Analysis of verification methods for indoor image matching

Nabeel Khan
Department of Computer Science
University of Otago, New Zealand
Email: nabeel@cs.otago.ac.nz

Brendan McCane
Department of Computer Science
University of Otago, New Zealand
Email: mccane@cs.otago.ac.nz

***Abstract*—This paper reports on experiments for indoor image-based location recognition. The basic method makes use of three stages: visual bag-of-words for ranking, a voting method, and a final verification method, if the voting method does not produce a consensus. Such a tiered approach is necessary when there are several visually similar locations in the image database, such as often occurs in office buildings. Three experiments are reported here. In the first, three common term-weighting schemes are compared: ntf, $ntfidf$ and $BM25$. Surprisingly ntf, the simplest scheme, is shown to be as accurate as $BM25$, and both are better than $ntfidf$. These results are surprising because $BM25$ has been experimentally shown to be one of the best weighting schemes for document information retrieval over many years, and $ntfidf$ has been the preferred weighting scheme for visual BoW in most other image retrieval work. In the second experiment, two verification methods are compared: one based on the fundamental matrix; and one based on a simpler homography computation. Again, surprisingly, the simpler and more efficient homography based method is shown to perform as well as the fundamental matrix method despite the fact that the fundamental matrix method is more physically plausible. The overall system achieves a recognition rate of approximately 80% with a wrong match rate of only 2% (no decision on 18%) on a very challenging office building data set. In the third experiment, the system is evaluated on the same office building dataset with more than one query image. A significant improvement is observed in localisation performance and the overall system achieves a recognition rate of 96% with only two wrong image matches.**

I. Introduction

Location recognition in an indoor environment is a challenging problem and has applications in robot navigation and as a navigation aid for the blind. The target application for this work is a navigation aid for the blind, and hence it is important that incorrect location returns are infrequent. For such applications it is preferable to return a "no match" result rather than an incorrect location, as it is always possible for the application to capture and use another query image.

Visual Bag of Words (BoW) has become a standard approach in image retrieval and image recognition problems, and this forms the basis for the work reported here. Sivic *et al.* [1] introduced the idea of visual BoW for the first time for object retrieval using inverse document frequency (idf) as the weighting scheme. Nister *et al.* [2] extended this work to large image databases using hierarchical clustering and term frequency-inverse document frequency ($tfidf$) weighting. The system is able to match an image in about one second. Filliat [3] used a two stage voting scheme for indoor matching. SIFT, hue and texture features are used for visual BoW and a room is recognized only if a quality threshold is reached. The work is tested on a small scale indoor environment and it has not been shown to work in office buildings where locations have similar color/texture schemes. Kang *et al.* [4] used visual BoW to perform matching on a large scale indoor office environment. The top eight images most similar to the query image are retrieved via visual BoW. A potential localisation is suggested if there is a cluster of pre-recorded images less than 3 meters from each other among the retrieved images. The system performance is evaluated on images from one floor of an office building. Robertson *et al.* [5] used homography and rectification for scene recognition in an outdoor environment. In their work, camera's are assumed calibrated (or at least approximately so), and database images are assumed rectified. Features are identified using the Harris corner detector and a RANSAC based algorithm for image registration is applied. The query image is matched against each database image and the closest match is returned as the location.

Aside from Kang *et al.* [4], few of these works focus on the difficult problem of image matching in large indoor environments with many visually similar locations. This paper is an extension of the work of Khan *et al.* [6] which proposed the use of a homography verification method with visual BoW. The current work extends that work by evaluating the $BM25$ term-weighting scheme, comparing the homography verification method to a more physically plausible fundamental matrix method and experimenting with multiple query images.

II. Overall System

In standard visual BoW, candidate images similar to the query image are retrieved and are ranked. The top ranked image is then selected as the best match. Visual BoW does not take into account the spatial configuration of features or other attributes (such as color) and this often leads to spurious matches. Nevertheless, the correct matching image is often in the top candidate matches, and incorporating a verification method in BoW should improve matching performance.

In the following, we assume that a suitable database of location-labeled images is available for training. From this trained database, we first extract 96D SIFT features [7] and use approximate K-means to cluster the features into cluster centers which are used as visual words. We have used seven different values of K in our experiments. The visual words make up the visual vocabulary and an inverted index is then developed to record which visual words occur in which images. All this is done off-line during the training phase.

For image classification, the inverted index is used to retrieve 200 trained images most similar to the query image. The histograms are then generated from the query and the top 200 candidate images using a suitable weighting scheme. The query and candidate histograms are compared to generate image rankings using the χ^2 measure [3] . If the top three ranked images refer to the same place, the location is simply returned without the need for verification. Otherwise, a verification method is used and the top 50 ranked images are evaluated one by one for a possible match. The verification method returns the location once it finds a match. However in case of no match, 'no decision' is returned.

