Structure from motion for scenes with large duplicate structures

Richard Roberts
Georgia Institute of Technology
Atlanta, GA
richard.roberts@gatech.edu

Sudipta N. Sinha, Richard Szeliski
Microsoft Research
Redmond, WA
{sudipsin, szeliski}@microsoft.com

Drew Steedly
Microsoft
Redmond,WA
steedly@microsoft.com

Abstract

Most existing structure from motion (SFM) approaches for unordered images cannot handle multiple instances of the same structure in the scene. When image pairs containing different instances are matched based on visual similarity, the pairwise geometric relations as well as the correspondences inferred from such pairs are erroneous, which can lead to catastrophic failures in the reconstruction.

In this paper, we investigate the geometric ambiguities caused by the presence of repeated or duplicate structures and show that to disambiguate between multiple hypotheses requires more than pure geometric reasoning. We couple an expectation maximization (EM)-based algorithm that estimates camera poses and identifies the false match-pairs with an efficient sampling method to discover plausible data association hypotheses. The sampling method is informed by geometric and image-based cues. Our algorithm usually recovers the correct data association, even in the presence of large numbers of false pairwise matches.

1. Introduction

Structure from Motion (SFM) is the problem of simultaneously estimating scene structure (3D points) and camera poses from an unordered set of images. Typical SFM methods first robustly match features in as many pairs of input images as possible, thereby recovering measurements of the relative rigid poses between camera pairs. Bundle adjustment [20] then computes a maximum likelihood estimate of the camera poses and point locations, after initialization using a subset of the pairwise measurements.

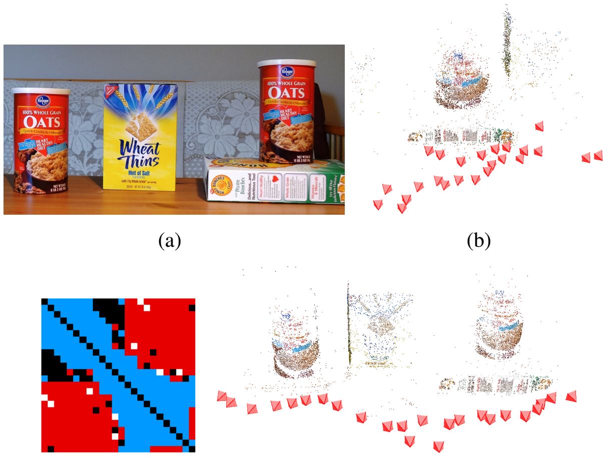
(a)
(b)
(c)
(d)

Figure 1: (a–b) Two identical objects in the scene result in a *folded reconstruction*. By inferring the erroneous matches (shown in red in the match graph adjacency matrix (c)), our method produces an accurate reconstruction shown in (d).

It turns out that in this traditional SFM pipeline, the implicit *data association*[1] method contains a fundamental assumption of there being only a single instance of any structure. When multiple large structures are similar, for example as shown in Figure 1, this assumption breaks down. This causes the pipeline to believe that two or more separate objects or structures are in fact the same, or to "mix and match" data associations between instances, which usually gives rise to folded or ghost structures. This is often a problem in architectural scenes.

In such cases, the relative pose estimates and data associations between the cameras involved in an erroneous match pair are incorrect. With large duplicate structures, the erroneous match pairs can form large, self-consistent sets, as shown in Figures 1 and 2. Without additional knowledge, there is no way to infer on a local scale that a particular small fraction of the match pairs is correct and the rest are incorrect. To avoid minor data association errors, state of the art SFM pipelines use smart heuristics for greedily choosing match pairs [17, 19, 13, 9, 18], though they cannot exclude large coherent sets of inter-instance matches.

Our goal in this paper is to determine the correct data

[1] In SFM, "data association" is the problem of determining correspondences, either between feature points or whole images. In the case of whole images, it can be seen as the validity of hypotheses that image pairs contain sets of matching features corresponding to the same 3D points.

association between pairs of images. We start with a set of geometrically consistent but potentially incorrect pairwise image matches and then determine which are *correct* and which are *erroneous*.

A potential approach is to exploit redundancy of the matches to detect which are inconsistent with the *majority* of the other matches (we describe this in Section 3). This approach labels the largest consistent set of edges as correct and the rest as erroneous, and is similar to previous approaches by Zach [23] and Govindu [7]. For reasons we discuss later on, this approach still contains the single instance assumption because it models the erroneous matches as statistically independent. Thus it can fail on multiple instances unless the erroneous matches are relatively few.

The contributions of this paper are first, to explain the fundamental difficulties that arise in structure from motion in the presence of duplicate structures and why current methods still often fail in such cases, and second, to propose a probabilistic model and method for inferring the correct data association that unify current approaches by combining geometric reasoning and non-geometric cues.

2. Related Work

The primary problem that occurs with duplicate structure is that large self-consistent sets of geometrically valid pairwise (or triplet-wise) image matches between instances are in fact incorrect. Previous work towards addressing this has used both geometric reasoning about camera poses and cues found in the images themselves.

