AN ALTERNATIVE TO SCALE-SPACE REPRESENTATION FOR EXTRACTING LOCAL FEATURES IN IMAGE RECOGNITION

Hans Jørgen Andersen and Giang Phuong Nguyen
Media Technology Section, Department of Architecture, Design and Media Technology, Aalborg University,
Niels Jernes Vej 14, Aalborg, Denmark

Keywords: Local Descriptors, Image Features, Triangular Representation, Image Retrieval, Image Recognition.

Abstract: In image recognition, the common approach for extracting local features using a scale-space representation has usually three main steps; first interest points are extracted at different scales, next from a patch around each interest point the rotation is calculated with corresponding orientation and compensation, and finally a descriptor is computed for the derived patch (i.e. feature of the patch). To avoid the memory and computational intensive process of constructing the scale-space, we use a method where no scale-space is required This is done by dividing the given image into a number of triangles with sizes dependent on the content of the image, at the location of each triangle. In this paper, we will demonstrate that by rotation of the interest regions at the triangles it is possible in grey scale images to achieve a recognition precision comparable with that of MOPS. The test of the proposed method is performed on two data sets of buildings.

1 INTRODUCTION

The use of local features has during the last years proven as an powerful method for recognition of objects, places and navigation (Zhou et al., 2009; Koeck et al., 2005; Zhi et al., 2009; Huang et al., 2009) The advantages of using local features lead to an increasing number of researches exploring these and a comprehensive overviews can be found in (Mikolajczyk and Schmid, 2005; Tuytelaars and Mikolajczyk, 2008).In this study we will concentrate on its use for building recognition, using the method recently introduced by the authors (Nguyen and Andersen, 2010). The ambition is to develop a method that is less computational expense compared to scale-space methods and hence suitable for implementation on resource limited devices as mobile phones or tablets.

Up to now, local features is mostly known as descriptors extracted from areas located at interesting points (Tuytelaars and Mikolajczyk, 2008). This means that, existing methods first detect interesting points, for example using Harris corner detector (Harris and Stephens, 1988). Then a patch is drawn which is centered at the corresponding interest point, and descriptors are computed from each patch. So, the main issue is how to define the size of the patch. In other words, how to make these descriptors scale invariant. To satisfy this requirement, these methods need to locate a given image at different scales, or so called the scale-space approach. A given image is represented in a scale-space using difference of Gaussian and down sampling (Lowe, 2004; Brown et al., 2005; Nguyen and Andersen, 2008). The size of a patch depends on the corresponding scale where the interest point is detected.

In short, the common approach for extracting local features using a scale-space representation have usually three main steps; first extract interest points at each scale, next from a patch around each interest point calculate the rotation and compensate for this, and finally compute a descriptor at each interest point (i.e. feature of the patch).

In the paper (Nguyen and Andersen, 2010) we proposed an alternative to the first step approximating the scale-space using bTree triangular encoding as introduced by (Distasi et al., 1997). This method divides the image into triangles of ”homogenous” regions. The process is done automatically and if an object appears at different scales, the triangular representation will adapt to draw the corresponding triangle size. Each triangle may be view upon as an interest region or ”point”. In our previous study, we investigated the use of gray scale or color information for triangle descriptors. We demonstrated that using color we could achieve an performance in accordance with the MOPS method (Brown et al., 2005) and that

Jørgen Andersen H. and Phuong Nguyen G..
AN ALTERNATIVE TO SCALE-SPACE REPRESENTATION FOR EXTRACTING LOCAL FEATURES IN IMAGE RECOGNITION.
DOI: 10.5220/0003836203410345
In *Proceedings of the International Conference on Computer Vision Theory and Applications* (VISAPP-2012), pages 341-345
ISBN: 978-989-8565-03-7

the method is comparable in terms of repeatability with other local feature detection methods using the test introduced by (Mikolajczyk and Schmid, 2005). In this paper we will demonstrate that by rotation of the interest regions it is possible even in grey scale images to achieve a recognition precision comparable with that of MOPS.

