Virtual Reference View Generation for CBIR-based Visual Pose Estimation

Robert Huitl, Georg Schroth, Sebastian Hilsenbeck,
Florian Schweiger and Eckehard Steinbach
Institute for Media Technology, Technische Universität München, Germany
{huitl, schroth, s.hilsenbeck, florian.schweiger, eckehard.steinbach}@tum.de

ABSTRACT

Determining the pose of a mobile device based on visual information is a promising approach to solve the indoor localization problem. We present an approach that transforms localized images along a mapping trajectory into virtual viewpoints that cover a set of densely sampled camera positions and orientations in a confined environment. The viewpoints are represented by their respective bag-of-features vectors and image retrieval techniques are applied to determine the most likely pose of query images at very low computational complexity. As virtual image locations and orientations are decoupled from actual image locations, the system is able to work with sparse reference imagery and copes well with perspective distortion. Experiments confirm that pose retrieval performance is significantly improved.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; I.3.3 [**Computer Graphics**]: Picture/Image Generation

Keywords

Virtual view, synthetic view, pose estimation, visual localization, image retrieval

1. INTRODUCTION

While smartphones routinely use various localization methods based on GPS, cellular networks and Wifi networks, none of the methods available today is able to determine a user's location inside buildings with meter-level accuracy. Using visual information available through a phone's camera is a promising approach to provide accurate localization in indoor environments without complex infrastructure. By comparing the camera image to geo-tagged reference images recorded previously during a mapping run, the location and orientation of the camera can be determined.

Many vision based localization systems like the one used in this work make use of local image features, organized in a searchable index using content-based image retrieval (CBIR) methods. Once trained on a set of reference images, CBIR systems are able to rapidly identify images similar in appearance to a query image. When applied to the problem of visual localization, two major problems arise which are addressed by the approach described in the following sections.

MM'12, October 29–November 2, 2012, Nara, Japan.

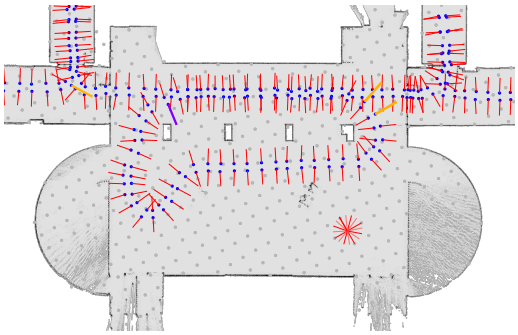


Figure 1: Images captured during mapping (blue dots, heading in red) and the virtual viewpoints created (grey dots). At each location, 16 views are computed (red compass rose).

Limited accuracy: In order to provide reference images for the image retrieval system, the environment needs to be mapped, i.e., images have to be captured at various locations and orientations, and corresponding map coordinates have to be stored. This can be achieved by a mapping trolley like the one described in [4], which automatically captures images and acquires a 3D point model as it is moved through the environment. Although automated to a large degree, mapping buildings on a large scale is time-consuming and tedious, and it is impossible to capture images at every location and orientation that can occur during localization. In practice, images are captured along a single trajectory only (see Fig. 1), drastically limiting the resolution of pose estimates returned by the image retrieval process.

Perspective distortion: The limited affine and perspective invariance of feature descriptors is a severe problem, as a location can be recognized only if a reference image with a pose similar enough to the query image exists. There has been extensive work on improving the robustness of feature descriptors under perspective distortion. As explained in the following section, robustness is gained at the expense of distinctiveness, hence these approaches tend to increase *recall* only, but not *precision*.

In order to overcome these problems and be able to use sparsely distributed reference images, we propose a novel method to extract local image features from virtual view-

points by identifying planar regions in the virtual images and applying corresponding homography transformations to reference images. By extracting local features from the generated image patches, and combining features from all patches for a viewpoint into a document that is indexed by the CBIR system, we are able to determine camera poses far off the mapper trajectory solely based on image retrieval results.

The remainder of this paper is structured as follows. In Section 2, we discuss previous work on visual localization and feature descriptor invariance under affine and perspective distortion. In Section 3, we introduce our approach for virtual view generation. In Section 4, we demonstrate the effectiveness of our approach on a realistic dataset for indoor localization. We conclude the paper in Section 5.