Reasoning about large-scale structure matches relies on data association errors being discoverable by globally-inconsistent matches. For example, in Figure 1, by looking at *small* neighborhoods of matches, it is unclear whether the matches between the two oat boxes are correct, or if the conflicting matches between the oat boxes and the yellow box are correct. Indeed, the former outnumber the latter. By considering all of the measurements, the conflict between these two sets of edges is revealed.

Govindu [6, 7] randomly samples spanning trees on a pairwise relative rotation graph to initialize camera orientations. The hypothesis maximizing a statistical measure based on the number of consistent matches and their quality is then selected. Zach *et al.* [23] infers the validity of matches by approximate inference over statistics on which matches are involved in inconsistent loops (which should be nearly closed) in the match graph. Klopschitz *et al.* [11] incrementally build up a reconstruction from multiple subsets of the cameras with the highest local connectivity.

Martinec and Pajdla [13] repeatedly discard the highest-residual matches to be robust to some incorrect matches. Their method can handle some amount of repeated structure, such as the similar structures on opposite sides of the St. Martin rotunda, where the erroneous matches constitute a minority (8%) of the measurements, and importantly, occur with approximate statistical independence.

Data association is also a critical problem in simultaneous localization and mapping (SLAM) in the context of robotics [1, 5, 16, 8]. When a set of newly observed features match to previously observed ones, the algorithm must decide whether a loop closure is occurring or the features just appear to be similar. However, the time-sequential nature of the *incremental* map-building process means that a new image can only match to features that can be observed near the camera's current location, making erroneous matches between instances of structure much less likely than for unordered images, as in SFM.

Earlier work by Zach *et al.* [22] proposed to resolve data association ambiguities using the features that are matched between two images but *not* detected in a third image. The intuition is that if a large fraction of image features match between two images but not a third in a camera triplet, it is likely that the third image observes a different instance than the first two. We integrate this powerful cue into our proposed method, which combines multiple image cues and global geometric reasoning.

3. Consistent Majority Optimization

In this section, we describe a basic probabilistic model for correct and erroneous pairwise image matches and an associated inference method. This serves as one component of the unified method we present in Section 5. This basic model labels as erroneous the matches that are geometrically inconsistent with the *majority* of the others.

In the next section, we show why this model by itself cannot solve any but the easiest cases of duplicate structure because it assumes that the erroneous matches are statistically independent of each other. Previous methods such as [23, 7, 13] also suffer from this limitation. Here we illustrate the cause with the help of a generative model.

3.1. Measurement Model

Given a set of putative image matches and their associated camera transforms, which are geometrically consistent with rigid camera transformations, we wish to infer which are correct. We introduce a generative model for these matches with hidden *correct/erroneous indicator variables*.

In our model, the i^{th} measurement z_i, supposing it is between the j^{th} and k^{th} cameras, is generated as the relative pose between them corrupted by Gaussian noise,

$$z_i \sim \mathcal{N}\left(x_j^{-1} x_k, C_{y_i}\right), \tag{1}$$

where x_j and x_k are the poses of the j^{th} and k^{th} cameras. The inverse is the inverse transformation, equivalent to the matrix inverse for pose and rotation matrices. Thus, $x_j^{-1} x_k$ is the predicted relative pose between cameras j and k. C_{y_i}

is the covariance matrix of the noise on the camera translation directions and rotation axes. We model the noise on each measurement as a mixture of Gaussian "inlier" and "outlier" densities. When y_i=1, C_1 is the inlier covariance, which be obtained from the pairwise reconstructions. When y_i=0, C_0 is the outlier covariance, which is chosen to be large (we use uniform $1\ \mathrm{rad}$ variance for rotation and 0.5 for the unit translations).

As we will describe in Section 5, in the first stage of our algorithm, we work only with the camera rotations to avoid scale ambiguity, in which case x_j, x_k, and z_i belong to the 3×3 matrix Lie group of rotations $\mathbb{SO}(3)$. In the second stage, we model full cameras poses by registering view triplets as we describe in Section 3.3, in which case x_j, x_k, and z_i are members of the 4×4 matrix Lie group of 3D rigid transformations $\mathbb{SE}(3)$.

The Gaussian mixture model makes the probability density on the camera poses in general non-Gaussian. However, when it is conditioned on the hidden indicator variables y, the distribution is Gaussian when linearized about an estimate of the camera poses x.

3.2. Iterative Inference Using EM

When all measurements are correct, the unknown poses can be recovered by minimizing least-squares error of the pose graph (i.e. match graph) [15, 6, 18, e.g.]. Since some are erroneous, we jointly infer the poses and the probabilities of each match being correct with an expectation-maximization (EM) algorithm [3]. This avoids combinatorial search over all y_i by finding a locally-optimal solution. In Section 5, we extend this method with random restarts to be more likely to find the global optimum.

