Building recognition for mobile devices: incorporating positional information with visual features

ROBIN J. HUTCHINGS AND WALTERIO W. MAYOL[1]

Department of Computer Science, Bristol University, Merchant Venturers Building, Woodland Road, BS8 1U, United Kingdom

robinhutchings@gmail.com, wmayol@cs.bris.ac.uk

Abstract: In this paper we describe a system that can accurately match buildings in photographs taken with mobile devices. Computer vision methods are enhanced by a novel use of the GPS (Global Positioning System) position of a user. With a query image's GPS location, an approximation of the view of each building from this position in world space can be calculated. Buildings are represented as planar surfaces and an efficient geographical database positions each building in world space. By using planar rectification of the building image, the effects of perspective change are accounted for. With such planar representations, accurate geometric analysis can then be used to help determine the correct building match. With these methods, a recognition system has been formed that has been shown to be accurate for a wide range of viewpoints and image scaling of a building.

Keywords: Building recognition, GPS, computer vision, mobile computing, tourism.

1 Introduction

The purpose of this work is to create a mobile application which will allow a user to discover information about a given building or landmark. After determining in which building the user is interested, specific information can be returned about this building and/or the surrounding area. In particular, this technology can be used as part of a tourist application based in a city space, which will provide both historical and current information about the area.

A number of mobile electronic tourist guides have been implemented elsewhere, with varying degrees of success in providing truly contextual information. The GUIDE system [1] for example, provides a contextual tourist service in the city of Lancaster. The closest

[1] Corresponding author: Walterio Mayol wmayol@cs.bris.ac.uk, Computer Science Department University of Bristol. Woodland Road, BS8 1UB, UK. Tel: (44) or (0) 1179545128.

base station in this network approximately determines a user's position in world space. With base stations around 200 metres apart, any contextual information returned to the user must cover this large area. Other electronic tourist guides have been developed that use the position of a user to find what attractions are in the surrounding area. A system known as LoL@ [2] provides an electronic guide of Vienna based on GPS position. Using the GPS location a general guide to the city is produced. The granularity and detail of any context aware service will ultimately depend on how precisely the location and interest of the user can be determined. A system based on building recognition can offer a high precision of contextual information.

In terms of the use of computer vision for building recognition, recently, in [17] a system is presented in which visual features are extracted in order to align buildings for matching and in [18] a system that uses colour based features to reduce the database of potential candidates is described. However, in these approaches, cues used to align and select views are purely dependent on visual features and moreover testing is limited in terms of the significant clutter and changes in scale that a city navigator system demands.

For a successful guide of this kind the accuracy the building recognition must be high and it has to account for a number of challenging image conditions. In this work we combine the conventional GPS-based localization approach with the refinements and accuracy that visual features can provide.

The paper is divided as follows. Section 2 describes the collection of images to form a database geographically organized, Section 3 describes the method used to extract and represent visual features. Sections 4 and 5 deal with the incorporation of the GPS data and the viewpoint variations respectively. Section 6 presents the methods for verifying that a true match has occurred before results are presented in Section 7. The paper ends with conclusions in Section 8.

2 Developing the image database

The equipment to capture photographs and sensor information has been provided by Hewlett Packard and the Mobile Bristol Project. Mobile Bristol is a mobile software platform that has been used for a range of applications, tourist based or otherwise. To be able to match a building inside a query photograph, the query image needs to be compared and then matched against a database of building images. When building this image

database an IPAQ rx3700 personal computer connected to a GPS unit and electronic compass is used, as shown in Figure 1.

Figure 1: The tools used to build the database. An IPAQ personal computer is shown with the compass attached so that a viewing direction can be obtained for the templates in the database. A Bluetooth GPS unit was used to provide the positional information for both queries and templates.

With these sensors, the position of where a photograph is taken from, as well as the viewing direction are known. The GPS position and compass direction are stored alongside each of these images. By adding an estimate for the distance to the building for each template, the position of the actual building in world space is calculated. The building images in the database and their associated positional information are referred to as templates. If the front of a building is considered as a planar surface then this plane should be perpendicular to the viewing direction. Conceptually, when all the buildings are loaded into the database in this way, a representation of the real world is formed by a series of planar surfaces as shown in Figure 2.

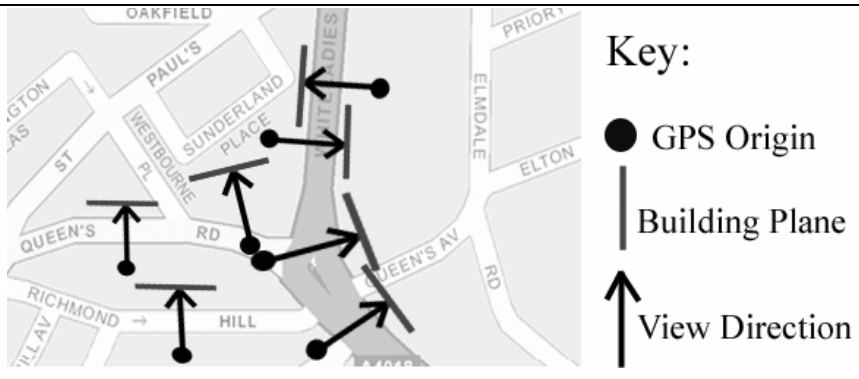

Figure 2: A template in the database is defined by the GPS position from where the image was taken, view direction and the distance to the plane. Now each building image is a defined plane in world space.

When a query photograph is taken, a GPS unit is also used to record the current real world position (note that a compass is not used when taking a query photograph.) Each building

in the database has a position in world space as shown in Figure 2, so buildings that are too far away from the query can be discounted. The amount of buildings that can be stored in the database is now limited only by the amount of memory, since the matching process can still run at speed by discounting many of these buildings for each query.

More complex 3D structures can be represented in the database by using images of the various planes that make up this building, all indexed to the same name. A rectangular building where all sides of the building are visible would therefore be represented with a database plane for each side of the building.

