Handling Urban Location Recognition as a 2D Homothetic Problem

Georges Baatz[1], Kevin Köser[1], David Chen[2], Radek Grzeszczuk[3], and Marc Pollefeys[1]

[1] Department of Computer Science, ETH Zurich, Switzerland
{gbaatz,kevin.koeser,marc.pollefeys}@inf.ethz.ch
[2] Department of Electrical Engineering, Stanford University, Stanford, CA, USA
dmchen@stanford.edu
[3] Nokia Research at Palo Alto, CA, USA
radek.grzeszczuk@nokia.com

Abstract. We address the problem of large scale place-of-interest recognition in cell phone images of urban scenarios. Here, we go beyond what has been shown in earlier approaches by exploiting the nowadays often available 3D building information (e.g. from extruded floor plans) and massive street-view like image data for database creation. Exploiting vanishing points in query images and thus fully removing 3D rotation from the recognition problem allows then to simplify the feature invariance to a pure homothetic problem, which we show leaves more discriminative power in feature descriptors than classical SIFT. We rerank visual word based document queries using a fast stratified homothetic verification that is tailored for repetitive patterns like window grids on facades and in most cases boosts the correct document to top positions if it was in the short list. Since we exploit 3D building information, the approach finally outputs the camera pose in real world coordinates ready for augmenting the cell phone image with virtual 3D information. The whole system is demonstrated to outperform traditional approaches on city scale experiments for different sources of street-view like image data and a challenging set of cell phone images.

1 Introduction

In recent years, due to the ubiquitousness of cell phones and cameras, the demand for real-time localization and augmentation of virtual (3D) information arose and several systems have been proposed to solve the location recognition problem [3, 1, 2, 6, 8–10] or the closely related image retrieval problem [4, 5, 16–18]. A commonly used scheme that we also follow extracts local features (e.g. [12, 11]) from a collection of reference images, vector-quantizes the feature descriptors to visual words and stores images as documents of these words in a database. Then for a query image techniques from web text search are applied to find the closest documents in the database, followed by a reranking of the result list based on geometric considerations.

We specifically look at the problem of place-of-interest recognition and camera pose estimation in urban scenarios, where we want to see how far we can get with visual information only. However, in contrast to general object recognition or image retrieval scenarios that cannot assume much about geometry and image content, we propose a tailored solution to the localization problem from cell phone images in a city. Here, often

- massive amounts of calibrated street level data are available for training[4]
- rough 3D city models exist[5]
- facades are planar and structures are vertically and horizontally aligned
- the camera's focal length is known approximately
- repetitive architectural elements appear that make 1-to-1 matching difficult

By projecting the offline training views to the surfaces, we can completely factorize out rotation from the recognition problem (in photometric matching and geometric verification). This enables the storage of gravity-aligned orthophotos (facade parts) in the database as opposed to densely sampling the space of all possible viewing poses. Query images can be transformed accordingly by finding the vertical and horizontal vanishing points of the given building. For recognition, matching and verification this reduces the problem to finding purely homothetic transformations, i.e. a scale and 2D offset on the building's surface. We show that this increases the discriminative power as compared to previous approaches on the one hand and allows to replace the computationally expensive RANSAC verification with a stratified homothetic parameter estimation, i.e. we perform three subsequent 1D estimates for distance, horizontal and vertical offset with respect to the building surface. Here the algorithm was designed in a way that e.g. window-to-window matches support the correct distance estimate through their scale ratio even if the match is from a different window instance on the facade's window grid. After having obtained the distance from the facade, horizontal and vertical offsets can be computed in the same way and we observe that using this reranking strategy is very effective in boosting the correct document to the first positions of the tested short list. As a side effect, we obtain the 6 DOF camera pose in absolute coordinates.

The key novel contributions are the orthophoto representation in the database allowing also for a more discriminative feature descriptor (upright SIFT), the homothetic verification scheme for repetitive structures and the exploitation of 3D building geometry so as to provide an absolute camera pose. In the next section we will relate the approach to previous work, before we go into details of the overall system and demonstrate its performance on different sources of cell phone and street level data.

[4] Nowadays several sources for image data taken from vehicles exist, e.g. Google's "Street View" or Microsoft's "Streetside". We use Earthmine's "3D street level imagery" for database creation and Navteq's "Enhanced 3D City Models" for testing.

