A Klein-Bottle-Based Dictionary for Texture Representation

(International Journal of Computer Vision)

Jose A. Perea*[1] and Gunnar Carlsson†[2]

[1]Department of Mathematics, Duke University, Durham NC 27708
[2]Department of Mathematics, Stanford University, Stanford CA 94503

Abstract

A natural object of study in texture representation and material classification is the probability density function, in pixel-value space, underlying the set of small patches from the given image. Inspired by the fact that small $n \times n$ high-contrast patches from natural images in gray-scale accumulate with high density around a surface $\mathscr{K} \subset \mathbb{R}^{n^2}$ with the topology of a Klein bottle [7], we present in this paper a novel framework for the estimation and representation of distributions around $\mathscr{K}$, of patches from texture images. More specifically, we show that most $n \times n$ patches from a given image can be projected onto $\mathscr{K}$ yielding a finite sample $S \subset \mathscr{K}$, whose underlying probability density function can be represented in terms of Fourier-like coefficients, which in turn, can be estimated from S. We show that image rotation acts as a linear transformation at the level of the estimated coefficients, and use this to define a multi-scale rotation-invariant descriptor. We test it by classifying the materials in three popular data sets: The CUReT, UIUCTex and KTH-TIPS texture databases.

1 Introduction

One representation for texture images which has proven to be highly effective in multi-class classification tasks, is the histogram of texton occurrences [8, 18, 21, 23, 25, 34, 35, 38]. In short, this representation summarizes the number of appearances in an image of either, patches from a fixed set of pixel patterns, or the types of local responses to a bank of filters. Each one of these pixel patterns (or filter responses, if that is the case) is referred to as a texton and the set of textons as a dictionary. Images are then compared via their associated histograms using measures of statistical similarity such as the Earth Mover's distance [29], the Bhattacharya metric [1], or the χ^2 similarity test as introduced by Leung and Malik [25].

For images in gray-scale and dictionaries with finitely many elements, the coding or labeling of $n \times n$ pixel patches, represented as column vectors of dimension n^2, can be seen as fixing a partition $\mathbb{R}^{n^2} = \mathscr{C}_1 \cup \cdots \cup \mathscr{C}_d$ of $\mathbb{R}^{n^2}$ into d distinct classes (each associated to a texton) and letting a patch contribute to the count of the i-th bin in the histogram if and only if it belongs to $\mathscr{C}_i$. For instance, if a patch is labeled according to the dictionary element to which it is closest with respect to a given norm, then the classes $\mathscr{C}_i$ are exactly the Voronoi regions associated to the textons and the norm. If labeling is by maximum response to a (normalized) filter bank, which amounts to selecting the filter with which the patch has largest inner product, then the classes $\mathscr{C}_i$ are the Voronoi regions associated to the filters with distances measured with the norm induced by the inner product used in the filtering stage. More generally, any partition of filter response (or feature) space induces a partition of patch space, by letting two patches be in the same class if and only if the same holds true for their filter responses (resp. associated features).

Several partition schemes have been proposed in the literature. Leung and Malik [25], and Varma and Zisserman [34, 35] have used k-means clustering on filter responses and patches from training images to derive textons, and let the partition be the associated Voronoi tessellation. [21] have implemented an adaptive density-driven tessellation of patch space, while [33] showed that it is possible to obtain high classification success rates with equally spaced bins in an 8-dimensional filter response space. The best classification results to date (and to our knowledge), consistently across widely different and chal-

*joperea@math.duke.edu
†gunnar@math.stanford.edu

lenging data sets, are those of [8]. They propose a geometrically driven partition of a feature space corresponding to multi-scale configurations of image primitives such as blobs, bars and edges.

While the relative merits of one partition over another have been extensively documented in the aforementioned studies, one idea emerges as the consensus: The frequency distribution of patches, in pixel-value space, is one of the most powerful signatures of visual appearance and material class for a texture image. Moreover, any reasonable estimate of the underlying probability density function (the different histograms above, arising from the different partitions, being examples), is highly effective in a variety of challenging classification tasks. One question that arises, however, is what the right space for fitting the distribution should be. That is, should one model how patches from a particular texture class are distributed in pixel-value space, or should the estimation problem be how they accumulate around predictable and low dimensional submanifolds. This question is motivated by the following observations:

1. [33] have noted that when using a regular grid-like partition of an 8-dimensional filter response space, with 5 equally spaced bins per dimension, most bins remain empty. As seen in Figure 8 [33], 5 bins per dimension maximizes classification accuracy in the CUReT database, and increasing or decreasing the number of bins hinders performance. One conclusion which can be drawn from this observation, is that it is likely that filter responses from patches do not fill out response space completely, but instead they populate a lower dimensional (and likely highly non-linear) submanifold.

2. State of the art classifiers require histograms with thousands of bins: 1,296 for the BIFs classifier of Crosier and Griffin [8], and 2,440 for the Joint Classifier of Varma and Zisserman [34]. This is a consequence of the dimension of the feature space where the distribution is being estimated. Since these histograms are usually compared using the nearest neighbor approach, which in high dimensions is computationally costly at best and meaningless at worst [3], then low dimensional representations are desirable.

3. [7] have shown that even when (mean-centered and contrast-normalized) high-contrast 3×3 patches from natural images populate their entire state space (of dimension 7), most of them accumulate with high-density around a 2-dimensional submanifold $\mathscr{K}$ with the topology of a Klein bottle [13]. A detailed explanation of these results will be provided in Section 2.

In short, we advocate for a representation of distributions of patches which is attuned to their likely form. While histograms derived from adaptive binning methods are, to a certain degree, tailored to the form of the distribution, they are likely to suffer from the curse of dimensionality and nonlinearity in the data. Indeed, if the support of the distribution (i.e. the set around which it concentrates) is highly nonlinear, it follows that bins need to be small in order to capture this complexity. Distributions of patches in pixel-value space and 3-jet representation have this property [27]. If in addition the ambient space of the distribution is high-dimensional, then sampling needs to be extremely large otherwise most bins will be empty, hindering estimation. This, we believe, is the main reason why studying distributions around low-dimensional (and highly-nonlinear) subsets of patch-space is relevant. The Klein bottle model provides a first approximation to this approach, but is by no means the complete picture. The classification results we obtain in this paper, however, suggest it is a good place to start.

We present in this paper a novel framework for estimating and representing the distribution around low dimensional submanifolds of pixel space. We show that:

- Propositions 3.1 and 3.3: Most patches from a texture image can be continuously projected onto a model of the Klein bottle, which we denote by K. The choice of this space comes from point 3 above. Thus, via this projection, the set of patches from a given image yields a sample $S \subset K$.

- Theorem 3.5: It is possible to construct orthonormal bases for $L^2(K,\mathbb{R})$, the space of square-integrable functions from K to $\mathbb{R}$, akin to step functions (which yield histogram representations) and Fourier bases. We actually show a bit more: Any orthonormal basis for $L^2(K,\mathbb{C})$ can be recovered explicitly from one of $L^2(T,\mathbb{C})$, where $T = S^1 \times S^1$ denotes the 2-dimensional torus. From here on we will use the notation $L^2(X) := L^2(X,\mathbb{C})$.

- Theorem 3.12: If $f : K \longrightarrow \mathbb{R}$ is the probability density function underlying the sample S, then its coefficients with respect to any orthonormal basis for $L^2(K,\mathbb{R})$ can be estimated from S.

- Theorem 3.15: Image rotation has a specially simple effect on the estimated coefficients of f with respect to a special trigonometric basis. We use this to define a multi-scale rotation-invariant representation for images, which we call the **EKFC-descriptor**.

- The EKFC-descriptor along with a metric learnt through the Large Margin for Nearest Neighbor approach introduced by [37], yields high classification success rates on the CUReT, KTH-TIPS and UIUCTex texture databases (Table 2).

The paper is organized as follows: In **Section 2** we review the results from [24] and [7] on the Klein bottle model $\mathscr{K}$, and discuss how the methods from computational topology [6] can be used to model low-dimensional subspaces of pixel-value space, parametrizing sets of relevant patches. In **Section 3** we develop most of the mathematical machinery. We describe a method for continuously projecting patches onto the Klein bottle $\mathscr{K}$, how to generate (all) orthonormal bases for $L^2(K,\mathbb{R})$, how to estimate the associated coefficients and the effect of image rotation on the resulting K-Fourier representation. **Section 4** begins with a description of the data sets used in the paper, the implementation details of our method, examples of computing the K-Fourier coefficients and the effects of image rotation. We end with the classification results obtained on the CUReT, KTH-TIPS and UIUCTex texture databases. The proofs of the mathematical results presented in this paper can be found in **Appendix A**.

2 Topological Dictionary Learning

The study of the global geometry and statistical properties of spaces of small patches extracted from natural images was undertaken in [24]. One of the goals was to understand how high-contrast 3×3 patches are distributed in pixel-value space, $\mathbb{R}^9$, and the extent to which there exists a global organization, with respect to their pixel patterns, around a recognizable structure. In order to model this distribution a data set $\mathscr{M}$ of approximately 4×10^6 patches from natural images was created, and its study yielded evidence for the existence of a 2-dimensional manifold around which the points in $\mathscr{M}$ accumulate with high density. Their model, which we refer to as the Lee model, consists of patches depicting step-edges at all orientations and distances from the center of the patch.

Subsequently [7], using the persistent homology formalism [6], confirmed the existence of a core high-density 2-dimensional manifold underlying the set $\mathscr{M}$. It was shown that this manifold has the topology of a Klein bottle, and that it consists of the patches from the Lee model as well as light and dark bars at all orientations. A parametrization for this space in terms of degree two polynomials was also provided.

The purpose of this section is to review the models by Lee and Carlsson, and also to make the case that the Klein Bottle is a relevant object when studying the distribution of patches from texture images. We will present several arguments as to why projecting onto the Klein bottle is a reasonable thing to do, and will end with a discussion of what we call Topological Dictionary Learning: The idea that sets of relevant patches from natural image often have global structure (e.g. geometry, topology, parametrizations), that in addition to being accessible yield clear benefits when understood.

2.1 Modeling the Distribution of Small Patches from Natural Images

The van Hateren Natural Image database[1] is a collection of approximately 4,000 monochrome calibrated images, photographed by Hans van Hateren around Groningen (Holland) in town and the surrounding countryside [31].

Figure 1: Exemplars from the van Hateren Natural Images Data base

[24] used this collection of images to model the distribution of 3×3 patches from natural images in their state space. From each image, five thousand 3×3 patches were selected at random and then the entry-wise logarithm of each patch was calculated. This "linearization" step is motivated by Weber's law, which states that the ratio $\Delta L/L$ between the just noticeable difference ΔL and the ambient luminance L is constant for a wide range of values of L. Next, out of the $5,000$ patches from a single image the top 20 percent with respect to contrast were selected. Contrast is measured using the D-norm $\|\cdot\|_D$, where an $n\times n$ patch $P=[p_{ij}]$ yields the vector $\mathbf{v}=[v_1,\ldots,v_{n^2}]^T$ given by $v_{n(j-1)+i}=p_{ij}$, and one lets

$$\|P\|_D^2=\sum_{r\sim t}(v_r-v_t)^2$$

where $v_r=p_{ij}$ is related ($\sim$) to $v_t=p_{kl}$ if and only if

$$|i-k|+|j-l|\leq 1.$$

[1] Available at `http://www.kyb.tuebingen.mpg.de/?id=227`

One thing to note is that the D-norm $\|\cdot\|_D$ can be seen as a discrete version of the **Dirichlet semi-norm**

$$\|I\|_D^2 = \iint_{[-1,1]^2} \|\nabla I(x,y)\|^2 dxdy$$

which is the unique scale-invariant norm on images I. The vectors from the log-intensity, high-contrast patches were then centered by subtracting their mean, and normalized by dividing by their D-norm. This yields approximately 4×10^6 points in $\mathbb{R}^9$, on the surface of a 7-dimensional ellipsoid.

The combinatorics of the relation $\sim$ on pixels from $n \times n$ patches can be described by means of a symmetric $n^2 \times n^2$ matrix D satisfying $\|P\|_D^2 = \langle \mathbf{v}, D\mathbf{v} \rangle$. When $n = 3$ there exists a convenient basis for $\mathbb{R}^9$ of eigenvectors of D called the DCT basis. This basis consists of the constant vector $[1,\ldots,1]^T$, which is perpendicular to the hyperplane containing the centered and contrast-normalized data, and eight vectors which in patch space take the form depicted in Figure 2.

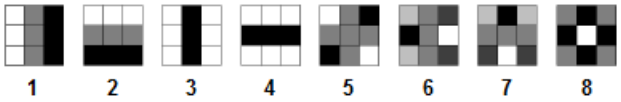

Figure 2: DCT basis in patch space.

By taking the coordinates with respect to the DCT basis of the points in the aforementioned 7-ellipsoid, we obtain a subset $\mathscr{M}$ of the 7-sphere

$$S^7 = \left\{ \mathbf{x} \in \mathbb{R}^8 \;\middle|\; \|\mathbf{x}\| = 1 \right\}$$

which models the distribution of small optical patches in their state space.

2.2 The Model of Lee et al.

The statistical analysis of $\mathscr{M}$ presented by [24] provided the following: High-contrast 3×3 patches from natural images accumulate with high density around a highly non-linear 2-dimensional submanifold of S^7. The way they arrive to this result is as follows: First they divide S^7 into (Voronoi) cells of roughly the same size, and analyze how points from $\mathscr{M}$ are distributed among them. The key observation is that half the points from $\mathscr{M}$ populate most of S^7 sparsely, and the other half are in a few cells which account for less than 6% of the total volume of the 7-sphere.

The (Voronoi) centers of the cells more densely populated by points from $\mathscr{M}$ correspond to patches which depict blurred step edges at all orientations. Since these edges can be described via two numbers: their orientation (i.e. the angle, with the horizontal, of the normal to the dividing boundary of the edge) and the distance from the edge to the center of the patch, it follows that this collection of high-density patches can be parametrized with a 2-dimensional submanifold of pixel-value space. We refer the reader to Figures 7 and 11 of [24], as well as to section 5.1 of the same paper for exemplary high-density patches and the relevant parametrization.

2.3 The Klein Bottle Model

The topological structure and a parametrization for the predicted surface was determined in [7], using the persistent homology formalism; an adaptation of tools from algebraic topology to the world of data [6]. Algebraic topology, an old branch of mathematics, studies the global organization of certain classes of mathematical structures by associating algebraic invariants such as vector spaces and linear maps between them. For closed surfaces without boundary that can be embedded inside a bounded subset of $\mathbb{R}^d$, $d \geq 3$, algebraic invariants which measure number of holes and orientability are complete. In other words, they are enough to determine the identity of a surface, up to continuous deformations which do not change the number of holes. Please refer to [17] for a thorough treatment of homology and the classification theorem of compact surfaces.

The main result of [7] was certainly unexpected: 50% of the points in $\mathscr{M}$ accumulate with high density around a surface $\mathscr{K}$ with the topology of a Klein bottle, which parametrizes both step-edges and bars. Neither the fact that it was a surface (i.e. a manifold of dimension 2) nor its high-density with respect to $\mathscr{M}$ were surprises; these points were already argued in [24]. The important observation is that its global topology is that of a Klein bottle, and that in addition to the step-edges from the Lee model, bars were also included

There are many descriptions and models for spaces with such topology; we recommend [13] as a good compilation. A simple model, which we will use throughout the paper, is the space K obtained from the rectangle $R = \left[\frac{\pi}{4}, \frac{5\pi}{4}\right] \times \left[\frac{-\pi}{2}, \frac{3\pi}{2}\right]$ when each point $\left(\alpha, \frac{-\pi}{2}\right)$ is identified (i.e. glued) with $\left(\alpha, \frac{3\pi}{2}\right)$ and each $\left(\frac{\pi}{4}, \theta\right)$ is identified with $\left(\frac{5\pi}{4}, \pi - \theta\right)$ for every $(\alpha, \theta) \in R$. See Figure 3 for a graphical representation.