The paper is organized as follows. In the next section, we will review a description of our approach using triangular representation for detection of homogenous regions, and how to compute local descriptors. In section 4, we introduce the proposed method for rotation of the interest regions before calculation of the descriptors. Experimental results in a image retrieval system are carried out in section 5.

2 INTEREST POINT DETECTOR

As mentioned above, the BTree triangular coding (BTTC) is a method originally designed for image compression (Distasi et al., 1997). For compression purpose, the method tries to find a set of pixels from a given image that is able to represent the content of the whole image. This means that given a set of pixels, the rest of the pixels can be interpolated using this set. In the reference, the authors divide an image into a number of triangles, where pixels within a triangle can be interpolated using information of the three vertices. The same idea can be applied to segment an image into a number of local areas, where each area is a homogenous triangle.

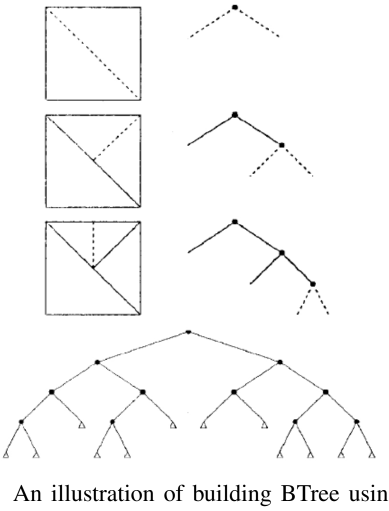

Figure 1: An illustration of building BTree using BTTC. The last figure shows an example of a final BTree.

A given image I is considered as a finite set of points in a 3-dimensional space, i.e. $I = \{(x,y,c)|c = F(x,y)\}$ where (x,y) denotes pixel position, and c is an intensity value. BTTC tries to approximate I with a discrete surface $B = \{(x,y,d)|d = G(x,y)\}$, defined by a finite set of polyhedrons. In this case, a polyhedron is a right-angled triangle (RAT). Assuming a RAT with three vertices (x_1,y_1), (x_2,y_2), (x_3,y_3) and $c_1 = F(x_1,y_1)$, $c_2 = F(x_2,y_2)$, $c_3 = F(x_3,y_3)$, we have a set $\{x_i,y_i,c_i\}_{i=1..3} \in I$. The approximating function $G(x,y)$ is computed by linear interpolation:

$$G(x,y) = c_1 + \alpha(c_2 - c_1) + \beta(c_3 - c_1) \tag{1}$$

where α and β are defined by the two relations:

$$\alpha = \frac{(x-x_1)(y_3-y_1)-(y-y_1)(x_3-x_1)}{(x_2-x_1)(y_3-y_1)-(y_2-y_1)(x_3-x_1)} \tag{2}$$

$$\beta = \frac{(x_2-x_1)(y-y_1)-(y_2-y_1)(x-x_1)}{(x_2-x_1)(y_3-y_1)-(y_2-y_1)(x_3-x_1)} \tag{3}$$

An error function is used to check the approximation:

$$\text{err} = |F(x,y) - G(x,y)| \leq \varepsilon, \varepsilon > 0 \tag{4}$$

If the above condition is not met then the triangle is divided along its height relative to the hypotenuse, replacing itself with two new RATs. The coding scheme runs recursively until no more division takes place. In the worst case, the process is stopped when it reaches to the pixel level, i.e. three vertices of a RAT are three neighbor pixels and err=0. The decomposition is arranged in a binary tree. Without loss of generality, the given image is assumed having square shape, otherwise the image is padded in a suitable way. With this assumption, all RATs will be isosceles. Finally, all points at the leave level are used for the compression process. Figure 1 shows an illustration of the above process.

