Disambiguating Visual Relations Using Loop Constraints

Christopher Zach
ETH Zürich, Switzerland
chzach@inf.ethz.ch

Manfred Klopschitz
TU Graz, Austria
klopschitz@icg.tugraz.at

Marc Pollefeys
ETH Zürich, Switzerland
marc.pollefeys@inf.ethz.ch

Abstract

Repetitive and ambiguous visual structures in general pose a severe problem in many computer vision applications. Identification of incorrect geometric relations between images solely based on low level features is not always possible, and a more global reasoning approach about the consistency of the estimated relations is required. We propose to utilize the typically observed redundancy in the hypothesized relations for such reasoning, and focus on the graph structure induced by those relations. Chaining the (reversible) transformations over cycles in this graph allows to build suitable statistics for identifying inconsistent loops in the graph. This data provides indirect evidence for conflicting visual relations. Inferring the set of likely false positive geometric relations from these non-local observations is formulated in a Bayesian framework. We demonstrate the utility of the proposed method in several applications, most prominently the computation of structure and motion from images.

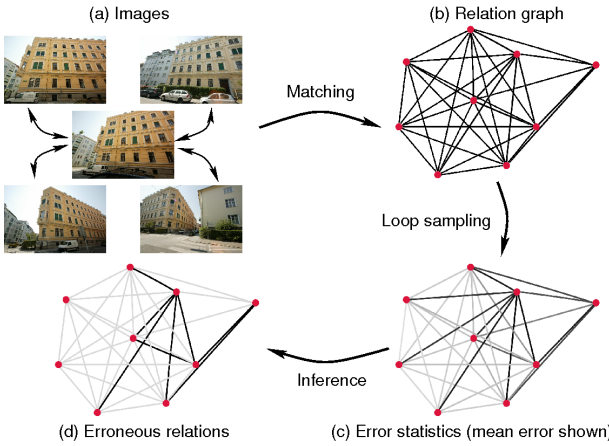
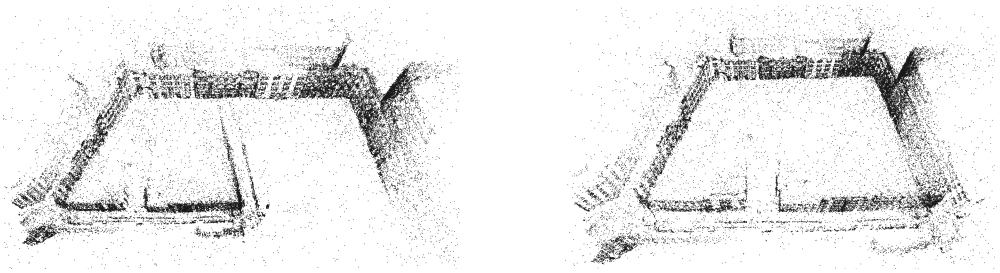


Figure 1. The original set of images (a) is robustly matched yielding a graph containing all potential pairwise relations (b). Acquisition of deviation statistics over loops results in non-local error observations (c), from which the incorrect relations are inferred (d). Large error and erroneous relations are indicated by dark edges. Observe that the central and the top right image look similar, but actually show different sides of the building.

1. Introduction

Computing the geometric relations from unorganized image sets purely from visual features is a difficult task. In order to obtain a tractable method, usually a pairwise matching procedure is applied first, which is followed by a fusion step to merge the initially obtained pairwise relations into some global reference frame. The approaches proposed in the literature vary widely in the details of this latter upgrade procedure. Since the first pairwise matching step uses only very limited information, the reported pairwise relations are susceptible to inconsistencies due to visual ambiguities, and the subsequent fusion method must be able to cope with such erroneous input. We do not restrict the notion of pairwise matching solely to images, but also consider e.g. mutual alignment of 3D point sets.

In this work we propose to detect and remove conflicting pairwise relations, and thereby cleaning the input for the subsequent upgrade step from incorrect data. The principal components are illustrated in Figure 1. The pairwise relations generated by the preceding matching stage are typically highly redundant, which enables checking the intrinsic geometric consistency of these relations. The set of reported pairwise relations corresponds directly to a graph structure associating its edges with the relations (Fig.1(b)). Most classes of pairwise relations relevant in computer vision applications—e.g. homographies, relative pose, Euclidean and similarity transformations—allow the concatenation of geometric relations to hypothesize new, potentially not directly observed relations. Large deviations between predicted (chained) and actually observed transformations indicate at least one conflicting edge among the involved relations. Under the weak assumption of invertible transformations we can restrict the focus on cycles in the graph structure. Concatenating the transformations along a loop in the graph should return the identity function in an ideal, noise-free setting. Again, the likelihood of having at least one incorrect edge in the loop is strongly related to the deviation of the chained transformation from



(a) 3D model using all visual relations (b) 3D model using only consistent relations

Figure 2. Distorted and correct output of a robust structure and motion pipeline using all geometrically verified epipolar relations (a) and using only those satisfying loop consistency (b).

the identity map. Collecting these statistics over loops indirectly points to potentially incorrect edges. The statistics for correct edges is generally contaminated by false positives also participating in the cycles, hence the conflicting edges cannot be read directly e.g. from the mean deviations (see Fig. 1(c)). Our proposed method uses a Bayesian network to infer the most likely set of incorrect transformations in the graph (Fig. 1(d)).

