperformed alternatives such as the RBF kernel in our tests.

$$K_{HIK}(X,Y) = \sum_{j=1}^{d} \min\left(x_j, y_j\right) \tag{4.2}$$

The MMD is then computed as in (Eqn. 3.15) in a sliding window fashion for each frame t, with the $w/2$ frames prior to t forming group P, and the next $w/2$ frames forming group Q. The necessary SIFT keys and scene descriptors may also be computed in small batches which lends itself to online computation and reduced memory requirements. The common kernel computations between windows may be reused; only the kernel distance between each new frame and the $w-1$ other frames in its window need to be computed.

Finally, a standard peak finding approach is used to find local MMD maxima within the window. This falls in the class of adaptive thresholding methods, where a scene change is declared for a frame if it is a local maxima, and was preceded in time by a MMD value lower by at least δ.

4.3 Experimental Setup and Results

We demonstrate our system on a 30 minute documentary video used in the NIST TRECVID 2001 shot boundary detection competition, 'Challenge at Glen Canyon', featuring complex natural outdoor scenes, compression artifacts and noise, and a wealth of object and camera motion. SIFT keys are computed at every 8 pixels in both dimensions, with each SIFT key having a spatial support of 16×16 pixels. A visual dictionary of $D = 768$ visual words was learned by applying k-means clustering over 25% of the video. Using a three-level pyramid ($L = 3$), as recommended in [15], this yields a $768 + 4 \cdot 768 + 4^2 \cdot 768 = 16,128$ dimension feature descriptor for each image.

The top graph of Fig. 4.3 shows sample scene changes and the output MMD values. Further examples of false positives, false negatives, and correctly detected changepoints are shown in Figure 4.4. For video summarization and change detection purposes, we observe that it is unnecessary to locate the shot boundary with frame accuracy; a latency may be specified, which allows us to reduce computational complexity significantly by temporal

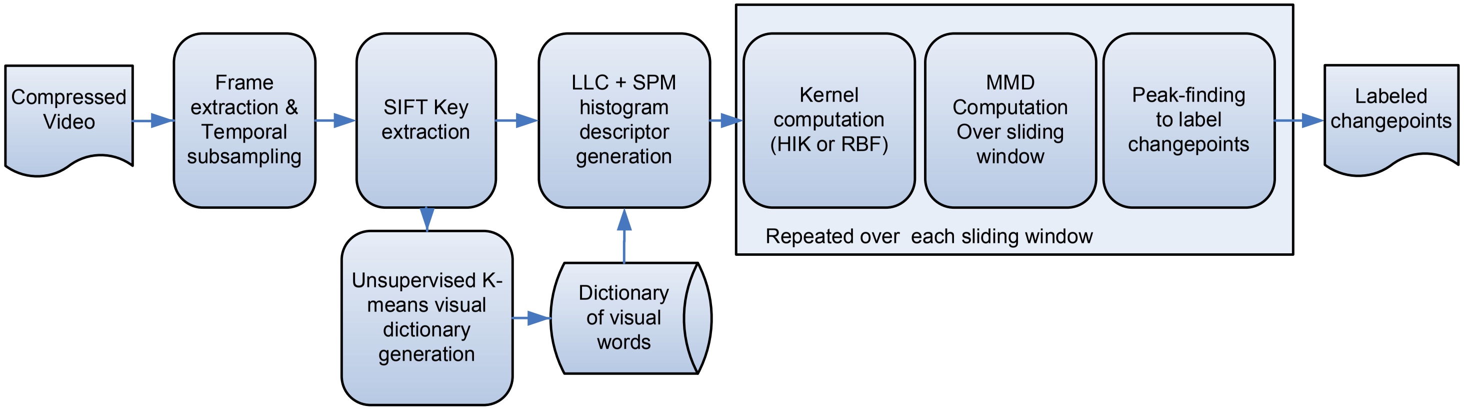

Compressed Video
Frame extraction & Temporal subsampling
SIFT Key extraction
LLC + SPM histogram descriptor generation
Kernel computation (HIK or RBF)
MMD Computation Over sliding window
Peak-finding to label changepoints
Labeled changepoints
Repeated over each sliding window
Unsupervised K-means visual dictionary generation
Dictionary of visual words

Figure 4.2: Visual summary of our visual changepoint detection system.

subsampling. We hypothesize that the similarity operators used in systems based on the motion field, edges, or colour histograms would likely produce spurious false peaks due to the scene discontinuities, in contrast with our system. To investigate this we have decimated the framerate from 29.97 fps to 1 fps.

4.3.1 Tested Alternatives

To put the performance of our proposed solution (entry #1 in Table 4.1, 'LLC + SPM + HIK + MMD') in context, we have also tested several combinations of the features and kernels individually, without MMD on the same data. We have also tested solutions proposed by other groups that do not rely on the bag-of-words approach at all.

In entry #2, we wished to isolate the contribution of MMD from the contribution of the features and kernels. To test this, we searched for minima of the similarity kernel between successive frames, and classified them as scene changes, without using MMD. This variant is labelled 'LLC + SPM + HIK'.

In entry #3, in §4.1 we hypothesized that by computing the difference of group means in a remapped feature space, increased discrimination could be obtained. This variant is denoted 'LLC + SPM (ℓ_2)', where the simple Euclidean distance (denoted by ℓ_2) is used as the distance measure. The same feature descriptor is used as our proposed solution.

Entry #4, 'BOW + HIK + MMD', serves to illustrate the improvement in performance gained from the LLC and SPM techniques, compared to the base bag of words (BOW) approach. It should be compared to 'LLC + SPM + HIK + MMD'.

Finally, the last two entries of Table 4.1 test two schemes representative of contemporary approaches to scene change detection on the same data, in order to illustrate their ability to adopt to different content genres and framerates. Entry #5, 'Local edge-based SAD', is a change-of-edge-intensity approach using fixed thresholds and local subregions compared with the sum of absolute differences (SAD) function, while the rank tracing approach, entry #6, is a shot boundary detector tuned for a particular genre, sports videos [85].

4.3.2 Ground truth data generation

The ground truth changepoints were obtained by manually reviewing the extracted frames of the video, and identifying the first frame of each semantically different scene. Different camera angles and lighting changes were not considered new scenes. It is a subjective process; for example, a slight zoom of a scene would not be considered as a new scene, but a large zoom, say from a full body shot to a close-up of the hand, would be. This annotation was done before the algorithm was run to avoid inadvertent bias. A total of 233 scene changes were identified in this manner from the entire set of 1615 frames.

4.3.3 Results

Results are shown in Table 4.1 using the well known precision and recall metrics, where Precision = TP/(TP + FP) and Recall = TP/(TP + FN). TP, FP, and FN are the number of true positives, false positives and false negatives, respectively. The table entries are ordered by overall harmonic mean of the two, $F1 = 2 \cdot \text{Precision} \cdot \text{Recall}/(\text{Precision} + \text{Recall})$. The first entry is the system as presented in §4.2. We set the window size to $w = 2$ experimentally with the aim of ensuring that we do not have any scenes shorter than this window. We examined the dynamic range R of the MMD over the video sequence and experimentally set the peak finding parameter $\delta = 0.05\,R$, but found the results are not too sensitive to this value.

We can see that entry #2 ('LLC + SPM + HIK') of Table 4.1, 'LLC + SPM + HIK', the variant without MMD, does not perform as well in terms of precision and recall to the MMD solution. The bottom graph of Fig. 4.3 shows its time-series behavior, and we see the primary issue is oversensitivity, leading to many spurious detections and oversegmentation.

The non-kernelized difference-of-means approach, 'LLC + SPM (ℓ_2)', entry #3 of Table 4.1 has 3.6% lower $F1$ than the proposed approach. One reason for its reduced performance can be visually seen in the middle plot of Fig. 4.3 — its dynamic range is much lower than that of the kernelized solutions, which makes it very sensitive to peak finding thresholds.

A 3.8% improvement in $F1$ was obtained by layering the LLC and SPM techniques over the base bag of words (BOW) approach, as seen by comparing entries one and four of Table 4.1. Finally, results for the edge-based SAD approach and the rank tracing approach are listed as entries five and six respectively. Settings were left at their default values for both systems, and we observe that they do not perform well in the domain outside one that they were trained in, which suggests there is a role for our unsupervised, non-parametric robust scene detection scheme to fill.

4.4 Conclusions

We have presented the design and experimental results of a scene detection system that models the concept of 'scene' at a higher level of abstraction than previous works, using

