Image Based Localization in Urban Environments

Wei Zhang and Jana Kosecka
Department of Computer Science
George Mason University
Fairfax, VA 22030
{wzhang2,kosecka}@cs.gmu.edu

Abstract

In this paper we present a prototype system for image based localization in urban environments. Given a database of views of city street scenes tagged by GPS locations, the system computes the GPS location of a novel query view. We first use a wide-baseline matching technique based on SIFT features to select the closest views in the database. Often due to a large change of viewpoint and presence of repetitive structures, a large percentage of matches ($> 50\%$) are not correct correspondences. The subsequent motion estimation between the query view and the reference view, is then handled by a novel and efficient robust estimation technique capable of dealing with large percentage of outliers. This stage is also accompanied by a model selection step among the fundamental matrix and the homography. Once the motion between the closest reference views is estimated, the location of the query view is then obtained by triangulation of translation directions. Approximate solutions for cases when triangulation cannot be obtained reliably are also described. The presented system is tested on the dataset used in ICCV 2005 Computer Vision Contest and is shown to have higher accuracy than previous reported results.

1 Introduction

The image based localization problem as considered in this application is comprised of three phases: location recognition, camera motion estimation between the query view and the closest reference views and final position triangulation. Thanks to recent advances in the areas of object recognition, wide baseline matching and structure and motion recovery, a variety of the techniques are currently available to tackle the individual subproblems. In order to integrate the individual components to the system and make it work in realistic setting, several challenges needs to addressed. One most notable challenge is related to the issue of obtaining reliable matches and subsequent correspondences between the query view and the reference views. Although in object recognition the true correspondences are often not essential they have been shown to improve recognition [1]. For localization the accurate correspondences are necessary for computation of the camera pose. In order to obtain reliable correspondences, we describe a modified wide-baseline matching scheme, which yields larger number of matches. The initial matching is then followed by a novel and efficient robust motion estimation technique capable with dealing with large number of outliers [2]. Integration of these two stages is crucial for obtaining accurate and repeatable results in difficult urban environments.

Overview The database of location views tagged with GPS data is initially acquired during the exploration of the environment. For the purpose of location recognition and wide-baseline matching we choose to represent each view by a set of SIFT keypoints. Given a new query view, the location recognition phase, described in Section 3, is accomplished by a voting scheme. Section 4 describes briefly a novel robust estimation technique used to identify correct matches and estimate motions between the query view and reference views. Two reference views are then selected for final GPS location triangulation. Depending on the quality of the matches, pose estimates and the amount of overlap between the reference views, the final GPS location of the query view is computed alternatively by interpolation between the two reference views. Individual steps of the approach are described in the following sections.

2 Related work

In the context of similar applications, the problem of location and building recognition has been addressed by several authors in the past, mostly considering outdoors scenes. In [3] authors used vertical vanishing direction for alignment of a building view in the query image to the canonical view in the database and proposed matching using point features followed by the relative pose recovery between the views. The alinement step relied on the presence of dominant plane and hence was applicable to scenarios with domi-

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

nant building facades. The actual triangulation using known GPS locations was not carried out. Authors in [4] focused mostly on the recognition aspects using local affine invariant regions and a set of color moment invariants to represent them. Recognition was based on the number of matched regions. In [5] the recognition was achieved by matching line segments and their associated descriptors. False matches were rejected by imposing epipolar geometry constraint. The relative pose was recovered using planar motion assumption between the views. Wide-baseline matching techniques were used for ordering of a set of widely separated views in [6]. The focus of this approach was on deciding how to 'stich' the unordered set of views assuming that they came from approximately same location. The approach was demonstrated on two different data with substantial overlap between the views. The GPS coordinates were not available, hence the location triangulation stage was not considered. In [7] the authors proposed to detect buildings using SIFT descriptors combined with the discriminative feature selection mechanism which reduced the overall complexity of the representation. Alternatively when dealing with large databases, it is desirable to use a global descriptor, to preselect small a number of candidates before carrying out recognition based on local features. For this stage color histograms were used, because of their simplicity and robustness to changes in object's scale, orientation and to some extent viewpoint. In [8] the authors described an approach to recognizing location from mobile devices using image-based Web search utilizing a hybrid color histogram and keyword search technique. However global color histograms are not very discriminative since images with similar color distributions but different content are often present.

