What Can Missing Correspondences Tell Us About 3D Structure and Motion?

Christopher Zach
VRVis Research Center/UNC Chapel Hill
cmzach@cs.unc.edu

Arnold Irschara, Horst Bischof
Institute for Computer Graphics and Vision, TU Graz
{irschara, bischof}@icg.tugraz.at

Abstract

Practically all existing approaches to structure and motion computation use only positive image correspondences to verify the camera pose hypotheses. Incorrect epipolar geometries are solely detected by identifying outliers among the found correspondences. Ambigous patterns in the images are often incorrectly handled by these standard methods. In this work we propose two approaches to overcome such problems. First, we apply non-monotone reasoning on view triplets using a Bayesian formulation. In contrast to two-view epipolar geometry, image triplets allow the prediction of features in the third image. Absence of these features (i.e. missing correspondences) enables additional inference about the view triplet. Furthermore, we integrate these view triplet handling into an incremental procedure for structure and motion computation. Thus, our approach is able to refine the maintained 3D structure when additional image data is provided.

1. Introduction

Fully automated computer vision methods for 3D scene reconstruction recently became robust enough to be used by non-vision experts. One can expect that topics like 3D modeling from publicly available photo collections [21, 9] and web-based reconstruction services [26] are the beginnings of further developments. Even if the mentioned approaches show excellent results for the demonstrated datasets, the experiences made in our work strongly suggest, that the following issues must be addressed in order to develop general purpose 3D reconstruction engines:
Scalability: A 3D modeling pipeline must cope with a large number of images. This requirement can be already fulfilled by using appropriate data structures and image retrieval techniques (e.g. [16]).

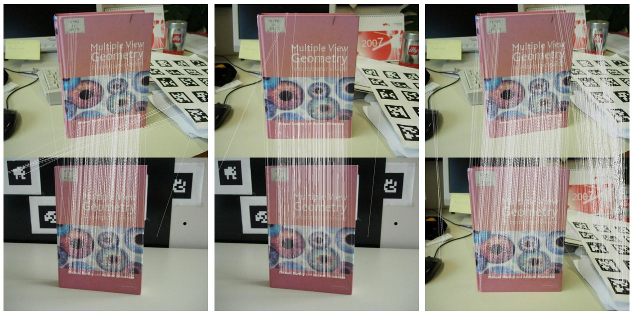
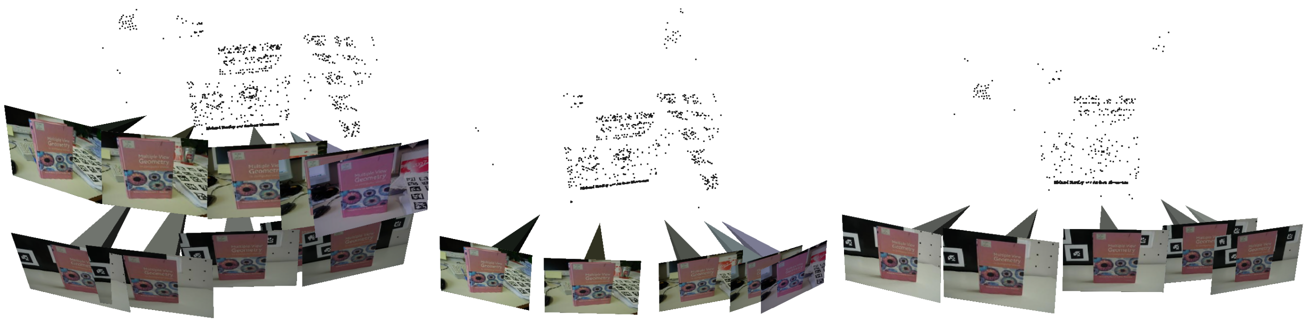

(a) $V_1 \leftrightarrow V_3$ (b) $V_2 \leftrightarrow V_3$ (c) $V_1 \leftrightarrow V_2$

Figure 1. Correspondences found for a view triplet (V_1, V_2, V_3).

Incremental reconstruction: New images may be added as they become available, hence an incremental approach is preferable. Methods like [21] for general view networks have all images available in advance to select relevant images and their motion paramaters. On the contrary, techniques developed for visual self localization and mapping (e.g. [4, 3]) maintain everything in an incremental manner, but are limited to sequences. It is desirable, that a 3D reconstruction approach has a model readily available anytime.
Order independence: It is another preferable aspect of an incremental reconstruction approach, that the order of supplied images is not relevant for the final result. One consequence of such independence is, that the reconstruction engine is able to recover from incorrect decision made earlier due to incomplete data. Although the proposed method in [21] is incremental to some extend, the order of images is encoded in the current 3D structure. Hence, it is very difficult in that approach to reestablish the correct model when additional images are supplied. These issues constitute the motivation for the work presented in this paper. At present, we are able to address only a subset of these challenges. Verifying two view geometries is fully incremental, but extracting the 3D structure from reliable epipolar geometries is not (although it can be performed quickly on request). Currently, rejecting incorrect two-view relationships is based solely on pure geometric arguments as described in the following.

