Viewpoint Invariant Features from Single Images Using 3D Geometry

Yanpeng Cao and John McDonald
Department of Computer Science
National University of Ireland, Maynooth, Ireland
{y.cao,johnmcd}@cs.nuim.ie

Abstract

In this paper we present a novel approach for generating viewpoint invariant features from single images and demonstrate their application for robust matching over widely separated views. The key idea consists of retrieving building structure from single images and then utlising the recovered 3D geometry to improve the performances of feature extraction and matching. Urban environments usually contain many structured regularities, so that the images of those environments contain straight parallel lines and vanishing points, which can be efficiently exploited for 3D reconstruction. We present an effective scheme to recover 3D planar surfaces using the extracted line segments and their associated vanishing points. The viewpoint invariant features are then computed on the normalized front-parallel views of the obtained 3D planes. The advantages of the proposed approach include: (1) the new feature is very robust against perspective distortions and viewpoint changes due to its consideration of 3D geometry; (2) the features are completely computed from single images and do not need information from additional devices (e.g. stereo cameras, or active ranging devices). Experiments are carried out to demonstrate the proposed scheme ability to effectively handle very difficult wide baseline matching tasks in the presence of repetitive building structures and significant viewpoint changes.

1. Introduction

Robust feature extraction is an essential functionality in many computer vision applications, such as Structure from Motion (SFM), pose estimation, and object recognition. Previously a number of successful techniques [2, 10, 4] have been proposed - a comprehensive review was given in [12]. The underlying principle is to normalize the extracted regions of interest to achieve invariance against the changes of illumination, scale, rotation and viewpoints. However, these methods only consider the 2D image texture and do not take advantage of important cues related to the scene's 3D geometry. These methods cannot produce reliable results of feature extraction and matching in the presence of repetitive structures and significant viewpoint changes. In this contribution we combine recent advances in 2D interest point detection and description with 3D viewpoint normalization to improve the descriptive ability of local features.

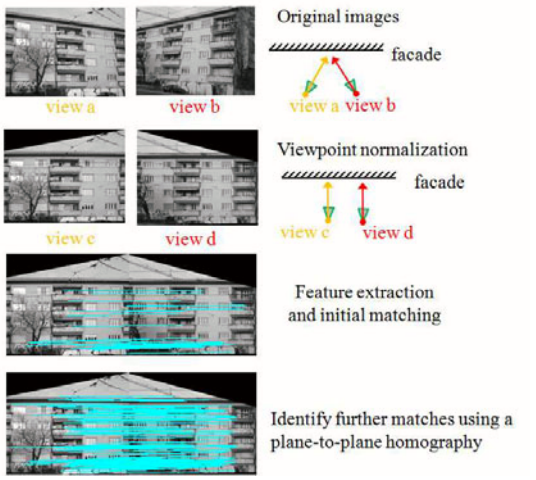


Figure 1. A demonstration of the extraction and matching of viewpoint invariant features from single images

Images taken in man-made environments usually possess many parallel lines which intersect in vanishing points. We propose an effective method to recover a number of 3D planes from single 2D images of urban environments by indentifying line segments within the image and then using them to represent the spatial layout of the environment. We take into account both the distribution of line segments and the possible structure of buildings to obtain reasonable 3D understanding of the scene. The viewpoint invariant features are then computed upon the normalized front-parallel views of 3D planes. The novel feature achieves better distinctiveness and robustness by considering 3D scene geometry. Features on the same 3D plane are related through a 3×3 homography which can further improve the results of feature matching. The key idea of the proposed method

978-1-4244-5498-3/09/$25.00 ©2009 IEEE

is schematically illustrated in Fig. 1. Compared with some previous efforts on combining 2D feature with 3D geometry [15, 7], our method requires only a single image, does not need information from additional devices and thus it offers wider applicability.

The rest of the paper is organized as follow. Section 2 reviews some existing approaches for feature extraction and 3D reconstruction. Our proposed method, which includes, line grouping, planar 3D reconstruction, viewpoint normalization, and feature extraction and matching, is explained in Section 3. The performance of the proposed method is evaluated and compared to the existing state of the art technique, SIFT [10], in Section 4. Finally, the conclusion is given in Section 5.

