Scene Dedicated Feature Descriptor with Random Forest Training for Better Augmented Reality Registration

Andras Takacs, Edgar A. Rivas-Araiza, and Jesus Carlos Pedraza-Ortega

Universidad Autónoma de Querétaro, Querétaro,
Mexico

Abstract. The most important part of an Augmented Reality system is the tracking system to support an accurate and robust registration. In outdoor environments, the continuously changing environmental characteristics and elements make hard to achieve this tracking process. The main point of this operation is that the descriptor has to work with great accuracy in all kind of situations. The most used descriptors have this distinctive capacity, but computers and mobile devices process them in a long time frame. This paper investigates a new trained, lighter, scene dedicated descriptor, which takes into account the scene characteristics. The descriptor is loaded with elements that can be computed faster and have distinctive information about the selected area. The complete descriptor is used for semantical feature extraction with the aid of a trained Random Forest classifier. For validation purposes, the descriptor was tested against the most used descriptors and in some cases it proved to be faster and equally reliable.

Keywords: Augmented reality, descriptors, random forest classifier

1 Introduction

In recent years, the use of Augmented Reality (AR) has been steadily growing. The stability of the AR applications is continuously improving, but in outdoor environments has lots of flaws due to the rapidly changing environmental factors and the mobile devices still limited storage capacity and processing power. The main challenge is to create a light and robust application for outdoor environments. The use of facade recognition and segmentation with a trained, environment dedicated descriptor is a possible way to stabilize these applications in built-up areas.

1.1 Related Work

In order to support digital data with real scenery it needs to be solved against other two crucial problems; Tracking and Registration. Tracking is the method how AR system specifies its position in the 3D environment and it is crucial

pp. 51–61; rec. 2015-03-28; acc. 2015-07-21

for the stable Registration. For this in outdoor AR different techniques were developed along the years, to deal with the changing light conditions, markerless environment and sparse areas among others. At the beginning of the last decade mainly magnetic sensors were used, such as the gyroscope, GPS, accelerometer or compass [2]. The development in the field of computer and mobile processing units facilitated to exploit more the video see-through AR applications with image processing functions. By the end of the decade even though the magnetic sensors were exploited, by using the gravitational force to get better tracking and registration [15, 16]or tracking the GPS position and cloud server for more stable outdoor tracking [22], the tracking is mainly done through the camera tracking environment features, or artificial markers. Computing frame by frame the spatial positions of this features specifies the systems status. In recent years in order to improve the video see-through AR tracking, the use of geo-tagged panoramic images increases the performance of AR systems [1]. However, those approaches are using a cloud-based processing unit to help the tracking system with a dependency to an internet connection. Also, there is a proposal [13] of a method that uses Random Forest to get a better feature tracking for PTAM [14]. However, they state that their system manages only about 650 scenes and both learning and recognition processes are implemented in online fashion.

According to [12]the automatic facade techniques were a response to the growing need of mass 3D reconstruction and modelling in city planning, geo-applications like "Google Earth" or "Microsoft Virtual Earth" and in 3D GPS navigation systems to reduce the reconstruction time and the storage size of the data. There are various techniques which were developed during the years, and there are differences not only in the technique used to extract features but the source data used also differs. Some researchers use input from terrestrial laser scanner like [18] where the features are detected from the point density, others use a mixed source, they gather the information from the laser scanner and images simultaneously for the reconstruction [4]. More techniques exist for reconstruction from images, by using different approaches to obtain necessary information. Also, there is the "bag of key point" method, which uses a general image categorization technique [10]. This method uses low-level SIFT image features as descriptors assigned to high-level image clusters called "vocabularies" for training a multi-class classifier. In [5] it is presented a technique for image parsing of architectural scenes. This is achieved via segmenting the images into visually recognizable regions (sky, foliage, building and street). Moreover, [11] developed further the technique of [10] using Opponent SIFT as descriptor and Randomized Decision Forest as classifier achieving a faster classifier than its predecessor.

1.2 Contribution

This paper presents the results a comprehensive performance evaluation of a specialized feature descriptor in terms of both computational efficiency and retrieval performance. The main purpose is to show that a specialized feature descriptor can produce better results in terms of performance and can be as efficient as the state of the art feature descriptors. The basic concept is to create lightweight descriptors with elements that specialized to the corresponding environment (buildings, green areas, and sparse areas) after training and empower them with machine learning techniques to be stable in the mentioned areas. The further goal is to create a new modular descriptor where the system can automatically detect the scenery and decide the composition of the descriptor.

