Using Photographs to Build and Augment 3D Models

Bernhard ZEISL, Olivier SAURER, Torsten SATTLER, Marc POLLEFEYS

Computer Vision and Geometry Group
ETH Zurich, Switzerland
{zeislb, saurero, sattlert, pomarc}@inf.ethz.ch

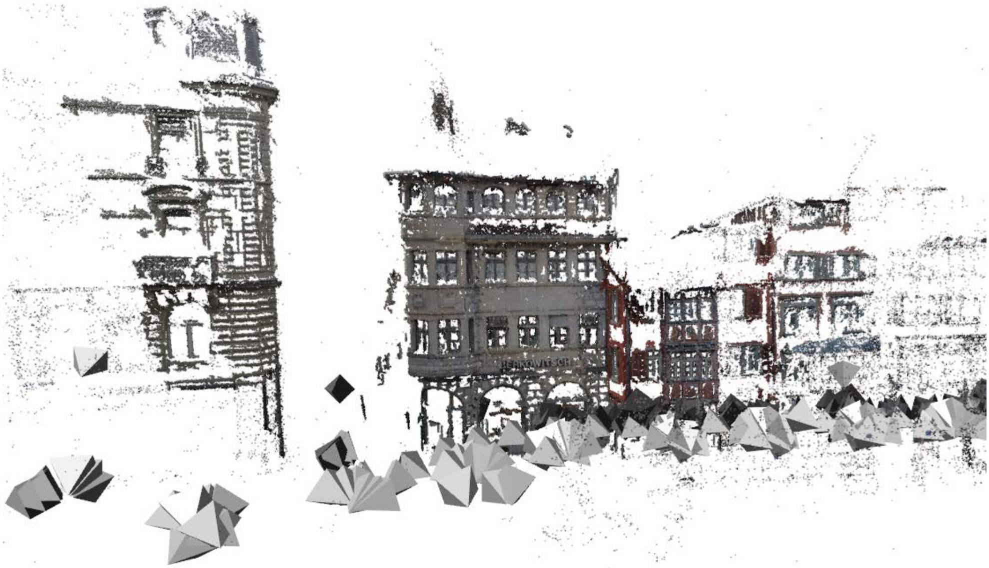


Fig. 1: Part of an urban reconstruction computed via Structure-from-Motion and dense reconstruction obtained solely from images. The gray pyramids visualize camera location an orientation for captured photographs.

Abstract

This paper presents an overview over existing techniques in the field of computer vision for building digital 3D models and for augmenting them with additional photographs. In terms of 3D modelling we illustrate a fully automatic approach for the alignment of scans without the need for any artificial markers or manual interaction. In addition we show how to create entire models solely from images (cf. Fig. 1) up to the scale of whole cities. For the task of image location wrt. an existing model, we differentiate between urban, man-made environments and landscapes. We describe approaches for both cases and demonstrate how novel photographs can augment the 3D model in order to create a richer representation of an environment.

We keep explanations at a higher level such that researchers from different fields are provided with a good overview; however, we reference numerous related works for the interested reader.

1 Introduction

Creating 3D models of urban areas and landscapes is an important step in documenting environments as the resulting 3D models can be used to analyze the scene and plan changes accordingly. In addition, such digital 3D models can be used to detect changes in a scene by comparing photos of the current state with an existing 3D model (TANEJA et al. 2013).

Laser scanners are the state-of-the-art technique to obtain highly detailed 3D models of an environment. However, scanning a scene from the ground level is usually very time-consuming, especially if the environment is not easily accessible by vehicles. In addition, registering individual scans to each other to obtain a single model usually requires manual work; either by carefully placing scan-targets in the scene or by providing point correspondences between scans during post-processing. Consequently it is expensive to capture models of larger scale. In order to create large-scale 3D models, it is therefore common to build upon airborne LIDAR systems (where the GPS sensor of the plane is utilized to register the scans) or to reconstruct the scene by means of images captured from airplanes. While these approaches easily collect data on a very large scale, they usually cannot capture the fine details that are visible only from the ground, e.g. details of a statue on a market place. In contrast, even fine structures can easily be captured in photos taken at ground level with standard consumer cameras. The missing details can thus be added by augmenting existing 3D models with registered images. If photos alone are not sufficient, it is possible to reconstruct a 3D model of the scene from the images taken at ground level - which nowadays can even be done on mobile phones (TANSKANEN et al. 2013; KOLEV et al. 2014) - and adding the model to the reconstruction obtained from the scanner data.

In this paper we survey different techniques that can be used to build and augment 3D models using image data. In Sec. 2, we present an approach that is able to fully automatically align scans even when they were taken from very different viewpoints with only limited overlap, as is the case for terrestrial scanners or when registering models obtained from the air and ground level. Since taking multiple terrestrial scans is cumbersome and time-consuming, we illustrate in Sec. 3 how to obtain a 3D model of a scene solely from images through Structure-from-Motion techniques. Sec. 4 then concentrates on augmenting existing 3D models with images. In Sec. 4.1, we consider the problem of registering photos taken at ground level against a model also obtained from ground level data, e.g. against models reconstructed via Structure-from-Motion. This problem is usually encountered when trying to localize images in urban environments and we emphasize this use case. Sec. 4.2 then details how to localize images with respect to aerial data through the example of localizing photos taken in mountainous terrain relative to a digital elevation model obtained from LIDAR data.