Let us describe the patches represented by $\mathscr{K}$ and how together they form a space with the topology of a Klein bottle. The starting point is the collection of patches which describe centered step-edges at all orientations. These patches have a well defined **direction**, by which we mean the angle $\alpha \in \left[\frac{\pi}{4}, \frac{5\pi}{4}\right)$ that the line normal to the boundary of the edge forms with the horizontal.

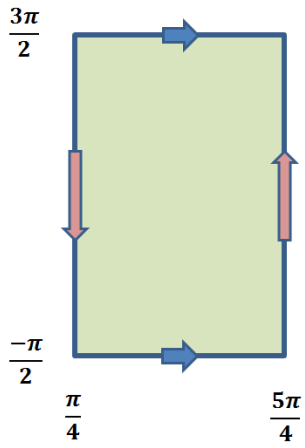

Figure 3: Klein bottle model K. $(\alpha, -\pi/2)$ is identified with $(\alpha, 3\pi/2)$ and $(\pi/4, \theta)$ is identified with $(5\pi/4, \pi-\theta)$ for every (α, θ). Arrows of the same color describe the gluing rules.

Figure 4 shows patches from centered step-edges at various orientations.

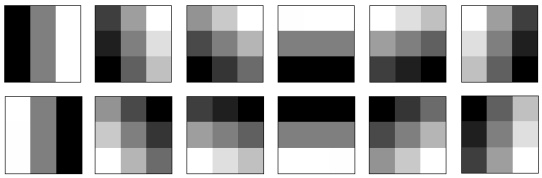

Figure 4: Patches from step-edges at various orientations. The two patches on the far left column have direction $\alpha = \pi$.

As the reader might guess, direction will correspond to the first coordinate in the model K. The next and final step is to add to the Klein bottle manifold those patches which correspond to non-centered step-edges and bars. Consider for a fixed $\alpha \in \left[\frac{\pi}{4}, \frac{5\pi}{4}\right)$ the centered step-edges E and $-E$, the centered dark bar B (i.e. the patch of the form light-strip |dark-strip | light-strip) and the light bar $-B$, all of which have direction α. The definition of direction for bars is analogous to that of step-edges. Moving along the vertical coordinate in the Klein bottle model K (Figure 3) corresponds to the transitions

$$(-B) \Longrightarrow E \Longrightarrow B \Longrightarrow (-E) \Longrightarrow (-B)$$

parametrized by

$$\cos(\theta)E + \sin(\theta)B, \quad \text{for } \theta \in \left[\frac{-\pi}{2}, \frac{3\pi}{2}\right).$$

We show in figure 5 these four transitions for the case $\alpha = \pi$.

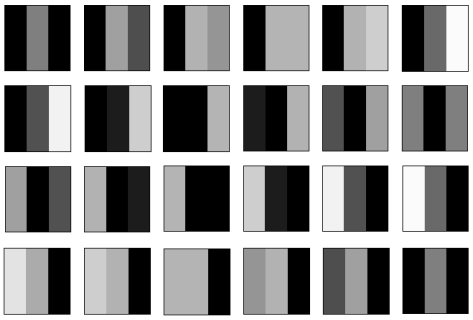

Figure 5: Transitions from bars to edges with direction $\alpha = \pi$. The transition is parametrized by $\cos(\theta)E + \sin(\theta)B$ for $\theta \in \left[\frac{-\pi}{2}, \frac{3\pi}{2}\right)$. The top row, from left to right, corresponds to θ going from $-\pi/2$ to 0. Subsequent rows correspond to the intervals $\left(0, \frac{\pi}{2}\right]$, $\left(\frac{\pi}{2}, \pi\right]$ and $\left(\pi, \frac{3\pi}{2}\right]$, respectively.

This completes the description of the Klein bottle manifold of frequently occurring patches. In summary, $\mathscr{K}$ is comprised of patches depicting step-edges and bars at all orientations. These patches can be parametrized via their direction α, and their edge/bar structure given by the transition angle θ. When we take into account the identifications, i.e. the fact that different pairs (α, θ) might describe the same patch (e.g. both $\left(\frac{\pi}{4}, 0\right)$ and $\left(\frac{5\pi}{4}, \pi\right)$ describe the centered step-edge with gradient in the northeast direction), then the model K (Figure 3) emerges. We show in Figure 6 a lattice of elements from $\mathscr{K}$, arranged according to their coordinates $(\alpha, \theta) \in K$. From this arrangement it readily follows that these patches fit together in a space with the topology of a Klein bottle.

Remark 2.1. For $(\alpha, \theta) \in K$ let $a + ib = e^{i\alpha}$ and $c + id = e^{i\theta}$. Let p be the polynomial

$$p(x,y) = c\frac{(ax+by)}{2} + d\frac{\sqrt{3}(ax+by)^2}{4} \tag{1}$$

and let $P = [p_{ij}]$ be the 3×3 patch obtained via local averaging

$$p_{ij} = \int_{\frac{-2i+3}{3}}^{\frac{-2i+5}{3}} \int_{\frac{2j-5}{3}}^{\frac{2j-3}{3}} p(x,y)\, dxdy, \qquad i,j = 1,2,3.$$

It follows that each $(\alpha, \theta) \in K$ determines a unique patch P, and that by taking entry-wise logarithms, centering and D-normalizing P we get a unique element in $\mathscr{K}$. This implies that $\mathscr{K}$ can be parametrized via the set of polynomials given by equation 1 satisfying $a^2 + b^2 = c^2 + d^2 = 1$.

Remark 2.2. Relevant to our discussion of the distribution of local features from natural images, is the solid of second order local image structure introduced by [15]. The aforementioned

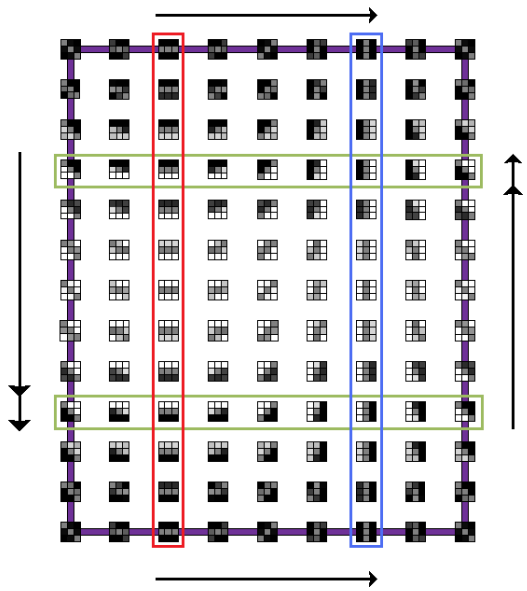

Figure 6: A lattice of patches in $\mathscr{K}$. The first coordinate α measures direction, and the second coordinate θ measures the bar/edge structure.

solid is a 3-dimensional orbifold (i.e. the set of orbits from a group acting on a manifold) constructed as follows: Let $I(x,y)$ be an image and consider, for a fixed scale $\sigma > 0$, the 2nd order jet at (x_0,y_0). That is, the vector

$$\begin{bmatrix} c_{00} & c_{10} & c_{01} & c_{20} & c_{11} & c_{02} \end{bmatrix} \tag{2}$$

resulting from the L^2-inner product of $I(x-x_0,y-y_0)$ with

$$G_\sigma^{(n,m)}(x,y) = \frac{1}{\sigma^2 2\pi}\left(\frac{d^n}{dx^n}e^{\frac{-x^2}{2\sigma^2}}\right)\left(\frac{d^m}{dy^m}e^{\frac{-y^2}{2\sigma^2}}\right)$$

for $n,m = 0,1,2$. The inner product with $G_\sigma^{(n,m)}$ is what we denote by c_{nm}, c_{10} and c_{01} encode the 1D local (blurred-edge-like) structure, while c_{20} and c_{02} measure the pure 2D (centered bar) behavior. The space of 2nd order jets is the set of vectors of the form (2) one obtains by varying both (x_0,y_0) and I. Now, every time an image is translated, rotated or reflected, an intensity constant is added or image intensity is multiplied by a positive factor, then a corresponding transformation (which one can write down explicitly) is applied to its jets. Thus, we can regard these transformations as acting on jet space itself. In this setting, two jets are said to be in the same orbit if one can be obtained from the other by a sequence of such transformations.

The solid defined by [15] is the set of orbits from the space of 2nd order jets via the action of the aforementioned group of transformations. In a nutshell, points in this space describe (up to 2nd order) the local structure of an image in a way which is invariant to rotations, translations, reflections, and alterations of intensity or contrast by a constant factor. A major feature of this highly-curved object is that it can be realized, in a volume-preserving manner, as a subset of $\mathbb{R}^3$, which in turn allows one to study the distribution of 2nd order features from ensembles of images. For natural images, in particular, it is shown that 1D local forms (i.e. edge-like) are overly represented: 50% of the distribution is concentrated in 20% of the volume of the solid.

2.4 The Relevance of Projecting onto the Klein Bottle

The prominence of bars and edges as image primitives has been extensively documented. At the physiological level it is known that the visual cortex in cats and macaque feature large arrays of neurons sensitive only to stimuli from straight lines at specific orientations [19, 20]. In information theory they arise as the filters which minimize redundancy in a linear coding framework [2]; while in feature learning for image representation, bars and edges are prevalent dictionary elements ([21] and [35]).

The Klein bottle model K can thus be regarded as a codebook for two of the most important features in natural images, not only from an statistical standpoint, but from a physiological and computational one. Its continuous character allows one to bypass the quantization problem[2], yielding accurate representations, while its low intrinsic dimensionality (a 2-manifold) provides the sparseness and economy desirable in any coding scheme. Moreover, the fact that the Klein bottle can be seen as a twisted version of the torus $S^1 \times S^1$ (see section 3), allows one to represent probability density functions on it with convenient bases for the space of square-integrable functions $L^2(K)$. This is highly non-trivial if K is replaced by other spaces.

When we project patches from an image onto their closest points in $\mathscr{K}$, we are essentially locally labeling the image according to its localized bar/edge-like structure, while retaining the directionality of the patch. Locally labeling images according to feature categories such as bars, blobs, edges, etc, has been highly successful in texture classification tasks, as shown by the work of [8]. Their success stems from having highly descriptive feature categories at several scales, with features ranging from those ubiquitous across texture images (e.g. edges and bars), to the ones which are more class-characteristic. What our framework provides is a way of concisely representing the ubiquitous part, using the low-dimensional manifold in patch space which best describes it. And what we show in this paper, is that even

[2]Provided one has a "continuous projection" such as the one described in subsection 3.1

with this limited vocabulary it is possible to achieve high classification rates. We do not claim that the Klein bottle manifold is the right space for studying distributions of patches from texture images, but rather that it is a good initial building block (as our results suggest) for a low-dimensional space (not necessarily a manifold) representing a richer vocabulary of features.

We close with a discussion of instances where it is to be expected that relevant portions of patch-space are described with low-dimensional submanifolds, and how one can determine their topology. It is to this framework that we refer to as Topological Dictionary Learning.

2.5 Learning the Topology of Relevant Subsets of Patch-Space

If a set of relevant patches, e.g. in high density regions of state-space or with high discriminative power, can be described with a few attributes having some form of symmetry, then the existence of a low dimensional geometric object which parametrizes them is to be expected. In this case, the topology of the underlying space can be uncovered using the persistent homology formalism [6], while the parametrization step can be aided with methods such as Circular Coordinates [10] or the Topological Nudged Elastic Band[3].

The Klein bottle model $\mathscr{K}$ presented by [7] was indeed the first realization of the Topological Dictionary Learning program, but other features can be also succinctly described. As an example, let us consider the set of $n \times n$ patches depicting bars as the ones in Figure 7.

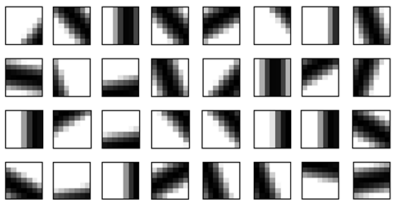

Figure 7: 7×7 patches in gray scale depicting bars at several positions and orientations.

One way of thinking about these patches is as the set of lines in $\mathbb{R}^2$, union an extra point representing the empty (all white) patch. Notice that each line is determined by its orientation and distance to the origin. For a parametrization, let us consider the upper hemisphere of S^2 in spherical coordinates

$$\begin{array}{rcll} x & = & \sin(\phi)\cos(\theta) & \\ y & = & \sin(\phi)\sin(\theta) & 0 \leq \phi \leq \frac{\pi}{2} \quad \text{and} \quad 0 \leq \theta < 2\pi \\ z & = & \cos(\phi) & \end{array}$$

and let (ϕ, θ) describe a patch P as follows : If $0 < \phi \leq \frac{\pi}{2}$, let P correspond to the line ℓ in $\mathbb{R}^2$ with parametric equation

$$\ell(t) \;=\; t \begin{bmatrix} -\sin(\theta) \\ \\ \cos(\theta) \end{bmatrix} \;+\; \cot(\phi) \begin{bmatrix} \cos(\theta) \\ \\ \sin(\theta) \end{bmatrix}.$$

If $\phi = 0$, then let P be the constant patch. If we flatten the upper hemisphere of S^2 so that we get a disk, and place on each polar coordinate (z, θ) the patch one obtains, then it follows that the constant patch will be at the center of the disk, and patches representing lines through the origin will be exactly at the boundary $z = 1$. Another thing to notice is that $(1, \theta)$ yields the same patch as $(1, \theta + \pi)$, and that no other duplications occur. In summary, the set of $n \times n$ patches depicting lines at different orientations and positions, along with the constant patch, can be parametrized via a 2-dimensional disk where boundary points have been identified with their antipodals. Thus, the underlying space is the real projective plane $\mathbb{R}P^2$, a 2-dimensional manifold inside pixel-value space $\mathbb{R}^{n^2}$. Notice that not only do we get a considerable reduction in dimensionality, but the fact that $\mathbb{R}P^2$ can be modeled as a quotient of S^2 allows one to apply geometric techniques from the 2-sphere, to sets of patches depicting lines.

As we have seen, sets of patches which can be described in terms of a few geometric features can often be para- metrized with low dimensional objects embedded in pixel-value space. Next we will concentrate on the Klein bottle dictionary $\mathscr{K}$, and on how to use it in the representation of distributions of patches from texture images.