Let us consider the somewhat artifical, but nevertheless illustrative example depicted in Figure 1(a): there is a substantial number of correspondences between the two views established on the common object appearing in both scenes, but the overall scenery is clearly different and the hypothesized epipolar geometry (EG) between these two views

978-1-4244-2243-2/08/$25.00 ©2008 IEEE



(a) Incorrectly merged model (b) Individual model #1 (c) Individual model #2

Figure 2. Generated 3D models by found correspondences only (a), and correctly separated models obtained by our proposed method (b) and (c).

is obviously wrong. Note, that from a geometric viewpoint we do not know the depth of the background, and the obtained EG indicated by the correspondences might be indeed right. Rejecting the EG solely on the basis of two views can be done by incorporating a prior assumption on the depths found in the scene. Such assumption limits the epipolar search from an infinite corresponding line to a bounded line segment. Alternatively, a higher understanding of the captured scene e.g. by estimating depths from monocular cues [25, 19] will allow reasoning about the validity of the hypothesized EG.

In this work we take a different route not requiring scene understanding or strong assumptions on the scene depths. Imagine, there is a third image available, which has a correct EG with one of the originally provided views. As indicated in Figure 1(b) and (c), there is again significant evidence for EGs between all three images. Nevertheless, the first and the third view share more and spatially better distributed correspondences. Under the assumption of correct EGs between all view pairs (and the correctness of the relative poses between the views) there is a substantial amount of correspondences found between the first and the third view not appearing in the predicted position in the second view. These “missing correspondences” provide a strong evidence, that there is something wrong with the EG between the first and the second view. Note, that the underlying reasoning is only performed on geometric relationships between multiple views. The camera poses and the respective 3D models are displayed in Figure 2.

This kind of reasoning about correct and false image relationships from additionally provided images is useful, if the target application is to obtain 3D models by a vision based structure from motion approach. The goal of this work is to augment a 3D reconstruction pipeline with the ability to detect and to recover previously incorrectly established EGs between images. The method proposed in this work addresses the detection of incorrect EG due to uniquely appearing but actually ambiguous objects. The proposed approach enables the seamless integration of non-monotone reasoning into structure from motion computation. Therein it is different from allmost all visual modeling approaches proposed so far.

This paper is organized as follows: Section 2 outlines relevant earlier work. Section 3 describes our approach to detect implausible two view geometries. The remaining two view relations are collected to constitute consistent reconstructions as described in Section 4. The 3D structure and motion computation for individual reconstruction is briefly sketched in Section 5, and Section 6 depicts experimental validations. Finally, Section 7 summarizes this work and indicates future research directions.

2. Related Work

Structure from motion methods can be classified into a hierarchy according to their universality. Especially early approaches (e.g. [17]) worked primarily on image and video sequences. The assumption in these methods is the direct relationship between temporal information and scene content. Improvements in wide-baseline matching [11, 6] and image retrieval [16] led to structure from motion approaches for general view configurations (e.g. [21]), where correspondences over unordered image sets are utilized. Still, there is a (hidden) assumption, that common features detected in several images induce a geometric relation between these views. This work relaxes this premise.

Recent and very inspiring work addressing structure from motion computation for unordered sets of images includes [12, 13, 22]. In [12] the authors propose a system to upgrade relative poses computed for image pairs to a full 3D reconstruction. They introduced the notions of *importance* and *reliability* of epipolar geometries between two images. The importance of an EG estimates the impact of the particular EG on the overall 3D geometry, whereas the reliability indicates the certainty about the EG. In their subsequent work [13] the separation of rotation registration and translation registration is made more explicit, and their approach was substantially accelerated by considering only appropriately selected 3D points. Identification and removal of nonexistent epipolar geometries is explicitly considered, but only by detecting image pairs with large error residua. We suppose that incorrect EG as depicted in Fig. 1 still remain unnoticed.

The approach presented in [22] shares several features with our proposed one: utilization of camera adjancency graphs and minimum spanning trees (MST), and the explicit validation step for camera poses. The camera adjacency graph is initially created by estimating the image similarities using a histogram measure. During MST construction the induced camera poses are verified by comparing dense depth maps. This is in contrast to our approach, where we first verify epipolar geometries using potential view triplets, and perform the MST construction step afterwards.

Schindler et al. [20] employ reasoning about missing (respectively invisible) feature matches to infer the temporal ordering of an unordered collection of images. Valid orderings of the images are those not violating a continuity constraint on the observed features. Thus, the main task is to solve a constraint satisfaction problem (CSP) in order to infer a suitable temporal ordering. The intractability of global CSP approaches is addressed by a greedy and local algorithm. Note that this approach is currently not fully automated: feature detection and matching is performed by a human operator.

Predicting correspondences in order to refine matches in a wide-baseline multiple view framework is part of the approach described in [6]. The aimed transitivity of matching relations guides the matching procedure, thus increasing the number of correspondences found in multiple views.