2 Utilizing Images for Automatic 3D Scan Alignment

When surveying construction sites, historical buildings or industrial facilities laser scanning is the state-of-the-art technique to obtain accurate three-dimensional models. To obtain a full 3D model, several 2.5D scans have to be aligned (cf. Fig. 3). Usually a scanner is positioned at different places in order to minimize scan shadows and to obtain a model as complete as possible. Since scanning is a time-consuming and therefore expensive task the

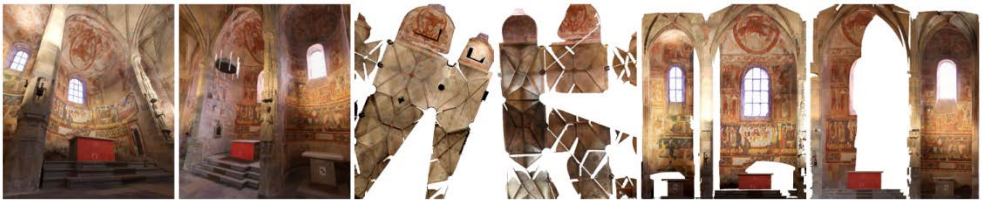


Fig. 2: (Left:) Images taken from 2 different scanner positions, which naturally exhibit a wide baseline. Feature matching and thus registration from these images fails in most cases. (Middle, right:) Generated salient direction rectified renderings which are subsequently used for our automatic registration.

number of scans is usually kept as small as possible, leading to a wide baseline setting between the scan positions. Not only scanning, but also the registration of individual scans takes a lot of time - either afterwards by manually aligning models, or on site by carefully positioning targets (artificial markers) in the scene, which are spotted and automatically detected from several scan positions. If one desires to rescan the facility at another point in time and align current data with an older model, exploiting artificial markers for registration is impossible. As a result there is a need for automatic registration methods which do not rely on any artificial landmarks or prediction on the relative motion, but can generate accurate registration results by exploiting the scan data itself.

Local alignment methods such as ICP (BESL et al. 1992) require a good initialization and are not applicable to wide baseline scenarios or when the relative rotation is unknown. GPS and magnetic compass can simplify the registration problem, but they fail under bridges, inside buildings, urban canyons, or close to metallic or electric installations. Modern laser scanners come with inbuilt or attachable cameras and deliver distance plus color information and we aim at exploiting this data jointly for fully automatic registration. We build upon image features rather than 3D geometry features, because they are plenty, well localized and much more discriminative. However, they suffer from viewpoint distortions and request for normalization, i.e. straight forward feature extraction and matching fails in most cases.

In our novel approach (ZEISL et al. 2013) we propose to become independent of the original sensor viewpoint (position and orientation of the camera) by exploiting characteristic geometric properties of the scene, namely salient directions, which are repeatable among different scans. Examples include peaks in the distribution of the surface normals, vanishing points, symmetry, gravity or other directions that can be reliably obtained from the sensor or the scene. Each salient direction is then exploited to render an orthographic view (cf. Fig. 2), and by this way removing the perspective effects that had been introduced by the particular scanner position. Importantly, for corresponding salient directions between scans the generated images are identical (for jointly seen Lambertian scene parts) up to a 2D similarity transformation. Thus, standard feature detection and description approaches can be employed and features are computed in a viewpoint normalized image representation. Compared to earlier approaches proposed for consumer depth cameras (ZEISL et al. 2012) or stereo systems (WU et al. 2008; CAO et al. 2011) our approach does not pose any requirements on the presence of particular geometric shapes. Moreover, we do not rely on features only on particular fitted models (planes, cylinders, cones), but match

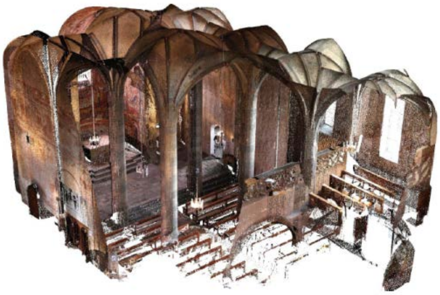
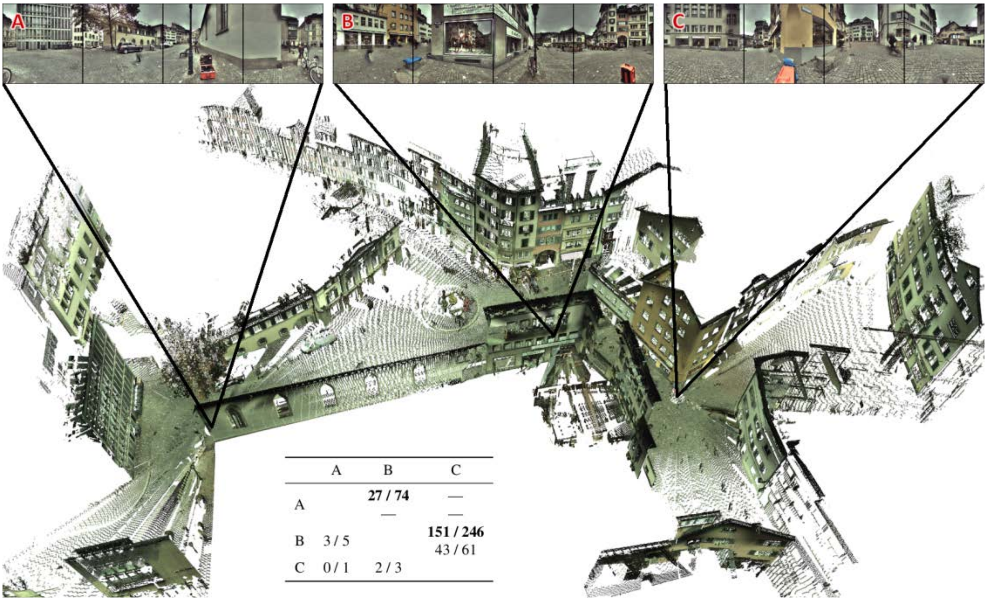


Fig. 3: Cut through a 3D model obtained by our alignment algorithm from 5 individual scans. We achieve entirely automatic registration from largely different viewpoints by exploiting depth and image data jointly.

the whole visible scene, this way significantly increasing the surface area where features can be extracted. This is an important aspect if the visible overlap between scans is small. Contrary to previous work where depth discontinuities can not be handled, our rectification approach generates images that consistently capture objects and features across different levels of depth. Such features at geometry boundaries and folds are among the most discriminative, as known e.g. from stereo.