3 Representing K-Distributions

Given a digital image in gray scale, it follows from the previous section that most of its $n \times n$ patches can be regarded as points in $\mathbb{R}^{n^2}$ close to $\mathscr{K}$. This is the case if n is small with respect to the scale of features in the image, and if "most" refers to the non-constant patches exhibiting a prominent directional pattern. These points, in $\mathbb{R}^{n^2}$, can be projected onto $\mathscr{K}$ yielding a finite set $S \subset K$, which can be interpreted as a random sample from an underlying probability density function $f : K \longrightarrow \mathbb{R}$. We will refer to such an S as a **K-sample**, and to its associated f as its **K-distribution**. Assuming that f is square integrable over K, and

[3] Preprint available at `http://arxiv.org/abs/1112.1993`

given an orthonormal basis $\{\phi_k\}_{k\in\mathbb{N}}$ for $L^2(K,\mathbb{R})$, there exists a unique sequence of real numbers $\{f_k\}_{k\in\mathbb{N}}$ so that

$$f = \sum_{k=1}^{\infty} f_k \phi_k. \tag{3}$$

Moreover since $\|f\|_{L^2}^2 = \sum_{k=1}^{\infty} f_k^2$, then given $\varepsilon > 0$ there exists $J \in \mathbb{N}$ so that $k \geq J$ implies $|f_k| < \varepsilon$, and therefore the vector $(f_1, \ldots, f_J)$ can be regarded as a way of representing (an approximation to) the distribution of $n \times n$ patches from the original image.

The representation for K-distributions we propose in this paper is an estimate of the vector $(f_1, \ldots, f_J)$ from the K-sample $S \subset K$. The basis for $L^2(K,\mathbb{R})$ will be derived from the trigonometric basis $e^{in\theta}$ for $L^2(\mathbb{T})$, where $\mathbb{T}$ denotes the circle $\mathbb{R}/2\pi\mathbb{Z}$. Notice that using approximations via step functions, instead of complex exponentials, recovers a histogram representation (see Remark 3.13).

We will show in this section how to go from an image to the K-sample $S \subset K$, how to construct bases for $L^2(K)$, and then how to estimate the coefficients $f_1, \ldots, f_J$ from S. We show that the estimators for the f_k's can be chosen to be unbiased and with convergence almost surely as $|S| \to \infty$ (Theorem 3.12); that taking finitely many coefficients for our representation is nearly optimal with respect to mean square error; and that when $\{\phi_k\}_{k\in\mathbb{N}}$ is derived from the trigonometric basis $e^{in\theta}$, then image rotation changes our representation via an specific linear transformation (Theorem 3.15).

3.1 From images to K-samples

Given an $n \times n$ patch $P = [p_{ij}]$ from a digital image in gray scale, we can think of it as coming from a differentiable (or almost-everywhere differentiable) intensity function

$$I_P : [-1,1] \times [-1,1] \longrightarrow \mathbb{R}$$

via local averaging

$$p_{ij} = \int_{1-\frac{2i}{n}}^{1-\frac{2i-2}{n}} \int_{-1+\frac{2j-2}{n}}^{-1+\frac{2j}{n}} I_P(x,y)\,dxdy$$

That is, each pixel is the result of integration of light intensity over some finite area. In this context, we will refer to I_P as **an extension** of P.

We say that a patch P is **purely directional** if there exist an extension I_P of P, a unitary vector $\begin{bmatrix} a \\ b \end{bmatrix} \in \mathbb{R}^2$ and a function $g : \mathbb{R} \longrightarrow \mathbb{R}$ such that $I_P(x,y) = g(ax+by)$ for every $x, y \in [-1,1]$. The Klein bottle model $\mathscr{K}$ we have presented, can therefore be understood as the set of purely directional patches in which g is a degree 2 polynomial without constant term, and so that I_P is unitary with respect to the Dirichlet semi-norm

$$\|I_P\|_D^2 = \iint_{[-1,1]^2} \|\nabla I_P\|^2 dxdy.$$

This interpretation allows us to formulate our two-step strategy for projecting a given patch onto K:

1. Find a unitary vector $\begin{bmatrix} a \\ b \end{bmatrix}$ which in average is the most parallel to ∇I_P. This way, I_P will be as constant as possible in the $\begin{bmatrix} -b \\ a \end{bmatrix}$ direction.

2. Let the projection of P onto $\mathscr{K}$ (hence K), be the patch corresponding to the polynomial

$$p(x,y) = c\frac{(ax+by)}{2} + d\frac{\sqrt{3}(ax+by)^2}{4}, c^2+d^2 = 1$$

which is closest to $I_P(x,y)$ as measured by $\|\cdot\|_D$.

Let us see how to carry this out. If I_P is an extension of P then we have the associated quadratic form

$$Q_P(\mathbf{v}) = \iint_{[-1,1]^2} \langle \nabla I_P(x,y), \mathbf{v} \rangle^2 dxdy. \tag{4}$$

Since $\langle \nabla I_P(x,y), \mathbf{v} \rangle^2$ is maximized as a function of $\mathbf{v}$ unitary, at (x,y), when $\mathbf{v}$ is parallel to $\nabla I_P(x,y)$, then maximizing Q_P is in fact equivalent to finding the vectors which in average are the most parallel to ∇I_P.

Proposition 3.1. *Let P be a patch, I_P an extension and let Q_P be the associated quadratic form (equation 4), which we write in matrix form as $Q_P(\mathbf{v}) = \mathbf{v}^T A_P \mathbf{v}$ for A_P symmetric.*

1. *If P is purely directional and $I_P(x,y) = g(ax+by)$, for $a^2+b^2 = 1$, then*

$$Q_P\left(\begin{bmatrix} a \\ b \end{bmatrix}\right) = \max_{\|\mathbf{v}\|=1} Q_P(\mathbf{v})$$

2. *Let $E_{max}(A_P) \subset \mathbb{R}^2$ be the eigenspace corresponding to the largest eigenvalue of A_P. If $S^1 \subset \mathbb{R}^2$ is the unit circle, then $E_{max}(A_P) \cap S^1$ is exactly the set of maximizers for Q_P over S^1.*

3. *If the eigenvalues of A_P are distinct, then*

$$E_{max}(A_P) \cap S^1 = \{\mathbf{v}_P, -\mathbf{v}_P\},$$

which determines a unique $\alpha_P \in [\pi/4, 5\pi/4)$ so that

$$\{\mathbf{v}_P, -\mathbf{v}_P\} = \{e^{i\alpha_P}, -e^{i\alpha_P}\}.$$

Remark 3.2. We would like to point out that Proposition 3.1 is closely related to the well-known corner detector by [16]. If we let $x_1 := x$ and $x_2 := y$, then the entries in the matrix A_P can be written explicitly as

$$A_P(i,j) = \iint_{[-1,1]} \frac{\partial I_P}{\partial x_i} \frac{\partial I_P}{\partial x_j} \, dxdy, \qquad i,j = 1,2.$$

This is exactly the matrix in the Harris-Stephens corner detector. Recall that according to [16], if λ_1, λ_2 are the eigenvalues of A_P then P is a flat region if both eigenvalues are small, it is a corner if both eigenvalues are large, and an edge if one of the eigenvalues is much larger than the other. Since

$$\lambda_1 + \lambda_2 = trace(A_P) = \|I_P\|_D^2 = 1 \quad \text{and} \quad \lambda_i \geq 0$$

then by requiring $\lambda_1 \neq \lambda_2$, we are essentially imposing the existence of a larger eigenvalue and thus we remove the patches depicting corners and flat regions, from the sample to be projected onto the Klein bottle.

Proposition 3.3. *Let $a + ib = e^{i\alpha}$ for $\alpha \in \mathbb{R}$, and let*

$$\langle f, g \rangle_D = \iint_{[-1,1]^2} \langle \nabla f(x,y), \nabla g(x,y) \rangle dxdy$$

denote the inner product inducing the Dirichlet semi-norm $\|\cdot\|_D$. If $u = \frac{(ax+by)}{2}$. Then the vector $\begin{bmatrix} c^* \\ d^* \end{bmatrix} \in S^1$ which minimizes the $\|\cdot\|_D$-error

$$\Phi(c,d) = \left\| I_P - \left(cu + d\sqrt{3}u^2 \right) \right\|_D \quad , \quad c^2 + d^2 = 1$$

is given by

$$c^* = \frac{\langle I_P, u \rangle_D}{\sqrt{\langle I_P, u \rangle_D^2 + 3\langle I_P, u^2 \rangle_D^2}}$$

$$d^* = \frac{\sqrt{3}\langle I_P, u^2 \rangle_D}{\sqrt{\langle I_P, u \rangle_D^2 + 3\langle I_P, u^2 \rangle_D^2}}$$

whenever

$$\varphi(I_P, \alpha) = \langle I_P, u \rangle_D^2 + \langle I_P, u^2 \rangle_D^2 \neq 0$$

and it determines a unique $\theta_P \in [-\pi/2, 3\pi/2)$ so that $c^* + id^* = e^{i\theta_P}$.

The association $P \mapsto (\alpha_P, \theta_P) \in K$, with α_P as in Proposition 3.1 and θ_P as in Proposition 3.3, is well-defined up to a consistent choice of an extension I_P, or rather that of its gradient ∇I_P. The choice we make in this paper is to let ∇I_P be piecewise constant and equal to the discrete gradient of P. The specifics on how to compute this gradient, how to deal with the case in which A_P is a scalar matrix, and what to do when the minimizer for $\Phi(c,d)$ is not well defined, will be discussed in the Implementation Details subsection (4.2).

Let us say a few words regarding the sense in which

$$P \mapsto (\alpha_P, \theta_P) \in K$$

is considered a projection. The orthogonal projection onto a linear subspace of an inner product space is uniquely characterized by its distance minimizing property. While $\mathscr{K}$ is not a linear subspace of $\mathbb{R}^{n^2}$, the approximation

$$I_P \approx c\frac{(ax+by)}{2} + d\frac{\sqrt{3}(ax+by)^2}{4}$$

attempts to minimize the $\|\cdot\|_D$-error as described in Proposition 3.3. It is because of this property that $P \mapsto (\alpha_P, \theta_P)$ is referred to as **the projection of P onto K.**

Our labelling scheme $P \mapsto (\alpha_P, \theta_P) \in K$ can be interpreted, at a given scale, as a continuous version of the MR8 (Maximum Response 8, see Figure 8) representation [35], that omits the Gaussian filter and its Laplacian.

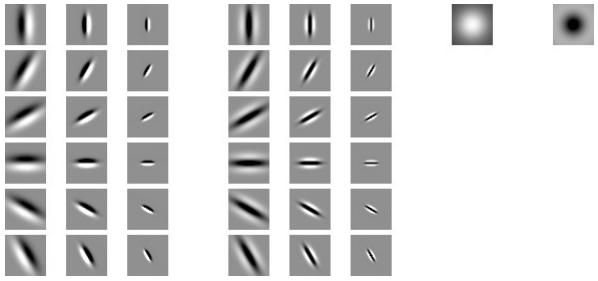

Figure 8: The MR8 (Maximum Response 8) filter banks. By taking the maximum filter response (or projection) across each column, the MR8 represents a patch in terms of eight real numbers.

Indeed, at each one of three scales, the MR8 representation registers the maximum projection (or filter response) of a patch onto six rotated versions of a linear (vertical edge) and a quadratic (vertical bar) filter, yielding the first six out of eight numbers in the descriptor. The last two, are the responses to a Gaussian and to the Laplacian of a Gaussian. The Klein bottle representation on the other hand, takes into account all possible rotations of the linear and quadratic filters, and also retains

the maximizing direction. In contrast to the MR8, we make our descriptor rotation invariant not by collapsing the directional coordinate, but by determining the effect on both the projected sample $S \subset K$ and the estimated Fourier-Like coefficients, as one rotates the image (Theorem 3.15). Once this is understood, one can make the distance between the descriptors invariant to the effects of rotation.

Another difference between our method and the MR8 representation is the way in which scale is handled. On the one hand, the MR8 filter keeps the size of the patch fixed, but changes the inner resolution via convolution with Gaussian kernels. Indeed, computing the filter response of a patch to the k-th derivative of a Gaussian is the same as first blurring the patch with the Gaussian, changing the scale, and then computing the L^2-inner product with a degree k Hermite polynomial. See for instance section 1.4 of [14]. We would like to point out that this falls within the multi-scale representation of Gaussian scale-space theory of [22].

[27] have studied the distribution in scale-space of small patches from natural images via 3-jet representations. Again, they provide evidence for the existence of a highly-nonlinear surface in jet-space, parametrizing multi-scale blurred step-edges. We believe the results of topological dictionary learning and density representation presented in this paper can be adapted to this scale-space framework. We hope to pursue these ideas in future work, but for the moment we use a simple multi-scale representation. In general terms, and in contrast to the MR8 representation, we increase the patch size; for each patch size we obtain a distribution, we estimate its coefficients, and then concatenate them to obtain a multi-scale representation. Please refer to section 4.2 for further details.

3.2 Constructing Bases for $L^2(K)$

The problem of finding good orthonormal bases for $L^2(M)$ when (M,η) is a general measure space, is in fact a very difficult one. For even when there are theoretical guarantees for orthonormal bases to exist, it is highly nontrivial to make principled choices. It is at this point that having the Klein bottle as underlying space ceases to be just an interesting feature and becomes a crucial component of our analysis. As we will see shortly, the Klein bottle is one space for which there are convenient choices of bases for $L^2(K)$.

One of the most fundamental results in the theory of square-integrable periodic functions is Fourier's theorem, stating that the set $\{e^{in\alpha} : n \in \mathbb{Z}\}$ of complex exponentials is an orthonormal basis for $L^2(S^1)$. Here $S^1 = \{z \in \mathbb{C} : |z| = 1\}$, and it readily follows that $\{e^{in\alpha+im\theta} : n,m \in \mathbb{Z}\}$ is an orthonormal basis for $L^2(T)$, where $T = S^1 \times S^1$ denotes the 2-dimensional torus. In other words, it is possible to choose convenient bases for $L^2(S^1)$, including but not limited to trigonometric exponentials, and these in turn yield bases for $L^2(T)$. The main point now is that since the Klein bottle K can be recovered from T via simple identifications based on symmetries, as we will see next, then finding orthonormal bases for $L^2(K)$ amounts to doing so for $L^2(T)$ while keeping track of said identifications.

The aforementioned symmetries arise naturally from the models we have for the Klein bottle of edge-like and bar-like frequently occurring patches. Indeed, it follows from the parametrization of $\mathscr{K}$ via the set of polynomials

$$p(x,y) = c\frac{(ax+by)}{2} + d\frac{\sqrt{3}(ax+by)^2}{4}$$

with $a^2+b^2 = c^2+d^2 = 1$ (see Remark 2.1), that each (z,w) in T determines a unique element in $\mathscr{K}$, and that (z,w) and (z',w') yield the same patch if and only if one has that $(z',w') = (-z,-\overline{w})$. Here $\overline{w}$ denotes the complex conjugate of w. That is, $\mathscr{K}$ (and therefore K) can be modeled as the space obtained from T by identifying (z,w) with $(-z,-\overline{w})$ for every $(z,w) \in T$. Thus, in essence, one can interpret K as

$$\Big\{\big\{(z,w),(-z,-\overline{w})\big\} : (z,w) \in T\Big\}$$

and think of elements in $L^2(K)$ as square-integrable functions $f : T \longrightarrow \mathbb{C}$ satisfying the identity

$$f(z,w) = f(-z,-\overline{w}) \tag{5}$$

for every $(z,w) \in T$. The reason why we think of these relations as symmetries is explained in Figure 9.

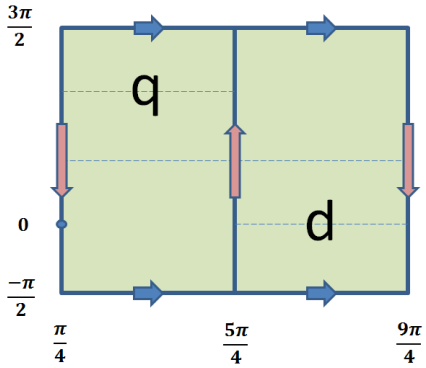

Figure 9: The transformation $\mu(z,w) = (-z,-\overline{w})$ can be written in polar coordinates as $\mu(\alpha,\theta) = (\alpha+\pi,\pi-\theta) \bmod 2\pi$. In particular, it maps the letter "q" on the upper left to the "d" on the lower right. The set $\big\{\{(z,w),\mu(z,w)\} \mid (z,w) \in T\big\}$ can be identified with $[\pi/4,5\pi/4] \times [-\pi/2,3\pi/2]$ along with the gluing rules depicted by the arrows. This recovers the model K.