3 Location recognition

The goal of the location recognition stage is to find the closest views from the model database to the given query view. We address this stage by means of wide-baseline matching techniques using local scale invariant features and their associated descriptors, followed by a voting stage. There are several representatives of local image features [9, 10, 11, 12] which have been shown to be robust with respect to changes in scale and/or affine transformations. In our work we use the SIFT features proposed by D. Lowe [9], which achieved best performance in the matching context based on comparison tests reported by Mikolajczyk and Schmid [13].

3.1 Matching SIFT keypoints

The SIFT keypoints correspond to highly distinguishable image locations which can be detected efficiently and have been shown to be stable across wide variations of viewpoint

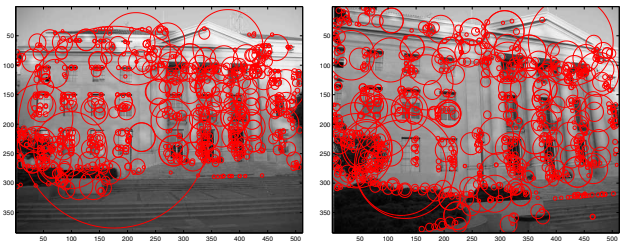


Figure 1: Examples of SIFT features. The size of the circle is proportional to scale of the feature.

and scale. Keypoints are detected by searching for peaks in the image $D(x, y, \sigma)$ which is obtained by taking a difference of two neighboring images in the scale space. Each detected keypoint has an associated descriptor, which characterizes the gradient distribution of the local image area around it. Candidate locations are obtained by searching for local extrema in pyramid $D(x, y, \sigma)$ obtained by taking a difference of two neighboring images in the scale space. Each region is endowed with a 128 dimensional descriptor f, which captures the gradient orientation information of the region, is rotationally invariant and has been shown to be robust with respect to large variations in viewpoint and scale. Figure 1 shows examples of detected SIFT keypoints. For each model image, the keypoints are extracted off line and saved in the database. After extracting keypoints from a query image, its descriptors are matched to those of the database views.

In the original matching scheme described in [9] a pair of keypoints is considered a match if the distance ratio between the closest match and second closest one is below some threshold τ_r:

$$\frac{d^2(f, f_{1^{st}})}{d^2(f, f_{2^{nd}})} < \tau_r^2, \qquad (1)$$

where $f \in \Re^n$ is the descriptor to be matched and $f_{1^{st}}$ and $f_{2^{nd}}$ are the nearest and the second nearest descriptors from the model database, with $d(.,.)$ denoting the Euclidean distance between two descriptors. The threshold $\tau_r = 0.8$ suggested in [9] was found effective for general object recognition. This ratio threshold is effective because correct discriminative keypoints often have the closest neighbor significantly closer than the closest incorrect match. In the context of buildings and street scenes, which contain many repetitive structures (e.g. windows), the above criterion will reject many possible matches, since often multiple nearest neighbors may have very close distances in the space of descriptors. Hence we chose to add another criterion, which considers two keypoints as matched, when the cosine of the angle between their descriptors f and g is above some

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

threshold τ_c:

$$\cos(f,g) = \frac{f^T g}{\|f\|_2 \|g\|_2} > \tau_c. \quad (2)$$

τ_c is set to be 0.97 in our experiments. The threshold is rather high to assure the outlier ratio would not increase, while more inliers are found. The fact that more correspondences are available makes the recognition more stable and motion estimation more accurate. This empirical value was based on the study of ROC curve associated with the threshold. The ROC curve was obtained using another database, but the threshold is general enough to ensure good performance on the ICCV database. In case multiple features pass τ_c (this happens because of repetitive structures), only the one with the highest cosine value is kept.

3.2 Coarse location recognition by voting

The closest reference views in the database are chosen by a simple voting scheme. The reference views with the largest number of matches with the query view will be selected. The locations of those reference views are likely to be close to the locations of the query view. In our experiments, top 5 views are retained. The best candidates among them will be further refined in the robust estimation stage.