In the reference (Distasi et al., 1997), experiments prove that BTTC produces images of satisfactory quality in objective and subjective point of view. Furthermore, this method is very fast in execution time, which is also an essential factor in any processing system. We note here that for encoding purpose, the number of points (or RAT) is very high (up to several ten thousand vertices depending on the image content). However, we do not need that detail level, so by increasing the error threshold in Eq.(4) we obtain fewer RATs while still fulfilling the homogenous region criteria. Examples using BTTC to represent image content with different threshold values are shown in figure 2.

3 DESCRIPTOR COMPUTATION

For description of the homogenous regions we use the gray-scale histogram. For estimation of the rotation

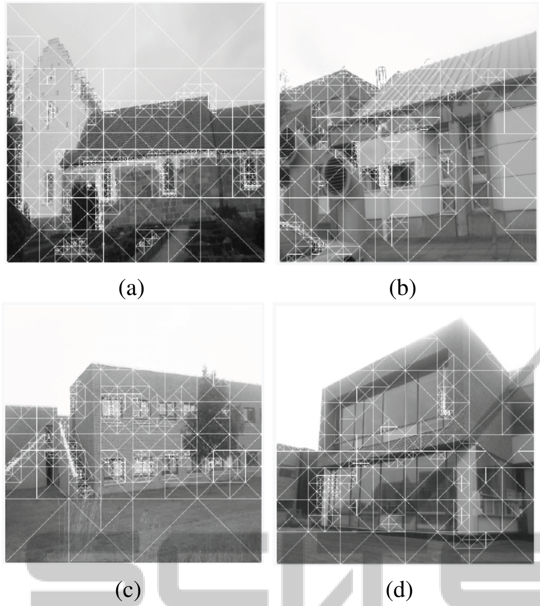

Figure 2: Examples of BTTC for extracting triangles.

a descriptor of each triangle is computed similar to method used in MOPS (Brown et al., 2005). First it computes the rotation of the patch using gradient values with a resolution of 10 orientation-bins. The patch is rotated according to the direction with the highest value of gradients. The rotated patch is then sampled into 8x8 areas. Accordingly, for each area the average gray scale value is computed, so finally we have a 64-dimensional feature vector.

4 ORIENTATION ESTIMATION

The BTTC encoding accounts for changes in scales but are sensitive to changes in rotation as the area used for calculation of the descriptor may change significantly. To compensate for this we will introduce a method inspired by that used in (Lowe, 2004; Brown et al., 2005).

For each triangle its center point μ is calculated by the average of the three vertices $\upsilon_{1,2,3}$ as described in Eq. 5. A square patch with its size determined by the maximum minus the minimum of the vertices Eq. 6 is then placed with its center at the triangles center point.

$$\mu = \frac{1}{3}\sum_{i=1}^{3}\upsilon_i \tag{5}$$

$$\text{size of patch} = \max(\upsilon_{1,2,3}) - \min(\upsilon_{1,2,3}) \tag{6}$$

Within the square the gradients is calculated and summed in bins with a resolution of 10 degrees. The maximum bin is found and the patch is rotated according to this, see figure 3. In this way the method is similar to the SIFT and MOPS approach (Lowe, 2004; Brown et al., 2005). The descriptor is now derived from the rotated patch.

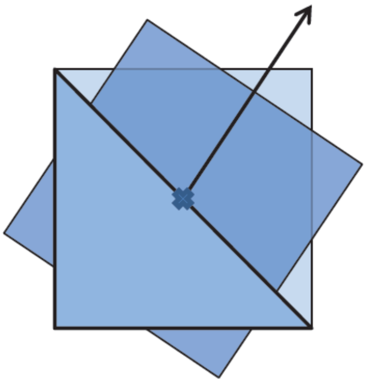

Figure 3: Triangle with initial square patch around μ the center point of the square defined by the vertices, $\upsilon_{1,2,3}$. In dark blue the rotated patch according to the orientation of the gradients in the initial square patch.

