Permutation Diffusion Maps with Application to the Image Association Problem in Computer Vision

Deepti Pachauri[†], **Risi Kondor**[§], **Gautam Sargur**[†], **Vikas Singh**[‡†]
[†]Dept. of Computer Sciences, University of Wisconsin–Madison
[‡]Dept. of Biostatistics & Medical Informatics, University of Wisconsin–Madison
[§]Dept. of Computer Science and Dept. of Statistics, The University of Chicago
pachauri@cs.wisc.edu risi@uchicago.edu gautam@cs.wisc.edu
vsingh@biostat.wisc.edu

Abstract

Consistently matching keypoints across images, and the related problem of finding clusters of nearby images, are critical components of various tasks in Computer Vision, including Structure from Motion (SfM). Unfortunately, occlusion and large repetitive structures tend to mislead most currently used matching algorithms, leading to characteristic pathologies in the final output. In this paper we propose a new method, Permutations Diffusion Maps (PDM), and a related new affinity measure, Permutation Diffusion Affinity (PDA), to solve this problem. PDM is inspired by Vector Diffusion Maps, recently introduced by Singer and Wu, and uses ideas from the theory of Fourier analysis on the symmetric group. We show that when dealing with difficult datasets, using PDM as a preprocessing step to existing SfM pipelines can significantly improve results.

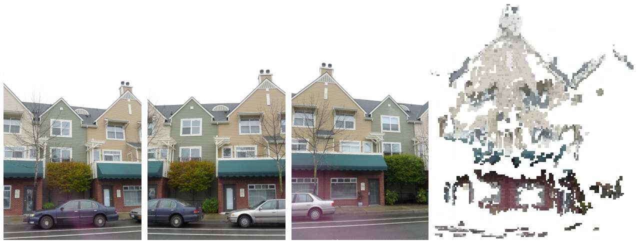
1 Introduction

Structure from motion (SfM) is the task of jointly reconstructing 3D scenes and camera poses from a set of images. Keypoints or features extracted from each image provide correspondences between pairs of images, making it possible to estimate the relative camera pose. This gives rise to an association graph in which two images are connected by an edge if they share a sufficient number of corresponding keypoints, and the edge itself is labeled by the estimated matching between the two sets of keypoints. Starting with these putative image to image associations, one typically uses the so-called bundle adjustment procedure to simultaneously solve for the global camera pose parameters and 3-D scene locations, incrementally minimizing the sum of squares of the re-projection error.

Despite their popularity, large scale bundle adjustment methods have well known limitations. In particular, due to the highly nonlinear nature of the objective function, they can get stuck in bad local minima. Therefore, starting with a good initial matching (i.e., an informative image association graph) is critical. Several papers have studied this behavior in detail [1], and conclude that if one starts the numerical optimization from an incorrect "seed" (i.e., a subgraph of the image associations), the downstream optimization is unlikely to ever recover.

Similar challenges arise in other fields, ranging from machine learning [2] to computational biology. For instance, consider the *de novo* genome assembly problem in computational biology [3]. The goal here is to reconstruct the original DNA sequence from fragments without a reference genome. Because the genome may have many repeated structures, the alignment problem becomes very hard. In general, reconstruction algorithms start with two maximally overlapping sequences, and proceed by selecting subsequents fragment using a process not unlike bundle adjustment, prone to similar issues with local minima [4]. In both cases it would be preferable to have a model that reasons globally over all pairwise information. In this paper, to make our presentation as concrete as possible, we restrict ourselves to describing such an algorithm in the context of Structure from Motion, while understanding that the underlying ideas apply more generally.

Several authors [5, 6, 7] have recently described situations in large scale structure from motion where setting up a good image association graph is difficult, and consequently a direct application of bundle adjustment yields unsatisfactory results. One such situation is when the scene depicted in the images involves a large number of duplicate structures (Figure 1). The preprocessing step in a standard pipeline will match visual features and set up the associations accordingly, but a key underlying assumption in most (if not all) approaches is that we observe only a single instance of any structure. This assumption is problematic when scenes have repeating architectural components or recurring patterns, such as windows, bricks, and so on.



(a) (b)

Figure 1: HOUSE sequence. (a) Representative images. (b) Folded reconstruction by traditional SfM pipeline [8, 9].

In Figure 1a views that look exactly the same do not necessarily represent the same physical structure. Some (or all) points in one image are actually occluded in the other image. Typical SfM methods will not work well when initialized with such image associations, regardless of which type of solver we use. In our example, the resulting reconstruction will be folded (Figure 1b). In other cases [5], we get errors ranging from phantom walls to severely superimposed structures yielding nonsensical reconstructions.