2. Related works

A large number of papers have been reported on robust 2D image feature extraction. For a detailed review, see [12]. Among them the SIFT scheme (Scale-invariant feature transform) [10] is the most widely used due to its superior invariance to changes in illumination, viewpoint, scale and orientation. The potential keypoints are firstly identified by searching for local extrema in a series of Difference-of-Gaussian (DOG) images. Next the local image patches at these location are normalized to achieve invariance up to a 2D similarity. Finally, a 128-element SIFT descriptor is computed based upon image gradients to characterize the local patch which can then be used in subsequent feature matching. In [11] the authors conducted a comprehensive evaluation of various feature descriptors and concluded that the SIFT descriptor outperforms other schemes. SIFT has been successfully applied for various computer vision tasks such as object recognition, 3D modeling, and pose estimation. However, the SIFT scheme cannot to produce satisfactory feature matching over widely separated views because perspective effect will add severe distortions to the resulting descriptors. Also it is difficult to identify a unique match for a feature in the presence of repetitive building structures.

Recently, many researchers considered the use of 3D geometry as an additional cue to improve 2D feature detection. A novel feature detection scheme, Viewpoint Invariant Patches (VIP), based on 3D normalized patches was proposed for 3D model matching and querying [15]. In [7], both texture and depth were exploited for computing a normal view onto the surface. In this way they kept the descriptiveness of similarity invariant features (e.g. SIFT) while achieving extra invariance against perspective distortion. In [16], 3D gradients and histograms were considered to generate 3D features which are invariant to changes in rotation, translation, and scale. However, in these methods 3D scene geometry was acquired using either multiple views (using SFM or stereo vision) or additional active sensors (Lidar or Radar). The goal of this paper is to retrieve the 3D spatial layout of buildings from single images and then use this information to generate viewpoint invariant features.

Previously a number of techniques have been developed for 3D understanding using monocular cues. Hoiem and his research group estimated the coarse geometric properties of a scene by learning appearance-based models of geometric classes, and then used the underlying 3D geometry to improve the performance of computer vision applications such as object detection and single view reconstruction [3, 6]. In [13], a supervised learning approach was proposed for 3D depth estimation based on Markov Random Fields. Usually architectural scenes are highly constrained, thus their images contains many regular structures including parallel lines, sharp corners, and rectangular planes. The presence of such structures suggests opportunities for constraining and therefore simplifying the reconstruction task. A number of techniques [8, 9] were proposed for detecting rectangles aligned with major directions using the recovered vanishing points. Such structures provide strong indications of the existence of co-planar 3D points. In [1], a visually pleasing urban 3D model is generated by solving the problem of model fitting. Assuming the environment is composed of a flat ground plane with vertical walls, they used a continuous polyline to parameterize ground-vertical boundary. The success of the above approaches inspired us to extend the conventional 2D image feature to the third dimension using the obtained 3D geometry from single images.

3. Our approach

Full 3D reconstruction from single images is a difficult task since the depth information is ambiguous given only local image features. In this work we propose to perform a partial 3D reconstruction in which the spatial layout of buildings is represented using a number of planes in 3D space. First, line segments are extracted and then grouped by identifying their associated vanishing points. In this step we include an important tilt rectification procedure to make the building structure more obvious. Then we use the line segments from different directions and the possible shape of building structure to obtain a planar 3D reconstruction. As the last step, the individual patches in the original image, each corresponding to an indentified planar region, are rectified to form front-parallel views of building facades. Viewpoint invariant features are then extracted from these rectified views to provide a basis for further matching. The details of each step are explained in the following sections.

3.1. Line grouping

Given an image taken in an urban environment, we apply the method described in [8] to extract a number of straight lines. Then we identify parallel world lines by grouping them into different principal directions. An effective approach is proposed to this problem by combining the elemetns of the method in [1] with the method in [8]. The approach includes the following steps:

(1) Select the line segments corresponding to the vertical vanishing direction ($\pm\pi/6$ vertical lines in the image) and compute a 3×3 matrix to transform them to appear vertical in the image. Apply this transformation to the original image to create a rectified view where the tilt effect is removed (i.e. vertical world lines become parallel to the image columns). In the rectified image, building boundaries will appear vertical, making the building structure more evident (see Fig. 2 (b)).

(2) Group the non-vertical lines into dominant vanishing directions by identifying their associated vanishing points. This is a difficult problem since any two lines will intersect and give a possible vanishing point. We proposed to restrict the search of possible vanishing points from the whole image to a horizontal strip by exploiting the fact that the "horizon" (the line connecting horizontal vanishing points) will appear horizontal in the rectified image. We apply the RANSAC algorithm to compute a vanishing point with the maximum line segment supports and take its y-coordinate as the "horizon" level. Next the x-coordinates of intersects between the "horizon" and the remaining line segments are computed and further clustered to generate the initial result of line grouping. Fig. 2 (c) shows an example of such grouping.