3 Image Features

In order to match objects (in this case buildings) between two photographs, both images are analysed and a feature-based representation formed for each image. The representations formed should be comparable, and an estimate found of the similarity between the two photographs (the template image and the query image) of the buildings. To be able to compare the images in this way, meaningful information needs to be extracted from both images. A model often used in computer vision, is where the pre-attentive stage detects key descriptive points of an object. In the attentive stage these points are then combined into a grouping or model of the object for purposes of recognition and matching [3]. An *interest point* or *corner* can be thought of as a small and descriptive pixel region of high image contrast in all directions. A given image can contain several of these interest points. If such points are to be used in a matching system then they must contain enough varied pixel information to give a level of certainty in the correctness of any matches, as well as being individual or rare. The use of interest points to match objects has been proven successful for a wide range of objects and conditions with many successful methods present in the literature (see [4,6] for a review).

One particularly relevant benefit of this approach is the robustness to clutter and occlusion for the matching process. In a system that matches buildings it is highly unlikely that the whole building will be visible in all photographs due to occlusion from people, cars, other buildings and so forth. If interest points are being used, then only a relatively small number of points between two images need to be matched for a positive identification.

3.1 Detecting interest points

Corner points often have the strongest intensity of pixel gradients that represent a unique structure around this pixel position. If the point is a strong corner (in terms of the corner detection metric used) then the corner should still be detectable from a wide range of different angles. These two facts mean that corners in images can potentially be used as part of a salient detection scheme when matching buildings in images.

To detect the interest points the Harris-Stephens detector [5] has been used. The Harris detector is a robust repeatable corner detector when considering a level of image rotation, perspective and illumination changes [6]. This detector analyses image derivatives in a windowed region to determine the corner strength at a given pixel. It is then possible to alter the sensitivity of the detector to locate more or less corner points, as shown in Figure 3. The sensitivity was set to find on average 1000 corner points in each building image.

Figure 3: The white squares show the detected interest points with different sensitivities of corner detection, with: a) 178 points detected, b) 788 points detected.

3.2 Representing interest points

In order to reliably match interest points in two different images of the same building, it is necessary to assign them a more unique descriptor. A truly successful descriptor should be invariant to the possible variables that can occur. For purposes of building matching the variants that need to be considered are:

1. Illumination from different lighting conditions
2. Scale change, from a change in distance to the building
3. Rotation caused by orientation of the camera
4. Affine and perspective transformations caused by changing viewpoints of the building.

Much work has been completed to represent interest points [7,8,9,10] with varied results of success. A traditional descriptor is simply an image patch centred on the point of interest [15]. With changing viewpoints and lighting conditions, the SIFT descriptor by Lowe [11] has been found to outperform a range of different local descriptors [4] and forms highly distinctive representations of interest points. This descriptor uses methods derived from biological vision and has been successfully used in a number of practical applications [12,13]. SIFT allows for significant pixel shifts and therefore changes in view point of a building and has been used for this work. This descriptor analyses a 16x16 pixel region around each interest point and forms a series of histograms describing the pixel content. These histograms are normalised and form a 128 parametre representation that uniquely describes the region surrounding the interest point.

4 Incorporating GPS to deal with scaling

With varying viewpoints and distances to a given building, the scaling of the building in the query photograph when compared against a template of the same building will alter.

In the literature [11,14], interest point detectors have been developed that search through scale space to detect a corner. If repeatable interest points can be found across different scales of images then this is particularly useful for object recognition where scaling is an issue. The Harris Laplacian detector by Schmid and Mikolajczyk [14] for instance extends the Harris detector to find interest points in scale space. However, in the case of this work, the GPS position of any query image can be used to determine the scaling needed for each database image. This in effect avoids the exhaustive searches through scale space, removing the need for more complex scale space detectors and improving recall time and saving computational resources, criteria of paramount relevance for mobile devices.

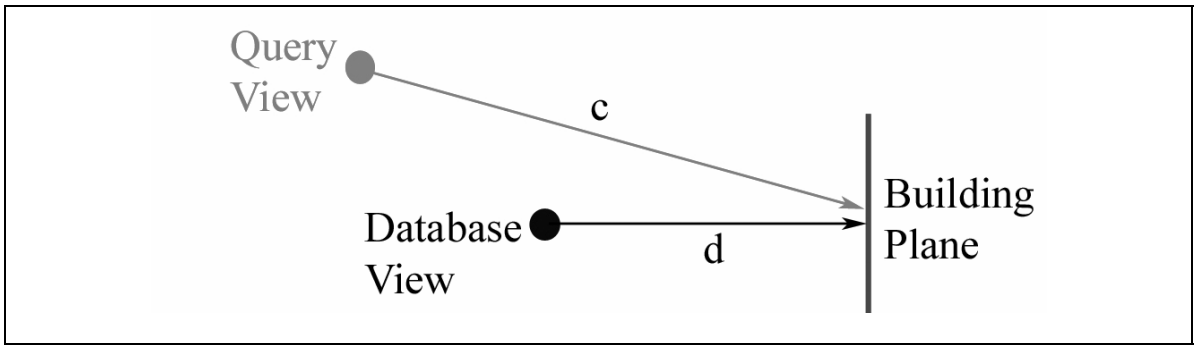

Figure 4: The position of the Query View is found from the GPS position. Since the building is represented in the database by a plane in world space the distance c can be found and the necessary scaling of the Database View needed for matching calculated.

As the distance c in Figure 4 grows, the scaling of the building in the Query View compared to the building in the Database View (a template) will increase. Without further changes, the descriptors that have been found in the template will no longer match against the corresponding points in the query view of the same building. Therefore, to be able to match the interest points for changing viewing distances it becomes necessary to build representations of these points at a range of scales.

For each interest point that has been found in the template image different scales of descriptors are created and stored. These scaled descriptors in effect represent how each interest point would appear in a photograph if taken from a greater or lesser distance than the original template image. The pixel regions around the interest points in the template image are bi-linearly interpolated into a 16x16 region as shown in Figure 5 and histograms built from this new pixel region.

