TOWARDS COMPLETE, GEO-REFERENCED 3D MODELS FROM CROWD-SOURCED AMATEUR IMAGES

W. Hartmann[a] *, M. Havlena[b], K. Schindler[a]

[a] ETH Zürich, Photogrammetry and Remote Sensing, 8093 Zürich, Switzerland
[b] ETH Zürich, Computer Vision Laboratory, 8092 Zürich, Switzerland

Commission III, WG III/1

KEY WORDS: structure from motion, crowd-sourcing, geo-reference

ABSTRACT:

Despite a lot of recent research, photogrammetric reconstruction from crowd-sourced imagery is plagued by a number of recurrent problems. *(i)* The resulting models are chronically incomplete, because even touristic landmarks are photographed mostly from a few "canonical" viewpoints. *(ii)* Man-made constructions tend to exhibit repetitive structure and rotational symmetries, which lead to gross errors in the 3D reconstruction and aggravate the problem of incomplete reconstruction. *(iii)* The models are normally not geo-referenced. In this paper, we investigate the possibility of using sparse GNSS geo-tags from digital cameras to address these issues and push the boundaries of crowd-sourced photogrammetry. A small proportion of the images in Internet collections ($\approx$ 10%) do possess geo-tags. While the individual geo-tags are very inaccurate, they nevertheless can help to address the problems above. By providing approximate geo-reference for partial reconstructions they make it possible to fuse those pieces into more complete models; the capability to fuse partial reconstruction opens up the possibility to be more restrictive in the matching phase and avoid errors due to repetitive structure; and collectively, the redundant set of low-quality geo-tags can provide reasonably accurate absolute geo-reference. We show that even few, noisy geo-tags can help to improve architectural models, compared to puristic structure-from-motion only based on image correspondence.

1. INTRODUCTION

Image-based 3D reconstruction of buildings and architectural monuments is a well studied problem in photogrammetry and computer vision. Traditionally, one would travel to the site and acquire the necessary images. Such a carefully planned recording will ensure complete coverage and sufficient pairwise overlap, such that camera orientation and subsequent (point-wise) reconstruction become easy – nowadays fully automatic reconstruction is available in many commercial systems.

The rapid development of image sharing sites and social networks (a staggering 1.9 billion images are uploaded to the Internet every day) has raised the possibility to find images for 3D reconstruction on the Internet, rather than go into the field. Millions of images of the human habitat are available on Internet photo sharing sites and social networks. The promise of crowd-sourced photogrammetry is to make use of this treasure trove, especially for city modelling, architecture, and heritage.

The crowd-sourcing approach to photogrammetry poses additional challenges, because the data are not recorded with photogrammetric requirements in mind. A key property of Internet image collections is their extremely uneven distribution. For some viewpoints of popular landmarks, which are most photogenic and easily accessible, there are thousands of nearly identical images. Other parts are only covered sparsely or not at all (e.g. less attractive viewpoints at the back of a building, and parts that require non-standard equipment like tripods or extreme wide-angle lenses). Moreover, the imaging conditions vary wildly, as individual photographs are taken with different cameras and/or in different weather conditions. Some images might be totally incompatible, e.g. pictures taken at night, or showing temporary objects on the facade like scaffolding or posters. All these cases must be filtered automatically when using large photo-collections, placing high demands on the orientation and reconstruction pipeline.

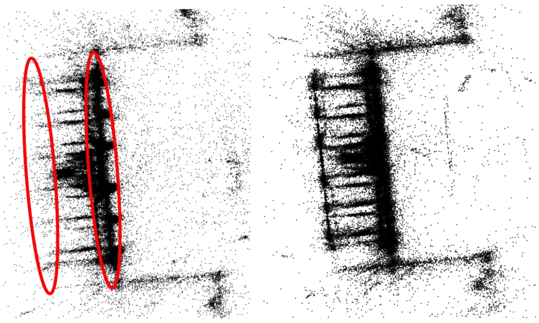


Figure 1. Reconstructions from uncontrolled images are sometimes compromised by duplicate structure and/or missing parts (left). Using the GNSS tags available for a small proportion of the images can mitigate the problem (right).

While modern software can deal with difficulties like unknown and varying focal length, lighting changes, and incompatible images, a number of common problems remain. Perhaps the most severe limitation is that more often than not the reconstructed 3D model will be incomplete. In some cases a part of the object of interest is really not covered by any images – that case cannot be solved without targetted recording. However, often the problem is more subtle: there would be suitable views in the data, but these are not found, because there are not enough tiepoint correspondences to connect them to the model. A frequent special case is that different sides of a building are reconstructed separately, but can, in the absence of tie points, not be recognised as belonging to the same object and therefore also not be transformed into a common coordinate system.

*Corresponding author: hartmann@geod.baug.ethz.ch

A second problem, which is less frequent, but, if undetected, leads to unusable models, is repetitive structure. In architectural reconstruction this seemingly exceptional situation is surprisingly frequent: identical copies of architectural elements such as windows, columns, etc. are routinely re-used multiple times on the same building, not only within a single facade, but also on different sides, giving rise to (partial) rotational symmetries. Such repetitive (or "duplicate") structures induce sets of multiple, mutually consistent but false correspondences between spatially distinct copies, and as a consequence grossly wrong camera poses. Their detrimental effect is two-fold. On the one hand, cameras that are attached to the model in a wrong pose are missing elsewhere, sabotaging the attempt to obtain a complete model. On the other hand, an orientation error in one camera propagates to further cameras that are (correctly) matched to it, leading to hallucinated 3D structure at a wrong location (see Fig. 1).