4 Motion Estimation

Given the top 5 closest views the goal of this stage is to compute camera motion between the query view and the reference views, assuming that there is sufficient amount of overlap between the views. The camera motion between the query view and the matched reference view is represented as $g = (R, T)$, where $R \in SO(3)$ is the rotation and $T = [t_x, t_y, t_z]^T \in \Re^3$ is the translation. The corresponding points obey the epipolar constraint and are related by so called essential matrix;

$$\mathbf{x}_2^T E \mathbf{x}_1 = 0, \quad (3)$$

where $\mathbf{x}_2$ and $\mathbf{x}_1$ are image coordinates of correspondences and $E = \hat{T}R$. Given the correspondences obtained in the feature matching stage, the essential matrix can be estimated using a standard eight point algorithm. Once E is estimated, it can be decomposed to 4 motions. One unique solution up to a scale can be obtained using positive depth constraint. The essential matrix model is suitable, when corresponding points are in general position. In case the correspondences come from a plane, the 8-point algorithm becomes degenerate and planar model needs to be used. In planar case the corresponding points are related by a homography model:

$$\mathbf{x}_2^T \propto (R + \frac{1}{d} T N^T)\mathbf{x}_1 \propto H\mathbf{x}_1 \quad (4)$$

where N is the plane normal and d is the distance of the plane from the origin. Given at least 4 correspondences, H can be estimated using 4-point linear algorithm. Given H there are two physically possible solutions for decomposition into R, T and N, d. The correct solution can be identified by either having some prior knowledge about the scene's plane normal or by using additional view (and its associated homography) and choosing the solutions where the two plane normals are consistent. The decomposition of the essential matrix and planar homography is a standard textbook material. We will refer the reader to Chapter 5 of [14] for the decomposition formulas.

Camera Calibration The above models assume that the intrinsic camera parameters are known. In the context of our application the camera is calibrated from a single view, using vanishing points information. Vanishing points can often be found reliably in man-made environments. In case three vanishing directions can be recovered, assuming zero skew and aspect ratio is 1, both the focal length and center of projection can be recovered [15] . In case one of the vanishing points lies at infinity, we assume that the center of the projection is known (and is in the center of the image) and estimate the focal length only. Alternatively, camera intrinsic parameters can be obtained from inter-image homographies relating different views of spherical panorama. In case camera is not calibrated the corresponding points are related to each other by so-called fundamental matrix F, with $F = K^{-T} E K^{-1}$, where K is the matrix of camera intrinsic parameters.

4.1 Robust Estimation

The matches used in the location recognition stage typically contain many mismatches. Although they are sufficient for selecting the likely reference views, the ranking of the views and the search for the true correspondences must be refined by imposing global geometric constraints. The RANSAC [16] algorithm is commonly used technique for robust estimation of the model parameters. The standard RANSAC algorithm first randomly selects M samples, for each sample estimates parameters of the model hypothesis and finds the support (typically, the number of inliers) for this hypothesis. In the second stage the hypothesis with the largest support is chosen as model and all its inliers are used to refine the model parameters. The number of samples M required to obtain a confidence ρ that at least one sample is outlier free can be computed as:

$$M = \left\lceil \frac{ln(1-\rho)}{ln(1-(1-\epsilon)^p)} \right\rceil \quad (5)$$

where p is the number of points per sample and ϵ is the fraction of outliers. Table 1 shows, that when the inlier ra-

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

Inlier ratio	60%	50%	40%	30%	20%
F	106	382	1827	13696	234041
H	22	47	116	369	1871

Table 1: The theoretical number of samples M required for RANSAC to ensure 95% confidence that one outlier free sample is obtained for estimation of F (seven-point sample) and H (four-point sample). The actual required number might be magnitude more.

tio is low a large number of samples is needed for estimation of fundamental (essential) matrix. In addition to the large number of samples, the standard RANSAC requires knowledge of the residual threshold for inliers and the ratio of inliers. Due to the viewpoint change and a large amount of repetitive structures, usually a large percentage ($> 50\%$) of matches is not correct. Here we describe a novel algorithm to deal with this problem. Similarly as in the standard RANSAC scheme we first use sampling to generate a set of hypotheses (i.e. fundamental matrices). This is achieved by sampling the set of correspondences by selecting 8-point samples and estimating F using the standard 8-point algorithm with normalization. At this stage our method dramatically departs from the previously proposed approaches. Instead of evaluating each hypothesis, we propose to evaluate residuals of each correspondence and classify the points as inliers/outliers directly.