Notice that this characterization allows one to interpret $L^2(K)$ as a linear subspace of $L^2(T)$. This leads one to consider the orthogonal projection

$$\Pi : L^2(T) \longrightarrow L^2(K)$$

of $L^2(T)$ onto $L^2(K)$, which in turn yields a recipe for producing orthonormal bases for $L^2(K)$ from those of $L^2(T)$. We describe this recipe in what follows. Firstly, an explicit formula for Π can be derived from equation 5. Indeed,

Proposition 3.4. *The function*

$$\begin{array}{rccc} \Pi: & L^2(T) & \longrightarrow & L^2(K) \\ & g & \mapsto & \frac{g(z,w)+g(-z,-\overline{w})}{2} \end{array} \tag{6}$$

is the orthogonal projection of $L^2(T)$ onto $L^2(K)$.

Notice that if $\{\phi_k | k \in \mathbb{N}\}$ is a basis (or a spanning set) for $L^2(T)$ then $\{\Pi(\phi_k) | k \in \mathbb{N}\}$ spans $L^2(K)$ and hence contains a basis. Moreover, applying Gram-Schmidt allows one to turn this basis into an orthonormal one. In other words, any orthonormal basis for $L^2(T)$ yields one for $L^2(K)$. A less obvious observation, whose proof can be found in the appendix, is that every orthonormal basis for $L^2(K)$ can be obtained from one of $L^2(T)$ in this fashion. We summarize the previous discussion in the following theorem.

Theorem 3.5. *If $\{\phi_k \mid k \in \mathbb{N}\}$ is a spanning set for $L^2(T)$ then*

$$\mathscr{S} = \left\{ \frac{\phi_k(z,w) + \phi_k(-z,-\overline{w})}{2} \;\middle|\; k \in \mathbb{N} \right\}$$

is a spanning set for $L^2(K)$.

If $\mathscr{B} \subset \mathscr{S}$ is a basis for $L^2(K)$, then applying Gram-Schmidt to $\mathscr{B}$ yields an orthonormal basis for $L^2(K)$. Moreover, any orthonormal basis for $L^2(K)$ can be recovered from one of $L^2(T)$ in this fashion.

Remark 3.6. Since bases for $L^2([0,1])$, such as Wavelets or Legendre polynomials, can be modified to yield bases for $L^2(S^1)$ and thus for $L^2(T)$, then the previous theorem gives a way of constructing lots of bases for $L^2(K)$.

In particular we have,

Corollary 3.7. *Let $\{\phi_n\}_{n\in\mathbb{N}}$ be a basis for $L^2(S^1)$. Then in polar coordinates*

$$\{\phi_n(\alpha)\phi_m(\theta) \mid n,m \in \mathbb{N}\}$$

is a basis for $L^2(T)$ and therefore

$$\left\{ \frac{\phi_n(\alpha)\phi_m(\theta) + \phi_n(\alpha+\pi)\phi_m(\pi-\theta)}{2} \;\middle|\; n,m \in \mathbb{N} \right\}$$

is a spanning set for $L^2(K)$.

The next two results are perhaps the most relevant for the rest of the paper; they yield Fourier-like bases for $L^2(K)$ and $L^2(K,\mathbb{R})$ by applying the previous corollary to the set $\{\phi_k(\alpha) = e^{ik\alpha}\}_{k\in\mathbb{Z}}$.

Corollary 3.8. *Let $n \in \mathbb{Z}$ and $m \in \mathbb{N} \cup \{0\}$. Then the set of functions*

$$\frac{e^{in\alpha+im\theta} + (-1)^{n+m} e^{in\alpha-im\theta}}{2}, \quad m = 0 \text{ implies } n \text{ is even.}$$

*is an orthonormal basis for $L^2(K)$. We refer to this set as the **trigonometric basis for** $L^2(K)$.*

When restricted to real valued functions $f : K \longrightarrow \mathbb{R}$, one can start with the basis $\mathscr{B}$ for $L^2(T,\mathbb{R})$ consisting of functions of the form $\phi(n\alpha)\psi(m\theta)$ where ϕ and ψ are either sines or cosines, and consider $\Pi(\mathscr{B})$. This yields

Corollary 3.9. *Let $n,m \in \mathbb{N}$ and let $\pi_{n,m} = \frac{(1-(-1)^{n+m})\pi}{4}$. Then the set of functions*

$$1,\ \sqrt{2}\cos(m\theta - \pi_{0,m}),\ \sqrt{2}\cos(2n\alpha),\ \sqrt{2}\sin(2n\alpha),$$

$$2\cos(n\alpha)\cdot\cos(m\theta - \pi_{n,m}),\ 2\sin(n\alpha)\cdot\cos(m\theta - \pi_{n,m})$$

*is an orthonormal basis for $L^2(K,\mathbb{R})$. We refer to this set as the **trigonometric basis for** $L^2(K,\mathbb{R})$.*

Remark 3.10. The symmetries which allow us to regard $L^2(K)$ as a subspace of $L^2(T)$ (equation 5) have a simplifying effect on the inner product $\langle\cdot,\cdot\rangle_T$ when restricted to $L^2(K)$. Indeed, if $f,g \in L^2(K)$ then one can check that

$$\begin{aligned} \langle f,g\rangle_T &= \frac{1}{(2\pi)^2}\int_{\frac{-\pi}{2}}^{\frac{3\pi}{2}}\int_{\frac{\pi}{4}}^{\frac{9\pi}{4}} f(\alpha,\theta)\overline{g}(\alpha,\theta)d\alpha d\theta \\ &= \frac{1}{2\pi^2}\int_{\frac{-\pi}{2}}^{\frac{3\pi}{2}}\int_{\frac{\pi}{4}}^{\frac{5\pi}{4}} f(\alpha,\theta)\overline{g}(\alpha,\theta)d\alpha d\theta \end{aligned}$$

and thus $L^2(K)$ can be endowed with its own inner product:

$$\langle f,g\rangle_K := \frac{1}{2\pi^2}\int_{\frac{-\pi}{2}}^{\frac{3\pi}{2}}\int_{\frac{\pi}{4}}^{\frac{5\pi}{4}} f(\alpha,\theta)\overline{g}(\alpha,\theta)d\alpha d\theta. \tag{7}$$

Definition 3.11. Let $f \in L^2(K,\mathbb{R})$, $\langle\cdot,\cdot\rangle_K$ as in equation 7, and let

$$\begin{aligned} a_m &= \langle f, \sqrt{2}\cos(m\theta - \pi_{0,m})\rangle_K \\ b_n &= \langle f, \sqrt{2}\cos(2n\alpha)\rangle_K \\ c_n &= \langle f, \sqrt{2}\sin(2n\alpha)\rangle_K \\ d_{n,m} &= \langle f, 2\cos(n\alpha)\cos(m\theta - \pi_{n,m})\rangle_K \\ e_{n,m} &= \langle f, 2\sin(n\alpha)\cos(m\theta - \pi_{n,m})\rangle_K \end{aligned}$$

be the coefficients of f with respect to the trigonometric basis for $L^2(K,\mathbb{R})$. If we order them with respect to their (total) frequencies and alphabetic placement as

$$a_1, \underbrace{a_2, b_1, c_1, d_{1,1}, e_{1,1}}_{\text{frequency=2}}, \underbrace{a_3, d_{1,2}, d_{2,1}, e_{1,2}, e_{2,1}}_{\text{frequency=3}}, a_4, b_2, \ldots$$

then we get the ordered sequence $K\mathscr{F}(f)$, which we will refer to as **the K-Fourier coefficients** of f. We will denote by $K\mathscr{F}_w(f)$ the truncated sequence of K-Fourier coefficients containing those terms with frequencies less than or equal to $w \in \mathbb{N}$.

3.3 Estimation in the Frequency Domain

We now turn our attention to the last step in the representation scheme: Given an orthonormal basis $\{\phi_k\}_{k\in\mathbb{N}}$ for $L^2(K)$, $f \in L^2(K)$ a probability density function on K, and $S \subset K$ a random sample drawn according to f, estimate the coefficients $f_k = \langle f, \phi_k \rangle_K$ from S. It turns out that this problem can be solved in great generality: For any measure space (M,η) so that $L^2(M)$ has orthonormal bases with at most countably many elements, one can choose unbiased estimators for the coefficients f_k. In very general terms, what one does is to put a Dirac mass at each point of S, interpret this construction as the Gaussian kernel estimator [30] with infinitely small width, and then calculate its coefficients with respect to the basis $\{\phi_k\}_{k\in\mathbb{N}}$.

Let us motivate the definition of these estimators with the case $M = \mathbb{R}^d$ and $\{h_\alpha \mid \alpha \in \mathbb{N}^d\}$ the harmonic oscillator wave functions on $\mathbb{R}^d$. Let $\mathscr{S}(\mathbb{R}^d)$ be the set of rapidly decreasing complex valued functions on $\mathbb{R}^d$ (see [28]), $\mathscr{S}'(\mathbb{R}^d)$ its algebraic dual (the space of tempered distributions) and let $h_\alpha : \mathbb{R}^d \longrightarrow \mathbb{R}$ be given by

$$h_\alpha(x_1,\ldots,x_d) = h_{\alpha_1}(x_1)\cdots h_{\alpha_d}(x_d).$$

Here $\alpha = (\alpha_1,\ldots,\alpha_d) \in \mathbb{N}^d$ and h_k is the k-th Hermite function on $\mathbb{R}$. It is known that $\{h_\alpha \mid \alpha \in \mathbb{N}^d\}$ is an orthonormal basis for $L^2(\mathbb{R}^d)$ (Lemma 3, page 142, [28]), and that for every $T \in \mathscr{S}'(\mathbb{R}^d)$ the series

$$\sum_\alpha T(\overline{h_\alpha}) h_\alpha$$

converges [4] to T. This suggests we define $T(\overline{\phi_k})$ as the coefficients of a tempered distribution T with respect to an arbitrary orthonormal basis $\{\phi_k\}_{k\in\mathbb{N}}$ for $L^2(\mathbb{R}^d)$.

Now, the Dirac delta function centered at $\mathbf{c} \in \mathbb{R}^d$

$$\begin{array}{rccc} \delta(\mathbf{x}-\mathbf{c}) & : \mathscr{S}(\mathbb{R}^d) & \longrightarrow & \mathbb{C} \\ & \psi & \mapsto & \psi(\mathbf{c}) \end{array}$$

is a continuous linear functional and therefore an element in $\mathscr{S}'(\mathbb{R}^d)$. An application of integration by parts shows that for every $\psi \in \mathscr{S}(\mathbb{R}^d)$ one has

$$\lim_{k\to\infty} \int_{\mathbb{R}^d} \psi(\mathbf{x}) \frac{k}{\sqrt{\pi^d}} e^{-k^2\|\mathbf{x}-\mathbf{c}\|^2}\, d\eta(\mathbf{x}) = \psi(\mathbf{c})$$

and therefore $\delta(\mathbf{x}-\mathbf{c})$ is a weak limit of Gaussians with mean $\mathbf{c}$, as the variance approaches zero. What this result tells us is that $\delta(\mathbf{x}-\mathbf{c})$ can be interpreted as the generalized function bounding unit volume, which is zero at $\mathbf{x} \neq \mathbf{c}$ and an infinite spike at $\mathbf{c}$. Moreover, if $\mathbf{X}_1,\ldots,\mathbf{X}_N$ are i.i.d. (independent and identically distributed) random variables with probability density function $f \in L^2(\mathbb{R}^d,\mathbb{R})$, then as a "generalized statistic"

$$\widehat{f_\delta}(\mathbf{X}) := \frac{1}{N}\sum_{n=1}^N \delta(\mathbf{X}-\mathbf{X}_n)$$

can be thought of as the Gaussian kernel estimator (see [30]) with infinitely small width. The coefficients of $\widehat{f_\delta}$ with respect to the Hermite basis can be computed as

$$\widehat{f_\delta}(\overline{h_\alpha}) = \frac{1}{N}\sum_{n=1}^N \overline{h_\alpha}(\mathbf{x}_n)$$

and motivate the following estimation theorem.

Theorem 3.12. *Let (M,η) be a measure space so that $L^2(M)$ has an orthonormal basis $\{\phi_k\}_{k\in\mathbb{N}}$. If $f : M \longrightarrow \mathbb{R}$ is a probability density function with*

$$f = \sum_{k=1}^\infty f_k \phi_k$$

and $X_1,\ldots,X_N$ are i.i.d. random variables distributed according to f, then

$$\widehat{f_k} := \frac{1}{N}\sum_{n=1}^N \overline{\phi_k}(X_n)$$

is an unbiased estimator for f_k. Moreover, $\widehat{f_k}$ converges almost surely to f_k as $N \to \infty$.

Proof. To see that the estimators are unbiased, notice that for each $k \in \mathbb{N}$ the expected value of $\widehat{f_k}$ is given by

$$\mathbb{E}[\widehat{f_k}] = \frac{1}{N}\int_M f \cdot \sum_{n=1}^N \overline{\phi_k}\, d\eta = \int_M f \cdot \overline{\phi_k}\, d\eta = f_k.$$

The convergence result is exactly the statement of the Strong Law of Large Numbers. □

[4] Convergence is with respect to the weak-* topology. This result is a consequence of the N-representation theorem (V.14, page 143, [28])

Remark 3.13. Our formalism of estimated coefficients also recovers histogram representations. Indeed, let $A \subset [0,1]$ and consider the (indicator) function $\mathbb{1}_A : [0,1] \longrightarrow \mathbb{R}$ which for $x \in [0,1]$ is defined as 1 if $x \in A$, and 0 otherwise. A **step function** is any expression of the form

$$\mathfrak{s}(x) = \sum_{i=1}^{m} r_i \mathbb{1}_{A_i}(x)$$

where $r_i \in \mathbb{R}$ is nonzero and the A_i's are disjoint subintervals of $[0,1]$. In the same way as the integral of a function can be approximated via Riemann sums, one has that the set of step functions is dense in $L^2([0,1])$. This means that for any $\varepsilon > 0$ and any $f \in L^2([0,1])$ there exists a step function $\mathfrak{s}$ so that $\|f - \mathfrak{s}\|_{L^2} < \varepsilon$. Fix $f \in L^2(K)$, $\varepsilon > 0$ and let $\mathfrak{s}$ be a step function no more than ε away from f. Let $A_1, \ldots, A_m$ be the subintervals of $[0,1]$ which define $\mathfrak{s}$. Since the A_i's are disjoint, it follows that the functions $\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_m}$ are mutually orthogonal in $L^2([0,1])$ and hence can be extended, as a linearly independent set, to an orthogonal basis $\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_m}, \phi_{m+1}, \phi_{m+2} \ldots$ of $L^2([0,1])$. Thus, we can write

$$f(x) = \sum_{i=1}^{m} f_i \frac{\mathbb{1}_{A_i}(x)}{\sqrt{|A_i|}} + \sum_{j=m+1}^{\infty} f_j \phi_j(x)$$

where $|A_i|$ denotes the length of the interval A_i and the ϕ_j can be assumed to have unit norm. If f is a density function on $[0,1]$ and $S = \{x_1, \ldots, x_N\} \subset [0,1]$ is drawn according to f, then Theorem 3.12 implies that one can estimate the first m coefficients $f_1, \ldots, f_m$ (which describe f with accuracy ε) as

$$\widehat{f_i} = \frac{1}{N} \sum_{n=1}^{N} \frac{\overline{\mathbb{1}_{A_i}}(x_n)}{\sqrt{|A_i|}} = \frac{n_i}{N\sqrt{|A_i|}} \ , \qquad i = 1, \ldots, m$$

Here n_i denotes the number of points from the sample S which lie inside the interval A_i. In other words, the vector of estimated coefficients $(\widehat{f_1}, \ldots, \widehat{f_m})$, with respect to the aforementioned basis, is exactly the normalized histogram on the intervals $A_1, \ldots, A_m$. Notice that approximation with step functions makes sense for any bounded space; in particular on the Klein bottle K.