2. RELATED WORK

Recently, content based image retrieval approaches have been successfully applied to location recognition in textured outdoor environments [1, 2, 11, 10]. Indoor environments, however, are more challenging, as only few distinctive features are available and perspective distortion is more pronounced in narrow corridors.

To increase robustness with respect to these distortions, Morel et al. [7] introduced ASIFT, an affine invariant version of SIFT. In addition to the scale parameter, their method also simulates a set of affine distortions over various tilt angles and computes conventional SIFT features on the resulting images. The approach requires sampling of a large space of transformations which renders it computationally expensive. Similarly, Chen et al. [3] propose augmentation of a product recognition database with synthetic views under various perspective distortions. As the training data comprises surface-frontal shots of planar CD covers only, perspective invariance is easily achieved and no complex geometry needs to be taken into account.

In location retrieval, information on the 3D structure of the scene is often available, e.g., via laser scans, and can be utilized to generate locally orthogonal projections. Chen et al. [2] combine conventional, perspective images with orthogonal projections of building facades to increase invariance with respect to the viewpoint. Increasing feature invariance, however, generally deteriorates distinctiveness, which is particularly unfavorable in texture-poor indoor environments.

Assuming knowledge of 3D geometry of both query and reference features, Wu et al. [13] propose viewpoint invariant patches (VIP) that achieve invariance to 3D camera motion and avoid the unguided sampling of the previous methods. For each feature, they generate a synthetic view, orthogonal to the local surface, and compute the descriptor for the resulting patch.

An approach to generate synthetic views for improved image registration in structure-from-motion is proposed by Irschara et al. in [5]. A major difference to our approach is that synthetic views are composed of features extracted from the original images, without transforming them to their appearance at the new location. Hence, only reference images with less than about 30° rotation between the cameras are eligible. This restricts the area that can be covered with sparse reference imagery. Further, occlusions are not handled, which is of particular importance in indoor environments where obstacles and walls restrict visibility.

Baatz et al. [1] generate orthogonal projections of building facades, similar to [2]. Query images, in contrast, are also normalized to surface-parallel views after analyzing them for vanishing points. This way, query features are extracted in a viewpoint invariant way, and 6-DOF pose estimation simplifies to the estimation of three parameters (distance and x,y-offsets). However, robust extraction of planes and the generation of frontal views at query time, as well as pairwise feature matching between query and database images, are expensive steps that are avoided by our approach.

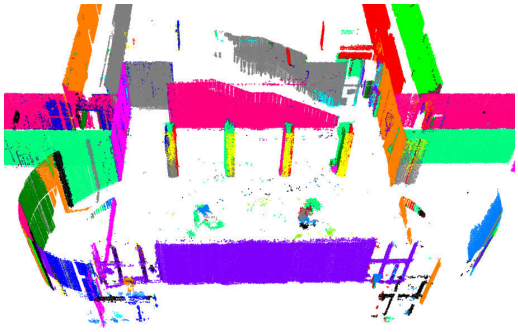


Figure 2: The point cloud acquired during mapping is segmented into planes (indicated by different colors) that provide the basis for projective transforms.

Liu et al. [6] describe an image based rendering approach similar to the one used in this paper. As their intention is to create visually pleasing views, a lot of effort is spent on properly aligning and blending the reference images. Our approach, in contrast, extracts local image features on image *patches*, hence it does not need to compute a perfectly stitched output image.

3. CREATING VIRTUAL VIEWS

The system consists of a reference database with virtual views storing the appearance at distinct locations in an environment, and an image retrieval engine that allows lookups in this database by using images as a query. The generation of the reference database is an offline process performed after an environment has been mapped. We assume that during the mapping phase, images have been captured and tagged with their 6-DOF pose (location and orientation) and that, in addition, a three-dimensional point cloud model has been acquired as described in [4]. A two-dimensional occupancy grid map is used as a reference coordinate system and to determine valid locations for virtual views.