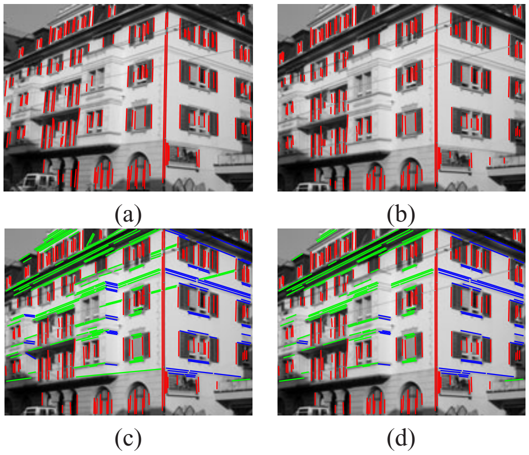


Figure 2. Line extraction and grouping. (a) original image with nearly vertical lines marked. (b) resulting image after tilt rectification. (c) the initial line grouping. (d) grouping result refined using the EM algorithm.

(3) Refine the result of line grouping and vanishing point estimation simultaneously by using the Expectation Maximization algorithm (EM) [8]. Here we iteratively estimate the coordinates of vanishing points as well as the probability of an individual line segment belonging to a particular vanishing point. The final result of line grouping is shown in Fig. 2 (d).

3.2. 3D geometry from line segments

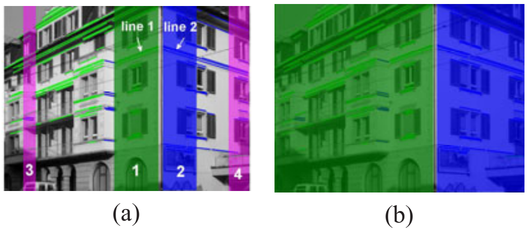


Figure 3. An example of image segmentation based on the distribution of horizontal lines. (a) each line segment gives a vertical strip to support a 3D plane in its corresponding direction. (b) segmentation result after assigning each strip a single direction.

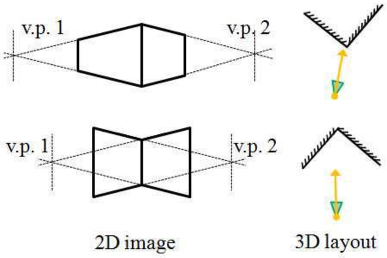


Figure 4. Types of corners and its associated 3D layout.

Next we divide the whole image into several vertical strips with each one corresponding to a different 3D plane (building walls) in the world. It's noted that each line segment from a horizontal vanishing direction (we only consider the horizontal directions since the wall boundaries will appear vertical after the tilt rectification) provides a strong indication of the existence of a 3D plane in that direction. For example, in Fig. 3 (a) the line 1 gives a vertical strip to support a 3D plane in its corresponding direction, and the line 2 suggests another plane in a different direction. Based on the distribution of line segments, the whole image can be divided into many strips. Overall, there are three kinds of strips: **(1)** strip covering lines segments from a single vanishing direction (i.e. strip 1 and 2 in Fig. 3 (a)); **(2)** strip covering line segments from multiple vanishing directions (i.e. strip 3 and 4); **(3)** strip covering no line segments. In our approach we assign to each strip of type **(1)** a direction corresponding to the vanishing point associated with the lines contained within it. All strips of type **(3)** are ignored. For a strip of type **(2)**, we assigned a single direction to it by referring to its neighbors (adjacent strips tend to belong to the same plane) and comparing the numbers of the lines it contains from different directions (the direction with the most line segment supports will be assigned to the strip).

Fig. 3 (b) shows an example of the final segmentation result. We consider the building facade is composed of a number of connected planes in the 3D space. Two walls meeting at one place in 3D space will form a corner. There are two types of corners - convex and concave (see the first column in Fig. 4). The type of a corner provides useful knowledge about the approximate 3D layout of building walls (see the second column in Fig. 4).

3.3. Viewpoint normalization

For better efficiency and accuracy, the initial line segments are refined and further merged by referring to their associated vanishing point. Each line segment contains two end points x_1 and x_2. We connect each end point to its associated vanishing points and then transform the obtained line to a point (ρ,θ) in the Hough space. The end points that fall into the same cell are merged to form a new line through an orthogonal linear least square fitting procedure. The refined line segment is more consistent with the vanishing direction. Also the number of lines is reduced to a small number after the merging.

