Learning Manifolds in the Wild

C. Hegde[‡], A. C. Sankaranarayanan[*], R. G. Baraniuk[†]
Contact: chinmay@csail.mit.edu (email) 281-804-5037 (phone)

[‡]*CSAIL, Massachusetts Institute of Technology*
[*]*ECE Dept., Carnegie Mellon University*
[†]*ECE Dept., Rice University*

Abstract

Despite the promise of low-dimensional manifold models for image processing, computer vision, and machine learning tasks, their utility has been hamstrung in practice by two fundamental challenges. First, practical image manifolds are *non-isometric* to their underlying parameter space, while the state-of-the-art manifold modeling and learning frameworks assume isometry. Second, practical image manifolds are strongly perturbed by *nuisance parameters* such as illumination variations, occlusions, and clutter.

In this paper, we develop new theory and practical algorithms for manifold modeling, learning, and processing that address these challenges. To address the isometry challenge, we show that the Earth Movers Distance (EMD) is a more natural metric for inter-image distances than the standard Euclidean distance, and use it to establish the isometry of manifolds generated by translations and rotations of a reference image. To the best of our knowledge, this is the first rigorous result on establishing manifold isometry for grayscale image families. To address the nuisance parameter challenge, we advocate an image representation based on local keypoint features, such as SIFT features, and use it to define a new *keypoint articulation manifold* (KAM). We introduce computationally efficient methods to perform manifold learning of the KAM and demonstrate their robustness.

We employ the KAM framework on a number of real-world image datasets acquired in the wild. As a particular application, we describe the utility of the KAM framework in the automatic organization of large-scale, unstructured collections of photographs gathered from the internet. Our intention is to demonstrate that manifold methods are not just elegant from a mathematical modeling perspective, but can also be of considerable utility in real applications.

1. Introduction

A host of problems in computer vision, machine learning, and pattern recognition involve efficient analysis, modeling, and processing of signal and image *ensembles*. Effective solutions to many such problems require exploiting the *geometric relationships* among the data in the ensemble of nterest. In classical signal processing and statistics, for example, the data form linear subspaces of the ambient space, which leads to simple, linear processing algorithms.

One important class of image ensembles arises in situations where there exists a parameter vector θ that controls the *appearance* of the objects within each image I_θ. Examples include: translation, specifying the location of an object in a scene; orientation, specifying its pose; or illumination, specifying the 3D location of the light source (or sources) present in a scene. Instead of the more prosaic linear subspaces, such image families form low-dimensional nonlinear manifolds in the high-dimensional ambient space. Under certain conditions, such a family forms an *image articulation manifold* (IAM). The dimension K of an IAM equals the number of free parameters in the articulation parameter θ. For example, the image translation manifold is two dimensional (2D), corresponding to horizontal and vertical translations) and can be interpreted very roughly as a two-dimensional "surface" in the high-dimensional ambient space $\mathbb{R}^N$. Hence an IAM can be a very concise model for the images it comprises.

Manifold-based models have long been used for applications involving data ensembles that can be described by only a few degrees of freedom. The promise of such models lies in their ability to potentially break the so-called "curse of dimensionality", a common problem in most practical machine learning tasks. Consequently, the last decade has witnessed great theoretical and algorithmic advances in this regard, and manifold models have been successfully applied to tasks such as data visualization, parameter estimation, transductive learning, and compact data representations [1, 2, 3, 4].

However, the significant theoretical advances in manifold-based image processing have not led to commensurate success in practice. The reasons for this stem from two fundamental challenges:

1. **Lack of isometry:** A common IAM desideratum is that the underlying manifold is *locally isometric* to the underlying parameter space, i.e., small changes in the articulation parameter θ generate images that are "nearby" in terms of Euclidean distance. Unfortunately, this assumption breaks down for anything except the simplest of IAMs. Donoho and Grimes [5] have

Figure 1: *An image ensemble gathered from the wild. Example images of the Notre Dame Cathedral gathered from FlickR [6]. Such a real-world image ensemble cannot be easily modeled via a strictly low-dimensional parametric representation; occlusions are significant, illlumination variations are dramatic, and imaging artifacts such as varying field-of-view, skew, and white balance abound. As a consequence, conventional manifold learning methods fail when applied to such ensembles.*

shown that for anything more complicated than a simple white object moving over a black background, local isometry does not hold.

2. **Nuisance variables:** In addition to the small number of degrees of freedom in the articulations of interest, real-world images ensembles often exhibit a potentially large number of other, nuisance articulations, such as illumination variations, changing backgrounds and clutter, and occlusions due to foreground objects. See Fig. 1 for an illustrative example.

This mismatch between theoretical assumptions and practical realities has diminished the impact of manifold models for real-world machine learning and vision problems.

In this paper, we propose a new approach for manifold-based image modeling, learning, and processing that addresses the two challenges. First, to address the isometry challenge, we rigorously prove that the classical *Earth Mover's Distance* (EMD) between images can be used to establish isometry for image ensembles generated by translations and rotations of a reference image. This result makes no restrictive assumptions and holds even when the images under consideration are highly textured grayscale images. To the best of our knowledge, this is the first analytical result proving the isometry of generic image manifolds.

Second, to address the nuisance variable challenge, we advocate a new image representation for manifold modeling, learning, and processing. Given a set of

articulating images, we represent each image using a set of *local features* (or keypoints). Such an approach is ubiquitous in practical computer vision approaches. A keypoint typically consists of a 2D *location* in the image domain and a higher-dimensional *descriptor* summarizing the local statistics of the grayscale values of the image. We will require that the keypoint locations and descriptors satisfy certain stability criteria (explained further in Section 4). Our running example will be the image features generated by the well-known Scale Invariant Feature Transform (SIFT) [7], but other image features are also possible within this framework. Under this new representation, we show that the transformed set of images can be viewed as a low-dimensional manifold that we dub the *keypoint articulation manifold* (KAM). In fact we prove that, under a suitable modification of the EMD metric, the KAM is smooth and isometric to the underlying parameter space. By moving to this alternate representation, we implicitly promote robustness to various nuisance parameters (such as varying illumination, backgrounds, occlusions, and clutter). Therefore, our proposed KAM modeling approach alleviates both of the challenges encountered in practical applications.

Third, to mitigate computational complexity concerns related to the EMD, we propose a fast EMD approximation based on similarity kernels between the keypoint representations. We validate the approximation on several real datasets and manifold-based learning problems and demonstrate improved manifold embeddings, improved parameter estimation on affine articulation manifolds using gradient descent, and a fast, efficient, and automatic organization of large unordered collections of photographs.

This paper is organized as follows. In Section 2, we review the existing literature on the nonlinear dimensionality reduction of image manifolds. In particular, we highlight some efforts geared towards addressing some of the fundamental challenges towards practical use of image manifolds, and discuss their limitations. In Section 3, we describe how the EMD ensures isometry of manifolds for simple classes of articulations. In Section 4, we extend the EMD to be applicable to a local feature-based image representation that enables robustness to undesirable articulations. In Section 5, we illustrate the performance of our approach on a range of manifold modeling and processing applications, and validate our technique on a number of image ensembles. In Section 6, we conclude with a discussion and highlight some directions for future research.

2. Background

2.1. Image articulation manifolds

In this paper, we are interested in image ensembles that are generated by varying an articulation parameter $\theta \in \Theta$. If Θ is a space of dimension K, then the ensemble of images forms a K-dimensional nonlinear image articulation manifold (IAM) $\mathcal{M} \subset \mathbb{R}^N$:

$$\mathcal{M} = \{I_\theta : \theta \in \Theta\}. \tag{1}$$

We adopt two complementary representations for images. First, we can model images as *continuous* functions on $\mathbb{R}^2$, i.e., $I : \mathbb{R}^2 \mapsto \mathbb{R}$. In such situations, if Θ is a space of dimension K, then the ensemble of images forms a K-dimensional *image articulation manifold* (IAM) $\mathcal{M} \subset L_2(\mathbb{R}^2)$. Second, we can model images as *discretized* functions defined over a domain of size $n \times n$. In such situations, the ensemble of images are modeled as points in $\mathbb{R}^N$, where $N = n^2$. We will use these two representations interchangeably when the context is clear.

Manifold learning is a *nonlinear* dimensionality reduction technique that aims to recover a faithful approximation to the underlying parameters $\{\theta_1, \theta_2, \ldots, \theta_M\}$ given example images $I_{\theta_1}, I_{\theta_1}, \ldots, I_{\theta_M}\} \subset \mathcal{M}$. Two common assumptions made by several practical manifold learning algorithms are that $\mathcal{M}$ is *smooth*, and that $\mathcal{M}$ is *isometric* to the underlying parameter space.

1. **Smoothness:** Informally, an IAM $\mathcal{M}$ is said to be smooth if a well-defined notion of *tangent space* exists at every point $I_\theta \in \mathcal{M}$. Formally, given an IAM, we can define tangent vectors at the point I_{θ_0} by studying curves passing through it. Let $\omega_\theta : [0, 1] \mapsto \mathcal{M}$ be a curve on the IAM such that $\omega_\theta(0) = I_{\theta_0}$ where θ is a K–dimensional parameter vector. The tangent vector associated with this curve at I_{θ_0} is given by

$$\frac{d}{dt}\omega_\theta(t)|_{t=0} = [\nabla_{\theta_1} I_{\theta_0} \ \cdots \ \nabla_{\theta_K} I_{\theta_0}] \, (\theta - \theta_0). \tag{2}$$

   The tangent space at I_{θ_0} is defined as the linear *span* of the gradient vectors $\nabla_\theta I_{\theta_0} = [\nabla_{\theta_1} I_{\theta_0}, \ \cdots \ , \nabla_{\theta_K} I_{\theta_0}]$. If this linear vector space is invariant to choice of the curve ω, then $\mathcal{M}$ is said to be be smooth at I_{θ_0}.
2. **Isometry:** The mapping $I : \theta \mapsto I_\theta$ is said to be *locally isometric* if Euclidean distances between images in a small neighborhood on the manifold $\mathcal{M}$ are proportional to the corresponding distances in the articulation space:

$$d_{\mathcal{M}}(I_{\theta_1}, I_{\theta_0}) \doteq \|I_{\theta_1} - I_{\theta_0}\|_2 = C\|\theta_1 - \theta_0\|_2. \tag{3}$$

If this property holds for all local neighborhoods on $\mathcal{M}$, then $\mathcal{M}$ is said to be isometric to Euclidean space.

A host of computational techniques for efficient nonlinear dimensionality reduction have been developed. Several of these techniques assume one or both of the above two assumptions. Some well-known techniques include Locally Linear Embedding [2], ISOMAP [1], Laplacian Eigenmaps [4], Hessian Eigenmaps [3], Maximum Variance Unfolding [8], and Diffusion Maps [9].