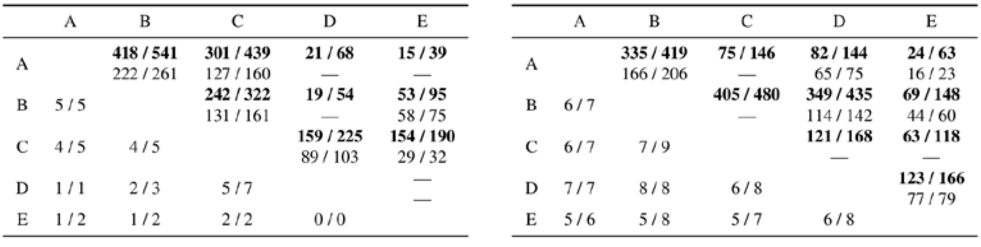
We evaluated our approach on three different datasets with different scene characteristics which are typical for laser scanning scenarios (such as historic sites or urban areas). For evaluation, we analyzed both the repeatability of salient directions and the registration performance itself. The former is essential for successful registration, since we need to detect at least one common salient direction from both viewpoints - which becomes more difficult with less overlap between regions. The latter is evaluated in terms of correct features matches and compared to planar rectification as in (WU et al. 2008; CAO et al. 2011).



	A	B	C
A		**27 / 74** —	— —
B	3 / 5		**151 / 246** 43 / 61
C	0 / 1	2 / 3	

Fig. 4: Registration result for scans within the city of Zurich. Images belonging to the different scan positions are visualized on top. Evaluation is equivalent to Tab. 1.

Table 1: Registration evaluation for 2 datasets (Castel and Church). (Upper right parts:) Relation between correct and tentative matches for our approach and planar rectification (WU et al. 2008). (Lower left parts:) Repeatability scores for salient directions (i.e. found vs. present salient directions in the overlap regions).



	A	B	C	D	E
A		**418 / 541** 222 / 261	**301 / 439** 127 / 160	**21 / 68** —	**15 / 39** —
B	5 / 5		**242 / 322** 131 / 161	**19 / 54** —	**53 / 95** 58 / 75
C	4 / 5	4 / 5		**159 / 225** 89 / 103	**154 / 190** 29 / 32
D	1 / 1	2 / 3	5 / 7		— —
E	1 / 2	1 / 2	2 / 2	0 / 0	

(a) CASTLE

	A	B	C	D	E
A		**335 / 419** 166 / 206	**75 / 146** —	**82 / 144** 65 / 75	**24 / 63** 16 / 23
B	6 / 7		**405 / 480** —	**349 / 435** 114 / 142	**69 / 148** 44 / 60
C	6 / 7	7 / 9		**121 / 168** —	**63 / 118** —
D	7 / 7	8 / 8	6 / 8		**123 / 166** 77 / 79
E	5 / 6	5 / 8	5 / 7	6 / 8	

(b) CHURCH

Our method generates more tentative and correct matches than existing methods, which enables us to register scan-pairs in cases in which these other approaches fail. As expected, this is particularly the case for scenes with numerous non-planar surfaces, where our approach is crucial for successful registration, as planar rectification requires textured planes -- which are often very small or non-existent. Finally, Fig. 3 and Fig. 4 illustrate the global registration results we obtain. Previously estimated relative poses connect pairs of scans with successful registration and by this form a graph over several scans. A solution for the absolute pose of each scans is obtained via extraction of a minimum spanning tree. A following refinement step doesn't improve the results noticeably, highlighting that estimated relative poses are already very precise.

3 Creating 3D Models from Images

Automatically registering multiple scans to form a consistent 3D model is an important step in obtaining large-scale 3D models. However, purely laser-based acquisition is often a cumbersome task due to long scan times and the tedious process of moving the scanner through the scene. In contrast, taking images from multiple scene positions is significantly easier, since consumer cameras are cheap and easy to handle and most modern smart-phones are already equipped with decent cameras. In addition, there is a vast source of imagery available on photo-sharing websites such as Flickr or Panoramio. Thus, we would like to use only sets of images for building a 3D model. The task is typically split into two consecutive steps: First, sparse reconstruction - commonly referred to as Structure-from-Motion (SfM) and second, dense reconstruction. We will briefly discuss these approaches in the following.

If one observes an object of interest with a camera from different viewpoints, a particular 3D point on the object will be projected to different (pixel-)locations in the images due to the camera movement. SfM tries to reverses this process: Given corresponding points over different images, the aim is to estimate their common 3D position and the original camera poses (position and orientation in space). Thus, the problem decomposes into (i) correspondence estimation between salient points in images, and (ii) 3D structure recovery from those estimated, tentative correspondences. For the first part, local interest point

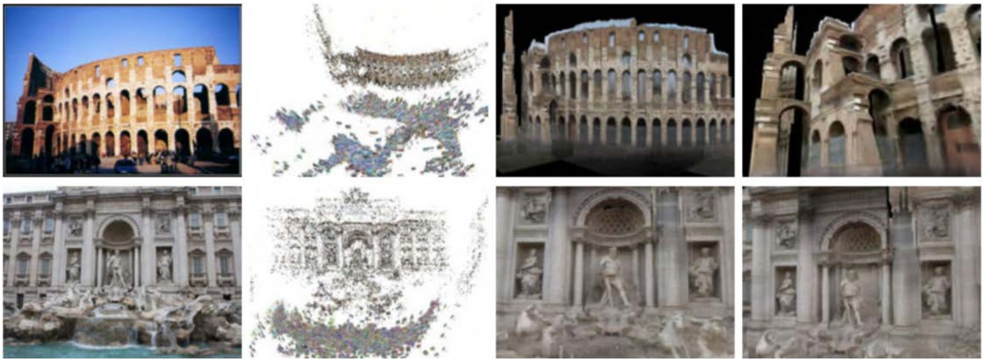


Fig. 5: (Left:) Sample input images obtained from Flicker; (2nd left:) Local sparse SfM reconstruction. (Right:) Dense, textured 3D model seen from 2 different views (FRAHM et al. 2010).

detectors -- such as Harris corners (HARRIS & STEPHENS 1988), SIFT (LOWE et al. 2004) or SURF (BAY et al. 2008) feature points -- are utilized to identify a set of discriminative features, where each is characterized by its local image neighborhood aggregated in a fixed size descriptor. Correspondence estimation then compares individual descriptors between different images and records the best matches. The second stage builds upon the sets of tentative matches between images, which typically also contain false matches. Between two images their epipolar geometry (HARTLEY & ZISSERMAN 2003) is estimated within a RANSAC framework to account for the present noise in the data. It describes the relative pose between two cameras and an initial estimate of the sparse 3D structure can be obtained via triangulating the matching points. For a collection of images, the sparse 3D structure and the camera poses are estimated jointly. This is achieved by first generating initial structure and pose hypothesis from image pairs[1]. It is followed by a global refinement step know as bundle-adjustment (TRIGGS et al. 2000) which minimizes the re-projection error between the projections of estimated 3D points and detected features points among all images.

