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

Anonymous Author(s)

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. The same problem, albeit in slightly different form, also occurs in statistical genomics, where partial fragments of the genome need to be globally assembled using pairwise information. In this paper we introduce a new method, Permutations Diffusion Maps (PDM), to solve the matching problem, as well as a related new affinity measure, derived using ideas from harmonic analysis on the symmetric group. We show that just by using it as a preprocessing step to existing SfM pipelines, PDM can greatly improve reconstruction quality on difficult datasets.

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 reprojection error.

Despite their popularity, large scale bundle adjustment methods have well known limitations. In particular, given the highly non-linear nature of the objective function, they can get stuck in a 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 sub-graph of the image associations), the downstream optimization is unlikely to ever recover.

Similar challenges arise commonly 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 sequence from fragments of DNA 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 the next fragment using a similar criterion. This procedure is bound to run into the same issues that are highlighted above [4]. It will be useful to have a model that reasons globally over all pairwise information to provide a more robust metric for association. The efficacy of global reasoning will largely depend on the richness of the representation used for
encoding putative pairwise information. The choice of representation is specific to the underlying
application. To make our exposition concrete, we will use SfM as the main motivating application,
and putative assignments between image pairs will give pairwise information, to describe and later
evaluate the performance of our proposed method.

To address the association problem in large scale structure from motion, recent efforts [5,
6, 7] have identified situations where setting up a good image association graph is partic-
ularly difficult, and therefore a direct application of bundle adjustment yields highly unsatis-
factory results. Consider a scene with a large number of duplicate structures (Fig. 1).
The preprocessing step in a stan-
dard pipeline will match visual
features and set up the asso-
ciations accordingly. A key
underlying assumption in most
(if not all) approaches is that
we observe only a single in-
stance of any structure. This
assumption is problematic where
scenes have numerous archi-
tectural components or recur-
rings patterns, such as windows,
bricks, and so on.

For example, Figure 1(a) shows an outdoor scene with three similar houses. 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 decide to use. In our example,
the resulting reconstruction will be folded (Fig. 1(b)). 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 dis-
ambiguation 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
[10, 11, 12, 13, 14], 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 re-
construction. 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 se-
quence 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 poten-
tially 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 asso-
ciations 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 disam-
biguation) problem argue for global geometric reasoning. In [15], 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 [16] seek consistency of
epipolar geometry constraints for triplets, whereas [17] expands it over larger consistent cliques.
The procedure in [15] 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 [18].

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

![Figure 1: HOUSE sequence. (a) Representative images. (b) Folded reconstruction by traditional SfM pipeline [8, 9].](image-url)
association conditioned on all other associations. The maximum likelihood expression does not factor out easily and explicit enumeration quickly becomes intractable. Our approach will make the group structure of image to image relationships explicit. 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. As an example, when the transformation is a matching, each edge in the graph is a permutation, i.e., a member of the symmetric group, $S_n$. It follows then that a special form of the Laplacian of this graph, derived from the representation theory of the group under consideration, encodes the symmetries of the functions on the group.

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 pre-conditioner to other approaches (especially, ones that incorporate timestamps and/or GPS data).

## 2 Synchronization

Consider a collection of $m$ images $\{I_1,I_2,\ldots,I_m\}$ of the same object or scene taken from different viewpoints and possibly under different conditions, and assume that a keypoint detector module has detected exactly $n$ landmarks (keypoints) $\{x_1, x_2, \ldots, x_n\}$ in each $I_i$. Given two images $I_i$ and $I_j$, the landmark matching problem consists of finding pairs of landmarks $x_p^i \sim x_q^j$ in the two images which correspond to the same physical feature. This is a critical component of several classical Computer Vision tasks, including Structure from Motion.

Assuming that both images contain exactly the same $n$ landmarks, the matching between $I_i$ and $I_j$ can be described by a permutation $\tau_{ji} : \{1, 2, \ldots, n\} \rightarrow \{1, 2, \ldots, n\}$ under which $x_p^i \sim x_{\tau_{ji}}(p)^j$. An initial guess for the $\tau_{ji}$ matchings is usually provided by local image features, such as SIFT descriptors. However, these matchings individually are very much prone to error, 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 $\tau_{ji}$ tells us that $x_p^i$ corresponds to $x_q^j$, and $\tau_{kj}$ tells us that $x_{\tau_{kj}}(p)^k$ corresponds to $x_{\tau_{kj}}(q)^k$, then the permutation $\tau_{ki}$ between $I_i$ and $I_k$ must assign $x_p^i$ to $x_{\tau_{ki}}(p)^k$.