Using view triplets for structure and motion computation is well established, e.g. by utilizing the trifocal tensor [10]. In [1] triplets are used to determine structure and motion for image sequences using a robust approach for trifocal tensor computation. In [7], this idea is extended further to merge overlapping and consecutive image triplets into longer subsequences using a hierarchical approach. The implicit assumption, that three consecutive views are good candidates for trifocal tensor estimation, was relaxed in [14], where "wide tensors" spanning between appropriate keyframes are employed. Since we assume calibrated cameras, and in order to avoid special handling of dominant planar scenes we utilize the five-point method [15] in conjuction with robust rotation and translation registration.

Explicit Bayesian reasoning for 3D reconstruction is typically encountered in two very different topics: most prominently, dense stereo computation incorporating a smooth shape prior has its roots in the Bayesian formulation of the dense stereo problem [8, 2, 23], where it naturally leads to Markov random field approaches. Moreover, probabilistic methods are employed for least-squared model estimation and selection in multiple view geometry [24, 18].

3. Reasoning About View Triplets

3.1. Basic Formulation

Let V_i be a set of views ($1 \leq i \leq n$). We denote the event that two particular views V_i and V_j are related by a *visually observable* epipolar geometry by $V_i \wedge V_j$. We will denote $V_i \wedge V_j = 1$, if there is a true epipolar relationship between these views, and $V_i \wedge V_j = 0$ otherwise. Establishing or rejecting this hypothesis is based on image observations and correspondence search.

Let C^+_{ij} denote the robustly determined inliers of the potential image correspondences between V_i and V_j, e.g. by using a RANSAC approach. We do not aim on predicting the exact positions of the inliers, hence we rather focus on the number of observed correspondences, $N^+_{ij} := |C^+_{ij}|$. Assume, we can estimate the prior probability $\mathbf{P}(N^+_{ij}|V_i \wedge V_j)$ that we observe those correspondences under the assumption of $V_i \wedge V_j$ (either 0 or 1).

Now, let us look at view triplets (V_i, V_j, V_k): First, we use the abbreviation $V_i \wedge V_j \wedge V_k = 1$ for $V_i \wedge V_j = 1$, $V_i \wedge V_k = 1$ and $V_j \wedge V_k = 1$, and $V_i \wedge V_j \wedge V_k = 0$ if any EG in this triplet is wrong. Under the premise of $V_i \wedge V_j \wedge V_k = 1$ (i.e. (V_i, V_j, V_k) forms a visually well-founded view triplet), we can take correspondences between e.g. V_i and V_j, and we expect to find the respective features again in V_k (since $V_i \wedge V_j \wedge V_k$ is assumed to be true). Examining regained features in V_k would only strengthen the belief in $V_i \wedge V_j \wedge V_k = 1$. More interesting are those correspondences between V_i and V_j which are *not* found in V_k. Observing many of these missing features consequently reduces the belief in $V_i \wedge V_j \wedge V_k = 1$. Denote the correspondences between V_i and V_j not detected in V_k by $C^-_{ij\to k}$. Again, we assume that the prior probability $\mathbf{P}(N^-_{ij\to k}|V_i \wedge V_j \wedge V_k)$ can be estimated (where $N^-_{ij\to k} = |C^-_{ij\to k}|$).

For simplicity, and because the third view V_k truly adds information to the view pair (V_i, V_j), we assume that the observable events N^+_{ij} for all i and j, and $N^-_{ij\to k}$ are pairwise independent. Of course, N^+_{ij} depends on the hidden variable $V_i \wedge V_j$, but not on the truth of epipolar geometries between other views. Likewise, $N^-_{ij\to k}$ only depends on the three latent variables $V_i \wedge V_j$, $V_i \wedge V_k$ and $V_j \wedge V_k$ constituting this view triplet.

These assumptions on the statistical independence result in a directed graphical model as depicted in Figure 3. Of course, the belief network can be extended to cover all view triplets. We do not examine this approach further for the following reasons:

- Firstly, the undirected belief network after moralization results in a loopy graph, hence exact inference is expensive. The mutually recursive dependence of the latent variables $V_i \wedge V_j$ on the other variables $V_i \wedge V_k$

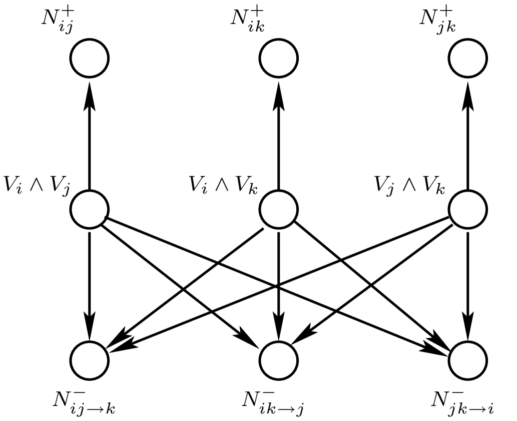


Figure 3. The Bayesian network for view triplet reasoning.

and $V_j \wedge V_k$ in the same triplet is apparent, since e.g. the belief in $V_i \wedge V_k$ depends *and* influences the belief in $V_i \wedge V_j$.