In this work we propose a solution towards resolving the two conflicting goals encountered in 3D computer vision: creating as connected results as possible (i.e. maximizing the recall), while simultaneously avoiding incorrectly merged components (maximizing the precision). We augment robust vision methods addressing these issues with explicit reasoning steps on the geometric consistency in order to increase the recall while maintaining the precision. The objective of the proposed method is avoiding distorted results as seen in Figure 2(a) by removing conflicting visual relations as preprocessing step (Figure 2(b)).

2. Related Work

Extracting information from erroneous data is generally the goal of robust estimation; in computer vision random sampling [7] and its subsequent extension are practical approaches to robustly estimate a small set of parameters from data contaminated with outliers. In certain problem setting the L^∞ cost function can be used to identify outliers in the given data [19], but usually robust cost functions like the Huber or Cauchy cost function are used in larger scale parameter estimation tasks (like bundle adjustment [22]). Inconsistencies in the visual relations are only implicitly addressed and may result in arbitrarily distorted outputs.

Correctly separating unrelated structure-from-motion models, which are otherwise merged into an incorrect single representation due to visual similarities, is addressed in [23, 15]. An explicit Bayesian framework to detect false positive epipolar relations from undetected features is proposed in [23], which uses belief networks for view triplets to assess the correctness of epipolar relations. The procedure to generate a 3D model is very conservative and potentially leads to unnecessarily many separate models by assuming that detected false positive edges always link completely unrelated models. False positives found by means of [23] may also refer to incorrect relations within the same model. A method to disambiguate visually similar copies of well-known landmarks reconstructed from community photo collections is presented in [15]. A combination of appearance-based clustering and geometric verification techniques is utilized to filter relevant images from unrelated ones, resulting in multiple unrelated instances or copies of widely known landmarks correctly being reconstructed.

In order to have an efficient method, we employ Bayesian inference on the abstract level of transformations between nodes (images, locally reconstructed models) and do not reconsider the association between e.g. image features and corresponding 3D points. In [2] also the correspondence between image observations and latent variables is re-evaluated and possibly reverted, but this is applied only on smaller sub-problems incorporating the recently observed data. The states of the latent variables and the associations are optimized by alternating minimization, therefore resembling the ICP method.

The method presented in [10] tries to identify consistent relative rotations before determining global camera orientations using a RANSAC scheme by sampling spanning trees from the epipolar graph. The estimated hypothesis parameters are the global orientations of all involved cameras. Evidently, the size of the epipolar graph that can be handled in such an approach is rather limited, and the author uses a sliding-window procedure to reduce the problem size. In this work we demonstrate that accumulating suitable statistics over cycles in the respective graph directly points to problematic edges, e.g. relative orientations. Further, Bayesian inference is much more tractable than random sampling for such a large hypothesis space.

Loops generated by linking smaller sub-maps are an important cue in robotics, in particular in simultaneous localization and mapping approaches. Upgrading the relative

orientations between sub-maps to absolute orientations in a common coordinate frame using explicit loop constraints is proposed in [5]. We utilize a different approach following [9, 16] to obtain globally consistent transformations from relative ones (see Section 5). Recently, the same authors suggest to consider only "compact" loop constraints derived from minimum cycle bases [6].

3. Our Method

This section describes the inference of false positive relationships between images from observation gathered by chaining local transformations. First, we describe the underlying generative model based on loop inconsistencies, followed by a depiction of how these loops are sampled.

Inference from Loop Inconsistencies Let i and j be indices of images (or some entity derived from images), and T_{ij} is a hypothesized geometric relation between i and j e.g. obtained by robust estimation from feature correspondences. T_{ij} might be the relative pose, a homography, or a similarity transformation between locally reconstructed models. We require that T_{ij} is invertible, i.e. for a given T_{ij} the reverse transformation T_{ji} can be determined. In principle, it is not necessary that $T_{ji} = T_{ij}^{-1}$ holds exactly (e.g. both directions can be estimated separately), but for the sake of simplicity we assume that T_{ji} is the exact inverse of T_{ij} in the following.