Mathematically, this is a reflection of the fact that if we define the product of two permutations $\sigma_1$ and $\sigma_2$ in the usual way as

$$\sigma_3 = \sigma_2 \sigma_1 \iff \sigma_3(i) = \sigma_2(\sigma_1(i)) \quad i = 1, 2, \ldots, n,$$

then the $n!$ different permutations of $\{1, 2, \ldots, n\}$ form a group. This group is called the symmetric group of order $n$ and denoted $S_n$. In group theoretic notation, the consistency conditions require that for any $I_i, I_j, I_k$, the relative matchings between them satisfy $\tau_{kj}\tau_{ji} = \tau_{ki}$. An equivalent condition is that to each $I_i$ we can associate a base permutation $\sigma_i$, so that $\tau_{ji} = \sigma_j \sigma_i^{-1}$ for any $(i, j)$ pair. Thus, the problem of finding a consistent set of $\tau_{ji}$‘s reduces to that of finding just $m$ 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 $(g_{ij})_{ij=1}^m$ of group elements obeying consistency relations are called synchronization problems. Starting with the seminal work of Singer et al. [19] on synchronization over the rotation group for aligning images in cryo-EM, followed by synchronization over the Euclidean group [20], and most recently synchronization over $S_n$ for matching landmarks [21][22], problems of this form have generated a lot of interest.
2.1 Vector Diffusion Maps

In the context of synchronizing three dimensional rotations for cryo-EM, Singer and Wu [23] have proposed a particular 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 $I_j$, the observed $O_{ij}$ relative rotation of $I_j$ to $I_i$ changes it to $O_{ij}Q_i$. If all the $(O_{ij})_{i,j}$ observations were perfectly synchronized, then no matter what path $(i_1, i_2, i_3, \ldots, j)$ we took from $i$ to $j$, the resulting rotation $O_{i_1i_2} \cdots O_{i_ni_j}Q_i$, would be the same. However, if some (in many practical cases, the majority) of the $O_{ij}$’s are incorrect, then different paths from one vertex to another contribute different rotations that need to be averaged out.

The natural choice for the loss function to describe the extent to which the $Q_1, \ldots, Q_m$ imputed base rotations (playing the role of the $\alpha_i$’s in the permutation case) satisfy the $O_{ij}$ observations is

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

where $w_{ij}$ is the edge weight. This can be written in the form $E(Q_1, \ldots, Q_m) = V^\top \mathcal{L}V$, where

$$V = \begin{pmatrix} Q_1 \\ \vdots \\ Q_m \end{pmatrix}, \quad \mathcal{L} = \begin{pmatrix} d_1 I & -w_{21} O_{21} & \cdots & -w_{m1} O_{m1} \\ -w_{1m} O_{1m} & \ddots & \vdots \\ \vdots & \ddots & \ddots \\ -w_{m1} O_{m1} & \cdots & d_m I \end{pmatrix},$$

and $d_i = \sum_{j \neq i} w_{ij}$. Note that since $w_{ij} = w_{ji}$, and $O_{ij} = O_{ij}^{-1}$, the matrix $\mathcal{L}$ is symmetric. Furthermore, the above is exactly analogous to the way in which in spectral graph theory, (see, e.g.,[24]) the functional $E(f) = \sum_{i,j} w_{ij}(f(i) - f(j))^2$ describing the “smoothness” of a function $f$ defined on the vertices of a graph with respect to the graph topology can be written as $f^\top \mathcal{L}f$ in terms of the usual graph Laplacian

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

The consequence of the latter is that (constraining $f$ to have unit norm), the function minimizing $E(f)$ is the eigenvector of $L$ with (second) smallest eigenvalue. Analogously, in synchronizing rotations, the steady state of the diffusion system, where (1) is minimal, can be computed by forming $V$ from the 3 lowest eigenvalue eigenvectors of $\mathcal{L}$, and then identifying $Q_i$ with $V(i)$, by which we denote its $i$’th $3 \times 3$ block. The resulting consistent array $(Q_jQ_i^{-1})_{i,j}$ of imputed relative rotations minimizes the loss (1).