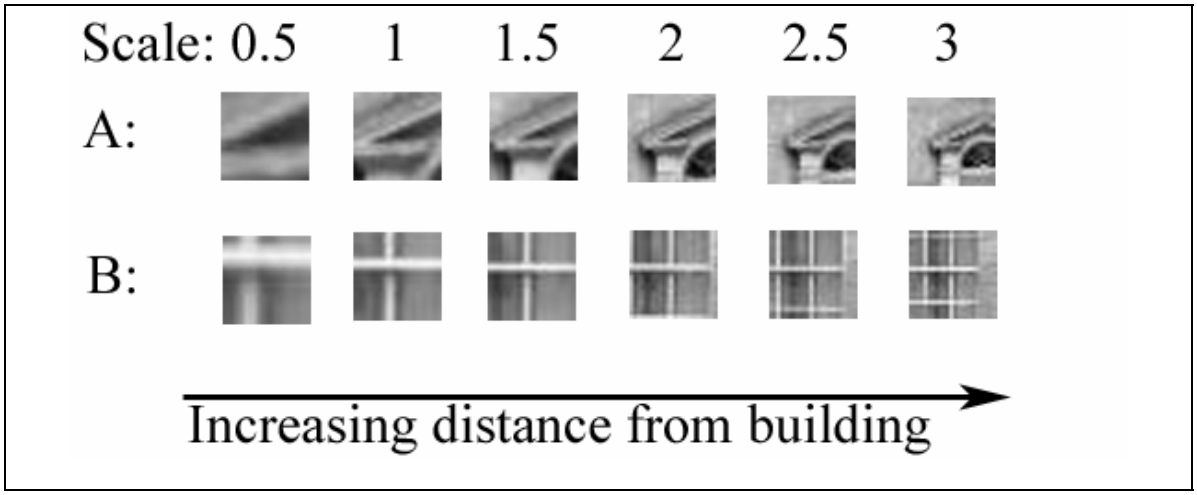

Figure 5: Two examples where descriptors are built for each scaled representation of a salient point in the template images. Scale of 2 is a 16x16 region formed by an interpolation of a 32x32 region centred on the point of interest, scale 3 is formed by a 48x48 region and so forth. The representation that is used for the matching process depends on the relative distances c and d from Figure 4.

Since the building is considered to be planar, the ratio of the distance c and d from Figure 4 is used to determine which scale of descriptors to search for a given query. For instance, if the query image is twice as far away from the building as the template then the scale 2 will be tested against. The chosen set of scaled descriptors is then used for this template image in the matching process. To test the effectiveness of this approach to scaling, the system was tested with and without the extra scales of descriptors, with the results shown in Figure 6 where:

$$Distance\ Ratio: \quad \frac{Distance\ of\ query\ to\ building}{Distance\ of\ template\ to\ building} \qquad (1)$$

Only one template image was used for the test building and scales were built up to a distance ratio of 7. The largest scale of descriptor is therefore built from a 112x112 pixel region about each interest point in the template image. Figure 7 gives an example of one of the test images used and shows the interest points that were matched.

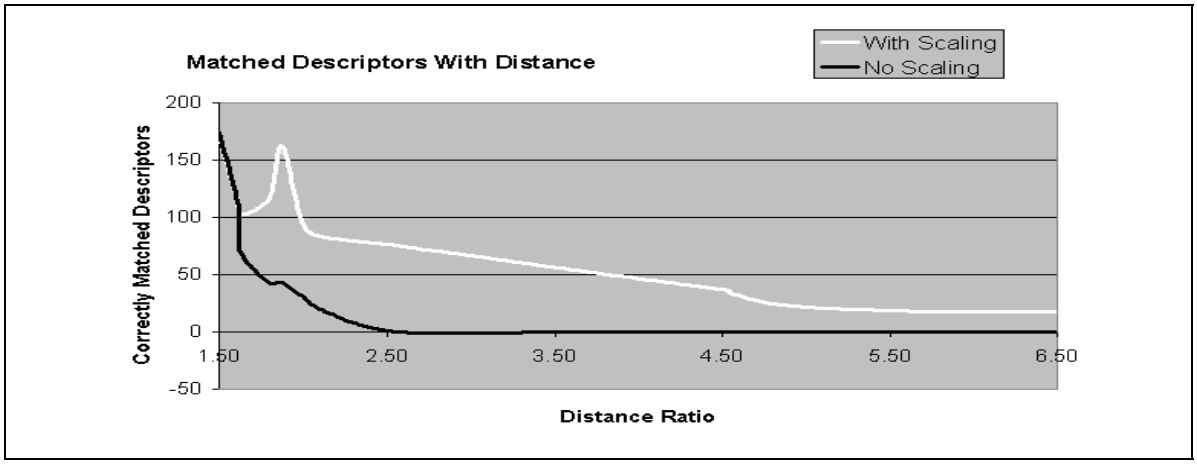

Figure 6: The graph shows the effectiveness of the scaling process. The scaled descriptors to match against for the "With Scaling" plot are calculated automatically from the GPS positions.

Figure 6 shows how without any scaling, the matching process completely fails at a ratio of 2.5 (a query at 50 metres from the building), where the system is unable to match any descriptors between the images. However, with scaling, the building can be matched at 130 metres and a distance ratio of 6.5. At this distance 18 points were matched between the two images. The peak that occurs with distance ratio 2 in the "With Scaling" plot occurs due to the non-continuous creation of scales. As the query image tends towards one of the created scales the scaled descriptors are more alike the query image descriptors, hence more matches are found.

Figure 7: Scaled descriptors were used and with a distance ratio of 3.67 the query image b) has been matched against the database image a). The white lines illustrate 3 interest points that have been matched between the two images.

5 Viewpoints and removal of outliers

By assuming that the front of a building is essentially a planar surface it becomes possible to use a homography (see [15] for an overview). For any two views of the same plane there is a 3x3 matrix known as a homography *H* that maps pixels in the first image to pixels in the second image, Figure 8. Where p is a pixel in image C, q is a pixel in image D and *H* is a 3x3 matrix, if C and D are both photographs of the same building then: p = *H*q.

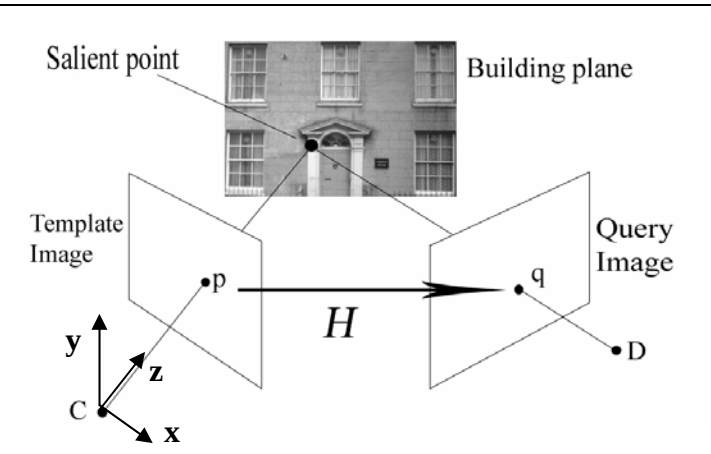

Figure 8: Two views of the same building will be related directly by a homography *H*. If *H* is known then the pixel q can be found from the pixel p.