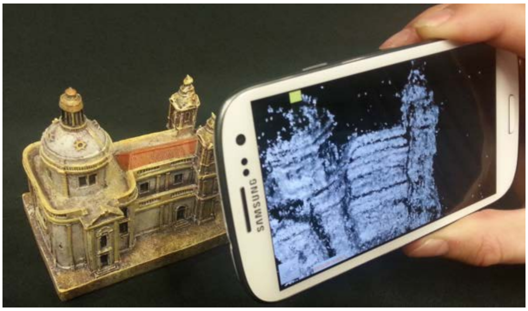


Fig. 6: 3D Reconstruction running in real-time on a consumer smart-phone.

Once the sparse 3D structure and camera poses are estimated, dense reconstruction aims at generating a dense, watertight model. For example, this can be achieved by first computing depth maps with a stereo correspondences algorithm (SCHARSTEIN & SZELISKI 2002) from neighboring images and then fusing them in a common volumetric representation encoding an

[1] More stable solutions can be obtained from images triplets via evaluation of the tri-focal tensor (HARTELY & ZISSERMAN 2003)

occupancy function for the 3D model. Note, that there exist various alternative approaches with different 3D shape representations such as voxels, level-sets, or polygon meshes and also numerous measures for evaluating the visual compatibility of a reconstruction with a set of images. The interested reader is referred to (SEITZ et al. 2006).

Building 3D models from real-world objects or scenes by means of these methods has been a very active and long-standing research topic in computer vision. Over the past decade tremendous progress has been made and led to impressive results. In (POLLEFEYS et al. 2004) we build visual, textured models from a sequence of uncalibrated images acquired with a hand-held camera. The system was shown to be able to reconstruct architectural sites with a relative accuracy of 1/500. In (POLLEFEYS et al. 2008) we aim for 3D reconstruction from video of urban scenes. Due to the large amount of captured data the system employs commodity graphics hardware to achieve real-time processing. Our depth map-based fusion approach is able to achieve a median error of only 3cm for 3D models of normal buildings. Since the volume-based representations used for dense reconstruction are very memory demanding and thus not applicable to larger scenes, we introduce a height-map based representation which is ideally suited to model building facades. Especially for touristic sites and other places of general interest, vast amounts of images can be found on the Internet, e.g., an image search on Google for the keyword “Rome” returns around 3 million photos. AGARWAL et al. (2009) and FRAHM et al. (2010) leverage this huge amount of image data in order to build large-scale 3D models (cf. Fig 5). The challenge hereby is to design matching and reconstruction algorithms such that they maximize computational parallelism and scale efficiently with the amount of available data. While the computations in (ARGARWAL et al. 2009) require a cluster infrastructure for computation, FRAHM et al. (2010) achieve the task of reconstructing Rome within 24 hours on a single workstation. With the increasing computational capabilities of current mobile phones, they are not only valuable devices for casually capturing images, but can nowadays also be used for directly building small models on the phone itself in real-time (TANSKANEN et al. 2013; KOVEL et al. 2014). As seen in Fig. 6 this enables an instantaneous feedback about the current model quality that can be used to guide the user to capture additional images where needed.

4 Geo-Localization of Images

Ideally, we would like to use both scan data and photos to obtain a single coherent 3D model, i.e., we would like to be able to register images against a 3D model. This can either serve the task of augmenting a 3D model with images (SNAVELY et al. 2006) or to leverage Structure-from-Motion techniques to grow the model. In this section, we will thus discuss how to localize a novel image with respect to a 3D model by computing the position and orientation, i.e., the camera pose, from which it was taken. We thereby distinguish between localization in urban and mountainous environments due to the different challenges associated with each type of scene. Notice that image localization also enables many other interesting applications. For example, one can register contemporary and historic photographs against the model to document changes over time in an environment (SCHINDLER et al. 2007).

Urban environments are an important use case for many interesting applications for image-based localization such as pedestrian navigation or touristic information. Since such scenes are often dominated by planar surfaces, localization approaches typically use local image

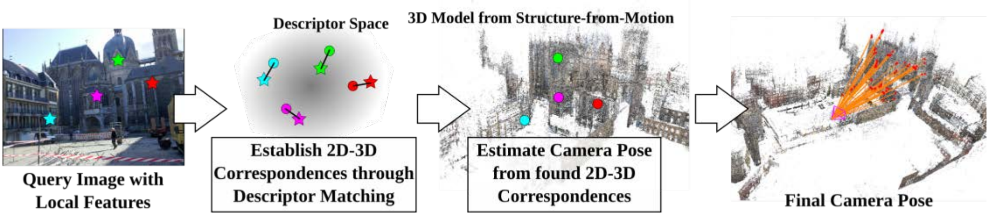

Descriptor Space
3D Model from Structure-from-Motion
Query Image with Local Features
Establish 2D-3D Correspondences through Descriptor Matching
Estimate Camera Pose from found 2D-3D Correspondences
Final Camera Pose

Fig. 7: The standard image-based localization pipeline: 2D-3D correspondences, established via descriptor matching, are used to estimate the camera pose, i.e. position and orientation, from which an image was taken relative to the 3D model.

features. Since the camera pose can be determined from three or more 2D-3D correspondences (HARALICK et al. 1994; LEE et al. 2013), the problem of estimating the pose becomes the problem of computing matches between 2D features in the query image and 3D points, which is solved by comparing local feature descriptors. Sec. 4.1 discusses multiple approaches to establish such 2D-3D matches.