- Secondly, we aim for a primarily incremental 3D reconstruction pipeline. Performing a full global reasoning after insertion of a new image would undo the advantages of an incremental approach.

Since we have only three binary hidden variables for a view triplet, we can perform the probabilistic inference efficiently by explicit calculation of the posterior probabilities. Given the observation of the actual number of epipolar correspondences N^+_{ij} and the missing features $N^-_{ij\to k}$ for all permutations of i, j and k, we phrase the joint probability density *pdf* according to the belief network in Figure 3:

$$\begin{aligned} \mathit{pdf}(\{V_i \wedge V_j\}, \{N^+\}, \{N^-\}) = \\ \prod \mathbf{P}\left(V_i \wedge V_j\right) \times \\ \prod \mathbf{P}\left(N^+_{ij} | V_i \wedge V_j\right) \times \\ \prod \mathbf{P}\left(N^-_{ij\to k} | V_i \wedge V_j \wedge V_k\right). \end{aligned} \tag{1}$$

The posterior probabilities $\mathbf{P}\left(\{V_i \wedge V_j\} | \{N^+\}, \{N^-\}\right)$ can be directly computed from the joint density *pdf*.

The posterior distribution provides additional information on the confidence of the most likely hypothesis. If all EGs in a view triplet are accepted (i.e. $V_i \wedge V_j \wedge V_k = 1$), then the ratio

$$\frac{\mathbf{P}\left(V_i \wedge V_j \wedge V_k = 1 | \{N^+\}, \{N^-\}\right)}{\max \mathbf{P}\left(V_i \wedge V_j = 0, V_i \wedge V_k, V_j \wedge V_k | \{N^+\}, \{N^-\}\right)} \tag{2}$$

assesses the confidence in that decision with respect to a particular EG $V_i \wedge V_j$. We use the logarithm of that ratio as the actual confidence value for $V_i \wedge V_j$ with respect to the triplet (V_i, V_j, V_k). The overall confidence of an EG participating in several view triplets is the minimum of those confidences. These values are later used as edge weights of the camera adjacency graph during the generation of the individual reconstructions (see Section 4).

3.2. Choice of Prior Probabilities

View Triplet Configuration Priors Basically, each of the latent variables $V_i \wedge V_j$, $V_i \wedge V_k$ and $V_j \wedge V_k$ can take either 0 or 1, resulting in 8 possible configurations. It turns out, that only four of those configurations are plausible: the case, that all epipolar geometries are discarded can be excluded, since it needs a likely EG between two views to verify the third one using the proposed reasoning. Likewise, the case that exactly one EG is rejected, is not plausible either, since such configuration would require very specific camera poses and image content.

Consequently, these undesirable configuration can be easily excluded by setting the prior probability of those cases to zero. We consider the remaining four configurations as equally likely.

Positive Support from Pairwise Correspondences This section describes the utilized choice for the prior probabilities. We employ point features to establish correspondences between images. In particular, DoG points with associated SIFT feature vectors [11] are extracted from the supplied images. Let N_i and N_j denote the number of features points detected in view V_i and V_j, respectively. If we presume, that (V_i, V_j) forms a visually related image pair (in terms of epipolar geometry), one can expect to recover a certain fraction of the features in V_i and V_j as correspondences. In order to obtain a symmetric model, we merge the features from V_i and V_j yielding $N_i + N_j$ items. The expected number of correspondences N^+_{ij} is now “close” to $(N_i + N_j)/2$ (since one correspondence represents two detected features). Of course, it is unlikely to find correspondences for all feature points. The repeatability of the feature point detector, image content overlap, occlusions due to the scene structure, perspective distortion etc. influences the number of recovered corrspondences. We simply accumulate these effects into one probability p_1, which is the likelihood of regaining a feature extracted in one view in the other view under the assumption $V_i \wedge V_j = 1$. Hence, recovering N^+_{ij} correspondences from $N_i + N_j$ features points is modeled by a binomial distribution with parameters $N_{ij} := (N_i + N_j)/2$ and p_1:

$$N^+_{ij} \sim B(N_{ij}, p_1) \;\text{ if } V_i \wedge V_j = 1.$$

If the two images V_i and V_j are visually unrelated (i.e. $V_i \wedge V_j = 0$), finding correspondences is just coincidental. We denote the probability of finding an incidental correspondence by p_0. Again, the observed number of correspondences in this case can be approximately modeled using a binomial distribution, but now with a much lower success probability $p_0 \ll p_1$:

$$N_{ij}^{+} \sim B(N_{ij}, p_0) \ \text{ if } V_i \wedge V_j = 0.$$

Negative Belief from Missing Correspondences This section addresses the estimation of $\mathbf{P}(C_{ij\to k}^{-}|V_i \wedge V_j)$. Note, that the role of the single views in a triplet is not symmetric: $\mathbf{P}(C_{ij\to k}^{-}|V_i \wedge V_j)$ is different from $\mathbf{P}(C_{ik\to j}^{-}|V_i \wedge V_k)$ and $\mathbf{P}(C_{jk\to i}^{-}|V_j \wedge V_k)$.