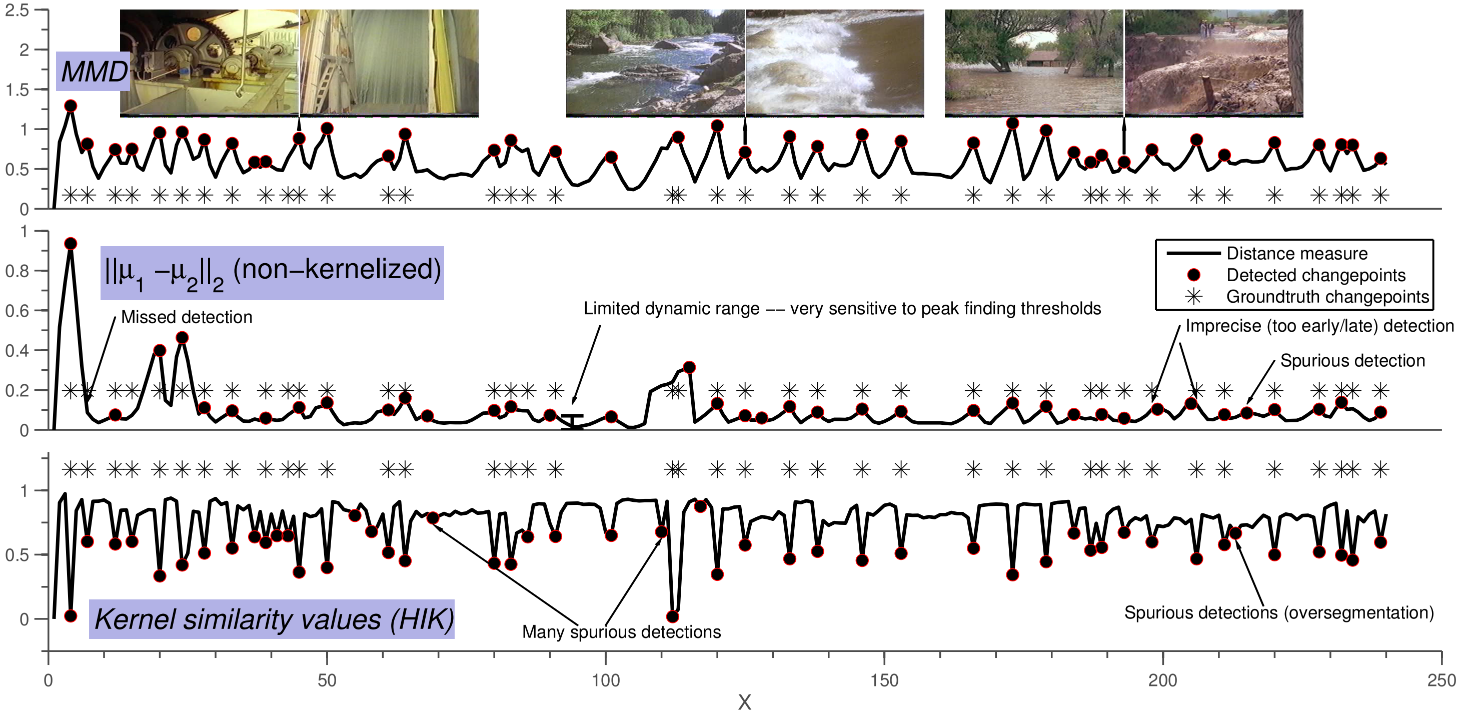
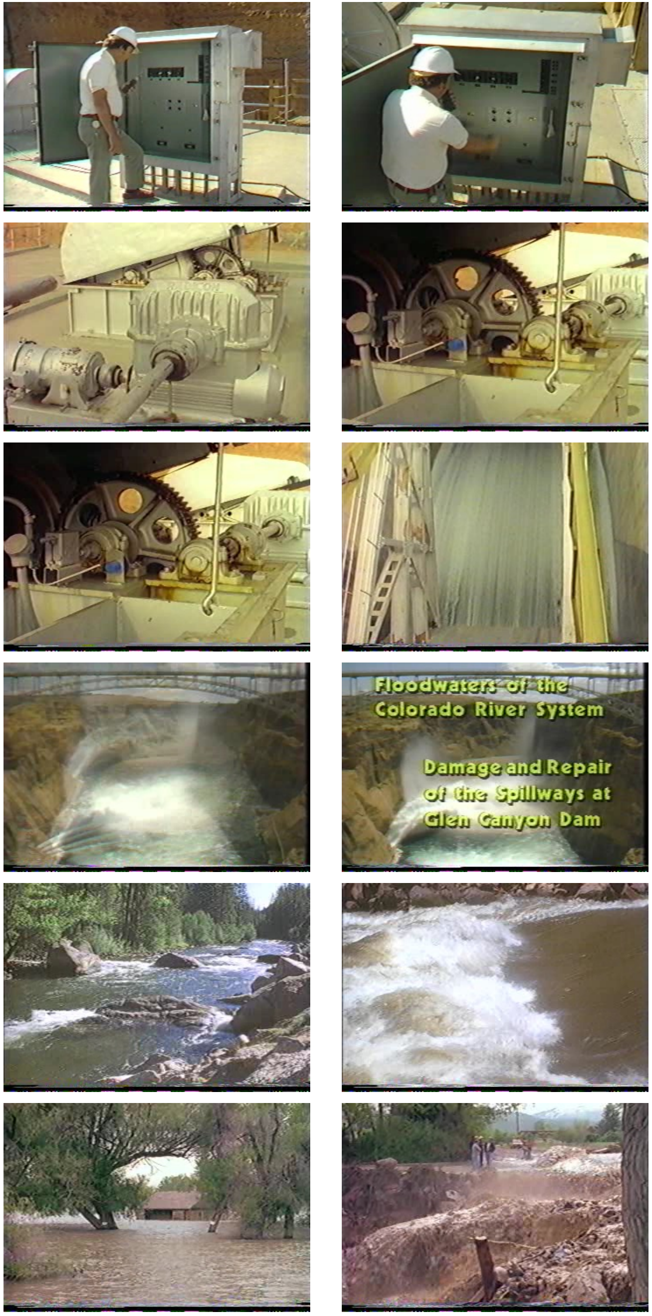


Figure 4.3: Changepoints detected in the first four minutes of 'Challenge at Glen Canyon' film

Figure 4.4: Examples of correctly and incorectly detected visual changepoints using our proposed scheme.


Floodwaters of the Colorado River System
Damage and Repair of the Spillways at Glen Canyon Dam

False positive.

False negative. The right-side frame represented a quick zoom in followed by a zoom out, resulting in a scene that was not longer than the window size w.

Correctly detected changepoint

False positive. Computer-generated overlays not present in the visual dictionary will present issues with our approach.

Correctly detected changepoint.

Correctly detected changepoint. Note the very similar color distributions and high-level scene arrangement.

Table 4.1: Visual Changepoint Detection results on 'Challenge at Glen Canyon' video.

Approach	# Detections	Precision	Recall	$F1$
1) LLC + SPM + HIK + MMD	247	**0.887**	**0.940**	**0.913**
2) LLC + SPM + HIK	247	0.879	0.931	0.904
3) LLC + SPM : $\langle \mu_1 - \mu_2, \mu_1 - \mu_2 \rangle_{\ell_2}$	255	0.839	0.919	0.877
4) BOW + HIK + MMD	245	0.853	0.897	0.875
5) Local edge-based SAD approach, $\lvert \mu_1 - \mu_2 \rvert > \tau$	730	0.315	**0.987**	0.478
6) Rank tracing approach [85]	158	0.544	0.369	0.440

the visual bag of words approach and linear combinations of densely sampled SIFT keys. In doing so, elements of a video that do not represent a true scene change, such as object motion or contrast changes that other feature representations would be sensitive to, are treated as 'noise', reducing the false positive rate. The maximum mean discrepancy kernelized distance was then used with the histogram intersection kernel, with the aim to use nonlinear basis function expansion to increase separability of group means. Results showed higher dynamic range, precision, and recall compared to conventional methods.

Our system works well with varying framerates, enabling a CPU vs detection latency tradeoff, whereas most other scene change detectors are tuned with framerate-dependent thresholds on color or intensity features. We also illustrated how many special types of editing transitions can be handled naturally by the framework instead of requiring special cases and detectors. Three directions for future work are to expand the range of content tested, to use the MMD statistical significance tests in [1] to further filter out false positives, and to utilize the visual changepoint information and temporal segmentation information in the supervised problem of *place recognition* in videos; that is, to enforce temporal consistency in classification results by averaging out individual false positive classifications over a contiguous segment.

The contribution of this work, then, has been to demonstrate the application of the MMD to a time series of visual data, and to 'marry' powerful feature descriptors from the world of computer vision with a problem based in the more traditional image processing world.

Chapter 5

Application 2: Computer Assisted Cancer Treatment Prognosis

Using MMD as a feature, on unordered groups of data

5.1 Introduction

Assessing the efficacy of cancer treatments on subjects in preclinical and clinical applications is presently a very slow affair; results may not be available to the researcher or clinician for weeks or even months. This can lead to ineffective cancer treatments continued needlessly as no faster feedback mechanisms have yet reached broad adoption. Quantitative ultrasound (QUS) methods [86] provide a promising alternative framework that can non-invasively, inexpensively and quickly assess tumor response to cancer treatments using standard ultrasound equipment.

In our work, we take steps towards the development of a computer-aided-prognosis system that uses 2-D QUS *parametric maps*, together with distances between pre- and post-treatment images computed using the maximum mean discrepancy (MMD) distance measure, to estimate the fraction of tumor cells that have died after drug treatment in a population of mice with breast cancer. QUS methods involve extracting features from the spectrum analysis of the backscattered radiofrequency signal samples over a region of interest, forming a QUS parametric map.

Using such an approach, cell death (apoptosis) signs, such as increased backscatter, may be detected on the timescale of hours to a few days, and are strongly linked [87] to gross future anatomical changes (which are detected on the timescale of weeks). This permits clinicians to receive feedback and switch to alternate treatments far earlier, in a step towards the goals of *personalized medicine*. This can save the costs of drugs and hospital care, while reducing patient recovery time and exposure to side effects. Conventional B-mode ultrasound images have difficulty being used for this purpose, because the operator- and machine-dependent settings of B-mode have a higher impact on repeatability and generalization of machine learning methods compared to QUS methods, which use the more machine- and operator-independent [86] calibrated backscattered RF values as their input.

This work has two main contributions. The first, as mentioned earlier, is a methodology to automatically give an early indication of the fraction of cells undergoing apoptosis (with respect to the original tumor cell density), using the QUS parametric maps obtained from coarse human-selected, B-mode ultrasound guided, regions of interest (ROI) windows. The second novel contribution in the medical imaging field is in our use of the Maximum Mean Discrepancy distance metric as a feature for classification in a dissimilarity-based framework. While MMD has also been previously used in medical imaging, such as in diffusion MRI [72], it was used as a biomarker feature selection method, i.e. statistical test, using two empirical PDF vectors as features. Our application differs substantially by using the MMD values themselves as features.

Using intensity histograms of the parametric maps as the feature descriptor, we compare our MMD-based method with three alternate feature representations that use the ℓ_2 distance. These feature representations are found in the QUS literature for detecting treatment response, and consist of a *texton* texture representation ('Texton'), the same intensity histogram used in the proposed MMD approach ('IntHist'), and a representation of each parametric map with its mean intensity ('MeanInt').

Performance is quantified by several measures, chiefly Pearson correlation $|r|$ to the groundtruth cell death fraction, statistical significance tests (t-test) between control and treatment groups, and binary classification accuracy in estimating the cell death fraction of a treated subject.

Using the MMD distance as a feature, statistically significant differences were detected using the unpaired t-test between treated animals and untreated controls within 24 h after treatment administration. Experimental results showed we were able to correctly classify a subject as having 'low' or 'high' cell death with 88.2% classification accuracy, and the MMD distance had a Pearson correlation coefficient to histologically-determined cell death ratios of $|r| = 0.76$.

5.1.1 Past Work

Imaging tumor response to treatment at cellular levels [88, 89] is a much younger sub-field of its vast parent area, medical imaging for cancer, which has been well established for research, clinical screening, and treatment planning purposes; see [90] for a comprehensive survey. The impetus for research in this area is to shift away from the conventional paradigm of using tumour size as a measure of treatment effectiveness. These changes in size can take weeks to months to become visible, and do not always occur even when the treatment is effective [88].

Both biochemical and morphological methods may be used to quantify tumor response. The goal of the biology-based methods typically involves measuring the differences in receptor expression or measuring metabolite levels [88]. Most of these biochemical methods, including magnetic resonance imaging (MRI), positron-emission tomography (PET), X-ray Computed Tomography (CT), have the disadvantage of requiring contrast agents to be administered to enhance the contrast from soft tissues. The agents' cost and potential for side effects and allergic reactions (for example, some of the agents are radioactive, albeit at low levels) limits the spread of the technology. Moreover, the cost of these devices is generally significantly greater than an ultrasound machine.

The morphological methods, which includes quantitative ultrasound (QUS), work on the principle of directly visualizing, or indirectly quantifying, tumor shape and structure at varying scales. Cancer therapies, such as photodynamic therapy, radiotherapy, or chemotherapy, generally function by creating a toxic environment for the tumors, or by inducing apoptosis, i.e. programmed cell death [91]. During this process, many compositional changes occur that will affect the viscoelastic properties of the tumor. These include nuclear condensation, cell swelling, fragmentation of the nucleus, and chromatin dissolution, which are hypothesized to directly or indirectly increase echogenity, i.e. the backscattered energy, of the tumor [86].

Simulation results from [92] demonstrate that a reduction in destructive interference patterns emanating from the cell nuclei, after changes such as nuclear condensation, can account for an increase in backscattered energy.