[5] In this contribution we use extruded building outlines from Sanborn data, for more info see http://www.sanborn.com/products/citysets.asp

2 Previous Work

Location recognition at the city scale is closely related to image search and large scale object recognition for which a huge amount of previous work exist. A commonly used approach builds on top of the bag-of-features approach of [4] and the scalable vocabulary trees (SVT) of [5]. In the image retrieval scenario, usually the camera intrinsics and object geometry are unknown. It can therefore be difficult to find strong geometrical constraints for filtering the initial visual-word based results, although recent approaches look at (locally) consistent orientations and feature shapes [16–18] and exploit that pictures are usually not taken upside down. Location recognition approaches [9, 8, 6] usually know the intrinsic parameters of the camera, but do not exploit dense 3D models of the scene since these are difficult to obtain for larger environments.

The closest earlier works to ours are probably by Robertson and Cipolla [3], Wu et al. [2] and Schindler et al. [1]. The first one uses vanishing points, but works purely in 2D with local patch matching on a relatively small set of images (<100) and does not obtain 6 DOF pose in the city coordinate system since 3D information is missing. The concept of rectifying features according to vanishing points has been presented recently in [10], where the authors focused on single images. Exploiting 3D geometry has been proposed in [13] and [14], however these approaches require depth information for both images to be matched. Building on top of that, [2] uses 3D information from local reconstructions of streets of houses for database creation, but can only handle query images taken at fronto-parallel perspective relative to the building and cannot cope with out-of-plane rotations. In the field of systems using image data only [1] presented a large scale recognition system with impressive results also based upon a vocabulary tree. However, only 2D image data is used and in our experiments we show that in urban scenarios with mainly building facades 3D rotation invariant matching and recognition outperforms 2D methods. Another difference is that both of the two latter methods need RANSAC for geometric verification which can become inefficient with repetitive urban structures and high fractions of mismatches. In contrast we provide a simple stratified voting scheme for verification.

While the trend in the last years went towards building bigger and bigger databases and generating even synthetic views to sample the space of all possible points of view [6], we go into a different direction and represent only the building facades (upright orthophotos). An interesting effect of the technique is that it enables the usage of upright features, for which the feature orientation is obtained from vertical building axes, avoiding multiple descriptors for the same keypoint, avoiding potential bias of standard SIFT descriptors towards the bins of canonical orientations and allows distinguishing local structures differing by rotation. It has already been observed in face recognition [15] that exploiting the knowledge of aligned patches and reducing the invariance requirements can increase the recognition performance. Already for the SURF detector [11], rotation invariance could be disabled, however this was mainly motivated by performance reasons, while we show that leveraging rotation information helps recognition.

3 Offline Creation of the Recognition System

Fig. 1. Left: Panoramic image near the San Francisco Ferry Building grabbed by Vehicle. Right: Extruded building outline of Ferry Building.

Data Acquisition and Selection: For creating the database we exploit two sources of information (see Figure 1):

- Calibrated image data: Panoramic images captured by a vehicle driving systematically through the streets. For each of these images camera position and orientation is known from GPS and sensor data.
- 2D Building floorplans as available from land registration or fire insurance companies as well as building heights. The 2D maps can be extruded to piecewise planar 3D models approximating the buildings (see Figure 1) and each of these buildings is assigned a place-of-interest ID.

For the dataset of San Francisco, panoramic images have been taken roughly every 10 meters and 14896 places of interest have been covered.

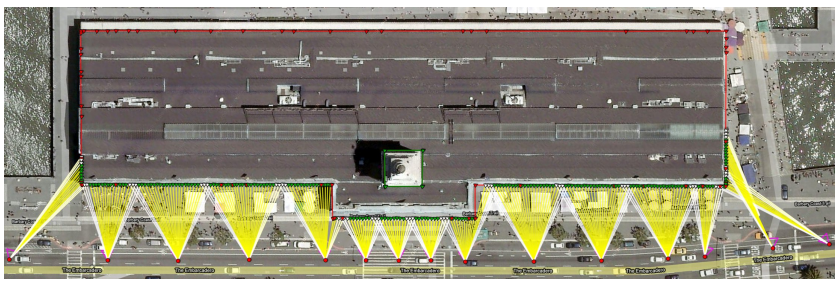

Fig. 2. Bird's eye view of Ferry Building. Portions of the panoramic images that are used to sparsely cover all facades of the POI are highlighted.

Sparse Representation of all Places-of-Interest of a City: Up to noise, resolution and model inaccuracies all panoramic images that see the same parts

of a facade should give rise to the same descriptors, so there is a huge redundancy in the captured panoramic images. While it might be beneficial to fuse multiple views of the same features, we leave the optimal redundant sampling of the facades from multiple overlapping panoramas for future work. Instead we use the following strategy to obtain a close to minimal representation of the buildings: For each POI, we find the panoramic images within 50m distance to the building outline and extract perspective images with a 60° field of view every 20°. We prune those that look away from the POI or see it at a very oblique angle. The others are selected or rejected so as to represent all the POI surface subject to minimal overlap and maximal orthophoto resolution, when projecting the view onto the facade (see Figure 2). We obtain 58601 perspective images on the San Francisco dataset.