1.3 Organization of this paper

Firstly the Random Forest classifier will be presented which will be followed by the state of the art descriptors which were used for the evaluation of dedicated descriptor. The second part of the paper will begin with an overview of the experimental setup, in Section V the results will be presented will finish with the Conclusion in Section VI.

2 Random Forest

The Random Forest [7] is a high-performance discriminative classifier, handling a large set of features without having difficulties due to the curse of dimensionality [11]. It is a supervised learning method that construct an ensemble of recursively created random binary decision trees (Fig.1) during the training period and learn more than one class at a time. In the classification process it returns the most voted class given a feature vector v_i by averaging the final probabilities p_τ of each tree.

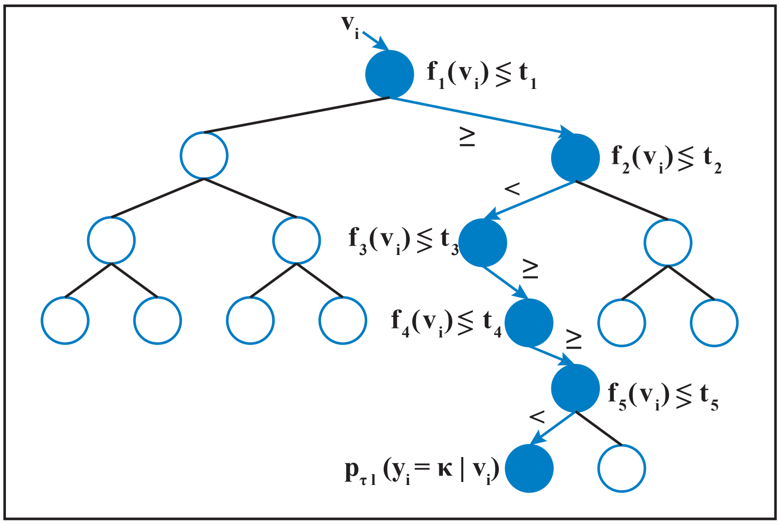

Fig. 1: Binary decision tree.

$$p(y_i = \kappa|v_i) = \frac{1}{T}\sum_{\tau=1}^{T} p_\tau(y_i = \kappa|v_i). \tag{1}$$

The strength of this process, compared to the random decision trees which may suffer from overfitting, that it has been aggregated randomly at two stages during the building of the forest in the training session. First at the Bootstrap Aggregation [7] where random subsets of data are created and from which the trees are learned, and second during the creation of the decision trees at the split functions using only a random fraction of all features.

3 The Descriptors

During the development, the two most used descriptor and their colour versions were tested for their characteristics in order to investigate the speed and accuracy performance of the dedicated descriptor.

3.1 State of the Art Descriptors

Scale Invariant Feature Transform (SIFT) Has been for the past 10 years the most used and referenced descriptor with 128 elements, which consist a set of orientation histograms on 4×4 pixel neighbours over a 16×16 region around the key point. The magnitudes are weighted by a Gaussian function afterward [17].

Speeded-Up Robust Features (SURF) It was built by [3] based on SIFT but, according to the authors it has a better performance. It is smaller in size a 64-dimension vector calculated from a squared region centred on the key point. The region is split into 4×4 subregions. They calculate a Gaussian weighted horizontal and vertical Haar wavelet, which are summed over the sub-regions, and also they calculate the absolute values of the same responses.

Opponent SIFT According to [19] this is the best performing SIFT descriptor on coloured images. It is calculated in the same way as the classical SIFT descriptor but for all the opponent color channel, where the color space contains one intensity and two chromatic channels. That adds up to a 384-dimension vector. These highly decorrelated channels were calculated in the following way.

$$\begin{pmatrix} O_1 \\ O_2 \\ O_3 \end{pmatrix} = \begin{pmatrix} \frac{R-G}{\sqrt{2}} \\ \frac{R+G-2B}{\sqrt{3}} \\ \frac{R+G+B}{\sqrt{3}} \end{pmatrix}. \tag{2}$$

Opponent SURF To retrieve the color information [9] created this descriptor by calculating the original SURF descriptors on the 3 opponent color spaces, which gives a 192-dimensional vector.