In the M step we find the maximum expected likelihood solution for the poses x given an estimate of the expected values of the indicator variables y,

$$x^t = \arg\max_x \sum_i \sum_{y_i} \left\langle \log p\left(x_j, x_k \,|\, z_i, y_i\right)\right\rangle_{y_i \,|\, x^{t-1}, z_i}. \tag{2}$$

The E step then estimates the expected value of each y_i, i.e., the probabilities of each of the edges being an inlier.

This leads to the M-step update equation[2]

$$x^t = \arg\max_x \log p(x) + \sum_i \lambda_i^t \left\| z_i^{-1} x_j^{-1} x_k \right\|^2_{C_1} + \left(1 - \lambda_i^t\right) \left\| z_i^{-1} x_j^{-1} x_k \right\|^2_{C_0}, \tag{3}$$

where $p(x)$ is a pose prior (we place a prior only on one of the cameras, to fix it at the origin), and $\|\cdot\|_C$ is the Mahalanobis distance with covariance matrix C. The term $z_i^{-1} x_j^{-1} x_k$ is the deviation of the measurement z_i from its prediction $x_j^{-1} x_k$ (see Equation 1). λ_i^t is the expectation of y_i, from the E-step update

$$\lambda_i^t = \frac{\mathcal{N}\left(z_i^{-1} x_j^{-1} x_k;\, \mathbf{0}, C_1\right) p\left(y_i{=}1\right)}{\sum_{y_{ij}=0,1} \mathcal{N}\left(z_i^{-1} x_j^{-1} x_k;\, \mathbf{0}, C_{y_i}\right) p\left(y_i\right)}. \tag{4}$$

Note that this is evaluated using the pose estimates from the previous iteration, though we omit $t-1$ superscripts on the poses for clarity. $p(y)$ is a prior on the probability of an edge being correct. We use an uninformative prior in experiments, but this could be specified beforehand or estimated online.

3.3. Inferring 6-DOF Camera Poses via Triplets

The inference method described above is defined and valid both for camera rotations in $\mathbb{SO}(3)$ and for full camera poses (rotation and translation) in $\mathbb{SE}(3)$. However, due to the inherent scale ambiguity in 3D reconstructions, the relative scale between any pair of pairwise relative translations is unknown. The optimal (MLE) way to handle this would be to use triplet measurements that constrain relative but not global scale. Because this would add a layer of complexity both to the problem definition and implementation, we instead opt for a simpler method of resolving scale that actually over-counts some measurements. Performing a full bundle adjustment (using the original feature measurements) after removing erroneous matches yields a final SFM solution that is not affected by this over-counting.

The approach we choose is to first perform triplet reconstructions, then choose a tree of triplet reconstructions spanning all of the cameras and traverse the tree while rescaling each child triplet to be consistent in scale with its parent. During the M-step update of x^t, we treat each triplet reconstruction as a set of three pairwise relative pose measurements (one from each pair in the triplet). Amongst multiple measurements for a camera pair, we use the one with the minimum residual. During the E-step, we compute a probability of being correct for each triplet, by first computing a similarity transformation that aligns each triplet reconstruction with the current pose estimates using the approach described in [10] and then evaluate Equation 4 with the camera projection centers of the scaled and current pose estimates.

4. Difficulties Caused by Multiple Instances

The model in the previous section assumes statistically independent outliers. Unfortunately, the erroneous match pairs that occur due to large duplicate structures can form large, coherent sets that can overwhelm the correct matches and appear as inliers, while the smaller sets of correct matches appear as outliers. In this section, we start with an example of how large coherent sets of erroneous image matches form, and show why modeling them as statistically

[2] Please see the supplementary material (http://www.cc.gatech.edu/~richard/cvpr11-supp/) for a derivation of these updates and brief explanation of Lie group notation.

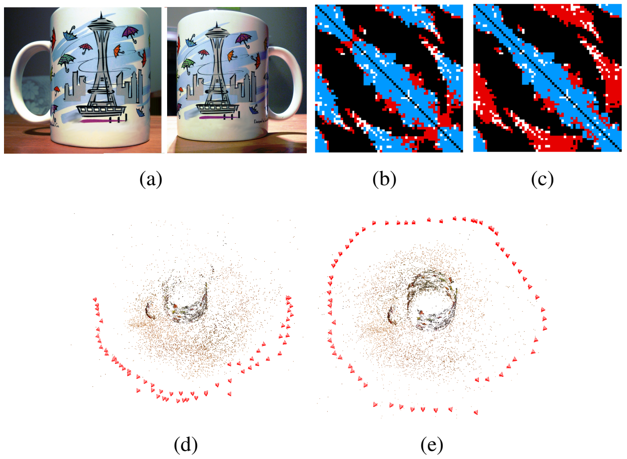

Figure 2: (a) The CUP sequence has a 180° symmetry. (b,d) A folded match matrix and reconstruction. Numerous erroneous matches between opposite sides of the cup (in blue in the matrix) outnumber the few matches at the "fold points" (in red in the matrix). Traditional SFM puts all cameras on one side, and the back of the cup is not reconstructed. (c,e) The correct data association matrix and reconstruction obtained using our approach (relying on the timestamps – see Section 5). Many more matches are labeled as outliers (red in the match matrices) to form the correct reconstruction.

independent leads to the implicit single instance assumption and an incorrect cost function. We then show that the large number of erroneous matches also makes it very difficult to discover the correct solution using naïve methods. In Section 5, we present our approach addressing these problems.