Let C_{ij}^{+} denote the correspondences between view i and j, and let N_{ij} be the number of triangulated points from C_{ij}^{+} lying inside the view frustum of V_k (i.e. actually visible in view V_k). Furthermore, N_{ijk} is the number of inlier correspondences across the whole triplet. As in the pairwise case, we expect N_{ijk} not to be much smaller than N_{ij}, if $V_i \wedge V_j \wedge V_k = 1$. One might use a binomial distribution again as described in the previous section for view pairs. But note that the considered view triplets already has some support from the correspondences over all three views, i.e. some image content is common in all three views. Hence, the binomial distribution parameters q_1 (in the case $V_i \wedge V_j \wedge V_k = 1$) and q_0 (if $V_i \wedge V_j \wedge V_k = 0$) are less distinct than the values p_1 and p_0 used for the pair prior, and the appropriate choice is rather critical. Therefore, we approximate the distribution of $N_{ij\to k}^{-}$ by a Poisson distribution:

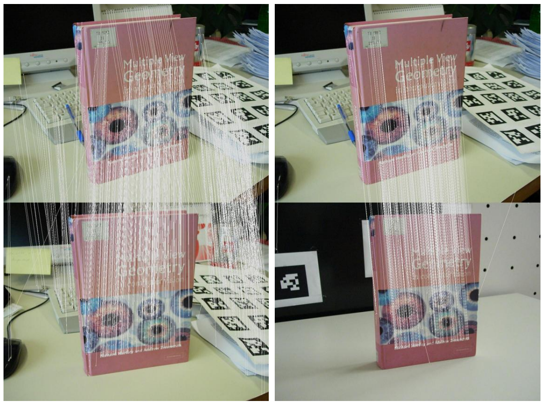
$$N_{ij\to k}^{-} \sim Pois(\lambda_1) \ \text{ if } V_i \wedge V_j \wedge V_k = 1 \tag{3}$$

$$N_{ij\to k}^{-} \sim Pois(\lambda_0) \ \text{ if } V_i \wedge V_j \wedge V_k = 0, \tag{4}$$

with $\lambda_1 \ll \lambda_0$.

3.3. Practical Considerations

Distribution of Feature Points In the discussion above we considered only the number of found or missing correspondences. The distribution of point features (or missing ones) in a particular image gives an additional cue about the prior probabilities. Consider two view pairs as depicted in Figure 4. The first image pair in Figure 4(a) (having a true underlying epipolar relation) not only has more correspondences, but these are better distributed over the image. Figure 4(b) shows the correspondences for a false epipolar view pair, where the correspondences are spatially concentrated on the "repetitive" scene content. A low number of found correspondences may indicate a false epipolar geometry or may be the result of little image structure. In order to partially disambiguate these possibilities, we replace the raw number of features by an *effective* quantity computed as follows: assume, the number of inlier point features is N. If these N points are uniformely distributed, the set of disks with centers at those points and appropriate radius r covers the whole image. If the N features are spatially grouped, those induced disks will cover only a fraction of the image. Using an estimate for r and the 2D Euclidean distance transform [5] this coverage can be easily calculated. The effective number of features is the original count weighted by their coverage. This procedure is applied to adjust the values of N_{ij}^{+} and $N_{ij\to k}^{-}$ as well.



(a) Good distribution (b) Concentrated distribution

Figure 4. Two correspondence distribution. In (a) the detected correspondences are better distributed over the image than in (b).

Suppresion of Missing Features In practical experiments it turned out, that the relatively strong assumption on the detector repeatability yields to false positive rejections of EGs. Particularly, the repeatability of the employed DoG points is low if there are substantial scale changes between the images. These incorrectly detected missing correspondences can be identified, since they are typically interspersed with found correspondences. Hence, missing correspondences are suppressed, if a found correspondence is spatially close. We use again the same estimate for the suppression radius as described in the previous paragraph.

4. Grouping of EGs

In the last section we described, how pairwise epipolar geometries are verified using a third view. It is not sufficient to collect those triplets containing only accepted pairs directly, since false epipolar geometries may still be included through undetected false epipolar pairs. Epipolar geometries are typically rejected only if there is a sufficiently strong indication for rejection. Hence, rejecting EGs is not a transitive operation.

The procedure to combine correct EGs is based on view triplets and performs several steps. First, all view triplets containing a rejected view pair are discarded. Afterwards, view triplets sharing a common view pair are collected to constitute individual reconstructions. This process can be seen as detecting connected components in a graph containing view triplets as nodes. Edges between nodes are present in this graph, if the respective view triplets have a view pair in common. Each of these resulting reconstructions can be easily registered into a common coordinate frame (see Section 5), but these reconstructions may still connect unrelated views. By adding a new image several new view triplets may be generated and the following procedure (and the steps outlined in Section 5) needs to be applied on the affected components.