In contrast to urban environments, localization in natural scenes is substantially more challenging due to drastic changes in vegetation between seasons, and the significant differences in scene appearance under different lighting and weather conditions, e.g., snow in winter. Consequently, approaches based on the type of features typically used for urban environments become impractical as local image features are not able to handle these strong appearance changes. In addition, the dense ground-level data commonly used is limited to cities and major roads. For mountains or countrysides only aerial image footage exists, which is much harder to relate with terrestrial imagery due to strong viewpoint changes. In Sec. 4.2, we therefore present an approach that is able to handle the challenging problem of localization in mountainous scenes.

4.1 Localization in Urban Environments

In the following, we assume that our scene is represented as a Structure-from-Motion model. This implies that each 3D point in the model was reconstructed from local features observed in at least two images, allowing us to associate the corresponding image descriptors with the 3D point. Notice that we can obtain a similar representation from laser data: If we are given both images and laser scans as in Sec. 2, then we can directly compute the 3D location for each extracted image feature. If only a (colored) point cloud is available, the method proposed by SIBBING et al. (2013) can be used to render synthetic images in which local features similar to those found in real photos can be extracted. As a result, we would again obtain a point cloud in which each point is associated with one or more descriptors.

Since each 3D point corresponds to at least one image feature, the 2D-3D correspondences required for camera pose estimation can be established by extracting local features in the query image and finding the nearest neighboring 3D point descriptors. The resulting pipeline is illustrated in Fig. 7. A simple strategy to implement the descriptor matching is to store all point descriptors in a tree and use tree-search to accelerate the matching of 2D features against the points. While it has been shown that this strategy is very effective in

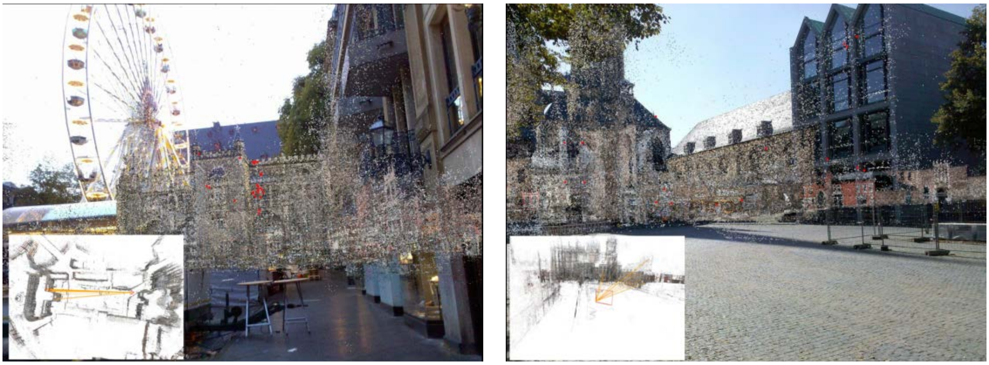


Fig. 8: Two examples for the localization results obtained with our approach (Sattler et al. 2012a). Our method successfully handles structures not in the model, such as the Ferris wheel, strong occlusions, and changes in illumination. The 3D model is projected into the images to demonstrate the quality of the estimated poses. The pose relative to the model is shown in the small inlays, where the lines connect the matching 3D points with the computed center of the camera.

terms of the number of images that can be localized (Sattler et al. 2011; Li et al. 2012), it is also rather slow as the high dimensionality of the descriptors results in search times of multiple seconds per image (Sattler et al. 2011). Thus, Li et al. (2010) propose a more efficient approach based on prioritized 3D-to-2D matching: Given an initial set of 3D points likely to be visible in a random query image, they try to find the corresponding nearest neighboring feature descriptors in the given query image. If a match is found, they increase the priorities of all other points that are visible together with the matching point. The next point they match against the image is then picked according to these priorities. The information about which points in the model are co-visible is thereby obtained from the Structure-from-Motion process. If two points are observed together in one of the database images used for the reconstruction, they are considered to be co-visible, a definition which approximates the true co-visibility relation. Choudhary & Narayanan (2012) propose a probabilistic version of this approach, where the next point is selected based on the probability of being co-visible with all matching points found so far.

While being significantly slower, both approaches are not as effective as the tree-based search approach. Recently, we have proposed a localization framework that is both efficient and effective (Sattler et al. 2011; Sattler et al. 2012a). It is based on a prioritized 2D-to-3D matching step. In a first step, we identify for each image feature a set of 3D points with similar local appearance without actually comparing their descriptors (Sattler et al. 2011). Features for which only a few points with similar appearance can be found are more unique and thus more informative than features with many similar points. Thus, we first try to find correspondences for these features by matching each of them against the set of similar points determined in the first step. While on average being one order of magnitude faster than the tree-based approaches, the drawback of this method is that we lose potential matches as we need to quantize the descriptor space in order to identify sets of similar points efficiently. In order to recover these lost matches, we exploit co-visibility information (Sattler et al. 2012a): Assuming that we have found a correct match, we can assume that 3D points close to the matching point are very likely to be visible in the query image. Thus, we try to match them against the image to obtain additional correspondences, enabling us to recover matches previously lost due to quantization. At the same time, we also show how to use co-visibility information to filter out wrong matches before attempting pose estimation, which can considerably accelerate this part of the pipeline. Our approach is the most efficient localization approach published so far, enabling us to find correspondences and the pose for a query image in around 260ms on average when matching against a model containing millions of 3D points. At the same time, we are at least as effective as standard tree-based approaches. Table 2 shows the localization accuracy of our method compared to other approaches on a standard benchmark dataset. As can be seen, we achieve a higher localization accuracy than these other methods and are also significantly better than GPS. Fig. 8 shows the quality of the estimated poses visually by projecting the point cloud into the query images through the pose computed by our algorithm[2].

Table 2 The localization accuracy obtained on the Dubrovnik dataset (Li et al. 2010).

Method	% Localized Images	25% Quantile [m]	50% Quantile [m]	75%Quantile [m]
Li et al. (2010)	94.1	7.5	9.3	13.4
CHOUDHARY & NARAYANAN (2012)	98.5	0.88	3.1	11.83
Sattler et al. (2012a)	99.5	0.4	1.4	5.3

Recently, LI et al. (2012) proposed a slightly modified version of tree-based search that is slightly more efficient than our approach. However, this effectiveness comes at the price of significantly longer run-times.