3 Permutation Diffusion

Its elegance notwithstanding, the vector diffusion formalism of the previous section seems ill suited for our present purposes of improving the SfM pipeline for two reasons: (1) synchronizing over $S_n$, which is a finite group, seems much harder than synchronizing over the continuous group of rotations; (2) rather than 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 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 fact (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 $\rho: G \to \mathbb{C}^{d_r \times d_r}$ which satisfies $\rho(g_2g_1) = \rho(g_2)\rho(g_1)$ is called a representation of $G$. The representation is unitary if $\rho(g^{-1}) = (\rho(g))^\dagger = \rho^\dagger$, 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 and $\rho: G \to \mathbb{C}^{d_r \times d_r}$ be a unitary representation of $G$. Then given an array of possibly noisy and unsynchronized group elements $(g_{ij})_{i,j}$ and corresponding positive confidence weights $(w_{ij})_{i,j}$, any synchronization loss function of the form

$$E(h_1, \ldots, h_m) = \sum_{i,j=1}^{m} w_{ij} \| \rho(h_jh_i^{-1}) - \rho(g_{ij}) \|_{\text{Frob}}^2$$
can be written in the form $\mathcal{E}(h_1, \ldots, h_m) = V^t \mathcal{L} V$, where

$$V = \begin{pmatrix} \rho(h_1) \\ \vdots \\ \rho(h_m) \end{pmatrix}, \quad \mathcal{L} = \begin{pmatrix} d_1 I & -w_{21} \rho(g_{21}) & \cdots & -w_{m1} \rho(g_{m1}) \\ \vdots & \ddots & \ddots & \vdots \\ -w_{1m} \rho(g_{1m}) & -w_{2m} \rho(g_{2m}) & \cdots & d_m I \end{pmatrix}. \quad (3)$$

To synchronize putative matchings between images, we just have to instantiate this proposition with the appropriate representation of the symmetric group. In particular, the obvious choice is the so-called defining representation, whose elements are the familiar permutation matrices

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

since the corresponding loss function is

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

The squared Frobenius norm in this expression simply counts the number of mismatches between the observed but noisy permutations $\tau_{ji}$ and the inferred permutations $\sigma_j \sigma_i^{-1}$. Furthermore, by the results of the previous section, letting $P_i \equiv P(\sigma(i))$ and $\widehat{P}_{ji} \equiv P(\tau_{ji})$ for notational simplicity, (4) can be written in the form $V^t \mathcal{L} V$ with

$$V = \begin{pmatrix} P_1 \\ \vdots \\ P_m \end{pmatrix}, \quad \mathcal{L} = \begin{pmatrix} d_1 I & -w_{21} \widehat{P}_{21} & \cdots & -w_{m1} \widehat{P}_{m1} \\ \vdots & \ddots & \ddots & \vdots \\ -w_{1m} \widehat{P}_{1m} & -w_{2m} \widehat{P}_{2m} & \cdots & d_m I \end{pmatrix}. \quad (5)$$

Therefore, similarly to the rotation case, synchronization over $S_n$ can be solved by forming $V$ from the first $d_{\text{pad}} = n$ lowest eigenvectors of $\mathcal{L}$, and extracting each $P_\sigma$, from its $i$'th $n \times n$ block. Here we must take a little care because unless the $\tau_{ji}$'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 [21], $V$ must be first be multiplied by $V(1)^t$, and then a linear assignment procedure used to find the permutation matrix most akin to each block. The resulting algorithm we call Synchronization by Permutation Diffusion.

4 Uncertain matches and diffusion distance

The obvious limitation of our framework, as described so far, is that it assumes that each keypoint in each image has a single counterpart in any other image. This assumption is far from being satisfied in realistic scenarios due to occlusion, repetitive structures, and noisy detections. Most algorithms, including [22] and [21], deal with this problem simply by setting the $P_{ij}$ entry of the Laplacian matrix in (5) equal to a weighted sum of all possible permutations. For example, if landmarks number 1...20 are present in both images, but landmarks 21...40 are not, then the effective $P_{ij}$ matrix will have a corresponding $20 \times 20$ block of all ones in it, rescaled by a factor of 1/20. The consequence of this approach is that each block of the $V$ matrix derived from $\mathcal{L}$ by eigendecomposition will also correspond to a distribution over base permutations.