In turn, apoptosis detected at early stages (as soon as 24 h) after treatment has been linked as a prognostic factor for treatment outcomes measured significantly later (7 to 21 days later) [87]; conversely, the inability of some cancerous cells to initiate apoptosis has been cited as a reason for variability in cancer treatment efficacy [88].

The first work to utilize ultrasound to detect apoptosis resulting from cancer treatment *in vivo* was reported by [93], which used high-frequency ultrasound (40 MHz f_c). This has recently been extended to conventional clinical US ranges (1-20 MHz) [94, 95], which enables much broader adoption of the technology by requiring fewer changes by manufacturers. It built on the much earlier theoretical work of [96], the seminal work behind QUS methods, that utilized spectrum analysis of the conventional low-frequency (5-15 MHz) radiofrequency (RF) signal to characterize tissues at much coarser scales, e.g. to distinguish between normal and detached retinal tissue.

Statistical methods, classifiers, and distances

Many works in the field of QUS analysis [97, 94], rely on 2D/3D feature plots and the t-test to demonstrate the statistical significance and discriminative power of their proposed system. The t-test may be paired or unpaired, depending on whether the comparison is being made between treated vs. non-treated animals (unpaired), or pre- and post-treatment (paired). Factorial analysis of variance (ANOVA) may be applied when multiple treatment options are being compared [98].

However, we note that comparisons of treated vs. non-treated animals assume effectiveness of the treatment in all animals, and therefore may be less helpful in clinical practice. Alternatively, the learning goal may be redefined to predict responding vs non-responding animals (defined, e.g. in [94] as those where a pathologist finds there was a decrease in tumor volume of 50% or more). When such studies obtain groundtruth labels via histopathologic assessment, such as post-treatment average cell death or average nuclei diameter, the Pearson correlation coefficient r may be computed, and appropriate supervised machine learning techniques applied. In the present work, our target is to predict responding vs non-responding subjects, defined as those with a cell death fraction $> 40\%$.

Relatively few works have attempted to use supervised learning to detect tumor response to treatment using medical imaging data, a logical next step towards developing decision support systems for clinicians. Larkin et al. [99] worked with contrast MRI to detect cell death, using the Support Vector Machine (SVM) classifier [17] and a 'Minkowski functional' feature to classify whether an image represented a treated tumor, or a non-treated control image. Classification accuracy after a 24 h period was 75% using 19 subjects. Histological analysis was used to confirm significant increases in cell death after treatment. A SVM classifier was also applied with the same goal of differentiating between pre- and post- treatment images of our dataset, using the QUS intercept data in [100]; classification accuracy was 87.3% when assessed 24 h after treatment.

Two works from a related field, computer-aided pathology detection, may also suggest appropriate supervised learning methods. Sørensen [101] used a fusion of rotation-invariant local binary pattern (LBP) features and intensity histograms in a dissimilarity-based classification approach to predict emphysema in CT imagery. The simple K-NN classifier and Euclidean distance were used, and the classifier's posterior probabilities were further used to compute the correlation to the groundtruth. Feleppa *et al.* [95] used the multi-layer perceptron (MLP, a type of artificial neural network) to distinguish between cancerous and non-cancerous tissues of the prostate at a pixel-by-pixel level using QUS midband and intercept features. They found SVMs to have similar accuracy to the MLP.

We note that fairly little comparative analysis appears to have been reported on the possible feature representations/transformations, similarity measures, and supervised classifiers that are essential components of computer-aided prognosis systems using QUS technology, which is on the road to commercialization and clinical use. This work aims to take a step forward in this direction.

5.2 Methods

5.2.1 Data collection and preparation

Experiments carried out at the Odette Cancer Centre, part of the Sunnybrook Medical Center (Toronto, Canada), on 17 severe combined immunodeficiency (SCID) mice formed the dataset for this study. The hind legs of the mice were injected with human breast cancer cells (cell line MDA-MB-231), where they grew into 7–9 mm sized xenograft tumours. All except two mice were then anaesthetized and given chemotherapy treatment (paclitaxel-doxorubicin) by way of intravenous tail vein injection. The last two mice served as a control group (labeled '0 h'), receiving a sham treatment.

Ultrasound imagery was taken of the tumour before treatment ('pre-treatment imaging'). The mice were divided into five groups, where each group was also imaged after a different, progressively increasing time interval: 0 (for the control group), 4, 12, 24, and 48 h after treatment.

Ground truth data generation

To obtain a ground truth for assessing the effectiveness of our methods, histological studies were conducted by the pathology department by excising the tumour after post-treatment imaging. The cells were sliced and stained, and then imaged at 20× and 40× magnification. Morphology-aware software (e.g. [102]) was used to count the cell area that had undergone *apoptosis*, or programmed cell death, by way of quantifying the image area that had responded to the staining agent. We refer to his groundtruth value as the 'cell death fraction'. For additional details on the medical aspects of the dataset preparation, we refer readers to [98, 100].

5.2.2 Quantitative ultrasound

A Sonix RP research ultrasound system (Ultrasonix, Vancouver, Canada), with center frequency of ˜7 MHz, focal depth 1.5 cm, and 40 MHz sampling rate was used with a L14-5/38 linear transducer. Multiple scan planes (11 to 15), centered on the tumour in a *region of interest* (ROI), were taken at progressively increasing depths, ˜0.5 mm apart.

Standard techniques of spectral estimation [103] were used to derive an estimate of the frequency spectrum. The transducer, ultrasound machine, and depth-related attenuation all introduced their own transfer functions, which were compensated for by normalizing with the power spectrum of an agar-embedded glass bead phantom model [104] (calibration target). Reported spectral values are therefore relative (dBr) to this target.

Primary Features

The Fourier transform of a sliding window of calibrated, backscattered ultrasound RF signals, 9–11 wavelengths in the axial direction, and 15 scan lines laterally, was first taken using the Hamming window function to obtain a power spectrum estimate at each point in the ROI.

Linear regression performed on this power spectrum, using a 6 dB bandwidth centered on the transducer's center frequency, yields three parameters, which we shall refer to as *primary features*: (a) the intercept of the fit line to the calibrated y-axis, termed the *intercept*; (b) the slope of the fit line, termed the *spectral slope*, and (c), the *midband fit (MBF)*, the power (in dBr) at the center frequency f_c. These parameters are extracted for each movement of the sliding window, thus forming three 2-D *parametric maps*.

It can be seen that the midband fit is not independent of the slope. The intercept is theoretically independent of attenuation, as it is a relative measurement with respect to a calibrated plate at the same depth. Very early work with QUS [93] used raw backscatter (RF) values directly instead of applying any linear regression; [96] used a simple model of ocular membrane scattering to derive the membrane's reflection coefficients directly from the power spectral density (PSD). Parametric maps of the MBF feature, pre- and post-treatment, are shown in Figure 5.1 over several time intervals.

The dataset was compiled and the features (2-D parametric maps of intercept, midband fit, and slope) extracted by our collaborators at Sunnybrook, while the data analysis has been my individual work.

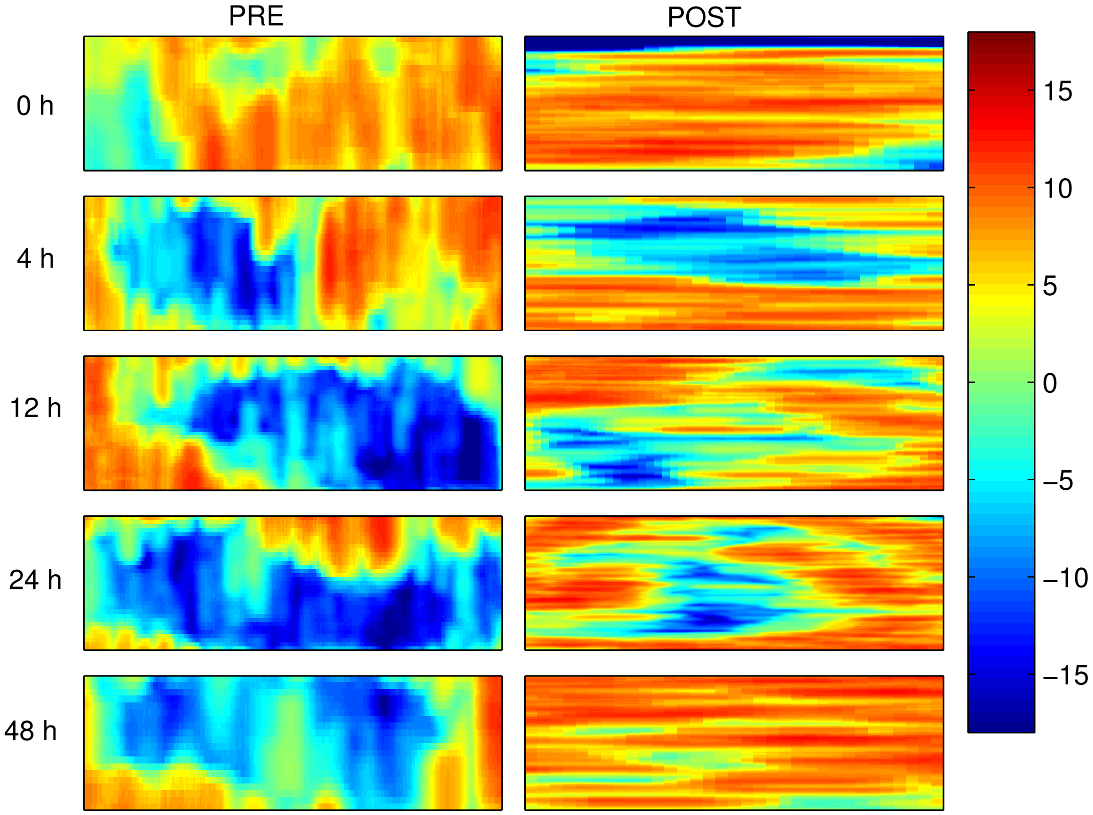


Figure 5.1: (Best viewed in colour.) Representative pre- and post-treatment QUS parametric maps of the midband feature for each exposure time group. The colour bar is in units of dBr, relative dB to the normalized calibration target.

The effectiveness of each primary feature appears to vary depending on many factors: the ultrasound frequencies used, the pathology type, the effectiveness of the applied therapy, and the test environment (*in vivo*/*in vitro*). Further discussion may be found in [98]. In working with high frequency (20-60 MHz) ultrasound, [98] used MBF and intercept; [105] worked with integrated backscatter spectrum values and SS, while [97] selected SS and MBF. Intercept alone was found most discriminative in lymph classification tests [95]. Using conventional ultrasound, [95] chose MBF and intercept values as did [94], while [100], using the same dataset as the present work, used intercept alone.

One consistent pattern was observed from this review of the QUS literature: the spectral slope was found to be discriminative in certain cases at high frequencies, but never in low frequency clinical ultrasound, a result which we also confirm in our work. We know that the echogenicity of backscattering surfaces varies with frequency, as the wavelengths grow much longer than the scale of the scattering structures. There may have been no lower frequency spectral signature (envelope) changes in the structures modified by the treatment. Another simple theory is that small changes in slope are more pronounced at frequencies further away from the y-axis.

Feature transformations

Several feature transformations have also been introduced that attempt to reduce the region of interest (ROI) dimensionality or represent it in a more discriminative fashion.