The main drawback of the localization approaches discussed above is that they need to keep all the point descriptors in memory, which becomes prohibitively expensive for very large models. An alternative to these approaches are methods that first try to identify and retrieve database images (used to reconstruct the model) taken from a similar viewpoint as the query image (IRSCHARA et al. 2009). The advantage of such methods is that they do not need the full descriptors for retrieval but can work on a fixed-size set of quantized descriptors, significantly reducing the memory requirements. While classical retrieval-based approaches (IRSCHARA et al. 2009) have been shown to be less effective (SATTLER et al. 2011), we were recently able to identify the algorithmic reasons for this behavior (SATTLER et al. 2012b). We proposed a slight modification that makes retrieval-based approaches as effective as methods based on directly matching descriptors without sacrificing their scalability. In addition, we have shown that we can exploit the fact that most surfaces in urban environments are planar to further improve the retrieval performance. We can rectify these planar regions and obtain more powerful descriptors that are more stable over a larger range of viewpoints, enabling us to localize more images taken under challenging conditions (BAATZ et al. 2010; CHEN et al. 2011). At the same

[2] Source code available: http://www.graphics.rwth-aachen.de/software/image-localization

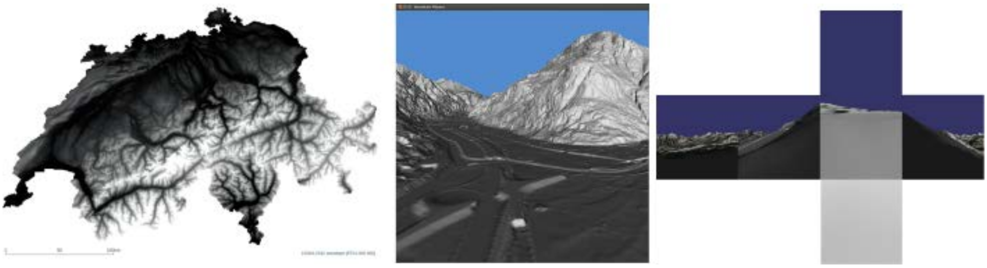


Fig. 9: (Left:) Digital elevation model (DEM) of Switzerland. Every 100m a synthetic view (middle) is rendered and converted to a cube map (right).

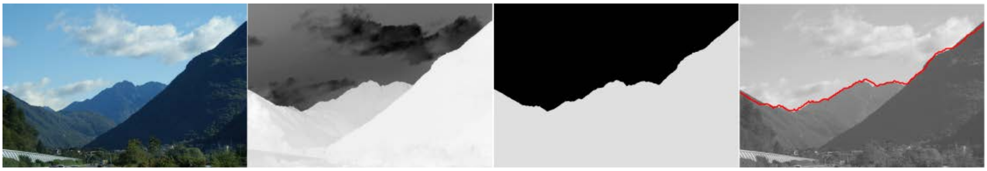


Fig. 10: The input query image (left) with the corresponding “pseudo depth map” (2nd left) obtained from image de-hazing. The “pseudo depth map” is then used together with gradient features to segment the image in ground and sky (2nd right). The best matching skyline stored in the database is overlaid onto the query image (right)

time, we can also simplify the pose estimation problem by exploiting the fact that the 3D points lie on planar surfaces (Baatz et al. 2011).

4.2 Localization in Mountainous Environments

Given a digital elevation model (DEM) of a country, or ultimately the world, we would like to tell where a given image was taken. In previous work Woo et al. (2007) and Stein et al. (1995) matched mountain peaks to a set of nearby mountains. We propose in (Baatz et al. 2011) to aggregate shape information across the whole skyline (not only the peaks) and search for a similar configuration of basic shapes in a large scale database that is organized to allow for query images of largely different fields of view. Our method first segments the skyline (either automatic or guided by an operator, for challenging images containing reflections and occlusions) and uses it to retrieve the most similar geo-localized skyline from a database.

The location recognition problem in its general form is six-dimensional, with three position and three orientation parameters, which have to be estimated. We make the assumption that the images are taken not too far from the ground and use the fact that people rarely twist the camera relative to the horizon (Brown et al. 2007) (i.e. small roll). In (Baatz et al. 2012) we propose a method to solve that problem using the outlines of mountains against the skyline (denoted as visible horizon).

For the visual database we seek a representation that is robust with respect to tilt of the camera which means that we are effectively left with estimating the 2D position on our

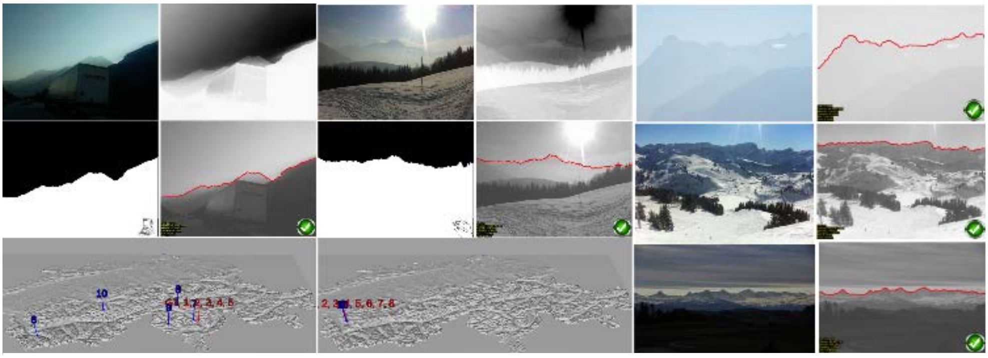


Fig. 11: (Left:) Each block of 4 images shows the input image, depth from de-hazing, the segmentation mask and the best matching contour overlaid (red) onto the original query image. The bottom row shows the 10 best matches retrieved from the database. (Right) more results showing the query image and the overlaid matching skyline.

DEM (latitude and longitude) and the viewing direction of the camera. The visible horizon of the DEM is extracted offline at regular grid positions (360 degree at each position), Fig. 9, and represented by a collection of vector-quantized local contourlets (contour words, similar in spirit to visual words obtained from quantized image patch descriptors (SIVIC et al. 2003)). In contrast to visual word based approaches, a viewing angle relative to the north direction is stored with each contourlet.