3.2 Environment Dedicated Descriptor

The newly proposed descriptor is an 113-dimension vector computed from a 9×9 patch selected around each key point. The size was chosen to be big enough to pick up edges, low-level changes on the image, and still reduce the saving time and size of data to the forest. The elements were chosen with the following characteristics:

Position - *2 values* - 2D image coordinates of the patch centres to separate points which are on the top (sky), on the bottom (street) and in the middle (facade).

Patch Mean - *6 values* - The mean of the Red, Green, and Blue (from the RGB channels) and Saturation values (from the HSV channels) over the patch are calculated to exploit the color changes on the images. Sine and Cosine of the mean of the Hue values over the patch are also estimated. Because the Hue is angular, it has a discontinuity. The red value at 0° is almost the same as the red at 360°. With the Sine and Cosine pair, we can make them equal.

The Third Order Central Moments - *24 values* - The third order central moments were generated to get distinctive shape description of the patch. The $\mu_{03}, \mu_{30}, \mu_{21}, \mu_{12}$ of the RGB and HSV channels over the patch measure the skew and the symmetry of the point spread around the mean of the patch. Firstly the M_{00} raw image moment calculated by

$$M_{ij} = \sum_x \sum_y x^i y^j I(x, y), \tag{3}$$

then the two components of the centroid:

$$\overline{x} = \frac{M_{10}}{M_{00}}, \quad \overline{y} = \frac{M_{01}}{M_{00}}. \tag{4}$$

Then the third order Central moments then defined as

$$\mu_{pq} = \sum_x \sum_y (x - \overline{x})^p (y - \overline{y})^q f(x, y). \tag{5}$$

The higher order moments describe more fine variations in the shape, but they are more sensitive to noise and left out for that reason.

Distance Transform - *81 values* - Distance transform measures the distance between the pixel and the nearest detected canny edge point. The values of the distance transfer are growing as the point is further away from the edge, in this way the values at the flat areas are at their maximum which is a good distinctive component in the descriptor to help the forest to separate the patches in flat areas from the patches from areas where lots of transition are located.

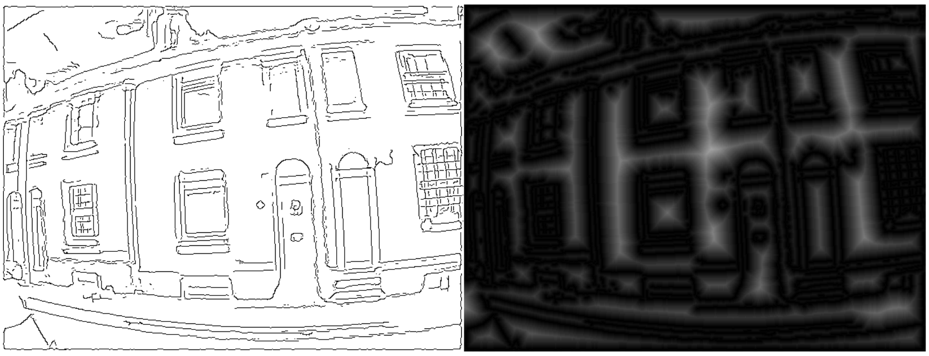

Fig. 2: Distance transform results.

4 Experiments

The steps of the experiment were identical for all the researched descriptors. First the interest points were specified then the descriptor values were calculated. The computed descriptor vectors were loaded to the Random Forest training method. In the final step, the database containing the trained decision trees was used in the classification method where the object features were segmented.

Descriptor Extraction To retrieve all possible information on the image, evenly spaced feature centre points were specified with the same distance to each other. The whole image has been blurred to remove the unnecessary edges and noise, and the Canny edges [8] were calculated for the distance transform image. The resulted images (blurred color and distance transform) was used for to extract the necessary information. The descriptor vectors we created out of the data extracted from the 9×9 patch area around the centre points.