2.2. Negative results for image manifolds

In spite of the elegance and promise of manifold modeling and learning, it has been rigorously shown that practical IAMs are neither smooth nor isometric. In most practical situations, the images under consideration have sharp edges that transform according to the articulation parameter θ. [5] show that such transformations induce a non-Lipschitz relationship between the distance metric $d_{\mathcal{M}}(\cdot,\cdot)$ and the Euclidean distance defined on vectors in Θ; specifically,

$$d_{\mathcal{M}}(I_{\theta_1}, I_{\theta_0}) = \|I_{\theta_1} - I_{\theta_2}\|_2 \geq C\|\theta_1 - \theta_2\|_2^{1/2}, \tag{4}$$

for a constant C independent of θ_1, θ_2. Due to the Lipschitz regularity exponent $1/2$ (instead of 1), the function $\theta \mapsto I_\theta$ is non-smooth everywhere. From a geometric perspective, the manifold of images containing moving edges is *nowhere differentiable*. This inherent non-smooth nature of image manifolds impedes the application of standard differential geometry-based tools used in nonlinear manifold modeling.

Efforts have been made to alleviate the non-differentiability of image manifolds [5, 10]. The basic approach is to define a *smoothing functional* that acts on the individual images I; for instance, this can be a 2D Gaussian kernel ϕ_s of scale s. By applying ϕ_s to all images in the manifold $\mathcal{M}$, we obtain a new set of images that do not contain any sharp edges; this results in a differentiable manifold $\mathcal{M}_s$ that is more amenable to analysis. The parameter s can be viewed as a scale parameter; computations can be performed at a sequence of different values for s, paving the way to *multiscale* numerical methods. This is particularly useful for common numerical tasks such as manifold-based parameter estimation using gradient descent [10].

While multiscale smoothing can render a manifold differentiable, it does not necessary lead to isometry. [5] have shown that isometry is guaranteed only for manifolds of black-and-white images exhibiting certain types of restrictive symmetries. However, for a pair of generic grayscale images $I_{\theta_1}, I_{\theta_2}$ belonging to $\mathcal{M}$,

the distance metric $d_{\mathcal{M}}(\theta_1, \theta_2)$ computed between the images smoothed at scale s is not necessarily proportional to $\|\theta_1 - \theta_2\|$ for any choice of the parameter s. This hampers the performance of any and all manifold-based algorithms that hinge upon the isometry assumption.

2.3. Local image features

Modern image processing and computer vision algorithms often eschew the pixel intensity representation for a more convenient, *feature-based* representation. Such a feature-based image modeling approach has found widespread use in a multitude of practical applications, including object recognition [11], multi-view 3D scene reconstruction [12], and manipulating and visualizing massive photo collections [6]. For an introduction to image features and their properties, see [13] and [14].

Perhaps the most popular feature-based representation of images is obtained by the *Scale Invariant Feature Transform* (SIFT) [7]. The core idea underlying the SIFT technique is the notion of *scale space* [15]. The scale space of an image I is the 3D scalar-valued function $L : \mathbb{R}^2 \times \mathbb{R} \mapsto \mathbb{R}$ obtained by convolving I with an isotropic Gaussian smoothing kernel of scale s so that

$$L(\boldsymbol{x}, s) = \phi_s * I. \tag{5}$$

Rich information about an image can be gleaned by analyzing the Laplacian of the scale space of the image, or $\nabla^2 L(\boldsymbol{x}, s)$. Indeed, extensive testing [13] has shown that the locations of maxima and minima of $\nabla^2 L(\boldsymbol{x}, s)$ (denoted by a list of 2D locations and scales $S = \{\boldsymbol{x}^i, s^i\}$) are extremely stable to small affine deformations of the image. The SIFT technique leverages this property of scale space to extract distinctive features from images.

Numerically, the SIFT technique proceeds as follows. An image I is operated upon to obtain a set of 2D locations called *keypoint locations* $\{\boldsymbol{x}^i, i = 1, \ldots, M\}$; these are precisely the extrema of the Laplacian of the scale space of I. Each keypoint location $\boldsymbol{x}^i$ is assigned a scale s^i, and an orientation θ^i. Once the set of keypoint locations are identified, certain special image statistics around each keypoint are computed and aggregated in the form of histograms. Such histograms are stored as high-dimensional vectors known as *keypoint descriptors* $\{\boldsymbol{f}^i, i = 1, \ldots, M\}$.

It has been both theoretically and empirically demonstrated that the SIFT keypoint locations are *covariant* to affine articulations, while the SIFT keypoint descriptors are *invariant* to a wide range of imaging parameters, including translations, in-plane rotations, scale, and illumination changes [7, 16]. Let I_A and I_B

be two images with keypoints given by $S(I_A) = \{\boldsymbol{x}_A^i\}$ and $S(I_B) = \{\boldsymbol{x}_B^j\}$, respectively. If the two images are related by an affine transformation (Z, t), then the keypoints are related by the same affine transformation (ignoring quantization and boundary artifacts):

$$I_B(\boldsymbol{x}) = I_A(Z\boldsymbol{x} + \boldsymbol{t}) \implies \forall i, \exists j \text{ such that } \boldsymbol{x}_B^j = Z\boldsymbol{x}_A^i + \boldsymbol{t}. \tag{6}$$

Therefore, by obtaining one-to-one correspondences between the keypoint descriptors of I_A and I_B, we can solve for the affine transformation (Z, t) linking the two images.

We have nominally chosen to focus on the SIFT as our flagship approach for generating image features, but other feature extraction techniques can also be applied in the framework developed below (for example, see [17, 18, 19]). In general, we will require that any such technique should yield image feature keypoints whose locations are covariant to the articulations of interest, and whose descriptors are invariant to the keypoint location, as well as other nuisance articulations.[1] The covariance-invariance properties help mitigate several phenomena such as unknown illuminations, occlusion, and clutter as detailed in Sections 4 and 5.

The large majority of manifold learning methods do not leverage the feature-based approach for representing images. To the best of our knowledge, the only reported manifold learning method that explicitly advocates feature-based image representations is the *Local Features* approach [20]. Given a collection of images, this approach extracts a set of local features from each of the images, and then learns a low-dimensional parametric embedding of *each* extracted feature. This embedding is constrained to preserve the spatial configuration of features. Further, similarity kernels are used to construct similarities on the keypoint locations and descriptors, and embeddings of the keypoints are learnt. This method has been shown to be robust to illumination, occlusions, and other artifacts, and thus shares many of the goals of our proposed approach. However, its theoretical development is somewhat ad hoc, its computational costs are potentially high, and the reported applications are mainly restricted to object detection and classification. We will discuss and compare our results to the Local Features approach in detail in Section 5.

[1] Naturally, the trivial (zero) feature descriptor also satisfies this invariance requirement. Our theoretical results below will continue to be valid for such degenerate cases; however, a meaningful feature descriptor that concisely represents local image statistics is obviously the better choice in practice.

3. Manifold Isometry via the Earth Mover's Distance

The central results of [5] advocating the multiscale smoothing approach for enabling manifold isometry were derived based on the assumption that images are modeled as functions defined on $\mathbb{R}^2$ equipped with the L_2-norm. However, this modeling assumption is comes up short in a key respect: L_2-distances between images are known to be poorly correlated with perceptual differences between images. For example, given images of a single translating white dot on a black background, the L_2-distance between any pair of images remains constant regardless of the translation parameters of the images.

3.1. The Earth Mover's Distance (EMD)

To address the pitfall caused by L_2-distances, researchers have proposed a multitude of alternate, perceptually meaningful distance measures on images. An important and useful metric used in image retrieval and analysis is the Earth Mover's Distance (EMD) [21]. Classically, the EMD is defined between distributions of mass over a domain, and represents the minimal amount of *work* needed to transform one distribution into another. In this context, the amount of work required to move a unit of mass from a point $\boldsymbol{x}_1 \in \mathbb{R}^2$ to a point $\boldsymbol{x}_2 \in \mathbb{R}^2$ is equal to the L_2-norm between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$.

For ease of exposition we will assume that images are defined over a discrete grid in $\mathbb{R}^2$, while noting that the results hold *mutatis mutandis* for continuous domain images. Formally, consider images I_1, I_2 as non-negative functions defined on a domain of size $n \times n$. Define a *feasible flow* as a function $\gamma : [n]^2 \times [n]^2 \to \mathbb{R}_+$ that satisfies the mass conservation constraints, i.e., for any pair of pixel locations $\boldsymbol{x}_i, \boldsymbol{y}_j \in [n]^2$,

$$\sum_{\boldsymbol{y}_k \in [n]^2} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_k) = I_1(\boldsymbol{x}_i), \qquad \sum_{\boldsymbol{x}_k \in [n]^2} \gamma(\boldsymbol{x}_k, \boldsymbol{y}_j) = I_2(\boldsymbol{y}_j).$$

Then, we define

$$\textit{EMD}(I_1, I_2) = \min_{\gamma} \sum_{\boldsymbol{x}_i, \boldsymbol{y}_j \in [n]^2} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2, \tag{7}$$

as the minimum cost flow from X to Y over all feasible flows. If the sum of the absolute values of the intensities of X and Y are equal, i.e., if $\|X\|_1 = \|Y\|_1$, then it can be shown that $\textit{EMD}(X, Y)$ is a valid metric on the space of images. In this section, we will assume the equality of the ℓ_1 norms of X and Y; however,

the metric property of the EMD holds even when this assumption is relaxed [21]. Unless otherwise specified we will assume that the EMD is always computed between images of equal ℓ_1 norm.

The EMD provides a powerful new angle for studying the geometric structure of image manifolds. As opposed to modeling images as functions in $L_2(\mathbb{R}^2)$, we instead represent images as elements of the normed space $L_{EMD}(\mathbb{R}^2)$. Under this geometry, we can prove the isometry of a much larger class of image ensembles; we discuss now some representative examples.

3.2. Case study: Translation manifolds

First, we prove the *global* isometry of image manifolds in $L_{EMD}(\mathbb{R}^2)$ formed by arbitrary translations of a generic image. Consider an image I_0, and denote $\mathcal{M}_{\text{trans}}$ as the IAM generated by 2D translations of I_0, where $\theta \in \Theta \subset \mathbb{R}^2$ represents the translation parameter vector:

$$\mathcal{M} = \{I : I(\boldsymbol{x}) = I_0(\boldsymbol{x} - \theta),\ \theta \in \Theta\}.$$

In order to avoid boundary and digitization effects, we will assume that the space of translation parameters Θ is compact, that the image has been sufficiently zero-padded, and that the images are of high resolution. It follows that the ℓ_1 norm of any image belonging to $\mathcal{M}_{\text{trans}}$ remains constant, i.e., $\|I_0\|_1$ is constant and well-defined.