In principle, this amounts to replacing the single observed matching $\tau_{ij}$ by an appropriate distribution $t_{ij}(\tau)$ over possible matchings, and concomitantly replacing each $\sigma_i$ with a distribution $p_i(\sigma)$. However, if some set of landmarks $\{u_1, \ldots, u_k\}$ are occluded in $I_i$, then each $t_{ji}$ will be agnostic with respect to the assignment of these landmarks, and therefore $p_i$ will be invariant to where they are assigned. Defining $\mu_{u_1, \ldots, u_k}$ as any permutation that maps $u_1 \mapsto 1, \ldots, u_k \mapsto k$, and regarding $S_k$ as the subgroup of permutations that permute $1, 2, \ldots, k$ amongst themselves but leave $k + 1, \ldots, n$ fixed, any set of permutations of the form $\{\nu \gamma \mu_{u_1, \ldots, u_k} \mid \gamma \in S_k\}$ for some $\nu \in S_n$ is called a (twisted) left $S_k$-coset, and is denoted $\nu S_k \mu_{u_1, \ldots, u_k}$. The above invariance property amounts to saying that $p_i$ is constant on each such coset.

When synchronized matchings by the diffusion map in the forms discussed above, these invariances will spontaneously appear in the $V$ matrix formed from the eigenvectors, and since they are related to what set of landmarks exactly are hidden or uncertain, are highly characteristic of the viewpoint that
the image was taken from. As noted earlier, an affinity score based on comparing images according
to this criterion is even more valuable than the synchronized matchings themselves.

If $p_i$ was a function given in explicit form, then the above type of invariance structure could be read
off easily from its so-called autocorrelation function

$$a_i(\sigma) = \sum_{\mu \in S_n} p_i(\mu)p_i(\mu\sigma),$$

since $a_i(\sigma)$ would be very high for any $\sigma$ in the $S_n\mu_n, \ldots, \mu_n$ coset. However, $V(i)$ (the $i'$th block of
$V$) only reveals a weighted sum $\hat{p}_i(\rho) := \sum_{\sigma \in S_n} p_i(\sigma)\rho(\sigma) = V(i)$ rather than the full function $p_i$.

Recent years have seen a series of papers involving Fourier transforms on the symmetric group,
which, given a function $f : S_n \rightarrow \mathbb{R}$, is of the form

$$\hat{f}(\lambda) = \sum_{\sigma \in S_n} f(\sigma)\rho_\lambda(\sigma), \quad \lambda \vdash n,$$

where the $\rho_\lambda$ are special representations of $S_n$ called irreducibles. Due to space restrictions, we leave
the details of this construction to the literature, see, e.g., [25, 26, 27]. Suffice to say that while $V(i)$
is not exactly a Fourier component of $p_i$, it can be expressed as a direct sum of Fourier components

$$V(i) = C^\dagger \left( \bigoplus_{\lambda \in \Lambda} \hat{p}_i(\lambda) \right) C$$

for some unitary matrix $C$ that only effects a basis transform. One of the properties of the Fourier
transform is that if $h$ is the correlation of two functions $f$ and $g$ (i.e., $h(\sigma) = \sum_{\tau \in S_n} f(\mu)g(\mu\sigma)$),
then $\hat{h}(\lambda) = \hat{f}(\lambda)^\dagger \hat{g}(\lambda)$. Consequently,

$$\hat{a}_i(\rho) := C^\dagger \left( \bigoplus_{\lambda \in \Lambda} \hat{a}_i(\lambda) \right) C = C^\dagger \left( \bigoplus_{\lambda \in \Lambda} \hat{p}_i(\lambda)^\dagger \hat{p}_i(\lambda) \right) C = V(i)^\dagger V(i)$$

it is very easy to compute matrix that captures essentially all the coset invariance structure encoded in
the inferred distribution over $\sigma_i$. To compute an affinity score between some $i$ and $j$, it remains to
calculate their coset invariance structures. The natural way is to compute the inner product $\langle a_i, a_j \rangle = \sum_{\sigma \in S_n} a_i(\sigma)a_j(\sigma)$, which, again using the correlation theorem, is found to be

$$\Pi(i,j) := \langle a_i, a_j \rangle = \text{tr}(\hat{a}_i(\rho)^\dagger \hat{a}_j(\rho)) = \text{tr}(V(i)^\dagger V(i) V(j)^\dagger V(j)).$$