5 EXPERIMENTAL SETUP

5.1 Databases

Our experiments were carried out with two datasets. The first one is called the AaU dataset, where images are captured different buildings/objects in the area of Aalborg University. This dataset contains of 410 images of 21 buildings, which is on average 20 images for each building. The reason for this rather large number per building is that we captured those buildings through the year including different imaging conditions such as weather and light sources. It should also be noted that all the buildings in the area are quite similar in their textures (for examples, brick wall or glasses). Moreover, images of the same building are taken at different viewpoints, rotations, and scales. In figure 4, we show example images of the same building.

The second dataset was taken from (Shao et al., 2003). This database contains images of different Zurich city buildings. We randomly select a set of 40 buildings, creating a database of 201 images. Each building was also taken with five different scales, viewpoints or orientations.

5.2 Evaluation

Evaluation of local features is based on image retri-

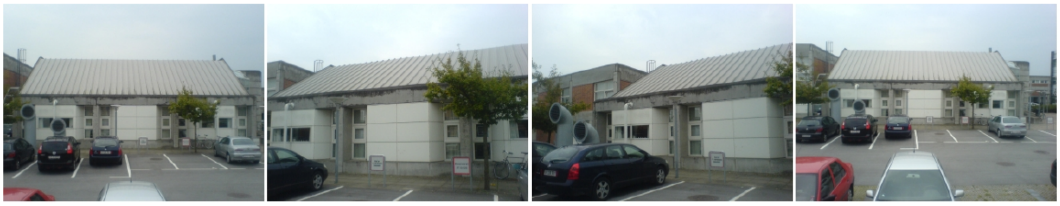

Figure 4: An example of images taken from the same building.

eval performance. To do so, we sequentially use images from the given dataset as a query image. Local features are extracted from the query image and compared to features of all other images in the dataset. The top 5 best matching are returned and groundtruth is manually assigned to corresponding building to compute the precision rate:

$$\text{precision} = \frac{\text{\# of correct matches}}{\text{\# of returned images}}(\%). \qquad (7)$$

6 RESULTS AND DISCUSSION

The average retrieval rate for the two data sets are shown in figure 5 and 6 We also report the retrieval performance using the multi-scale oriented patches (MOPS) (Brown et al., 2005), which applies the common approach using a 5 scale-space representation, for comparison. From the figures it is evident that the rotation of the local interest points increase the performance by 10% for the first image decreasing to below 5% for the fifth image.

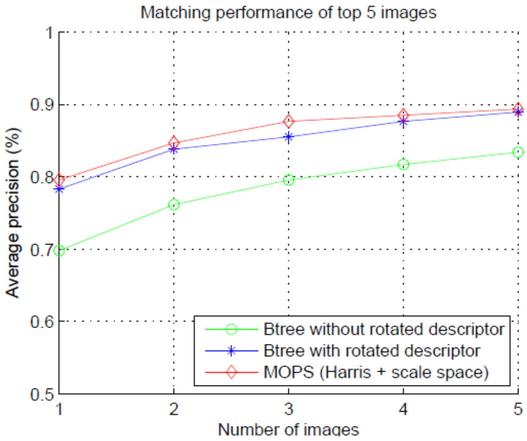

Figure 5: Precision vs. number of returned images for the AaU data set.

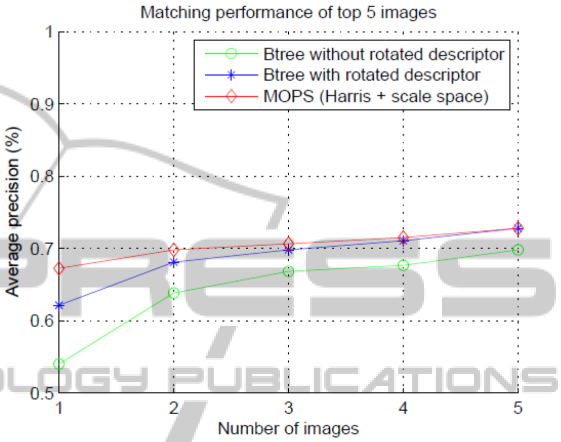

Figure 6: Precision vs. number of returned images for the Zurich data set.