Geometric Rectification: Using the building height information we extrude the building outlines to 3D. We then project the reference images onto these 3D surfaces and render synthetic orthoviews. Since the scene geometry is roughly

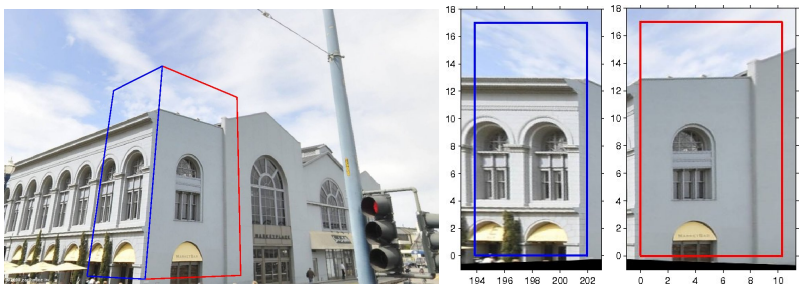

Fig. 3. Left: Building geometry projected into an image. Right: Two orhtophotos genarated from this image with overlaid geometry. The axes show the known scale in meters.

known for each of the calibrated panoramas, the image data can be projected onto the approximate geometry (see Figure 3). For each of the planar facade parts we generate orthophotos and use GPU-SIFT[6] to extract DoG keypoints and SIFT descriptors. Generally, for descriptor computation, previous approaches estimate keypoint orientations from the local gradient histogram. Rotating the local patch however in a way that the dominant peak is in the zero degree direction potentially makes the descriptors less discriminative, since all of them might have now significant mass in the zero degree descriptor bins and purely rotated local patches can no longer be distinguished. Instead, we project the gravity direction onto the facade and align the keypoints with this direction (upright SIFT). Effectively, by computing a gravity-compatible orthophoto, we

[6] C. Wu: "SiftGPU" (Version 0.5.360) http://cs.unc.edu/ccwu/siftgpu

remove all effects of 3D rotation and perspective from the image data[7]. Matching such features reduces the 6 DOF perspective recognition problem to a homothetic problem involving only scale and offset ambiguities in the 2D plane.

Scalable Vocabulary Tree Indexing: Based upon the extracted descriptors we use hierarchical k-means clustering to learn a vector quantization and build a visual vocabulary. We choose a random subset of 16M descriptors from the whole set of about 130M. We build a tree with the following parameters: split factor $k = 10$, depth $d = 6$ which leads to one million leaf nodes. We then index the bags of features using an inverted file system (IFS) for fast retrieval.

4 Recognition of Places of Interest

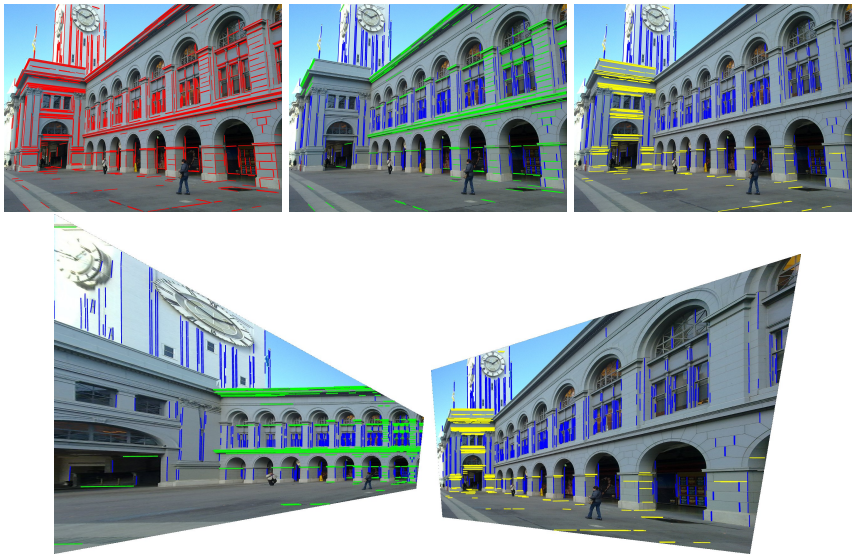

Fig. 4. Top row. Left: Query image with detected line segments. Middle and right: Lines belonging to the same vanishing point have been given the same color. Each image shows only the lines corresponding to one pair of orthogonal vanishing points. Bottom row: Two rectifications of the query image according to the two chosen pairs of vanishing points.