Using the average scatterer diameter (ASD) and average acoustic concentration (AAC), which are two parameters of a spherical backscattering model, [106] was able to differentiate between breast carcinomas and sarcomas in mice. These features are both derived quantities based on the SS and intercept features. ASD and AAC were evaluated along with intercept and slope features in [95] for detecting tumorous lymph tissue, but this did not result in higher classification accuracy than using the spectral intercept alone.

More recently, several authors have proposed treating parametric maps of these backscatter-derived features as images, and to apply traditional methods of texture analysis on them. These include grey-level co-occurrence matrices (GLCMs) of the MBF and intercept maps [94], from which several image moments were computed: correlation, contrast, and homogenity. Comparison of breast tumors after one week of treatment showed statistically significant differences between pre- and post-treatment populations using these features. A trained dictionary of texture atoms (textons) was applied by [100] to distinguish between pre- and post-treatment images, using the same breast cancer dataset as our work; each ROI is described by a histogram of these textons. We implemented these two methods for comparisons to our proposed method of representing each ROI with an intensity histogram.

5.2.3 Maximum mean discrepancy

The inspiration for this study arose from our previous work using maximum mean discrepancy in an unsupervised scenario, for time-series changepoint detection [4]. We realized that the problem of detecting semantic scene changes in videos had parallels to detecting significant changes in tumour properties from a series of ultrasound parametric images. The analogy is not exact: in the ultrasound case, adjacent images in a population are taken from discontinuous spatial vantage points (ROIs), and the pre- and post-treatment populations are temporally separated by a large amount of time.

We expect that such a measure will be useful in exploiting the intra-group variance information available from multiple samples/instances taken of each of the pre- and post-treatment populations.

5.2.4 Computer aided prognosis

After the 2-D parametric maps have been prepared for each ROI, the intercept and MBF values for each region of interest image are separately gathered into normalized (to unit area) histograms of intensity values. These uniformly spaced histograms have D bins, and constitute a rotation- and scale-invariant nonparametric density estimate (§3.3) of the feature. All of the $(m_1 + m_2)^2$ pairwise similarities between the m_1 ROIs in the pre-treatment group, and the m_2 ROIs in the post-treatment group are then computed using a similarity kernel K, generating the *joint distance matrix* K_{joint}. The *self-distance matrices* K_{m_1} and K_{m_2} are similarly generated using a kernel matrix computed within each group. We have chosen the histogram intersection kernel (HIK), a parameter-free kernel that has been used to good effect in many image analysis applications (§2.2.4).

$$k_{HIK}(X, Y) = \sum_{j=1}^{d} \min\left(x_j, y_j\right) \tag{5.1}$$

We next compute the maximum mean discrepancy (MMD^2) distance using the kernel matrices, using Eqn. 3.15.

$$M\hat{M}D_b^2(X, Y) = \frac{1}{m_1^2} \sum_{i,j=1}^{m_1} K_{m_1}(\mathbf{x_i}, \mathbf{x_j}) - \frac{2}{(m_1 + m_2)^2} \sum_{i,j=1}^{m_1+m_2} K_{joint}(\mathbf{x_i}, \mathbf{y_j}) + \frac{1}{m_2^2} \sum_{i,j=1}^{m_2} K_{m_2}(\mathbf{y_i}, \mathbf{y_j}) \tag{5.2}$$

The last stage is to train a k-NN classifier using the MMD values as features, and the groundtruth values as labels in a dissimilarity-based classification scheme (§3.1, [68]).

5.3 Experimental Setup and Results

The mice were divided into five exposure time groups, as shown in Table 5.2, and multiple ultrasound images were taken just prior to treatment with chemotherapy (the 'pre' population), and after exposure (the 'post' population) of the specified duration. This represented a total of 443 QUS ROIs with three parametric primary feature maps each – one for each of the intercept, midband fit point, and the slope. The size of the parametric maps varied depending on the operator-selected ROIs, but were on the order of 250 × 30 'pixels', or feature values.

We selected several main evaluation criteria : t-test, Pearson correlation to cell death, and ability to predict cell death, and report on each one in a separate subsection. The naming scheme used throughout our figures, charts and text is

[QUS Feature]-[Representation]-[Distance Measure]-[Kernel (if used)]
e.g. *Intercept-IntHist-MMD-HIK*.

To set the parameters of the system, a grid search was performed on histograms of sizes {2-200}, and the histogram size yielding lowest K-NN classification error was selected. This was 9 bins for MBF, and 10 bins for the intercept feature.

The slope primary feature performed poorly across all feature representations and evaluation metrics tested, and so its results have been omitted. Development took place on a contemporary Windows Core i5-2520M machine with 4 GB of RAM, using MATLAB.

5.3.1 Alternative solutions tested

We compare our proposed approach with three alternative systems published in the recent literature on detecting tumor response changes using QUS.

Gangeh *et al.* [100] proposed treating the parametric images of QUS primary features (e.g. Figure 5.1) as textures that can be analyzed using the bag-of-textons approach, which is one of the state-of-the-art texture representation methods. Working with each parametric map in turn, we implemented this approach, extracting 500 randomly chosen textons of size 5 × 5 from each image, and computed the dictionary used to form the bag-of-textons by using $k-$means clustering separately over the set of each subject's pre-treatment and post-subject images. The per-subject codebooks are then concatenated together. Thus, if there are N subjects and k atoms per set, the final dictionary is of size $2Nk$. We set $k = 10$, based on a classification-error-minimizing grid search of k values {5-25} and patch

Table 5.1: Pearson correlation coefficients $|r|$ for the linear least-squares fit of computed distances to cell death

		Midband	Intercept
MMD	Intensity Hist.	**0.76**	0.67
	Texton Hist.	0.75	0.64
L2	Intensity Hist., $\|\|\mu_{post} - \mu_{pre}\|\|_{\ell 2}$	0.68	0.57
	Mean Intensity, $\mu_{post} - \mu_{pre}$	0.65	0.65

sizes from $\{1 \times 1\}$ to $\{12 \times 12\}$. The bag-of-textons histogram descriptor is then formed for each ROI using the final codebook. We refer readers to [100] for a detailed description.

The second comparison uses the same feature representation as our MMD approach, intensity histograms ('IntHist'), but computes the distance using the Euclidean distance instead of MMD.

The third and final comparison is made by representing each ROI using the mean value ('MeanInt') (a scalar value, $\mu_{post1}, \mu_{post2}, \cdots$) of its parametric feature map, and computing the group means μ_{pre}, μ_{post} separately for each exposure time. This approach is implicitly used in the majority of QUS works that compute 2D feature plots of primary features, e.g. [105], or report the average pre- and post-treatment differences of a primary feature, e.g. [97]. As the polarity of the difference is informative, we abuse the notation slightly and do not take the absolute difference of means (as we see in Fig. 5.3, this affects the 0 h case only; the other exposure groups are still computed using the proper ℓ_2 distance).

5.3.2 Correlation to cell death

Table 5.1 lists the Pearson correlation coefficients r of the different distances under comparison after a linear regression fit to the groundtruth cell death fraction. While a low absolute value of r may simply indicate a nonlinear relationship of the imagery to cell death, relative differences in r between data analysis methods, for the same underlying data and treatment, can give us meaningful performance data on the features and distances. Here, a 16.9% improvement over the traditional method of comparing mean intensities ($r = 0.65$) is observed compared to the highest-performing combination, Midband-IntHist-MMD-HIK ($r = 0.76$). The linear fit to the MMD values is plotted in Figure 5.2; a nonlinear cell death to exposure time relation can be seen.

Figure 5.3 compares the average cell death, Midband-IntHist-MMD-HIK distances,

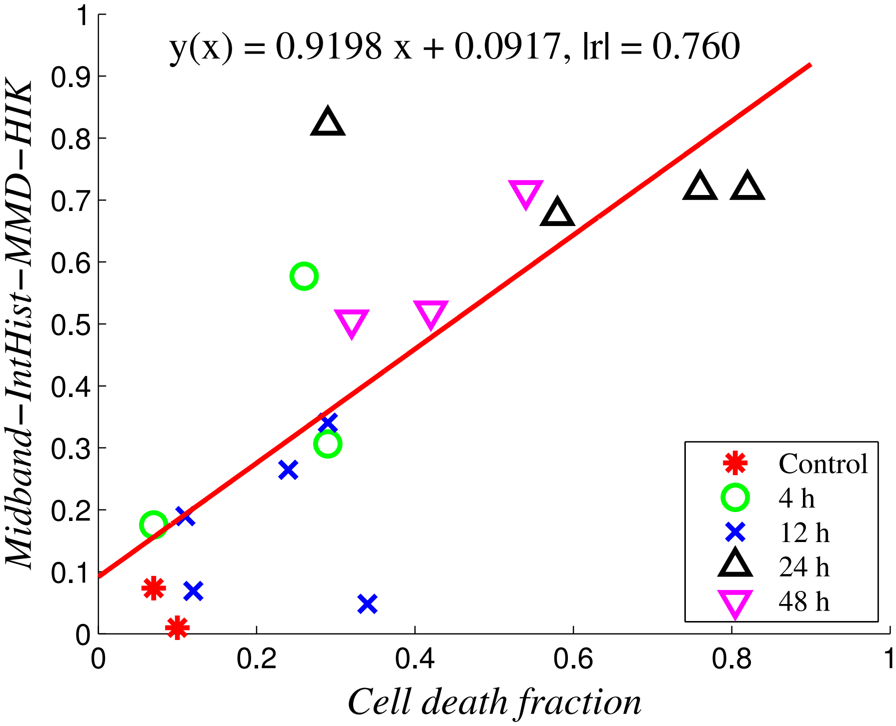


Figure 5.2: Linear fit of MMD vs cell death fraction to the midband fit (MBF) feature data, using the HIK kernel

and Midband-MeanInt-L2 distances by exposure time group. Due to the different vertical scales, it is difficult to compare MMD and ℓ_2 directly, although we can see that the pseudo-ℓ_2 method (the discussion in 5.3.1 explains why the value is negative) actually reported a negative value for the control group, which correlates poorly to the near 10% cell death experienced. To quantitatively compare the histograms in Figure 5.3, we again turn to a kernel designed for histograms, the HIK (Eqn. 2.37). The MMD and L2 distances are normalized to have the same value as the cell death fraction at 24 h. Comparisons against the groundtruth histogram yielded a similarity score of 1.433 with Midband-IntHist-MMD-HIK, and 1.303 for Midband-MeanInt-L2 (higher scores indicate higher similarity), a 10% higher similarity value for MMD.