5.1 Changing viewpoints and planar rectification

The homography *H* that maps pixels in the Template image to pixels in the Query in Figure 8 can be calculated using the translation and rotation of the camera D with respect to C:

$$H = K(R - \frac{tn^T}{d})K^{-1} \qquad (2)$$

Where:

R = rotation matrix of D, *n* = plane normal, *t* = translation of D

K = camera calibration matrix, *d* = distance to plane of C

In equation (2) the camera C is set as the world origin and C is aligned with the world axis, so the view direction of C is pointing down the z-axis of the world. This co-ordinate frame will be referred to as Camera Space. By calculating different values for the rotation matrix R and translation T of the second camera D it becomes possible to move camera D around in Camera Space and thus determine how the building would appear from different viewpoints. To rectify an image of a building in this way, the building plane is assumed to be perpendicular to the viewing direction of C. The plane that defines the building therefore has a normal *n* of (0,0,-1) with respect to the view C.

To create the extra views the second camera is translated and rotated for a range of viewing angles where the distance to the plane is constant and the camera is always focused on the centre of the original image. Figure 9 shows two extra views that have been formed by rotating the view of the building by a given angle.

Since the position of the camera C and its viewing direction in terms of the real world is known, the Camera Space that has been used to form the rectified views can be translated into real world space. Now the real positions and viewing angles of where these rectified views are "taken" from in world space can be calculated. In effect these views represent how the building would appear in a photograph from a defined position in the real world. The extra views for a building are formed automatically when the original view is added to the database.

Figure 9: a) The original view. b) Change in viewing angle of 15 degrees. c) Change in viewing angle of 30 degrees. These views are created automatically when a building is added to the database.

Once added to the database one building may now have several views, with the position of each view defined in world space. These views can either be rectified planes as above or potentially extra photographs that have been taken directly. When matching against a query image one of these views will be the most likely match for the query. The viewing vector is used to determine which of these views to use for the matching process as show in Figure 10.

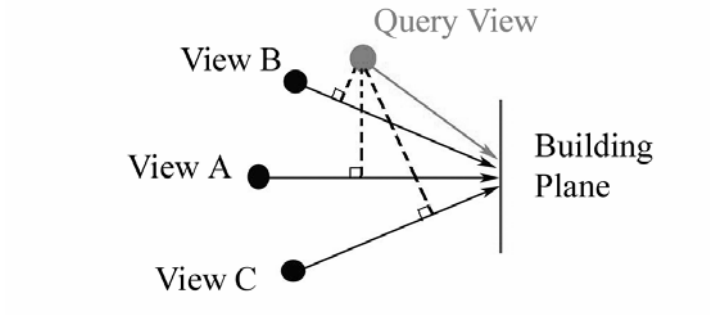

Figure 10: The views B and C have been created by planar rectification and given their respective positions in world space. The dashed lines are used to find the distance of the query view to a particular view in the database. In this case View B will be used to match against the query view since it is "closest."

5.2 Use of RANSAC to estimate the homography

In the same way as a homography can be used to map pixels from one image of a building to another view, a homography also maps the positions of interest points between the two views. For each candidate template image a homography *H* is estimated and the number of interest points that lie within the homography between the query and template is found. If a number of points lie in a given homography then the candidate template may be a match

and should be analysed further. To estimate *H* between two images Random Sample And Consensus [16] or RANSAC is used.

Once complete, the homography that covers the most inliers is found and the matches that are covered by this homography are used for further analysis. Two cases apply when using homographies to match a query image against a template image: if the template image is an incorrect match or if in-fact a correct match has been found.

5.3 A correct match and removal of outliers

If the database image is the correct building, then finding the homography between the query image and database image can remove outliers and mismatches. Any matches that are covered by the homography are now considered as correct matches, a process shown in Figure 11 and 12.

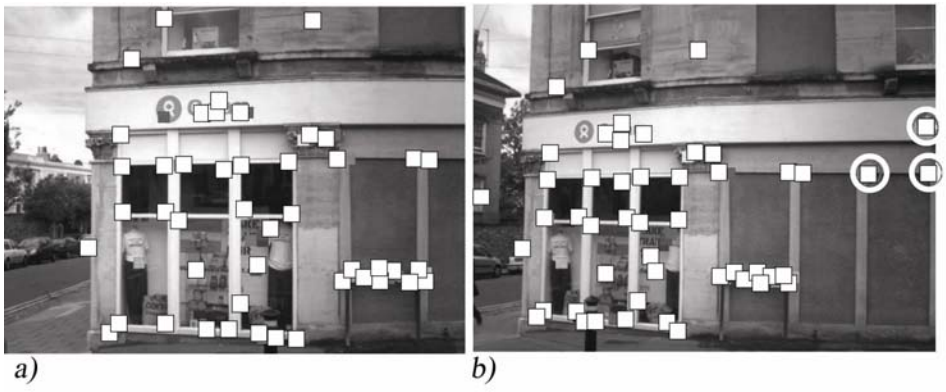

Figure 11: Before applying RANSAC, 52 matches are found between the two images a) and b) as shown by the white squares. Three definite outliers or mismatches are circled in the query image b).