4.1. Coherent Sets of Erroneous Matches

In order to understand how the modeling of erroneous edges as statistically independent leads to the implicit single instance assumption, we start with an example. Figure 2 shows an "orbit" sequence with a 180° radial symmetry. Matches between nearby cameras are correct, but matches across the circle confuse the two sides of the cup as the same structure and cause the reconstruction to fold in half.

The assumption of statistically-independent outliers in this model implicitly results in the assumption of only one instance of any scene structure. Some matches must be ignored to produce a consistent reconstruction. In order to fold the reconstruction, only the few matches at the "fold points" of the folded reconstruction need to be ignored (marked as erroneous), as seen in the red entries along the diagonal of the match matrix in Figure 2b. To unfold the reconstruction, all of the many matches across the circle must be ignored, as seen in the red entries of the match matrix in Figure 2c. According to the independent-outlier model, each match is ignored at a constant cost. Thus, the largest coherent set of edges overwhelms any edges not consistent with it. This model is only suitable when erroneous edges occur randomly due to match errors, degenerate point configurations, or other uncorrelated random processes.

4.2. Combinatorial Search for the Correct Matches

Correlated outliers are not the only problem to be addressed. In addition to requiring a *scoring* function that behaves correctly, any inference method must also *discover* the correct solution. Exhaustive search is of course intractable, and unfortunately, local search methods, such as the EM algorithm presented in Section 3, as well as most previous research, are susceptible to local minima.

Although stochastic sampling methods are generally useful for solving problems with local minima, the coherent erroneous matches again cause a problem for naïve sampling methods. For sampling random spanning trees, for instance, it becomes extremely unlikely to sample a tree with no erroneous matches. We can see an example of this from the ratio of red erroneous edges to blue correct edges along any given row of the right-hand match matrix in Figure 2. In order to choose a correct hypothesis, the sampler must roughly choose two correct matches for every image, a probability that decreases roughly inverse-exponentially as the number of images increases.

5. Proposed Method

Our approach combines image cues with global geometric reasoning to label pairwise image matches as correct and erroneous. We sample minimal configurations of data associations, and from these samples perform a local search for complete match validity and camera pose configurations. To address the issue of incorrect solutions appearing more likely than the correct one, as described in Section 4, we formulate a likelihood function that leverages image cues. To efficiently discover the correct configuration, we also employ the cues in a heuristic to guide the sampling process.

For structure from motion, we apply the proposed technique in two stages. First, we estimate global camera orientations using only pairwise relative rotation measurements. Subsequently, we simultaneously estimate rotation and translation using precomputed camera triplets and the extension described in Section 3.3. For computing relative pairwise pose and triplet reconstructions, we employ known techniques described in [14, 17, 13, 18, 19, 2, 9]. Finally, we use only the match pairs inferred as correct as input to a state of the art structure from motion pipeline with standard bundle adjustment [20] to compute the final reconstruction.

5.1. Sampling Minimal Hypotheses

The proposed algorithm is similar to RANSAC [4]. We sample spanning trees, which are minimal hypotheses from

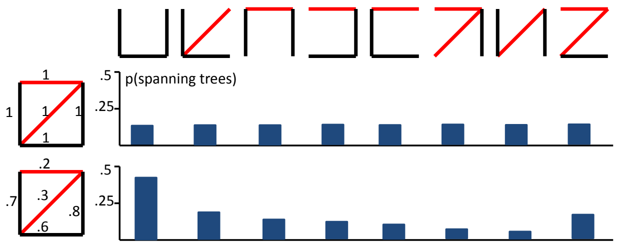

Figure 3: (Left) A match graph with 4 cameras and 5 pairwise matches (three correct and two incorrect edges colored black and red respectively) is shown. Red edges would suggest different relative camera positions than shown. When randomly sampling the spanning trees of this graph (shown at the top) from a uniform probability density, the probability of sampling the left-most spanning tree (all its edges are black) is quite low (upper histogram). By weighting the edges using evidence of their correctness, the probability of the left-most spanning tree becomes higher (lower histogram).

which we estimate all camera poses. A spanning tree containing no erroneous matches is sufficient to generate a complete correct solution after including other matches that are consistent with it. As we show in Figure 3, the probability of naïvely (i.e. uniformly) sampling such a spanning tree is very low, and becomes exponentially more so as the number of matches increases, even if the fraction of correct edges remains the same. The key to sampling a correct spanning tree in a reasonable amount of time is to define a probability density over spanning trees in which correct trees are more likely, and to sample from this density.

Generating random spanning trees according to a specific distribution is a well-studied problem in random graph theory. Here we use an efficient algorithm by Wilson [21]. The distribution over spanning trees is defined by a weight on each edge, and the probability of each spanning tree is proportional to the product of its edge weights.

