RATIO VOTING: A NEW VOTING STRATEGY FOR LARGE-SCALE IMAGE RETRIEVAL

Yusuke Uchida, Koichi Takagi, Shigeyuki Sakazawa

KDDI R&D Laboratories, Inc., Japan

ABSTRACT

We propose a new voting strategy referred to as ratio voting to improve bag-of-visual words-based image retrieval. It limits the number of votes in proportion to the number of features in visual words, while conventional schemes use (estimated) distances or rank information as a filtering criterion. Ratio voting realizes adaptive thresholding that captures the density of feature vectors. In experiments, we adopt two different distance estimation methods in the post-filtering step and show that ratio voting achieves a considerable improvement in spite of its simplicity in both cases. Furthermore, we perform exhaustive experiments in combining ratio voting with multiple assignment approaches and show that choosing a multiple assignment approach also has a remarkable impact on accuracy.

***Index Terms*—** Specific object recognition, visual words, hamming embedding, product quantization, inverted index

1. INTRODUCTION

With the advancement of both stable interest region detectors [1] and robust and distinctive descriptors [2], local feature-based image or object retrieval has attracted a great deal of attention. Particularly, it has become applicable to large-scale databases with a bag-of-visual words (BoVW) framework [3]. Figure 1 illustrates a standard framework of BoVW-based image retrieval system. In the BoVW framework, local feature points or regions are detected from an image, and feature vectors are extracted from them. These feature vectors are quantized into visual words (VWs) using a visual codebook (visual vocabulary), resulting in a histogram representation of VWs. Image similarity is measured by L_1 or L_2 distance between the normalized histograms. As VW histograms are generally very sparse, an inverted index data structure and a voting function enables an efficient similarity search. The equivalency between L_2 distances and scores obtained with the voting function is described in [4] in detail. A tf-idf weighting scheme [3] is naturally integrated with the voting function. Finally, geometric verification [5] is performed to refine the results obtained with the voting function.

Although the BoVW framework realizes efficient retrieval, there is some room for improvement in terms of accuracy. One significant drawback of VW-based matching is that two features are matched if and only if they are assigned to the same VW [4]. There are two major expansions of voting function to alleviate this problem: post-filtering approaches [4, 6] and multiple assignment approaches [7, 4]. In post-filtering approaches, after VW-based matching, unreliable matches are filtered out according to (estimated) distances between query and reference features. In multiple assignment approaches, query features vote not only for reference features in the nearest VW but also for reference features in the k-nearest VWs.

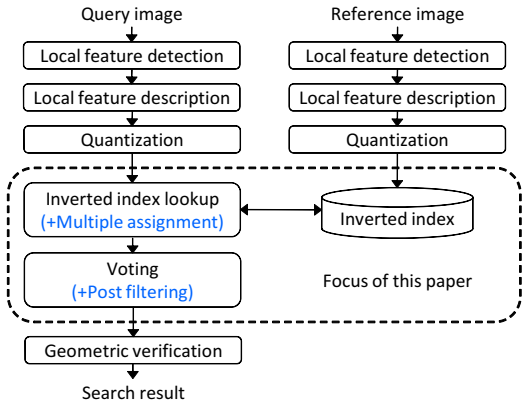


Fig. 1. Standard framework of bag-of-visual words-based image retrieval system.

To date, local feature detectors/descriptors have been investigated quite well in the literature [8, 1, 2, 9]. Although a choice of voting strategies has also considerable impact on the accuracy of search results as shown in Section 4.2, only a few studies have focused on voting processes [4, 7]. In this paper, focusing on voting processes, we propose a new voting strategy referred to as ratio voting; it limits the number of votes in proportion to the number of features in VWs, while conventional schemes use (estimated) distances or rank information as a filtering criterion. Ratio voting realizes adaptive thresholding that captures the density of feature vectors. In experiments, we show that ratio voting achieves a considerable improvement in spite of its simplicity. Furthermore, we perform exhaustive experiments wherein ratio voting is combined with multiple assignment approaches and show that the choice of a multiple assignment approach also has a remarkable impact on accuracy.