Related Work. The issue described above is variously known in the literature as the SfM disambiguation problem or the data/image association problem in structure from motion. Some of the strategies that have been proposed to mitigate it impose additional conditions, such as in [10, 11, 12, 13, 14, 15], but this also breaks down in the presence of large coherent sets of incorrectly matched pairs. One creative solution in recent work is to use metadata alongside images. "Geotags" or GIS data when available have been shown to be very effective in deriving a better initialization for bundle adjustment or as a post-processing step to stitch together different components of a reconstruction. In [6], the authors suggest using image timestamps to impose a natural association among images, which is valuable when the images are acquired by a single camera in a temporal sequence but difficult to deploy otherwise. Separate from the metadata approach, in controlled scenes with relatively less occlusion, missing correspondences yield important local cues to infer potentially incorrect image pairs [6, 7]. Very recently, [5] formalized the intuition that incorrect feature correspondences result in anomalous structures in the so-called visibility graph of the features. By looking at a measure of local track quality (from local clustering), one can reason about which associations are likely to be erroneous. This works well when the number of points is very large, but the authors of [5] acknowledge that for datasets like those shown in Fig. 1, it may not help much.

In contrast to the above approaches, a number of recent algorithms for the association (or disambiguation) problem argue for *global* geometric reasoning. In [16], the authors used the number of point correspondences as a measure of certainty, which was then globally optimized to find a maximum-weight set of consistent pairwise associations. The authors in [17] seek consistency of epipopolar geometry constraints for triplets, whereas [18] expands it over larger consistent cliques. The procedure in [16] takes into account loops of associations concurrently with a minimal spanning tree over image to image matches. In summary, the bulk of prior work suggests that locally based statistics over chained transformations will run into problems if the inconsistencies are more global in nature. However, even if the objectives used are global, *approximate* inference is not known to be robust to *coherent* noise, which is exactly what we face in the presence of duplicate structures [19].

This paper. If we take the idea of reasoning globally about association consistency using triples or higher order loops to an extreme, it implies deriving the likelihood of a specific image to image association conditioned on *all* other associations. This joint likelihood does not factor and explicit enumeration quickly becomes intractable. Our approach will make the *group* structure of image to image relationships explicit. Similarly to prior approaches, we will also operate on the association graph derived from image pairs but with a key distinguishing feature. The association relationships will now be denoted in terms of a 'certificate', that is, the *transformation* which justifies the relationship. The transformation may denote the pose parameters derived from the correspondences or the matching (between features) itself. Other options are possible — as long as this transformation is a *group action* from one set to the other. If so, we can carry over the intuition of consistency over larger cliques of images desired in existing works and rewrite those ideas as invariance properties of functions defined on the *group*. In particular, when the transformation is a matching, each edge in the graph is a permutation, i.e., a member of the symmetric group $\mathbb{S}_n$, and a generalization of the Laplacian related to the representation theory of $\mathbb{S}_n$ encodes the associations. In this regard, the present paper owes the most to the literature of synchronization problems, specifically [20][21][22][23][24].

The key contribution of this paper is to show that the global inference desired in many existing works falls out nicely as a diffusion process using such a Laplacian. We show promising results demonstrating that for various difficult datasets with large repetitive patterns, results from a simple decomposition procedure are, in fact, competitive with those obtained using sophisticated optimization schemes with/without metadata. Finally, we note that the proposed algorithm can either be used standalone to derive meaningful inputs to a bundle adjustment procedure or as a preprocessing step to other approaches (especially ones that incorporate timestamps and/or GPS data).

2 Synchronization by Vector Diffusion

Consider a collection of m images $\{\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_m\}$ of the same object or scene taken from different viewpoints and possibly under different conditions, and assume that in each image $\mathcal{I}_i$, a keypoint detector has detected n landmarks (keypoints) $\{x_1^i, x_2^i, \ldots, x_n^i\}$. Given two images $\mathcal{I}_i$ and $\mathcal{I}_j$, the landmark matching problem consists of finding pairs of landmarks $x_p^i \sim x_q^j$ (with x_p^i coming from image $\mathcal{I}_i$ and x_q^j coming from $\mathcal{I}_j$) which correspond to the same underlying physical feature.

Assuming that both images contain exactly the same n landmarks, the matching between $\mathcal{I}_i$ and $\mathcal{I}_j$ may be described by the unique permutation $\tau_{ji} \colon \{1, 2, \ldots, n\} \to \{1, 2, \ldots, n\}$ under which $x_p^i \sim x_{\tau_{ji}(p)}^j$. Typically, local image features, such as SIFT descriptors, can provide an initial guess for each τ_{ji}, but by itself each of these individual image-to-image matchings is highly error prone, especially in the presence of occlusion and repetitive structures. A major clue to correcting these errors is the constraint that matchings must be consistent, i.e., if τ_{ji} tells us that x_p^i corresponds to x_q^j, and τ_{kj} tells us that x_q^j corresponds to x_r^k, then the permutation τ_{ki} between $\mathcal{I}_i$ and $\mathcal{I}_k$ should assign x_p^i to x_r^k. Mathematically, this is a reflection of the fact that, defining the product of two permutations σ_1 and σ_2 in the usual way as

$$\sigma_3 = \sigma_2 \sigma_1 \qquad \Longleftrightarrow \qquad \sigma_3(i) = \sigma_2(\sigma_1(i)) \qquad i = 1, 2, \ldots, n,$$