Removing 3D Rotation Effects from Query Image: The incoming query image is assumed to come from a calibrated camera for which we expect to roughly know whether it was held more in landscape or in portrait orientation,

[7] apart from image resolution issues due to interpolation

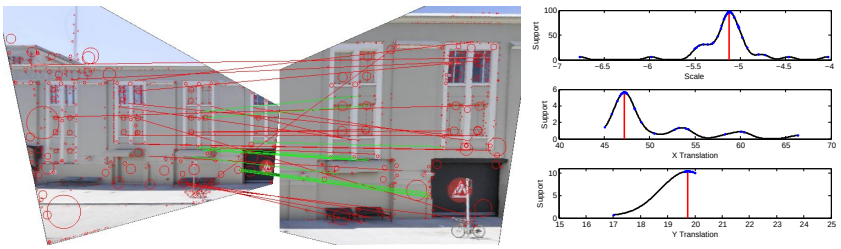

Fig. 5. Our voting scheme is illustrated using two images of Academy of Art University. Red circles indicate the scale of features, red lines are the raw correspondences and green lines are the final inliers. In the X Translation plot, note the secondary local maxima occurring at a 6m interval. They correspond to the repeating window structure. In the Y Translation plot, there is only one local maximum, since there is no vertical repetition. Also note that all but one scale inlier support the right y-offset, even though some of them vote for the wrong x-offset

so that we can correctly assign vanishing points to real-world directions. We detect line segments in the image using a method based on [19], estimate vanishing points as intersections of these lines, followed by a subsequent refinement step. Since the camera calibration is known, we can backproject the presumed vanishing points to rays in 3D space, which should be orthogonal. Every pair of points that does not fulfill this orthogonality constraint is no longer considered for rectification.

In case there are still multiple pairs of vanishing points left, we try to reduce the number of candidate pairs further. We estimate the importance of a plane by taking into account the number of lines on it and the closeness of lines corresponding to different vanishing points. We stretch the lines by 15% on both ends and then count the number of intersecting lines. For the plane with the highest number and all those within 95% of it, we generate an orthoview while discarding all the other planes.

The vertical of the rectified images (see Figure 4) becomes the vanishing point (interpreted as a ray) which is closest to the known gravity vector. On these images, we then compute upright SIFT features which are used to query the vocabulary tree. The top 50 candidates are further examined by geometric verification.

Geometric Verification Voting Scheme: So far, ranking only used frequencies of visual words for POI identification. As usual, geometrical verification of the feature configurations can be used to improve the ranking. Unlike previous approaches, who usually perform RANSAC, we leverage the fact that we are solving a homothetic problem.

Since we are matching orthophotos, we may observe differences in scale and offset that translate to the camera distance and position with respect to the facade. First we observe that for all true correspondences $\{(S_{\text{facade},j}, S_{\text{query},j})\}$

the scale ratios $\rho_i := \sigma_{\text{query},i}/\sigma_{\text{facade},i}$ should be equal up to some tolerance. When swapping the roles of the images, the same argument applies for the inverse ratios, since the problem is symmetric. Consequently, we transfer it to the logarithmic domain, and require the differences of logarithmic scale ratios to agree up to a threshold $\log t$ that depends on the expected scale estimation uncertainty of the SIFT detector:

$$|\log \rho_i - \log \rho_j| \leq \log t\,. \tag{1}$$

In order to determine the scale ratio with the most support, we use a technique inspired by kernel density estimation [20]: every scale ratio contributes a Gaussian probability density function with mean $\log \rho_i$ and standard deviation $\log t$. We then consider the sum of all these contributions and find its maximum (more precisely, the argmax). All the datapoints within a certain distance (e.g. $2 \log t$) are considered inliers.

Using the estimated scale ratio, we transform the feature coordinates of both images to a common scale. Since we know the true scale of the database image, we can have all the coordinates expressed in meters. Truly matching feature points now differ only by a global translation. The x and y components of this translation are estimated independently. We define the coordinate differences $\xi_i := x_{\text{query},i} - x_{\text{facade},i}$ and $\nu_i := y_{\text{query},i} - y_{\text{facade},i}$. As before, true correspondences should exhibit a consistent coordinate difference:

$$|\xi_i - \xi_j| \leq d \qquad \text{and} \qquad |\nu_i - \nu_j| \leq d\,. \tag{2}$$

Since all of the coordinates are expressed in terms of a known unit, we can again derive in a principled way a reasonable value for translation tolerance d, completely independently of image resolutions. We vote for x- and y-displacement separately using the same scheme as before (without transforming to log-space). The intersection of the two resulting inlier sets constitutes the final inlier set of the geometric verification (see Figure 5) and its cardinality is used to generate a new ranking of all the candidates under consideration.