2. IMPROVING VW-BASED IMAGE RETRIEVAL

There are two major approaches used to improve the performance of VW-based image retrieval in voting: post-filtering approaches and multiple assignment approaches. Both approaches are reviewed in this section.

2.1. Post-filtering approaches

As the naive BoVW framework suffers from many false matches of local features, post-filtering approaches are proposed to suppress unreliable feature matches [4]. In this section, an overview of post-filtering approaches to improve naive VW-based image retrieval is presented. There are two important parts to post-filtering approaches: distance estimation and filtering criteria.

2.1.1. Distance estimation

As mentioned previously, after VW-based matching, distances between a query feature and reference features that are assigned to the same visual word as the query feature are estimated for post-filtering. As exact distance calculation is undesirable in terms of computational cost and memory requirement to store raw feature vectors [10], short code-based methods are used for this purpose [4, 10]: feature vectors are encoded into short codes and distances between feature vectors are approximated by distances between the short codes. In this paper, we adopt a product quantization (PQ)-based method [10]. It has been shown to outperform other short codes like spectral hashing (SH) [11] or a transform coding-based method [12] in terms of the trade-off between code length and accuracy in approximate nearest neighbor search. In the PQ method, a reference feature vector is decomposed into low-dimensional subvectors. Subsequently, these subvectors are quantized separately into a short code, which is composed of corresponding centroid indices. The distance between a query vector and a reference vector is approximated by the distance between a query vector and the short code of a reference vector. Distance calculation is efficiently performed with a lookup table. This distance is used to filter out unreliable feature matches according to filtering criteria introduced in Section 2.1.2, which considerably improves the precision of matching with only slight degradation of recall.

2.1.2. Filtering criteria

Based on the estimated distances described above, unreliable feature matches are filtered out. There is room for discussion on how to utilize the distances. To date, several criteria are used for filtering.

- Distance criterion: The most straightforward way is to filter out reference features with larger (approximated) distances than the predefined threshold [4, 6].

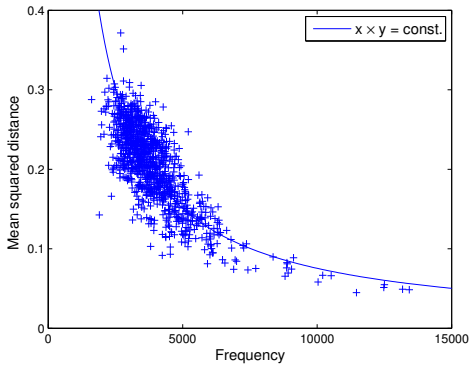
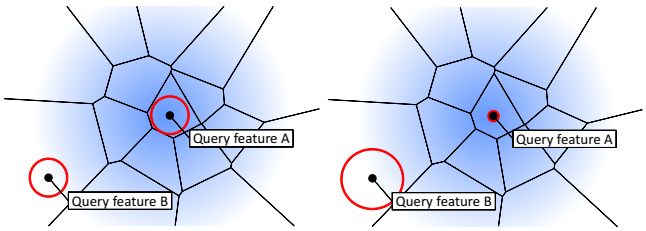


Fig. 2. For each VW, the number of feature vectors in the VW and the mean squared distance between the feature vectors and its centroid is plotted. A visual codebook with the size of 20K is created from 4M SIFT feature vectors introduced in Section 4.1. Here we plot randomly selected 2K VWs out of 20K VWs.

- Rank criterion: The alternative is to use the k-nearest neighbor features in voting and to filter out the others [10]. In this case, for each feature vector in a query image, reference features are sorted according to distances between the query feature and the reference features in ascending order, and corresponding top-k reference features are used in voting.