the $n!$ different permutations of $\{1, 2, \ldots, n\}$ form a group. This group is called the symmetric group of order n and is denoted $\mathbb{S}_n$. In group theoretical notation the consistency conditions reduce to requiring that given any three images $\mathcal{I}_i, \mathcal{I}_j$ and $\mathcal{I}_k$, the relative matchings between them must satisfy $\tau_{kj}\tau_{ji} = \tau_{ki}$. An equivalent condition is that it must be possible to associate to each $\mathcal{I}_i$ a "base permutation" σ_i so that $\tau_{ji} = \sigma_j \sigma_i^{-1}$ for any (i, j) pair. Thus, the problem of finding a consistent set of τ_{ji}'s is reduced to finding the m base permutations $\sigma_1, \ldots, \sigma_m$.

Problems of this general form, where given some (finite or continuous) group G, one must estimate a matrix $(g_{ji})_{j,i=1}^m$ of group elements obeying $g_{kj} g_{ji} = g_{ki}$ are called synchronization problems. Starting with the seminal work of Singer et al. [20][21] on synchronization over the rotation group for aligning images in cryo-EM, followed by synchronization over the Euclidean group [25], and most recently synchronization over $\mathbb{S}_n$ for matching landmarks [23][24], such problems have recently generated a lot of interest. Some of the newest and most promising approaches involve semi-definite programming [15][24][26].

In the context of synchronizing three dimensional rotations for cryo-EM, Singer and Wu [22] proposed a particularly elegant formalism, called Vector Diffusion Maps, which conceives of synchronization as diffusing the base rotation Q_i from each image to its neighbors. However, unlike in ordinary diffusion, as Q_i diffuses to $\mathcal{I}_j$, the observed O_{ji} relative rotation of $\mathcal{I}_j$ to $\mathcal{I}_i$ changes Q_i to $O_{ji} Q_i$. If all the $(O_{ji})_{i,j}$ observations were perfectly synchronized, then no matter what path

$i \to i_1 \to i_2 \to \ldots \to j$ we took from i to j, the resulting rotation $O_{j,i_p} \ldots O_{i_2,i_1} O_{i_1,i} Q_i$ would be the same. However, if some (in many practical cases, the majority) of the O_{ji}'s are incorrect, then different paths from one vertex to another contribute different rotations, which one then needs to average in some appropriate sense.

A natural choice for the loss that describes the extent to which the $Q_1, \ldots, Q_m$ imputed base rotations (playing the role of the σ_i's in the permutation case) satisfy the O_{ji} observations is

$$\mathcal{E}(Q_1, \ldots, Q_m) = \frac{1}{2} \sum_{i,j=1}^{m} w_{ij} \| Q_j - O_{ji} Q_i \|^2_{\text{Frob}} = \frac{1}{2} \sum_{i,j=1}^{m} w_{ij} \| Q_j Q_i^\top - O_{ji} \|^2_{\text{Frob}}, \qquad (1)$$

where the w_{ij} edge weight descibes our confidence in rotation O_{ji}. A crucial observation is that this loss can be rewritten in the form $\mathcal{E}(Q_1, \ldots, Q_m) = V^\top \mathcal{L} V$, where

$$V = \begin{pmatrix} Q_1 \\ \vdots \\ Q_m \end{pmatrix}, \qquad \mathcal{L} = \begin{pmatrix} d_i I & -w_{1,2} O_{1,2} & \ldots & -w_{1,m} O_{1,m} \\ \vdots & \ddots & & \vdots \\ -w_{m,1} O_{m,1} & -w_{m,2} O_{m,2} & \ldots & d_m I \end{pmatrix}, \qquad (2)$$

and $d_i = \sum_{j \neq i} w_{ij}$. Note that since $w_{ij} = w_{ji}$, and $O_{ij} = O_{ji}^{-1} = O_{ji}^\top$, the matrix $\mathcal{L}$ is symmetric. Furthermore, the above is exactly analogous to the way in which in spectral graph theory, (see, e.g.,[27]) the functional $\mathcal{E}(f) = \frac{1}{2} \sum_{i,j} w_{ij} (f(i) - f(j))^2$ describing the "smoothness" (with respect to the graph topology) of a function f defined on the vertices of a graph can be written as $f^\top L f$ in terms of the usual graph Laplacian

$$L_{ij} = \begin{cases} -w_{ij} & i \neq j \\ \sum_{k \neq i} w_{ik} & i = j. \end{cases}$$

As it is well known, constraining f to have unit norm and excluding the subspace of constant functions, the function minimizing $\mathcal{E}(f)$ is the eigenvector of L with (second) smallest eigenvalue. Analogously, in synchronizing rotations, the steady state of the diffusion system, which minimizes (1), can be computed by forming the $3m \times 3$ dimensional matrix V from the 3 lowest non-zero eigenvalue eigenvectors of $\mathcal{L}$, and appropriately rounding each 3×3 block V_i of V to the nearest orthogonal matrix Q_i. The resulting array $(Q_j Q_i^\top)_{i,j}$ of imputed relative rotations is guaranteed to be consistent, and minimizes the loss (1).