Definition 3.14. Let $\widehat{K\mathscr{F}}(f)$ denote the sequence of estimators obtained from Theorem 3.12 and the trigonometric basis for $L^2(K, \mathbb{R})$. Let $\widehat{K\mathscr{F}}_w(f)$ denote the truncated sequence corresponding to frequencies less than or equal to $w \in \mathbb{N}$. Given a random sample $S = \{(\alpha_1, \theta_1), \ldots, (\alpha_N, \theta_N)\} \subset K$ drawn according to $f \in L^2(K, \mathbb{R})$, we let $\widehat{K\mathscr{F}}(f, S)$ denote the sequence of **estimated K-Fourier coefficients**

$$\begin{aligned}
\widehat{a}_m &= \frac{1}{N} \sum_{k=1}^{N} \sqrt{2} \cos(m\theta_k - \pi_{0,m}) \\
\widehat{b}_n &= \frac{1}{N} \sum_{k=1}^{N} \sqrt{2} \cos(2n\alpha_k) \\
\widehat{c}_n &= \frac{1}{N} \sum_{k=1}^{N} \sqrt{2} \sin(2n\alpha_k) \\
\widehat{d}_{n,m} &= \frac{1}{N} \sum_{k=1}^{N} 2 \cos(n\alpha_k) \cos(m\theta_k - \pi_{n,m}) \\
\widehat{e}_{n,m} &= \frac{1}{N} \sum_{k=1}^{N} 2 \sin(n\alpha_k) \cos(m\theta_k - \pi_{n,m})
\end{aligned}$$

and let $\widehat{K\mathscr{F}}_w(f, S)$ be its truncated version.

Now that we have established some properties of the estimators, let us end this section with a discussion on why truncating the sequence of coefficients is justified. It is the case that while there are reasonably good estimators for the coefficients of f in very general settings, their dependency on the choice of basis makes it hard to identify just how much information is effectively encoded in this representation. Let us try to make this idea more transparent. The first thing to notice is that the sequence of estimators

$$S_J(\widehat{f_\delta}) := \sum_{k=1}^{J} \widehat{f_k} \phi_k$$

does not converge to f but to $\widehat{f_\delta}$ (in the weak-* topology) as $J \to \infty$, regardless of the sample size N, and provided we have pointwise convergence

$$\sum_k \psi_k \phi_k(\mathbf{x}) = \psi(\mathbf{x}), \quad \text{for all } \psi \in \mathscr{S}(\mathbb{R}^d) \text{ and all } \mathbf{x} \in \mathbb{R}^d.$$

This is true for instance in the case of Hermite functions and trigonometric polynomials. More specifically, what Theorem 3.12 provides is a way of approximating a linear combination of delta functions instead of the actual function f, and thus even when the $\widehat{f_k}$ are almost surely correct for large N, the convergence in probability is not enough to make them decrease fast enough as k gets larger. In short, taking more coefficients is not necessarily a good thing, so one should choose bases in which the first coefficients carry most of the information making truncation not only necessary but meaningful. Moreover, it is known [36] that truncation is also nearly optimal with respect to mean integrated square error (MISE) within a natural class of statistics for f. Indeed, if within the family of estimators

$$f_N^* := \sum_{k=1}^{\infty} \lambda_k(N) \widehat{f_k} \phi_k$$

one looks for the sequence $\{\lambda_k(N)\}_{k\in\mathbb{N}}$ that minimizes

$$\text{MISE} := \mathbb{E}\left[\, \|f_N^* - f\|_2^2 \,\right]$$

then provided the f_k for $k \leq J(N)$ are large compared to $\frac{var(\phi_k)}{N}$ and the f_k for $k > J(N)$ are negligible, then letting $\lambda_k(N) = 1$ for $k \leq J(N)$ and zero otherwise achieves nearly optimal mean square error.

The previous discussion yields yet another reason for considering trigonometric polynomials: The high-frequency K-Fourier coefficients tune the fine scale details of the series approximation. Thus, most of the relevant information in the probability density function is encoded in the low frequencies, which can be easily approximated via estimators that converge almost surely as the sample size gets larger, and achieve nearly optimal mean square error within a large family of estimators.

3.4 The Effect of Image Rotation

We will show in what follows that image rotation has a particularly simple effect on the distribution of patches on the Klein bottle and that, with respect to the trigonometric basis, it has an straightforward interpretation in the frequency domain: A linear transformation depending solely on the angle of rotation.

Let us consider an $n \times n$ patch P from an image I. If I is rotated in the counterclockwise direction with respect to its center by τ degrees, $\tau \in [-\pi, \pi]$, then one obtains a new image I^τ and P will be mapped to a patch P^τ. While this new patch is not exactly the same as rotating P by τ degrees, since we are taking square patches instead of disks, it follows that if I has a predominantly directional pattern with angle α in a pixel neighborhood of P, then P^τ will have predominant direction $\alpha + \tau$. Moreover, the edge-like or bar-like structure of P^τ (i.e. the vertical coordinate in the Klein bottle model; see Figure 6) will be roughly that of P. In summary, given that the horizontal coordinate (α) in the model K encodes the predominant direction of a patch, and the vertical coordinate (θ) parametrizes its edge/bar-like structure, then if (α_P, θ_P) are the coordinates of P when projected onto K, we have that $(\alpha_P + \tau, \theta_P)$ approximate those of P^τ.

It is clear that taking $(\alpha_P + \tau, \theta_P)$ as an approximation for the projection of P^τ onto K could be inaccurate for the patches in I which do not have a neighborhood where the image is predominantly directional. What one expects, however, is that by taking the collection of all patches from I^τ projected onto K, the local inaccuracies will be negligible when considering the global behavior of the distribution. In other words, if $S \subset K$ is the sample obtained from patches of I and we let

$$S^\tau = \{(\alpha + \tau, \theta) \mid (\alpha, \tau) \in S\}$$

then if $\widetilde{S} \subset K$ is the sample from projected patches of I^τ we have that S^τ and $\widetilde{S}$ can be thought of as sampled from the same distribution. Let $f : K \longrightarrow \mathbb{R}$ be the density function underlying S. Since S^τ is obtained from S by horizontal translation on K, we have that S^τ (and therefore $\widetilde{S}$) has distribution

$$f^\tau(\alpha, \theta) := f(\alpha - \tau, \theta). \tag{8}$$

What we will see now is that the trigonometric basis on K is specially well suited for understanding the effect that image rotation has on the frequency domain. Indeed,

Theorem 3.15. *Let $a_m, b_n, c_n, d_{n,m}, e_{n,m}$ be the K-Fourier coefficients of $f \in L^2(K, \mathbb{R})$, and let $a_m^\tau, b_n^\tau, c_n^\tau, d_{n,m}^\tau, e_{n,m}^\tau$ be those of $f^\tau(\alpha, \theta) = f(\alpha - \tau, \theta)$. Then we have the identities*

$$a_m^\tau = a_m \tag{9}$$

$$\begin{bmatrix} b_n^\tau \\ c_n^\tau \end{bmatrix} = \begin{bmatrix} \cos(2n\tau) & -\sin(2n\tau) \\ \sin(2n\tau) & \cos(2n\tau) \end{bmatrix} \begin{bmatrix} b_n \\ c_n \end{bmatrix} \tag{10}$$

$$\begin{bmatrix} d_{n,m}^\tau \\ e_{n,m}^\tau \end{bmatrix} = \begin{bmatrix} \cos(n\tau) & -\sin(n\tau) \\ \sin(n\tau) & \cos(n\tau) \end{bmatrix} \begin{bmatrix} d_{n,m} \\ e_{n,m} \end{bmatrix} \tag{11}$$

for every $\tau \in \mathbb{R}$ and every $n, m \in \mathbb{N}$. Thus, there exists a linear map $T_\tau : \ell^2(\mathbb{R}) \longrightarrow \ell^2(\mathbb{R})$ which preserves lengths, depends solely on τ, and so that $K\mathscr{F}(f^\tau) = T_\tau(K\mathscr{F}(f))$ for every $f \in L^2(K, \mathbb{R})$. Here $\ell^2(\mathbb{R})$ denotes the set of square summable sequences of real numbers.

This discussion can be summarized as follows: Rotating an image in the counterclockwise direction by τ degrees with respect to its center, operates as a translation of its distribution of patches on the Klein bottle model K horizontally by τ units. Please refer to figure 10 for an example.
This shift of the sample translates into the K-Fourier coefficients domain as a linear transformation T_τ, which can be described as a block diagonal matrix whose blocks are either the number 1, or rotation matrices which act by an angle $2n\tau$ or $n\tau$ depending on the particular frequencies.

This description allows us to define a distance between K-Fourier coefficients which is invariant under image rotation. Indeed, consider two images I and J, and let $S, S' \subset K$ be the samples obtained from projecting their $n \times n$ patches onto the Klein bottle. Moreover, let f and g in $L^2(K, \mathbb{R})$ be the probability density functions underlying S and S', respectively. Recall that $\widehat{K\mathscr{F}}_w(f, S)$ and $\widehat{K\mathscr{F}}_w(g, S')$ denote the truncated K-Fourier coefficients of f and g estimated from the samples S and S'. Each rotation I^τ of I by τ, has the effect of shifting S

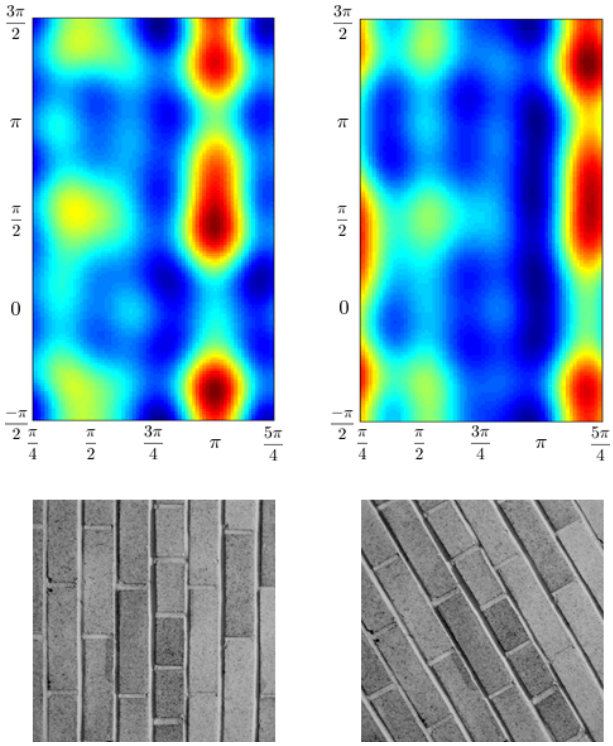

Figure 10: **Top:** distribution on the Klein bottle model K of projected 13×13 patches from the images on the bottom. High and low density are represented using the colors red and blue, respectively. Notice that the one on the right is roughly a translated version of the one on the left by $\frac{\pi}{6}$, from left to right. For the figure on the top right, the heat appearing on the left side is explained by the boundary identifications which give rise to the Klein bottle. **Bottom:** Two images of bricks.

in the horizontal direction in K by τ units yielding the sample S^τ. Since S^τ has underlying distribution f^τ then we can also compute $\widehat{K\mathscr{F}}_w(f^\tau, S^\tau)$, which is in essence an estimate of the K-Fourier coefficients corresponding to the rotated image I^τ. By considering the expression

$$d_R(I,J) := \inf_\tau \left\| \widehat{K\mathscr{F}}_w(f^\tau, S^\tau) - \widehat{K\mathscr{F}}_w(g, S') \right\|_2 \qquad (12)$$

we are in fact taking all possible rotated versions of I and J, and computing the smallest distance between their estimated K-Fourier coefficients. In other words, among all possible pairings of rotated versions of I and J, we compute the smallest distortion as measured with estimated K-Fourier coefficients. It follows that this distance is invariant under rotations of I and J. We would like to note that images of rotated textured objects are also affected by the roughness of the material in the form of local shading. For this type of rotation d_R is invariant only up to a certain degree. We use a real data set (the SIPI database) of images of rotated textured objects in order to illustrate the behavior of d_R. Please refer to Figure 17 in Section 4.2. It is also important to note that this distance is not equipped, in principle, to deal with 3d motions and non-rigid deformations.

3.5 Implementing the d_R Distance

The most important point is not that d_R exists, but that it can be efficiently computed for truncated sequences of K-Fourier coefficients: Using the linear transformation T_τ from Theorem 3.15 we get that $T_\tau\left(\widehat{K\mathscr{F}}_w(f,S)\right) = \widehat{K\mathscr{F}}_w(f^\tau, S^\tau)$ and therefore equation 12 can be rewritten as

$$d_R(I,J) = \inf_\tau \left\| T_\tau\left(\widehat{K\mathscr{F}}_w(f,S)\right) - \widehat{K\mathscr{F}}_w(g,S') \right\|_2 \qquad (13)$$

Moreover, since T_τ (thought of as a matrix) is made up of blocks of rotation matrices, then it does not change lengths (it is an isometry), and hence computing $d_R(I,J)$ is equivalent to maximizing the inner product

$$\left\langle T_\tau\left(\widehat{K\mathscr{F}}_w(f,S)\right),\ \widehat{K\mathscr{F}}_w(g,S') \right\rangle \qquad (14)$$

with respect to τ. Using the fact that the entries of T_τ, as a matrix, are sines and cosines of angles of the form $m\tau$, then computing the derivative of equation 14 with respect to τ and making it equal to zero yields the following result:

Theorem 3.16. *There is a nonzero complex polynomial $p(z)$ of degree at most $2w$, with coefficients depending solely on $\widehat{K\mathscr{F}}_w(f,S)$ and $\widehat{K\mathscr{F}}_w(g,S')$, and so that if τ^* is a minimizer for*

$$\Psi(\tau) = \left\| \widehat{K\mathscr{F}}_w(f^\tau, S^\tau) - \widehat{K\mathscr{F}}_w(g,S') \right\|_2$$

then $\xi_* = e^{i\tau^*}$ is a root of $p(z)$.

We describe the construction of this polynomial in the proof of the Theorem, but what is important is that from S, S', w and τ one can compute $d_R(I,J)$ by finding the roots of $p(z)$. These roots can, in principle, be computed as the eigenvalues of the companion matrix of $p(z)$, for which fast algorithms exist. This is how the MATLAB routine `roots` is implemented. One can then look among the arguments of the unitary complex roots of $p(z)$ for the global minimizers (in $[-\pi,\pi]$) of equation 12.