3.1 Plane segmentation

Rendering (partial) images from arbitrary viewpoints in the 3D scene is the key component of the proposed approach. In order to keep the mapping phase and rendering of virtual views as lightweight as possible, we want to avoid triangulation of points to meshes and instead confine ourselves to simple geometric models such as planes, which are a good representation of most regions found in building interiors. Further, the projections of a plane into the image space of two cameras are related by a homography, which makes it easy to compute new views from existing images. Hence, we first identify planes in the point cloud model by robustly fitting horizontal (floors and ceilings) and vertical planes (walls) using a sample consensus method ([9], see Fig. 2). We proceed by computing a mapping M from 3D points to plane identifiers. Further, for each point P in the segmented

cloud, we determine the set of reference images I_p that depict the given point by checking whether it lies inside the viewing frustum. Raycasting from the point towards the camera centers is used to detect occlusions.

3.2 View generation

3.2.1 Identification of visible planes

First, the major planes visible in the virtual view (see Fig. 3a) are determined by casting rays from the center of the virtual camera through pixels in its image plane into the scene. When a ray hits a scene point, the map M is used to find the plane identifier. This step is performed for all pixels of the virtual view (although spatial sub-sampling in the pixel domain can be used for efficiency), and the resulting list of planes is sorted by the number of pixels that belong to each plane. Note that, for each plane, the algorithm keeps track of the pixels that are part of the plane (see Fig. 3b).

3.2.2 Image assignment

At this point, each plane is processed separately in order to find the reference images with a good view on its 3D points. In its simplest form, the algorithm combines the image lists I_p for all plane points into a single list and applies histogram binning to determine the reference image which covers the plane best. In the following step, this image is warped to the virtual viewpoint and its pixels are removed from the current plane's pixel mask (see Fig. 3c). The image assignment process is repeated until the number of pixels remaining falls below a threshold or no more reference images are available for the plane.

As the correct selection of reference images is essential for optimal results, we add two constraints to the image selection algorithm. First, an upper limit on the angle between the reference image's normal and the plane normal avoids using low-resolution views of a plane. Second, when multiple reference images cover approximately the same number of plane pixels, we pick the one closest to the virtual view's location. This avoids low resolution warping results and prefers reference images with similar perspective.

3.2.3 Image warping and feature extraction

The camera pose of the reference image is denoted by a homogeneous 4×4 matrix $\mathbf{T}_{ref}$, the pose of the virtual image is denoted by $\mathbf{T}_{virt}$. The relative transformation between both views follows as

$$\mathbf{T} = \mathbf{T}_{ref}^{-1} \cdot \mathbf{T}_{virt} = \begin{bmatrix} \mathbf{R} & \mathbf{t} \\ \mathbf{0}^T & 1 \end{bmatrix}. \quad (1)$$

With a plane defined in Hessian normal form $\mathbf{x}^T \cdot \mathbf{n} = d$, the distance between the plane and the reference image is

$$\Delta = \mathbf{t}_{ref}^T \cdot \mathbf{n} - d\,. \quad (2)$$

The homography $\mathbf{H}$ relating coordinates in the reference image to coordinates in the virtual image is then given by

$$\mathbf{H} = \mathbf{K}_{virt} \left(\mathbf{R} - \mathbf{t} \cdot \left(\mathbf{T}_{ref}^{-1} \cdot \mathbf{n} \right)^T \cdot \frac{1}{\Delta} \right) \mathbf{K}_{ref}^{-1}\,, \quad (3)$$

where $\mathbf{K}_{ref}$ and $\mathbf{K}_{virt}$ are the camera calibration matrices for the reference image and the virtual image, respectively.

Using Equation 3, the reference image is warped to the virtual viewpoint and local image features are extracted from the resulting image patch (see Fig. 4). For any non-trivial scene, the generated patch contains areas where the plane-to-plane homography is inadequate to express viewpoint change. For this reason, all features outside the pixel mask (see Sec. 3.2.1) are discarded. Finally, the features extracted from all the planes in a virtual view are combined into a single bag-of-features vector that is indexed by a CBIR system for retrieval during localization.

	P @ 1	P @ 3	P @ 5
Reference Views ($r = 5\,m$)	0.33	0.28	0.25
Virtual Views ($r = 3\,m$)	0.46	0.43	0.41
Virtual Views ($r = 5\,m$)	0.57	0.57	0.56

Table 1: Mean precision at cutoff ranks 1, 3 and 5. Relevant views are within radius r around the query location.