2.2. Multiple assignment approaches

While post-filtering approaches try to improve the precision of feature matches with only slight degradation of recall, multiple assignment approaches improve recall at the cost of the precision of feature matches. The basic idea here is, at a search step, to assign a query feature not only to the nearest VW but to the several nearest VWs. This technique alleviates the problem of quantization error; sometimes, similar features are assigned to different VWs. In [7], each query feature is assigned to the fixed number of the nearest VWs and the influence of a matched feature to image similarity is weighted according to the distance between the query feature and the assigned VWs. In [4], the distance d_0 to the nearest VW from a query feature is used to determine the number of multiple assignments, where the query feature is assigned to the VWs such that the distance to the VWs is smaller than αd_0 ($\alpha = 1.2$ in [4]). This approach adaptively changes the number of assigned VWs according to ambiguity of the feature. In this paper, we refer the former approach as fixed number multiple assignment (Fixed MA) and the latter as adaptive number multiple assignment (Adaptive MA). As post-filtering approaches and multiple assignment approaches are complementary, it is desirable to use multiple assignment in conjunction with post-filtering.



(a) Distance criterion (b) Rank criterion

Fig. 3. Querying range. The shade of a color represents the density of feature vectors. Red circles show the maximum range of features that can be matched with the query feature depending on the two criteria. In the case of (a), the range is explicitly defined, while, in (b), the range is implicitly determined by the distance between the query feature and the k-th nearest reference feature.

3. PROPOSED VOTING STRATEGY

In this section, we describe the problem of unfairness among VWs caused by conventional post-filtering criteria that were described in Section 2.1.2. Accordingly, a new post-filtering criterion is proposed to alleviate this problem. Finally, two strategies are explained wherein the proposed filtering criterion is combined with the multiple assignment approaches.

3.1. Problem in conventional criteria

Although it is well-known that the frequency of VWs satisfies Zipf's law [13], the relationship between the frequency and the size of a visual word cell has not been comprehensively investigated. Figure 2 shows the relationship between the frequency of VWs and the mean squared distance between feature vectors in a VW and its centroid. The mean squared distance roughly corresponds to the size of the Voronoi cell of a VW in feature space. It can be seen that VWs with a larger number of features tend to have relatively small Voronoi cell sizes. In other words, the density of feature vectors is quite different from one VW to another, which should be considered when designing post-filtering approaches.

Figure 3 illustrates the problem associated with conventional criteria in terms of feature density. In the case of the criterion based on raw distance (Figure 3 (a)), query features in frequent VWs (e.g., feature A) cause a large number of votes, while query features in infrequent VWs (e.g., feature B) cause a small number of votes. In the case of the criterion based on rank (Figure 3 (b)), despite the divergence in feature density, the number of votes is the same in all VWs. If we convert the rank criterion into distance, a query feature in frequent (e.g., feature A) VWs can be matched only with those features that are very near to the query feature, while a query feature in infrequent VWs (e.g., feature B) can be matched with features far from the query feature. This is in contrast to the distance criterion.

To summarize, for both criteria, some VWs have a significant impact in voting and others have little impact. This unfairness among VWs degrades final accuracy after voting and the criteria do not make the best of the post-filtering approach.

3.2. New filtering criterion

To alleviate the unfairness among VWs, we propose a new filtering criterion that restricts the number of votes in proportion to the number of features in a VW.

- Ratio criterion: Instead of using top-k reference features, top-p *proportion* of reference features in a corresponding VW are used in voting.

This is a natural extension of VW-based matching. According to the ratio criterion, filtering is always performed in all VWs if the threshold p is smaller than 1.0. Where $p = 1.0$, retrieval results become identical to those obtained by naive VW-based matching. With the distance and rank criterion, there is a threshold where filtering is not performed in some VWs and in the other VWs filtering is performed, which is an extreme example of unfairness. The ratio criterion can be regarded as a modified version of the rank criterion whereby the threshold k is adaptively changed according to the number of features in the same VW as a query feature: let N_v denote the number of features assigned to v-th VW, $k = p \times N_v$. As the number of features in each VW is known before querying, sorting by distances can be performed efficiently with a fixed-size heap.