3 Permutation Diffusion

Its elegance notwithstanding, the vector diffusion formalism of the previous section seems ill suited to our present purposes of improving the SfM pipeline for two reasons: (1) synchronizing over $\mathbb{S}_n$, which is a finite group, seems much harder than synchronizing over the continuous group of rotations; (2) rather than getting an actual synchronized array of matchings, what is critical to SfM is to estimate the association graph that captures the extent to which any two images are related to one-another. The main contribution of the present paper is to show that both of these problems have natural solutions in the formalism of group representations.

Our first key observation (already alluded to in [21]) is that the critical step of rewriting the loss (1) in terms of the Laplacian (2) does not depend on any special properties of the rotation group other than the facts that (a) rotation matrices are unitary (in fact, orthogonal) (b) if we follow one rotation by another, their matrices simply multiply. In general, for any group G, a complex valued function $\rho\colon G \to \mathbb{C}^{d_\rho \times d_\rho}$ which satisfies $\rho(g_2 g_1) = \rho(g_2)\rho(g_1)$ is called a representation of G. The representation is unitary if $\rho(g^{-1}) = (\rho(g))^{-1} = \rho^\dagger$, where $M^\dagger$ denotes the Hermitian conjugate (conjugate transpose) of M. Thus, we have the following proposition.

Proposition 1. *Let G be any compact group with identity e and $\rho\colon G \to \mathbb{C}^{d_\rho \times d_\rho}$ be a unitary representation of G. Then given an array of possibly noisy and unsynchronized group elements, $(g_{ji})_{i,j}$, and corresponding positive confidence weights $(w_{ji})_{i,j}$, the synchronization loss (assuming $g_{ii} = e$ for all i)*

$$\mathcal{E}(h_1, \ldots, h_m) = \frac{1}{2} \sum_{i,j=1}^{m} w_{ji} \left\| \rho(h_j h_i^{-1}) - \rho(g_{ji}) \right\|^2_{\mathit{Frob}} \qquad h_1, \ldots, h_m \in G$$

can be written in the form $\mathcal{E}(h_1,\dots,h_m)=V^\dagger\mathcal{L}V$, where

$$V=\begin{pmatrix}\rho(h_1)\\ \vdots\\ \rho(h_m)\end{pmatrix},\qquad \mathcal{L}=\begin{pmatrix} d_i\, I & -w_{1,2}\,\rho(g_{1,2}) & \dots & -w_{1,m}\,\rho(g_{1,m})\\ \vdots & & \ddots & \vdots\\ -w_{m,1}\,\rho(g_{m,1}) & -w_{m,2}\,\rho(g_{m,2}) & \dots & d_m\, I\end{pmatrix}. \quad (3)$$

To synchronize matchings between images using this proposition, one plugs in the approriate unitary representation of the symmetric group. The simplest choice is the so-called defining representation, whose elements are the familiar permutation matrices

$$\rho_{\text{def}}(\sigma)=P(\sigma)\qquad [P(\sigma)]_{q,p}=\begin{cases}1 & \sigma(p)=q\\ 0 & \text{otherwise,}\end{cases}$$

since the corresponding loss function is

$$\mathcal{E}(\sigma_1,\dots,\sigma_m)=\frac{1}{2}\sum_{i,j=1}^{m} w_{ji}\,\|P(\sigma_j\sigma_i^{-1})-P(\tau_{ji})\|^2_{\text{Frob}}. \quad (4)$$

The squared Frobenius norm in this expression simply counts the number of mismatches between the observed but noisy permutation τ_{ji}, and the inferred permutation $\sigma_j\sigma_i^{-1}$. For this choice of ρ, letting $P_i := P(\sigma(i))$ and $P^{\text{obs}}_{ji} := P(\tau_{ji})$, $\mathcal{E}(\sigma_1,\dots,\sigma_m)=V^\top\mathcal{L}V$, with

$$V=\begin{pmatrix}P_1\\ \vdots\\ P_m\end{pmatrix},\qquad \mathcal{L}=\begin{pmatrix} d_i\, I & -w_{1,2}\,P^{\text{obs}}_{1,2} & \dots & -w_{1,m}\,P^{\text{obs}}_{1,m}\\ \vdots & & \ddots & \vdots\\ -w_{m,1}\,P^{\text{obs}}_{m,1} & -w_{m,2}\,P^{\text{obs}}_{m,2} & \dots & d_m\, I\end{pmatrix}. \quad (5)$$

Consequently, just as in the rotation case, synchronization over $\mathbb{S}_n$ can be solved by forming V from the first $d_{\rho_{\text{def}}}=n$ lowest eigenvectors of $\mathcal{L}$, and extracting each P_i from its i'th $n\times n$ block, V_i. Here we must take a little care because unless the τ_{ji}'s are already synchronized, it is not a priori guaranteed that the resulting block will be a valid permutation matrix. Therefore, analogously to the procedure described in [23], we first multiply V_i by $V_1^\top$, and then use a linear assignment procedure to find the permutation $\widehat{\sigma}_i$, whose permutation matrix is closest to $V_iV_1^\top$. The resulting algorithm we call Synchronization by Permutation Diffusion.