A reconstruction containing the views $V_1, \ldots, V_N$ naturally induces an undirected camera adjacency graph with the edges being the verified two-view geometries (e.g. [22]). We designate such a graph as *consistent*, if there is no path from V_i to V_j for any rejected EG between V_i and V_j, i.e. $V_i \wedge V_j = 0$ implies that V_i and V_j are in different components. This definition of consistency is quite conservative, since views with incorrect EGs might be correctly part of the same reconstruction. In the current framework this definition of consistency potentially results in too many small individual reconstructions, but none of these will include a rejected EG. Relaxing this strong condition will be addressed in future work.

Splitting an inconsistent graph into several consistent subgraphs is performed using a modified version of Kruskal's algorithm for minimum spanning tree computation. Essentially, we extent the test for cycle prevention with additional checks for consistency. Two disjoint sets (i.e. individual reconstructions) are not merged, if this would yield an inconsistent tree (see Algorithm 1). The employed edge weights are just the EG confidence values calculated from the posterior probabilities (recall Section 3), thus highly reliable view pairs are merged first. Of course, our algorithm delivers a forest instead of a spanning tree.

Algorithm 1 Modified Kruskal's method

```
Procedure  $F$ = Modified MST
Input: A potentially inconsistent weighted graph $G = (V, E)$
  $F := \emptyset$; $\forall i$ : MAKE-SET($DS, i$)
  for each edge $(i, j) \in E$ in order of nonincreasing weight
  do
    $r_i \leftarrow$FIND-SET($DS, i$); $r_j \leftarrow$FIND-SET($DS, j$)
    if $r_i \neq r_j$ and
        $\forall k \in$ SET($DS, i$), $\forall l \in$ SET($DS, j$): $V_k \wedge V_l = 1$
    then
      UNION-SET($DS, i, j$)
      $F \leftarrow F \cup (i, j)$
    end if
  end for
```

5. 3D Structure Extraction

After the verified pairwise epipolar geometries are collected into a set of consistent reconstructions, the initial structure and motion remains to be determined. Currently, we follow an approach inspired by [13] to obtain the extrinsic camera parameters, which relies only on the robustly estimated relative poses. If a newly added image does not change the topology of the EGs (i.e. two or more reconstructions are not merged and no additional EG is rejected), an initial estimate of its pose and respective 3D points can be immediately determined (e.g. by perspective pose computation). In all other cases the structure and motion parameters of affected individual reconstructions are determined as follows:

First, the given relative rotations $\{R^{ij}\}$ between two views are upgraded into a consistent set of rotations $\{R_i\}$ by solving the overdetermined system of equations, $R_j = R^{ij} R_i$. As described in [13] we solve the system initially for approximate rotation matrices and subsequently enforce the orthonormality of R_i using the SVD. The registered translations are computed using a two-step procedure in order to always obtain physically meaningful results. At first, the global scales are determined using a linear approach. Separating scale and translation estimation has the advantage, that positive scales can be easily enforced.

With the knowledge of the registered rotations and scales, the coordinate frames of view triplets differ only by translational offsets, which are determined linearly as well. Algebraic least squares solutions for the offsets and the camera centers are obtained using the respective normal equation. The initial 3D structure is created by triangulation of the inlier correspondences. Finally, a metric bundle adjustment is applied.

6. Results

This section provides additional real-world examples, where incorporating the proposed approach employing missing correspondences results in substantially enhanced reconstructions. In these experiments the basic probability parameters $(p_0, p_1, \lambda_0, \lambda_1)$ are set to $(0.001, 0.1, 0.95, 0.2)$, respectively. Figures 5(a) and (b) depict example views of highly similar, but nevertheless different facades. Without the incorporation of our proposed method all views are incorrectly combined into one common reconstruction. Enabling the rejection of two view geometries results in splitted 3D models as illustrated in Figure 5(c) and (d). The second example is an indoor environment with similar fire extinguisher appearing in the images (Figure 6(a)–(c)). This common object acts as an "visual anchor" linking all views into a common frame (Figure 6(d)). Separation of individual reconstructions is not perfect in this case, since a few images actually belonging to scenery depicted in Figure 6(c) are attached to the middle one. This example shows, that a sufficient number of absent features is required for perfect reasoning.

Note that the purpose of these examples is to provide evidence, that incorrect EGs can be detected and handled even with very limited and visually misleading image data.

7. Conclusion and Future Work

In this work we argue, that finding correspondences and detecting outliers among those are generally not sufficient to obtain a correct 3D reconstruction. Hence, we propose an approach for structure and motion computation that is able to detect incorrect two view geometries by reasoning about missing correspondences retrieved from view triplets. Generally, our method can be used to augment existing 3D reconstruction pipelines with little additional costs. Moreover, our proposed approach naturally fits into incremental systems for online 3D reconstruction.

Future work will address relaxing some of the strong assumptions utilized in the proposed framework. The employed prior probabilities on the number of detected correspondences is based on a rather simplistic model. We intend to refine this model in future research. Additionally, rejected epipolar geometries need not to reside in disjoint reconstructions. Adding a reasoning framework for possible and prohibited camera poses will enable the generation of better connected reconstructions and a improved handling of more general repetitive patterns.

