Image retrieval with the use of Bag of Words and structural analysis

R Malashin¹
¹National Research University of Information Technology, Mechanics and Optics, 197101, 49 Kronverkskiy av., St. Petersburg, Russia
Email: malashinroman@mail.ru

Abstract. The paper considers task of image retrieval with the use of Bag of Words (BoW). The significant improvement in robustness is achieved by using fast and robust algorithm of structural verification based on Hough transform of parameters of keypoint matches.

1. Introduction
One of important tasks in computer vision is image retrieval. There exist many factors that make the problem challenging: illumination change, viewpoint change, content variety. Target independent analysis based on coarse structural elements [1-3] can provide robust results but these algorithms consumes too much computational resources for most of practical applications. Methods based on local features such as SIFT [4] or SURF [5] are much more effective especially when combined with Bag of Words (BoW) [6], because histogram of visual words is very compact representation of an image. The major drawback is that original BoW concept doesn't take size, orientation and location of features into account. In this paper spatial information is utilized by very fast and robust procedure of matches clustering.

2. Proposed methods
The basic idea of the structural (or spatial) verification is similar to the idea of Philbin et al. [6]. Let us consider that BoW histograms are precomputed for every image in database and there is query image. Then the retrieval algorithm consists of the following steps:

- Compute BoW histogram for the query image.
- Find $N$ BoW histograms closest to the query histogram according to tf.idf [6].
- For every $i$-th of $N$ best image candidates: a) generate keypoint-to-keypoint matches of $i$-th and query image; b) eliminate outliers using scene geometry constraint.
- Re-rank search results according to number of geometrically consistent matches.

Below there is a brief description of the proposed algorithms.

2.1. Algorithm to generate keypoint matches
Local matches are generated by interpreting corresponding histogram bins as a set of keypoint matches. Matches that arise from bins with several features are less reliable than those derived from bins with single feature. This information is considered by assigning weights to generated matches. Total number of generated matches is $m_1 \cdot m_2$, where $m_1$, $m_2$ – number of features assigned to specified visual word that are detected in two images (sizes of corresponding bins of two BoW histograms). The weight of every $i$-th match is computed as follows:
\[ w_i = \frac{\min(m_1, m_2)}{m_1 \cdot m_2} \]  

(1)

Despite the fact that corresponding features may not be detected or can be assigned to different visual words, \( w_i \) equals probability of match \( i \) to be correct. The details on usage of these weights are discussed in the next subsection.

2.2. Scene geometry constraint

Due to different causes most of generated keypoint matches are incorrect. The general spatial constraint on image pair of static 3D scene implies usage of epipolar geometry. Unfortunately RANSAC algorithms are not able to recover fundamental matrix when number of outliers is more than 70-75%. In this work spatial information is utilized by Hough transform that is used to find clusters of matches that agree on the same parameters of similarity transform [4]. That allows successful matching with less than 5% of inliers.

Hough transform is affected by boundary effect. To avoid the problem in [4] each keypoint match votes for the 2 closest bins in each dimension, giving total of 16 entries for each hypothesis [4]. The alternative solution proposed in this paper is to organize several Hough tables that are displaced against each other in each dimension by half of bin size, every match increments appropriate bin in every table. Figure 1 illustrates this approach for 1- and 2-dimensional parameter space (similarity transform parameter space is 4D). The two approaches (several votes in single table and single vote to several tables) were compared through experiments on matching image pairs (similar to experiments described in section 3). The proposed approach with only two displaced Hough tables provide almost the same matching quality as voting for 16 hypotheses in one table (0.3% decrease in number of correctly matched images). Usage of only two tables means that boundary effect is partially ignored (figure 1b) but total number of entries to the tables decreases almost by a magnitude. That leads to a magnitude decrease of clusters found by Hough transform (see figure 2) and allows significant speedup of the program.

![Figure 1](image1.png)  

Figure 1. Usage of two Hough tables A1 and A2 to avoid boundary effect. (a) – in case of 1-dimensional parameters space, (b) – in case of 2-dimensional parameter space. In (b) circles mark points in parameter space where clusters can be split.

![Figure 2](image2.png)  

Figure 2. Number of bins with more than two votes relative to total number of votes. The number is significantly reduced by partially ignoring boundary effect using two Hough tables. The graph was generated by matching hundreds of images having non-overlapping content.

According to equation (1) different matches have different probability to be correct. Only clusters that contain at least 1% of total match weights were considered, others are eliminated. The verification
procedure that is able to remove rest of ‘false’ clusters and incorrect matches from ‘correct’ clusters was developed. The procedure is based on RANSAC in similarity transform parameter space. Two matches are “randomly” picked from cluster and form a sample subset that is used to recover unknown parameters. The probability of picking a match equals its weight (equation (1)) divided by sum of weights of all matches in cluster. The latter was efficiently implemented by “cloning” indexes of more probable matches. Hypothesis that do not agree with parameters are rejected without verification of the cluster. RANSAC is followed by least square solution for the best parameters and then outlier elimination takes into account all parameters of similarity transform.

Note that the proposed algorithm described in section 2.1 generates 1-to-n matches that have higher probability to form incorrect (false) clusters of matches with relevant mutual parameters [7]. In this work we had to bear upon this fact but matches with $w_i < 0.1$ are eliminated before the Hough transform.

3. Experimental results

Experiments were carried out on the database of manually labelled indoor images with resolution 640×480. The very compact vocabulary was used (with only 2000 visual words). The experiment on image matching has shown that the proposed methods significantly outperforms baseline method (SURF [7] + epipolar geometry constraint). The results is summarized in table 1.

| Detector | Descriptor | Geometrical constrains | Correct keypoint matches, % | Number of pairs with more than 50% of inliers |
|----------|------------|------------------------|----------------------------|---------------------------------------------|
| SURF     | SURF       | Epipolar geometry      | 53,76                      | 205                                         |
| SURF     | SURF + BoW | Cluster analysis        | 88,37                      | 273                                         |

In the second experiment the developed algorithms were applied directly to the task of image retrieval (for database with several hundred images). It appeared that the proposed structural verification of 10 image candidates decreases false positives by more than a magnitude and false negatives by 36%. At the same time structural verification of one image pair consumes only 3 ms in average in PC environment.

Acknowledgements
This work was supported by the Ministry of Education and Science of the Russian Federation, and by the Government of the Russian Federation, Grant 074-U01.

References
[1] Lutsiv V, Malyshev I, Pepelka V and Potapov A 2002 Proc. of SPIE 4741 351–62
[2] Lutsiv V, Malyshev I and Potapov A 2003 Proc. of SPIE 5238 164–75
[3] Lutsiv V, Malyshev I and Pepelka V 2001. Proc. of SPIE 4380 174–83
[4] Lowe D 2004 Int. J. of Computer Vision 60 91-110
[5] Bay H, Tuytelaars T and Van Gool L 2006 SURF: Speeded Up Robust Features Proc. 9th European Conf. on Computer Vision (Graz, Austria, May 7-13) pp 404–417
[6] Philbin J, Chum O, Isard M, Sivic J, Zisserman A 2007 Object retrieval with large vocabularies and fast spatial matching Conf. on CVPR (Minneapolis, USA, June 18–23) pp 1–8
[7] Malashin R 2014 J. Phys.: Conf. Ser 536 012018