It is known [26] that `roots` produces the exact eigenvalues of a matrix within roundoff error of the companion matrix of p. This, however, does not mean that they are the exact roots of a polynomial with coefficients within roundoff error of those of p, but the discrepancy is well understood. Indeed, let $\widetilde{p}$ be the

monic polynomial having $\texttt{roots}(p)$ as its exact roots, let ε be the machine precision and let $E = [E_{ij}]$ be a perturbation matrix, with E_{ij} a small multiple of ε, and so that if A is the companion matrix of p then $A+E$ is the companion matrix of $\widetilde{p}$. It follows [12] that if

$$p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1 z + a_0$$

then to first order the coefficient of z^{k-1} in $\widetilde{p}(z) - p(z)$ is

$$\sum_{m=0}^{k-1} a_m \sum_{i=k+1}^{n} E_{i,i+m-k} - \sum_{m=k}^{n} a_m \sum_{i=1}^{k} E_{i,i+m-k}.$$

Hence even when $\texttt{roots}(p)$ does not return the exact roots of p, but those of a slight perturbation $\widetilde{p}$, their arguments can be used as the initial guess in an iterative refinement such as Newton's method. This is how we have implemented the d_R distance: We compute the eigenvalues of the companion matrix for p, and then use their arguments (as complex numbers) in a refinement step via Newton's method on equation 14.

3.6 Rotation-Invariant K-Fourier Coefficients

We end this section with a rotation-invariant version of our sequence of estimated K-Fourier coefficients. In a way, it can be thought of as a canonical representation, with respect to image rotation, for estimated K-Fourier coefficients. Let $S \subset K$ be a random sample drawn according to $f : K \longrightarrow \mathbb{R}$, and let $\widehat{K\mathscr{F}}_w(f,S)$ be the estimated K-Fourier coefficients. Recall that this is (Definition 3.14) a truncation of the ordered sequence

$$\widehat{a}_1\,,\ \widehat{a}_2\,,\ \widehat{b}_1\,,\ \widehat{c}_1\,,\ \widehat{d}_{1,1}\,,\ \widehat{e}_{1,1}\,,\ \widehat{a}_3\,,\ \widehat{d}_{1,2}\,,\ \widehat{d}_{2,1}\,,\ \widehat{e}_{1,2}\,,\ \ldots$$

Let $\widehat{\mathbf{v}} = \begin{bmatrix} \widehat{d}_{1,1} \\ \widehat{e}_{1,1} \end{bmatrix}$; we say that $\widehat{K\mathscr{F}}_w(f,S)$ is in **canonical form** if $\widehat{d}_{1,1} > 0$ and $\widehat{e}_{1,1} = 0$. The first point is that for estimated K-Fourier coefficients there is, in general, a value of τ so that $T_\tau\left(\widehat{K\mathscr{F}}_w(f,S)\right)$ is in canonical form. Indeed, since $\|\widehat{\mathbf{v}}\| = 0$ is a zero probability event, then there exists (with probability one) $\sigma \in [0, 2\pi)$ so that

$$\begin{bmatrix} \|\widehat{\mathbf{v}}\| \\ 0 \end{bmatrix} = \begin{bmatrix} \cos(\sigma) & -\sin(\sigma) \\ \sin(\sigma) & \cos(\sigma) \end{bmatrix} \widehat{\mathbf{v}}$$

It follows from Theorem 3.15 (equation 11) that the vector $T_\sigma\left(\widehat{K\mathscr{F}}_w(f,S)\right)$ is in canonical form. Let us change the notation from σ to $\sigma(f)$ to indicate the dependence on f. The second, and perhaps most important point, is that if we consider the estimated K-Fourier coefficients from rotated versions of the same image, then they all have the same canonical form. Indeed,

Proposition 3.17. *Let $\sigma(f)$ be as in the previous paragraph. Then $\sigma(f^\tau) \equiv \sigma(f) - \tau \pmod{2\pi}$, and therefore*

$$T_{\sigma(f^\tau)}\left(\widehat{K\mathscr{F}}(f^\tau, S^\tau)\right) = T_{\sigma(f)}\left(\widehat{K\mathscr{F}}(f,S)\right)$$

for every τ.

This proposition implies that the following makes sense.

Definition 3.18. Let $\widehat{\mathbf{v}} = \begin{bmatrix} \widehat{d}_{1,1} \\ \widehat{e}_{1,1} \end{bmatrix}$ and let $\sigma(f) \in [0, 2\pi)$ be so that $T_{\sigma(f)}\left(\widehat{K\mathscr{F}}(f,S)\right)$ is in canonical form. We let this vector be the estimated sequence of **rotation-invariant** K-Fourier coefficients, for a sample $S \subset K$ drawn according to a distribution $f \in L^2(K, \mathbb{R})$.

That is, it is well defined to regard the canonical form as a rotation-invariant version of the estimated K-Fourier coefficients.

4 Results

The purpose of this section is to apply the theory we have developed so far, to the problem of classifying materials from texture images photographed under various poses and illuminations. We begin by describing the data sets which we will use: The CUReT, KTH-TIPS and UIUCtex databases for classification, and the SIPI data set to illustrate the rotation invariant properties of our descriptor.

Next we give a detailed description of the numerical implementation, as well as the results on the SIPI data set. The last two parts deal with the metric learning approach we use for classification, and a summary (as well as comparisons with state-of-the-art methods) of our results.

4.1 The Data Sets

SIPI[5] The SIPI rotated textures data set, maintained by the Signal and Image Processing Institute at the University of Southern California, consists of 13 of the Brodatz images [5], digitized at different rotation angles: 0, 30, 60, 90, 120, 150 and 200 degrees. The 91 images are 512×512 pixels, 8 bits/pixel, in TIFF format.

[5] `http://sipi.usc.edu/database`

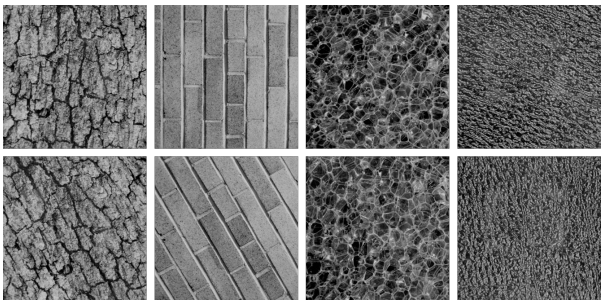

Figure 11: Exemplars from the SIPI Rotated Texture data set

CUReT[6] The Columbia-Utrecht Reflectance and Texture database [9], is one of the most popular image collections used in the performance analysis of texture classification methods. It features images of 61 different materials, each photographed under 205 distinct combinations of viewing and illumination conditions, providing a considerable range of 3D rigid motions. Following the usual setup in the literature, the extreme viewing conditions are discarded, leaving 92 gray scale images per material, of size 200×200 pixels, 8 bits/pixel, in PNG format. What makes this a challenging data set, is both the considerable variability within texture class, and the many commonalities across materials. One drawback, in terms of scope, is its lack of variation in scale.

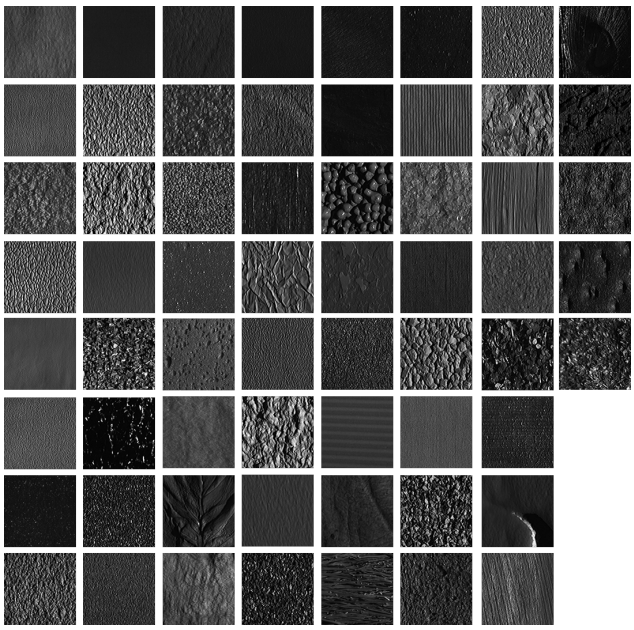

Figure 12: The 61 materials in the CUReT data set

[6] http://www.cs.columbia.edu/CAVE/software/curet

KTH-TIPS[7] The KTH database of Textures under varying Illumination Pose and Scale, collected by Mario Fritz under the supervision of Eric Hayman and Barbara Caputo, was introduced in [18] as a natural extension to the CUReT database. The reason for creating this new data set was to provide wide variation in scale, in addition to changes in pose and illumination. The database consists of 10 materials (already present in the CUReT database), each captured at 9 different scales and 9 distinct poses. That is, the KTH-TIPS data set has 810 images, 200×200 pixels, 8 bits/pixel in PNG format.

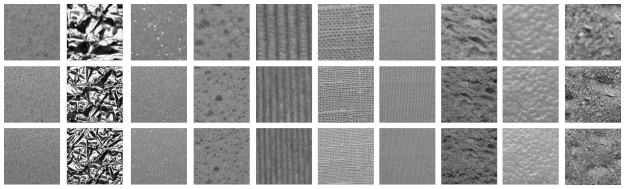

Figure 13: Three scales of each of the ten materials in the KTH-TIPS data set.

UIUCTex[8] The UIUC Texture database, first introduced in [23], and collected by the Ponce Research Group at the University of Illinois at Urbana-Champaign, features 25 texture classes with 40 samples each. The challenge of this particular data set lies in its wide range of viewpoints, scales, and non-rigid deformations. All images are in gray scale, JPG format, 640×480 pixels, 8bits/pixel.

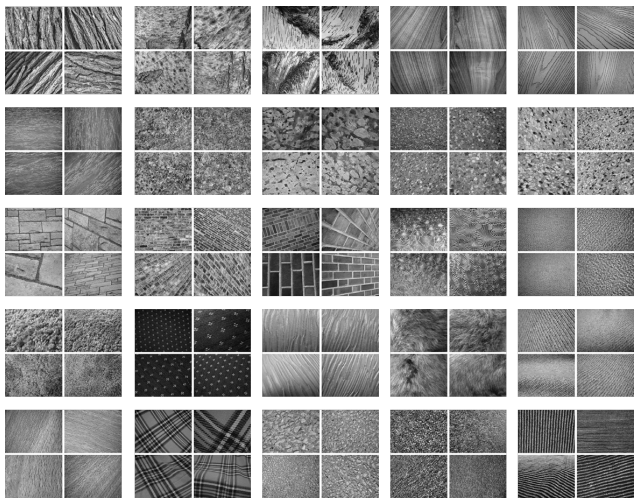

Figure 14: Four exemplars from each of the twenty five classes in the UIUC Texture data set.

[7] http://www.nada.kth.se/cvap/databases/kth-tips

[8] http://www-cvr.ai.uiuc.edu/ponce_grp/data

4.2 Implementation Details

Using the mathematical framework developed in section 3, we now give a detailed description of the steps involved in computing the estimated K-Fourier coefficients of a particular image, as well as the form of the EKFC descriptor.

Preprocessing Given a digital image in gray scale and an odd integer n, we extract its $n \times n$ patches. The choice of n odd makes the implementation simpler, but is somewhat arbitrary and has no bearing on the results presented here. If the image has pixel depth d, then each patch is an $n \times n$ integer matrix with entries between 0 and $2^d - 1$. We add 1 to each entry in the matrix, and following Weber's law take entry-wise natural logarithms. Each log-matrix is then centered by subtracting its mean (as a vector of length n^2) from each component, and D-normalized provided its D-norm is greater than 0.01. Otherwise the patch is discarded for lack of contrast.

The Projection Let P be an $n \times n$ patch represented by the centered and normalized log-matrix $[p_{ij}]$. As described in the discussion following Proposition 3.3, we let ∇I_P be piecewise constant and equal to the discrete gradient of P, which we now define. If $2 \leq i, j \leq n-1$ then we let

$$\nabla P(i,j) = \frac{1}{2}\begin{bmatrix} p_{i,j+1} - p_{i,j-1} \\ p_{i-1,j} - p_{i+1,j} \end{bmatrix}$$

$$HP(i,j) = \begin{bmatrix} p_{i,j+1} - 2p_{i,j} + p_{i,j-1} & H_{xy}P(i,j) \\ H_{xy}P(i,j) & p_{i+1,j} - 2p_{i,j} + p_{i-1,j} \end{bmatrix}$$

where

$$H_{xy}P(i,j) = \frac{p_{i-1,j+1} - p_{i-1,j-1} + p_{i+1,i-1} - p_{i+1,j+1}}{4}.$$

If either $r \in \{1,n\}$ or $t \in \{1,n\}$, then there exists a unique $(i,j) \in \{2,\ldots,n-1\}^2$ which minimizes $|i-r|+|j-t|$, and we let

$$\nabla P(r,t) = \nabla P(i,j) + HP(i,j)\begin{bmatrix} r-i \\ j-t \end{bmatrix}.$$

That is, we approximate the gradient of P at a location (r,t) near (i,j) using the first order Taylor expansion. As for the gradient of the extension I_P, we let $\nabla I_P(x,y) = \nabla P(i,j)$ if

$$\left|x - \left(-1 + \frac{2j-1}{n}\right)\right| + \left|y - \left(1 - \frac{2i-1}{n}\right)\right| < \frac{1}{n}$$

for some $(i,j) \in \{1,\ldots,n\}^2$, and $\mathbf{0}$ otherwise. This definition captures directionality more accurately than simply calculating finite differences. This can be seen, for instance, in a 3×3 patch with extension $I_P(x,y) = \frac{\sqrt{3}(x+y)^2}{8}$.

If the eigenvalues of the associated matrix A_P are distinct (see Proposition 3.1), then we let $\alpha_P \in [\pi/4, 5\pi/4)$ be the direction of the eigenspace corresponding to the largest eigenvalue. Since A_P is a 2×2 matrix, α_P can be computed explicitly in terms of ∇P. If A_P has only one eigenvalue, then the patch P is discarded due to its lack of a prominent directional pattern.

If $\Phi(c,d)$ is constant (see Proposition 3.3) then P is discarded, since it does not have linear nor quadratic components. Otherwise we let $\theta_P \in [-\pi/2, 3\pi/2)$ correspond to the minimizer (c^*,d^*). Next we compute $\Phi(c^*,d^*)$, which can be thought of as the the distance from P to K. Notice that the triangular inequality with respect to $\|\cdot\|_D$ implies that $\Phi(c^*,d^*) \leq 2$. We include (α_P, θ_P) in the sample $S \subset K$ if $\Phi(c^*,d^*) \leq r_n$, for r_n as reported in Table 1.

Table 1: Maximum distance to Klein bottle

Patch size (n)	3	5	7	9	$n \geq 11$
r_n	1.2247	1.3693	1.4031	1.4114	1.4135

The rationale behind this choice is as follows:

$$\left\| I_P - \left(c^* u + d^* \sqrt{3} u^2\right) \right\|_D = \sqrt{2\left(1 - \varphi(I_P, \alpha_P)\right)}$$

and the values of r_n are set so that $\varphi(I_P, \alpha_P) \geq \frac{1}{2^{n-1}}$ is the inclusion criterion. Here $\sqrt{\varphi(I_P, \alpha_P)}$ can be thought of as the size of the contribution of $Span\{(ax+by), (ax+by)^2\}$ if one were to write a series expansion for $I_P(x,y)$ in terms of the (rotated) monomials $(ax+by)^k(ay-bx)^m$, $k,m \in \mathbb{N}$.