Finally, puristic reconstruction only from image data can by design only solve for relative camera poses, so the resulting model is scaled arbitrarily and not geo-referenced. Going to the field to measure ground control points would defeat the purpose of crowd-sourcing, so it would be desirable to also recover absolute orientation from the crowd-sourced data.

Which additional information beyond the pixel intensities is available in crowd-sourced imagery, that could be used to solve this problem? Many digital cameras (including those on cell phones) have a cheap, consumer-grade single-frequency GNSS receiver, and sometimes the GNSS coordinate at the time of taking the picture is stored as meta-data in the image's EXIF header. These coordinates, sometimes referred to as *geo-tags* are currently not available for the majority of images[1], and they are not accurate enough to be used directly – errors in some cases exceed 10 m. But, just like the approximate focal length that is also often part of the meta-data, they are good enough as initial values, and can constrain the problem sufficiently to overcome some of the problems described above.

The contributions of this paper are *(i)* we show that even sparse and noisy geo-tags are in some cases sufficient to approximately align partial reconstructions of the same building, such that image-based methods can again take over and connect the images with corresponding points. *(ii)* we analyse the pairwise image correspondences in a set of crowd-sourced images in order to find and sever weak connections that could possibly be due to repetitive structure. This leads to separate partial reconstruction, which can again be joined using GNSS geo-tags. *(iii)* we empirically investigate the possibility to geo-reference the reconstruction with the help of the available GNSS tags. While individual tags are too inaccurate for absolute orientation at the scale of single buildings, the greatly redundant set of tags gives visibly better geo-reference. Here it seems to help, rather than hurt, that each image was recorded at a different time, reducing the correlation in the GNSS errors.

2. RELATED WORK

Research into photogrammetric 3D reconstruction from personal photographs on the Internet (a.k.a. "neo-photogrammetry" Leberl (2010)) was triggered by Snavely et al. (2006). That work was the starting point for a flurry of activity, mainly aiming to render the initial approach more efficient Snavely et al. (2008); Li et al. (2008); Agarwal et al. (2009); Frahm et al. (2010). Moreover, dense matching and surface reconstruction at large scale were also integrated into the original, sparse reconstruction pipeline Goesele et al. (2007); Frahm et al. (2010).

Camera pose estimation and 3D reconstruction from arbitrary, uncontrolled (perspective) images has reached a certain maturity, as evidenced by several open-source and commercial software packages (e.g. *Bundler, VisualSFM, Pix4D, Acute3D*). Nevertheless, a number of important open problems remain, which is also why crowd-sourced photogrammetry has not yet been widely adopted for surveying purposes. The main issues in our view are *(i)* that more often than not the models are incomplete and show only some of the sides of a building or scene; *(ii)* that repetitive structures and symmetries lead to incorrect correspondences and gross errors (hallucinated and/or missing parts) of the 3D models; and *(iii)* that, even if several useful parts of a larger scene have been reconstructed correctly, the models are not geo-referenced with appropriate accuracy.

To start out, we note that crowd-sourcing in the strict sense, without any mechanism to ensure complete coverage, will always only be applicable for a limited number of objects that are frequently photographed and easily accessible from all sides – as long as no images exist for any important part of a scene the reconstruction will remain incomplete, and dedicated field-work or complementary data sources are required. We point out that our definition of crowd-sourcing does not include image acquisition as a side-product of other commercial activities, such as mounting cameras on taxis, delivery vans, public transport vehicles, etc. While that strategy certainly can provide better coverage, we see it as a form of (semi-)systematic mobile mapping, given that the camera system is designed and dedicated for the purpose of photogrammetric reconstruction, and there will have to be a economic incentive for its installation and use. To encourage complete recording without a monetary reward Tuite et al. (2011) proposed to embed mapping in an online game, in such a way that views which cover previously unmapped parts receive higher scores. While the gaming approach could be very interesting for specific situations, it appears that in general maintaining and supervising it may be as time consuming as photogrammetric recording. The situation is similar for users who take pictures with the explicit aim of photogrammetric modelling, but without training ("map lovers" or "casual mappers" Heipke (2010)). While they do have an intrinsic motivation to obtain complete models, they would have to be supported with carefully interfaces, as can be seen from the large number of unsuitable image sets uploaded to web-based structure-from-motion services such as *Arc3D* (`www.arc3d.be`, Tingdahl and Van Gool (2011)) or *CMP SfM* (`http://ptak.felk.cvut.cz/sfmservice`, Heller et al. (2015)).

Other work has proposed to use high-level information such as orthogonality, symmetry or known architectural patterns to guess the best reconstruction for unseen object parts Dick et al. (2002); Mathias et al. (2011); Cohen et al. (2015). While the resulting models are visually pleasing and useful for graphics and visualisation, they remain an "informed guess" and cannot replace actual measurements.

There is however also a more subtle cause for incompleteness: in quite a few cases the necessary images to cover a certain part are actually present in the data, but are missed at the stage where the reconstruction pipeline discovers the correspondences and builds up the camera network. Also frequent is the case that groups of overlapping views for different parts (e.g. opposite sides of a building) are found, but reconstructed independently, because

[1] Note, map coordinates *not* measured by GNSS but obtained interactively from the user, e.g. by clicking on a map, are also often called "geo-tags". In our experience, these are too inaccurate to be used in a meaningful way.