4 Uncertain Matches and Permutation Diffusion Affinity

The limitation of our framework, as described so far, is the assumption that each keypoint in each image will have a single counterpart in every other image that the local matching procedure with some error can identify. In realistic scenarios this is far from satisfied, due to occlusion, repetitive structures, and noisy detections. Most algorithms, including [24] and [23], deal with the problem simply by turning the P_{ij} block in (5) into a weighted sum of all possible permutations. For example, if landmarks number $1\dots 20$ are present in both images, but landmarks $21\dots 40$ are not, then the P_{ij} block in (5) will have a corresponding 20×20 block of all ones, rescaled by a factor of $1/20$.

This approach effectively amounts to replacing τ_{ji} by an appropriate *distribution* $t_{ji}(\tau)$ over matchings. Correspondingly, when we form V from the first d_ρ eigenvectors of $\mathcal{L}$, each resulting V_i block will stand for a distribution $p_i(\sigma)$, rather than a single base permutation σ_i. Moreover, if some set of k landmarks $U=(u_1,\dots,u_k)$ are occluded in $\mathcal{I}_i$, then t_{ij} (for any j) will be agnostic to their assignment, and consequently p_i will be invariant to what is mapped to $u_1,\dots,u_k$. Let $\sigma\sim_U\sigma'$ denote the relation that two permutations σ and σ' differ *only* in what numbers they map to $u_1,\dots,u_k$, but fully agree on what they assign to any landmark *not* in U (i.e., $\sigma(i)=\sigma(j)\ \ \forall\, i\not\in U$). Clearly, $\sim_U$ is an equivalence relation on $\mathbb{S}_n$, and it is not difficult to see that letting μ_U be some reference permutation that maps $1\mapsto u_1,\dots,k\mapsto u_k$, and $\mathbb{S}_k$ be the subgroup of permutations that permute $1,2,\dots,k$ amongst themselves but leave $k+1,\dots,n$ fixed, the equivalence classes of $\sim_U$ are the sets

$$\mu_U\mathbb{S}_k\nu := \{\mu_U\gamma\nu \mid \gamma\in\mathbb{S}_k\}\qquad \nu\in\mathbb{S}_n. \quad (6)$$

These sets are called (two-sided) $\mathbb{S}_k$–cosets. Note that while $|\mathbb{S}_n|=n!$, there are only $n!/k!$ distinct equivalence classes, so not all possible values of ν yield a distinct coset.

What is important is that uncertainty in the synchronization process with respect to a given set of landmarks $\{u_1,\dots,u_k\}$ (typically due to occlusion) has a clear algebraic signature, namely the inferred p_i being constant on each of the cosets in (6). Conversely, if we find that p_i is constant on these cosets, that is a strong indication that $u_1, \ldots, u_k$ are occluded, which is an important clue to estimating $\mathcal{I}_i$'s viewpoint, sometimes even more informative than the synchronized matchings themselves.

The invariance structure of p_i is most easily detected from its so-called autocorrelation function

$$a_i(\sigma) = \sum_{\omega \in \mathbb{S}_n} p_i(\sigma\omega)\, p_i(\omega). \tag{7}$$

Clearly, (7) attains its maximum at the identity permutation, where $a_i(e) = \sum_{\omega \in \mathbb{S}_n} p_i(\omega)^2$. However, when p_i has invariances, the same maximum will be attained over a wider plateau of permutations. Note, in particular, that ω and $\sigma\omega$ always fall in the same $\mu_U\, \mathbb{S}_k\, \nu$ coset when $\sigma \in \mu_U\, \mathbb{S}_k\, \mu_U^{-1}$. Therefore, if p_i happens to be a function that is constant on $\mu_U\, \mathbb{S}_k\, \nu$ cosets, then any $\sigma \in \mu_U\, \mathbb{S}_k\, \mu_U^{-1}$ will maximize $a_i(\sigma)$.

Of course, in synchronization problems p_i is not directly accessible to us, rather we only have access to the weighted sum $\widehat{p}_i(\rho) := \sum_{\sigma \in \mathbb{S}_n} p_i(\sigma)\, \rho(\sigma) = V_i\, V_1^\top$. Recent years have seen the emergence of a number of applications of a generalized notion of Fourier transformation on the symmetric group, which, given a function $f \colon \mathbb{S}_n \to \mathbb{R}$, is defined

$$\widehat{f}(\lambda) = \sum_{\sigma \in \mathbb{S}_n} f(\sigma)\, \rho_\lambda(\sigma), \qquad\qquad \lambda \vdash n,$$

where the ρ_λ are special, so-called irreducible, representations of $\mathbb{S}_n$, indexed by the λ integer partitions. Due to space restrictions, we leave the details of this construction to the literature, see, e.g., [28, 29, 30]. Suffice to say that while $\widehat{p}_i(\rho)$ is not exactly a Fourier component of p_i, it can be expressed as a direct sum of Fourier components