The K-Fourier Coefficients We select the cut-off frequency $w \in \mathbb{N}$ so that $\widehat{K\mathscr{F}}_w(f,S)$ includes the coefficients with large variance, and excludes the first regime where the $\widehat{f}_k$ tend to zero. Recall that after this regime the $\widehat{f}_k$ oscillate with large variance since $S_J(\widehat{f}_\delta) \to f_\delta$. For the results presented here we let $w = 6$, and fix it for once and for all. While similar cut-off frequencies have been obtained with maximum likelihood methods, see for instance [11], other choices might improve classification results. It follows that $\widehat{K\mathscr{F}}_w(f,S)$ is a vector of length 42. We illustrate the computation of the estimated K-Fourier coefficients. Let us consider the K-sample from 3×3 patches in the image depicted in Figure 15.

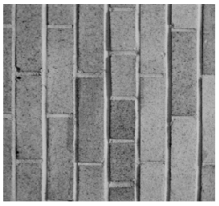

Figure 15: Bricks, included in the **SIPI** rotated texture data set, was extracted from page 94 of [5] and digitized at 0 degrees.

We compute the estimated K-Fourier coefficients and the corresponding estimate $\widehat{f}$ for the underlying probability density function $f : K \longrightarrow \mathbb{R}$. Here

$$\widehat{f}(\alpha,\beta) = \sum_{k \in N_w} \widehat{f}_k \cdot \phi_k(\alpha,\beta)$$

where $\{\phi_k\}_{k\in\mathbb{N}}$ is the trigonometric basis for $L^2(K,\mathbb{R})$, $\widehat{f}_k$ is as in Definition 3.14, and $N_w \subset \mathbb{N}$ is so that ϕ_k has frequency less than or equal to $w = 6$ whenever $k \in N_w$. We summarize the results in Figure 16.

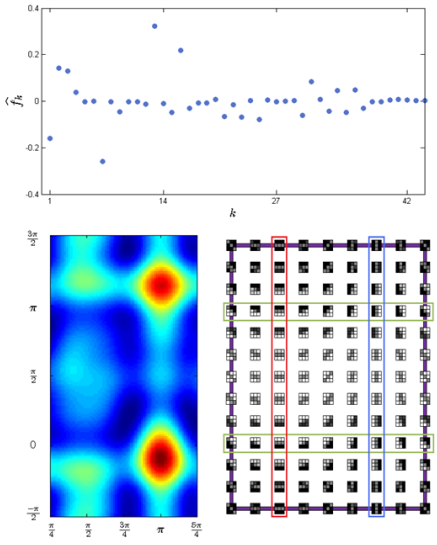

Figure 16: Estimated K-Fourier coefficients for the texture of Bricks. **Top**: First 44 estimated K-Fourier coefficients. **Bottom left**: Heat-map representation for the estimate $\widehat{f}$. $(\alpha,\beta) \in K$ is colored according to high density (red) or low density (blue). **Bottom right**: A lattice of patches in the Klein bottle model $\mathscr{K}$. It follows that the hotter spots in the heat-map representation of $\widehat{f}$ correspond to vertical patterns.

Rotation Invariance We use the **SIPI** database to illustrate the rotation invariance properties, as presented in section 3.4, of estimated K-Fourier coefficients. Our goal is to show that in this database rotated versions of the same image are clustered if we measure distance with the d_R-metric (see equation 12). This data set is specially relevant for this task, since the type of rotation it describes, i.e. planar, is exactly the type we have modeled.

First we compute the estimated K-Fourier coefficients of the distribution of projected 3×3 patches for each one of the images in the **SIPI** data set. Next, for each pair of images I, J we calculate the d_R distance

$$d_R(I,J) := \min_{\tau} \left\| \widehat{K\mathscr{F}}_w(f,S) - T_\tau\left(\widehat{K\mathscr{F}}_w(g,S')\right) \right\|_2$$

between them as described in the discussion following Theorem 3.16. In order to see that rotated versions of the same image are clustered when distance is measured as above, we use classical Multi-Dimensional Scaling (MDS). The goal of MDS is, given points $\mathbf{x}_1,\ldots,\mathbf{x}_N$ in $\mathbb{R}^D$, a measure of distance $d(i,j)$ between each pair $\mathbf{x}_i$ and $\mathbf{x}_j$, and an integer $n \leq D$, to find a collection of points $\mathbf{y}_1,\ldots,\mathbf{y}_N$ in $\mathbb{R}^n$ so that if we measure distance with Euclidean norm then $\|\mathbf{y}_i - \mathbf{y}_j\|$ is as close as possible to $d(i,j)$, for all $i,j = 1,\ldots,N$.

We applying MDS to the **SIPI** data set using the distance d_R and $n = 3$. The result is summarized in Figure 17.

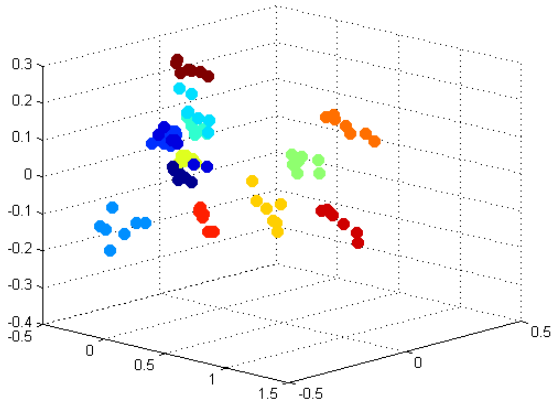

Figure 17: Multi-Dimensional Scaling for d_R-distance matrix on the **SIPI** data set. Each dot represents an image, and rotated versions of the same texture share the same color. For colors, please refer to an electronic version of the paper.

As one can see, rotated versions of the same image are clustered when distance is measured with d_R. Notice that no learning step was necessary in this experiment.

The EKFC Descriptor Let $M+N$ be the total number of $n \times n$ patches in a given image, with M being the number of patches which are too far from the Klein bottle to be projected, and let S be the resulting K-sample from its $n \times n$ patches. We let the EKFC (estimated K-Fourier coefficients) descriptor at scale n be

$$EKFC_n = \left[1 - \frac{1}{M}, \frac{T_{\sigma(f)}\left(\widehat{K\mathscr{F}}_w(f,S)\right)}{M} \right]$$

where $T_{\sigma(f)}\left(\widehat{K\mathscr{F}}_w(f,S)\right)$ is the truncated sequence of estimated rotation-invariant K-Fourier coefficients. Please refer to Definition 3.18.

We obtain the (**multi-scale**) EKFC descriptor by concatenating the $EKFC_n$'s for several values of n. For the results presented here we used $n = 3, 7, 11, 15, 19$ and

$$EKFC = [EKFC_3,\ EKFC_7,\ EKFC_{11},\ EKFC_{15},\ EKFC_{19}]$$

a vector of length 215.

4.3 LMNN for Metric Learning

One of the most basic example-based classification techniques is the k-th nearest neighbor method, in which a new instance is classified according to the most common label among its k nearest neighbors. Needless to say, this method depends largely on the notion of distance used to compare data points. For the special case of distances between probability density functions, many dissimilarity measures exist: The χ^2-statistic between histograms, the earth mover's distance, L^p norms, etc. While distances defined through various heuristics are useful, metrics learnt from the data often outperform them in classification tasks.

One of the most popular metric learning algorithms is the Large Margin Nearest Neighbor (LMNN) introduced by [37]. In this algorithm, where data points are represented as vectors $\mathbf{x} \in \mathbb{R}^D$, the goal is to learn a linear transformation

$$L : \mathbb{R}^D \longrightarrow \mathbb{R}^D$$

which rearranges the training set so that, as much as possible, the k nearest neighbors of each data point have the same label, while vectors labeled differently are separated by a large margin. If $\mathbf{M} = L^T L$ then $d_{\mathbf{M}}(\mathbf{x},\mathbf{y}) = \|\mathbf{M}(\mathbf{x}-\mathbf{y})\|_2$ is referred to as the learnt Mahalanobis distance.

4.4 Classification Results

We evaluate the performance of the EKFC descriptor by classifying the images in the CUReT, KTH-TIPS and UIUCTex databases according to their material class. For each data set, the evaluation consists of a metric learning step on a randomly selected portion of the data, followed by a labeling of the remaining images. We report the mean classification success rate and standard deviation on each database, from 100 random splits training/test set.

Metric Learning For each texture class, we randomly select a fixed number of exemplars which we label, and aggregate to form a training set. We then learn a Mahalanobis metric from these labeled examples using the LMNN algorithm[9]. For the results presented here, we used 3 nearest neighbors to learn a global metric, as well as 20% of the training set for cross validation. This, to prevent the algorithm from going into over-fitting.

Labeling Step We use the energy-based classification as described in [37], which has been observed to produce better results than k-th nearest neighbor labeling. In this framework, a test example is assigned the label which minimizes the same loss function that determines the Mahalanobis metric $d_{\mathbf{M}}$, if the example along with the proposed label were to be added to the training set.

Classification Once the labels have been determined via the energy-based method, we compute the percentage of test images which have been labeled correctly. The mean and variance of these percentages is computed over 100 random splits training/test set. We summarize our classification results on Table 2.

5 Discussion and Final Remarks

We have presented in this paper a novel framework for estimating and representing the distribution, around low dimensional submanifolds of pixel-space, of patches from texture images. In particular, we have shown that most $n \times n$ patches from an image can be continuously projected onto a surface $\mathscr{K} \subset \mathbb{R}^{n^2}$ with the topology of a Klein bottle, and that the resulting set can be interpreted as sampled from a probability density function $f : \mathscr{K} \longrightarrow \mathbb{R}$. The K-Fourier coefficients of f can then be estimated from the sample, via unbiased estimators which converge almost surely as the sample size goes to infinity. The choice of the model $\mathscr{K}$, comes from the fact that small high-contrast patches from natural images accumulate around it with high density [7]. In addition, we show that the estimated K-Fourier coefficients of a rotated image can be recovered from

[9] `http://www.cse.wustl.edu/~kilian/code/code.html`

Table 2: Classification scores on the CUReT, UIUCTex and KTH-TIPS data sets. We include, for comparison, the results (as reported in Table 2 of [8]) from state of state-of-the-art methods for material classification from texture images.

	CUReT 43 images per class in training set	**UIUCTex** 20 images per class in training set	**KTH-TIPS** 40 images per class in training set
our EKFC descriptor + LMNN learnt metric	$95.66 \pm 0.45\%$	$91.23 \pm 1.13\%$	$94.77 \pm 1.3\%$
Crosier & Griffin [8]	$98.6 \pm 0.2\%$	$98.8 \pm 0.5\%$	$98.5 \pm 0.7\%$
Varma & Zisserman - MR8 [34]	97.43%		
Varma & Zisserman - Joint [35]	98.03%	$97.83 \pm 0.66\%$	$92.4 \pm 2.1\%$
Hayman et al. [18]	$98.46 \pm 0.09\%$	$92.0 \pm 1.3\%$	$94.8 \pm 1.2\%$
Lazebnik et al. [23]	$72.5 \pm 0.7\%$	96.03%	$91.3 \pm 1.4\%$
Zhang et al. [38]	$95.3 \pm 0.4\%$	$98.70 \pm 0.4\%$	$95.5 \pm 1.3\%$
Broadhurst [4]	$99.22 \pm 0.34\%$		
Varma and Ray [32]		$98.9 \pm 0.68\%$	

those of the original one, via a linear transformation which depends solely on the rotation angle.

Concatenating the first 43 rotation-invariant K-Fourier coefficients from $n \times n$ patches at 5 different scales ($n = 3, 7, 11, 15, 19$), we obtain the multi-scale rotation-invariant EKFC-descriptor. Note that its dimension, 215, is considerable lower than that of popular histogram representations (2,440 [33] and 1,296 [8]), while still achieving high classification rates on the CUReT ($95.66 \pm 0.45\%$), UIUCTex (91.23 ± 1.13) and KTH-TIPS ($94.77 \pm 1.3\%$) data sets.

As we have seen, the Klein bottle dictionary represents patches in terms of bars and edges, which are not expected to be enough to efficiently capture all primitives in a texture image. It is thus encouraging that we have obtained such high classification results, though not better than the state-of-the-art, across challenging texture data sets. It is important to note that the UIUC-Tex data (in which we obtained the lowest performance) is different from the CUReT and KTH-TIPS collections in that in addition to 3d rotations it also includes non-rigid deformations. For 3d rigid motions one can use the learning step as a copping strategy if the invariant, as is our case, only accounts for planar rotations. For non-rigid deformations, however, one needs richer features (e.g. blobs) and include deformation invariance in the representation. This takes us back to the point of enlarging the Klein bottle model.

The success of state-of-the-art methods stems from having highly descriptive feature categories at several scales, with features ranging from those prevalent across textures (e.g. edges and bars), to the ones which are more class-characteristic. Our framework provides a way of concisely representing the ubiquitous part using the low-dimensional manifold in patch-space which best describes it, and what we show in this paper, is that even with this limited vocabulary it is possible to achieve high classification rates. We do not claim that the Klein bottle manifold is the right space for studying distributions of patches from texture images, but rather that it is a good initial building block (as our results suggest) for a low-dimensional space (not necessarily a manifold) representing a richer vocabulary of features.

A fundamental question going forward, is how to enlarge the Klein bottle model in a way which is motivated by the distribution of patches from texture images, so that it includes other important image primitives while keeping a low intrinsic dimensionality, and so that the framework of estimated coefficients can be applied. One way of doing so is by following a CW-complexes philosophy [17], in which a space is built in stages by pasting disks of increasing dimension along their boundaries. We hope to explore this avenue in future work.

Acknowledgements. Jose Perea was partially supported by the National Science Foundation (NSF) through grant DMS 0905823. Gunnar Carlsson was supported by the NSF through grants DMS 0905823 and DMS 096422, by the Air Force Office of Scientific Research through grants FA9550-09-1-0643 and FA9550-09-1-0531, and by the National Institutes of Health through grant I-U54-ca149145-01.

References

[1] Aherne, F.J.; Thacker, N.A., and Rockett, P.I. The bhattacharyya metric as an absolute similarity measure for fre-

quency coded data. *Kybernetika*, 34(4):363–368, 1998.

[2] Bell, A.J. and Sejnowski, T.J. The "independent components" of natural scenes are edge filters. *Vision research*, 37(23):3327, 1997.

[3] Beyer, K.; Goldstein, J.; Ramakrishnan, R., and Shaft, U. When is "nearest neighbor" meaningful? *Database Theory-ICDT'99*, pages 217–235, 1999.

[4] Broadhurst, R.E. Statistical estimation of histogram variation for texture classification. In *Proc. Intl. Workshop on Texture Analysis and Synthesis*, pages 25–30, 2005.

[5] Brodatz, P. *Textures: a photographic album for artists and designers*, volume 66. Dover New York, 1966.

[6] Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255, 2009.

[7] Carlsson, G.; Ishkhanov, T.; De Silva, V., and Zomorodian, A. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1): 1–12, 2008.

[8] Crosier, M. and Griffin, L.D. Using basic image features for texture classification. *International journal of computer vision*, 88(3):447–460, 2010.

[9] Dana, K.J.; Van Ginneken, B.; Nayar, S.K., and Koenderink, J.J. Reflectance and texture of real-world surfaces. *ACM Transactions on Graphics (TOG)*, 18(1):1–34, 1999.

[10] De Silva, V.; Morozov, D., and Vejdemo-Johansson, M. Persistent cohomology and circular coordinates. *Discrete & Computational Geometry*, 45(4):737–759, 2011.