The results show that adding the rotation to descriptors, the proposed method has a compatible performance with the MOPS. Besides, the complexity of the method is much less compared to the scale-space approaches, since it does not require the scale-space structure for extracting descriptors at each scale level.

The two methods was implemented in C++ and their processing time logged using a desktop Pc with an AMD Athlon 2.20 GHz processor with 4G RAM. The result from a paired t-test of this initial study showed that the Btree method in terms of processing time per feature is significant $(p < 0.001)$ faster than MOPS. The increase in per feature point processing time is between 10-15%. However, because the Btree method use more feature points than MOPS, it is only significant faster for the Zurich dataset $(p < 0.001)$ in terms of the total processing time per image, whereas MOPS is faster for the AaU data set $(p < 0.01)$. The average time processing per image was for the AaU data set 483 msec. for the Btree method and 444 msec. for MOPS, whereas for the Zurich data set the average processing time per image was 506 msec. for the Btree method and 567 msec. for MOPS. So, to draw any firm conclusions a more thorough performance evaluation has to be done, including aspects as varying threshold for homogenous regions etc.

7 CONCLUSIONS

In this paper, we propose bTree triangular encoding as introduced by (Distasi et al., 1997) for detection of local interest points. This method divides the image into triangles of homogenous regions. The process is done automatically and if an object appears at different scales, the triangular representation will adapt to draw the corresponding triangle size. Each triangle may be view upon as an interest region or ”point”. We demonstrate that by rotation of the interest regions it is possible in grey scale images to achieve a recognition precision comparable with that of MOPS. Without the scale-space construction, the method is simpler and less computational expense. Therefore, it is suitable for usage on platforms with limited resources as mobile phones or tablets.

REFERENCES

Brown, M., Szeliski, R., and Winder, S. (2005). Multi-image matching using multi-scale oriented patches. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, volume 1, pages 510–517.

Distasi, R., Nappi, M., and Vitulano, S. (1997). Image compression by B-Tree triangular coding. *IEEE Transactions on Communications*, 45(9):1095–1100.

Harris, C. and Stephens, M. (1988). A combined corner and edge detector. *Proceedings of the Alvey Vision Conference*, pages 147–151.

Huang, S., Cai, C., Zhang, Y., He, D. J., and Zhang, Y. (2009). An efficient wood image retrieval using surf descriptor. *2009 International Conference on Test and Measurement*, 2:5558.

Koeck, J., Li, F., and Yang, X. (2005). Global localization and relative positioning based on scale-invariant keypoints. *Robotics and Autonomous Systems*, 52(1):27 – 38.

Lowe, D. (2004). Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110.

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

Nguyen, G. and Andersen, H. (2008). Urban building recognition during significant temporal variations. In *Proceedings of the IEEE Workshop on Applications of Computer Vision*, pages 1–6.

Nguyen, G. P. and Andersen, H. J. (2010). A new approach for detecting local features. In *International Conference on Computer Vision Theory and Applications*, pages 1–6. Institute for Systems and Technologies of Information, Control and Communication.

Shao, H., Svoboda, T., and Gool, L. V. (2003). Zubud zurich buildings database for image based recognition. Tech report, Swiss Federal Institute of Technology. http://www.vision.ee.ethz.ch/showroom/zubud/.

Tuytelaars, T. and Mikolajczyk, K. (2008). Local invariant feature detectors. *Foundations and Trends in Computer Graphics and Vision*, 3(3):177–280.

Zhi, L. J., Zhang, S. M., Zhao, D. Z., Zhao, H., and Lin, S. K. (2009). Medical image retrieval using sift feature. In *Proceedings of the 2009 2nd International Congress on Image and Signal Processing CISP09*.

Zhou, H., Yuan, Y., and Shi, C. (2009). Object tracking using sift features and mean shift. *Computer Vision and Image Understanding*, 113(3):345 – 352.