$$\widehat{p}_i(\rho) = C^\dagger \Big[\bigoplus_{\lambda \in \Lambda} \widehat{p}_i(\lambda) \Big] C$$

for some unitary matrix C that is effectively just a basis transform. One of the properties of the Fourier transform is that if h is the cross-correlation of two functions f and g (i.e., $h(\sigma) = \sum_{\mu \in \mathbb{S}_n} f(\sigma\mu)\, g(\mu)$), then $\widehat{h}(\lambda) = \widehat{f}(\lambda)\, \widehat{g}(\lambda)^\dagger$. Consequently, assuming that V_1 has been normalized to ensure that $V_1^\top V_1 = I$, and using the fact that in our setting all matrices are real,

$$\widehat{a}_i(\rho) := C^\dagger \Big[\bigoplus_{\lambda \in \Lambda} \widehat{a}_i(\lambda) \Big] C = C^\dagger \Big[\bigoplus_{\lambda \in \Lambda} \widehat{p}_i(\lambda)\, \widehat{p}_i(\lambda)^\dagger \Big] C = (V_i\, V_1^\top)\, (V_i\, V_1^\top)^\top = V_i\, V_i^\top$$

is an easily computable matrix that captures essentially all the coset invariance structure encoded in the inferred distribution p_i.

To compute an affinity score between two images $\mathcal{I}_i$ and $\mathcal{I}_j$ reflecting how many occluded landmarks they share, it remains to compare their coset invariance structures, for example, by computing $(\sum_{\sigma \in \mathbb{S}_n} a_i(\sigma)\, a_j(\sigma))^{1/2}$. Omitting certain multiplicative constants arising in the inverse Fourier transform, again using the correlation theorem, one finds that this reduces to

$$\Pi(i,j) = \operatorname{tr}\,(V_i\, V_i^\top V_j\, V_j^\top)^{1/2}, \tag{8}$$

which we call Permutation Diffusion Affinity (PDA). Remarkably, PDA is closely related to the notion of diffusion similarity derived in [22] for rotations, using entirely different, differential geometric tools. Our experiments show that PDA is surprisingly informative about the actual distance between image viewpoints in physical space, and, as easy it is to compute, can greatly improve the performance of the SfM pipeline.

5 Experiments

Our experiments focus on challenging image association problems from the literature, where geometric ambiguities due to large duplicate structures are present in up to 50% of the matches, so even sophisticated SfM pipelines run into difficulties [6]. Rather than replacing the standard SfM pipeline with Permutations Diffusion Maps (PDM) altogether, our general approach is to use PDM as a preprocessing step to compute (8) for every image pair, and then feed these PDA scores into the SfM pipeline to improve its performance. More information on the experiments, including videos of 3D reconstructions, and an additional experiment on scene summarization rather than SfM [31], can be found on the project website: `http://pages.cs.wisc.edu/~pachauri/pdm/`.

In the SfM experiments we used PDM to generate an image match matrix which is then fed to a state-of-the-art SfM pipeline for 3D reconstruction [8, 9]. The baseline was a Bundle Adjustment procedure which uses visual features for matching and has a built-in heuristic outlier removal module. Several other papers have used a similar comparisons [6]. For each dataset, SIFT was used to detect and characterize landmarks [32, 33]. We compute putative pairwise matchings $(\tau_{ij})_{i,j=1}^{m}$ by solving $\binom{m}{2}$ linear independent assignments [34] based on their SIFT features. The permutation matrix representation is used for putative matchings $(\tau_{ij})_{i,j=1}^{m}$ as in (5). Here, n is relative large, on the order of 1000. Ideally, n is the total number of distinct keypoints in the 3D scene, but is not directly observable, so we set n to be the maximum number of keypoints detected in any single image in the dataset. Eigenvector based procedure computes weighted affinity matrix. We used a binary match matrix as the input to an SfM library [8, 9]. Note that we only provide this library the image association hypotheses, leaving all other modules unchanged. With (potentially) good image association information, the SfM modules can sample landmarks more densely and perform bundle adjustment, leaving everything else unchanged. The baseline 3D reconstruction is performed using the same SfM pipeline without intervention.

The "HOUSE" sequence has three instances of similar looking houses (Figure 1). The diffusion process accumulates evidence and eventually provides strongly connected images in the data association matrix (Figure 2a). Warm colors correspond to high affinity between pairs of images. The binary match matrix was obtained by applying a threshold on the weighted matrix (Figure 2b). We used this matrix to define the image matching for feature tracks. This means that features are *only* matched between images that are connected in our matching matrix. The SfM pipeline was given these image matches as a hypotheses to explain how the images are "connected". The resulting reconstruction correctly gives three houses (Figure 2c). In contrast, the same SfM pipeline when allowed to track features automatically with an outlier removal heuristic resulted in a folded reconstruction (Figure 1b). One may ask if more specialized heuristics will do better, such as time stamps, as suggested in [6]. However, experimental results in [5] and elsewhere strongly suggests that these datasets still remain challenging.