This scheme has several advantages over previous approaches: RANSAC on top of an essential matrix, affine or projective transformation estimates 5, 6 or 8 parameters respectively. In contrast, our approach only needs to determine three degrees of freedom total, which means that the search space is smaller. On top of that, each degree of freedom is estimated separately further reducing the search space, which increases reliability and efficiency. In fact, we can afford exhaustively testing every hypothesis rather than sampling just some of them.

Every feature correspondence provides three constraints (scale, x- and y-coordinate). Thus, a single correspondence is enough to generate a complete hypothesis. Earlier, RANSAC-based approaches usually ignore scale and require outlier-free subsets of 5, 3 or 4 correspondences respectively. In order to hit such a set reliably, one needs to draw a number of samples which is essentially exponential in the number of required correspondences.

Finally, even wrong correspondences can still contain partial information about the solution. For instance, if one window in an image gets matched to

the wrong window in the other image, this correspondence will likely vote for the right scale ratio and possibly for one correct coordinate.

Pose Estimation from 2D-2D Correspondences: Since we used vanishing points to rectify the original query image, we obtain the camera orientation with respect to the facade directly from the vanishing points. Since the rectified image plane is parallel to the facade, the only remaining parameters are those obtained in the previous section: Since we know the facade texture in meters the scale ratio can directly be used to compute a (perpendicular) distance pos_z of the camera from the facade. Assuming the camera is calibrated with focal length 1 pixel and principal point at zero, then

$$\mathrm{pos}_\mathrm{z} = \mathrm{res}_\mathrm{facade} \cdot \sigma_\mathrm{facade}/\sigma_\mathrm{query}, \tag{3}$$

where $\mathrm{res}_\mathrm{facade}$ represents the resolution of the orthophoto in pixel/meter. The cell phone's pos_x-offset (parallel to the facade) can directly be computed from the feature position

$$\mathrm{pos}_\mathrm{x} = \mathrm{res}_\mathrm{facade} \cdot (x_\mathrm{facade} - \sigma_\mathrm{facade}/\sigma_\mathrm{query} \cdot x_\mathrm{query}), \tag{4}$$

and pos_y in an analogous way. The local camera orientation with respect to the wall is simply the inverse vanishing point rotation. Finally, the relative coordinates with respect to the facade can be converted to absolute world coordinates using the facade's pose in the world.

5 Experiments

Upright SIFT versus Traditional SIFT: In order to test whether the SIFT descriptor's discriminative power improves if we do not rotate it into the dominant gradient orientation a simple experiment has been run (see Figure 6) on the image sequences for descriptor evaluation provided by [7]. Here we warp all 5 images of such a sequence to the first image, so that orientations are the same for corresponding SIFT keypoints.[8] Features at the same position ±50% feature size, same scale ±20% and same orientation ±30° are assumed to be a geometrical ground truth correspondence, other features are assumed to be not in correspondence. By comparing every descriptor of image 1 to every descriptor in the other images we generate the precision-recall diagram for the three sequences bark, wall and graffiti (see Figure 6) as has been done in [7]. In all of these sequences upright produces a significantly higher precision for a given recall fraction of the geometrical ground truth matches. A possible explanation is that when rotating the SIFT descriptor to the dominant orientation some gradient orientation histogram entries are more likely to obtain responses than others (e.g. those of the dominant orientation). This makes it more difficult to distinguish local regions that mostly differ by a rotation whereas this is possible using upright SIFT.

[8] For this experiment, we used A. Vedaldi and B. Fulkerson's vlfeat (v0.94 available from http://vlfeat.org) for detector and descriptor in this experiment.