3.3 Localization

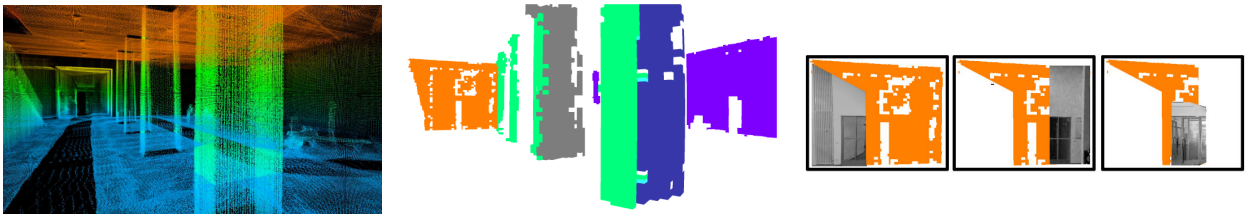
With the reference database prepared as described above, finding the position and orientation of a camera is achieved by extracting features from the query image and retrieving the most similar virtual views from the CBIR database. This step can be performed very quickly using an inverted index and has been shown to scale well up to millions of documents. Please refer to [12] and [8] for the basic concepts.

4. EXPERIMENTS

We evaluate our approach experimentally using the publicly accessible image and point cloud dataset described in [4]. The whole dataset contains more than 40,000 images of the corridors and halls of a public building. For this evaluation we used the 3,146 high-resolution close-ups of the *2011-11-28* subset, captured along a trajectory of more than one kilometer. The area shown in Fig. 1 is a small portion of this subset. The floorplan is sub-sampled to a resolution of one meter per pixel, and a virtual location is created for each "free" pixel. The height of the virtual camera is fixed at 1.50 m above ground. To simulate different orientations, virtual views are generated for yaw angles advancing in steps of $\frac{\pi}{8}$, creating 16 views per location. In total we get 6,352 locations and 101,632 views.

The image retrieval system is trained on 24.8 million SIFT features extracted from the image patches for the virtual views (see Fig. 4). We use an approximate k-means (AKM) quantizer with a vocabulary size of 200,000 visual words and TF-IDF weighting. The query time per image on a single thread is around 200 ms, however AKM can easily be configured to perform considerably faster.

We query the system using images captured at various locations in the mapped environment. To demonstrate that the system is capable of inferring the appearance at arbitrary locations, we pay attention to keeping a distance to the mapper trajectory. Four query images and the corresponding results are shown in Fig. 5. The proposed method is able to robustly determine the correct orientation (quantized to intervals of $\frac{\pi}{8}$). The top-ranked image retrieval results concentrate in the immediate vicinity of the query location in almost all cases, the orientation is determined correctly in all cases. If no virtual views were used, the accuracy of position and orientation would be limited to the pose of the reference images. Further, if the perspective distortion between query view and reference views is too large, no correct reference image might have been found at all.



(a) Points visible in virtual view (b) Planes in virtual view (c) Image assignment for left plane in (b)

Figure 3: The point cloud from the virtual view's perspective (a) is used to lookup visible planes in a pre-computed point-to-plane map (b). Images are assigned and warped to each plane (c). The mask keeps track of unassigned plane parts.

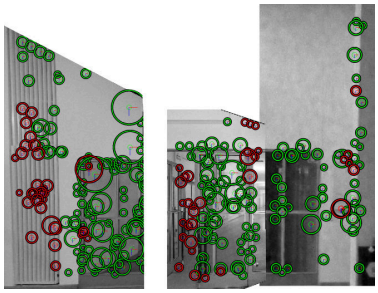


Figure 4: Warped image patches for the plane in Fig. 3c. Features inside the mask are green, outside red.

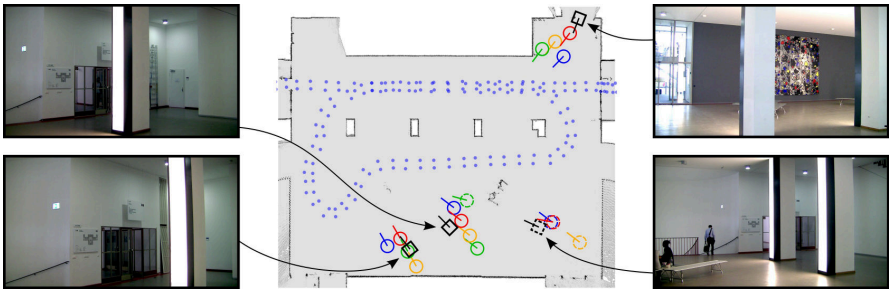


Figure 5: Top-ranked retrieval results for 4 query images (black square is the ground truth pose). Location and orientation are drawn as circles (rank 1: red, 2: green, 3: blue, 4: orange). No post-processing of image retrieval results has been applied.