If a set of transformations $\{T_{ij}\} = \{T_e\}$ is given, such that the underlying undirected graph $G = (V, E)$ with $E = \{e = (i, j)\}$ has cycles, then chaining all transformations along a loop should result into the identity transformation (if one ignores noisy measurements for now). Let $L = (e_1, e_2, \ldots, e_{|L|})$ denote an arbitrary loop in G with length $|L|$ and starting with edge e_1, and T_L the accumulated transformation, $T_L = T_{e_{|L|}} \circ \cdots \circ T_{e_1}$. If the transformations T_e are subject to measurement noise, then the deviation between T_L and the identity I follows some noise characteristic, which can be modeled for particular problem instances. We will measure the discrepancy between T_L and I using a non-negative function $d(T_L)$.

Observe that with G being a loopy graph in most applications, there is some redundancy in the set of hypothesized transformations $\{T_e\}$. If T_L deviates substantially from the identity map for a loop L, this strongly suggest that at least one of the individual transformations T_e in the loop is incorrect and should be discarded. By accumulating these deviations over a large set of loops one can obtain the statistics needed to infer the the set of false positives. If we visualize the mean deviations for a small example (recall Figure 1(c)), then one observes that one incorrect edge influences all loops containing this particular edge, and the mean error attributed to all edges in the graph is "blurred." The main question is now how to infer the false positives from observation over cycles?

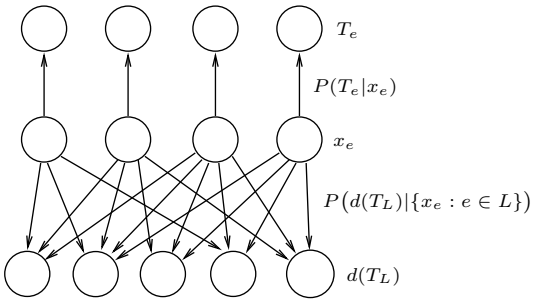


Figure 3. The Bayesian network for cycle inference.

Obviously this problem can be casted as a Bayesian inference task. We introduce latent binary variables x_e for every edge, such that $x_e = 1$ indicates a false positive edge. The event that at least one of the loop edges is a false positive is abbreviated by $x_L = 1$, i.e. $x_L = \max_{e \in L} x_e$. We have to model two prior probabilities:

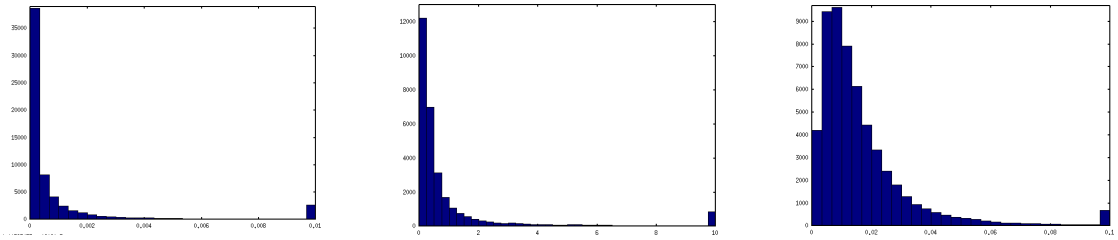
- The likelihood observing the deviation $d(T_L)$ for a loop under the assumption that none of the edges in the loop is incorrect, $P(d(T_L)|x_L = 0)$. This distribution is induced by the assumed noise model.
- The probability measuring $d(T_L)$ if at least one of the edges is a false positive, $P(d(T_L)|x_L = 1)$. As commonly employed in the literature, we generally model this likelihood by a uniform, least informative distribution. In our applications the range of $d(\cdot)$ can be easily bounded, and therefore $P(d(T_L)|x_L = 1)$ has finite support.

Optionally, a prior likelihood $P(x_e)$ can be provided for every edge, which can be determined e.g. from the confidence in the estimated transformation T_e. In our experiments we did not use the prior likelihoods (corresponding to a uniform prior on the unknowns). The structure of this belief network is illustrated in Figure 3. We are interested in an assignment for all the edge variables $x_e \in \{0, 1\}$ maximizing the joint probability

$$\prod_{L \in \mathcal{L}} P\big(\{x_e\}_{e \in L} | d(T_L)\big) \propto \prod_e P(x_e) \prod_{L \in \mathcal{L}} P\big(d(T_L)|x_L\big). \tag{1}$$

We have several options to perform (approximate) inference in this network. First, the Bayesian network can be directly converted to a factor graph representation by introducing factor nodes corresponding to the loops (and optionally factors for the unary priors). Loopy belief propagation (LBP) [14] is an efficient method for approximate inference in such graphs. We utilize LBP implementation provided by the libDAI library [17].