For each data point (e.g. correspondence) we study the distribution of the errors with respect to all hypotheses. For each hypothesis F_j instead of considering residual error $r^2 = (\mathbf{x}_2^{iT} F_j \mathbf{x}_1^i)^2$ we use the so called Sampson error. Sampson distance is a first order approximation of the reprojection error. Given a fundamental matrix F, the Sampson error for i^{th} correspondence is defined as:

$$E_s = \frac{(\mathbf{x}_2^{iT} F \mathbf{x}_1^i)^2}{(F\mathbf{x}_1^i)_1^2 + (F\mathbf{x}_1^i)_2^2 + (F^T\mathbf{x}_2^i)_1^2 + (F^T\mathbf{x}_2^i)_2^2} \qquad (6)$$

where $\mathbf{x}_1^i$ and $\mathbf{x}_2^i$ are the image coordinates of corresponding points and $(F\mathbf{x})_k^2$ represents the square of the k-th entry of the vector $F\mathbf{x}$.

As Figure 2 illustrates, residual distributions of the inliers typically have strong peaks close to 0, while residuals of the outliers are more spread out. Although the outliers also have a high value for the first bin, because some hypotheses are generated using samples which contain the outliers, it is considerably lower than the inlier case. These qualitatively different properties of residual distributions for inliers and outliers can be captured by lower order statistics computed from the distributions. We used skewness and kurtosis, which yield qualitatively different values for inliers and outliers. Skewness γ measures the asymmetry of the data around the sample mean μ and kurtosis β is the

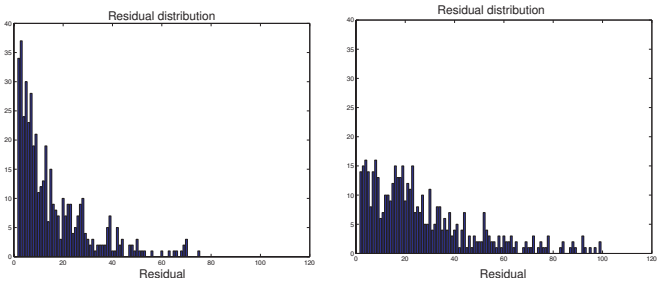


Figure 2: Error distribution for a true inlier (left) and a true outlier (right), when data contains 50% outliers.

degree of peakedness of a distribution. They are defined as:

$$\gamma = \frac{E(x-\mu)^3}{\sigma^3} \quad \text{and} \quad \beta = \frac{E(x-\mu)^4}{\sigma^4} \qquad (7)$$

Skewness of the normal distribution (or any perfectly symmetric distribution) is zero. If the value of skewness is positive, the data are spread out more to the right of the mean than to the left. Given these features the problem of inlier identification is then formulated as a classification problem. Note that the computation of histogram analysis is trivial in comparison with the computations in the sampling stage. The approach outlined above can successfully classify outliers and inliers with only a fraction of the computational cost of the standard RANSAC. This is due to the fact that our approach does rely on the existence of an outlier free sample, since it does not use the hypothesis evaluation stage. It is the entire ensemble of hypotheses which determines, whether the point is an inlier or an outlier. More details about the approach can be found in [2]. The computational requirement of the standard RANSAC algorithm to estimate H is relatively low (see Table 1), thus we used it for homography estimation while retaining the efficiency of the entire system.

4.2 Motion model selection

In the context of this application, the general motion model captured by fundamental matrix F is often not appropriate and the homography model is favored for the following reasons:

1. Buildings are dominant in city scenes. It is likely that corresponding points are located in planar building facades. Even the case when correspondences are in general position, usually a large number of points come from the same plane.

2. Inliers decision of the fundamental matrix is based on residuals measured by the Sampson distance. Thus the error along the epipolar line is not accounted for.