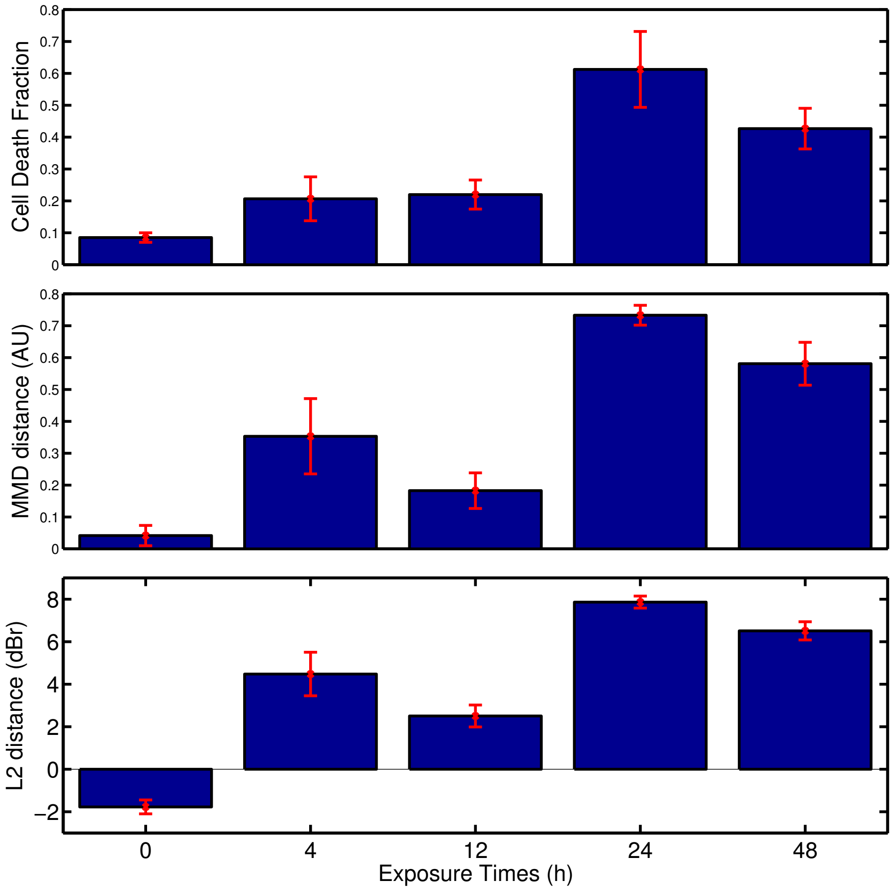


Figure 5.3: Cell death, MMD (arbitrary units), and difference of means (relative dB, dBr) vs Exposure Time. The error bars show the ± standard error of the mean (SEM).

5.3.3 Statistically significant differences from control group

To compare the statistical significance of the different computed distances, we computed Welch's unpaired two-sample t-test with unequal variance, using a significance level of $\alpha = 0.05$. The unequal variance t-test is used as the sample population varies significantly between groups. The first group consisted of all of the control subjects' distances, while the second group contained an exposure group's distances. Results are reported in Table 5.2 for the different feature-distance combinations, and are roughly arranged in order of ascending p values.

All of the feature-distance combinations have $p < 0.05$ (the standard threshold of significance) after 24 h of treatment, and $p < 0.01$ when using the MBF. The results are most interesting for the (Intercept/Midband)-MeanInt-L2 combinations which stand almost alone in reporting significant differences for the 4 h and 12 h groups (the exception is Midband-Texton-MMD-HIK for the 12 h group). As a t-test between control group and non-control group groundtruth cell death fractions (Table 5.2) indicates that the control and 4 h group do not possess statistically significant differences, this suggests that the mean intensity feature and ℓ_2 distance are overly sensitive and are reporting a false positive, in contrast to MMD.

As a practical matter, we note that a clinician is unlikely to change treatment protocols after such a short period, when it is known that the peak response is typically seen 24 h after treatment administration (Figure 5.3). Previous QUS works have also not made claims about changes being detected so quickly after treatment, with the earliest cited detection time being 24 h to our knowledge [105, 98]; [97] imaged after intervals as short as 1 h but did not report the p values per exposure group.

5.3.4 Predicting cell death over a threshold

Next, we train a supervised binary classifier, using the distance value between pre- and post-treatment populations as the sole feature, and cell death fractions as groundtruth. The target is to predict whether or not a subject will have cell death greater than a threshold τ. Ten-fold leave-one-case-out (LOCO) cross-validation was used with the K-nearest neighbours (K-NN) classifier to successively test the instances, as it is the most common choice in dissimilarity representation [68] and able to represent complex, multimodal classification surfaces. We performed classification using two cell death thresholds, 20% and 40%. Based on the distribution of cell deaths in the population (Figure 5.2), we observed noticeable gaps around the 20% and 40% levels, and we have therefore hypothesized that these levels are less susceptible to misclassifications caused by noise in the feature values. The classification accuracy, area under curve (AUC) of the Receiver-Operator Curve, and Type I (1 - Sensitivity) and Type II (1 - Specificity) error rates are summarized in Tables 5.3 and 5.4 for the 20% and 40% thresholds, respectively. Entries are sorted in order of ascending test error. Class priors were set to their observed frequencies, and class-weighted classification error results are reported.

Generally, the MBF feature proved more discriminative in terms of classification error compared to the intercept. MMD with the histogram of midband values had the lowest error at both threshold levels.

Table 5.2: Test of statistical significance (unpaired, two-sample, unequal variance t-test). p-values shown for each exposure group vs. the control group. * denotes $p < 0.05$; ** denotes $p < 0.01$.

Feature	Group	Control (0 h)	4 h	12 h	24 h	48 h
	Number of subjects m	2	3	5	4	3
	Percentage of subjects with > 20% cell death	0%	60%	66%	100%	100%
	Avg. cell death (%)	8.5 ± 2.1	20.7 ± 11.9	22.0 ± 0.10	61.3 ± 23.8	42.7 ± 11.0
	Groundtruth p-values against control		0.2158	0.0405*	0.0204*	0.0278*
Intercept	Intercept-MeanInt-L2, $\mu_{post} - \mu_{pre}$		0.0265*	0.0446*	0.0011**	0.0032**
	Intercept-IntHist-MMD-HIK		0.2184	0.1419	0.0015**	0.0271*
	Intercept-Texton-MMD-HIK		0.2017	0.0659	0.0015**	0.0403*
	Intercept-IntHist-L2, $\|\|\mu_{post} - \mu_{pre}\|\|$		0.2539	0.1649	0.0163*	0.0503**
Midband	Midband-MeanInt-L2, $\mu_{post} - \mu_{pre}$		0.0186*	0.0011**	0.0005**	0.0006**
	Midband-Texton-MMD-HIK		0.1077	0.0172*	0.0002**	0.0062**
	Midband-IntHist-MMD-HIK		0.1109	0.0812	0.0007**	0.0074**
	Midband-IntHist-L2, $\|\|\mu_{post} - \mu_{pre}\|\|$		0.1264	0.1210	0.0031**	0.0033**

Table 5.3: Classification error percentages and ±1 σ, predicting whether cell death is greater than $\tau = 20\%$. The K-nearest-neighbour (K-NN) classifier k parameter is optimized by leave-one-out cross-validation over each fold). Ten-fold cross-validation is used; results shown are averaged over ten runs. Area Under (the Receiver-Operator) Curve (AUC) and Type I/II error are shown.

Feature	***k*-NN (20%) error**	**ROC AUC (%)**	**Type I error (%)**	**Type II error (%)**
Midband-IntHist-MMD-HIK	**0.1529 ± 0.030**	**0.877**	0.14	0.18
Intercept-IntHist-MMD-HIK	0.1765 ± 0.000	0.812	0.25	**0.00**
Midband-MeanInt-L2	0.1824 ± 0.033	0.730	0.17	0.22
Midband-Texton-MMD-HIK	0.2176 ± 0.028	0.831	0.16	0.36
Intercept-Texton-MMD-HIK	0.2235 ± 0.037	0.749	0.31	0.02
Midband-IntHist-L2	0.2588 ± 0.030	0.755	0.24	0.30
Intercept-MeanInt-L2	0.2647 ± 0.042	0.825	0.23	0.36
Intercept-IntHist-L2	0.2941 ± 0.039	0.805	0.25	0.40

Table 5.4: Classification error percentages and ±1 σ, predicting whether cell death is greater than $\tau = 40\%$. The k-nearest-neighbour (k-NN) classifier k parameter is optimized by leave-one-out cross-validation over each fold). Ten-fold cross-validation is used; results shown are averaged over ten runs. Area Under (the Receiver-Operator) Curve (AUC) and Type I/II error are shown.

Feature	***k*-NN (40%) error**	**ROC AUC (%)**	**Type I error (%)**	**Type II error (%)**
Midband-IntHist-MMD-HIK	**0.1176** ± 0.000	0.770	0.20	**0.08**
Midband-IntHist-L2	0.2294 ± 0.033	0.858	**0.18**	0.25
Intercept-IntHist-L2	0.2294 ± 0.019	0.807	0.20	0.24
Midband-MeanInt-L2	0.2353 ± 0.028	0.663	0.24	0.23
Midband-Texton-MMD-HIK	0.2882 ± 0.019	**0.882**	0.40	0.24
Intercept-IntHist-MMD-HIK	0.3294 ± 0.050	0.757	0.52	0.25
Intercept-Texton-MMD-HIK	0.3765 ± 0.057	0.741	0.68	0.25
Intercept-MeanInt-L2	0.4294 ± 0.056	0.684	0.66	0.33

5.3.5 Discussion

Overall, the Midband-IntHist-MMD-HIK combination had the strongest performance amongst the different feature-distance combinations on the evaluated metrics and thresholds. Figure 5.4 visually compares the MMD versus the Euclidean distances in dissimilarity space. With MMD, we can observe that low cell death subjects are clustered more tightly into a corner compared to ℓ_2, and the inter-class distance is increased. This reduces the likelihood of k-NN errors as well as the errors of classifiers employing linear discriminant functions. We had also tested combinations of features in early feature fusion configurations, as well as late classifier-level fusion (using Intercept-IntHist-MMD together with Midband-IntHist-MMD, for example), but found they did not improve classification error over using individual features. Using this plot, we can see why feature fusion may not be more effective – a diagonal discriminant line is not able to cleanly separate the classes more than a straight line perpendicular to one of the axes.

The intensity information embodied implicitly in the intercept, and more explicitly in the midband fit parameter is important to ensure good discrimination, as shown by the poor performance of the slope, which is insensitive to absolute backscattered energy levels.

The intensity histograms outperformed the texture based methods (textons), which raises the question of whether the QUS parametric maps are best treated as texture images, and under which conditions. Texture methods look for familiar, regular patterns of physical structure, but in this case, the underlying scattering structures (of nuclei fragmenting, cell walls disintegrating, etc.) are occurring at much smaller scales than the ultrasound wavelength, which will cause a speckle pattern to appear superimposed on the B-mode image. Under this hypothesis, we would expect texture representation to show better relative performance when tested with higher-frequency ultrasound, which has higher spatial resolution, or for coarser, anatomical-level tasks such as emphysema detection. Conversely, we would also expect other texture based approaches, such as the Local Binary Pattern feature as in [101] to fare poorly on the present dataset.

We found the selection of the kernel has strong impact on kernel-based learning methods. Experiments (results omitted) were conducted using the Gaussian radial basis function (RBF) kernel (using heuristics for the tuning of its σ parameter), the Hellinger kernel (related to the Bhattacharyya distance), χ^2 kernel, and correlation kernel in conjunction with MMD and intensity histograms. No alternative kernel tested consistently outperformed HIK on different primary features and both cell death thresholds.