Since the node variables and factors and are tightly connected, and LBP does not provide quality guarantees, we



(a) Homography deviations (b) Rotation deviations (in degrees) (c) Alignment errors

Figure 4. Histograms of empirical deviations from the respective identity map. The inlier portion of the histogram roughly follows an exponential distribution.

also explored inference directly based on optimizing the energy functional corresponding to the joint probability Eq. 1. The log-likelihood of Eq. 1 reads as

$$\begin{aligned} E(\{x_e\}) &:= \sum_e l(x_e) + \sum_L l\big(d(T_L)|x_L\big) \\ &= \sum_e \Big(x_e\, l(x_e = 1) + \big(1 - x_e\, l(x_e = 0)\big)\Big) \\ &+ \sum_L x_L\, l\big(d(T_L)|x_L = 1\big) \\ &+ \sum_L (1 - x_L)\, l\big(d(T_L)|x_L = 0\big), \end{aligned}$$

with $x_L := \max_{e \in L}\{x_e\}$. We abbreviate the cost coefficients of x_e and x_L by

$$\begin{aligned} \rho_e &:= l(x_e = 1) - l(x_e = 0) \quad \text{and} \\ \rho_L &:= l\big(d(T_L)|x_L = 1\big) - l\big(d(T_L)|x_L = 0\big), \end{aligned}$$

which leads to

$$E(\{x_e\}, \{x_L\}) = \sum_e \rho_e x_e + \sum_L \rho_L x_L + const \qquad (2)$$

subject to $x_e \in \{0, 1\}$ and $x_L = \max_{e \in L}\{x_e\}$. In order to obtain a convex problem, we replace the non-convex constraints $x_e \in \{0, 1\}$ by $x_e \in [0, 1]$. Next, the (non-convex) constraint set

$$C := \big\{(x_L, x_{e_1}, \ldots, x_{e_{|L|}}) \in [0, 1]^{|L|+1} : x_L = \max_e\{x_e\}\big\},$$

linking x_L and the edge variables x_e, is replaced by the convex constraints

$$x_L \geq x_e \;\forall e \in L, \quad x_L \leq \sum_{e \in L} x_e, \quad x_L \in [0, 1], x_e \in [0, 1].$$

Overall, determining the optimal x_e in the convex relaxation setting is now a linear program, for which efficient solvers are available. In our experiments we observed that most or even all variables x_e are either 0 or 1, and only a few variables attain fractional values. Hence, a branch and bound method is also a viable (and exact) inference procedure for this set of problems.

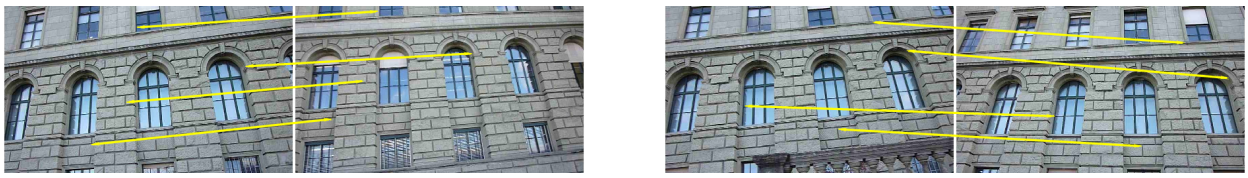
One interesting aspect of Eq. 2 is that the global solution (e.g. found by branch and bound) explains all inconsistent loops and there is no need to iterate the inference to detect additional conflicting edges. This can be seen as follows: Let $x^* = (\{x_e^*\}, \{x_L^*\})$ be the optimal solution of Eq. 2, and $\mathcal{I}$ be the indices of inconsistent edge and loop variables, i.e. $x_k = 1$ iff $k \in \mathcal{I}$. The energy Eq. 2 can be split into two parts, $E(\{x_k\}) = \sum_{k \in \mathcal{I}} \rho_k x_k + \sum_{k \notin \mathcal{I}} \rho_k x_k$. Iterating the inference procedure corresponds to fixing x_k to one for all $k \in \mathcal{I}$ and only optimizing over the unknowns $\{x_k\}_{k \notin \mathcal{I}}$. Clearly,

$$E(\{x_k^*\}) \leq \min_{\{x_k\}_{k \notin \mathcal{I}}} \left(\sum_{k \notin \mathcal{I}} \rho_k x_k \right) + \sum_{k \in \mathcal{I}} \rho_k,$$

since x_k^* is the global minimizer of $E(\cdot)$ and $x_k^* = 1$ for $k \in \mathcal{I}$. Equality is attained by setting $x_k = 0$ for $k \notin \mathcal{I}$, hence no additional conflicting edges are reported by repeating the inference.[1] Since loopy belief propagation does generally not report global solutions, repeating the inference procedure may label additional edges as conflicting. We observed only minimal changes after the first inference pass.