For each detected plane in the last step, we choose four line segments (two from the vertical direction and two from a horizontal direction) and compute their points of intersection to construct a quadrilateral. We then compute the homography, $H \in \Re^{3\times3}$, which relates the obtained quadrilateral to a rectangle. Without loss of generality we can assume that a corner of the rectangle in the 3D world is denoted by homogeneous coordinates $\mathbf{X} = [h, s.h, 0, 1]$, where h is the height of 3D rectangle and s is the aspect ratio between the width and the height and its corresponding image coordinates are $\mathbf{x} = [x, y, 1]$. Then the mapping between 3D world coordinates and 2D image coordinates satisfy the following relations:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} \sim K\,[R_1\ R_2\ R_3\ t] \begin{bmatrix} h \\ s.h \\ 0 \\ 1 \end{bmatrix} \sim K\,[R_1\ R_2\ t] \begin{bmatrix} h \\ s.h \\ 1 \end{bmatrix}$$

$$\sim K\,[R_1\ R_2\ t]\,diag(1, s, 1) \begin{bmatrix} h \\ h \\ 1 \end{bmatrix} \quad (1)$$

Denote H_s which maps the quadrilateral patch to a unit square (since four corners of the quadrilateral is known, H_s can be solved in a close form) and substitute it into Eq. 1 to obtain:

$$H_s = K\,[R_1\ s.R_2\ t] \quad (2)$$

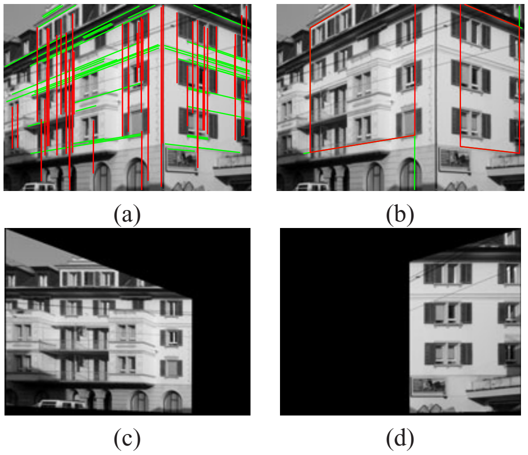
It's noted that R_1 and R_2 are columns of a rotation matrix and should have unit normal. Thus the aspect ratio s can be recovered as follows:

$$s = \frac{\|h_2\|}{\|h_1\|} \quad (3)$$

where h_1 and h_2 are the first and second columns of matrix $K^{-1}H_s$. Once the aspect ratio s is recovered, the homography H is then computed as:

$$H = H_s diag(1, 1/s, 1) \quad (4)$$

The recovered homography H enables us to warp the original image to a normalized front-parallel view where perspective distortion is removed. Fig. 5 shows an example of such viewpoint normalization. It's noted that a rectangular window in the 3D world will also appear rectangular in the normalized image.


(a) (b)

(c) (d)

Figure 5. An example of viewpoint normalization. (a) the initial line segments are refined and merged. (b) four segments from two different directions construct a quadrilateral. (c)-(d) the normalized front views of seperate building facades.

3.4. Viewpoint invariant feature extraction and matching

Within the warped front-parallel views of building facades, where perspective effects have been removed, viewpoint invariant features are computed in a same manner to the SIFT scheme [10]. The proposed method is similar to [15] in spirit, instead we use single images for both viewpoint normalization and feature extraction. Moreover, the detected 3D planes allow us to identify sets of co-planar features which can be related through a 3×3 homography. Considering this point-to-point mapping relation allows us to restrict the correspondence search in the second image so that it's easier to identify a distinctive match there. It leads to significant improvement for feature matching in the presence of repetitive building structures. The matching procedure contains the following steps:

(1) Obtain an initial set of feature correspondences. A pair of features is considered as a correspondence if the distance ratio between the closest match and the second closest

one is below a predefined threshold. The recovered building 3D layout is also considered to improve performance. For instance, features on a plane in the first image will only be matched to features on another single plane in the second image.

(2) Identify inliers consistent with a homography from all the correspondences using the RANSAC technique [5].

(3) Find more feature correspondences using the estimate homography to define a small search region about the transferred feature position.