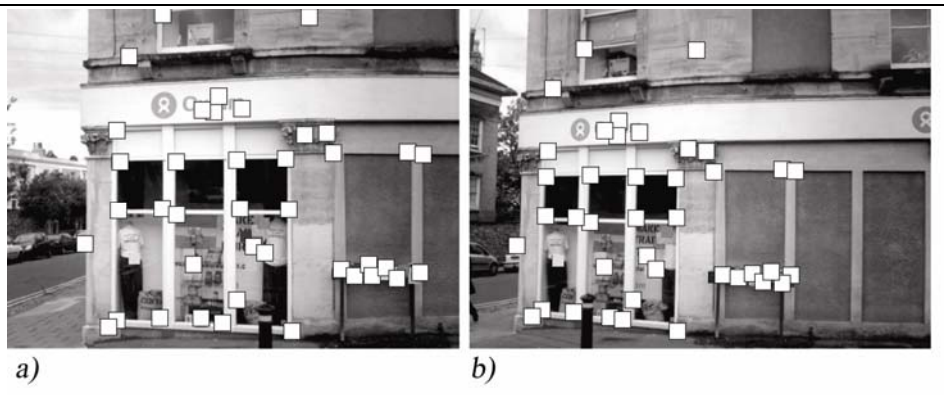

Figure 12: After four iterations of RANSAC the number of matches is reduced to 44 and all outliers have been removed with each match in image a) corresponding to the same point in image b).

All the matches shown in Figure 12 are covered by one single homography and hence are considered correct. This assumption however, becomes problematic when it is not the same building in both images.

5.4 Using RANSAC to identify an incorrect match

If a database image does not contain the correct building then it is unlikely that any of the mismatched points between the two images will form a homography. For this to occur the descriptors in both images must lie on a plane. The use of RANSAC can therefore be used to discount a large proportion of the mismatched descriptors and incorrect building matches. However, there are cases when purely by chance, a number of the mismatched descriptors are covered by a homography as shown in Figure 13.

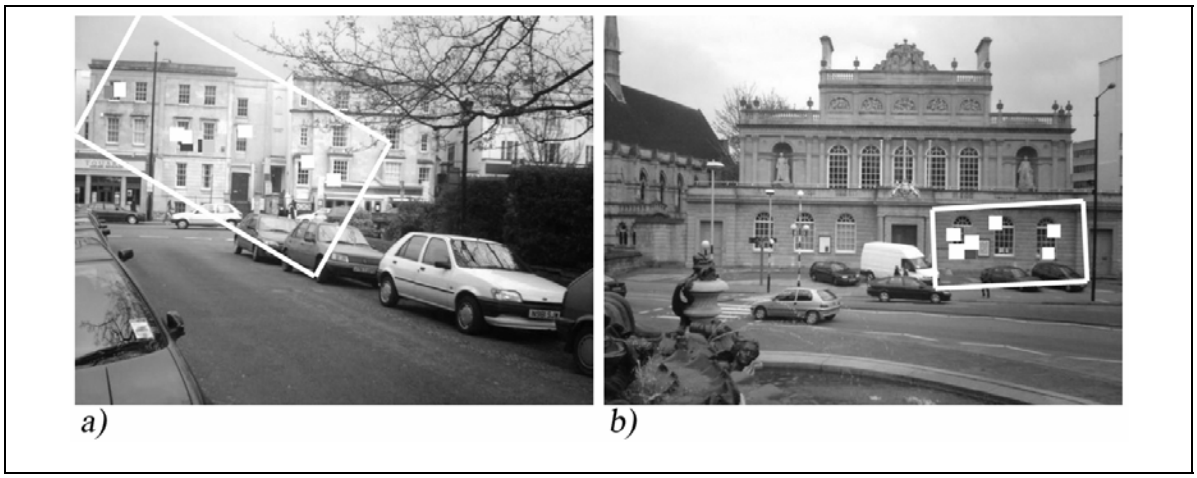

Figure 13. The points in images a) and b) have been incorrectly matched due to the similar structure in the windows. The matched points have been found to lie in a homography with a mapping H existing between the two planes marked in white. This error can be detected as explained in the following section.

In the case of Figure 13, there were 62 mismatches to start with, 53 of them were discounted by RANSAC. But a homography was found mapping 6 points on the plane in the template image to the 6 points on the plane in the query image. Figure 13 illustrates how the matching descriptors returned after RANSAC need to be analysed further to ensure that they do indeed represent a correct building match. A simplistic approach would be to decide incorrect building matches by a threshold on the number of matched points, however, this would require threshold tuning and a better approach was developed as explained in the next section.

6 Structural analysis of matches

At this stage, matches have been found between the query image and the template image that are covered by a homography. The structure of the matches in the template image needs to be compared to the structure of the matches in the query image to ensure that a correct match has been found. A constraint is added to the system where the query images must be a landscape orientation. With this constraint in place it becomes possible to analyse the matches in terms of the vectors between the descriptors in each image, as illustrated in Figure 14.