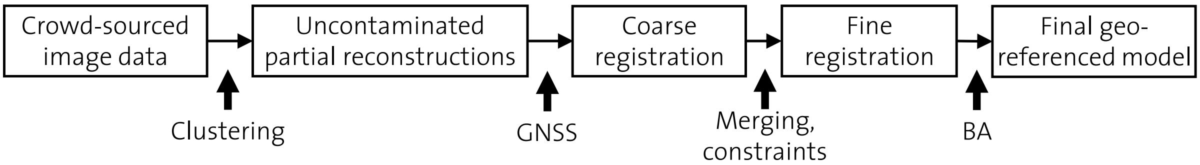


Figure 2. Overview of the proposed method.

there are not enough correspondences to reliably connect them. Lowering the threshold so that partial networks are joined more aggressively unfortunately is not *per se* a viable solution, because then the matches on repetitive visual structures induce wrong connections. These two causes of incompleteness can be mitigated with the methods described in this paper: weakly connected fragments can be merged into bigger models with the help of (approximate) geo-reference. The possibility to repair broken connections in turn makes it possible to detect and remove matches caused by duplicate structures more aggressively.

Duplicate structures due to repeated architectural elements are rather frequent in urban scenes. Their effect on structure-from-motion systems is that some relative orientations, which are based on points on those duplicate structures, are geometrically consistent but nevertheless wrong. As a consequence different (internally correct) parts of the camera network are attached to each-other in an incorrect manner, leading to “ghost structures” that appear at the wrong location in the object coordinate system, and to unwarranted holes in the reconstruction at the position where those “ghosts” should have been placed. The problem has already received some attention Wilson and Snavely (2013); Heinly et al. (2014). The principle to deal with duplicate structure is always the same: inspect the distribution of corresponding points, including the unambigious ones, to identify potentially wrong matches. Then, break the camera network and either discard the smaller part, or explore where else it could be connected, in the hope that this will fix the reconstruction. This works quite well, as long as the overlap between individual parts in the correct orientation is sufficient to find that arrangement, and merge the parts accordingly. We go one step further and use the available geo-tags to find a plausible orientation of the parts, so that one can actively search for correct correspondences.

Finally, we also use the geo-tags as evidence for absolute geo-referencing. Positioning cameras with GNSS is an obvious practice, but the locations recorded by consumer cameras and mobile phones are too inaccurate: empirically, one must expect errors on the order of 5 m, which is about 10-25% of a the size of a typical building. However, for geo-referencing purposes crowd-sourced image collections have an advantage over dedicated recordings: the pictures have been taken across different times, days and seasons. Therefore, the geo-tags, while individually noisy, are only weakly correlated (although some degree of correlation of course remains due to the fixed geographic latitude, neighbouring objects, etc.) Including them as soft constraints in the final bundle adjustment, with high uncertainty, therefore cancels out a lot of the individual error and empirically seems to achieve usable geo-reference. Our approach is related to others that also aim for absolute geo-reference of crowd-sourced models Strecha et al. (2010); Üntzelmann et al. (2013). However, they rely mostly on 2D building outlines from a GIS to geo-register partial 3D models, and only use GNSS as an optional additional cue. Possibly their GNSS coordinates had strong systematic errors because they operated in narrow streets. Other researchers prefer to only use GNSS coordinates (and other geo-tags) as initial values for either geo-registration to aerial images Kaminsky et al. (2009); Wang et al. (2013) or image orientation Crandall et al. (2011), but do not include them in the actual adjustment. In our view consumer-grade GNSS tags are a weaker requirement than accurate ortho-images or map layers of sufficient resolution. More and more photos are geo-tagged – often without the user even noticing it – while high-quality geo-data are not available in many parts of the world. If available, they do potentially allow for more accurate geo-referencing than consumer-grade GNSS alone.

3. METHOD

3.1 Crowd-sourcing image data

Many different image repositories are available on the Internet, which vary greatly in terms of the number of images for a given location, the image quality, and also how easy it is to identify and download the relevant images. In this work (see method overview in Fig. 2) we collect images from Flickr, which has been the major data source for crowd-sourced photogrammetry Snavely et al. (2006); Frahm et al. (2010). Flickr offers a reasonable compromise: a sufficient number of images are available for many locations, at the same time the quality and resolution of the photographs is a lot better than the average Internet picture. Also, the number of unusable images unrelated to the geographic location (portraits and pictures showing concerts, parties and similar events, which happen to be taken at the location of interest) is comparatively small. Moreover, querying Flickr is straightforward, as there is an API to download images, either by location or with textual keywords. The (hand-clicked) majority of image locations as well as the keywords are added by the users who upload the images, and both options have been shown to work for 3D reconstruction. We prefer to query by location, using a modified version of the script of Hays and Efros (2008). This strategy in our experience returns a larger number of relevant images, and avoids problems with keywords in different languages, as well as ambiguous names (e.g., the keyword “Nevsky Cathedral” will return images from over ten different cities that possess a church with that name). We also note that the time interval for the query must be specified, which can be useful in case of architectural changes, construction works, etc.

Besides the pixel values, crowd-sourced images do often come with a small amount of meta-data, stored in their EXIF header. E.g., most SfM methods parse the header to obtain an approximate value for the focal length. Many cameras, from mobile phones to professional DSLRs, have an integrated (single frequency) GNSS antenna and can also store the location where the picture was taken in the header. Processing must not rely on always having meta-data. In fact only about half of all images on Flickr have an EXIF header, e.g. some image editing packages remove the header, and $<$ 10% of all images have a GNSS coordinate. Nevertheless, one can expect to find at least one GNSS location in most image networks of 50 or more images, and in the future that number will increase, as more and more cameras are produced with built-in GNSS antennas.