Cycle Generation Generating all cycle in a loopy graph is obviously intractable, hence we need to restrict the number of inspected loops to a more manageable amount. In several graph-related applications the notion of cycle bases (optionally also augmented with minimality in some sense, see e.g. [13]) plays an important role. Cycle bases allow the generation of all loops in a graph by simple vector arithmetic in $\mathbb{Z}_2$. We use the exhaustive set of cycles with length three together with loops induced by a so-called spanning tree bases. These cycle bases are derived from spanning trees of a (connected) graph by forming loops using non-tree edges. Thus, every edge in a graph not appearing in the spanning tree creates a loop together with the unique path on the tree between the respective nodes. In order to avoid explicit modeling of the transformation uncertainties

[1] Here we ignore the possibility of different, equally global solutions.



(a) Unrelated images, 228 matches

(b) Snapped to the wrong repetition, 331 matches

Figure 5. Rejected image pairs passing geometric verification using homographies. The yellow arrows indicate a few matching positions to assist the interpretation.

with respect to the cycle length, we limit the maximal loop length to six.

Since in our application we strive for redundancy in the loop statistics, we use a sequence of spanning trees to gather more cycles in the graph. The first spanning tree is a minimum spanning tree induced by estimated edge uncertainties (i.e. derived from the number of inlier correspondences). Drawing loops using a spanning tree cycle basis leads to very uneven sampling of edges in the graph, since tree edges are part of cycles much more frequently than non-tree edges. Hence, we assign the weights used to determine the subsequent spanning trees inversely proportional to the number of sampled loops containing the respective edge. This approach ensures, that loop statistics over edges are acquired roughly uniformly. If several components of a graph are connected by only a few edges, data for these edges is still sampled very frequently. But these edges are usually very important e.g. to connect weakly linked parts of a 3D reconstruction, and acquiring well supported statistics for those links is a welcome feature.

4. Application: Homography Matching

This section discusses the specific details of our approach, when the geometric transformation between nodes (i.e. images) is described by homographies. In contrast to the fundamental or essential matrix used as primary transformation associated with edges described in the next section, homographies provide a very strong cue to verify their mutual consistency via chaining transformations along loops. If a set of hypothesized homographies H_e between two images i and j forming the edge $e = (i, j)$ in the graph network is given, then we have $H_L = H_{e_{|L|}} \circ \cdots \circ H_{e_1} \propto I$, $e_i \in L$, in the ideal, noise-free case. A simple, but effective way to measure the deviation of H_L from the identity matrix I is

$$d(H_L) := \min_{\alpha} \|\alpha H_L - I\|_F = \|\tilde{H}_L - I\|_F$$

with α determined as $\alpha = \mathrm{tr}(H_L)/\|H_L\|_F^2$, and $\tilde{H}_L := \alpha H_L$. Observe that $d(H_L)$ is bounded by $\sqrt{3}$, since the elements of $\tilde{H}$ have magnitude less or equal one. In order to obtain numerically stable results, all H_e's are computed from normalized feature positions in $[-1, 1]^2$ (i.e. translated with respect to the image center and scaled by the reciprocal image width). This ensures roughly equal magnitudes for all the elements of H_e.

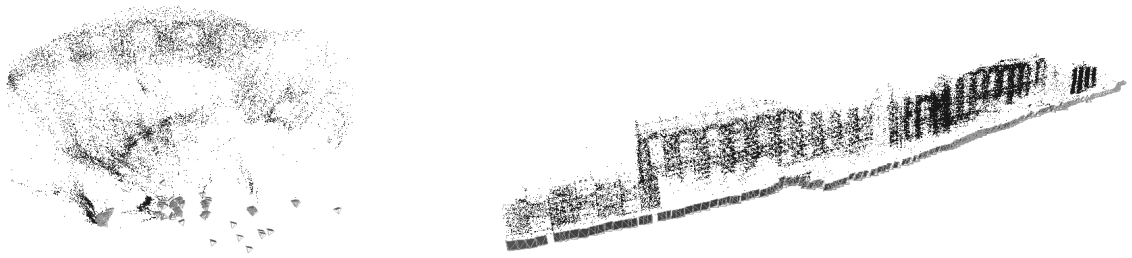
For real data, the deviations $d(H_L)$ between the concatenated homographies and the identity map is a sharply decreasing function for correct transformations (see Fig. 4(a)). Hence, we observe that $d(H_L)$ is much smaller than $\sqrt{3}$ for inliers, and the prior likelihood $P(d(H_L)|x_L = 0)$ can be modeled by an exponential distribution (we set its mean to 0.01). The observations $d(H_L)$ under presence of erroneous edges in the loop is generated by the least informative, uniform distribution, i.e. $P(d(H_L)|x_L = 1) \sim U[0, \sqrt{3}]$.

Figure 5 displays a few image pairs passing the geometric verification, but failing the stronger consistency check proposed in this section. This image sequence shows a highly repetitive, roughly planar facade (see Figure 6(b) for the 3D structure and camera path).