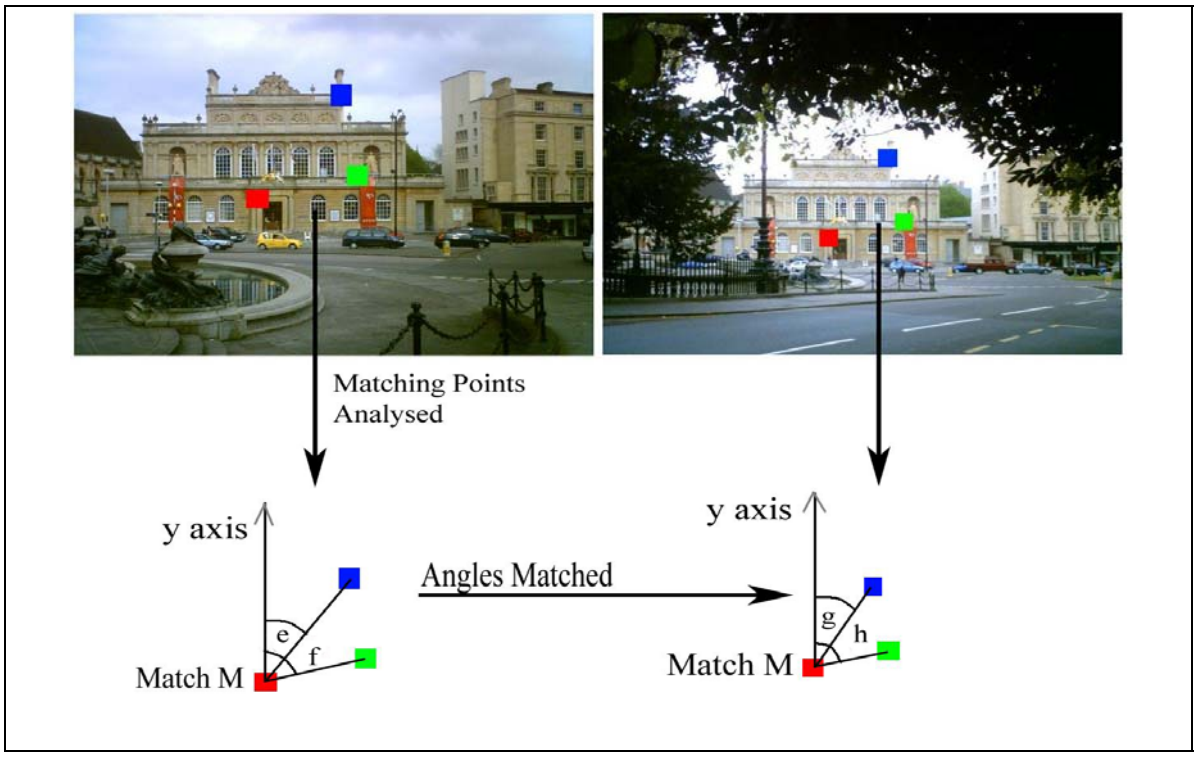

Figure 14: Three points are found to match between two images. If this in fact a correct building match then the arrangement of the points should be similar. To analyse this similarity the angle e is compared to g and the angle f to h.

The algorithm that is used to analyse the structural composition is given here:

REPEAT 10 TIMES:

Choose a feature match M at random.

FOR ALL MATCHES WHERE MATCH P != MATCH M

i) Find the angle Ø between the y-axis and the vector that joins M to P. Ø is found in both query image and database image for comparison

ii) Calculate the error E between Ø in database image and Ø in query image.

iii) Accumulate the errors of E.

Find the average error for E and test against a threshold. **END**

If the two images are of the same building then the spatial arrangement of the matches in both images should be alike. By comparing the angle e to g and the angle f against h in Figure 14, a measure of the resemblance in structure of the matches is obtained as an angle error. With this structural test the system would now discard the template image shown in Figure 13 as an incorrect match for this query.

For a false positive to occur a number of matches need to both lie in a homography and also pass the test for structural resemblance. The use of these two tests in this way greatly reduces the chances of falsely matching a building.

6.1 Determining the final match

If a building is the focus of a photograph then generally the building will be central to the photograph. For any candidate buildings that pass all structural tests, the average distance between the centre of the query image and the position of the matched descriptors for this particular building is found. A scoring system is formed using this distance and the number of matches found for a building and subject of the photograph is determined.

There are many correct interest point matches shown in Figure 15, but the correct match in terms of the focus of the photograph is the match shown in Figure 16. The system determines that this is the case by analysing the average distance of a match to the centre of the query image to determine the focus of the photograph.

a) Query image b) Incorrect database match

Figure 15: The focus for the query image a) is the "Oxfam" building, but 72 points have been matched against the database template "Dancewell" in image b).

a) Query image c) Correct database match

Figure 16: The correct match has been found through spatial weighting since the matches in the query **a)** are more centrally located than the matches shown in Figure 15 a)

7 Results

To test the system, a range of different types of about 30 buildings from the city of Bristol, UK, were included in the database. Query photographs were then taken at different times of the day as well as with significant clutter, distance from building and occlusions. In general, the buildings with the most individual structure have the most matched points. When testing a query image for these results, the image was tested against all buildings in the database, so each query was matched correctly when tested against the 30 different facades. Buildings have been matched despite significant changes in scale, lighting and angle of view. All matching interest points are shown by coloured squares and some with lines connecting a few features to verify that the correct building corners have been matched between the two views. Results are shown in the following Figures.

In Figures 17 and 18, it is shown how query images are matched with rectified views of the facade stored when the database was created. Figures 19 to 24 show further examples of different scales, time of the day and occlusions that the system is able to handle.

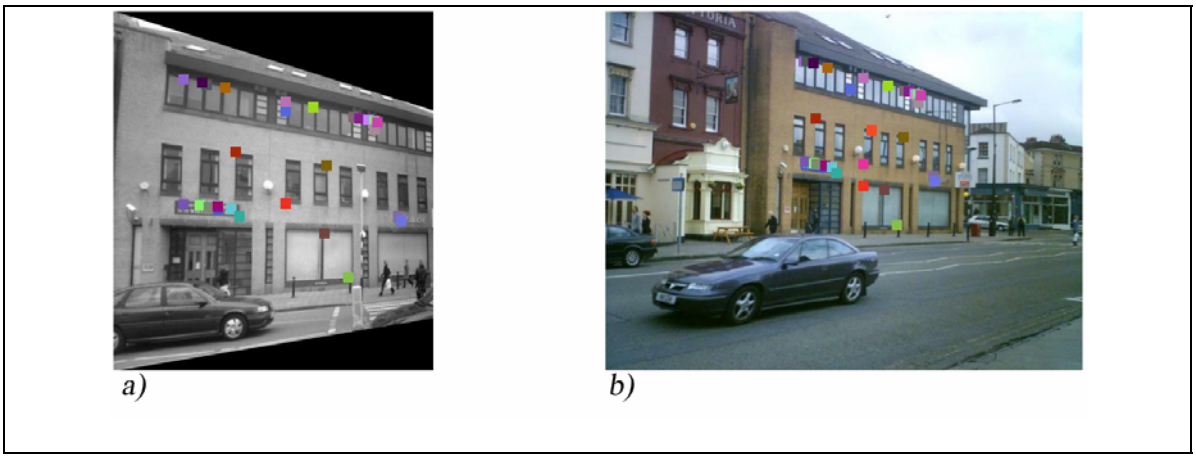

Figure 17: The query image b) has been matched against a rectified view a) of the building. The rectified view was selected automatically by the GPS position of the query and 30 features have been matched

Figure 18: It can be seen by comparing the views that the images have been taken at very different times of day. The database image is a rectified view and the lettering in the shop signs has matched many points.