At query time, a sky segmentation technique is applied that copes with the often present haze. Subsequently the extracted contour is robustly described by a set of local contourlets plus their relative angular distance with respect to the optical axis of the camera, Fig. 10. Then, we use an inverted file system for the contour words to find the most promising location and simultaneously vote for the viewing direction, which is an integrated geometric verification, already during the bag-of-words search.

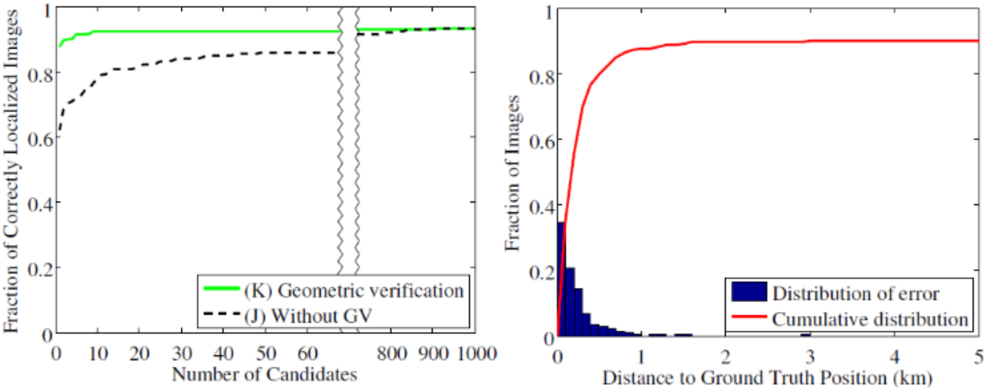


Fig. 12: (Left:) Retrieval performance before and after geometric verification (GV). (Right:) Fraction of queries having a given distance to the ground truth position. About 89% of the images were correctly localized within an error distance of 1km to the ground truth.

We validate the proposed approach using a public digital elevation model of Switzerland (obtained from Swiss Topo[3]) that covers more than 40 000km2 and a set of more than 200 query images from different sources with ground truth position. Also, we demonstrate that the horizon is highly informative and can be used effectively for localization. On the 200+ query images 49% were segmented fully automatically while the others required some human interaction. Of all the query images 89% were correctly localized within a distance of 1km to the ground truth, Fig. 11 and Fig. 12. Querying the database containing 3M panoramic contours takes in average 10 sec per query.

5 Conclusion

In this work we have given an overview over different methods in computer vision which leverage images to build and augment digital 3D models. We believe that modeling from images represents an interesting alternative to state-of-the-art laser scanning techniques in situations where scanning is cumbersome or simply too expensive due to the scale of the scene. Since not only scanning but also the alignment of scans so far needs manual interaction, we have presented a fully automatic approach for scan alignment. Additionally, the discussed image localization methods can be utilized to complement present 3D models of urban areas or landscapes with additional image data in order to obtain a richer model.

In the future we will continue our work in 3D modeling from images and aim at jointly exploiting scan and image data to obtain even more accurate and detailed models. In terms of localization we would for example like to target the problem of spatially organizing contemporary and historic image collections.

References

AGARWAL, S., SNAVELY, N., SIMON, I., SEITZ, S. M., SZELISKI, R. (2009). Building Rome in a Day. International Conference on Computer Vision.

ARTH, C., REITMAYR, G., SCHMALSTIEG, D. (2012), Full 6DOF Pose Estimation from Geo-Located Images. Asian Conference on Computer Vision.

BAATZ, G., KOESER, K., GRZESZCZUK, R., POLLEFEYS, M. (2010), Handling Urban Location Recognition as a 2D Homothetic Problem. European Conference on Computer Vision.

BAATZ, G., KOESER, K., CHEN, D., GRZESZCZUK, R., POLLEFEYS, M. (2011), Leveraging 3D City Models for Rotation Invariant Place-of-Interest Recognition. International Journal of Computer Vision, 96(3), p. 315-334.

BAATZ, G., SAURER, O., KOESER, K., POLLEFEYS, M. (2012), Large Scale Visual Geo-Localization of Images in Mountainous Terrain.

BAY, H., ESS, A., TUYTELAARS, T., VAN GOOL, L. (2008), Speeded-Up Robust Features (SURF), Computer Vision and Image Understansing, 110(3), pp. 346–359, 2008.

BESL, P., MCKAY, N. (1992), A Method for Registration of 3D shapes. IEEE Transactions on Pattern Analysis and Machine Intelligence.

[3] www.swisstopo.admin.ch

BROWN, M., LOWE, D.G. (2007), Automatic panoramic image stitching using invariant features. International Journal on Computer Vision.

CAO, S., SNAVELY, N. (2013), Graph-Based Discriminative Learning for Location Recognition. IEEE Conference on Computer Vision and Pattern Recognition.

CAO, Y., YANG, M., MCDONALD, J. (2011), Robust Alignment of Wide Baseline Terrestrial Laser Scans via 3D Viewpoint Normalization. Workhsop on Applications of Computer Vision

CHEN, D., BAATZ, G., KOESER, K., TSAI, S. , VEDANTHAM, R., PYLAENAEINEN, T., ROMEILA, K., CHEN, X., BACH, J., POLLEFEYS, M., GIROD, B. GRZESZCZUK (2011), City-scale landmark identification on mobile devices. IEEE Conference on Computer Vision and Pattern Recognition.

CHOUDHARY, S., NARAYANAN, P. J. (2012), Visibility Probability Structure from SfM Datasets and Applications. European Conference on Computer Vision.

FISCHLER, M. A. & BOLLES, R. C. (1981), Random Sample Consensus: A paradigm for model fitting with applications to image analysis and automated cartography. Communications of the ACM, 24 (6), 381-395.

FRAHM, J. M., FITE-GEORGEL, P., GALLUP, D., JOHNSON, T., RAGURAM, R., WU, C., JEN, Y.-H., DUNN, E., CLIPP, B., LAZEBNIK, S., POLLEFEYS, M. (2010), Building Rome on a cloudless day. European Conference on Computer Vision.