References

[1] P. A. Beardsley, P. H. S. Torr, and A. Zisserman. 3D model acquisition from extended image sequences. In *European Conference on Computer Vision*, pages II:683–695, 1996.

[2] P. N. Belhumeur. A bayesian approach to binocular stereopsis. *Int. Journal of Computer Vision*, 19(3):237–260, 1996.

[3] A. J. Davison, I. Reid, N. Molton, and O. Stasse. MonoSLAM: Real-time single camera SLAM. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 29(6):1052–1067, 2007.

[4] E. Eade and T. Drummond. Scalable monocular SLAM. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 469–476, 2006.

[5] P. F. Felzenszwalb and D. P. Huttenlocher. Distance transforms of sampled functions. Technical report, Cornell Computing and Information Science, 2004.

[6] V. Ferrari, T. Tuytelaars, and L. Van Gool. Wide-baseline multiple-view correspondences. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 718–725, 2003.

[7] A. W. Fitzgibbon and A. Zisserman. Automatic camera recovery for closed or open image sequences. In *European Conference on Computer Vision*, page I: 311, 1998.

[8] D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *International Journal of Computer Vision*, 14:211–226, 1995.

[9] M. Goesele, N. Snavely, B. Curless, H. Hoppe, , and S. Seitz. Multi-view stereo for community photo collections. In *IEEE International Conference on Computer Vision (ICCV)*, 2007.

[10] R. Hartley. A linear method for reconstruction from points and lines. In *IEEE International Conference on Computer Vision (ICCV)*, pages 882–887, 1995.

[11] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int. Journal of Computer Vision*, 60(2):91–110, 2004.

[12] D. Martinec and T. Pajdla. 3d reconstruction by gluing pairwise euclidean reconstructions, or "how to achieve a good reconstruction from bad images". In *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, 2006.

[13] D. Martinec and T. Pajdla. Robust rotation and translation estimation in multiview reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2007.

[14] D. Nistér. Reconstruction from uncalibrated sequences with a hierarchy of trifocal tensors. In *European Conference on Computer Vision (ECCV)*, pages I: 649–663, 2000.

[15] D. Nistér. An efficient solution to the five-point relative pose problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 26(6):756–770, 2004.

[16] D. Nistér and H. Stewenius. Scalable recognition with a vocabulary tree. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2161–2168, 2006.

[17] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *Int. Journal of Computer Vision*, 59(3):207–232, 2004.

[18] M. Pollefeys, F. Verbiest, and L. Van Gool. Surviving dominant planes in uncalibrated structure and motion recovery. In *European Conference on Computer Vision (ECCV)*, pages 837–851, 2002.

[19] A. Saxena, S. H. Chung, and A. Y. Ng. Learning depth from single monocular images. *NIPS*, 18, 2005.

[20] G. Schindler, F. Dellaert, and S. B. Kang. Inferring temporal order of images from 3d structure. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–7, 2007.

[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] K. L. Steele and P. K. Egbert. Minimum spanning tree pose estimation. In *International Symposium on 3D Data Processing, Visualization and Transmission (3DPVT)*, pages 440–447, 2006.

[23] J. Sun, H. Y. Shum, and N. N. Zheng. Stereo matching using belief propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 25(7):787–800, 2003.

[24] P. H. S. Torr. Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *Int. Journal of Computer Vision*, 50(1):35–61, 2002.

[25] A. Torralba and A. Oliva. Depth estimation from image structure. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 24(9):1226–1238, 2002.

[26] M. Vergauwen and L. Van Gool. Web-based 3D reconstruction service. *Mach. Vision Appl.*, 17(6):411–426, 2006.