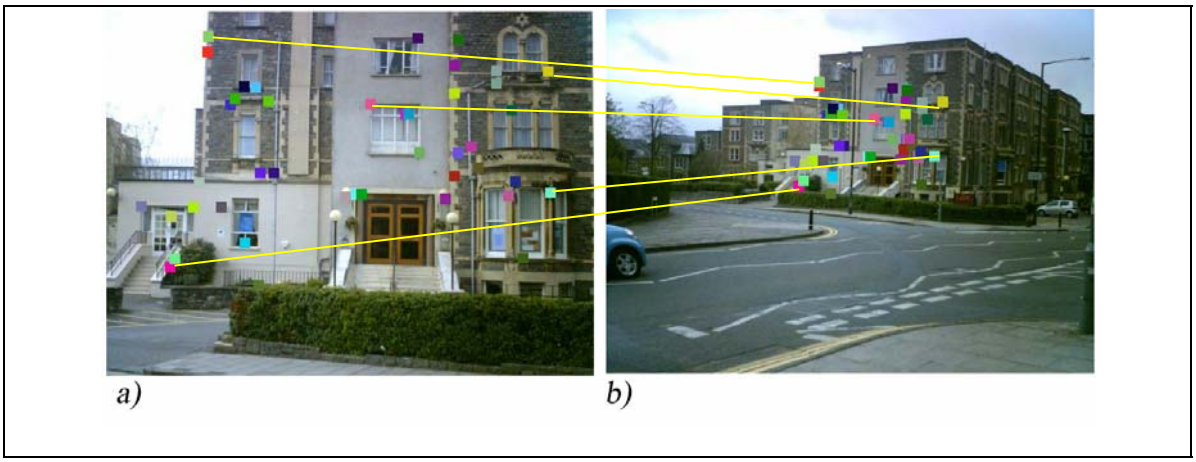

Figure 19: From a large change in angle of view as well as scaling the system has matched 48 features between images a) and b).

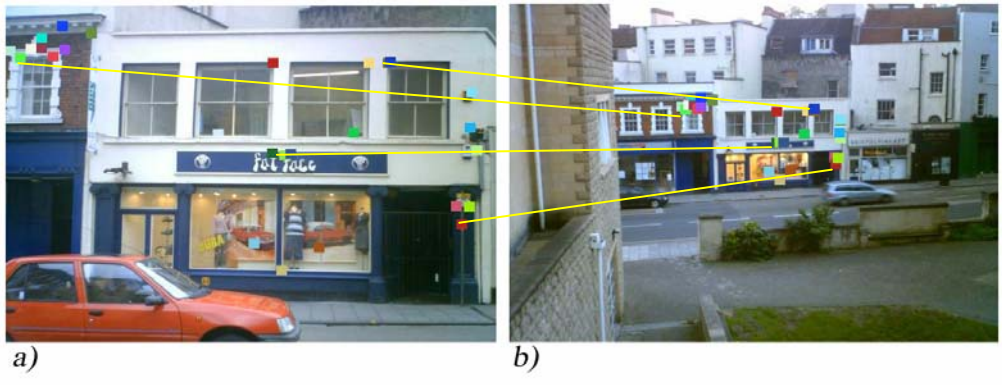

Figure 20: Query image b) was taken from a higher elevation, as well as at distance. The time and light of the day is different and 32 features have been correctly matched.

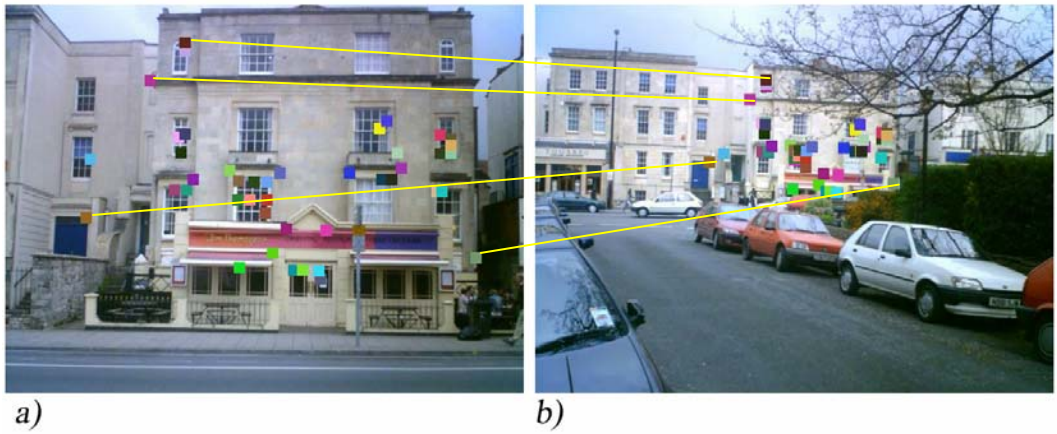

Figure 21: The query image b) shows significant distance to facade and occlusion and has matched 52 features with a).

Figure 22: The building in the query image is only a small fraction of the whole image. So with significant scaling and angle change, 25 interest points have been successfully matched.

Figure 23: The images have been taken at different times of the day. Also, the database building is relatively simple without much structure in its facade. The building has been identified out of the 30 other buildings by matching 18 points.

Figure 24: 124 points have been matched under viewpoint, distance change and large occlusion.

7.1 Errors in the system

The largest source of error occurs with very large changes in viewing direction of a given building. By using the electronic compass[2] with the IPAQ the matching process was evaluated using query images with increasing angles of viewing direction of a given building. This test was carried out both with the rectified views and also with only the one view of the front of the building. When evaluating the synthetic views, two rectified images were used, at an angle of 15 degrees and an angle of 30 degrees. Which of these views to use was calculated automatically by the GPS position of the query images and the results are shown in Figure 25.

[2] Note that the compass was not part of the setup used in the results presented above.