3.2 Finding uncontaminated (partial) reconstructions

In the crowd-sourced setting the primary input consists of all images found for a certain geographic area or keyword, but it is not guaranteed, and in fact very unlikely, that all those images form a single, connected camera network. Dividing the data into smaller, connected networks (and a trash bin for irrelevant images) is part of the reconstruction process. To that end one uses heuristic criteria that determine when an image can be added to the network. Roughly speaking the two main criteria are whether enough interest point matches could be found to a (small) number of other images, and whether these matches allow for relative orientation with low residuals. These heuristics can sometimes be too strict, so that although the required images are among those fed into the algorithm, the reconstruction remains incomplete, meaning that a significant part of the "good" images could not be connected to the network.

But the decision can also go wrong in the other direction and accept relative orientations that are geometrically consistent, but nevertheless wrong. The main reason why configuration of multiple tie-points can be matched although they are not images of the same object points are duplicate structures that appear multiple times, e.g. on different walls of a building. Note that the duplicate structure itself is in fact reconstructed correctly: if, say, multiple copies of a window exist that really are identical within the measurement accuracy, then using all of them to reconstruct the window in 3D is actually a good thing, because the redundancy will suppress noise. Unfortunately, many images depict not only the duplicate part[2], but also other structures that are not symmetric, and on which tie-points will be found that (correctly) connect to further images, resulting in "ghosting" effects where object appear in the wrong place, such as walls that penetrate eachother.

At the origin of the problem are unwarranted network connections due to wrong tie-point matches. A natural way to overcome the problem is thus to detect weak connections supported by an unusually small number of tie-points, and sever those connections to instead obtain smaller, but correct networks. In fact, a related step is built into most large-scale SfM systems anyway: to bring down the computational cost, the input images are clustered into subsets such that each image in a subset has sufficient overlap with a few other cluster members. In the recent work of Havlena and Schindler (2014) the overlap between two images' fields of view is approximated by counting (putative) interest point matches. This suggests a simple trick to find weak connections: apply the same idea with a stricter threshold, so that connections with too few supporting feature matches never make it to the relative orientation procedure.

In more detail the *VocMatch* method Havlena and Schindler (2014) quantizes feature descriptors into a vocabulary that is so large (16 Mio. words) that visual words appear at most once in a large majority of images. In that way the quantization directly induces multi-ray correspondences and no explicit descriptor comparison is required. By counting the number of correspondences between all pairs of images one can easily obtain a (symmetric, integer-valued) matching matrix Q. That matrix is then inspected to find image pairs for which $q_{ij}/\min(q_i, q_j) > q_{\min}$. Recording those image pairs produces a binary matrix B, with entries $b_{ij} = b_{ji} = 1$ for image pairs that have enough tie-points to be part of the same camera network. The original *VocMatch* method recommended $q_{\min} = 1.5\%$.

For our task we simply raise $q_{\min}$, to the point where the b_{ij} indicate that images i and j can not only be oriented w.r.t. eachother, but are also unlikely to contain duplicate structure. The underlying assumption is that duplicate structures will not fill the entire image. As a consequence the fraction $q_{ij}/\min(q_i, q_j)$ of potential tie-points will be smaller.[3] Using a stricter threshold will lead to more and smaller clusters – empirically, we find that connections due to duplicate structure are among the first ones to break, such that the smaller clusters are indeed more correct, as expected, see Section 4. Too small clusters with < 50 images are simply discarded. [4]

The small, clean clusters are individually reconstructed with standard methods. For our experiments we use Bundler Snavely et al. (2008), which is slow, but rather reliable due to its conservative strategy where the growing camera network is frequently rebundled. (For computational efficiency we replace the bundle adjustment module with multicore bundle adjustment Wu (2013).) The price to pay for clusters that are with high probability clean is a fragmented reconstruction consisting of multiple partial 3D models (or, equivalently, camera networks). We go on to explain how these are merged into a complete model using auxiliary GNSS information.

3.3 GNSS-assisted model merging

Each 3D model found in the previous step consists of camera orientations and 3D structure points and represents a part of the same landmark, in its own, local object coordinate system. Generating such error-free sub-models can be a goal in itself Wilson and Snavely (2013), but we want to go one step further and reconnect them into a complete reconstruction that covers the entire object. What makes this task challenging is that by construction different models have little overlap (respectively, few common points) – otherwise they would have not been separated in the first place. More aggressive matching between images from different clusters could potentially discover the necessary matches, but will also bring back the duplicate structure problem, since ambiguous elements will be present in different clusters, if the previous step was successful.

Here, we come back to the GNSS location information that comes for free with some of the images. About 5-20% of the images (depending on the region) will have GNSS coordinates in their EXIF headers. Those coordinates are recorded with extremely cheap consumer-grade antennas and have absolute (x, y)-errors on the order of 5-10 meters, depending on the hardware, but also on the satellite constellation at the time of recording, and environmental effects such as atmospheric conditions and multi-path effects.

Still, even if the relative position of two models can only be determined up to a few meters, that is sufficient to rule out most ambiguities due to duplicate structures. The first step is to do a coarse absolute orientation, by transforming the photogrammetric models onto the GNSS coordinates (we use Cartesian UTM coordinates). For tourist photographs, which dominate on Flickr, we found it advantageous to operate in 2D at this point. Most of the images are taken from the same height above ground (within the measurement accuracy of the GNSS tags, which as usual is

[2] If exclusively duplicate structure is visible, a "wrong" relative orientation will have no negative effect and will go undetected.

[3] To see this, raise $q_{\min}$ until almost all image pairs are lost. The remaining "networks" will be stereo pairs in which almost *all* available interest points match, hence they cannot contain a mixture of duplicate structure and correct matches.

[4] We note that stricter clustering also has side-effects, which can be wanted or unwanted. E.g., it tends to separate daytime and night images from the same viewpoint.