GONZALEZ, R. C., WOODS, R. E. & EDDINS, S. L. (2004), Digital image processing using MATLAB. Prentice Hall, 624 p.

HARALICK, R., LEE, C. N., OTTENBERG, K., NOELLE, M. (1994), Review and and analysis of solutions of the three point perspective pose estimation problem. International Journal on Computer Vision.

HARTLEY, R., ZISSERMAN, A. (20043, Multiple View Geometry in Computer Vision. 2nd Edition. Cambridge University Press.

HARRIS, C., STEPHENS, M. (1988), A combined corner and edge detector. Alvey Vision Conference, vol. 15, pp. 147–151.

HECHT, R., MEINEL, G., BUCHROITHNER, M. F. (2006), Estimation of urban green volume based on last pulse lidar data at leaf-off aerial flight times. Proc. 1st EARSeL Workshop on Urban Remote Sensing. Humboldt-University, Berlin, Germany, 2-3 March 2006 (CD-ROM).

KOLEV, K., TANSKANEN, P., SPECIALE, P., POLLEFEYS, M. (2014), Turning Mobile Phones into 3D Scanners. Inernational Conference on Computer Vision.

LEE, G. H., LI, Bo., POLLEFEYS, M., FRAUNDORFER, F. (2013), Minimal Solutions for Pose Estimation of a Multi-Camera System. International Symposium on Robotics Research, 2013.

LI, Y., SNAVELY, N., HUTTENLOCHER, D. (2010), Location Recognition using Prioritized Feature Matching. European Conference on Computer Vision.

LI, Y., SNAVELY, N., HUTTENLOCHER, D., FUA, P. (2012), Worldwide pose estimation using 3d point clouds. European Conference on Computer Vision.

LOWE, D. G. (2004), Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision. 60(2), pp. 91–110, 2004.

MCGLONE, C., MIKHAIL, E. & BETHEL, J. (Eds.) (2004), Manual of photogrammetry. American Society for Photogrammetry and Remote Sensing, 5th edition, 1151 p.

POLLEFEYS, M., VAN GOOL, L., VERGAUWEN, M., VERBIEST, F., CORNELIS, K., TOPS, J., KOCH, R. (2004), Visual modeling with a hand-held camera. International Journal of Computer Vision, *59*(3), 207-232.

POLLEFEYS, M., NISTÉR, D., FRAHM, J. M., AKBARZADEH, A., MORDOHAI, P., CLIPP, B., ENGELS, C., GALLUP, D., KIM, S.-J., MERRELL, P., SALMI, C., SINHA, S., TALTON, B., WANG, L., YANG, Q., STEWENIUS, H., YANG, R., WELCH, G., TOWLES, H. (2008), Detailed real-time urban 3d reconstruction from video. International Journal of Computer Vision, *78*(2-3), 143-167.

SATTLER, T., LEIBE, B., KOBBELT, L. (2011), Fast Image-based Localization using Direct 2D-to-3D Matching. International Conference on Computer Vision.

SATTLER, T., LEIBE, B., KOBBELT, L. (2012), Improving Image-Based Localization by Active Correspondence Search. Europena Conference on Computer Vision.

SATTLER, T., WEYAND, T., LEIBE, B., KOBBELT, L. (2012), Image Retrieval for Image-Based Localization Revisited. British Machine Vision Conference.

SCHARSTEIN, D., SZELISKI, R. (2002), A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Cmputer Vision, 47(1-3), 7-42.

SCHINDLER, G., BROWN, M., SZELISKI, R. (2007), City-Scale Location Recognition. IEEE Conference on Computer Vision and Pattern Recognition.

SCHINDLER, G., KRISHNAMURTHY, P., DELLAERT, F. (2007), Inferring Temporal Order of Images From 3D Structure. IEEE Conference on Computer Vision and Pattern Recognition.

SEITZ, S. M., CURLESS, B., DIEBEL, J., SCHARSTEIN, D., & SZELISKI, R. (2006), A comparison and evaluation of multi-view stereo reconstruction algorithms. Conference on Computer Vision and Pattern Recognition.

SIBBING, D., SATTLER, T., LEIBE, B., KOBBELT, L. (2013), SIFT-Realistic Rendering. International Conference on 3D Vision.

SIVIC, J., ZISSERMAN, A. (2003), Video Google: A text retrieval approach to object matching in videos. International Conference on Computer Vision.

SNAVELY, N., SEITZ, S. M., & SZELISKI, R. (2006), Photo tourism: exploring photo collections in 3D. ACM Transactions on Graphics, SIGGRAPH. 25(3), 835-846.

STEIN, F., MEDIONI, G. (1995), Map-based localization using the panoramic horizon.

TANEJA, A. BALLAN, L., Pollefeys, M. (2013), City-Scale Change Detection in Cadastral 3D Models using Images. IEEE Conference on Computer Vision and Pattern Recognition.

TANSKANEN, P., KOLEV, K., MEIER, L., CAMPOSECO, F., SAURER, O., POLLEFEYS, M. (2013), Live Metric 3D Reconstruction on Mobile Phones. International Conference on Computer Vision.

TRIGGS, B., MCLAUCHLAN, P. F., HARTLEY, R., FITZGIBBON, A. W. (2000). Bundle Adjustment - A Modern Synthesis. Vision Algorithms: Theory and Practice (pp. 298-372). Springer Berlin Heidelberg.

WOO, J., SON, K., LI, T, KIM, G.S, KWEON I.S. (2007), Vision-based UAV navigation in mountain area.

WU, C., CLIPP, B., LI, X., FRAHM, J.-M., POLLEFEYS, M. (2008), 3D Model Matching with Viewpoint-Invariant Patches (VIP). IEEE Conference on Computer Vision and Pattern Recognition.

ZEISL, B., KOESER, K., POLLEFEYS, M. (2012), Viewpoint Invariant Matching via Developable Surfaces. Workshop on Consumer Depth Cameras, European Conference of Computer Vision.

ZEISL, B., KOESER, K., POLLEFEYS, M. (2013), Automatic Registration of RGB-D Scans via Salient Directions. International Conference on Computer Vision