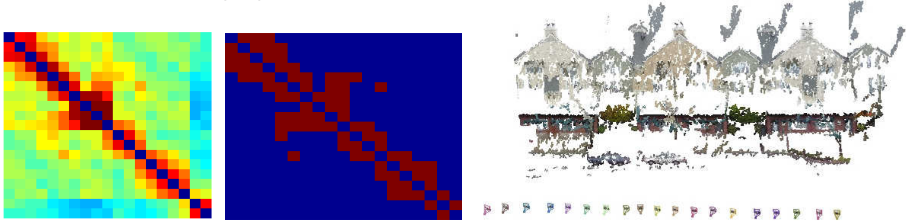

(a) (b) (c)

Figure 2: House sequence: (a) Weighted image association matrix. (b) Binary image match matrix. (c) PDM dense reconstruction.

The "CUP" dataset has multiple images of a 180 degree symmetric cup from all sides (Figure 3a). PDM reveals a strongly connected component along the diagonal for this dataset, shown in warm colors in Figure 3b. Our global reasoning over the space of permutations substantially mitigates coherent errors. The binary match matrix was obtained by thresholding the weighted matrix (Figure 3c). As is evident from the reconstructions, the baseline method only reconstruct a "half cup". Due to the structural ambiguity, it also concludes that the cup has two handles (Figure 4b). In contrast, the PDM reconstruction gives a perfect reconstruction of the full cup with a single handle (Figure 4a).

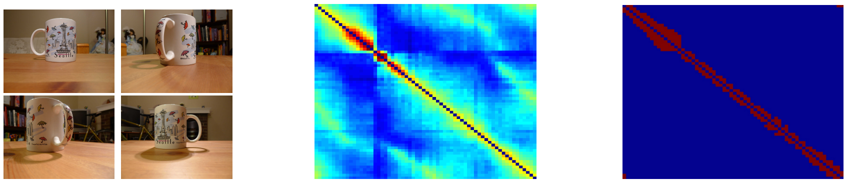
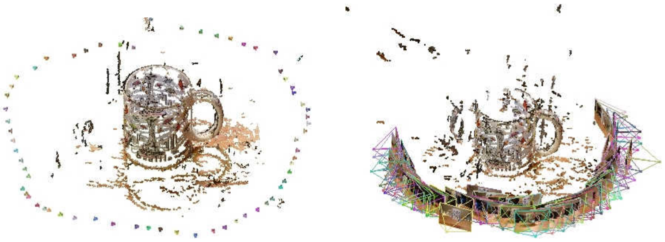

(a) (b) (c)

Figure 3: (a) Representative images from CUP dataset. (b) Weighted data association matrix. (c) Binary data association matrix.

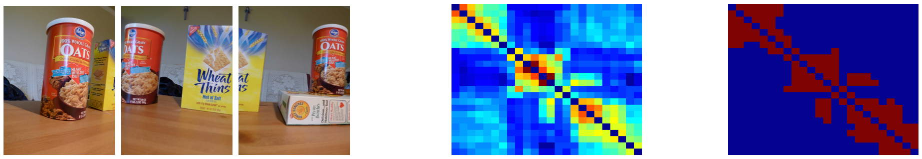
The "OAT" dataset contains two instances of a red oat box, one on the left of a box of "Wheat Thins", and another on the right (Figure 5a). The PDM weighted match matrix and binary match



(a) (b)

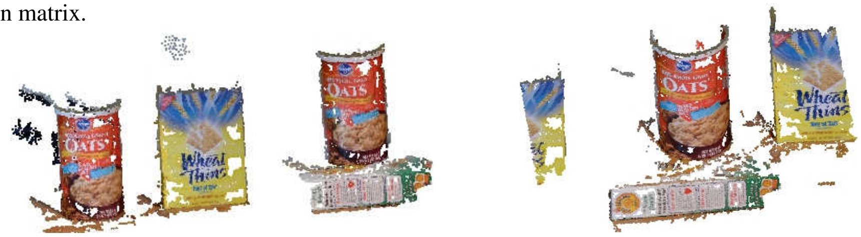
Figure 4: CUP dataset. (a) PDM dense reconstruction. (b) Baseline dense reconstruction.

matrix successfully discover strongly connected components, (Figures 5b, 5c). The baseline method confused the two oat boxes as one, and reconstructs only a single box, (Figure 6b). Moreover, the structural ambiguity splits the Wheat Thins into two pieces. On the other hand, PDM gives a nice reconstruction of the two oat boxes with the entire wheat things in the middle, Figure 6(a). Several more experiments (with videos), can be found on the project website.



(a) (b) (c)

Figure 5: (a) Representative images from OAT dataset. (b) Weighted data association matrix. (c) Binary data association matrix.



(a) (b)

Figure 6: OAT dataset. (a) PDM dense reconstruction. (b) Baseline dense reconstruction.