Proposition 1. *For an arbitrary base image I_0, the translation manifold $\mathcal{M}_{\text{trans}}$ is* globally isometric *to the parameter space Θ under the EMD metric.*

Proof: Consider any pair of images

$$I_1(\boldsymbol{x}) = I_0(\boldsymbol{x} - \theta_1),\ \ I_2(\boldsymbol{x}) = I_0(\boldsymbol{x} - \theta_2)$$

that are elements of $\mathcal{M}_{\text{trans}}$. We will prove that $EMD(I_1, I_2)$ is proportional to the ℓ_2 distances between the corresponding parameter vectors $\|\theta_1 - \theta_2\|_2$. Let $\check{\boldsymbol{x}}$ denote the *center of mass* of the image $I(\boldsymbol{x})$:

$$\check{\boldsymbol{x}} = \frac{1}{\|I\|_1} \sum_{\boldsymbol{x}_k \in [n]^2} \boldsymbol{x}_k I(\boldsymbol{x}_k).$$

Then, we have the following relations between the centers of mass of I_1, I_2 and *any* feasible flow f:

$$\begin{aligned}
\|\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2\|_2 &= \left\| \frac{\sum_i \boldsymbol{x}_i I_1(\boldsymbol{x}_i)}{\|I_1\|} - \frac{\sum_j \boldsymbol{y}_j\, I_2(\boldsymbol{y}_j)}{\|I_2\|_1} \right\|_2 \\
&= C \left\| \sum_i \boldsymbol{x}_i I_1(\boldsymbol{x}_i) - \sum_j \boldsymbol{y}_j I_2(\boldsymbol{y}_j) \right\|_2 \\
&= C \left\| \sum_i \boldsymbol{x}_i \sum_k \gamma(\boldsymbol{x}_i, \boldsymbol{y}_k) - \sum_j \boldsymbol{y}_j \sum_k \gamma(\boldsymbol{x}_k, \boldsymbol{y}_j) \right\|_2 \\
&= C \left\| \sum_{i,k} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_k)\boldsymbol{x}_i - \sum_{j,k} \gamma(\boldsymbol{x}_k, \boldsymbol{y}_j)\boldsymbol{y}_j \right\|_2 \\
&= C \left\| \sum_{i,j} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j)(\boldsymbol{x}_i - \boldsymbol{y}_j) \right\|_2 \\
&\leq C \sum_{i,j} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) \left\| \boldsymbol{x}_i - \boldsymbol{y}_j \right\|_2,
\end{aligned}$$

where the last inequality is a consequence of the triangle inequality. Taking the infimum over all possible feasible flows, we have that

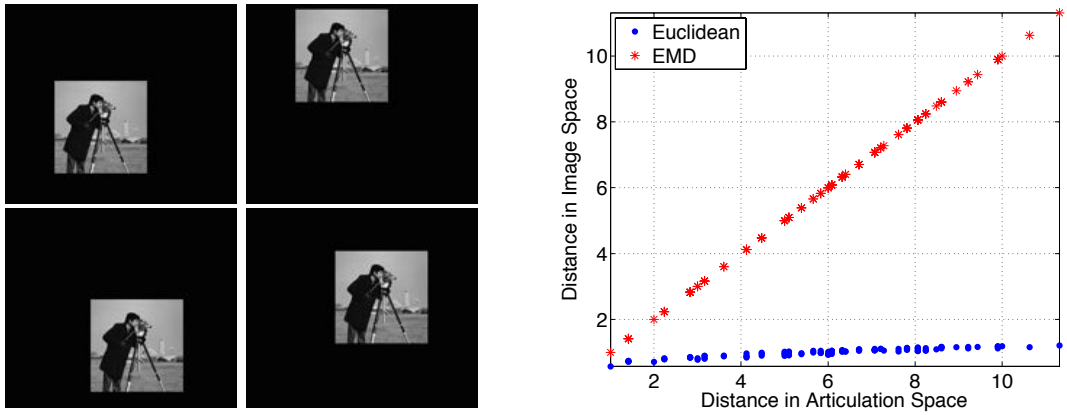
$$\|\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2\|_2 \leq C \cdot \textit{EMD}(I_1, I_2). \tag{8}$$

However, in the case of images that are 2D translations of one another, there always exists a feasible flow that *achieves* this infimum. This can be represented by the set of flows parallel to $\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2$ originating from the pixel $\boldsymbol{x}_i$ and terminating at the corresponding $\boldsymbol{y}_j$. We simply rewrite the vector $\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2$ as the difference in translation vectors $\theta_1 - \theta_2$, thereby implying that

$$\textit{EMD}(I_1, I_2) \propto \|\theta_1 - \theta_2\|_2.$$

Global isometry of $\mathcal{M}_{\text{trans}}$ is an immediate consequence. □

We numerically illustrate the validity of Proposition 1 in Fig. 2. Figure 2(a) displays several sample images from the manifold formed by translations of the well-known *Cameraman* test image. We form 100 example pairs of such images, record the distance between the translation parameter vectors (the "distance in articulation space"), and compute the Euclidean (ℓ_2) distance and EMD between



(a) (b)

Figure 2: *(a) Sample images from a translation manifold. (b) Variation of the Euclidean distance and the EMD as a function of the distance in the articulation space. The EMD correlates linearly with articulation distance for the entire range of articulations (global isometry).*

the corresponding images. We compute the EMD using the FastEMD solver [22]. Figure 2(b) clearly indicates that the ℓ_2 distance is largely uninformative with respect to the articulation distance, while the EMD almost perfectly correlates with the articulation distance over the entire range of translations (global isometry).

3.3. Case study: Rotation manifolds

Next, we prove the *local* isometry of image manifolds formed by rotations of a generic image. The IAM $\mathcal{M}_{\text{rot}}$ is generated by pivoting an image I_0 by an angle $\theta \in \Theta \subset [-\pi, \pi]$, around a fixed point in $\mathbb{R}^2$. We assume for simplicity that the pivot point is the origin. Then, the manifold $\mathcal{M}_{\text{rot}}$ is given by

$$\begin{aligned} \mathcal{M} &= \{I : I(\boldsymbol{x}) = I_0(R_\theta \boldsymbol{x}),\ \theta \in \Theta\}, \quad \text{where} \\ R_\theta &= \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \end{aligned}$$

i.e., R_θ is a orthonormal rotation matrix. Once again, we assume that the images are sufficiently zero padded and that their ℓ_1 norms remain constant.

Proposition 2. *For an arbitrary base image I_0, the rotation manifold $\mathcal{M}_{\text{rot}}$ is* locally isometric *to the parameter space Θ under the EMD metric.*

Proof: Consider any pair of images

$$I_1(\boldsymbol{x}) = I_0(R_{\theta_1}\boldsymbol{x}), \;\; I_2 = I_0(R_{\theta_2}\boldsymbol{x})$$

that are elements of $\mathcal{M}_{\text{rot}}$. Since the set of rotations in $\mathbb{R}^2$ forms a group (called the special orthogonal group $SO(2)$), we have the relation

$$I_2(\boldsymbol{x}) = I_1(R_{\theta_1-\theta_2}\boldsymbol{x}) = I_1(R_{\Delta\theta}\boldsymbol{x}). \tag{9}$$

Once again, we denote the locations of the centers of mass of I_1 and I_2 as $\check{\boldsymbol{x}}_1$ and $\check{\boldsymbol{x}}_2$ respectively. Observe that the centers of mass of I_1 and I_2 also obey the relation $\check{\boldsymbol{x}}_2 = R_{\Delta\theta}\check{\boldsymbol{x}}_1$. Hence, we have

$$\|\check{\boldsymbol{x}}_2 - \check{\boldsymbol{x}}_1\|_2 \;\; = \;\; \|R_{\Delta\theta}\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_1\|_2 = \left\| \left(\begin{bmatrix} \cos\Delta\theta & -\sin\Delta\theta \\ \sin\Delta\theta & \cos\Delta\theta \end{bmatrix} - \mathbb{I}_{2\times 2} \right) \check{\boldsymbol{x}}_1 \right\|_2 .$$

To establish local isometry, we need to show that the EMD between a pair of images exhibits a linear relationship with the magnitude of the distance in articulation space $\Delta\theta$ in the regime where $\Delta\theta$ is small. In such a regime, we can perform a first-order Taylor series expansion to obtain

$$\begin{aligned} \|\check{\boldsymbol{x}}_2 - \check{\boldsymbol{x}}_1\|_2 \;\; &\approx \;\; \left\| \left(\begin{bmatrix} 1 & -\Delta\theta \\ \Delta\theta & 1 \end{bmatrix} - \mathbb{I}_{2\times 2} \right) \check{\boldsymbol{x}}_1 \right\|_2 = \left\| \begin{bmatrix} 0 & -\Delta\theta \\ \Delta\theta & 0 \end{bmatrix} \check{\boldsymbol{x}}_1 \right\|_2 \\ &= \;\; |\Delta\theta| \; \|\check{\boldsymbol{x}}_1\|_2. \end{aligned}$$

However, the quantity $\|\check{\boldsymbol{x}}_1\|$ represents the distance of the center of mass of image I_1 from the origin, which is constant for images belonging to $\mathcal{M}_{\text{rot}}$. Further, we established in (8) that the distance between the centers of a pair of images is upper bounded by a constant times the EMD between the images. Hence, for some constant $\alpha > 0$, we have the following lower bound:

$$\textit{EMD}(I_1, I_2) \geq \alpha|\Delta\theta|. \tag{10}$$

We now prove a similar upper bound on the EMD. By definition, the EMD is calculated by considering the minimum over all feasible flows from I_1 to I_2. Consider the (feasible) flow f corresponding to the bijective mapping between $I_1(\boldsymbol{x}) \doteq I_1(R_{\Delta\theta}\boldsymbol{y})$ and $I_2(\boldsymbol{y})$, i.e.,

$$\gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) = \begin{cases} I_1(\boldsymbol{x}_i), & \boldsymbol{x}_i = R_{\Delta\theta}\boldsymbol{y}_j \\ 0, & \text{otherwise.} \end{cases}$$

For small values of $\Delta\theta$, the magnitude of the displacement of the pixel $\boldsymbol{x}_i$ induced by this flow can be approximated as

$$\|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2 \approx |\Delta\theta| \; \|\boldsymbol{x}_i\|_2,$$

and hence the cost of the flow f can be computed by evaluating the right hand side of (7). This quantity provides an upper bound for the EMD between images I_1 and I_2 as follows:

$$\begin{aligned} \textit{EMD}(I_1, I_2) &\leq \sum_{i,j} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2 \\ &= \sum_i I(\boldsymbol{x}_i) \; |\Delta\theta| \; \|\boldsymbol{x}_i\|_2. \end{aligned}$$