Table 1 shows the mean precision over 252 queries (six frames at 42 locations) achieved by the first result, by the top-3 results, and by the top-5 results, respectively. A precision of 1.0 is achieved if all top-ranked results are relevant. Clearly, the virtual view approach outperforms the unprocessed reference images. In 56% of all cases, the top-ranked result is a correct location with our virtual view approach, compared to 33% when only reference images are used.

5. CONCLUSION

We have proposed and evaluated an approach that enables rapid visual pose estimation using sparse reference images. By generating virtual views for all potential locations and orientations within a given area, we are able to determine the pose of a query image even when no nearby reference images are available.

The approach relies on knowledge about planar regions present in the scene, and projectively transforms reference images to the virtual view's location. Query images are matched to virtual views by applying state-of-the-art image retrieval techniques. We experimentally show how a localization system based on this approach is able to robustly determine the orientation and position on a meter-level within fractions of a second.

6. ACKNOWLEDGMENTS

This work is supported by the space agency of the German Aerospace Center with funds from the Federal Ministry of Economics and Technology on the basis of a resolution of the German Bundestag under the reference 50NA1107.

7. REFERENCES

[1] G. Baatz, K. Köser, D. Chen, R. Grzeszczuk, and M. Pollefeys. Leveraging 3D city models for rotation invariant place-of-interest recognition. *International Journal of Computer Vision*, 96(3):315–334, Feb. 2012.

[2] D. Chen, G. Baatz, K. Köser, S. Tsai, R. Vedantham, T. Pylvanainen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk. City-scale landmark identification on mobile devices. In *CVPR*, pages 737–744, Colorado Springs, USA, June 2011.

[3] D. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, J. Singh, and B. Girod. Robust image retrieval using multiview scalable vocabulary trees. In *Proc. of SPIE*, number 1, pages 72570V–9, San Jose, USA, Jan. 2009.

[4] R. Huitl, G. Schroth, S. Hilsenbeck, F. Schweiger, and E. Steinbach. TUMindoor: An extensive image and point cloud dataset for visual indoor localization and mapping. In *IEEE ICIP*, Orlando, USA, Sept. 2012.

[5] A. Irschara, C. Zach, J.-M. Frahm, and H. Bischof. From structure-from-motion point clouds to fast location recognition. In *CVPR*, pages 2599–2606, Miami, USA, June 2009.

[6] T. Liu, M. Carlberg, G. Chen, J. Chen, J. Kua, and A. Zakhor. Indoor localization and visualization using a human-operated backpack system. In *Indoor Positioning and Indoor Navigation (IPIN)*, pages 1 –10, Sept. 2010.

[7] J.-M. Morel and G. Yu. Asift: A new framework for fully affine invariant image comparison. *SIAM Journal on Imaging Sciences*, 2(2):438–469, Apr. 2009.

[8] D. Nistér and H. Stewénius. Scalable recognition with a vocabulary tree. In *CVPR*, pages 2161–2168, New York, USA, June 2006.

[9] R. B. Rusu and S. Cousins. 3D is here: Point Cloud Library (PCL). In *IEEE International Conference on Robotics and Automation (ICRA)*, Shanghai, China, May 9-13 2011.

[10] G. Schroth, R. Huitl, M. Abu-Alqumsan, F. Schweiger, and E. Steinbach. Exploiting prior knowledge in mobile visual location recognition. In *ICASSP*, Kyoto, Japan, Mar. 2012.

[11] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, and E. Steinbach. Mobile visual location recognition. *IEEE Signal Processing Magazine*, 28(4):77–89, July 2011.

[12] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *International Conference on Computer Vision*, volume 2, pages 1470–1477, Beijing, Oct. 2003.

[13] C. Wu, B. Clipp, X. Li, J.-M. Frahm, and M. Pollefeys. 3D model matching with viewpoint-invariant patches (VIP). In *CVPR*, Anchorage, USA, June 2008.