Remarkably, this affinity measure, which we call Permutation Diffusion Affinity (PDA), is closely
related to the notion of diffusion similarity derived in [23] 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

We used the proposed Permutation Diffusion Maps to extract image association matrices for various
datasets described in the literature. Geometric ambiguities due to large duplicate structures are
evident in each of these datasets, in up to 50% of the matches [6, 28]. As we will show shortly, even
sophisticated SfM pipelines run into difficulties for such datasets. Our goal here is to precede entire
SfM engine with one simple preprocessing step. If our preprocessing step generates good image
association information, an existing SfM pipeline which is a very mature software with several linear
algebra toolboxes and vision libraries integrated together, can provide good reconstructions. While
our primary interest is SfM, to illustrate the utility of PDM, we also present experimental results for
scene summarization for a set of images [29, 30], see supplement.

Structure from Motion (SfM). 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]. As a baseline, we provide these
images to a Bundle Adjustment procedure which uses visual features for matching and already
has a built-in heuristic outlier removal module. Several other papers have used a similar set of
comparisons [6]. For each dataset, SIFT was used to detect and characterize landmarks [31, 32]. We compute putative pairwise matchings \( (\tau_{ij})_{i,j=1}^{n} \) by solving \( \binom{n}{2} \) linear independent assignments [33] based on their SIFT features. **Image Match Matrix:** Permutation matrix representation is used for putative matchings \( (\tau_{ij})_{i,j=1}^{n} \). Here, \( n \) is relative large, on the order of 1000. Eigenvector based procedure computes weighted affinity matrix. While specialized methods can be used to extract a binary image matrix (such that it optimizes a specified criteria), we used a simple thresholding procedure. **3D reconstruction:** We used binary match matrix as an input to a 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, just as if nothing has changed. The baseline 3D reconstruction is performed using the same SfM pipeline without such intervention.

The **HOUSE** sequence has three instances of similar looking houses, see Figure 1. The diffusion process accumulates evidences and eventually provides strongly connected component, see Figure 2(a). Warm colors correspond to high affinity between a pair of images. The binary match matrix was obtained by applying a threshold on the weighted matrix, see Figure 2(b). We used this matrix to define the image matching for feature tracks. This means, feature are only matched between images that are connected in our match matrix, otherwise not. The SfM pipeline was given these image matches as a hypotheses to explain how the images are “connected”. The resultant reconstruction correctly gives three houses, see Figure 2(c). The same SfM pipeline when allowed to track features automatically with an outlier removal heuristic, resulted in a folded reconstruction, see Figure 1(b). One may ask if more specialized heuristics will do better. Note that the literature suggests various heuristics (e.g., - time stamps) for such problems. But experimental results in [6] and others, strongly suggest that these datasets still remain challenging.

![Figure 2: House sequence](image)

The **CUP** dataset has multiple images of a 180 degree symmetric cup from all sides, Figure 3(a). PDM reveals a strongly connected component along the diagonal for this dataset, shown in warm colors in Figure 3(b). Our global reasoning over the space of permutations substantially mitigates coherent errors. The binary match matrix was obtained by thresholding the weighted matrix, see Figure 3(c). As it 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 4(b). PDM reconstruction gives a perfect reconstruction of the “full cup” with one handle as expected, see Figure 4(a). The **OAT** dataset contain two instances of red oat box. One on the left of the

![Figure 3: CUP dataset](image)

**WHEAT THINS** and another on the right of it, see Figure 5(a). The PDM weighted match matrix and binary match matrix successfully discover strongly connected components, see Figure 5(b-c). The baseline method confused the two instances of oat box as one and reconstruct only one oat box, see Figure 6(b). 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 THINS in the
middle, Figure 6(a). For better visualization and many more experiments (with videos), please see the supplement.

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

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

6 Conclusions

We have shown that the proposed Permutation Diffusion Maps algorithm on a special type of image association graph, in which edge attributes pertain to matchings between images, helps find truly global solutions to the image association problem. Even when a large number of visual feature matches are erroneous due to repetitive structures, performing a random walk on this graph in most cases reveals the correct associations. Our experiments on a variety of challenging datasets from the literature give strong evidence supporting the hypothesis that deploying the proposed formulation, even as a pre-conditioner, can significantly mitigate problems encountered in performing Structure from Motion on scenes with repetitive structures. The proposed model can easily generalize to other applications outside computer vision where unordered matching instances are common.

References