The ℓ_2-norm of $\boldsymbol{x}$ is invariant with respect to rotation, and hence the quantity $\sum_i I(\boldsymbol{x}_j) \; \|\boldsymbol{x}_j\|_2$ is constant across all images I belonging to $\mathcal{M}_{\text{rot}}$. Therefore, for some constant $\beta > 0$, we have the following upper bound:

$$\textit{EMD}(I_1, I_2) \leq \beta |\Delta\theta|. \tag{11}$$

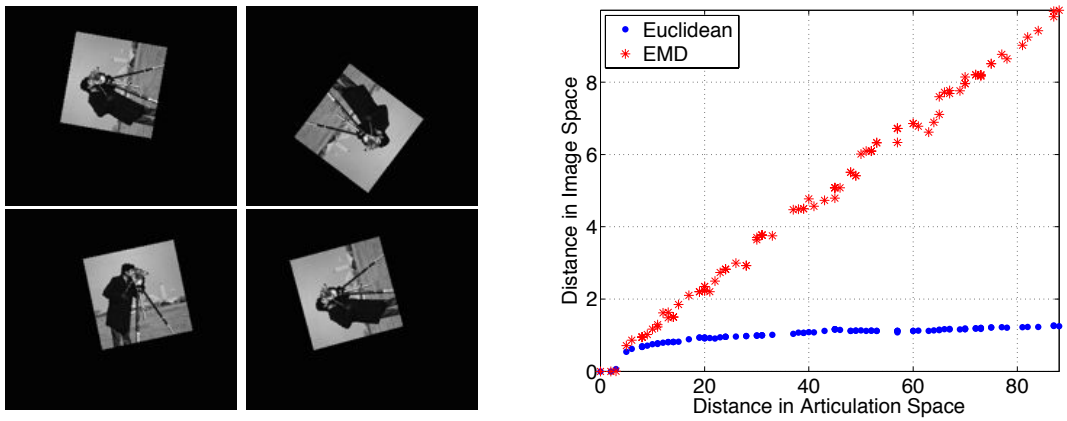
Combining (10) and (11), we obtain that the manifold $\mathcal{M}_{\text{rot}}$ is approximately isometric to Θ under the EMD metric. □

We numerically illustrate the validity of Proposition 2 in Fig. 3. Figure 3(a) displays several sample example images formed by rotations of the *Cameraman* test image. As above, we form 100 example pairs of such images, record the distance between the rotation parameter vectors (the "distance in articulation space"), and compute the Euclidean (ℓ_2) distance and EMD between the corresponding images. Figure 2(b) clearly indicates that the ℓ_2 distance is largely uninformative with respect to the articulation distance, while the EMD closely correlates with the articulation distance (local isometry).

4. Keypoint Articulation Manifolds

Thus far, we have formally proved — for arbitrary translation and rotation manifolds containing images with sharp edges and complex textures — that replacing the ℓ_2 with the EMD surmounts the non-isometry challenge that has plagued manifold frameworks to date. [2] We now turn to the second challenge of nuisance

[2] A natural question arises whether we can extend the results of the preceding two sections to general affine articulations. However, since the EMD is technically defined on pairs of images of equal mass, a similar (local) isometry argument does not seem to apply to the case where the images undergo non-uniform scaling.



(a) (b)

Figure 3: *(a) Sample images from a rotation manifold. (b) Variation of the Euclidean distance and the EMD as a function of the distance in the articulation space. The EMD correlates nearly linearly with articulation distance for the entire range of articulations (local isometry).*

variables caused by real-world artifacts in the imaging enterprise, such as varying illumination, non-stationary noise, unknown backgrounds, and occlusions.

Consider the set of images generated by a translating a white disk in front of a black background under an unknown, spatially varying illumination. Because of the varying illumination, the pixel intensities of the disk will not be constant across the images. In this case, the minimum-cost flow in (7) will not be mass-preserving, and the EMD will not be isometric to the translation parameter distance. The standard practical approach to handling illumination variations is to transform the image into a feature-based representation that is robust to such variations. In this section, we propose a systematic framework for analyzing families of articulating images not in terms of their pixel intensities but rather in terms of their *local features)*. As we will see, a number of theoretical and practical advantages result.

4.1. Feature-based representations for images

We consider local feature representations that consist of a set of image *keypoints* and a corresponding set of *descriptors*. Given an image I defined as a real-valued function over a domain $\Omega \subset \mathbb{R}^2$, we compute the set of *keypoint locations* $X(I) = \{\boldsymbol{x}_i, i = 1, \ldots, N\} \subset \Omega$ using a local feature extraction algorithm $\mathcal{A}$. At the computed keypoint locations, we compute *keypoint descriptors* $F(I) = \{\boldsymbol{f}_i, i = 1, \ldots, N\}$; each $\boldsymbol{f}_i \in F$ typically can be described as a vector

in high-dimensional space $\mathbb{R}^D$. Thus, instead of representing an N-pixel image I as a vector in $\mathbb{R}^N$, we represent it as a set of keypoint location-descriptor pairs $I \sim \{(\boldsymbol{x}_i, \boldsymbol{f}_i), i = 1, \ldots, N\}$, or informally, a "bag of keypoints." Each keypoint location-descriptor pair is an element of an abstract space $\mathcal{X}$ that can be identified with $\mathbb{R}^2 \times \mathbb{R}^D$. Note that $\mathcal{X}$ in itself does not constitute a normed vector space, primarily because the space F is typically not closed under the usual operations of addition and scalar multiplication.

We require that the local feature extraction algorithm $\mathcal{A}$ possess the following properties:
(P1) The keypoint locations are *covariant* to the articulation parameters of interest. For example, in the case of translation, a global translation applied to an image must induce an equivalent, global translation in *every* computed keypoint location.
(P2) The keypoint descriptors are *invariant* to the image articulation parameters of interest.
(P3) The keypoint extraction is *stable*, i.e., no spurious keypoints are detected or missed across different images on the manifold.

Of course, a keypoint extraction algorithm $\mathcal{A}$ exactly satisfying these three properties is hypothetical and may not exist in practice. However, several efficient feature extraction methods have been extensively explored and shown to possess **(P1)–(P3)** to a close approximation. The most celebrated is the *Scale Invariant Feature Transform* (SIFT) [7], which approximately possesses **(P1)–(P3)** for the case of affine articulations [16] . We will focus on this technique in our computations below without loss of generality.

Definition 1. *Given a keypoint extraction algorithm $\mathcal{A}$ that satisfies properties* **(P1)–(P3)** *and an IAM $\mathcal{M} = \{I_\theta : \theta \in \Theta\}$, the* keypoint articulation manifold *(KAM) is defined as $\mathcal{K} = \{I_\theta \sim \{(\boldsymbol{x}_i, \boldsymbol{f}_i)\}_{i=1}^M : I_\theta \in \mathcal{M}\}$.*

We seek an appropriate metric on the set $\mathcal{K}$. Consider a grayscale image $I_0(\boldsymbol{x}) \sim \{(\boldsymbol{x}_i, \boldsymbol{f}_i)\}_{i=1}^M$. Define the *keypoint location image* as

$$K_0(\boldsymbol{x}) = \sum_{i=1}^{M} \delta(\boldsymbol{x} - \boldsymbol{x}_i),$$

where $\delta(\cdot)$ is the Kronecker delta function. The keypoint location image can be viewed as a non-negative function over the discrete domain, i.e., $K \in \mathbb{R}_+^N$. Therefore, it is possible to define the EMD between any pair of keypoint location images, which induces a metric on the KAM $\mathcal{K}$. That is, for any pair of images

$I_{\theta_1}, I_{\theta_2} \in \mathcal{M}$, we define the *keypoint distance* d_κ as the EMD between their corresponding keypoint location images:

$$d_\kappa(I_{\theta_1}, I_{\theta_2}) = \textit{EMD}(K_{\theta_1}, K_{\theta_2}).$$

It should be obvious from the properties **(P1)–(P3)** that the KAM generated by an ideal keypoint extraction algorithm $\mathcal{A}$ is smooth and globally isometric to any parameter space for which the covariance property **(P1)** holds. We now showcase the power of the invariance property **(P2)**.

4.2. Case study: Illumination variations

We prove the following proposition about the geometry of the KAM generated by applying an idealized SIFT-like transformation.

Proposition 3. *Consider an IAM $\mathcal{M}$ generated by images of an arbitrary object as it undergoes 2D translations and in-plane rotations and is then illuminated by an unknown spatially varying illumination. Let $\mathcal{K}$ be the KAM generated by applying a keypoint extraction algorithm $\mathcal{A}$ that is covariant to translation and in-plane rotation, and invariant to illumination. Then $\mathcal{K}$, endowed with the keypoint distance d_κ, is globally isometric to the parameter space Θ.*

Proof: We will describe the case where the articulations comprise 2D translations; the extension to in-plane rotations is straightforward and mirrors the derivation in Proposition 2. Any image $I \in \mathcal{M}$ corresponding to the translation parameter θ can be expressed in terms of a base image I_0 as

$$I(\boldsymbol{x}) = L_\theta I_0(\boldsymbol{x} - \theta),$$

where L_θ represents an unknown linear operator representing the illumination corresponding to θ. Consider any pair of images

$$I_1(\boldsymbol{x}) = L_{\theta_1} I_0(\boldsymbol{x} - \theta_1), \quad I_2(\boldsymbol{x}) = L_{\theta_2} I_0(\boldsymbol{x} - \theta_2),$$

that are elements of $\mathcal{M}$. Denote the keypoint location image of I_0 as $K_0(\boldsymbol{x}) = \sum_{i=1}^{M} \delta(\boldsymbol{x} - \boldsymbol{x}_i)$. By assumption, the algorithm $\mathcal{A}$ stably extracts keypoint locations in a covariant manner, and also is invariant to the illumination operators $L_{\theta_1}, L_{\theta_2}$. Therefore,

$$K_1(\boldsymbol{x}) = K_0(\boldsymbol{x} - \theta_1), \;\; K_2(\boldsymbol{x}) = K_0(\boldsymbol{x} - \theta_2),$$

where K_1, K_2 are the corresponding keypoint location images of I_1, I_2. The keypoint distance $d_\kappa(I_1, I_2)$ is equal to the EMD between K_1 and K_2, computed using (7). However, in this case the minimum cost flow γ is nothing but a permutation (since K_1, K_2 are the superposition of an identical number M of Kronecker delta functions). Denote $\pi : X(I_{\theta_1}) \to X(I_{\theta_2})$ as a feasible permutation. Therefore,

$$\begin{aligned} \textit{EMD}(K_1, K_2) &= \min_{\gamma} \sum_{\boldsymbol{x}_i, \boldsymbol{y}_j \in [n]^2} \gamma(\boldsymbol{x}_i, \boldsymbol{y}_j) \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2 \qquad (12) \\ &= \min_{\pi} \sum_{i=1}^{M} \|\boldsymbol{x}_i - \pi(\boldsymbol{x}_i)\|_2. \qquad (13) \end{aligned}$$