The key observation is that to specify a distribution over spanning trees that is more likely to include correct matches, we simply specify edge weights according to how likely each edge is to be correct. We now describe two image cues that we combine to form the edge weights.

Missing Correspondence Cue: For image pairs observing the same structure instance, portions of the rest of the scene, such as the background, are also likely to match. Otherwise, it is possible that the match is between separate instances, as shown in Figure 4a. We use the *missing correspondence cue* proposed by [22] slightly modified to discount missing correspondences nearby to matched ones.

For the j-th image, the feature points matched to any other image are denoted by $\mathbf{X}_j$, out of which those matched

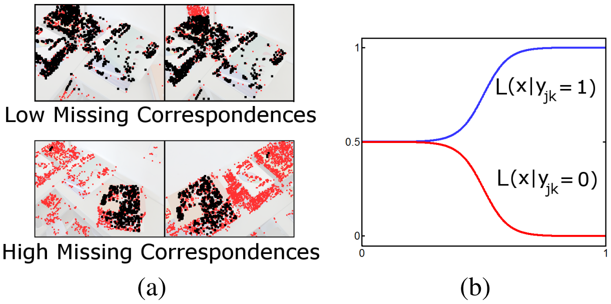

Figure 4: (a) The missing correspondences (red points) are visualized for a correct pair (top) and erroneous pair (below). Low missing correspondences indicate that the match pair is more likely to be correct. However no conclusions can be drawn when the degree of missing correspondences is high. (b) Likelihood $L(y_{jk}; M)$ on the binary indicator y_{jk}, which is 1 if the pairwise match between image j and k is correct, and 0 if it is erroneous – see Eqs (5) and (6).

in the k-th image are denoted by $\mathbf{X}_{jk}$. The fraction of *matched correspondences* from image j to k is thus $\frac{|\mathbf{X}_{jk}|}{|\mathbf{X}_j|}$.

More matched correspondences increase the likelihood that a match is correct, though we found it beneficial to *emphasize* the effect of missing correspondences that occur spatially far in the image from matched correspondences. The intuition is that missing correspondences near to matched ones are likely to be simply matching failures instead of missing structure due to an inter-instance match.

Thus, instead of defining a likelihood in terms of the matched correspondence fraction, we use a *background-adjusted* matched correspondence measure $f_{jk} = \frac{|\mathbf{X}_{jk}|}{n^*}$. n^* is a background-adjusted measure of the points in image j. When there are many missing correspondences far from the matched ones, n^* is large (decreasing f_{jk} and thus the likelihood of a correct match) and approaches the total number of point in image j. However, it decreases (increasing the likelihood of a correct match) when there are many missing correspondences near to the matched ones. We define it as $n^* = |\mathbf{X}_{jk}| + \sum_{u \notin \mathbf{X}_{jk}} b_u$, where b_u is a normalized measure of how far a point u is to each of its R (=20) nearest-neighbors in $\mathbf{X}_{jk}$, defined as $b_u = (\frac{1}{n}) \sum_r^R (1 - exp(-d_{ur}/\sigma))$.

Using the background-adjusted matched correspondence measure, the overall matched correspondence measure for an image pair j, k is $M_{jk} = \max(f_{jk}, f_{kj})$. Then, the likelihood on the binary indicator y_{jk}, which is 1 if the pairwise match between image j and k is correct, and 0 if it is erroneous (see Figure 4b), is

$$L(y_{jk}{=}1; M_{jk}) = \tfrac{1}{2}(1 + (1 + exp(-\alpha(M_{jk} - \beta)))^{-1}) \tag{5}$$

$$L(y_{jk}{=}0; M_{jk}) = 1 - L(y_{jk}{=}1; M_{jk}). \tag{6}$$

We use $\alpha = 20$ and $\beta = 0.5$.

Image timestamp cue: Modern digital cameras record timestamps in their EXIF tags. In a single photographer scenario, this provides approximate sequence information, and pairwise matches relatively close in time are less likely to be erroneous than those far in time. This is not true in other cases such as Internet photos [19, 12].

We calculate a likelihood that a match is correct based on the timestamp cue according to the ratio between the time difference of the match and the smallest time difference of any match involving one of its cameras, i.e., we compute $q_{jk} = \frac{\min_l \{t_{jl}\}}{t_{jk}}$. The time cue for pair jk is then $T_{jk} = \max(q_{jk}, q_{kj})$ and we model $L(y_{jk}; T_{jk})$ as described above (5,6) but with α=10 and β=0.25.

Given both cues, edge weights for pair jk are computed as

$$w_{jk} = \frac{L(y_{jk}{=}1; M_{jk}) L(y_{jk}{=}1; T_{jk})}{\sum_{y_{jk}=0,1} L(y_{jk}; M_{jk}) L(y_{jk}; T_{jk})}. \qquad (7)$$

5.2. Completing the Match Labeling

Given each sampled minimal hypothesis, we find the complete set of matches consistent with the hypothesis and then refine the estimated camera poses. From these hypotheses, we then select the one with the highest score. For efficiency, we remove duplicate spanning trees from the sample set using a binary hashing scheme prior to scoring.