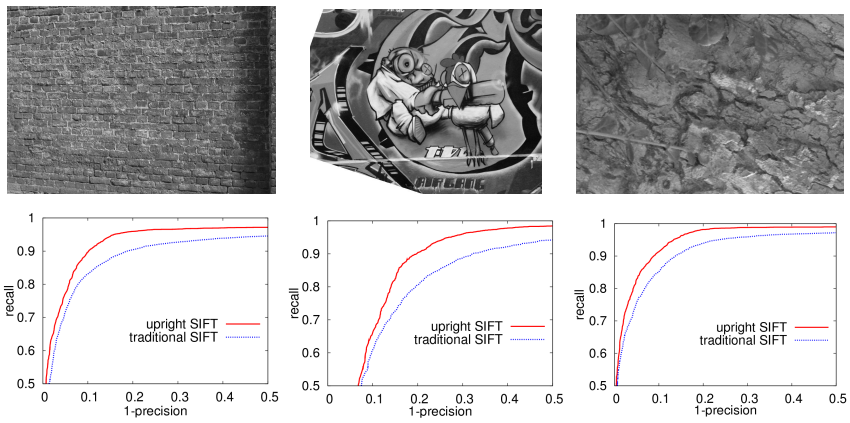

Fig. 6. Upright-SIFT vs. traditional SIFT with orientation estimation: All 5 images of the wall, graffiti and bark sequences [7] are warped to the first image of their sequence before DoG keypoints are extracted. We now compare the descriptiveness of upright-SIFT (with zero-orientation) and standard SIFT which estimates orientation from the local gradient histogram [12]. For a given precision (fraction of correct matches within all obtained matches) we get a higher recall rate (fraction of correct matches with respect to the set of geometrical ground truth correspondences.

Vanishing Point Detection: For 31034 Earthmine images, we ran the vanishing point detection algorithm. In order to measure the error, we computed the angles between the directions that were found and the horizontals/verticals of known building surfaces. The distribution of these angles is shown in Figure 7. 75% of the time, the vanishing points are estimated correctly up to 2 degrees, the median error is 0.9°.

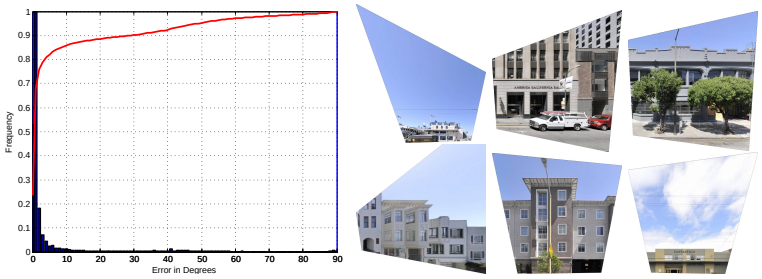

Fig. 7. Left: Histogram of orientation errors from vanishing points in degrees (blue) and cumulative curve (red), histogram scaled to the range $[0, 1]$. Right: Some rectified cell phone images

Recognition: Different variants of recognition pipelines are compared:

- **Affine** This is our reference implementation. The SVT and IFS are trained and built on the raw survey images. As feature descriptor we use standard SIFT. For geometric verification we use the affine model.
- **Masked** Same as before, except that for survey images we use geometric models to discard all features that do not lie on a building. This variant uses the same regions of the original images as the following variants. Its interest lies in testing how discarding background features affects recognition.
- **Rectified** Survey images are rectified using known 3D models of the buildings and query images are rectified using estimated vanishing points. The feature descriptor is still standard SIFT. Geometric verification is our proposed 3-degrees-of-freedom plane alignment using stratified histogram voting.
- **Upright** Survey and query images are rectified as before, but in addition we use upright SIFT. Geometric verification is again 3DOF plane alignment.

We evaluated each of these four implementations on three different query sets:

- **Earthmine** This dataset consists of 31,034 Earthmine images that were *not* selected for the training set. However, they stem from the same day and have been taken under the same conditions as the training set so that they must be considered as very easy. The images were automatically chosen such that they point towards a building. Whether or not this building is partially or completely occluded by vegetation was not a factor.
- **Navteq** This dataset consists of 182 images, sampled at angles of 70° to 120° degrees (with respect to driving direction) and 0° to 20° (tilt) from panoramic image data from Navteq, where panoramic images have been chosen such that buildings could be seen reasonably well. This data has been taken more than one year later than the Earthmine training data and with different equipment.
- **Cellphone** This dataset consists of 1180 images taken by various people with different camera phones (Nokia N95, N97, N900, N86) having between 5 and 8 megapixel resolution. These images are from pedestrians' perspective partially under extreme angles and constitute the most challenging dataset.

We examined how frequently a correct building is returned as one of the top n candidates for n ranging from 1 to 50. This information was recorded for both the ranking before and after geometric verification and for all combinations of implementations and query sets. The results are shown in Figure 8. Since we are targeting augmented reality applications, we are mainly interested in the percentages for the top ranked image. These numbers are summarized in Table 1.