The currently dominant application for homography-based image alignment is the generation of panoramic images. The requirement of zero baseline between the image leads to a restricted class of homographies, for which specific minimal solvers and refinement procedures exist [4]. In turns out, that the zero-baseline constraint is already quite strong, since it essentially rules out matching e.g. repetitive visual structures residing on the same facade. Future applications of homography verification are the extension of [1] from captured videos to unorganized image collections, and the enhancement of relative pose verification discussed in the following section.

5. Application: Screening the Epipolar Graph

The prototypical example for removing incorrect pairwise geometric relations between images is computing a 3D model from visual input. In [20] several failure cases of the widely known Photo Tourism software [21] are presented and discussed. In particular, the confusion of the structure and motion pipeline due to similar visual structures and scene repetitions is addressed. Due to the incremental structure of the Photo Tourism approach, failure cases are not solely induced by erroneous relations between images, but can also be the result of drift in the camera poses, or due to numerous outliers at the feature correspondence level.



(a) W/o edge filtering (143 views registered) (b) With edge filtering (all 189 views registered)

Figure 6. Model generated by Bundler for a facade with highly repetitive elements (a) without using epipolar graph filtering, and (b) with epipolar filtering solely using relative rotations.

This section discusses the detection of conflicting edges in the epipolar graph obtained by pairwise image matching and subsequent geometric verification. By assuming (potentially only roughly) calibrated cameras, each edge $e = (i, j)$ in the epipolar graph is associated with a relative transformation $(R_{ij}, t_{ij}) = (R_e, t_e)$ relating the coordinate frames of views i and j. Since merely the direction, but not the length of the baseline t_e is known, only the relative rotations R_e can be directly chained along a path. Similar to image-to-image homographies we have a consistency criterion over loops L,

$$R_L = R_{e_{|L|}} \times \cdots \times R_{e_1} = I \qquad e_i \in L \tag{3}$$

in a noise-free setting. In principle, there are (relatively weak) consistency conditions for the translation vectors t_e (e.g. [8, 3]), but we restrict the discussion in this section to the rotation component. The next section describes stronger verification criteria, if the (relative) lengths of the baselines are known.

In a noisy setting, Eq. 3 holds only approximately, and the deviation of R_L from I is related to the likelihood for the correctness of the loop L. We use the rotation angle α_L of R_L, i.e. $\cos(\alpha_L) = (\mathrm{tr}(R_L) - 1)/2$, as the observed quantity for the sampled loops in the epipolar graph. For the inference procedure we need to model $P(\alpha_L|x_L)$. If the loop L is contaminated by an incorrect epipolar edge ($x_L = 1$), then we adopt $\alpha_L \sim U(0, \pi)$, since arbitrary accumulated rotations R_L can be generated in this case. If the loop is presumed to contain no error ($x_L = 0$), then the empirically observed angular errors can be approximately modeled by an exponential distribution (see Fig. 4(b)). We discovered in our experiments, that the exact choice of parameters for the fitted distribution (from a reasonable range) has a minor effect on the result of the inference. In our experiments we choose the mean to be 2 degrees.

In addition to extending our own structure and motion pipeline with loop consistency checks (see Section 6), we incorporated a "black-list" feature to the freely available Bundler software[2], which discards epipolar matches found to be incorrect by the proposed epipolar graph verification. Due to the specific incremental structure and motion approach employed in the Bundler software, the provided epipolar black-list is not fully utilized avoiding erroneous results. Nevertheless, solely verifying the epipolar graph can improve the resulting 3D model drastically as shown in Fig. 6. The reconstruction of a highly ambiguous facade is severely distorted without filtering the pairwise image matches (Fig. 6(a)) and correctly modeled otherwise (Fig. 6(b)). Figure 7 and 8 illustrate the identified "shortcuts" visible in the epipolar graph due to incorrect matching of repeating visual structures.

Using a holistic, non-incremental approach for structure and motion computation as described in the next section directly benefits from epipolar graph screening. We utilize a more global approach in order to avoid the problem of failed loop closure due to accumulated drift in the camera poses frequently affecting increment reconstruction methods.

6. Application: Structure and Motion

While filtering the epipolar graph solely based on the consistency of relative rotation is already a powerful tool, additional inference steps can be applied subsequently. In particular, merging partial reconstructions obtained in initial steps of a structure and motion pipeline (e.g. as proposed in [12, 11]) can benefit from verification of loop consistencies. In the following we briefly summarize our framework for structure and motion (SaM) computation[3].