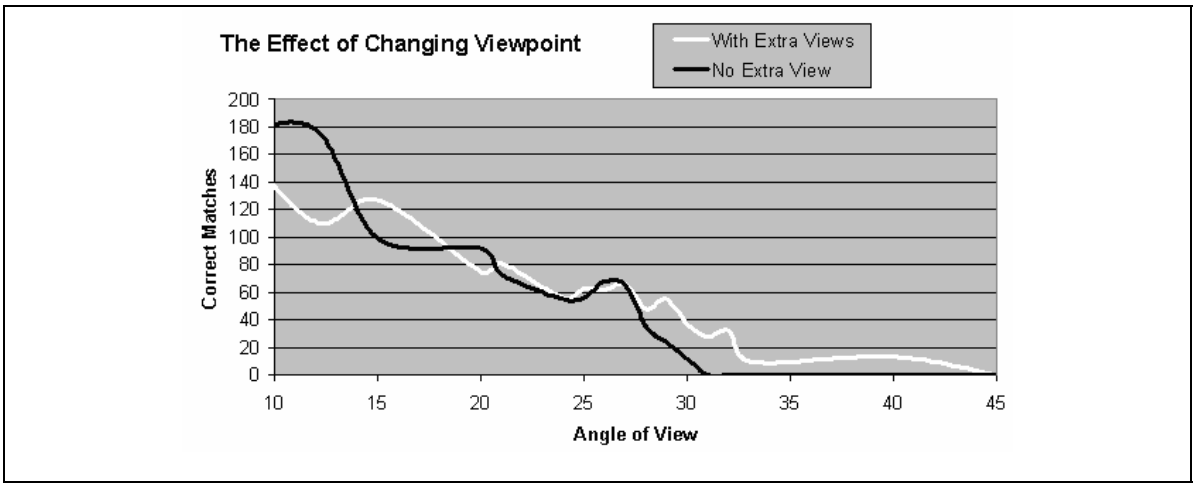

Figure 25: Comparing the effectiveness of the system with and without automatically generated extra views.

Figure 25 shows how without the extra synthetic views, the building cannot be matched passed angles of 30 degrees. The fact that 70 points can still be matched at viewing angles of 25 degrees illustrates the robustness of the SIFT descriptor in coping with changes in viewpoints. With the use of the rectified views the building can be matched up to a 40-degree change in viewing direction.

Examples of when the system fails are shown in Figure 26 and 27.

Figure 26: The building has not been matched due a change in view of over 40 degrees as well as severe occlusion occurring in the query image.

Figure 27: The change in viewing angle meant the system was unable to match the query shown in b) to a)

8 Conclusions

A system that can accurately match buildings between a range of views and distances has been proposed. By using SIFT descriptors [11] and Harris-Stephens corner detector [5], interest points are located and represented from images of buildings. The interest points of a query are then accurately matched against the interest points of database images.

A novel solution for the problems that occur with scaling has been implemented using the GPS position. The GPS is used to simplify the search through scale space when matching descriptors and the tests show encouraging results under real and varied conditions. Additionally, the use of GPS, helps to delimitate the subset of the building's database significantly reducing computational cost.

A system of planar rectification has been developed which accounts for perspective changes that occur with a large change in viewpoint. These extra views are created automatically when an image is loaded, and the most likely match for being selected automatically by the query's GPS position. The buildings in the database can be represented by more than one plane, allowing for more complex structures to be represented and there can be any number of views of each of these planes in the database.

The system has been tested for a wide range of different buildings, conditions and viewpoints. In nearly all cases the system can correctly identify the query building, except in extreme changes of viewpoints or lighting. This technology could then form the basis of a tourist guide that helps the user to learn about the buildings and city around them.

Acknowledgements

To HPLabs Bristol and the Mobile Bristol Project for facilitating the development of this work.

References

[1] Cheverst K, Davies N, Mitchell K, Friday A, Efstratiou C (2000) Developing a Context-aware Electronic Tourist Guide: Some Issues and Experiences. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, The Hague, pp 17 – 24

[2] Pospischil G, Umlauft M, Michlmayr E (2002) Designing LoL@, a Mobile Tourist Guide for UMTS. In: Proceedings Mobile Human-Computer Interaction, Berlin, Heidelberg, 2002, pp 140–154

[3] Kadir T, Brady M (2001) Saliency, Scale and Image Description. International Journal of Computer Vision, Volume 45, Issue 2, pp 83 – 105

[4] Mikolajczyk K, Schmid C (2004) A performance evaluation of local descriptors. Technical Report, accepted to PAMI, October 6, 2004

[5] Harris C, Stephens M (1988) A Combined Corner and Edge Detector. In: Proceedings of the fourth Alvey Vision Conference, 1988, pp 153-158

[6] Schmid C, Mohr R, Bauckhage C (1998) Comparing and evaluating interest points. ICCV, pp 230-235, 1998

[7] Lazebnik S, Schmid C, Ponce J (2003) A Sparse Texture Representation Using Affine-Invariant Regions. In: Proc. CVPR, Volume 2, pp 319-324, 2003

[8] Bauer J, Bischof H, Klaus A, Karner K (2004) Robust and Fully Automated Image Registration Using Invariant Features. In: Proceedings International Society for Photogrammetry and Remote Sensing, Istanbul, Turkey, 2004

[9] Tuytelaars T, Gool L V (2004) Matching Widely Separated Views Based on Affine Invariant Regions. International Journal of Computer Vision, Volume 59, Issue 1 (August 2004), pp 61 – 85

[10] Mikolajczyk K, Schmid C (2004) Scale and Affine Invariant Interest Point Detectors. International Journal of Computer Vision 60(1), 63–86

[11] Lowe D (2004) Distinctive image features from scale-invariant keypoints. Int. J. of Computer Vision, 60(2):91--110, 2004. ACM

[12] Sivic J, Zisserman A (2003) Video Google: A Text Retrieval Approach to Object Matching in Videos. In: Proceedings of the International Conference on Computer Vision, 2003

[13] Se S, Lowe D, Little J (2002) Mobile Robot Localization and Mapping with Uncertainty using Scale-Invariant Visual Landmarks. The International Journal of Robotics Research, Vol. 21, No. 8, 735-758, 2002

[14] Mikolajczyk K, Schmid C (2001) Indexing based on scale invariant interest points. In: 8th International Conference on Computer Vision, pp 525-- 531, 2001

[15] Hartley R, Zisserman A (2003) Multiple View Geometry in Computer Vision. Cambridge University Press, Cambridge

[16] Fischler M A, Bolles R C (1981) Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography, In: Communications of the ACM 26, pp 381-395, 1981

[17] Robertson D and Cipolla R (2004) An Image-Based System for Urban Navigation. British Machine Vision Conference (BMVC), United Kingdom, 2004

[18] Zhang W and Košecká J (2005) Localization Based on Building Recognition. Workshop on Applications for Visually Impaired, IEEE Computer Vision and Pattern Recognition (CVPR), 2005