[11] De Wit, T.D. and Floriani, E. Estimating probability densities from short samples: a parametric maximum likelihood approach. *Physical Review E*, 58(4):5115, 1998.

[12] Edelman, A. and Murakami, H. Polynomial roots from companion matrix eigenvalues. *Mathematics of Computation*, 64(210):763–776, 1995.

[13] Franzoni, G. The klein bottle: Variations on a theme. *Notices of the AMS*, 59(8), 2012.

[14] Griffin, Lewis D. Feature classes for 1d, 2nd order image structure arise from natural image maximum likelihood statistics. *Network: Computation in Neural Systems*, 16(2-3):301–320, 2005.

[15] Griffin, Lewis D. The second order local-image-structure solid. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(8):1355–1366, 2007.

[16] Harris, Chris and Stephens, Mike. A combined corner and edge detector. 1988.

[17] Hatcher, A. *Algebraic topology*. Cambridge UP, Cambridge, 2002.

[18] Hayman, E.; Caputo, B.; Fritz, M., and Eklundh, J.O. On the significance of real-world conditions for material classification. *Computer Vision-ECCV 2004*, pages 253–266, 2004.

[19] Hubel, D.H. and Wiesel, T.N. Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3):574–591, 1959.

[20] Hubel, D.H. and Wiesel, T.N. Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, 195(1):215–243, 1968.

[21] Jurie, F. and Triggs, B. Creating efficient codebooks for visual recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 604–610. IEEE, 2005.

[22] Koenderink, Jan J. The structure of images. *Biological cybernetics*, 50(5):363–370, 1984.

[23] Lazebnik, S.; Schmid, C., and Ponce, J. A sparse texture representation using local affine regions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8): 1265–1278, 2005.

[24] Lee, A.B.; Pedersen, K.S., and Mumford, D. The nonlinear statistics of high-contrast patches in natural images. *International Journal of Computer Vision*, 54(1):83–103, 2003.

[25] Leung, T. and Malik, J. Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision*, 43(1): 29–44, 2001.

[26] Moler, C. Cleve's corner: Roots-of polynomials, that is. *The MathWorks Newsletter*, 5(1):8–9, 1991.

[27] Pedersen, Kim S and Lee, Ann B. Toward a full probability model of edges in natural images. In *Computer Vision-ECCV 2002*, pages 328–342. Springer, 2002.

[28] Reed, M. and Simon, B. *Methods of Modern Mathematical Physics: Vol.: 1.: Functional Analysis*. Academic press, 1972.

[29] Rubner, Y.; Tomasi, C., and Guibas, L.J. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000.

[30] Silverman, B.W. *Density estimation for statistics and data analysis*, volume 26. Chapman & Hall/CRC, 1986.

[31] van Hateren, J.H. and van der Schaaf, A. Independent component filters of natural images compared with simple cells in primary visual cortex. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 265(1394):359–366, 1998.

[32] Varma, M. and Ray, D. Learning the discriminative power-invariance trade-off. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[33] Varma, M. and Zisserman, A. Unifying statistical texture classification frameworks. *Image and Vision Computing*, 22(14):1175–1183, 2004.

[34] Varma, M. and Zisserman, A. A statistical approach to texture classification from single images. *International Journal of Computer Vision*, 62(1):61–81, 2005.

[35] Varma, M. and Zisserman, A. A statistical approach to material classification using image patch exemplars. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(11):2032–2047, 2009.

[36] Watson, G.S. Density estimation by orthogonal series. *Ann. Math. Statist*, 40(4):1496–1498, 1969.

[37] Weinberger, K.Q. and Saul, L.K. Distance metric learning for large margin nearest neighbor classification. *The Journal of Machine Learning Research*, 10:207–244, 2009.

[38] Zhang, J.; Marszalek, M.; Lazebnik, S., and Schmid, C. Local features and kernels for classification of texture and object categories: A comprehensive study. *International Journal of Computer Vision*, 73(2):213–238, 2007.

Appendix A. Main Results and Proofs

Proposition 3.1.

1. The Cauchy-Schwartz inequality implies that for every $\mathbf{v} \in S^1$ and every almost-everywhere differentiable I_P (not necessarily purely directional) one has that

$$Q_P(\mathbf{v}) \leq \iint_{[-1,1]^2} \|\nabla I_P\|^2 dxdy.$$

Since the equality holds when $I_P(x,y) = g(ax+by)$ and $\mathbf{v} = \begin{bmatrix} a \\ b \end{bmatrix}$, the result follows.

2. Let $\lambda_{max} \geq \lambda_{min} \geq 0$ be the eigenvalues of A_P, and let B_P be a unitary matrix such that

$$A_P B_P = \begin{bmatrix} \lambda_{max} & 0 \\ 0 & \lambda_{min} \end{bmatrix}.$$

If $\mathbf{v} \in S^1$ and $\mathbf{v} = B_P \begin{bmatrix} v_x \\ v_y \end{bmatrix}$, then $1 = \|\mathbf{v}\|^2 = v_x^2 + v_y^2$, $Q_P(\mathbf{v}) = \lambda_{max} v_x^2 + \lambda_{min} v_y^2$ and therefore

$$\max_{\|\mathbf{v}\|=1} Q_P(\mathbf{v}) = \lambda_{max}.$$

Finally, since $Q_P(\mathbf{v}) = \lambda_{max}$ for $\mathbf{v} \in S^1$, if and only if $\mathbf{v} \in E_{max}(A_P)$, we obtain the result.

3. If the eigenvalues of A_P are distinct, then $E_{max}(A_P)$ is one dimensional and thus intercepts S^1 at exactly two antipodal points.

□

Proposition 3.3. Since u and $\sqrt{3}u^2$ are orthonormal with respect to $\langle \cdot, \cdot \rangle_D$, then it follows that

$$\begin{aligned} \operatorname*{argmin}_{c^2+d^2=1} \Phi(c,d) &= \operatorname*{argmin}_{c^2+d^2=1} \Phi(c,d)^2 \\ &= \operatorname*{argmax}_{c^2+d^2=1} \left\langle I_P, cu + d\sqrt{3}u^2 \right\rangle_D \\ &= \operatorname*{argmax}_{c^2+d^2=1} \left\langle \begin{bmatrix} c \\ d \end{bmatrix}, \begin{bmatrix} \langle I_P, u \rangle_D \\ \langle I_P, \sqrt{3}u^2 \rangle_D \end{bmatrix} \right\rangle \end{aligned}$$

which can be found via the usual condition of parallelism, provided $\varphi(I_P, \alpha) \neq 0$. □

Proposition 3.4. Let $g \in L^2(T)$ and consider

$$g^{\perp}(z,w) = \frac{g(z,w) + g(-z,-\overline{w})}{2}$$

It follows that $g^{\perp}$ is square-integrable on T, and that:

1. $g^{\perp}(-z,-\overline{w}) = g^{\perp}(z,w)$ for every $(z,w) \in T$. Hence $g^{\perp} \in L^2(K)$ for every $g \in L^2(T)$.
2. If $g_1, g_2 \in L^2(T)$ and $a_1, a_2 \in \mathbb{C}$ then
$$(a_1 g_1 + a_2 g_2)^{\perp} = a_1(g_1^{\perp}) + a_2(g_2)^{\perp}$$
3. $g^{\perp} = g$ for every $g \in L^2(K)$.

We claim that $g^{\perp}$ is the orthogonal projection of $g \in L^2(T)$ onto $L^2(K)$, and all we have to check is that $g - g^{\perp}$ is perpendicular to every $h \in L^2(K)$. To this end, let us introduce the notation

$$g^*(z,w) = g(-z,-\overline{w}).$$

By writing the inner product of $L^2(T)$ in polar coordinates, using the substitution $(\alpha,\theta) \mapsto (\alpha+\pi, \pi-\theta)$, and the fact that h satisfies equation 5, one obtains that

$$\begin{aligned}
\langle g^*, h\rangle_T &= \frac{1}{(2\pi)^2}\int_{-\frac{\pi}{2}}^{\frac{3\pi}{2}}\int_{\frac{\pi}{4}}^{\frac{9\pi}{4}} g(\alpha+\pi,\pi-\theta)\overline{h}(\alpha,\theta)\,d\alpha d\theta \\
&= \frac{1}{(2\pi)^2}\int_{-\frac{\pi}{2}}^{\frac{3\pi}{2}}\int_{\frac{\pi}{4}}^{\frac{9\pi}{4}} g(\alpha,\theta)\overline{h}(\alpha+\pi,\pi-\theta)\,d\alpha d\theta \\
&= \frac{1}{(2\pi)^2}\int_{-\frac{\pi}{2}}^{\frac{3\pi}{2}}\int_{\frac{\pi}{4}}^{\frac{9\pi}{4}} g(\alpha,\theta)\overline{h}(\alpha,\pi)\,d\alpha d\theta \\
&= \langle g, h\rangle_T
\end{aligned}$$

Therefore $2\langle g - g^{\perp}, h\rangle_T = \langle g, h\rangle_T - \langle g^*, h\rangle_T = 0$ and we get the result. □

Theorem 3.5. By continuity, Π takes spanning sets to spanning sets which is the first part of the theorem. Now, if we consider the decomposition

$$L^2(T) = L^2(K) \oplus L^2(K)^{\perp}$$

where $L^2(K)^{\perp}$ denotes the orthogonal linear complement of $L^2(K)$ in $L^2(T)$, and $\mathscr{B}$ is an orthonormal basis for $L^2(K)$, then for any orthonormal basis $\mathscr{B}'$ of $L^2(K)^{\perp}$ we have that $\mathscr{B} \cup \mathscr{B}'$ is an orthonormal basis for $L^2(T)$. The thing to notice is that since $\ker(\Pi) = L^2(K)^{\perp}$ and Π restricted to $L^2(K)$ is the identity, then $\Pi(\mathscr{B}') = \{\mathbf{0}\}$ and therefore $\Pi(\mathscr{B} \cup \mathscr{B}') = \mathscr{B} \cup \{\mathbf{0}\}$. It follows that the only subset of $\mathscr{B} \cup \{\mathbf{0}\}$ which can be a basis for $L^2(K)$ is $\mathscr{B}$, and that it is invariant when applying Gram-Schmidt. □

Theorem 3.15. If $f, g \in L^2(K, \mathbb{R})$ then from a change of coordinates it follows that $\langle f^{\tau}, g\rangle_K = \langle f, g^{-\tau}\rangle_K$, and therefore

$$\begin{aligned}
a_m^{\tau} &= a_m \\
b_n^{\tau} &= \langle f^{\tau}, \sqrt{2}\cos(2n\alpha)\rangle_K \\
&= \langle f, \sqrt{2}\cos(2n(\alpha+\tau))\rangle_K \\
&= \cos(2n\tau)b_n - \sin(2n\tau)c_n \\
c_n^{\tau} &= \langle f, \sqrt{2}\sin(2n(\alpha+\tau))\rangle_K \\
&= \cos(2n\tau)c_n + \sin(2n\tau)b_n
\end{aligned}$$

Similarly

$$\begin{aligned}
d_{n,m}^{\tau} &= \cos(n\tau)d_{n,m} - \sin(n\tau)e_{n,m} \\
e_{n,m}^{\tau} &= \cos(n\tau)e_{n,m} + \sin(n\tau)d_{n,m}.
\end{aligned}$$

□

Corollary .1. *Let $T_{\tau} : \ell^2(\mathbb{R}) \longrightarrow \ell^2(\mathbb{R})$ be as in Theorem 3.15. Then*

$$T_{\tau+\beta} = T_{\tau} \circ T_{\beta}.$$

Theorem 3.16. From the description of T_{τ} as a block diagonal matrix whose blocks are rotation matrices (Theorem 3.15), it follows that

$$\begin{aligned}
\operatorname*{argmin}_{\tau} \Psi(-\tau) &= \operatorname*{argmax}_{\tau} \langle K\mathscr{F}_w(f), T_{\tau}(K\mathscr{F}_w(g))\rangle \\
&= \operatorname*{argmax}_{\tau} \sum_{n=1}^{w} x_n\cos(n\tau) + y_n\sin(n\tau)
\end{aligned}$$

where x_n and y_n depend solely on $K\mathscr{F}_w(f)$ and $K\mathscr{F}_w(g)$. By taking the derivative with respect to τ, we get that if τ^* is a minimizer for $\Psi(\tau)$ then we must have

$$\sum_{n=1}^{w} n y_n\cos(n\tau^*) - n x_n\sin(n\tau^*) = 0$$

which is equivalent to having

$$Re\left(\sum_{n=1}^{w} n(y_n + i x_n)\xi_*^n\right) = 0.$$

Let $q(z) = \sum_{n=1}^{w} n(y_n + i x_n)z^n$, $\bar{q}(z) = \sum_{n=1}^{w} n(y_n - i x_n)z^n$ and

$$p(z) = z^w \cdot \left(q(z) + \bar{q}\left(\frac{1}{z}\right)\right).$$

It follows that $p(z)$ is a complex polynomial of degree less than or equal to $2w$, and so that

$$\begin{aligned} p(\xi_*) &= \xi_*^w \left(q(\xi_*) + \bar{q}\left(\frac{1}{\xi_*}\right)\right) \\ &= \xi_*^w \left(q(\xi_*) + \bar{q}\left(\overline{\xi_*}\right)\right) \\ &= 2\xi_*^w \cdot Re\left(q(\xi_*)\right) \\ &= 0. \end{aligned}$$

□

Corollary .2. *The vector $T_\tau\left(\widehat{K\mathscr{F}}(f)\right)$ is a componentwise unbiased estimator for $K\mathscr{F}(f^\tau)$, which converges almost surely as the sample size tends to infinity.*

Proposition 3.17. Let $\widehat{\mathbf{v}}^\tau$ be the vector with entries $\widehat{d}^\tau_{1,1}$ and $\widehat{e}^\tau_{1,1}$ for $\widehat{K\mathscr{F}}(f^\tau, S^\tau)$. It follows from Theorem 3.15 and Corollaries .1 and .2 that

$$\begin{aligned} \widehat{\mathbf{v}}^\tau &= \begin{bmatrix} \cos(\tau) & -\sin(\tau) \\ \sin(\tau) & \cos(\tau) \end{bmatrix} \widehat{\mathbf{v}} \\ &= \begin{bmatrix} \cos\left(\tau - \sigma(f)\right) & -\sin\left(\tau - \sigma(f)\right) \\ \sin\left(\tau - \sigma(f)\right) & \cos\left(\tau - \sigma(f)\right) \end{bmatrix} \begin{bmatrix} \|\widehat{\mathbf{v}}\| \\ 0 \end{bmatrix} \\ &= \begin{bmatrix} \cos\left(\tau - \sigma(f)\right) & -\sin\left(\tau - \sigma(f)\right) \\ \sin\left(\tau - \sigma(f)\right) & \cos\left(\tau - \sigma(f)\right) \end{bmatrix} \begin{bmatrix} \|\widehat{\mathbf{v}}^\tau\| \\ 0 \end{bmatrix} \end{aligned}$$

from which we get $\sigma(f^\tau) \equiv \sigma(f) - \tau \pmod{2\pi}$, and

$$\begin{aligned} T_{\sigma(f^\tau)}\left(\widehat{K\mathscr{F}}(f^\tau, S^\tau)\right) &= T_{\sigma(f)-\tau} \circ T_\tau \left(\widehat{K\mathscr{F}}(f, S)\right) \\ &= T_{\sigma(f)}\left(\widehat{K\mathscr{F}}(f, S)\right) \end{aligned}$$

as claimed. □