We observe that the performance is generally better on the Earthmine query set than on the other two, which is to be expected since these images come from the same source as the database images.

We notice that *Affine* generally outperforms *Masked*. The difference between the two is that the database for the former contains features from both buildings

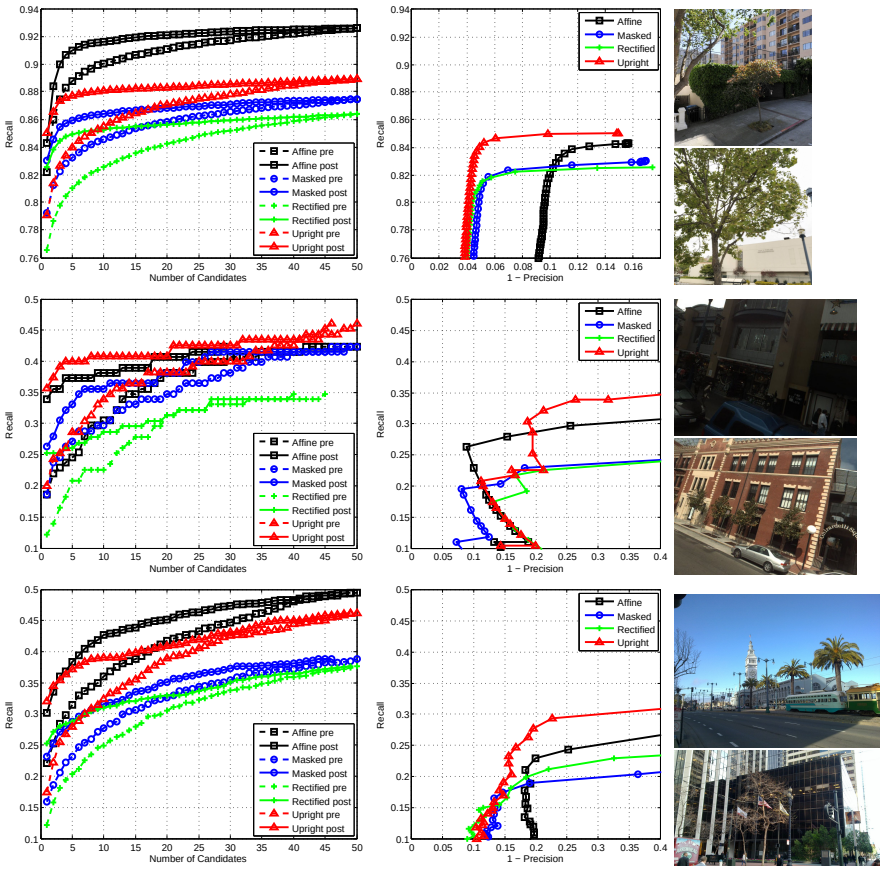

Fig. 8. Left column: Frequency of correct building being among top n candidates. Middle column: Precision-vs.-recall curve based on the number of inliers for accepting a candidate answer. Right column: Sample query images. Top row: Earthmine. Middle row: Navteq. Bottom row: Cellphone

and surroundings, while the latter uses only features from buildings. This indicates that features from the surroundings help recognition rather than distract. This is probably the main reason why the pre-verification curves of the other two methods are lower than *Affine*. They suffer from the same disadvantage as *Masked*: having ignored the features from the surroundings.

With respect to the pre-verification curves, *Rectified* does slightly worse than *Masked*. On the other hand, the post-verification curve for *Rectified* is flatter. This means that rectifying the images may hurt performance in the SVT part, but it allows for a stronger geometric verification (3DOF homothetic vs. affine).

It also paves the way for using upright SIFT. As already stated before, upright SIFT is more discriminative because it can distinguish image patches that differ only by a rotation. We see that already the pre-verification curve for our proposed

Table 1. Frequency of the top-ranked image being correct. For each dataset the best percentage has been highlighted

	Affine	Masked	Rectified	Upright
Earthmine	84.3%	83.0%	82.6%	**85.0%**
Navteq	33.9%	26.3%	25.2%	**35.7%**
Cellphone	30.2%	23.2%	25.2%	**32.1%**

method (*Upright*) is higher than for *Masked* and *Rectified*. Combined with the strong 3DOF verification, it outperforms the other methods on all three datasets with respect to the top-ranked candidate (see Table 1). On top of that, this advantage gets bigger on the more challenging datasets.

We have seen that *Affine* has the highest pre-verification curve due to the inclusion of background features. Even though *Upright* is the better overall system, combining the advantages of both methods might yield even better results. We plan to address this in future work.