3.3. Combination with multiple assignment approaches

There are several options in combining the ratio criterion and multiple assignment approaches. In this paper, we explore following two strategies.

- Strategy A: apply post-filtering independently in each of the multiply assigned VWs.
- Strategy B: apply post-filtering after merging all features in multiply assigned VWs.

Figure 4 depicts the difference between strategy A and strategy B, assuming that a query feature is assigned to two VWs, X including 400 features and Y including 100 features, and that the threshold for ratio voting is set to 0.1. In the case of strategy A (Figure 4 (b)), the reference features in VW X and VW Y are sorted independently according to distances from the query feature. Subsequently, the top 10% of features in each VW, 40 features from VW X and 10 features from VW Y, are used for voting. In the case of strategy B (Figure 4 (c)), the reference features in VWs X and Y are put together and sorted according to distances from the query feature. Thereafter, the top 10% of features (50 features) in VW X or Y are used for voting. Note that strategy B is applicable

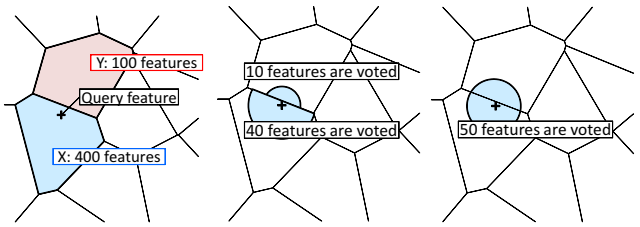


Fig. 4. Two strategies in combination with multiple assignment. (a) Assuming a query feature is assigned to two VWs X and Y: VW X includes 400 reference features and VW Y includes 100 reference features. (b) Strategy A: $400\times0.1=40$ features are voted from VW X and $100\times0.1=10$ features are voted from VW Y, respectively. (c) Strategy B: $(400+100)\times0.1=50$ features are voted from VW X and Y.

(reasonable) only in the case where estimated distances are comparable among different VWs.

4. EXPERIMENTAL EVALUATION

In this section, different voting strategies are compared using a publicly available dataset. First of all, the distance, rank, and ratio criterion are compared in terms of image retrieval accuracy. Second, rank and ratio criterion are evaluated in terms of the unfairness discussed in Section 3.1. Third, the ratio voting is combined with four types of multiple assignment approaches for further improvement. A weighted voting technique is also introduced.

4.1. Experimental setup

Experiments are performed on a publicly available dataset provided by [14]: the University of Kentucky recognition benchmark dataset[1]. The dataset includes 2,550 different objects or scenes. Each of these objects is represented by four images taken from four different angles, giving a total of 10,200 images. These images are used as both reference and query images. Mean average precision (MAP) [14, 4] is used as an indicator of performance. As SIFT feature vectors extracted from the test dataset and also other datasets (e.g., "cd training dataset") are available[2], these feature vectors are used in the experiments for reproducibility. We use the first 4M feature vectors from the "cd training dataset" to create a visual codebook and codebooks for product quantization. Note that these features are extracted from completely different images from the 10,200 test images. Standard parameter settings [4, 10] are used in our experiments. The size of the visual codebook is set to 20K. For the PQ method, reference features are divided into 8 16-dimensional subvectors, and encoded by a product quantizer with 256 centroids, resulting in 8×8 bit codes[3].