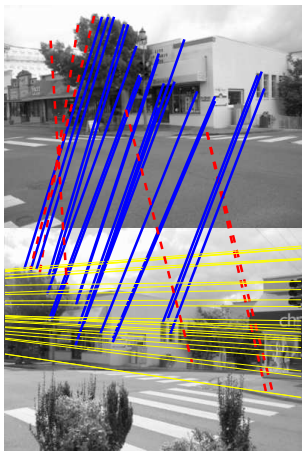
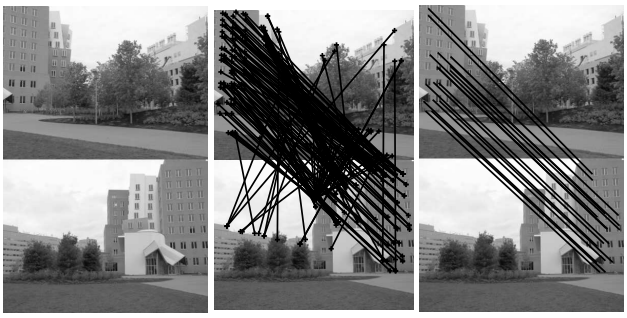


Figure 3: The matching points connected by dashed red line are not correct correspondences, yet they are chosen as inliers because their errors are along the epipolar lines.

This makes the process of inlier identification for fundamental matrix more difficult. It is often the case that features are matched to wrong locations, but their Sampson errors are small and hence they are considered as inliers, as shown in Figure 3. This is more likely to happen when there are many repetitive structures in the scene.

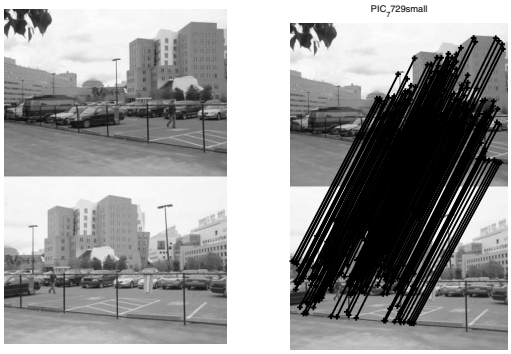
In the presented system we always attempt to fit the homography to the correspondences set first. In case homography model is selected, in addition to the *gross outliers* (the incorrect correspondences), those correct correspondences which does not comply with the homography constitute the *pseudo outliers*. Even though the inlier ratio would be less than the actual percentage of correct correspondences, a correct homography can still be obtained as long as enough points come from one plane. As shown in Table 1, even with only 20% inliers, theoretically 1871 samples are enough obtain the correct estimate. Of course, more samples are needed in practice. We use 5000 samples to ensure that correct homography can be obtained. As Figure 4 demonstrates, even though building facade only takes a small portion of the image, correct correspondences can still be obtained based on the homography model.

For the scenes where there are no major planes as shown in Figure 5, a homography model does not have sufficient support. In those cases, a fundamental matrix model is chosen. The system determines whether the fundamental matrix is needed by checking the homography fitting result. If the number of inliers of the estimated homography is lower than some threshold τ_N, a fundamental matrix will be fitted instead; τ_N is set to be $0.2N$, where N is the number of putative matches.



(a) (b) (c)

Figure 4: (a) original image pair. (b) correspondences found by matching of SIFT features. (c) correct correspondences, which lie in one plane, are identified using homography model.



(a) (b)

Figure 5: Fundamental matrix based inlier identification.

5 Final localization

Based on the motion estimation results, we re-rank the top 5 reference views based on the number of correctly identified matches. Top 3 views are selected, then two *best* reference views are chosen for the final localization. Notice they are not simply the top 2 views. In this case, not only the two reference views need to be close neighbors of the query view (so that motion between query view and them can be reliably estimated), but also the motion between the two reference views needs to be reliably estimated. It may happen that viewpoint change between the top 2 views is quite large and correspondences can not be set up reliably. To address the problem, we select from 2^{nd} and 3^{rd} views the closest view to the 1^{st} nearest neighbor. This selected view, together with the 1^{st} nearest neighbor, constitute the two reference views.

If all three camera motions can be recovered, the location of the query view can be obtained by triangulation. Let (R_{1q}, T_{1q}) be the motion between the first reference view and the query view, (R_{2q}, T_{2q}) be the motion between the

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

second reference view and the query view and (R_{12}, T_{12}) be the motion between the first and the second reference view. Note T_{12} and T_{1q} are with respect to the coordinate system of the first view, while T_{2q} is not. For the triangulation to proceed, they need to be represented in same coordinate system as the other two translations. Hence we set $T'_{2q} = R_{12}^{-1} T_{2q}$. The translation vectors are then projected to the plane, disregarding their vertical component, since we only need plane coordinates. Now the three translation vectors, all in the coordinate system of the first reference view, determine the shape of a triangle with three vertexes being the three positions (query, first and second reference). Since the positions of the first and second reference are known, the size of the triangle is also fixed. Thus the location of the query view can be determined.