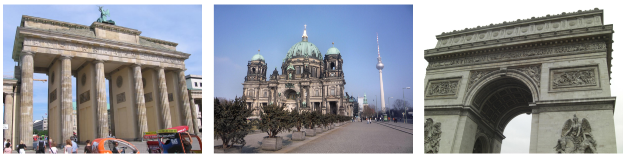
by a factor of 2 or so worse in vertical direction). It is therefore sufficient to project the camera centers onto a best-fitting horizontal plane, found with 2-point RANSAC followed by least-squares fitting to the inliers. For each model k we get a similarity transformation T_k from local object coordinates to UTM coordinates, so that all models are (roughly) co-registered in a planimetric UTM frame.

Recall that the absolute orientation is *not* accurate enough to stop at this point: the errors reach 5 m or more, compared to a typical building size of 25-100 m. Hence, we need to go back to the models and find corresponding 3D structure points for a better alignment. At first glance it might not seem difficult to establish correspondences, given that we have access to the images and the interest points from which the 3D point was triangulated. Unfortunately, by the nature of the problem, different models tend to correspond to distinct viewpoints with large (angular) baselines between them, and thus push descriptor-based matching to the limits. The vocabulary-based approach used above is not sufficient, and we need to go back to the original images. Luckily this is computationally much less costly now, because fewer images and fewer points are involved. We use standard SIFT comparison as implemented in VLFeat Vedaldi and Fulkerson (2008), restricted to pairs of images from the two different models, and to interest points that successfully could be triangulated to 3D structure points in their respective model. Given the *a-priori* constraints imposed by the coarse registration, one can now avoid most false matches, including those from duplicate structure. First, matches are only allowed between points whose 3D reconstructions lie within a plausible distance from eachother in the common UTM object space. This distance should reflect the accuracy of the coarse registration. Second, the 2D alignment is again estimated with RANSAC, but constrained to rotations below 15°, so as to stay close enough to the rough initialization. The magnitude of the translation need not be restricted, because this is already implicit in the maximal distance between corresponding object points. The transformation is declared valid if it supported by enough inliers. Too high thresholds must be avoided at this stage because of the difficult wide-baseline situation. We chose a support set of 15 inliers, which still provides a healthy redundancy. In a nutshell, our strategy can be described as follows: (1) break models apart at unreliable connection with too few tie-point matches; (2) find more reliable matches with the help of the rough GNSS alignment, and stitch the model parts back together.

3.4 Re-adjustment and geo-referencing

Now that both point-to-point correspondences across different partial reconstructions and redundant, albeit inaccurate, GNSS positions are available, it is natural to fuse all available information in a final bundle adjustment. All camera parameters (interior as well as exterior) and 3D tie-point coordinates are estimated in one big least-squares problem, so as to jointly minimize the reprojection errors in image space and the deviations from the observed GNSS positions. The camera intrinsics are three parameters, one for focal length and two for radial distortion. The camera extrinsics are six parameters, three for rotation which is parametrized as Rodriques axis-angle vector and three for translation. The GNSS positions have two horizontal parameters. In the absence of better estimates, the weights $1/\sigma^2_{GNSS}$ of the camera positions are set empirically to $\sigma_{GNSS} = \pm 10$ m. Together with $\sigma_{xy} = \pm 2$ pixels for the image coordinates, this means that the relative alignment between the models is dominated by the more accurate point correspondences, whereas the absolute datum is determined jointly by a redundant set of GNSS coordinates.

We implement the adjustment with the *CERES* solver Agarwal and Mierle (2012). The software, available as open-source *C++* code, is numerically stable and efficient for large networks with thousands of cameras, and it is convenient to use, requiring the user only to specify the cost functions for all groups of observations.



(a) BRATOR (b) BERDOM (c) TRIOMPHE

Figure 3. Sample images for each of the three datasets. *(a)* BRATOR – Bradenburger Tor, *(b)* BERDOM – Berliner Dom, and *(c)* TRIOMPHE – Arc de Triomphe.

	no. images	no. EXIF	no. EXIF GNSS
BRATOR	25,298	16,337	5,174
BERDOM	15,890	7,213	394
TRIOMPHE	30,668	20,624	6,412

Table 1. Number of images in each dataset.

4. EXPERIMENTS AND RESULTS

To evaluate the proposed ideas we present experiments in three different datasets of crowd-sourced pictures. As a baseline we use the standard approach without auxiliary GNSS tags. In order to maximise the coverage of a given object, the standard method tries to split only geographically distinct sites or objects into separate clusters (for efficiency), but to keep all photos of the same building in a single cluster, so as to avoid unwanted fragmentation. For our baseline results we implement that approach by running *VocMatch* with the default parameters. For our example buildings, which do contain duplicate elements, the corresponding reconstructions suffer from reduced coverage and ghost structures. We then go on to show how the proposed modifications improve correctness and completeness of the models. Finally, we also show the potential of geo-referencing with consumer-grade GNSS locations.

4.1 Datasets

The selected datasets depict three well-known landmarks that are known to cause problems because of symmetric structures, e.g. Heinly et al. (2014). The datasets were downloaded from the photo sharing platform Flickr Yahoo! (2005), by querying for all images from the time interval between January 2007 and January 2015 (Fig. 3). We tried both natural query modes: BRATOR and TRIOMPHE were specified via a planimetric bounding box, whereas for BERDOM we submitted a text query. The TRIOMPHE dataset contains a total of 30,668 images, BERDOM contains 15,890 images, and BRATOR contains 25,298 images. Table 1 also shows the number of images in each dataset that have an EXIF header, and the number of images whose header includes a GNSS location. The latter varies from 21% to as little as 2.5%.

4.2 Reconstruction without GNSS support

For the purpose of this paper we selected landmarks that are known to be problematic, so it was expected that the standard reconstruction pipeline would lead to (partially) incorrect models. Of course many buildings can nevertheless be reconstructed correctly from crowd-sourced images without further precautions.