[1] http://www.vis.uky.edu/~stewe/ukbench/

[2] http://vis.uky.edu/~stewe/ukbench/data/

4.2. Impact of post-filtering approaches

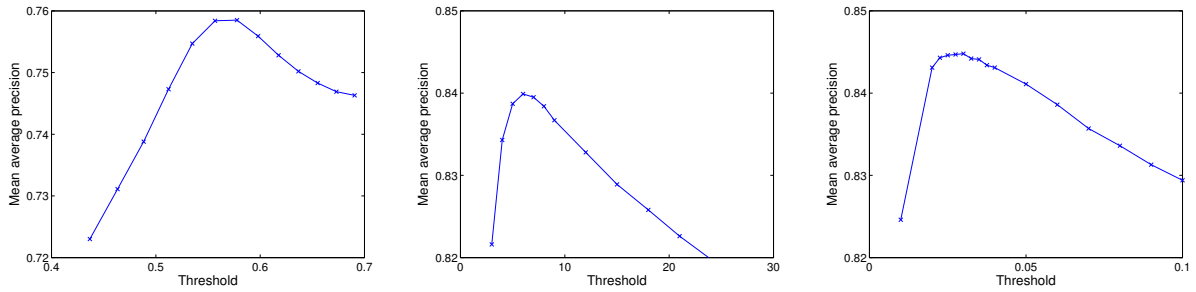
In this section, the impact of different post-filtering approaches is explored. Figure 5 shows image retrieval accuracy (MAP) for different thresholds. Using distances between a query feature and reference features estimated by the PQ method, filtering is performed based on (a) distance, (b) rank, and (c) ratio criterion. Figure 5 (b) corresponds to the state-of-the-art scheme described in [10]. Post-filtering based on Euclidean distance does not work well due to diversity in the size of each VW cell. A larger threshold cannot filter out any matches in most of the VWs with small Voronoi cells, while a lower threshold filters out few matches in VWs with large Voronoi cells. The rank criterion achieves better performance over the distance criterion by alleviating unfairness. This is because the distance from the feature vector to its k-th nearest neighbor can capture the local scale in a feature space [15]. However, it is not the best approach because in the case of VW-based image retrieval, the number of features in each VW is also quite different. The ratio criterion achieves the best performance by automatically adjusting the threshold according to the number of features in the VWs. Mean processing time required in voting increases from 8 [msec] to 14 [msec] when we adopt ratio voting. This overhead is negligible compared with the processing time required for feature detection and extraction. We also tried a Hamming embedding method [4], where feature vectors are embedded into the Hamming space and distances between them are estimated by the Hamming distance. It is found that the best performance (a MAP score of 0.835) obtained by using the rank criterion with $th = 7$ is outperformed by the best performance (a MAP score of 0.846) obtained by using the ratio criterion with $th = 0.036$.

4.3. Comparison in terms of feature repeatability

In this section, the unfairness among VWs is explored for the rank and ratio criteria. We introduce a new measurement, called a repeatability score S, defined by the average number of ground truth images that get at least one vote from each query feature. The score reflects the repeatability of query features in voting. Let Q denote a set of query features from all query images, $\mathrm{NN}(q)$ denote a set of reference image identifiers voted from query feature q, and $I_{\mathrm{GT}}(q)$ denote a set of ground truth image identifiers corresponding to q; the repeatability score S is defined by

$$S = \frac{1}{|Q|} \sum_{q \in Q} \sum_{i \in I_{\mathrm{GT}}(q)} R(i, \mathrm{NN}(q)), \qquad (1)$$

[3] This corresponds to the notation $m = 8$ and $k^* = 256$ in [10].



(a) Distance as a threshold: The best performance (MAP = 0.759) is achieved with th = 0.58.

(b) Rank as a threshold: The best performance (MAP = 0.840) is achieved with $th = 6$.

(c) Ratio as a threshold: The best performance (MAP = 0.845) is achieved with $th = 0.03$.