To complete the labeling and refine the camera poses, we use the EM algorithm described in Section 3 with a slight modification. Since we are interested in finding the match pairs consistent with the sampled spanning tree, we fix the indicator variables y_{jk} for the spanning tree edges to 1 so they are always inliers in the EM solution. In the case of pure rotations, we initialize EM by composing relative rotations along the spanning tree.

Initialization in the case of triplets is similar. First, we compute a minimal subset of edges in the original match graph such that every edge in the subset is shared by at least two different triplets, also ensuring that every node in the graph is covered by the subset. We sample random spanning trees from this graph, with the same edge weights as above. For each spanning tree, we find the subset of all *tree-induced* triplets, those with two edges in the the spanning tree. Before chaining triplets to initialize EM, if the set of triplets is not connected, additional triplets must be selected in order to join the disconnected triplets. The two largest disconnected sets are iteratively merged by computing a *loop erased random walk* [21] between triplets in these two sets. The random walk is performed on a trifocal graph [2], in which triplets form nodes and triplets that share edges (in the original match graph) are connected via edges. This random walk uses the same edge weights as the random spanning tree generator in Sec. 5.1.

Dataset	BLDG	DESK	BOOKS	OATS	BOXES	CUP
#Images	76	31	21	24	25	64
#Pairs	889	265	180	250	265	990
#Erroneous	161 (18%)	30 (11%)	50 (28%)	125 (50%)	115 (43%)	459 (46%)
Trad.SFM	yes	no	no	no	no	no
Ours-M	yes	yes	yes	no	no	no
Ours-T	yes	yes	yes	yes	yes	yes
Ours-B	yes	yes	yes	yes	yes	yes

Table 1: For each dataset, the number of images, match pairs and erroneous pairs found by our method are listed. We report whether a correct reconstruction was produced with traditional SFM, and when we use missing correspondence (Ours-M), timestamps (Ours-T) or both (Ours-B).

5.3. Scoring Hypotheses

The solution of the EM algorithm initialized from each unique spanning tree hypothesis generates an inlier probability $p(y_{jk}{=}1)$ for each edge of the pairwise match-graph in the rotation case. When sampling triplets, $p(y_{jk}) = \max_{i \in T_{jk}} \{p(y_i)\}$, where $p(y_i)$ is the inlier probability of the i-th triplet (using the notation from Sec. 3). We assign y_{jk}=1 (i.e. mark edge jk as an inlier) when $p(y_{jk}{=}1) > .9$ and otherwise assign y_{jk}=0. This binary assignment of the variables in Y is called a *configuration*. From among all the configurations sampled by our approach, we choose the one with the highest log-likelihood,

$$L(Y) = \sum_{jk} \log(L(y_{jk}; M_{jk}) L(y_{jk}; T_{jk})). \qquad (8)$$

In ambiguous cases, where the cues are weak, the best k configurations can be computed as well.

6. Results

In this section, we evaluate our inference method on datasets that either contain duplicate structures or a large object moved in the scene while the images were being taken. The latter case produces coherent sets of erroneous matches in the same way as duplicate structures. These datasets contain very high fractions of coherent erroneous matches (46-50%), and existing methods fail on most of them. We compare to a state-of-the-art SFM pipeline incorporating smart initialization and outlier removal heuristics, similar to [19]. Our method produces correct reconstructions while in all cases but one the traditional pipeline folds the reconstructions. Further comparisons with BUNDLER [19] are shown in the supplementary material.

In Figures 5 – 9, the color-coded matrices indicate the validity of each pairwise image match according to the missing correspondences cue, the timestamp cue, the hand-labeled ground truth, and the labels inferred by our method using only pairwise rotations, and using triplet reconstructions as well. In each match matrix, entries are labeled as "correct" (blue), "erroneous" (red), "unmatched" (black), or discarded due to high residual errors in the match pair (red). Red and blue matches are all interpreted as valid by traditional SFM pipelines, unless pruned during initialization.

Figure 5 shows one sequences with two identical books, and another where a pile of books was moved partway through the collection of the dataset. The traditional SFM pipeline "folds" both reconstructions, i.e. only one instance of the duplicate structure is reconstructed. Our method successfully infers the matches between instances as erroneous to produce correct reconstructions.

Figure 6 shows an outdoor scene with three similar facades on a street, on which the traditional method fails whereas our method recoveres the correct structure and camera motion.

Figure 7 shows the BOXES sequence, which contains two identical cups and two identical cereal boxes. The traditional pipeline reconstructs a single cereal box and produces a "ghost" instance of the textured ground plane pitched to 70°, due to the second cereal box leaning at 70°.

Figures 1 and 8 show the OATS sequence with two identical boxes in the scene and the incorrect reconstruction obtained by the traditional method. In this sequence, the motion of the camera was mostly translational, as seen by the camera trajectory in Figure 1(d). Due to this, inference using only camera rotations did not discard all the erroneous matches (see Figure 8) whereas triplet-based inference correctly discarded most of the erroneous match pairs.