The optimization (13) can be calculated, for example, via the Hungarian algorithm [23]. However, note that, for any permutation π,

$$\begin{aligned} \sum_i \|\boldsymbol{x}_i - \pi(\boldsymbol{x}_i)\| &\geq \left\| \sum_i \boldsymbol{x}_i - \sum_i \pi(\boldsymbol{x}_i) \right\|_2 \\ &= M \left\| \frac{\sum_i \boldsymbol{x}_i}{M} - \frac{\sum_i \pi(\boldsymbol{x}_i)}{M} \right\|_2 \\ &= M \|\check{\boldsymbol{x}}_1 - \check{\boldsymbol{x}}_2\|_2, \end{aligned}$$

where $\check{\boldsymbol{x}}_1, \check{\boldsymbol{x}}_2$ are the centers of mass of the keypoint location images K_1, K_2. Repeating the argument in the proof of Proposition 1, we have that this minimum cost permutation is achieved by mapping the keypoint in K_1 at location $\boldsymbol{x}_i$ to the corresponding keypoint in K_2 at location $\boldsymbol{y}_i \sim \boldsymbol{x}_i + \theta_1 - \theta_2$ (due to the covariance property, this correspondence always exists).

Therefore, $\textit{EMD}(K_1, K_2)$ is proportional to the distance between the centers of mass of K_1 and K_2, which equals $\theta_1 - \theta_2$. The isometry of the KAM is an immediate consequence. □

4.3. Practical computation of the keypoint distance

In order to realize the promise of Proposition 3 in practice, we must address three practical concerns:

1. Noise and numerical errors will render properties **(P1)**–**(P3)** approximations, at best.
2. Real-world phenomena such as occlusions and clutters will also invalidate **(P1)**–**(P3)**. Indeed, accurate detection and filtering of spurious keypoints reduces to establishing exact correspondences between the keypoints, which remains a highly challenging problem in machine vision.

3. The computational complexity of the EMD computation (12) is *cubic* in the number of extracted keypoints M, and real-world high-resolution images typically yield several hundreds or even thousands of keypoints [7].

In order to address these challenges, we now propose a computationally efficient approximation to the EMD-based keypoint distance d_κ in (12) between any pair of images. We leverage the fact that the keypoint *descriptors*, $\{\boldsymbol{f}_i\}_{i=1}^M \subset \mathbb{R}^D$, calculated from an image I_θ are (approximately) *invariant* to the articulation parameter θ (recall property **(P2)**). By evaluating a suitably defined *similarity kernel*, $S : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}$, on every pair of keypoint descriptors, we can rapidly establish approximate correspondences between the keypoints. A weighted average of the distances between the corresponding keypoint locations yields the EMD approximation.

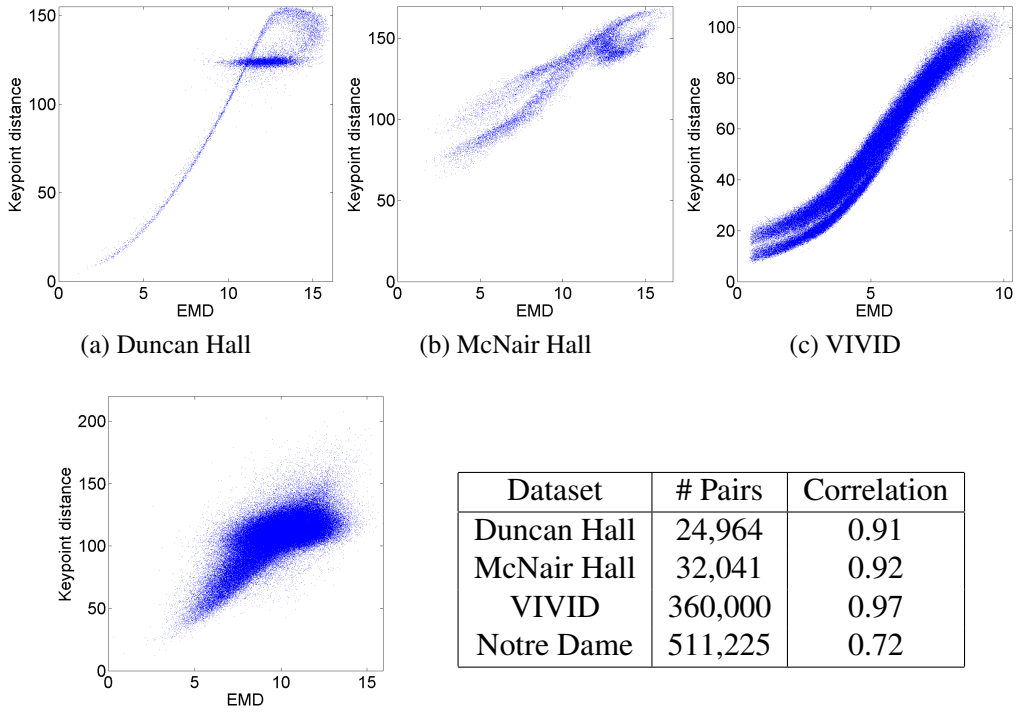
The full calculation proceeds as follows. Given images $I_1 \sim \{(\boldsymbol{x}_i, \boldsymbol{f}_i), i = 1, \ldots, M_1\}$ and $I_2 \sim \{(\boldsymbol{y}_j, \boldsymbol{g}_j), j = 1, \ldots, M_2\}$, we define the *approximate keypoint distance* between I_1 and I_2 as:

$$\begin{aligned} \widetilde{d}_\kappa(I_1, I_2) &= \alpha^{-1} \sum_{i,j=1}^{M_1, M_2} S(\boldsymbol{f}_i, \boldsymbol{g}_j) \|\boldsymbol{x}_i - \boldsymbol{y}_j\|_2, \quad \text{where} \\ \alpha &= \sum_{i,j} S(\boldsymbol{f}_i, \boldsymbol{g}_j). \end{aligned} \tag{14}$$

The normalization factor α ensures that the approximate keypoint distance does not depend on the *number* of detected keypoint pairs $M_1 \times M_2$. The ideal similarity kernel would yield a value of 1 for every pair of corresponding keypoint locations and zero for all other pairs. In the case when all the keypoint descriptors of the reference image I_0 are distinct, the similarity kernel $S(\boldsymbol{f}_i, \boldsymbol{g}_j)$ would be nonzero only when $\boldsymbol{f}_i \approx \boldsymbol{g}_j$, thereby efficiently *approximating* the minimum cost flow $\gamma(\boldsymbol{x}_i, \boldsymbol{y}_j)$ in (12) without an explicit minimization. Consequently, the complexity of evaluating the approximate keypoint distance can be reduced from $\mathcal{O}(M^3)$ to $\mathcal{O}(M^2)$, a significant advantage for practical real-world calculations. We demonstrate this computational advantage numerically in Section 5.

The choice of similarity kernel $S(\cdot, \cdot)$ is somewhat flexible. However, to account for numerical discrepancies in the descriptors extracted by the algorithm $\mathcal{A}$, we will focus on the Gaussian radial-basis kernel for $S(\cdot, \cdot)$. For any descriptor pair $(\boldsymbol{f}, \boldsymbol{g})$ and bandwidth parameter $\sigma > 0$, the similarity kernel $S(\cdot, \cdot)$ is given by

$$S(\boldsymbol{f}, \boldsymbol{g}) = e^{-\frac{\|\boldsymbol{f} - \boldsymbol{g}\|^2}{\sigma^2}}. \tag{15}$$



(a) Duncan Hall (b) McNair Hall (c) VIVID

(d) Notre Dame

Dataset	# Pairs	Correlation
Duncan Hall	24,964	0.91
McNair Hall	32,041	0.92
VIVID	360,000	0.97
Notre Dame	511,225	0.72

(e) Statistics

Figure 4: *Empirical comparisons between the EMD in (13) and the kernel-based approximate keypoint distance in (14) for image-pairs sampled from several datasets. (a-d) We plot the the corresponding pair/tuple of EMD and keypoint distance for all image-pairs from each dataset. (e) Statistics showing the size of the dataset in terms of number of image pairs and the correlation between the EMD and the approximate keypoint distance* (14).

The optimal value of σ in (15) depends on the numerical stability of the algorithm $\mathcal{A}$ used to extract feature keypoints from the images. In practice (and for all the experiments below) with SIFT feature keypoints, the value $\sigma = 150$ gave excellent numerical results; moreover, performance is stable to small changes around this value for σ.

How accurate is the quality of approximation achieved by the approximate keypoint distance (14) with respect to the (ideal) EMD keypoint distance (13)? We do not provide a rigorous analytical characterization, and instead experimentally justify that the approximation is faithful. Figure 4 displays the results of comparing the calculated pairwise distances (13) and (14) between thousands of

image pairs, sampled from several real-world datasets that we discuss in Section 5. Each panel in Fig. 4 corresponds to a particular dataset, and each blue dot in the scatter plot represents the calculated distances for a specific pair of images in the dataset. We also report correlation coefficients for the different plots (a perfectly linear relationship would correspond to a coefficient of 1.) We observe a very close correlation, both qualitatively and quantitatively, between the distances for three of the four datasets. The obvious outlier is the Notre Dame Cathedral dataset (refer to the example images in Fig. 1), which is challenging for several reasons; we discuss it in detail below in our experiments). Therefore, the approximate keypoint distance closely represents the ideal keypoint distance (while being significantly easier to compute.)

Before concluding this section, we observe that other choices of similarity kernels $S(\cdot, \cdot)$ in (14) are also possible. There exist several extensive surveys in the literature on the efficient design of similarity kernels based on local image features [24, 25]. We touch on this topic further in Section 6.