We also examined the precision-recall trade-off. The number of inliers for the top candidate is compared to a threshold. If the number is below, the system returns "no-answer", otherwise it returns the top candidate. By setting this threshold to lower values, one achieves a higher recall (how often a query gets a correct answer), but also lower precision (how often an answer is actually correct). By choosing a higher threshold these spurious matches can be reduced at the cost of losing some correct matches as well.

For all three query sets *Masked* and *Rectified* share a similar precision-recall curve with a better precision than *Affine*, but a worse recall. For the Earthmine and Cellphone datasets, *Upright* is clearly the better choice, while for Navteq it depends on how one wants to trade precision for recall.

6 Conclusion

We presented an approach for recognizing places of interest in cell phone images. By exploiting approximate 3D city models it was possible to convert street level data to an orthophoto-like representation of the facades of the city. In this representation also the gravity direction is known which enabled the use of upright SIFT features which have been proven more discriminative than classical SIFT on the standard feature descriptor test sets as well as in the location recognition pipeline. The given system can be seen as 3D rotation invariant matching and allowed for estimating homothetic transformations between a rectified cell phone image and a building facade, where the parameters scale and 2D offset of the homothetic transformation can be estimated separately. This allows for an efficient 1D voting scheme related to kernel density estimation and the resulting reranking has been shown to be very effective in boosting the true image to a top position in the reranked list.

Acknowledgments. The authors would like to thank Friedrich Fraundorfer for valuable and helpful discussions as well as Ramakrishna Vedantham and Sam Tsai for help with the software and database infrastructure. They also would like to thank the following peolple from Navteq: Bob Fernekes, Jeff Bach, Alwar Narayanan and from Earthmine: John Ristevski, Anthony Fassero.

References

1. G. Schindler, M. Brown and R. Szeliski: "City-Scale Location Recognition." CVPR 2007
2. C. Wu, F. Fraundorfer, J. Frahm and M. Pollefeys: "3D model search and pose estimation from single images using VIP features." Workshop on Search in 3D, CVPR 2008
3. D. Robertson and R. Cipolla: "An image based system for urban navigation." BMVC 2004
4. J. Sivic and A. Zisserman: "Video Google: A Text Retrieval Approach to Object Matching in Videos." ICCV 2003
5. D. Nistér and H. Stewénius: "Scalable recognition with a vocabulary tree." CVPR 2006
6. A. Irschara, C. Zach, J.-M. Frahm and H. Bischof: "From structure-from-motion point clouds to fast location recognition." CVPR 2009
7. K. Mikolajczyk and C. Schmid: "A performance evaluation of local descriptors." IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 10, 2005
8. W. Zhang and J. Kosecka: "Image based localization in urban environments." 3DPVT 2006
9. Z. Zhu, T. Oskiper, S. Samarasekera, R. Kumar and H. S. Sawhney: "Real-time global localization with a pre-built visual landmark database." CVPR 2008
10. Y. Cao and J. McDonald: "Viewpoint Invariant Features from Single Images using 3D Geometry." IEEE Workshop on Applications of Computer Vision 2009
11. H. Bay, A. Ess, T. Tuytelaars and L. Van Gool: "SURF: Speeded Up Robust Features." Computer Vision and Image Understanding, Vol. 110, No. 3, 2008
12. D. G. Lowe: "Distinctive image features from scale-invariant keypoints." International Journal of Computer Vision, Vol. 60, No. 2, 2004
13. K. Köser and R. Koch: "Perspectively Invariant Normal Features." Workshop on 3D Representation for Recognition, ICCV 2007
14. C. Wu, B. Clipp, X. Li, J.-M. Frahm and M. Pollefeys: "3D Model Matching with Viewpoint Invariant Patches (VIPs)." CVPR 2008
15. P. Dreuw, P. Steingrube, H. Hanselmann and H. Ney: "SURF-Face: Face Recognition Under Viewpoint Consistency Constraints." BMVC 2009
16. H. Jegou, M. Douze and C. Schmid: "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search." ECCV 2008
17. J. Philbin, O. Chum, M. Isard, J. Sivic and A. Zisserman: "Object Retrieval with Large Vocabularies and Fast Spatial Matching." CVPR 2007
18. M. Perdoch, O. Chum and J. Matas: "Efficient Representation of Local Geometry for Large Scale Object Retrieval." CVPR 2009
19. J. Kosecka and Wei Zhang: "Video Compass." ECCV 2002
20. C. M. Bishop: "Pattern Recognition and Machine Learning." ISBN 0-387-31073-8, Section 2.5.1, page 123, 2006