We compared our approach to [23] on the CUP sequence. The match matrix inferred by [23] (in Figure 9a) labels many erroneous edges as correct and labels edges at the weakest part of the match pair graph as erroneous. In contrast, when we use timestamps (Figure 9(b,c)), many erroneous edges are filtered and a correct reconstruction is obtained. We found the missing correspondence cue to be relatively less discriminative for this dataset (see Figure 9d) as there are fewer background features in this dataset.

Our method took 44 and 90 minutes on the CUP (64 imgs.) and BLDG (76 imgs.) datasets respectively, in addition to the baseline SFM pipeline which took 9 and 12 minutes respectively. Bundler took 24 and 56 minutes on CUP and BLDG. Our EM implementation uses a dense solver for the M-step. Currently, one iteration of Rotation EM takes 3 seconds and Triplet EM takes 30 seconds for CUP. Dense solving is currently a large bottleneck but could easily be replaced with a sparse solver. The algorithmic complexity of our method depends on the strength of the cues. Stronger cues mean fewer samples are required to reach the same degree of statistical certainty of sampling a valid hypothesis. In our experiments, we sampled 200 times (Rotation EM) and 50 times (Triplet EM). To find the number of samples needed to ensure with a certain probability that a valid hypothesis was found is an interesting open question.

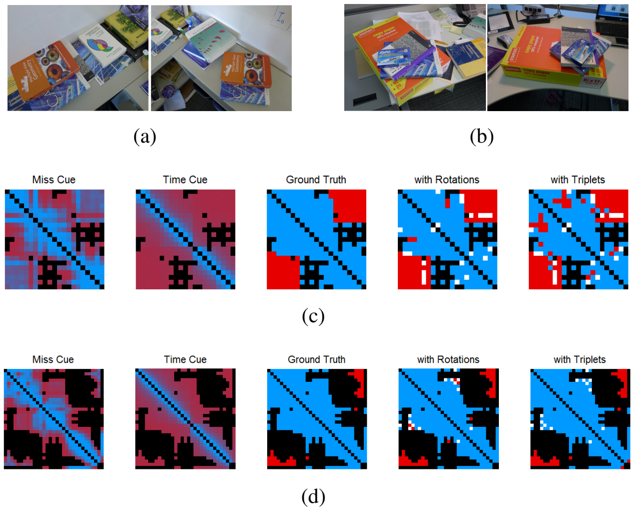

Figure 5: (a) BOOKS sequence – two identical books in the scene. (b) DESK sequence – a pile of books have been moved to create a second virtual instance. (c) For the BOOKS sequence, the image cues, the ground truth labeling and the labeling computed by our method is shown. (d) The same information is shown for the DESK sequence.

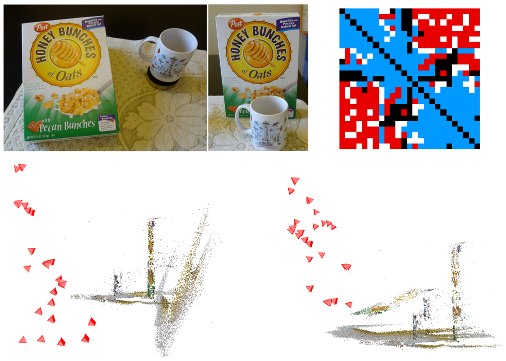

Figure 7: BOXES sequence: The left point cloud was obtained from the comparison pipeline, whereas the one on the right was obtained using our method.

Our algorithm can fail if the cues are weak or misleading, when either a good hypothesis is never sampled, or the correct reconstruction is sampled but scores lower than an incorrect one. For the latter case, we can compute a family of plausible top-scoring candidates for manual selection, whereas competing methods have no way to achieve this.

7. Conclusion

In this paper, we have demonstrated the difficulty and ambiguity of inferring data associations for SFM on scenes

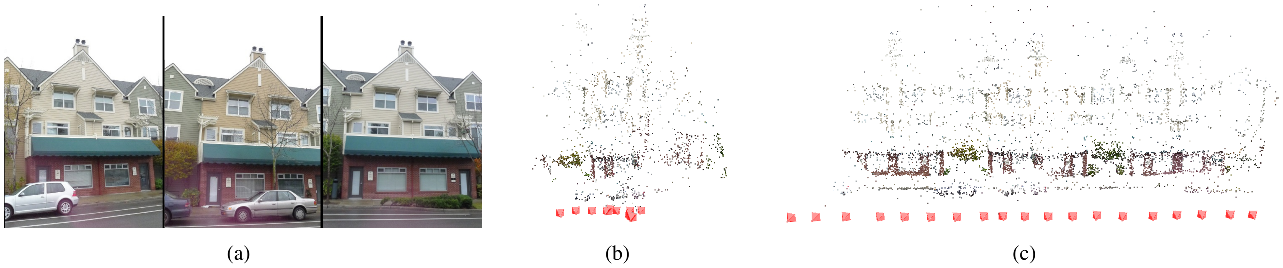

Figure 6: HOUSE sequence: (a) Identical building facades in separate locations. (b) The duplicate structures collapse into a single instance in the baseline reconstruction. (c) Our method correctly reconstructs the facade and linear camera motion.