At the 40% threshold, which we believe to be more clinically relevant, sensitivity was 80% and specificity 91.7%. As one approximate point of reference, we note that [107] reported tumor response detection sensitivity and specificity of 78% and 86%, respectively, using spectrum analysis of proton magnetic resonance spectroscopy (MRS) data taken one *week* after at least three rounds of chemotherapy. That group also studied human breast cancer tissue treated *in vivo* with chemotherapy, but with the very significant difference of using human patients instead of mice.

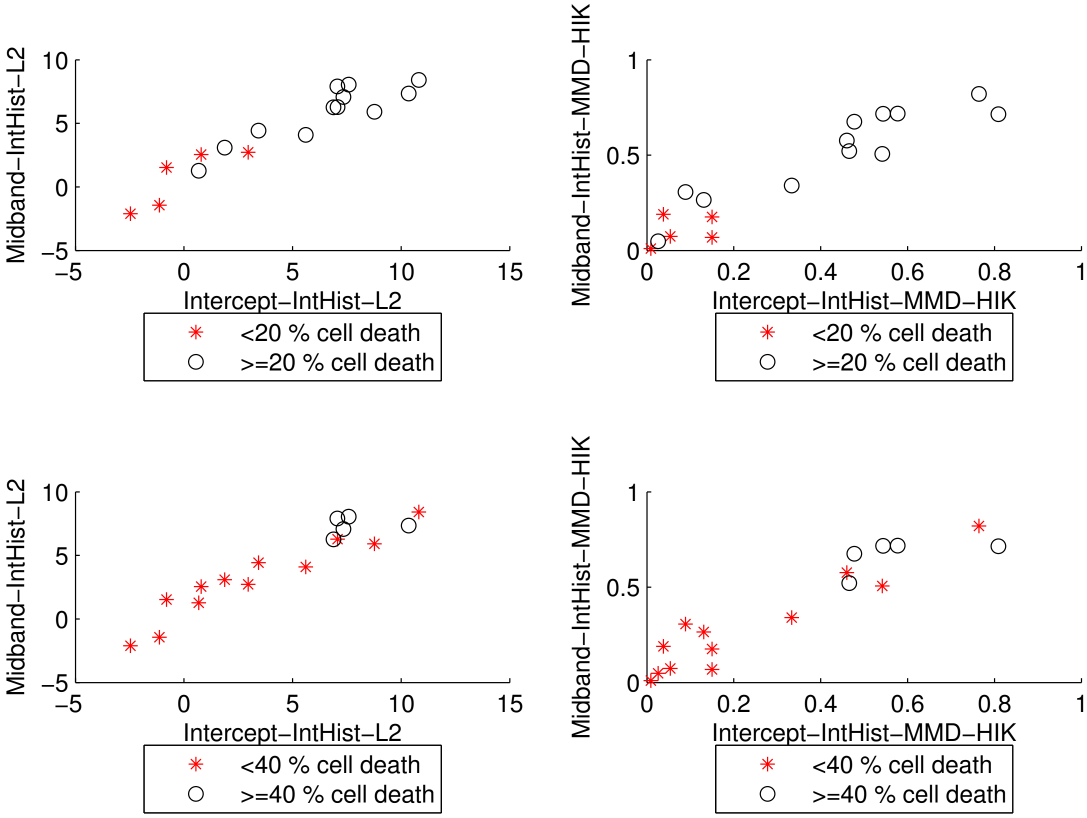


Figure 5.4: Midband distances plotted against intercept distances. Difference of means on left, MMD on right.

5.4 Conclusions

We presented a system for noninvasive tumor response assessment of a system using QUS parametric maps, containing several aspects novel to the analysis of QUS imagery: a dissimilarity-based classification scheme employing the MMD distance measure as features, the introduction of the MMD value as a proxy for cell death fraction, and the use of intensity histograms of primary features. Three alternative, commonly used feature representation and distance schemes were implemented for comparison purposes. While all showed statistically significant differences between pre- and post-treatment groups, significant improvements in both correlation to histologically determined cell death ratios and classification accuracy were observed using MMD and intensity histograms, for both the MBF and intercept features. The system has classification accuracy of 84.7% when predicting cell death <20%, and 88.2% for cell death <40%. Our proposed approach has just one parameter to set, the number of intensity histogram bins, which is optimized automatically during the learning phase.

The techniques utilized in this work can be applied to other treatments and pathologies, not just for tumor response, but for the broader problem of pathology detection or treatment response monitoring using medical imaging. Our work may contribute to one possible path forward for a fast, noninvasive and inexpensive computer aided diagnosis system, which can fuse together other metadata and predictors about the patient, such as age, gender, and family history to assist clinicians. The MMD can be used in an additional context in such a setting, that is, to identify the additional features and metadata that will be statistically discriminative between populations.

Chapter 6

Conclusions

6.1 Summary of accomplishments

Kernel methods based on nonlinearly mapping data to a Reproducing Kernel Hilbert Space have made significant impacts in many areas of machine learning, thanks to their computational efficiency, increase in group discrimination, and a wide range of kernels that may be selected according to the task and data types.

Maximum mean discrepancy is a distance measure between probability distributions, based on the idea of using the squared sum of probability density differences as a metric. It utilizes kernels to compute the unnormalized Fisher ratio, or squared difference of means, taken in a Reproducing Kernel Hilbert Space. MMD has three desirable dissimilarity properties identified through our literature review: it is a metric; it has relatively low computational complexity, and it is able to exhibit selective invariance properties via the designer's selection of kernels. For example, the histogram intersection kernel ignores dimensions, with no matching values, making it less prone to noise in high-dimensional situations.

This thesis has concerned itself with whether MMD-based discrimination of two sets of image objects may outperform its non-kernel equivalent, the unnormalized Fisher ratio, computed in the original feature space using the ℓ_2 distance, $d_{\ell_2}(X,Y) = \sqrt{(\mu_{\mathbf{x}} - \mu_{\mathbf{y}})^2}$. We investigated this hypothesis by applying MMD on two different problems in the area of image analysis. After investigation of past approaches and experimental testing, we presented system designs of feature descriptors, data transformations and kernels that yield promising results in two quite different types of data – time series of multivariate data, and unordered grouped data, where the ordering of instances is not information bearing.

The first problem, video scene change detection, involved a time series of ordered objects, where we do not know the group membership of an object. We computed SIFT keys for each image on a dense grid based on a greyscale input, and built a very high dimensional bag-of-visual-words histogram by matching these keys to the closest *visual atom*, obtained through unsupervised clustering. We then further enhanced precision and recall by adding the Spatial Pyramid Matching technique to capture larger-scale spatial dependencies, and the Locally-constrained Linear Coding technique to find linear combinations of each visual atom to represent each SIFT key. The MMD is then computed over the frames of the video sequence in an overlapping sliding window fashion, successively forming 'current' and 'next' groups of frames. A standard peak finding routine is used on the MMD sequence to find local maxima, which are interpreted as scene change points.

Experimental results showed that MMD had improved precision and recall, compared to the difference of means computed with ℓ_2 (which we have often abbreviated to 'ℓ_2'), or using the raw kernel values alone. Several reasons for this were seen from a visual inspection of the time-domain behavior of these three measures. The ℓ_2 distance, operating in the original feature space, had very low dynamic range, making peak-finding noise prone. The HIK value, on the other hand, had high dynamic range, but many spurious peaks, leading to oversegmentation.

Our second application is in the biomedical engineering field. We test the ability of MMD to discriminate between two groups of 2-D parameter maps, obtained through spectrum analysis of backscattered ultrasound radiofrequency data. One group represents ultrasound scans of human breast cancer tissue xenografted into a mouse, taken before chemotherapy treatment, while the other group represents scans for the same mouse, taken at a fixed time after treatment. Current research practice involves computing the difference, in the original feature space, of the sample means to test for statistical significance between pre- and post-treatment images. We instead represented each parameter map as an intensity histogram, and computed the MMD distance on these unordered objects with the HIK, where the size and membership of each group was known *a priori*.

Experimental results showed that the MMD distance had notably higher Pearson correlation to the groundtruth cell death ratios, compared to ℓ_2. In addition, we had proposed dissimilarity-based classification using MMD as a feature, which is one of the main machine learning contributions of this work. Tests with the k-NN classifier showed substantially lower classification error using this method. Quantitative ultrasound methods, to which this work contributes, has significance for the early treatment assessment, or prognosis, of cancer patients as soon as 24 h after treatment, instead of the present standard of weeks and months.

Significance

Our experimental results confirm our hypothesis that MMD, paired with task-appropriate feature representations and kernels, can provide an improvement in statistical pattern recognition applications compared to the difference of means computed using the ℓ_2 distance.

We proposed two ways that MMD can be applied to areas beyond its original remit as a two-sample test for unordered, grouped data: as a feature in a dissimilarity-based classification system, and as a distance measure for changepoint detection in ordered data with unknown group membership. This provided additional experimental confirmation to existing results [8] on the effectiveness of MMD for changepoint detection.

Consequently, MMD may be evaluated as a replacement wherever the distance and dissimilarity measures used in Table 2.1 are employed, for example in visual object tracking, as well as applications outside the field of image processing and analysis.

6.2 Limitations of the work

When comparing results against conventional, non-kernelized algorithms, a difficulty arises in attributing impacts on performance between MMD and the choice or tuning of the selected kernel. When the combination of MMD, kernel, and feature representation works well and outperforms a benchmark, all is well; however, if and when the converse occurs, it becomes difficult to debug the problem.

This remark is a good backdrop for our next comment – the choice of data representations and kernels is very important, as our experiments with scene change detection and quantitative ultrasound data showed. While MMD itself is parameterless, this should be kept in mind by designers, as it means there will be more potential variables and decisions to be optimized compared to a kernel-free design.

Obviously, when the data can already be cleanly separated in the original feature space, the potential gains are very limited, and MMD may not outperform other distances in applications. This could be seen in the results of the ultrasound data, in the case where each parametric map was represented by a single scalar mean; MMD had marginal benefit over ℓ_2, but showed significant gains when a histogram descriptor was used.

Larger group sample sizes may be expected to improve MMD performance. Sample sizes were limited in both applications, as may be expected from early-stage exploratory research. Our review of MMD variants, which are largely reliant on eigendecomposition-based dimensionality reduction, suggested that they will be most effective with larger group sizes m_1, m_2 than we have tested in our experiments.

Finally, the definition of MMD assumes independent and identically distributed (i.i.d.) samples ([57], Lemma 6). In the case of our scene change application, each frame is ordered; frame $t+2$ gives more information about the contents of frame t than frame $t+10$. Despite this apparent theoretical contradiction, MMD still worked well, outperforming the peak-finding result obtained by using the same features and kernel, but without MMD. We note that the other work to address time-series changepoint detection with MMD, [8] does not appear to have explicitly addressed this point. Therefore, this may be one potential avenue for future theory-building.

6.3 Future Work

In closing, we list several ideas for future work below:

1. MMD may be applied for changepoint detection on other time series of multimedia data. Audio segmentation of sounds, speakers, and emotions are potential candidates. Previously difficult domains, such as those described by graphs or trees, may potentially give good results when combined with appropriate features and kernels.