In some cases, even though two reference views can be found, the motion between them can't be reliably estimated because they are widely separated and have a small overlap. Our solution to this situation is to interpolate the positions of the two reference views. Assuming that closer view would have more correspondences with the query view, the query position is computed as:

$$\frac{N_{ref1} P_{ref1} + N_{ref2} P_{ref2}}{N_{ref1} + N_{ref2}},$$

where N_{ref} represents the number of correspondences and P_{ref} represents the location of reference view. The system would choose this solution when the number of identified inliers is less than some threshold (16 in our experiments). Note that there are cases when only one reference view can be found. The reason is that some reference views are widely separated (having little overlap) and the query view happened to be close to one of them, or the query view is "extrapolated" (outside the location range of reference views). In the system, if the number of identified inliers for the second reference view is too limited (8 in our experiments), only one reference view will be used and its position is assigned to the query view.

6 Experiment

Our experiments were based on the dataset provided by 2005 ICCV Vision Contest, taken by Richard Szeliski (http://research.microsoft.com/iccv2005/Contest/). The images in the datasets are not ordered, making the multi-view relationships difficult to obtain. The GPS position of a subset of images are provided and the locations of the unlabeled images need to be estimated. The ground truth position of those unlabeled images are also provided. Prior to feature extraction, we sub-sampled the images and alleviated the radial distortion using a fixed coefficient for all images. The GPS positions are in the spherical coordinates (latitude/longitde) of the Earth, not the Euclidean coordinates (meter) which is needed here. The conversion to meters is given by:

$$\begin{aligned} d_{long} &= R \cos((Lat_0 + Lat_1)/2) \sin(Long_1 - Long_0) \\ d_{lat} &= R \sin(Lat_1 - Lat_0) \\ d &= \sqrt{(d_{long})^2 + (d_{lat})^2} \end{aligned} \quad (8)$$

where $R = 6.3713 \times 10^6$. When the Euclidean coordinates are obtained, they can be converted into spherical coordinate by inverting Equation 8.

Figure 6 and Figure 7 demonstrate two examples where the triangulation of motion vectors is used to calculate position. In comparison with ground truth, the localization error is within 4 meters for Figure 6 and within 2 meters for Figure 7. Figure 8 shows an example when interpolation was used for localization. The localization error is 8 meters. Figure 9 shows an example where only one reference view can be identified. In this case, the location of the reference view is assigned to the query view. Table 2 summarizes the experimental results for the dataset which was used to evaluate contest results. Based on the scoring criteria of the contest, the average score of the results is 3.68. This result is better than our submission for the contest and other reported results, which can be found in (http://research.microsoft.com/iccv2005/Contest/Results /Results5Final.htm). The total execution time of computing locations of all the test views is 24 minutes, which is shorter than most reported execution times. This is mostly due to the efficiency of the robust motion estimation stage.

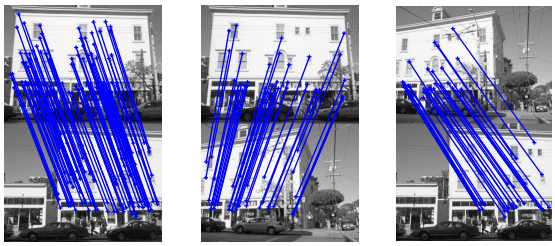


Figure 6: An example where localization is based on the triangulation and all three motions can be estimated using the homography model. Left: Identified correspondences between the query view (top) and the first reference view (bottom). Center: Identified correspondences between the query view (top) and the second reference view (bottom). Right: Identified correspondences between the second view (top) and the first reference view (bottom).

7 Conclusions

In this paper we present a prototype system for image based localization in urban environments. The system is com-

Figure 7: Another example where localization is based on the triangulation and all three motions can be estimated based on fundamental matrix. Left: Identified correspondences between the query view (top) and the first reference view (bottom). Center: Identified correspondences between the query view (top) and the second reference view (bottom). Right: Identified correspondences between the second view (top) and the first reference view (bottom).