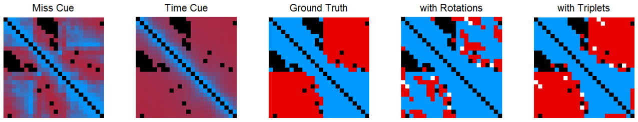

Figure 8: Cues and inferred labeling for the OATS sequence.

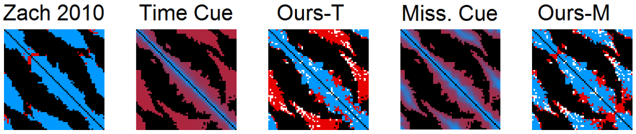

Figure 9: (a) The match matrix obtained using [23] on CUP. (b) Timestamp matrix and (c) match matrix inferred from it using our method (the reconstruction is shown in Figure 2e). (d) Missing correspondence matrix and (e) match matrix inferred from it using our method (reconstruction was folded).

with duplicate structure instances. For such cases, we have proposed a new approach for inferring and removing erroneous match pairs, which can occur when the different structure instances are matched based on visual similarity.

In summary, our main contribution lies in characterizing the underlying geometric ambiguity in the problem and in a new algorithm based on a unified probabilistic model and sampling-based inference method that incorporates global geometric reasoning with evidence from pairwise image cues. We demonstrate results on challenging datasets with up to 50% erroneous matches, on which a state-of-the art SFM method produces incorrect "folded" reconstructions, but our method produces correct reconstructions.

References

[1] C. Bibby and I. Reid. Simultaneous localisation and mapping in dynamic environments (SLAMIDE) with reversible data association. In *Proc. of Robotics: Science and Systems*, 2007.

[2] J. Courchay, A. S. Dalalyan, R. Keriven, and P. F. Sturm. Exploiting loops in the graph of trifocal tensors for calibrating a network of cameras. In *ECCV (2)*, pages 85–99, 2010.

[3] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[4] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM*, 24(6):381–395, 1981.

[5] A. Gil, Ó. Reinoso, Ó. M. Mozos, C. Stachniss, and W. Burgard. Improving data association in vision-based SLAM. In *IROS*, pages 2076–2081, 2006.

[6] V. M. Govindu. Lie-algebraic averaging for globally consistent motion estimation. In *Proc. CVPR (1)*, pages 684–691, 2004.

[7] V. M. Govindu. Robustness in motion averaging. In *Proc. ACCV (2)*, pages 457–466, 2006.

[8] D. Hähnel, S. Thrun, B. Wegbreit, and W. Burgard. Towards lazy data association in slam. In *ISRR'03*, pages 421–431, 2003.

[9] M. Havlena, A. Torii, J. Knopp, and T. Pajdla. Randomized structure from motion based on atomic 3d models from camera triplets. In *Proc. CVPR*, pages 2874–2881, 2009.

[10] B. K. P. Horn, H. Hilden, and S. Negahdaripour. Closed-form solution of absolute orientation using orthonormal matrices. *Journal of the Optical Society America*, 5(7):1127–1135, 1988.

[11] M. Klopschitz, A. Irschara, G. Reitmayr, and D. Schmalstieg. Robust incremental structure from motion. In *Proc. 3DPVT*, 2010.

[12] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *Proc. ECCV*, pages 427–440, 2008.

[13] D. Martinec and T. Padjla. Robust rotation and translation estimation in multiview reconstruction. In *Proc. CVPR*, 2007.

[14] D. Nister. An efficient solution to the five-point relative pose problem. *PAMI.*, 26:756–777, June 2004.

[15] E. Olson, J. Leonard, and S. Teller. Fast iterative optimization of pose graphs with poor initial estimates. In *Proc. ICRA*, pages 2262–2269, 2006.

[16] A. Ranganathan, E. Menegatti, and F. Dellaert. Bayesian inference in the space of topological maps. *IEEE Transactions on Robotics*, 22:92–107, 2006.

[17] F. Schaffalitzky and A. Zisserman. Multi-view matching for unordered image sets, or "How do I organize my holiday snaps?". In *Proc. ECCV*, pages 414–431, 2002.

[18] S. N. Sinha, D. Steedly, and R. Szeliski. A multi-staged linear approach to structure from motion. In *RMLE-ECCV workshop*, 2010.

[19] N. Snavely, S. M. Seitz, and R. Szeliski. Photo Tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006.

[20] B. Triggs, P. Mclauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – a modern synthesis. In *Vision Algorithms: Theory and Practice, LNCS*, pages 298–375, 2000.

[21] D. B. Wilson. Generating random spanning trees more quickly than the cover time. In *Proc. STOC*, pages 296–303, 1996.

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

[23] C. Zach, M. Klopschitz, and M. Pollefeys. Disambiguating visual relations using loop constraints. In *CVPR*, pages 1426–1433, 2010.