4. Experiments

In this section we test the performance of the proposed viewpoint invariant features and subsequently apply them for wide baseline matching. We used the benchmark testing image dataset - ZuBuD [14]. The dataset consists of multiple images of 201 buildings taken in the Zurich city centre. For each building 5 images were acquired at significantly different viewpoints, in different seasons, weather and illumination conditions. To demonstrate the improvement of the proposed method we compared it with the most widely used feature extraction technique, SIFT [10].

4.1. Descriptiveness evaluation

After the viewpoint normalization, corresponding scene elements will have more similar appearance, so that the resulting features suffer less from perspective distortions and shows better descriptiveness. In this section, we test how well two features of a correspondence can match with each other in terms of the Euclidean distance between their corresponding descriptors. We followed the method described in [12] to find the ground truth for correct matches. The detected features in the first image are projected onto the second one using the homography relating the images (we manually select 4 well-conditioned correspondences to calculate the homography). A pair of features are considered matched if the overlap error of their corresponding regions in minimal and less than a threshold [12]. We adjusted the threshold value to change the number of generated feature correspondences and calculated the average Euclidean distance of their descriptors. The result was shown in Fig. 6. Note that our proposed method can generate feature correspondences with higher similarity levels (lower Euclidean distance between descriptors).

In Fig. 7, two pairs of matched features (covering the same scene patch) were found in the original images (the first column in Fig. 7 (a)) and the normalized images (the first column in Fig 7 (b)). Their corresponding descriptors were shown in the second column in Fig. 7 (a) and (b). The orientation histogram values associated with each spatial bin were depicted by lines of different lengths for each one of the 8 quantized gradient directions. The matched feature extracted on the normalized views have more similar descriptors - the Euclidean distance between descriptors decreased from 0.4854 (on the original images) to 0.3153 (on the normalized front-parallel views). Also, the distance ratio between the best match and the second best match [10] decreased from 0.6751 (on the original images) to 0.553 (on the normalized front-parallel views). The improvements enable more robust matches over widely separated views.

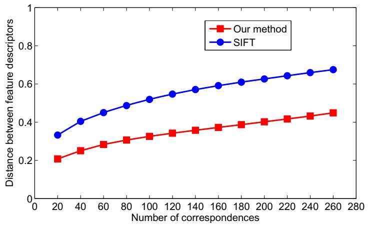
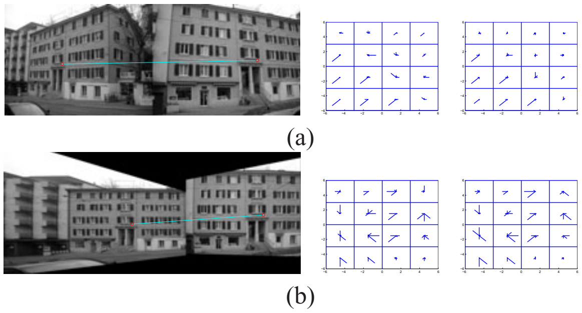


Figure 6. Matched features and the Euclidean distances between their corresponding descriptors.

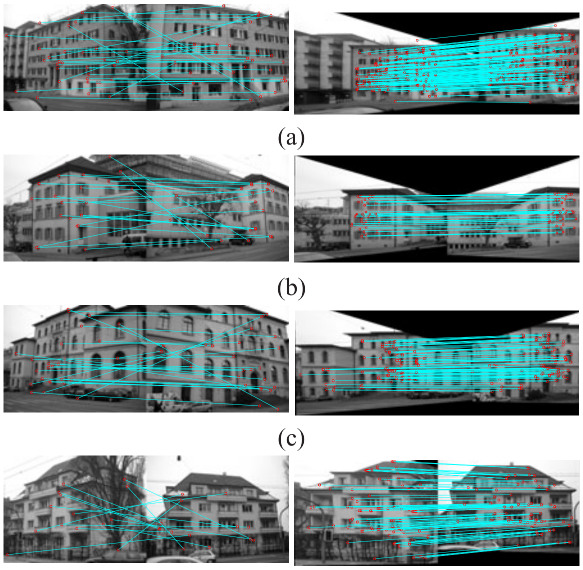

(a)

(b)

Figure 7. Matched features found in the original image and the normalized image, along with the corresponding descriptors.