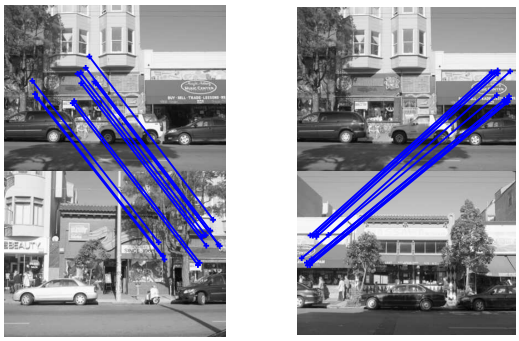


Figure 8: Two reference views can be found, yet they have no overlap. Therefore the motion estimation between reference views is not possible and the interpolation of these positions is used instead. Left: Identified correspondences between the query view (top) and the first reference view (bottom). Right: Identified correspondences between the query view (top) and the second reference view (bottom).

prised of three phases: coarse location recognition, camera motion estimation between query and reference views and final position triangulation. The coarse location recognition is base on the matches of SIFT keypoints. A novel robust estimation technique is used for motion estimation based on the putative matches which usually contains large portion of outliers. Thus the efficiency of the system is ensured in addition to the accuracy. Appropriate motion model is automatically selected for different scenes. The system is shown to be both accurate and efficient based on the dataset provided by ICCV Computer Vision Contest. Further improvements can be obtained by carrying out additional search for matches, in case there is little overlap between the reference views. The accuracy could be further improved by non-linear refinement of the motion estimates and more accurate off-line calibration stage. The accuracy of system depends

Figure 9: Only one reference view (bottom) can be found for the query image (top), because the query view is "extrapolated".

Error range	$< 2m$	$< 4m$	$< 8m$	$< 16m$	$> 16m$
Distribution	7	3	6	6	0

Table 2: Distribution of the localization errors.

on the density of the model views available in the database. Therefore, selection of optimal set of model views requires further investigation.

References

[1] T. Berg A. Berg and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *CVPR*, 2005.

[2] W. Zhang and J. Kosecka, "A new inlier identification procedure for robust estimation problems," in *Robotics: Science and Systems*, 2006.

[3] D. Robertson and R. Cipolla, "An image-based system for urban navigation," in *BMVC*, 2004.

[4] H. Shao, T. Svoboda, T. Tuytelaars, and L. Van Gool, "Hpat indexing for fast object/scene recognition based on local appearance," in *Computer Lecture Notes on Image and Video Retrieval*, July 2003, pp. 71–80.

[5] T. Goedeme and T. Tuytelaars, "Fast wide baseline matching for visual navigation," in *CVPR'04*, 2004, pp. 24 – 29.

[6] F. Schaffalitzky and A. Zisserman, "Multi-view matching for unordered image sets," in *ECCV'02*, 2002, pp. 414–431.

[7] Lucas Paletta and Gerald Fritz, "Urban object detection from mobile phone imagery using informative sift descriptors," in *SCIA*, 2005.

[8] T. Yeh, K. Tollmar, and T. Darrell, "Searching the web with mobile images for location recognition," in *CVPR*, 2004.

[9] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, 2004.

[10] C. Schmid and R. Mohr, "Local greyvalue invariants for image retrieval," *Pattern Analysis and Machine Intelligence*, vol. 19, pp. 530–535, 1997.

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE

[11] T. Tuytelaars and L. Van Gool., "Matching widely separated views based on affine invariant regions," *IJCV*, vol. 59, 2004.

[12] Jiri Matas, Ondrej Chum, M. Urban, and Tomas Pajdla, "Robust wide baseline stereo from maximally stable extremal regions," in *BMVC'02*, 2002, pp. 384–393.

[13] K. Mikolajczk and C. Schmid, "A performance evaluation of local descriptors," in *CVPR 2003*, 2003.

[14] Yi Ma, Stefano Soatto, Jana Kosecka, and Shankar Sastry, *An Invitation to 3D Vision: From Images to Models*, Springer Verlag, 2003.

[15] J. Kosecka and W. Zhang, "Video compass," in *Proceedings of European Conference on Computer Vision*, 2002, pp. 657 – 673.

[16] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," in *ECCV'96*, 1996, pp. 683 – 695.

Proceedings of the Third International Symposium on
3D Data Processing, Visualization, and Transmission (3DPVT'06)
0-7695-2825-2/06 $20.00 © 2006 IEEE