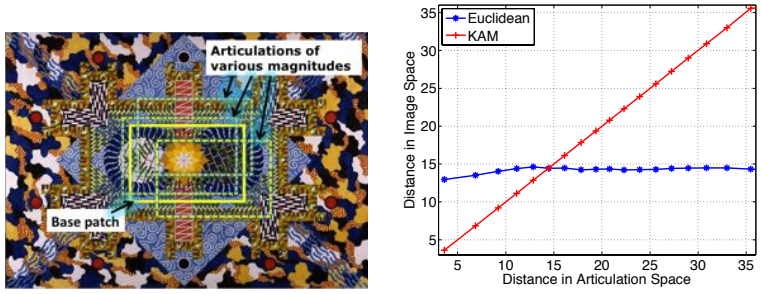
5. Experiments

This experimental section has dual aims. First, we back up the theoretical results on KAM smoothness and isometry using several real-world datasets. Second, we push the KAM technique out of its theoretical comfort zone with new, challenging applications involving a number of real-world datasets acquired "in the wild." Our intention is to convincingly demonstrate that manifold methods are not just elegant but also of considerable practical utility in real applications.

For all experiments that follow, we use the fast approximation to the EMD proposed in (14). We use SIFT as the keypoint extraction algorithm $\mathcal{A}$, the Gaussian radial basis function with $\sigma = 150$ for the similarity kernel S, and ISOMAP [1] with $k = 8$ neighbors for obtaining the low-dimensional embeddings from the distances computed from (14). We will call this procedure the "KAM approach".

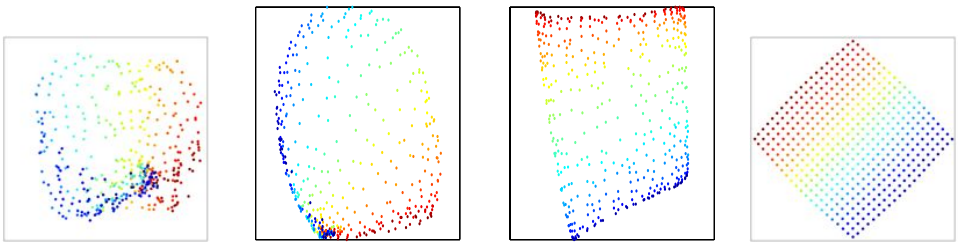
5.1. Confirming isometry

Figure 5 extends the synthetic experiment in Fig. 2 by using both the approximate EMD from (14) and real data. We extracted 400 patches of size 80×80 centered at points of a grid of *uniformly-spaced* locations in the highly textured photograph in Fig. 5(a) and replicated the experimental steps of Fig. 2. Figure 5(b) clearly indicates that the Euclidean (ℓ_2) inter-image distance is largely uninformative with respect to the articulation distance, while the approximate EMD almost



(a) Images (b) Distances

Figure 5: *Confirmation of the results of Fig. 2 using both the approximate EMD from (14) and real data. (a) Sample images from a translation manifold. (b) Variation of the Euclidean distance and the approximate EMD as a function of the distance in the articulation space. The approximate EMD correlates linearly with articulation distance for the entire range of articulations (practically confirming global isometry).*



(a) ISOMAP (b) Lap. Eigenmaps (c) Local Features (d) KAM

Figure 6: *Continuation of the experiment in Fig. 5, plotting 2D embeddings of the translated images using various state-of-the-art manifold learning methods. The KAM approach recovers the underlying parametrization perfectly (modulo a rotation).*

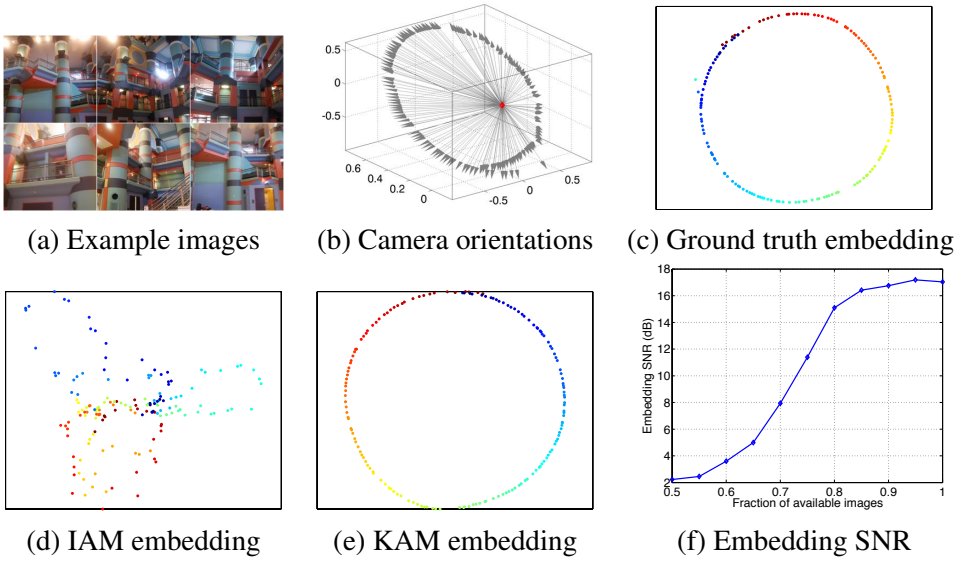
perfectly correlates with the articulation distance over the entire range of translations (practically confirming global isometry).

5.2. Manifold embedding

We now showcase the invariance and stability properties of the KAM approach with a number of challenging manifold learning (nonlinear dimensionality reduction) examples.

5.2.1. Highly textured translation manifold

Figure 6 continues the example of Fig. 5 from Section 5.1. Given the sampling of 400 highly textured, translated test images, we ran three state-of-the-art



(a) Example images (b) Camera orientations (c) Ground truth embedding

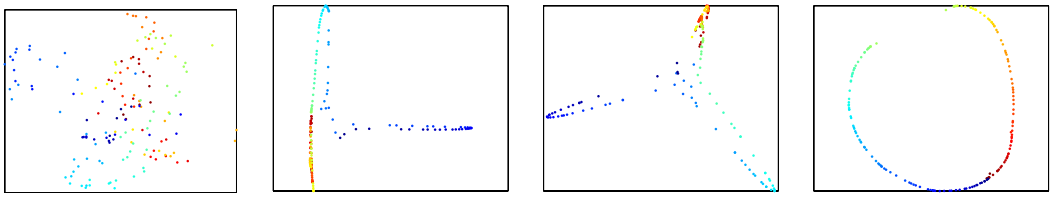
(d) IAM embedding (e) KAM embedding (f) Embedding SNR

Figure 7: *Manifold learning in the wild I: Duncan Hall indoor scene. (a) Samples from a set of 160 images obtained from an approximately static hand-held camera that span a* 360° *panorama of an indoor scene. (b) Camera orientation vectors obtained from a state-of-the-art SfM algorithm [6] to provide precise camera orientation vectors (grey arrows) for each of the images; these can be considered as the "ground truth" values of the underlying parameter space. (c) 2D ISOMAP embedding of the ground truth camera orientation vectors. (d) 2D ISOMAP embedding of the IAM using the* ℓ_2 *metric. (e) 2D KAM embedding; it is virtually equivalent to the optimal embedding using the ground truth (up to an arbitrary rotation). (f) Embedding SNR vs. fraction of available images, indicating that the performance of the KAM approach degrades gracefully with manifold subsampling.*

manifold learning algorithms (ISOMAP, LLE, and Local Features). None of them is able to recover the nonlinear projection into the 2D parameter space as well as our KAM-based ISOMAP.

5.2.2. Duncan Hall indoor scene

Using a static handheld camera, we collected a set of 160 high-resolution indoor photographs that formed a 360° panoramic view of the walls and ceiling of a large atrium in Rice University's Duncan Hall (see Fig. 7). The images are governed not only by an underlying dominant articulation parameter (the viewing angle of the camera), but also by several other degrees of freedom (camera shake and significant lighting variations, including bright sunlight glints). We applied the state-of-the-art structure-from-motion (SfM) Bundler algorithm [6] to



(a) Diffusion Maps (b) LLE (c) Lap. Eigenmaps (d) Local Features

Figure 8: *Manifold learning in the wild I: Duncan Hall indoor scene. Additional results for the dataset in Figure 7. (a, b, c) State-of-the-art manifold learning algorithms based on ℓ_2-distances between images perform poorly. (d) The Local Features (LF) approach fares better. However, the KAM approach (Fig. 7(e)) still significantly outperforms the Local Features approach in terms of fidelity to the parameter space.*

estimate, up to an arbitrary rotation, the 3D camera orientation vector for each sample image. We will regard these vectors as the "ground truth" articulation parameters for each image.

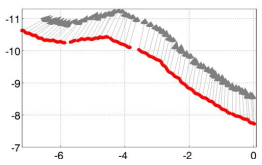
Figure 7 displays the low-dimensional (2D) embeddings obtained by ISOMAP using both the classical IAM (using the Euclidean inter-image distance) and the proposed KAM approach (using the approximate EMD inter-image distance). We note that the KAM embedding recovers a near-perfect approximation (modulo a rotation) of the underlying parametrization, whereas the IAM approach yields poor quality results. Figure 8 displays additional embeddings produced by four other maninfold learning algorithms, including the Local Features approach [20]. Clearly the KAM approach is much improved over all of these techniques. This demonstrates that the KAM approach is robust to camera jitter and changing lighting conditions.

We now demonstrate that the KAM approach is robust to the sampling of the manifold. Define the *embedding signal-to-noise-ratio (SNR)* as the negative logarithm of the L_2-error of the 2D KAM embedding measured with respect to the ground truth. Figure 7(f) shows that the embedding SNR degrades gracefully even when the KAM-based manifold learning algorithm is presented with only a random fraction of the 160 available images.

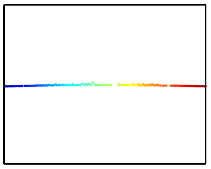
5.2.3. McNair Hall outdoor scene

We collected a set of 180 images of the front facade of Rice University's McNair Hall by walking with a handheld camera in an approximately straight trajectory; therefore, the underlying parameter space is topologically equivalent to a subset of the real line $\mathbb{R}^1$. Several sample images are shown in Fig. 9(a). We

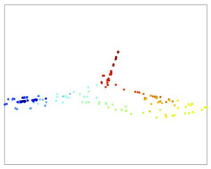
(a) Sample images



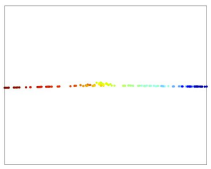
(b) Bundler estimates



(c) Bundler



(d) IAM



(e) KAM

Figure 9: *Manifold learning in the wild II: McNair Hall outdoor scene. (a) Samples from a set of 180 images obtained by moving a hand-held camera in an approximately straight trajectory. The image ensemble is topologically equivalent to a 1D manifold. (b) Camera location ground truth obtained from the SfM Bundler algorithm ([6]). Camera locations are noted in red and their orientations with grey arrows. (c) 2D ISOMAP embedding of the ground truth camera orientation vectors. (d) 2D ISOMAP embedding of the IAM using the ℓ_2 metric. (e) 2D KAM embedding is a close approximation to the ground truth embedding.*

used the SfM Bundler software to estimate the camera locations and orientations; the results are displayed in Fig. 9(b). As above, we computed low-dimensional embeddings of the images using ISOMAP on the set of pairwise Euclidean and approximate EMD image distances. The embedding obtained using the KAM approach closely resembles the "ground truth" embedding and successfully recovers the 1D topology of the image dataset.