4.2. Wide baseline matching

Next we demonstrated the advantages of the proposed viewpoint invariant features by applying it to very difficult wide baseline matching tasks. We test the proposed method on the 1st view and the 5th view of a building which have the largest viewpoint change (In many cases the view angles changed more than 90 degrees). We first found a number of putative correspondences based on the distance ratio test (the ratio threshold was set at 0.7), then applied the RANSAC algorithm to estimate a homography and identified inliers. Note that in many image pairs used in our experiments, SIFT cannot generate enough accurate correspondences to compute the correct homography in this step. Finally, we identified more correspondences through the homography-guided matching procedure. The number of putative correspondences and correct ones were counted manually. The results and evaluation were reported in Fig. 8 and Tab. 1. It's noted that the viewpoint invariant feature can handle the large viewpoint changes which SIFT is difficult or even impossible.



(a)

(b)

(c)

(d)

Figure 8. The examples of widebaseline matching (the left result is using SIFT on the original image and the right one is using our proposed method).

Image pairs	SIFT	Initial results (VI)	Final result (VI)
a	9(T)/35(N)	43(T)/61(N)	166(T)/169 (N)
b	1(T)/22(N)	21(T)/39(N)	46(T)/46(N)
c	4(T)/27(N)	27(T)/49(N)	80(T)/84(N)
d	1(T)/14(N)	20(T)/35(N)	94(T)/100(N)

Table 1. The result of wide baseline matching and evaluation (N - number of generated correspondences, T - number of correct ones). Note, the viewpoint invariant features generate enough correct correspondences to compute the homograhpy. This cannot always be achieved using the SIFT features.

5. Summary and conclusions

In this paper we proposed an effective method for extracting and matching viewpoint invariant features from single images. The key idea is to use the 3D geometry as an additional cue to improve the performance of 2D feature. The new features are very robust against perspective distortions and viewpoint changes. Compared with some previous works on combining 2D feature with 3D geometry, our method works completely on single images and hence is more widely applicable. We have demonstrated the use of this novel features in the context of wide baseline matching tasks. However, the proposed method only works well on the images taken in an urban environment, where a number of 3D planes can be used to represent the scene. In the future, we will further extend the method to some more complex and larger scale environments.

6. Acknowledge

Research presented in this paper was funded by a Strategic Research Cluster grant (07/SRC/I1168) by Science Foundation Ireland under the National Development Plan. The authors gratefully acknowledge this support.

References

[1] O. Barinova, V. Konushin, A. Yakubenko, K. Lee, H. Lim, and A. Konushin. Fast automatic single-view 3-D reconstruction of urban scenes. *ECCV*, 2008.

[2] H. Bay, A. Ess, T. Tuytelaars, and L. Van Gool. Speeded-up robust features (SURF). *Comput. Vis. Image Underst.*, 110(3):346–359, 2008.

[3] H. Derek, A. A. Efros, and M. Hebert. Putting objects in perspective. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2006.

[4] M. Donoser and H. Bischof. Efficient maximally stable extremal region (MSER) tracking. *IEEE Conf. on Computer Vision and Pattern Recognition*, 1, 2006.

[5] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, New York, NY, USA, 2003.

[6] D. Hoiem, A. A. Efros, and M. Hebert. Geometric context from a single image. *IEEE International Conf. on Computer Vision*, 2005.

[7] K. Koeser and R. Koch. Perspectively invariant normal features. *IEEE International Conf. on Computer Vision*, 2007.

[8] J. Kosecka and W. Zhang. Video compass. *ECCV*, 2002.

[9] J. Kosecka and W. Zhang. Extraction, matching, and pose recovery based on dominant rectangular structures. *Comput. Vis. Image Underst.*, 100(3):274–293, 2005.

[10] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[11] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(10):1615–1630, 2005.

[12] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. Van Gool. A comparison of affine region detectors. *Int. J. Comput. Vision*, 65(1-2):43–72, 2005.

[13] A. Saxena, S. H. Chung, and A. Y. Ng. 3-D depth reconstruction from a single still image. *Int. J. Comput. Vision*, 76(1):53–69, 2008.

[14] T. S. H. Shao and L. Van Gool. Zubud-zurich buildings database for image based recognition. Technical Report 260, Swiss Federal Institute of Technology, 2004.

[15] C. Wu, B. Clipp, X. Li, J. Frahm, and M. Pollefeys. 3d model matching with viewpoint-invariant patches (VIP). *IEEE Conf. on Computer Vision and Pattern Recognition*, 2008.

[16] A. Zaharescu, E. Boyer, K. Varanasi, and R. P. Horaud. Surface feature detection and description with applications to mesh matching. *IEEE Conf. on Computer Vision and Pattern Recognition*, 2009.