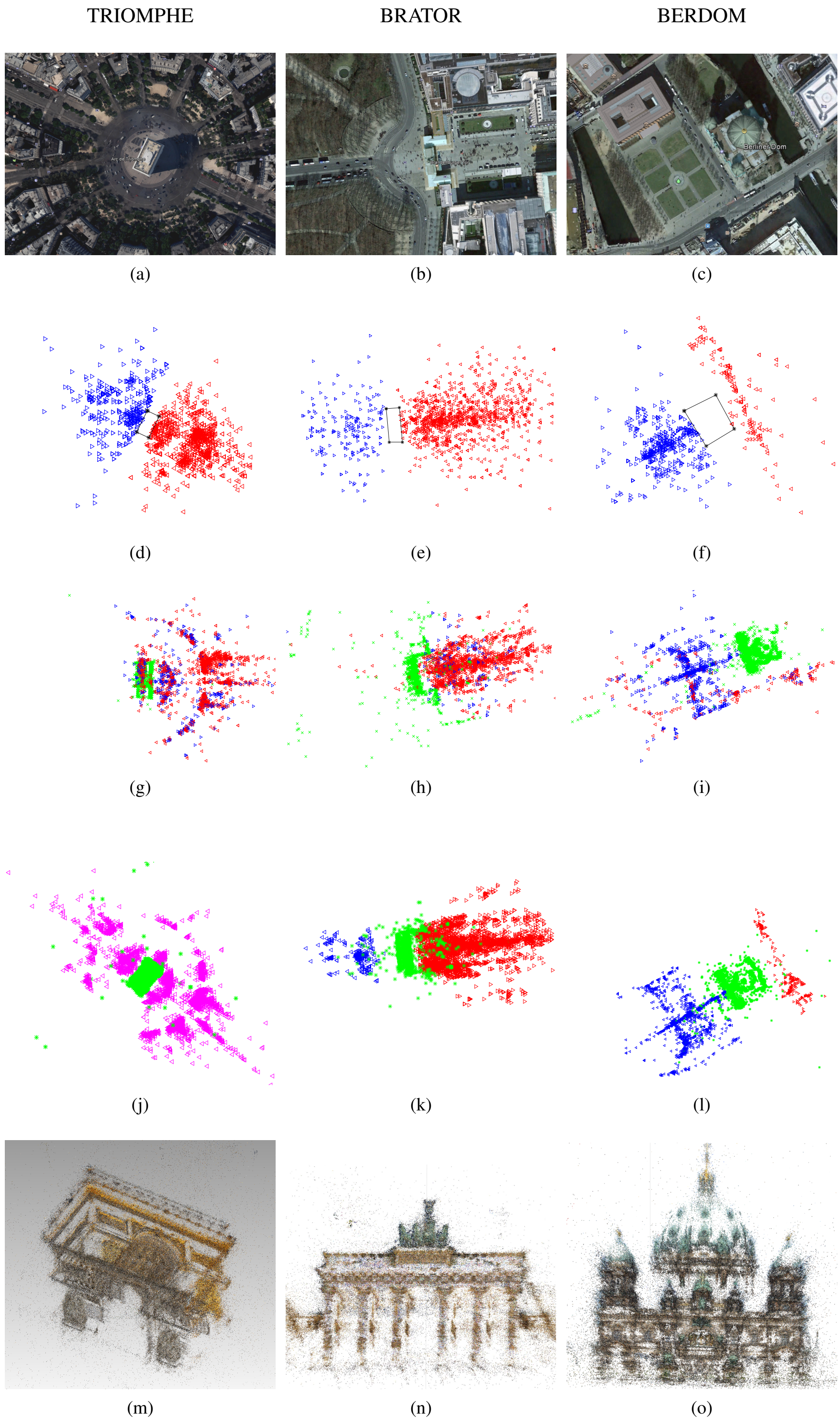
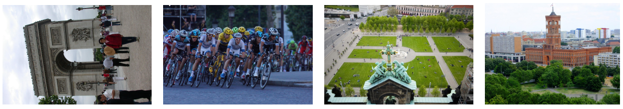


Figure 4. The different steps of our method. (a)-(c) top view of the landmark (courtesy of Google Earth). (d)-(f) cameras (red/blue triangles) with GNSS information. (g)-(i) reconstruction with Bundler when using no GNSS and standard VocMatch parameters. The cameras have the same color as in (d)-(f). (j)-(l) results using the proposed method (different colors for each cluster). (m)-(o) example views of the 3D point clouds.



(a) TRIOMPHE (b) TRIOMPHE (c) BERDOM (d) BERDOM

Figure 5. Representive images which are not part of the largest cluster. *(a)* rotated landmark, *(b)* Tour de France *(c)* and *(d)* rooftop view.

The problem is that it is not known in advance which ones will cause problems.

The experimental setup for each dataset is the same. The complete image set is clustered using the original *VocMatch* approach with default parameters, which are designed to filter out unrelated images, but not to break the pictures covering the landmark into smaller clusters. Too small clusters (< 50 images) are discarded. For BERDOM there is only a single large enough cluster, whereas BRATOR has one dominant and one smaller cluster. TRIOMPHE has one dominant and five smaller clusters. Three of the four small clusters in Fig. 5 show views in which the building of interest is hardly visible, e.g. the Tour de France passing the Arc de Triomphe or scenic views from the roof of the Berliner Dom. The fourth small cluster seems to isolate images of the Arc the Triomphe which have not been rotated from landscape to portrait view (Fig. 5a), because the variant of SIFT without rotation normalisation was used. By and large the results confirm that *VocMatch* works as planned: clustering only discards images that are irrelevant for the subsequent reconstruction.