5.3. Parameter estimation

We study the effectiveness of the KAM approach for articulation parameter estimation. Given a sample image $I_\theta \in \mathcal{M}$, $\theta \in \Theta$, our aim is to estimate the underlying vector θ. The non-differentiability of IAMs of images with sharp edges renders IAM-based approaches ineffective for this problem. However, limited progress to date has been made using multiscale smoothing and gradient descent [10]; our goal here is to demonstrate the robust performance of a simple and direct KAM-based estimate.

We consider the 400-image translation manifold dataset from Section 5.1 and Fig. 5 as a "training set". Then, we select a target image patch at random and attempt to estimate its 2D translation parameters by finding the closest among

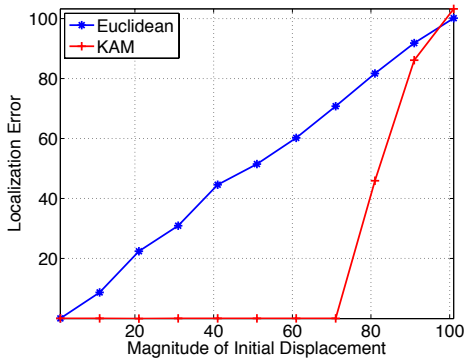


Figure 10: *Parameter estimation performance for the translation manifold in Fig. 5. The x axis corresponds to the 2D Euclidean distance between the initial translation parameters of the gradient descent and those of the target image. The y axis corresponds to the magnitude of the error between the estimated and target articulations. Gradient descent on the KAM converges accurately for a wide range of initial displacement magnitudes, while gradient descent on the IAM does not yield accurate results for even small values of initial displacement.*

the training set images via a multiscale gradient descent method; the technique used is similar to the method proposed in Section 6.4.1 of [10]. The articulation parameters of the retrieved training image serve as the estimate. We repeat this procedure using both the Euclidean (IAM) and approximate EMD (KAM) distances and record the magnitude of the error between the true and estimated target translation parameters.

Figure 10 displays the results of a Monte-Carlo simulation over 40 independent trials. Thanks to the smooth and isometric structure of the KAM, we obtain accurate estimation results even when initializing the gradient descent method far from the target translation value (over 70 pixels, which is significant considering that the images are of size 80×80 pixels). In contrast, the IAM approach suffers from large estimation errors even then starting relatively close to the target value.

We stress that we do not claim our method of estimating the translation parameters via gradient descent on the KAM as constituting a state-of-the-art image registration algorithm. Rather, our aim is merely to show that the smoothness and isometry of the KAM support even naïve information extraction algorithms, in contrast to IAMs.

5.4. Organizing photo collections

We now explore how KAMs can be used to automatically organize large collections of images, particularly collections that can be well-modeled by an essen-

tially small number of parameters. An example is the set of photos of a tourist landmark captured by different individuals at different times. The intrinsic variability of this set of photos might be extremely high, owing to occlusions (trees, vehicles, people), variable lighting, and clutter. However, the essential parameters governing the images can be roughly identifed with the 3D camera position, orientation, and zoom. We postulate that the KAM approach will help enforce this intrinsic low-dimensionality of the photos and thus provide a meaningful organization. In colloquial terms, we are organizing the photographs by solving a complicated "image jigsaw puzzle" in high-dimensional space by exploiting its low-dimensional geometry

One approach to organize photo colections is the *Photo Tourism* method [6], which runs the SfM Bundler algorithm to accurately estimate the position of each 3D point in the scene and then infers the 3D camera locations and orientations corresponding to each photograph. Unfortunately, while powerful, this algorithm is computationally very demanding and takes several days to execute for a dataset comprising even just a few hundred images.

As an alternative, we propose a far simpler approach: simply extract the keypoints from each of the images, compute the keypoint distances between all pairs of images, and then estimate the geodesics along the KAM. If the low-dimensional manifold assumption holds, then the images corresponding to the nearest neighbors along the geodesics will be semantically meaningful.

5.4.1. Notre Dame Cathedral

We test our hypothesis on the well-known Notre Dame Cathedral dataset, a collection of 715 high-resolution images of the popular Parisian tourist trap chosen randomly from FlickR. From each photo, we extract SIFT keypoint locations and descriptors. Using the approximate keypoint distance (14), we construct the matrix of pairwise keypoint distances. As in the ISOMAP algorithm, we use this matrix to construct a $k = 12$-nearest neighbor graph, which we use to estimate the geodesic between any given pair of query images.

Figure 11(a) demonstrates the promise of this proposed technique. We display the seven (geodesic) nearest neighbors for four different query images, and it is visually clear that the retrieved nearest neighbors are closely semantically related to the query image. For comparison purposes, we performed an identical experiment by computing pairwise image distances using the Local Features method [20] and display the results in Fig. 11(b). Evidently, the KAM approach results in more semantically meaningful groupings than the Local Features method. In the supplementary material, we include a much larger gallery of (geodesic) nearest neigh-

bors for a diverse set of query images. These experimental comparisons are meant to demonstrate that the performance of the approximate keypoint distance seems to be fairly robust in practice, despite the fact that the correlation between the exact and approximate distances for this dataset is not perfect (as discussed in Fig. 4).

Going a step further, given a pair of starting and ending images, we display the intermediate images along the estimated KAM geodesic in Fig. 12. Once again, we observe that the estimated "path" between the photos is both intuitive and interpretable. For example, the images in the bottom row of Fig. 12 can be interpreted as zooming out from the inset sculpture to the cathedral facade. Our method took less than 3 hours to execute in MATLAB.

5.4.2. Statue of Liberty

We repeat the Notre Dame experiment on a database of 2000 images comprising the Statue of Liberty [29] chosen randomly from FlickR. Once again, we extract local image features from each photo and estimate a nearest-neighbor graph using the approximate keypoint distance. Figure 13 illustrates that the estimated geodesics between starting and ending images are again semantically meaningful. For example, the images in the top row of Fig. 13 can be interpreted as zooming in and panning around the face of the monument.

Of course, our manifold-based method does not produce a full 3D reconstruction of the scene and thus cannot be considered as an alternative to the full 3D modeling technique employed in Photo Tourism [6]. Nevertheless, it can be viewed as a new and efficient way to discover intuitive relationships among photographs. These relationships can potentially be used to improve the performance of algorithms for applications like camera localization and multi-view 3D reconstruction.

6. Discussion

Image manifolds have largely been studied from a theoretical standpoint. In this paper, we have taken some initial steps to bridge the chasm between theory and applications. We have advocated the need for improved distance measures that provide meaningful distances between image pairs, and improved image representations that are robust to nuisance variations. To this end, we have proposed an EMD-based metric on local image features that yields a smooth and isometric mapping from the articulation parameter space to the image feature space.

A first key aspect of our approach is its *simplicity*. In contrast with the current state-of-the-art methods in SfM, calculating distances in our framework does not involve complicated physics-based modeling of relationships between images, such as epipolar geometry or multi-view stereo. Instead, we merely exploit the low-dimensional manifold geometry inherent in large image ensembles.

A second key aspect of our approach is its *computational efficiency*. By avoiding explicit correspondence computations between keypoints and image registration, we save significantly on computational complexity. This is reflected in a number of our experiments. The SfM bundler approach [6] greedily establishes correspondences and extracts considerable 3D geometric information from the input images. Yet, it takes several hours, or even days, to produce meaningful results. In contrast, our KAM-based method runs in the order of minutes for data sets of about 150 images and a few hours for a larger dataset of $700+$ images.

The ideas we have developed here can be immediately extended to more general settings. For example, the pyramid match kernel [25] is an efficient, robust similarity measure between image pairs that is tailored to object detection. Such a kernel can conceivably be used to induce interesting geometrical structures on IAMs in the same manner as our EMD-based approach. We have largely focused on affine articulations in the object or camera, hence motivating our choice of SIFT [7] as the feature extraction algorithm $\mathcal{A}$. But this choice can be flexible; for example, a problem involving the manifold of all possible illuminations of an object would likely involve a pose-invariant descriptor. The KAM approach could be extended to such problems in a principled manner, including proving analytical results along the lines of Propositions 1–3.

We have chosen to demonstrate via extensive numerical experiments that the KAM approach offers practical robustness to nuisance phenomena such as background clutter and foreground occlusions. However, modeling such phenomena in a theoretically principled fashion is a very difficult task. Particularly challenging scenarios arise in the adversarial setting, where the nuisance clutter and occlusions are deliberately chosen to be perceptually similar to the actual scene of interest. In such a scenario, large, unpredictable errors in the distance computation (14) are possible. We defer the precise characterization of the performance of the KAM approach in such challenging circumstances to future work.

The primary computational bottleneck in our framework is the calculation of pairwise keypoint distances between images, which scales as $\mathcal{O}\left(M^2\right)$, where M is the number of images. To enable M to scale to tens, or hundreds, of thousands of images or more, we plan to explore the Nyström method [26, 27], which approximates the unknown pairwise distance matrix as low rank and attempts to recover

it from a small set of rows and columns of the matrix. Under the same low-rank assumption, a host of techniques from the matrix completion literature [28] can also potentially be applied to recover the pairwise distance matrix from randomly sampled entries. Recently, adaptive selection schemes have been proposed [30] that show improved performance over random selection strategies. All of these schemes can potentially be deployed in conjunction with our proposed framework.

Acknowledgements

This work was supported by grants NSF CCF-0431150, CCF-0926127, and CCF-1117939; DARPA N66001-11-C-4092 and N66001-11-1-4090; ONR N00014-10-1-0989, N00014-11-1-0714, and N00014-12-10579; AFOSR FA9550-09-1-0432; ARO MURI W911NF-07-1-0185 and W911NF-09-1-0383; and the Texas Instruments Leadership University Program.

References

[1] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.

[2] S. Roweis, L. Saul, Nonlinear dimensionality reduction by local linear embedding, Science 290 (2000) 2323–2326.

[3] D. Donoho, C. Grimes, Hessian Eigenmaps: Locally linear embedding techniques for high dimensional data, Proc. Natl. Acad. Sci. 100 (10) (2003) 5591–5596.