Fig. 5. MAP with various thresholds. Distances are estimated via the PQ method and post-filtering is performed based on (a) Distance, (b) Rank, and (c) Ratio criterion. The best MAP and corresponding threshold is also shown. Where $th = \infty$, all schemes become identical to the pure VW-based method and the MAP score declines to 0.7455.

where

$$R(i, \mathrm{NN}(q)) = \begin{cases} 1 & \text{if } i \in \mathrm{NN}(q) \\ 0 & \text{else.} \end{cases} \quad (2)$$

The set of image identifiers $\mathrm{NN}(q)$ is determined by a filtering criterion and a threshold. As there are four ground truth images in the dataset, the inequality $0 \leq S \leq 4$ is satisfied. The repeatability score S monotonically increases as the threshold is increased. If we set the threshold to 0, S becomes identical to 0. Note that, even if we set the threshold to ∞, S does not become identical (or even close) to 4. This is because, in the VW-based framework, each query feature can be matched with only reference features in the same VW.

To see the differences among VWs with different frequencies, all VWs are classified into four groups: the top 25% most frequent VWs, the next 25% most frequent VWs and so on in descending order of frequency. The repeatability score S is calculated independently for features in the four groups. Figure 6 shows the repeatability scores of four VW groups, where distances are estimated by the PQ method and both the rank and ratio criteria are adopted. It is shown that, in the case of the rank criterion, the repeatability score is quite different depending on the VWs. Query features in frequent VWs are matched less with correct reference features than features in infrequent VWs as discussed in Section 3.1. In contrast, in the case of the ratio criterion, the unfairness is alleviated at all thresholds; features in different VWs have similar repeatability scores. This indicates that all features can equally contribute to scores in feature-level matching, resulting in considerable improvement in image search accuracy compared to the other criterion.

4.4. Impact of multiple assignment approaches

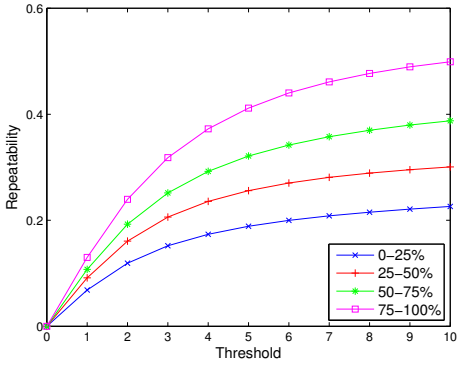
In this section, we evaluate combinations of the ratio criterion and multiple assignment approaches. The strategies A and B described in Section 3.3 are combined with both Fixed MA and Adaptive MA. Figure 7 shows the impact of the choice of multiple assignment approaches. We can see that the combination of Adaptive MA and strategy B is the best choice for ratio voting. In ratio voting, the weighting method [4] is also applicable according to the relative rank of reference features sorted by distances. A weight $\exp(-(t/N)^2/\sigma^2)$ is assigned to the t-th nearest reference feature in voting, where N denotes the number of features in the same VW as a query feature and σ denotes an adjustable parameter. Figure 7 also shows the effectiveness of the weighting method that is combined with Adaptive MA + B. We achieved a MAP score of 0.891 and Kentucky Score (KS)[4] of 3.47 where $th \geq 0.01$ and $\sigma = 0.05$, while a MAP score of 0.878 and KS of 3.42 were reported in [4]. We also confirmed the scalability of our system by adding the MIRFLICKR-1M dataset[5] as a distractor set, which includes 1 million images. The system based on the ratio criterion achieved a MAP score of 0.757 with $th = 0.005$, while the system based on the rank criterion achieved a MAP score of 0.752 with $th = 10$, where multiple assignment and weighting were not adopted for a fair comparison. Finally, we implemented a server-client system for mobile phones, where the mobile phone sends the captured image to the server and the server returns search results with geometric verification [5]. Figure 8 shows an example of the recognition results to a query. Our system accurately recognizes and localizes multiple images.