Following the standard pipeline, the dominant cluster of each dataset was fed into structure-from-motion computation. For our experiments we use Bundler. As usual, geometric verification of the putative matches discards further images that cannot be reliably attached to the model. Of the 16'767 images in the main TRIOMPHE cluster, 10'898 are used for the final 3D model. For BERDOM the number is 3'353 out of 4'376, for BRATOR 8'976 out of 11'275.

All three reconstructions exhibit problems due to duplicate structure. For TRIOMPHE and BRATOR only one side of the building is reconstructed, whereas the other one is missing. The cameras in blue in Fig. 4g and 4h should be on the other side of the building, i.e. these cameras and structure points were reconstructed incorrectly. As a consequence for the BRATOR dataset also the nearby "Hotel Adlon" building is placed on the wrong side of the monument, where in fact there is a park. For BERDOM (Fig. 4i) one of the side walls is mistakenly matched to the other one and thus missing in the reconstruction. Because of the wrong correspondence a neighbouring building is also grossly misplaced.

4.3 Reconstruction with GNSS support

The reconstruction was repeated as proposed in this paper, i.e. first a stricter clustering that severs weak connections, then an approximate alignment based on GNSS tags, followed by a refinement based on guided 3D correspondence search. In detail, we increase the threshold $q_{\min}$ for the (relative) number of matched points, assuming that duplicate structure will not occur across the entire image set and thus get a lower number of matches. Compared to the original VocMatch method we also use a stricter word rarity constraint to remove potentially spurious connections: the word must appear in ≤ 40 database images. The statistics for the clustering for each dataset are shown in (Tab. 2). As expected the stricter clustering creates some fragmentation, but all duplicate structure problems are resolved. BRATOR only separates into two clusters (front and back view), which are successfully reconstructed and reconnected. For BERDOM more clusters are found, but only the two largest ones are usable – the remaining clusters contain night images, pictures of the interior, pictures from a nearby festival, etc., which cannot be connected to the largest cluster. For TRIOMPHE despite applying a stricter threshold (up to $q_{\min} = 2.5\%$) the large cluster did not split into smaller clusters which contain the different sides of the building. The other large cluster listed in the table contains images of a different landmark, the Eiffel Tower. We note that the partial models are smaller than the monolithic image networks found before (Sec. 4.2), so the reconstruction is also faster.

Dataset	$q_{\min}$	no. images in each cluster ($>$ 100 images)
BRATOR	1.5%	5537 277
BERDOM	1.8%	1427 293 287 254 231 224 155 133
TRIOMPHE	2.5%	5040 678 196 112

Table 2. The dataset dependent threshold $q_{\min}$ is listed for each dataset. Based on this threshold a different number of clusters having more than 100 images are created. For each dataset the size of the different clusters is specified.

Next, these partial reconstructions are roughly registered based on the available GNSS tags, as explained in Sec. 3.3. The coarse registration is followed by a constrained 3D point matching (see Sec. 3.3). With the newly found point correspondences and all available GNSS camera locations, the complete model is then computed in a final bundle adjustment.

The BRATOR and TRIOMPHE models are more or less complete, i.e. both sides are reconstructed.[5] BERDOM is now also covered from the front and back and is visibly more complete (and correct, since the two sides that were wrongly matched are not completely identical). Nevertheless, a part of the northern wall is still missing. Here, we face the fundamental limitation of online photo collections. The dataset contains too few pictures from that side.

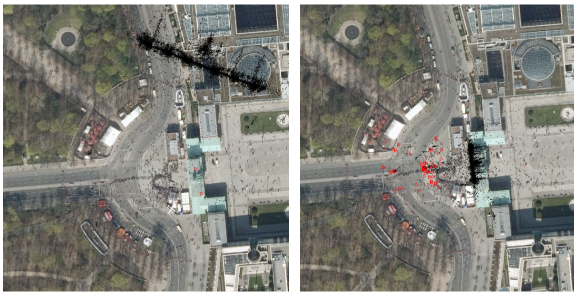
4.4 Georeferencing

For all three models, we also included the absolute GNSS coordinates as (uncertain) observations in the final bundle adjustment, in order to obtain geo-referenced models. Unfortunately, we do not have ground truth coordinates for any of the cameras or structure points, so we can only qualitatively check the accuracy of the absolute geo-reference. To that end we overlay the model with freely available orthophotos (for Berlin from the Berlin Senate Department for Urban Development and the Environment and for Paris from Google Earth). Here the adjusted UTM coordinates are used and the alignment is verified visually (Fig. 7). For comparison, we also show the alignment based on minimal sets of 2 randomly selected GNSS tags (Fig. 6). The comparison confirms that individual geo-tags are too noisy to be trusted at the scale of buildings. Nevertheless, the large set of GNSS locations together allows for reasonable geo-referencing. We estimate the localisation accuracy (uncertainty plus systematic errors) of the final models to be approximately $\pm 1 - 2$ m. In case of small models and therefore few GNSS tags this can go down to ± 10 m.

5. CONCLUSIONS

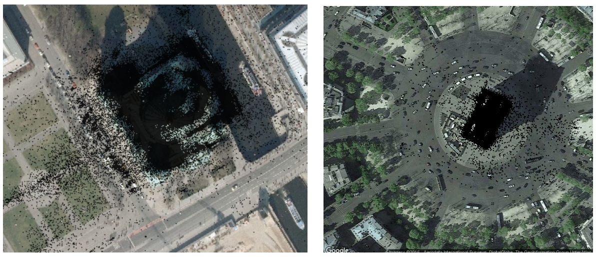
We have investigated the possibility to overcome problems of crowd-sourced photogrammetric reconstruction that are caused

[5] Note, the completeness of the final model also depends greatly on which dense matching method is used and how it is tuned. This is beyond the scope of the present work.