2. In the scene change detection system, our system should be tolerant, in theory, to rotations, fast movement, and other non-semantic changes due to the invariance inherited from the choice of SIFT and LLC descriptors. However, this has not yet been rigorously confirmed in experimental testing. To do this, we have already obtained past NIST TRECVID shot boundary detection datasets, representing hundreds of hours of video, for comparison against published results.

3. Another logical step would be to directly compare against the other test statistics and divergences in Table 2.1, in order to make a stronger statement about the performance of MMD. We note that some of these dissimliarity measures may be turned into kernel measures of similarity, so it may be a matter of cooperation, and not competition, with MMD.

4. One exciting possibility is to use MMD for visual change detection on spatial series of data, as opposed to time series, as a type of novelty detection scheme. Descriptors suited to intra-frame analysis, such as texture and colour descriptors, may be applied to this task.

5. Finally, we have demonstrated promising dissimilarity-based classification results using MMD. A technique that can be used to further improve performance is combining multiple distance measures (e.g. MMD and Mahalanobis distances) for classification, as demonstrated in [69], for MRI-based schizophrenia detection.

References

[1] A. Gretton, K. Borgwardt, M. Rasch, B. Schölkopf, and A. Smola, "A Kernel Method for the Two-Sample-Problem," in *Adv. in Neural Information Processing Systems*, vol. 19. MIT Press, 2006, pp. 513–520.

[2] T. Peeters, P. R. Rodrigues, A. Vilanova, and B. M. Haar Romeny, "Analysis of Distance/Similarity Measures for Diffusion Tensor Imaging," in *Visualization and Processing of Tensor Fields*, D. Laidlaw and J. Weickert, Eds. Springer Berlin Heidelberg, 2009, vol. II, pp. 113–136.

[3] A. Müller, "Integral probability metrics and their generating classes of functions," *Advances in Applied Probability*, pp. 429–443, 1997.

[4] M. Diu, M. Gangeh, and M. S. Kamel, "Unsupervised Visual Changepoint Detection using Maximum Mean Discrepancy (accepted)," in *Proc. of the International Conference on Image Analysis and Recognition.* Springer, 2013.

[5] G. Csurka and C. Dance, "Visual categorization with bags of keypoints," in *Workshop on Statistical Learning in Computer Vision, European Conf. on Computer Vision*, 2004, pp. 1–22.

[6] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, 2006, pp. 2169–2178.

[8] M. Sinn, A. Ghodsi, and K. Keller, "Detecting Change-Points in Time Series by Maximum Mean Discrepancy of Ordinal Pattern Distributions," in *Conf. on Uncertainty in Artificial Intelligence (UAI)*, 2012, pp. 786–794.

[9] R. Szeliski, *Computer Vision: Algorithms and Applications.* Springer, 2010.

[10] G. P. Penney, J. Weese, J. A. Little, P. Desmedt, D. L. Hill, and D. J. Hawkes, "A comparison of similarity measures for use in 2-D-3-D medical image registration." *IEEE Transactions on Medical Imaging*, vol. 17, no. 4, pp. 586–95, Aug. 1998.

[11] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[12] A. R. Webb, *Statistical Pattern Recognition*, 2nd ed. Wiley, 2003.

[13] Y. Rubner, J. Puzicha, C. Tomasi, and J. M. Buhmann, "Empirical evaluation of dissimilarity measures for color and texture," *Computer Vision and Image Understanding*, vol. 84, no. 1, pp. 25–43, 2001.

[14] C. C. Aggarwal, "Re-designing distance functions and distance-based applications for high dimensional data," *ACM SIGMOD Record*, vol. 30, no. 1, pp. 13–18, Mar. 2001.

[15] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Proc. of the 2010 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR).* IEEE, 2010, pp. 3360–3367.

[16] Z. Wang and A. C. Bovik, "Mean squared error: love it or leave it? A new look at signal fidelity measures," *Signal Processing Magazine, IEEE*, vol. 26, no. 1, pp. 98–117, 2009.

[17] T. Hastie, R. Tibshirani, and J. J. H. Friedman, *The Elements of Statistical Learning*, 1st ed. Springer New York, 2001.

[18] S. Kullback and R. A. Leibler, "On information and sufficiency," *The Annals of Mathematical Statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[19] P. E. Hart, R. O. Duda, and D. G. Stork, *Pattern Classification*, 2nd ed. Wiley Chichester, 2001.

[20] J. Puzicha, T. Hofmann, and J. M. Buhmann, "Non-parametric similarity measures for unsupervised texture segmentation and image retrieval," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE Computer Society, 1997, pp. 267–272.

[21] F. J. Massey Jr, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951.

[22] T. W. Anderson, "On the distribution of the two-sample Cramer-von Mises criterion," *The Annals of Mathematical Statistics*, pp. 1148–1159, 1962.

[23] T. Kailath, "The divergence and Bhattacharyya distance measures in signal selection," *IEEE Transactions on Communication Technology*, vol. 15, no. 1, pp. 52–60, 1967.

[24] T. Ojala, M. Pietikäinen, and D. Harwood, "A comparative study of texture measures with classification based on featured distributions," *Pattern Recognition*, vol. 29, no. 1, pp. 51–59, Jan. 1996.

[25] S. Kullback, *Information Theory and Statistics.* Courier Dover Publications, 1968.

[26] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[27] J. Hafner, H. S. Sawhney, W. Equitz, M. Flickner, and W. Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 7, pp. 729–736, 1995.

[28] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Proc. of the Sixth International Conference on Computer Vision.* IEEE, 1998, pp. 59–66.

[29] H. Motulsky, *Intuitive Biostatistics: a Nonmathematical Guide to Statistical Thinking.* Oxford University Press, 2010.

[30] S. Amari, H. Nagaoka, and D. Harada, *Methods of Information Geometry.* American Mathematical Society, 2000.

[31] D. Comaniciu, V. Ramesh, and P. Meer, "Real-time tracking of non-rigid objects using mean shift," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2. IEEE, 2000, pp. 142–149.

[32] J. Zhang, M. Marszaek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: A comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, 2007.

[33] O. Pele and M. Werman, “Fast and robust earth mover’s distances,” in *Proc. of the 12th Int’l Conf. on Computer Vision (ICCV).* IEEE, 2009, pp. 460–467.

[34] C. Cortes and V. Vapnik, “Support-vector networks,” *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995.

[35] H. Daumé III, “From zero to reproducing kernel hilbert spaces in twelve pages or less,” [online, cited Apr. 11, 2013] http://www.umiacs.umd.edu/~hal/docs/daume04rkhs.pdf, 2004.

[36] P. Fieguth, *Statistical Image Processing and Multidimensional Modeling.* Springer Science+Business Media, 2011.

[37] M. Welling, “Lecture notes on kernel ridge regression,” [online, cited Apr. 11, 2013] http://www.ics.uci.edu/~welling/classnotes/papers_class/Kernel-Ridge.pdf, 2005.

[38] K. B. Petersen and M. S. Petersen, “The matrix cookbook,” [online, cited Apr. 21, 2013] http://orion.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf, p. 17, 2008.

[39] T. Hofmann, B. Schölkopf, and A. J. Smola, “Kernel methods in machine learning,” *The Annals of Statistics*, pp. 1171–1220, 2008.

[40] A. Ghodsi, “Dimensionality Reduction: A Short Tutorial,” [online, cited Apr. 11, 2013] http://www.math.uwaterloo.ca/~aghodsib/courses/f06stat890/readings/tutorial_stat890.pdf, 2006.

[41] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K. R. Mullers, “Fisher discriminant analysis with kernels,” in *Proc. of the 1999 IEEE Signal Processing Society Workshop.* IEEE, 1999, pp. 41–48.

[42] C. K. I. Williams and D. Barber, “Bayesian classification with Gaussian processes,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 12, pp. 1342–1351, 1998.

[43] Y. Freund and R. E. Schapire, “Large margin classification using the perceptron algorithm,” *Machine learning*, vol. 37, no. 3, pp. 277–296, 1999.

[44] I. S. Dhillon, Y. Guan, and B. Kulis, “Kernel k-means: spectral clustering and normalized cuts,” in *Proc. of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data mining.* ACM, 2004, pp. 551–556.

[45] A. Nosedal-Sanchez, C. B. Storlie, T. C. M. Lee, and R. Christensen, "Reproducing kernel hilbert spaces for penalized regression: A tutorial," *The American Statistician*, vol. 66, no. 1, pp. 50–60, 2012.

[46] C. E. Rasmussen, *Gaussian Processes for Machine Learning.* The MIT Press, 2005.

[47] L. Rosasco and G. Durrett, "Reproducing Kernel Hilbert Spaces," [online, cited Apr. 11, 2013] http://www.mit.edu/~9.520/spring10/scribe-notes/class03-rkhs-scribe.pdf, 2010.

[48] I. Steinwart, D. Hush, and C. Scovel, "An explicit description of the reproducing kernel Hilbert spaces of Gaussian RBF kernels," *IEEE Transactions on Information Theory*, vol. 52, no. 10, pp. 4635–4643, 2006.

[49] A. Gretton, S. Dino, B. Sriperumbudur, and D. Silver, "Advanced Topics in Machine Learning," [online, cited Apr. 11, 2013] http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/rkhscourse.html.

[50] C.-W. Hsu, C.-C. Chang, and C.-J. Lin, "A practical guide to support vector classification," [online, cited Apr. 11, 2013] www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf, 2003.

[51] C. M. Bishop, *Pattern Recognition and Machine Learning.* Springer, New York, 2006.

[52] J. Wu, "A fast dual method for HIK SVM learning," in *Proc. of the 11th European Conference on Computer Vision (ECCV).* Springer, 2010, pp. 552–565.

[53] R. P. W. Duin, "On the choice of smoothing parameters for Parzen estimators of probability density functions," *IEEE Transactions on Computers*, vol. 100, no. 11, pp. 1175–1179, 1976.

[54] G. R. G. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan, "Learning the kernel matrix with semidefinite programming," *The Journal of Machine Learning Research*, vol. 5, pp. 27–72, 2004.

[55] S. Danafar, A. Gretton, and J. Schmidhuber, "Characteristic kernels on structured domains excel in robotics and human action recognition," in *Machine Learning and Knowledge Discovery in Databases.* Springer, 2010, pp. 264–279.

[56] A. Smola, A. Gretton, L. Song, and B. Schölkopf, "A Hilbert space embedding for distributions," in *Proc. of the 18th International Conference on Algorithmic Learning Theory*, 2007, pp. 13–31.

[57] A. Gretton, "A Kernel Two-Sample Test," *Journal of Machine Learning Research*, vol. 13, pp. 723–773, 2012.

[58] K. M. Borgwardt, A. Gretton, M. Rasch, H.-P. Kriegel, B. Schölkopf, and A. Smola, "Integrating structured biological data by Kernel Maximum Mean Discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. 49–57, 2006.

[59] H. W. Kuhn, "The Hungarian method for the assignment problem," *Naval Research Logistics Quarterly*, vol. 2, no. 12, pp. 83–97, 1955.

[60] F. Desobry, M. Davy, and C. Doncarli, "An online kernel change detection algorithm," *IEEE Transactions on Signal Processing*, vol. 53, no. 8, pp. 2961–2974, 2005.