6 Conclusions

Inspired by the Vector Diffusion formalism of [22], we have proposed a new algorithm called Permutation Diffusion Maps for solving permutation synchronization problems, and an associated new affinity measure called Permutation Diffusion Affinity (PDA). Experiments show that the latter, in particular, can significantly improve the quality of Structure from Motion reconstructions of difficult scenes. Interestingly, PDA has an interpretation in terms of the inner product between two autocorrelation functions expressed in Fourier space, which, we believe, is a new approach to detecting hidden symmetries, with many potential applications even outside the realm of permutation problems.

Acknowledgments

This work was supported in part by NSF–1320344, NSF–1320755, and funds from the University of Wisconsin Graduate School. We thank Charles Dyer, Li Zhang, Amit Singer and Qixing Huang for helpful discussions and suggestions.

References

[1] D. Crandall, A. Owens, N. Snavely, and D. P. Huttenlocher. Discrete-continuous optimization for large-scale structure from motion. In *CVPR*, 2011.

[2] A. Nguyen, M. Ben-Chen, K. Welnicka, Y. Ye, and L. Guibas. An optimization approach to improving collections of shape maps. In *Computer Graphics Forum*, volume 30, 2011.

[3] R. Li, H. Zhu, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20, 2010.

[4] M. Pop, S. L. Salzberg, and M. Shumway. Genome sequence assembly: Algorithms and issues. *IEEE Computer*, 35, 2002.

[5] K. Wilson and N. Snavely. Network principles for SfM: Disambiguating repeated structures with local context. In *ICCV*, 2013.

[6] R. Roberts, S. Sinha, R. Szeliski, and D. Steedly. Structure from motion for scenes with large duplicate structures. In *CVPR*, 2011.

[7] N. Jiang, P. Tan, and L. F. Cheong. Seeing double without confusion: Structure-from-motion in highly ambiguous scenes. In *CVPR*, 2012.

[8] C. Wu. Towards linear-time incremental structure from motion. In *3DTV-Conference, International Conference on*, 2013.

[9] C. Wu, S. Agarwal, B. Curless, and S. M. Seitz. Multicore bundle adjustment. In *CVPR*, 2011.

[10] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or ”how do I organize my holiday snaps?”. In *ECCV*. 2002.

[11] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3D. In *ACM transactions on graphics (TOG)*, volume 25, 2006.

[12] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *CVPR*, 2007.

[13] M. Havlena, A. Torii, J. Knopp, and T. Pajdla. Randomized structure from motion based on atomic 3d models from camera triplets. In *CVPR*, 2009.

[14] S. N. Sinha, D. Steedly, and R. Szeliski. A multi-stage linear approach to structure from motion. In *Trends and Topics in Computer Vision*. 2012.

[15] O. Ozyesil, A. Singer, and R. Basri. Camera motion estimation by convex programming. *CoRR*, 2013.

[16] O. Enqvist, F. Kahl, and C. Olsson. Non-sequential structure from motion. In *ICCV Workshops*, 2011.

[17] C. Zach, A. Irschara, and H. Bischof. What can missing correspondences tell us about 3D structure and motion? In *CVPR*, 2008.

[18] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, 2010.

[19] V. M. Govindu. Robustness in motion averaging. In *Computer Vision–ACCV 2006*, pages 457–466. Springer, 2006.

[20] A. Singer and Y. Shkolnisky. Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming. *SIAM Journal on Imaging Sciences*, 4, 2011.

[21] A Singer. Angular synchronization by eigenvectors and semidefinite programming. *Applied and computational harmonic analysis*, 30, 2011.

[22] A. Singer and H.-T. Wu. Vector diffusion maps and the connection Laplacian. *Communications of Pure and Applied Mathematics*, 2011.

[23] D. Pachauri, R. Kondor, and V. Singh. Solving the multi-way matching problem by permutation synchronization. *NIPS*, 2013.

[24] Qixing Huang and Leonidas Guibas. Consistent shape maps via semidefinite programming. *Computer Graphics Forum*, 2013.

[25] M. Cucuringu, Y. Lipman, and A. Singer. Sensor network localization by eigenvector synchronization over the Euclidean group. *ACM Transactions on Sensor Networks (TOSN)*, 8, 2012.

[26] Y. Chen, L. Guibas, and Q. Huang. Near-optimal joint object matching via convex relaxation. In *ICML*, 2014.

[27] F. R. K. Chung. *Spectral graph theory (CBMS regional conference series in mathematics, No. 92)*. American Mathematical Society, 1996.

[28] J. Huang, C. Guestrin, and L. Guibas. Fourier theoretic probabilistic inference over permutations. *JMLR*, 2009.

[29] R. Kondor. A Fourier space algorithm for solving quadratic assignment problems. In *SODA*, 2010.

[30] D. Rockmore, P. Kostelec, W. Hordijk, and P. F. Stadler. Fast Fourier transforms for fitness landscapes. *Appl. and Comp. Harmonic Anal.*, 2002.

[31] S. Zhu, L. Zhang, and B. M Smith. Model evolution: An incremental approach to non-rigid structure from motion. In *CVPR*, 2010.

[32] D.G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60, 2004.

[33] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60, 2004.

[34] H.W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2, 1955.