[4] M. Belkin, P. Niyogi, Laplacian Eigenmaps for dimensionality reduction and data representation, Neural Comp. 15 (6) (2003) 1373–1396.

[5] D. Donoho, C. Grimes, Image manifolds which are isometric to Euclidean space, J. Math. Imaging and Vision 23 (1).

[6] N. Snavely, S. M. Seitz, R. Szeliski, Photo tourism: Exploring photo collections in 3D, ACM Trans. Graph. 25 (3) (2006) 835–846.

[7] D. G. Lowe, Distinctive image features from scale-invariant keypoints, Intl. J. Comp. Vision 60 (2) (2004) 91–110.

[8] K. Q. Weinberger, L. K. Saul, Unsupervised learning of image manifolds by semidefinite programming, Intl. J. Comp. Vision 70 (1) (2006) 77–90.

[9] R. R. Coifman, S. Lafon, Diffusion maps, Appl. Comp. Harm. Anal. 21 (1) (2006) 5–30.

[10] M. B. Wakin, D. L. Donoho, H. Choi, R. G. Baraniuk, The multiscale structure of non-differentiable image manifolds, in: SPIE Optics and Photonics, Vol. 5914, 2005, pp. 413–429.

[11] J. Sivic, A. Zisserman, Video Google: A text retrieval approach to object matching in videos, in: IEEE Intl. Conf. Comp. Vision, 2003.

[12] Y. Furukawa, J. Ponce, Accurate, dense, and robust multiview stereopsis, IEEE Trans. Pattern Anal. Mach. Intell. 32 (8) (2010) 1362–1376.

[13] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, Intl. J. Comp. Vision 60 (1) (2004) 63–86.

[14] T. Tuytelaars, K. Mikolajczyk, Local invariant feature detectors: A survey, Found. Trends in Comp. Graphics and Vision 3 (3) (2008) 177–280.

[15] A. P. Witkin, Scale-space filtering, Readings in Comp. Vision 2 (1987) 329–332.

[16] J. M. Morel, G. Yu, Is SIFT scale-invariant?, Inverse Problems and Imag. 5 (1) (2011) 91–110.

[17] C. Wallraven, B. Caputo, A. Graf, Recognition with local features: the kernel recipe, in: IEEE Intl. Conf. Comp. Vision, 2003, pp. 257–264.

[18] J. Zhang, M. Marszałek, S. Lazebnik, C. Schmid, Local features and kernels for classification of texture and object categories: A comprehensive study, Intl. J. Comp. Vision 73 (2) (2007) 213–238.

[19] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in: IEEE Conf. Comp. Vision and Pattern Recog., 2005.

[20] M. Torki, A. Elgammal, Putting local features on a manifold, in: IEEE Conf. Comp. Vision and Pattern Recog., 2010.

[21] Y. Rubner, C. Tomasi, L. J. Guibas, The earth mover's distance as a metric for image retrieval, Intl. J. Comp. Vision 40 (2) (2000) 99–121.

[22] O. Pele, M. Werman, A linear time histogram metric for improved SIFT matching, in: Euro. Conf. Comp. Vision, 2008.

[23] H. Kuhn, The hungarian method for the assignment problem, Naval Research Logistics 2 (1-2) (1955) 83–97.

[24] S. Lyu, Mercer kernels for object recognition with local features, in: IEEE Conf. Comp. Vision and Pattern Recog., 2005.

[25] K. Grauman, T. Darrell, The pyramid match kernel: Discriminative classification with sets of image features, in: IEEE Intl. Conf. Comp. Vision, 2005.

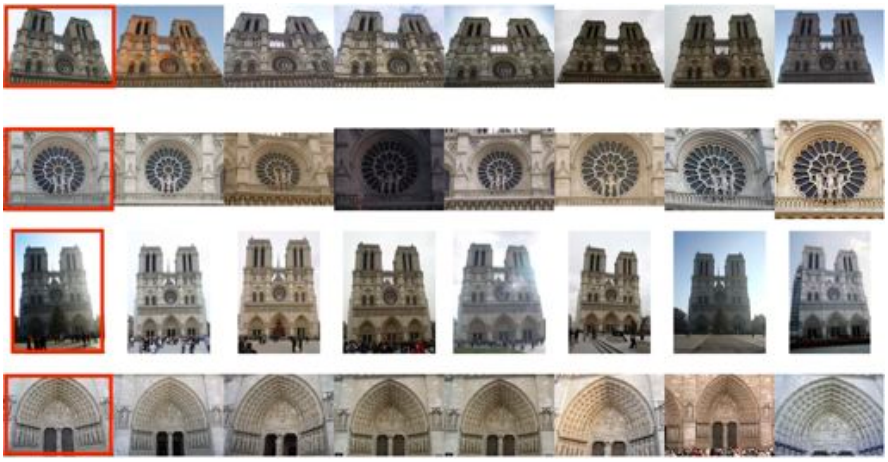
[26] C. Williams, M. Seeger, Using the Nyström method to speed up kernel machines, in: Adv. Neural Inf. Proc. Sys., 2001.

[27] C. Fowlkes, S. Belongie, F. Chung, J. Malik, Spectral grouping using the Nyström method, IEEE Trans. Pattern Anal. Mach. Intell. 26 (2) (2004) 214–225.

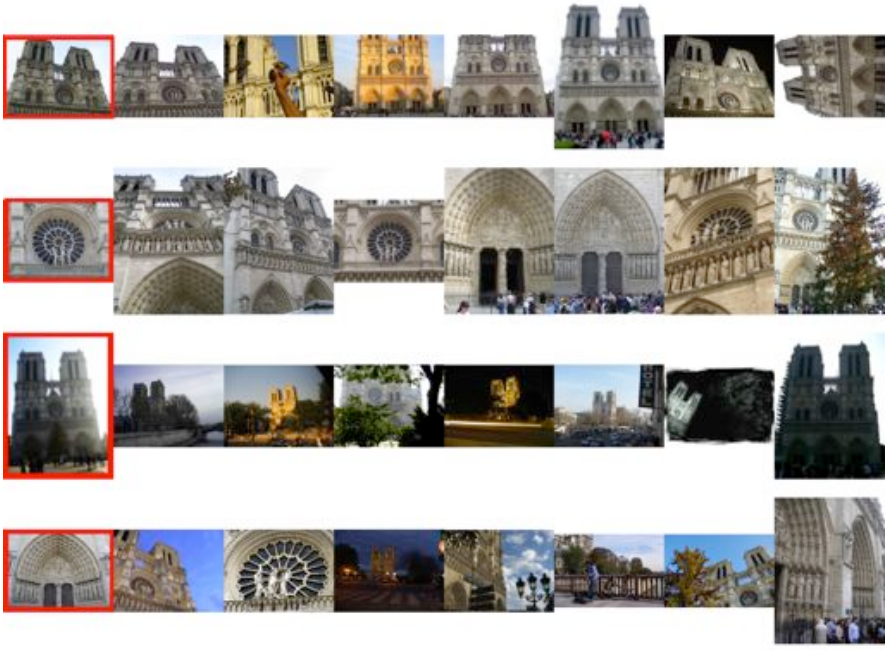
[28] E. Candès, B. Recht, Exact matrix completion via convex optimization, Found. Comp. Math. 9 (6) (2009) 717–772.

[29] X. Li, C. Wu, C. Zach, S. Lazebnik, J. M. Frahm, Modeling and recognition of landmark image collections using iconic scene graphs, in: Euro. Conf. Comp. Vision, 2008.

[30] B. Eriksson, G. Dasarathy, A. Singh, R. Nowak, Active clustering: Robust and efficient hierarchical clustering using adaptively selected similarities, Arxiv preprint arXiv:1102.3887.



(a) KAM nearest neighbors



(b) Local Features nearest neighbors

Figure 11: *Automatic photo organization using (a) our proposed KAM embedding approach and (b) an approached based on Local Features [20]. The leftmost image in each row (marked in red) indicates the query image, and we retrieve the seven geodesic nearest neighbor images for each query image. In contrast to the Local Features approach, the KAM approach provides more semantically meaningful nearest neighbors.*

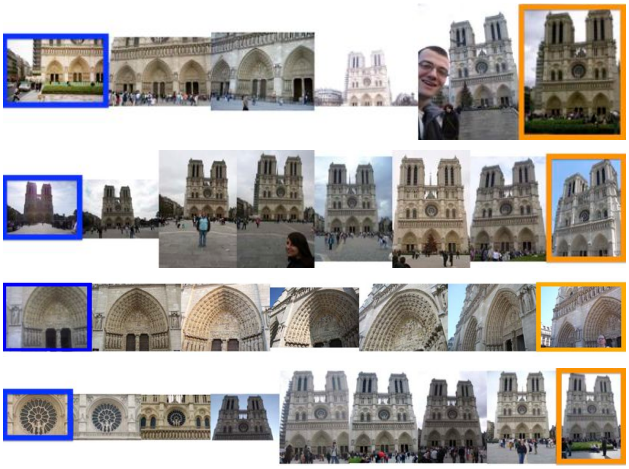


Figure 12: *Geodesic paths between images in the Notre Dame dataset. Shown are images along the estimated geodesic for four different choices of start images (marked in blue) and end images (marked in orange).*

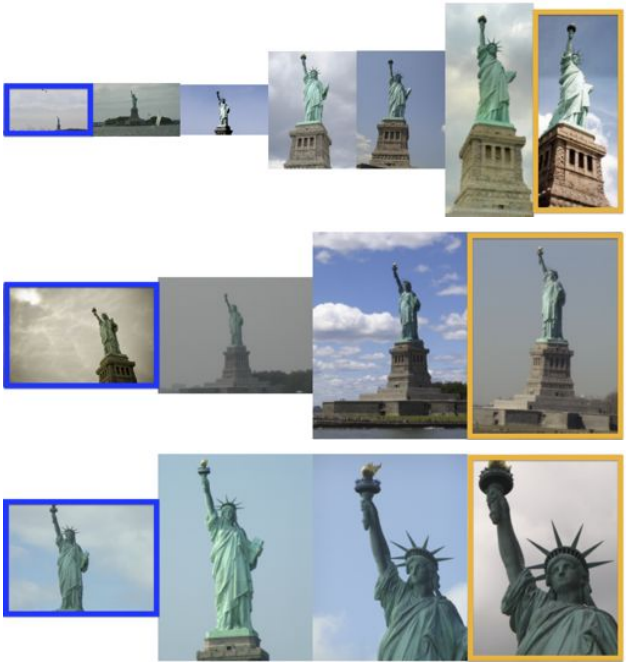


Figure 13: *Geodesic paths between images in the Statue of Liberty dataset. Shown are images along the estimated geodesics for four different choices of start images (marked in blue) and end images (marked in orange).*