SaM Computation Overview SIFT features extracted from the images are fed into a generic vocabulary tree in order to obtain a set of potentially matching images. Geometric verification based on essential matrix computation [18] is applied on these image pairs using either the known intrinsics or approximate values from the EXIF tags. These steps required to generate the epipolar graph consume about 70% of the processing time, and the subsequent stages are computationally cheaper. The epipolar graph is filtered us-

[2] http://phototour.cs.washington.edu/bundler/

[3] Corresponding software will be made publicly available at http://www.inf.ethz.ch/personal/chzach.

ing the method discussed in Section 5, and the remaining epipolar edges are used to generate image triplets. These triplets are geometrically verified. In order to be able to robustly handle undetected incorrect triplets, our approach is based on generating a set of small submodels (at most 15 views) first. These submodels are generated by random growth from a starting view and are highly redundant, such that every image participates in 10 submodels. Similarity transformations between triplets belonging to the same submodel are robustly determined, and the consistency of these transformations is verified as follows.

Triplet Verification Screening the homographies (Section 4) and epipolar relations (Section 5) identifies erroneous transformations, i.e. relations between visual entities. Inconsistent loops in the triplet graph indicate incorrectly established image triplets rather than erroneous transformations between triplets. Hence, we modify the interpretation of the latent variables (now x_k, where k is a triplet) to represent the validity of image triplets in contrast to edges/transformations between triplets. The generative model Eq. 1 remains the same otherwise. This conversion also lowers the number of latent nodes by orders of magnitude, since the number of edges in a triplet graph grows combinatorially with the connectivity in the epipolar graph.

It remains to discuss the utilized deviation $d(T_L)$ for chained similarity transformations $T_L = T_{e_{|L|}} \circ \cdots \circ T_{e_1}$. T_L reads as 4-by-4 matrix, $T_L = \left(\begin{smallmatrix} {}^{s_L}R_L & {}^{t_L} \\ \mathbf{0} & 1 \end{smallmatrix}\right)$, with $s_L = \prod_k s_k$, $R_L = \prod_{|L|}^{1} R_k$, and

$$t_L = \sum_k t_k \left(\prod_{j=k}^{|L|} s_j \right) \left(\prod_{j=|L|}^{k+1} R_j \right), \qquad (4)$$

where s_k, R_k, and t_k are the scale, rotation and translation components of T_{e_k}. As in Section 4 we use essentially $d(T_L) = \|T_L - I\|_F$ to quantify the geometric inconsistency. Since the uncertainty (variance) in the relative translations t_k is multiplied by the respective factor in Eq. 4, we scale t_L by $\left(\sum_k \prod_{j=k}^{|L|} s_j\right)^{-1/2}$ to bring the translation component to a normalized range. The empirical distribution of $d(T_L)$ is illustrated in Fig. 4(c).

After verification of triplet correctness based on their relative transformations in the same submodel, the image triplets are upgraded into a common coordinate frame and a few (at most 10) iterations of a local bundle adjustment are performed. Subsequently, similarity transformations between the submodels can be hypothesized using 3D point correspondences, which can be filtered by repeating the loop inference.

Results Upgrading the triplets in a submodel into a common coordinate frame can still fail due to undetected erroneous visual relations. Such cases are discovered if a substantial fraction (25%) of the triangulated points are outliers (with respect to the reprojection error) or very few 3D points are visible in one of the cameras. Such submodels are discarded and not considered in the final model generation. Table 1 summarizes the performance figures for several datasets with varying complexity. Adding this triplet verification step raises the number of submodels passing this criterion, and thus increases the size of the largest connected component in the final result.

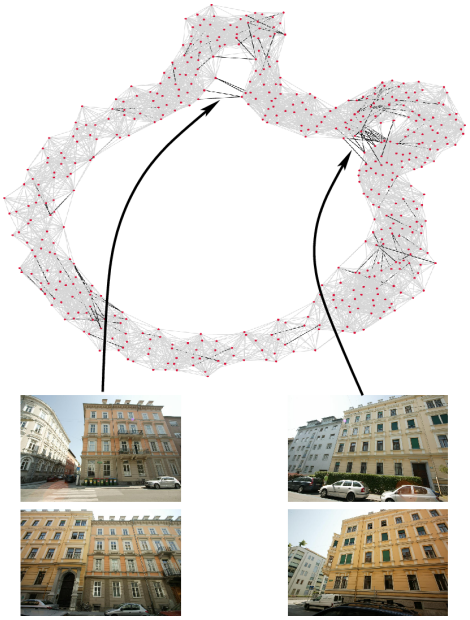


Figure 7. The epipolar graph with verified (light gray) and discarded (black) edges for the "Block" dataset (see also Figure 2), and selected image pairs corresponding to discarded edges.