5. CONCLUSIONS

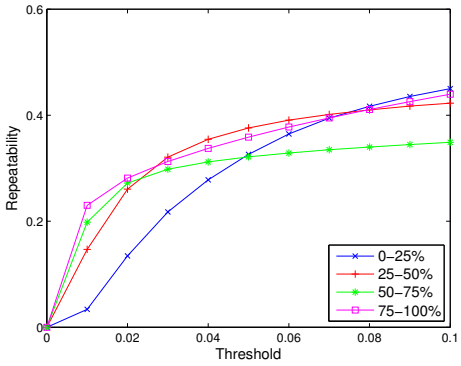
In this paper, we proposed a new voting strategy referred to as ratio voting to improve bag-of-visual words-based image retrieval. It limits the number of votes in proportion to the number of features in VWs, while conventional schemes use

[4]KS is the average number of relevant images ranked in top four positions when search results are sorted by scores [14].

[5]http://press.liacs.nl/mirflickr/



(a) Rank as a threshold



(b) Ratio as a threshold

Fig. 6. Repeatability scores for four VW groups.

(estimated) distances or rank information as a filtering criterion. Ratio voting realizes adaptive thresholding that captures the density of feature vectors. In experiments, we showed that ratio voting achieves a considerable improvement in spite of its simplicity.

6. REFERENCES

[1] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L. V. Gool, "A comparison of affine region detectors," *IJCV*, vol. 60, no. 1-2, pp. 43–72, Nov. 2005.

[2] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on PAMI*, vol. 27, no. 10, pp. 1615–1630, Oct. 2005.

[3] J. Sivic and A. Zissermane, "Video google: A text retrieval approach to object matching in videos," in *Proc. of ICCV*, 2003, pp. 1470–1477.

[4] H. Jégou, M. Douze, and C. Schmid, "Improving bag-of-features for large scale image search," *IJCV*, vol. 87, no. 3, pp. 316–336, 2010.

[5] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Proc. of CVPR*, 2007, pp. 1–8.

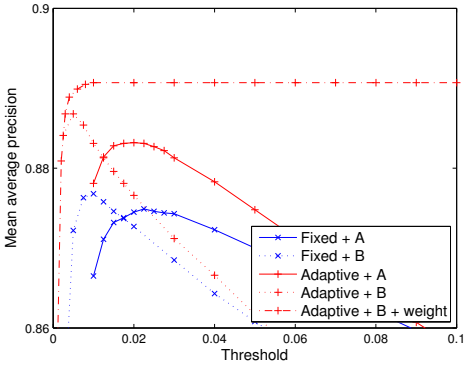


Fig. 7. MAP of different combination of strategies. Note that the best MAP without multiple assignment is 0.845.

Fig. 8. From left to right: a querying environment, a query image captured by a cell phone, and a recognition result.

[6] Y. Uchida, M. Agrawal, and S. Sakazawa, "Accurate content-based video copy detection with efficient feature indexing," in *Proc. of ICMR*, 2011.

[7] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Proc. of CVPR*, 2008.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[9] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool, "Surf: Speeded up robust features," *CVIU*, vol. 110, no. 3, pp. 346–359, 2008.

[10] H. Jégou, M. Douze, and C. Schmid, "Product quantization for nearest neighbor search," *IEEE Trans. on PAMI*, vol. 33, no. 1, pp. 117–128, 2011.

[11] Y. Weiss, A. Torralba, and R. Fergus, "Spectral hashing," in *Proc. of NIPS*, 2008, pp. 1753–1760.

[12] J. Brandt, "Transform coding for fast approximate nearest neighbor search in high dimensions," in *Proc. of CVPR*, 2010, pp. 1815–1822.

[13] J. Yang, Y. G. Jiang, A. G. Hauptmann, and C. W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proc. of MIR*, 2007.

[14] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. of CVPR*, 2006, pp. 2161–2168.

[15] L. Zelnik-Manor and P. Perona, "Self-tuning spectral clustering," in *Proc. of NIPS*, 2004, pp. 1601–1608.