[61] Z. Harchaoui, F. Vallet, A. Lung-Yut-Fong, and O. Cappe, "A regularized kernel-based approach to unsupervised audio segmentation," in *Proc. of the IEEE International Conference on Acoustics, Speech, and Signal Processing.* IEEE Computer Society, 2009, pp. 1665–1668.

[62] O. Arif and P. A. Vela, "Robust density comparison for visual tracking," in *Proc. of the British Machine Vision Conference*, 2009, pp. 45.1–45.10.

[63] A. Gretton, O. Bousquet, A. Smola, and B. Schölkopf, "Measuring Statistical Dependence with Hilbert-Schmidt Norms," *Biological Cybernetics*, vol. 3734, no. 140, pp. 63–77, 2005.

[64] O. Arif and P. A. Vela, "Robust Density Comparison Using Eigenvalue Decomposition," in *Principal Component Analysis*, P. Sanguansat, Ed. InTech, 2012, ch. 11.

[65] N. H. Anderson, P. Hall, and D. M. Titterington, "Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates," *Journal of Multivariate Analysis*, vol. 50, no. 1, pp. 41–54, 1994.

[66] E. Pękalska and R. P. W. Duin, *The Dissimilarity Representation for Pattern Recognition: Foundations And Applications.* World Scientific, Singapore, 2005.

[67] E. Pękalska, P. Paclik, and R. P. W. Duin, "A Generalized Kernel Approach to Dissimilarity-based Classification," *Journal of Machine Learning Research*, vol. 2, no. 2, pp. 175–211, 2002.

[68] E. Pękalska and R. P. W. Duin, "Dissimilarity representations allow for building good classifiers," *Pattern Recognition Letters*, vol. 23, no. 8, pp. 943–956, Jun. 2002.

[69] A. Ula, R. P. W. Duin, U. Castellani, M. Loog, P. Mirtuono, M. Bicego, V. Murino, M. Bellani, S. Cerruti, and M. Tansella, "Dissimilarity based detection of schizophrenia," *International Journal of Imaging Systems and Technology*, vol. 21, no. 2, pp. 179–192, 2011.

[70] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[71] S. Jegelka and A. Gretton, "Generalized clustering via kernel embeddings," in *Proc. of the 32nd German Conf. on Advances in Artificial Intelligence*, 2009, pp. 144–152.

[72] Y. Rathi, J. Malcolm, O. Michailovich, J. Goldstein, L. Seidman, R. W. McCarley, C.-F. Westin, and M. E. Shenton, "Biomarkers for identifying first-episode schizophrenia patients using diffusion weighted imaging." *Proc. of the International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, vol. 13, no. 1, pp. 657–65, Jan. 2010.

[73] L. Duan, I. W. Tsang, D. Xu, and S. J. Maybank, "Domain Transfer SVM for video concept detection," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition.* IEEE, Jun. 2009, pp. 1375–1381.

[74] S. J. Russell and P. Norvig, *Artificial Intelligence: a Modern Approach*, 2nd ed. Prentice Hall, 2010.

[75] D. W. Scott, *Multivariate density estimation: theory, practice, and visualization.* Wiley, 2009, vol. 383.

[76] K. Pearson, "Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material," *Philosophical Transactions of the Royal Society of London. A*, vol. 186, pp. 343–414, 1895.

[77] C. Snoek and M. Worring, "A review on multimodal video indexing," in *Proc. of the IEEE Intl. Conf. on Multimedia and Expo*, vol. 2. IEEE, 2002, pp. 21–24.

[78] C. Cotsaces, N. Nikolaidis, and I. Pitas, “Video shot detection and condensed representation. a review,” *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 28–37, 2006.

[79] A. Yanagawa and S. Chang, “Columbia University’s baseline detectors for 374 LSCOM semantic visual concepts,” Columbia University ADVENT Technical Report #222-2006-8, 2007.

[80] W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, “A Survey on Visual Content-Based Video Indexing and Retrieval,” *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 41, no. 6, pp. 797–819, 2011.

[81] A. F. Smeaton, P. Over, and A. R. Doherty, “Video shot boundary detection: Seven years of TRECVid activity,” *Computer Vision and Image Understanding*, vol. 114, no. 4, pp. 411–418, 2010.

[82] “Autonomy Virage,” [online, cited Apr. 11, 2013] http://www.virage.com.

[83] M. Omidyeganeh, S. Ghaemmaghami, and S. Shirmohammadi, “Video Keyframe Analysis Using a Segment-Based Statistical Metric in a Visually Sensitive Parametric Space,” *IEEE Transactions on Image Processing*, vol. 20, no. 10, pp. 2730–2737, 2011.

[84] A. Ranganathan, “PLISS: labeling places using online changepoint detection,” *Autonomous Robots*, vol. 32, no. 4, pp. 351–368, 2012.

[85] W. Abd-Almageed, “Online, simultaneous shot boundary detection and key frame extraction for sports videos using rank tracing,” in *Proc. of the 15th IEEE International Conference on Image Processing*. IEEE, 2008, pp. 3200–3203.

[86] G. J. Czarnota and M. C. Kolios, “Ultrasound detection of cell death,” *Imaging in Medicine*, vol. 2, no. 1, p. 12, 2010.

[87] J. Chang, M. Ormerod, T. J. Powles, D. C. Allred, S. E. Ashley, and M. Dowsett, “Apoptosis and proliferation as predictors of chemotherapy response in patients with breast carcinoma.” *Cancer*, vol. 89, no. 11, pp. 2145–52, Dec. 2000.

[88] K. Brindle, “New approaches for imaging tumour responses to treatment.” *Nature Reviews Cancer*, vol. 8, no. 2, pp. 94–107, Feb. 2008.

[89] A. Sadeghi-Naini, O. Falou, J. M. Hudson, C. Bailey, P. N. Burns, M. J. Yaffe, G. J. Stanisz, M. C. Kolios, and G. J. Czarnota, “Imaging innovations for cancer therapy response monitoring,” *Imaging*, vol. 4, no. 3, pp. 311–327, 2012.

[90] L. Fass, "Imaging and cancer: A review," *Molecular Oncology*, vol. 2, no. 2, pp. 115–152, Aug. 2008.

[91] R. Gerl and D. L. Vaux, "Apoptosis in the development and treatment of cancer." *Carcinogenesis*, vol. 26, no. 2, pp. 263–70, Feb. 2005.

[92] J. W. Hunt, A. E. Worthington, A. Xuan, M. C. Kolios, G. J. Czarnota, and M. D. Sherar, "A model based upon pseudo regular spacing of cells combined with the randomisation of the nuclei can explain the significant changes in high-frequency ultrasound signals during apoptosis," *Ultrasound in Medicine & Biology*, vol. 28, no. 2, pp. 217–226, Feb. 2002.

[93] G. J. Czarnota, M. C. Kolios, J. Abraham, M. Portnoy, F. P. Ottensmeyer, J. W. Hunt, and M. D. Sherar, "Ultrasound imaging of apoptosis: high-resolution non-invasive monitoring of programmed cell death in vitro, in situ and in vivo." *British Journal of Cancer*, vol. 81, no. 3, pp. 520–7, Oct. 1999.

[94] A. Sadeghi-Naini, O. Falou, and G. J. Czarnota, "Quantitative ultrasound spectral parametric maps: Early surrogates of cancer treatment response." in *Proc. of the 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Aug. 2012, pp. 2672–2675.

[95] E. J. Feleppa, J. Mamou, C. R. Porter, and J. Machi, "Quantitative ultrasound in cancer imaging." *Seminars in Oncology*, vol. 38, no. 1, pp. 136–50, Feb. 2011.

[96] F. L. Lizzi, M. Greenebaum, E. J. Feleppa, M. Elbaum, and D. J. Coleman, "Theoretical framework for spectrum analysis in ultrasonic tissue characterization." *The Journal of the Acoustical Society of America*, vol. 73, no. 4, pp. 1366–73, Apr. 1983.

[97] B. Banihashemi, R. Vlad, B. Debeljevic, A. Giles, M. C. Kolios, and G. J. Czarnota, "Ultrasound imaging of apoptosis in tumor response: novel preclinical monitoring of photodynamic therapy effects." *Cancer Research*, vol. 68, no. 20, pp. 8590–6, Oct. 2008.

[98] J. Lee, R. Karshafian, N. Papanicolau, A. Giles, M. C. Kolios, and G. J. Czarnota, "Quantitative ultrasound for the monitoring of novel microbubble and ultrasound radiosensitization." *Ultrasound in Medicine & Biology*, vol. 38, no. 7, pp. 1212–21, Jul. 2012.

[99] T. J. Larkin, H. C. Canuto, M. I. Kettunen, T. C. Booth, D.-E. Hu, A. S. Krishnan, S. E. Bohndiek, A. A. Neves, C. McLachlan, M. P. Hobson, and K. M. Brindle,

"Analysis of image heterogeneity using 2D Minkowski functionals detects tumor responses to treatment," *Magnetic Resonance in Medicine [online, cited Apr. 11, 2013] http://onlinelibrary.wiley.com/journal/10.1002/(ISSN)1522-2594*, Feb. 2013.

[100] M. J. Gangeh, A. Sadeghi-Naini, M. S. Kamel, and G. J. Czarnota, "Assessment of cancer therapy effects using texton-based characterization of quantitative ultrasound parametric images," in *Proc. of the International Symposium on Biomedical Imaging (accepted)*, 2013.

[101] L. Sørensen, S. B. Shaker, and M. de Bruijne, "Quantitative analysis of pulmonary emphysema using local binary patterns." *IEEE Transactions on Medical Imaging*, vol. 29, no. 2, pp. 559–69, Feb. 2010.

[102] I. M. Helmy and A. M. A. Azim, "Efficacy of ImageJ in the assessment of apoptosis," *Diagnostic Pathology*, vol. 7, no. 1, pp. 1–6, 2012.

[103] B. Porat, *Digital Processing of Random Signals: Theory and Methods.* Dover Publications, 1994.

[104] F. Dong, E. L. Madsen, M. C. MacDonald, and J. A. Zagzebski, "Nonlinearity parameter for tissue-mimicking materials," *Ultrasound in Medicine & Biology*, vol. 25, no. 5, pp. 831–838, 1999.

[105] R. M. Vlad, S. Brand, A. Giles, M. C. Kolios, and G. J. Czarnota, "Quantitative Ultrasound Characterization of Responses to Radiotherapy in Cancer Mouse Models," *Clinical Cancer Research*, vol. 15, no. 6, pp. 2067–2075, Mar. 2009.

[106] M. Oelze, W. O'Brien, and J. Zachary, "Quantitative Ultrasound Assessment of Breast Cancer Using a Multiparameter Approach," in *Proc. of the 2007 IEEE Ultrasonics Symposium.* IEEE, Oct. 2007, pp. 981–984.

[107] N. R. Jagannathan, M. Kumar, V. Seenu, O. Coshic, S. N. Dwivedi, P. K. Julka, A. Srivastava, and G. K. Rath, "Evaluation of total choline from in-vivo volume localized proton MR spectroscopy and its response to neoadjuvant chemotherapy in locally advanced breast cancer." *British Journal of Cancer*, vol. 84, no. 8, pp. 1016–22, Apr. 2001.