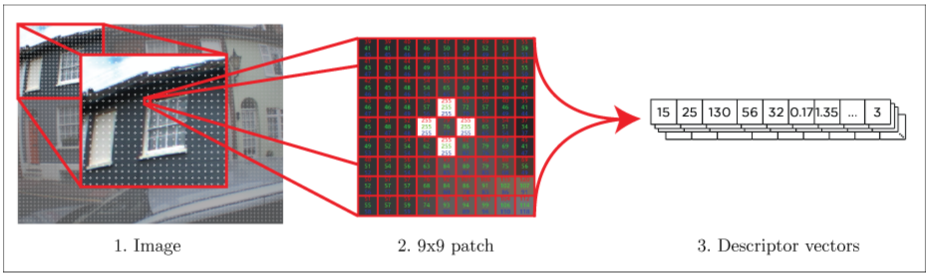

Fig. 3: Descriptor extraction.

Random Forest training The creation of Random forest follows the standard framework developed by [6]. The 90% of the images were used to create for each descriptor a database with 100 trees. At every non-leaf node in each tree, a binary test was assigned which chooses 4 variables in order to find the best split. The tree growing stops when it reached its maximum depth (15) or the maximum number of training images were reached. The forest was saved for the later classification and test.

Classification The final function is a feature segmentation algorithm. The saved trained Random Forest was used for a pixel-by-pixel classification on the test image to detect finer details on the facade. The results of this classification can be observed on Fig. 4. where each color represents a class, the yellow circles belong the wall class, the red the window or door class, the blue circles marking the roof class and the green circles showing the other class. We can corroborate that the descriptor is strong in the door and window areas and produce good results on the roof areas. The result images were turned into a binary image in order to eliminate the fine noise with morphological operations. This step reduces the irregularities around features which could cause a problem in the rectangle fitting (for example it creates a connection with other window areas). After detecting the contours of the segmented area, the algorithm utilizes topological analysis method [20] which counts all the non-zero components and extracts the boundaries on the binary image. To fit the rectangle around the area the Ramer-Douglas-Peucker algorithm (which recursively divides the line between the given first and last point) was used to approximate the polygon enclosed by the previously detected contour using with another polygon with less vertices. After the bounding box is detected, the results are re-projected to the original image.

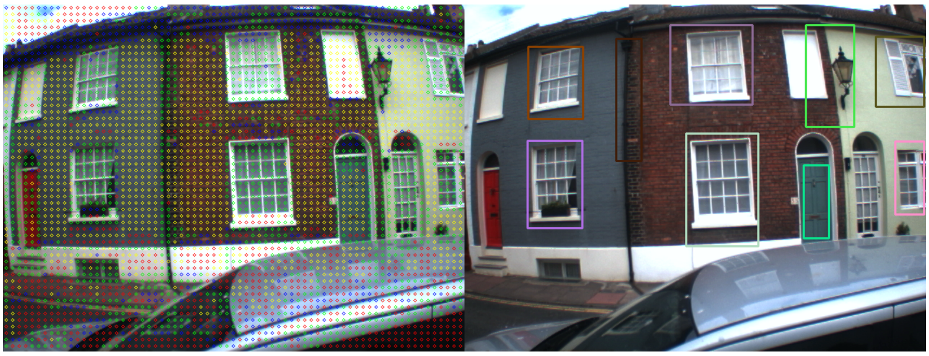

Fig. 4: Project descriptor classification results.

5 Results

The training and test images were saved from a video recording. It was recorded in two different occasions with different lighting conditions in Queens Gardens street in Brighton, United Kingdom. These videos were stored frame by frame and from this large set of images a dataset was selected. To test and evaluate the performance of the different descriptors, the results of the predictions from each image were compared to the ground truth of the same image. A confusion matrix was set from that information for each descriptor, which was the base for performance evaluation. On Table 1. we can see throughout the testing the Opponent SIFT and SIFT descriptors provide the most reliable performance. Above 70% were the correct detection rate. Interestingly the Opponent SIFT descriptor was designed for a coloured environment, but in the tests it did not give better performance as the simple SIFT descriptor. On the contrary, the descriptor which was designed for greyscale imagery, had a better performance throughout. On the other hand, the performance of the proposed descriptor designed for this project is giving the third best performance throughout the tests, and most importantly in the test where the window bounding rectangles were extracted, the speed of the proposed descriptor called Project is for long the fastest.