(a) BRATOR (b) BRATOR

Figure 6. Result for BRATOR using the minimum number of 2 GNSS tags for georeferencing the model *(a)*. Using all available GNSS tags *(b)* the model is georeferenced correctly.



(a) BERDOM (b) TRIOMPHE

Figure 7. Result for BERDOM and TRIOMPHE dataset after geo-referencing (TRIOMPHE image courtesy of Google Maps).

by duplicate scene structures. By using only the automatically added GNSS tags that are available in the meta-data of some Internet photographs, we were able to reconstruct, in an automatic fashion, architectural 3D models which are more correct, more complete, and geo-referenced with an accuracy that we consider sufficient for many applications.

Although the results are very encouraging, the proposed method is to a certain extent empirical and heuristic, and there are certainly instances where it will not work. Possible failure cases include: duplicate structures which are so dominant that they remain in the same cluster until the image set is too fragmented – here it would potentially help to also use the GNSS coordinates during clustering; small clusters without GNSS tags – this is difficult to solve, but hopefully will become increasingly rare; and a failure of the 3D correspondence search needed to re-combine the partial models – it may be possible to revert to purely geometric matching methods as they are used for tasks such as LiDAR scan registration.

It will be interesting to see whether the "global lightfield" captured by everyday pictures, cameras on vehicles, drones, etc. will one day be so dense that completely crowd-sourced reconstruction becomes possible beyond selected touristic landmarks.

References

Acute3D, accessed 04/2016. `http://www.acute3d.com`.

Agarwal, S. and Mierle, K., 2012. Ceres Solver: Tutorial & Reference. Google Inc.

Agarwal, S., Snavely, N., Simon, I., Seitz, S. and Szeliski, R., 2009. Building Rome in a day. In: ICCV.

Bundler, accessed 04/2016. `https://github.com/snavely/bundler_sfm`.

Cohen, A., Sattler, T. and Pollefeys, M., 2015. Merging the unmatchable: Stitching visually disconnected SfM models. In: ICCV.

Crandall, D., Owens, A., Snavely, N. and Huttenlocher, D., 2011. Discrete-continuous optimization for large-scale structure from motion. In: CVPR.

Dick, A. R., Torr, P. H. S. and Cipolla, R., 2002. A bayesian estimation of building shape using MCMC. In: ECCV.

Frahm, J.-M. et al., 2010. Building Rome on a cloudless day. In: ECCV.

Goesele, M., Snavely, N., Curless, B., Hoppe, H. and Seitz, S. M., 2007. Multi-view stereo for community photo collections. In: ICCV.

Havlena, M. and Schindler, K., 2014. VocMatch: Efficient multiview correspondence for structure from motion. In: ECCV.

Hays, J. and Efros, A., 2008. IM2GPS: estimating geographic information from a single image. In: CVPR.

Heinly, J., Dunn, E. and Frahm, J.-M., 2014. Correcting for duplicate scene structure in sparse 3D reconstruction. In: ECCV.

Heipke, C., 2010. Crowdsourcing geospatial data. ISPRS Journal 65, pp. 550–557.

Heller, J., Havlena, M., Jancosek, M., Torii, A. and Pajdla, T., 2015. 3D reconstruction from photographs by CMP SfM web service. In: IAPR MVA.

Kaminsky, R., Snavely, N., Seitz, S. and Szeliski, R., 2009. Alignment of 3D point clouds to overhead images. In: CVPR Workshops.

Leberl, F., 2010. Time for neo-photogrammetry? GIS Development 14(2), pp. 22–24.

Li, X., Wu, C., Zach, C., Lazebnik, S. and Frahm, J.-M., 2008. Modeling and recognition of landmark image collections using iconic scene graphs. In: ECCV.

Mathias, M., Martinovic, A., Weissenberg, J. and Van Gool, L., 2011. Procedural 3D building reconstruction using shape grammars and detectors. In: 3DIMPVT.

Pix4D, accessed 04/2016. `https://pix4d.com`.

Snavely, N., Seitz, S. and Szeliski, R., 2006. Photo tourism: Exploring image collections in 3D. In: SIGGRAPH.

Snavely, N., Seitz, S. and Szeliski, R., 2008. Modeling the world from Internet photo collections. IJCV 80(2), pp. 189–210.

Strecha, C., Pylvänäinen, T. and Fua, P., 2010. Dynamic and scalable large scale image reconstruction. In: CVPR.

Tingdahl, D. and Van Gool, L., 2011. A public system for image based 3D model generation. In: MIRAGE.

Tuite, K., Snavely, N., Hsaio, D.-Y., Tabing, N. and Popovic, Z., 2011. PhotoCity. In: CHI.

Üntzelmann, O., Sattler, T., Middelberg, S. and Kobbelt, L., 2013. A scalable collaborative online system for city reconstruction. In: ICCV Workshops.

Vedaldi, A. and Fulkerson, B., 2008. VLFeat: An open and portable library of computer vision algorithms. `http://www.vlfeat.org`.

VisualSFM, accessed 04/2016. `http://ccwu.me/vsfm/`.

Wang, C.-P., Wilson, K. and Snavely, N., 2013. Accurate georegistration of point clouds using geographic data. In: 3DV.

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

Wu, C., 2013. Towards linear-time incremental structure from motion. In: 3DV.

Yahoo!, 2005. Flickr: Online photo management and photo sharing application – `http://www.flickr.com`.