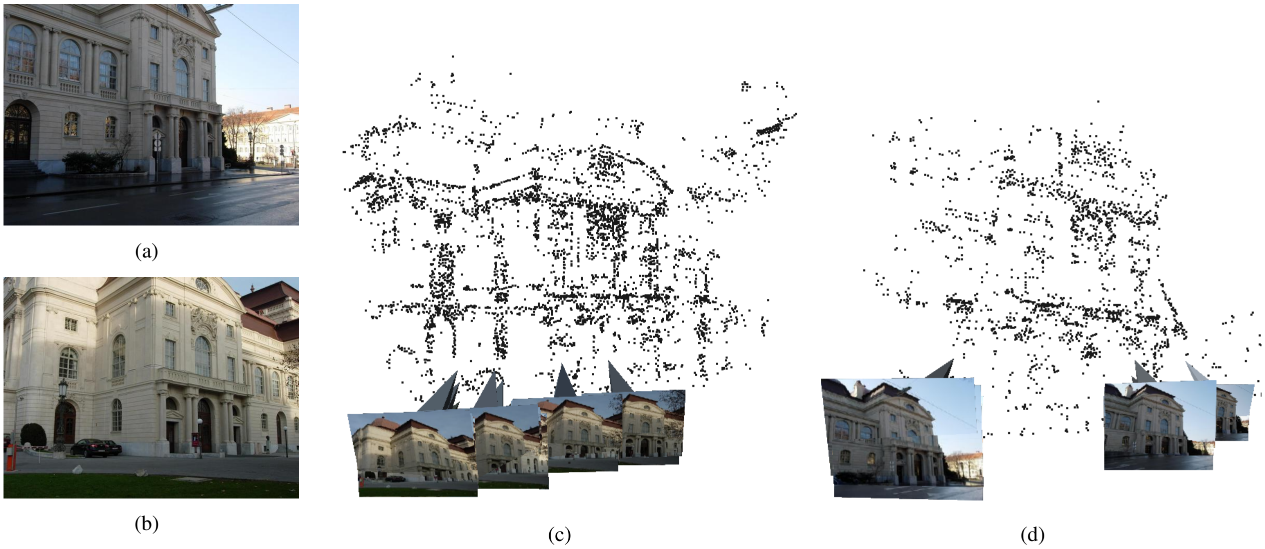


Figure 5. (a)–(b) Source images showing symmetric facades representing opposite sides of the building (23 in total). Small panoramas are used to capture the full height. (c)–(d) The two separated reconstructions are shown. The 3D reconstruction obtained without EG verification incorrectly merges all views (not shown here).

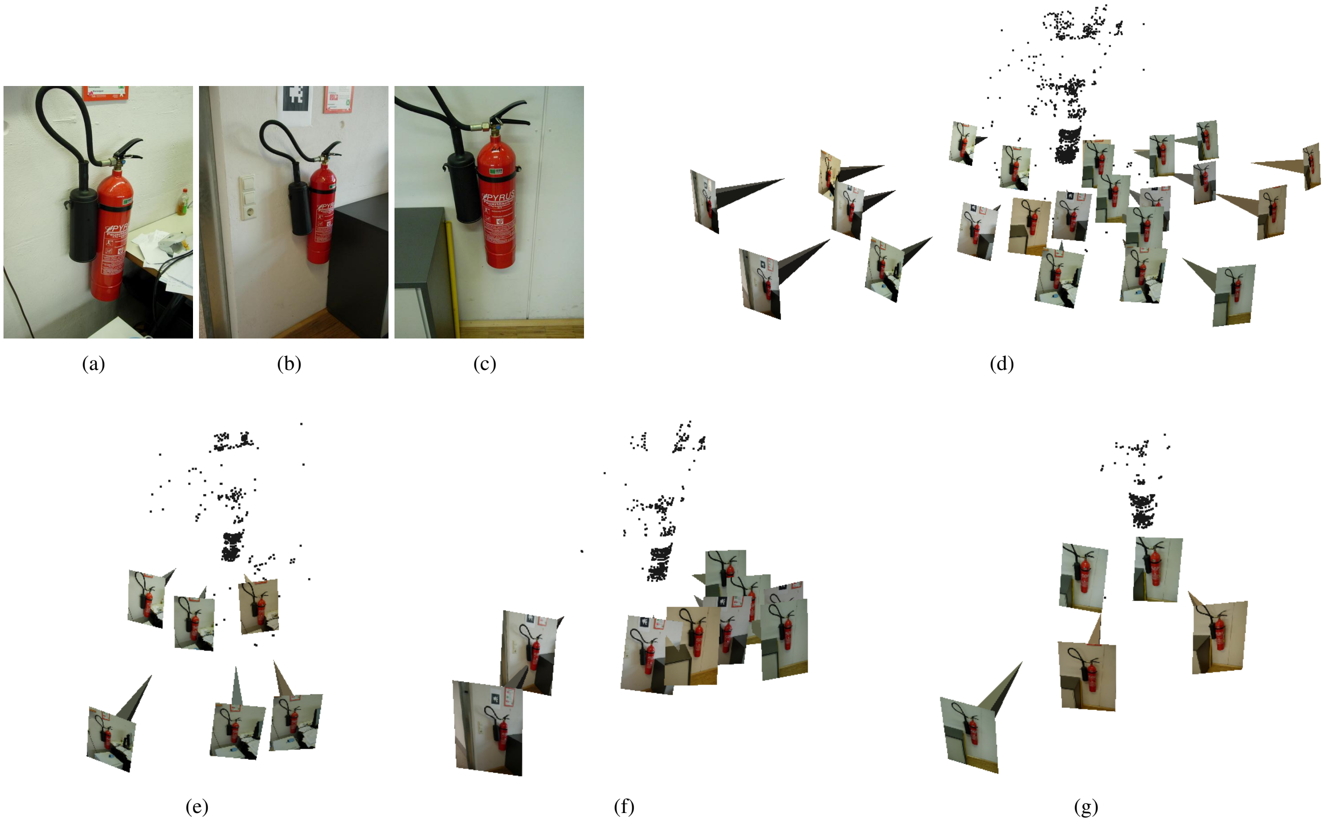


Figure 6. (a)–(c) Source images showing the same kind of fire extinguisher at different places (out of 24). (d) Incorrectly fused result of structure and motion without EG verification. (e)–(g) The three individual components obtained by our method. The separation between (f) and (g) is not perfect due to insufficient background structure.