Table 1: Total true positives.(%)

	Number of training images				
Descriptor	**9**	**20**	**30**	**40**	**52**
SIFT	74.71	77.52	73.58	71.46	72.02
Opponent SIFT	72.52	75.47	70.61	69.15	70.85
SURF	57.85	57.19	55.17	55.35	54.72
Opponent SURF	57.27	56.26	55.89	55.49	55.97
Project	67.42	63.89	58.17	59.77	61.92

The Fig 5. shows the details of the precision of each tested descriptor in each segmentation category. The values of true positive points in each category show the level of accuracy of each descriptor. In each class, the Project descriptor was operating with high exactness even in sparse areas like the other areas where the efficiency of the SIFT and Opponent SIFT descriptors gave a poorer performance. In Fig 4. the distribution of classified points are displayed where we can observe the Project descriptor's results.

In Fig 6. it can be observed the time being spent by the computer to reading, calculating the necessary data for the descriptor and based on the outcome of the classification, segmenting the window and door areas. The result data shows that while the most effective SIFT descriptor needed 223 seconds, the Opponent SIFT descriptor for the same work needed three times as much effort in time as in this case the computer has to do the same calculation on the three color channels.

The Opponent SURF descriptor occupied 298 seconds for the work which is three times slower than the Project descriptor and its performance in detection were inferior to this descriptor. These results correlate with the finding of [21] where the SIFT descriptor was more accurate in feature matching but in a considerable longer time frame. Fig 4. shows the final outcome of the segmentation algorithm where it can be noticed the strength of window detection and the weakness of the descriptor vector in terms of distortion and rotation.

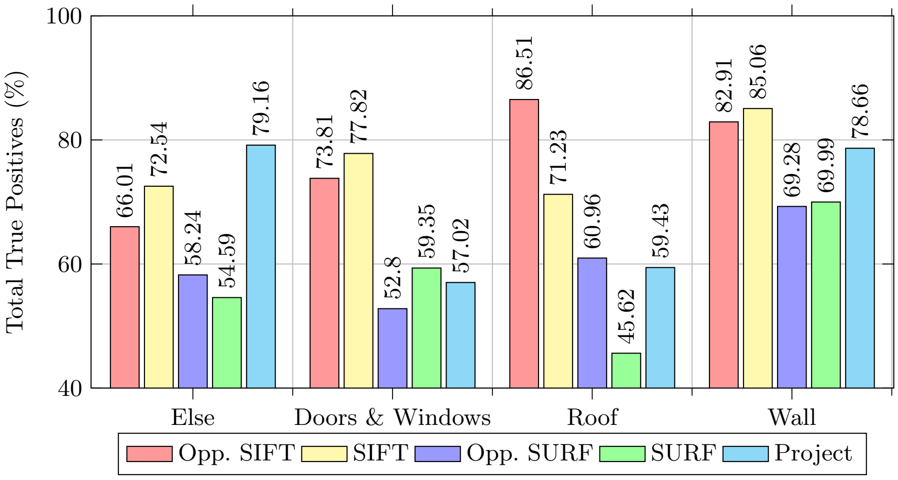

Fig. 5: Correctly qualified point ratio in category groups after training descriptors with 52 images

6 Conclusions

In this paper feature extraction with a new scene dedicated descriptor were studied based on speed and accuracy. The results show that although the mainstream descriptors have reliable performance in detect image features, a descriptor which is created specifically for a certain environment can have similar accuracy but in a shorter time period. This projects a new path to investigate a trained dynamic descriptor which can adjust characteristics of the retrieved information according to the environment. Based on the results it is also planned to stabilize the Project descriptor for rotation, light change and distortion, and to create another version for different environmental characteristics.

Acknowledgement. We would like to thank the Consejo Nacional de Ciencia y Tecnologa through the project number 340519 without whom this paper could not have been completed. Also, I would like to thank the Universidad Autnoma de Quertaro for its facilities and support.