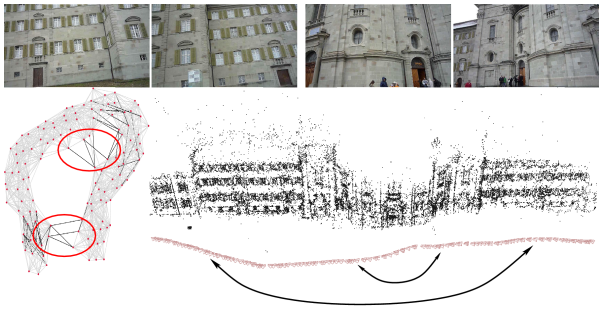


Figure 8. Epipolar graph with filtered edges (bottom left) and the reconstructed model (bottom right). The arrows indicate the approximate position of the erroneously matched images (top row, at the wings and at the central structure, respectively).

7. Discussion and Future Work

We demonstrate that enforcing the consistency of geometric relations estimated from visual input identifies conflicting relations and assists in generating improved final results in several applications. An interesting extension of

Dataset	#views	#submodels	#components	largest component	inference time (B&B)
Abbey	126	81/84	1	126 views / 29229 pts	0.6s
		84/84	1	126 views / 29250 pts	0.6s + 10s
Block	479	176/229	3	238 views / 23126 pts	17s
		215/229	1	476 views / 56230 pts	17s + 27s
Block2	3482	786/1152	26	395 views / 42686 pts	256s
		982/1152	29	550 views / 56233 pts	256s + 80s

Table 1. The effect of additional screening of similarity transformations between submodels. The first row for each dataset displays the characteristics only with epipolar graph filtering, and the second one shows the figures with all verification steps enabled. The inference times (last column) are provided separately for epipolar screening (first number) and triplet verification (second value).

this work is the propagation of faulty relations detected by some other method (e.g. using the one proposed in [23] or even by user interaction) through consistent loops. This allows to infer additional erroneous visual relations by identifying only a very small set of incorrect transformations. The method proposed in this work specifically exploits the redundancy in the matching graph, but future research will address other sources of redundant information, e.g. visibility constraints that need to be satisfied when merging separately reconstructed 3D models.

Acknowledgements: We are grateful to Arnold Irschara for providing some of the image datasets, and to Noah Snavely for releasing his Bundler software. Manfred Klopschitz is supported by FWF contract W1209 and the Christian Doppler Laboratory for Handheld Augmented Reality.

References

[1] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski. Photographing long scenes with multi-viewpoint panoramas. In *Proceedings of SIGGRAPH 2006*, pages 853–861, 2006.

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

[3] M. Brand, M. Antone, and S. Teller. Spectral solution of large-scale extrinsic camera calibration as a graph embedding problem. In *Proc. ECCV*, 2004.

[4] M. Brown, R. Hartley, and D. Nister. Minimal solutions for panoramic stitching. In *Proc. CVPR*, 2007.

[5] C. Estrada, J. Neira, and J. D. Tardós. Hierarchical slam: real-time accurate mapping of large environments. *IEEE Transactions on Robotics*, 21:588–596, 2005.

[6] C. Estrada, J. Neira, and J. D. Tardós. Finding good cycle constraints for large scale multi-robot SLAM. In *IEEE Int. Conf. Robotics and Automation*, 2009.

[7] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with application to image analysis and automated cartography. *Communication Association and Computing Machine*, 24(6):381–395, 1981.

[8] V. M. Govindu. Combining two-view constraints for motion estimation. In *Proc. CVPR*, pages 218–225, 2001.

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

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

[11] 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.

[12] A. Irschara, C. Zach, and H. Bischof. Towards wiki-based dense city modeling. In *Workshop on Virtual Representations and Modeling of Large-scale environments (VRML)*, 2007.

[13] T. Kavitha, C. Liebchen, K. Mehlhorn, D. Michail, R. Rizzi, T. Ueckerdt, and K. Zweig. Cycle bases in graphs: Characterization, algorithms, complexity, and applications. *Computer Science Reviews*, 2009.

[14] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, 2001.

[15] 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*, 2008.

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

[17] J. M. Mooij. libDAI 0.2.2: A free/open source C++ library for discrete approximate inference methods, 2008. http://www.libdai.org.

[18] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(6):756–770, 2004.

[19] K. Sim and R. Hartley. Removing outliers using the L_∞ norm. In *Proc. CVPR*, 2006.

[20] N. Snavely. *Scene Reconstruction and Visualization from Internet Photo Collections*. PhD thesis, University of Washington, 2008.

[21] N. Snavely, S. Seitz, and R. Szeliski. Photo tourism: Exploring photo collections in 3D. In *Proceedings of SIGGRAPH 2006*, pages 835–846, 2006.

[22] B. Triggs, P. McLauchlan, R. Hartley, and A. Fitzgibbon. Bundle adjustment – A modern synthesis. In *Vision Algorithms: Theory and Practice*, volume 1883 of *LNCS*, pages 298–372, 2000.

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