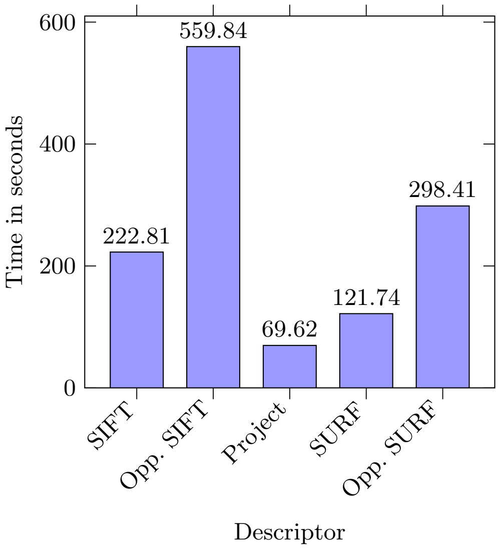

Fig. 6: Descriptor performance in time

References

1. Arth, C., Klopschitz, M., Reitmayr, G., Schmalstieg, D.: Real-time self-localization from panoramic images on mobile devices. In: Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on. pp. 37–46 (2011)
2. Azuma, R., Baillot, Y., Behringer, R., Feiner, S., Julier, S., MacIntyre, B.: Recent advances in augmented reality. IEEE Comput. Graph. Appl. 21(6), 34–47 (2001)
3. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Comput. Vis. Image Underst. 110(3), 346–359 (jun 2008)
4. Becker, S., Haala, N.: Refinement of building fassades by integrated processing of lidar and image data. International Archives of Photogrammetry, Remote Sensing and Spatial Information Science 36, 7–12 (2007)
5. Berg, A., Grabler, F., Malik, J.: Parsing images of architectural scenes. In: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on. pp. 1–8 (2007)
6. Breiman, L.: Bagging predictors. Mach. Learn. 24(2), 123–140 (aug 1996)
7. Breiman, L.: Random forests. Mach. Learn. 45(1), 5–32 (oct 2001)
8. Canny, J.: A computational approach to edge detection. Pattern Analysis and Machine Intelligence, IEEE Transactions on PAMI-8(6), 679–698 (1986)
9. Chu, D.M., Smeulders, A.W.M.: Color invariant surf in discriminative object tracking. In: ECCV Workshop on Color and Reflectance in Imaging and Computer Vision (2010)
10. Csurka, G., Dance, C.R., Fan, L., Willamowski, J., Bray, C.: Visual categorization with bags of keypoints. In: In Workshop on Statistical Learning in Computer Vision, ECCV. pp. 1–22 (2004)
11. Frohlich, B., Rodner, E., Denzler, J.: A fast approach for pixelwise labeling of facade images. In: Pattern Recognition (ICPR), 2010 20th International Conference on. pp. 3029–3032 (2010)

12. Gool, L.V., Zeng, G., den Borre, F.V., Müller, P.: Towards mass-produced building models. In: Stilla, U., Mayer, H., Rottensteiner, F., Heipke, C., Hinz, S. (eds.) Photogrammetric Image Analysis. pp. 209–220. Institute of Photogrammetry and Cartography, Technische Universitaet Muenchen (sep 2007)
13. Guan, T., Wang, C.: Registration based on scene recognition and natural features tracking techniques for wide-area augmented reality systems. Multimedia, IEEE Transactions on 11(8), 1393–1406 (2009)
14. Klein, G., Murray, D.: Parallel tracking and mapping for small ar workspaces. In: Mixed and Augmented Reality, 2007. ISMAR 2007. 6th IEEE and ACM International Symposium on. pp. 225–234 (2007)
15. Kurz, D., Benhimane, S.: Gravity-aware handheld augmented reality. In: Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on. pp. 111–120 (Oct 2011)
16. Kurz, D., Benhimane, S.: Augmented reality: Handheld augmented reality involving gravity measurements. Computers & Graphics 36(7), 866–883 (2012)
17. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. J. Comput. Vision 60(2), 91–110 (nov 2004)
18. Pu, S.: Automatic building modeling from terrestrial laser scanning. In: Oosterom, P., Zlatanova, S., Penninga, F., Fendel, E.M. (eds.) Advances in 3D Geoinformation Systems, pp. 147–160. Lecture Notes in Geoinformation and Cartography, Springer Berlin Heidelberg (2008)
19. Van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. vol. 32, pp. 1582–1596 (2010)
20. Suzuki, S., Be, K.: Topological structural analysis of digitized binary images by border following. Computer Vision, Graphics, and Image Processing 30(1), 32–46 (1985)
21. Valgren, C., Lilienthal, A.J.: Sift, SURF & seasons: Appearance-based long-term localization in outdoor environments. Robotics and Autonomous Systems 58(2), 149–156 (2010)
22. Wu, Y., Choubassi, M.E., Kozintsev, I.: Augmenting 3d urban environment using mobile devices. In: Mixed and Augmented Reality (ISMAR), 2011 10th IEEE International Symposium on. pp. 241–242 (2011)