A. Weighting Schemes

Reliable image rankings depend on effective histograms generated from the query and candidate images. The histograms represent the frequency information of the visual words in the image. A weighting scheme is normally used to generate these histograms from the images and the choice of weighting scheme varies from application to application. We have compared the following weighting schemes:

1) **Normalised Term Frequency (ntf) :** In ntf, each histogram bin refers to the normalised frequency count of visual words in an image [8].
2) **Normalised Term Frequency-Inverse Document Frequency ($ntfidf$):** In $ntfidf$, visual words which appear in many images are penalized and more weight is given to those words which appear in few images. For a vocabulary of size M, $ntfidf$ is computed as follows [8]:

$$t_{di} = \frac{n_{di}}{n_d}.\log\frac{N}{n_i}, \tag{1}$$

Where n_d is the total number of visual words in the image d, N is the total number of images and n_i is the number of images with visual word i.
3) **Okapi BM25 ($BM25$)**
Okapi BM25 is the best known probabilistic weighting scheme in information retrieval [9] and is commonly used for document retrieval [10], [11]. Given a query Q image with words $q_1, ..., q_n$, and a training image D, the BM25 score is computed as follows [9]:

$$R = \sum_{i=1}^{n} IDF(q_i)\frac{f(q_i, D).(k_i + 1)}{f(q_i, D) + k_1.(1 - b + b.\frac{|D|}{avgdl})} \tag{2}$$

where $f(q_i, D)$ is q_i's term frequency in the image D, $|D|$ is the number of visual words in image D and $avgdl$ is the average number of visual words per image in the collection. k_1 and b are free parameters and are set at $k_1 = 2$ and $b = 0.75$ for this work. IDF is the inverse document frequency of the query term q_i.

III. Verification Methods

Verification can be performed on every candidate image retrieved by visual BoW, but verification is an expensive operation. Hence the proposed system only performs verification if visual BoW fails to identify a consistent match in the top three ranked images. We have experimented with different numbers for the voting scheme, but no advantage was gained in using more than three.

Khan *et al.* compared a planar homography scheme against several local image information based verification methods and found planar homography method to be the superior method [6]. However, planar homography is not a good model of the world in general as many sets of points in a scene won't be planar. Here we introduce several variants of a more physically plausible fundamental matrix based method. We hypothesise that the fundamental matrix methods will be more accurate but may be less efficient than the homography method. First we describe the homography method and then the fundamental matrix methods.

A. Homography based (p-BoW)

The planar homography method is described in Figure 1. SIFT correspondences and RANSAC are both used to produce potential homographies which are verified using the homography transformation [6], [12] on the remaining SIFT keypoints. The justification for the algorithm is that although a planar homography is not expected to work for all feature correspondences, it should work for several correspondences especially indoors. The algorithm and its constants are chosen the same as in [6].

1: Find 10 best SIFT correspondences against query.
2: $numPerspective = 0$.
3: Use RANSAC for random picking of 4 SIFT correspondences 15 times.
4: **for all** sets of 4 SIFT correspondences **do**
5: Compute the transformation matrix.
6: Transform all candidate features to new locations.
7: Count the number of transformed features coming within a 3×3 window that also correspond (within a threshold distance of 150).
8: **if** COUNT$>= 3$ **then**
9: $numPerspective++$
10: **end if**
11: **end for**
12: **return** $numPerspective >= 3$

Fig. 1: Algorithm for *p-BoW* based on planar homography

B. Fundamental matrix based (fm-BoW)

The homography method uses structural information for planar elements of the scenes. *fm-BoW* on the other hand uses a full structure match between the query and candidate images. The algorithm starts by identifying a number of the best SIFT correspondences between the query and the candidate image followed by the computation of the fundamental matrix ([13]) via RANSAC. The fundamental matrix is a 3x3 matrix which relates corresponding points in two images of the same scene. The general algorithm estimates the fundamental matrix and corresponding inlier points from the candidate points determined by SIFT correspondences. Inlier points are those points which match according to the fundamental matrix equation:

$$\mathbf{x}'^T\mathbf{F}\mathbf{x} < \epsilon, \tag{3}$$

where x' and x are points in two images, and ϵ is an empricially determined threshold. A larger number of inliers indicates a more reliable image match. The *fm-BoW* variants tested are described below:

fm1-BoW
1) Find top 25% SIFT correspondences between query and the candidate image.
2) If query and candidate have at least 25% inliers; consider it a best match.

fm2-BoW
1) Find top 25% SIFT correspondences between query and the candidate image.
2) If query and candidate have at least 40% inliers; consider it a best match.

fm3-BoW
1) Find top 50% SIFT corres. against query.
2) At least 20% inliers; consider it a best match.

fm4-BoW
1) Find top 10 SIFT corres. against query.
2) At least 20% inliers; consider it a best match.

fm5-BoW
1) Find top 10 SIFT correspondences against query.
2) At least 75% inliers; consider it a best match.

fm6-BoW
1) Find top 30 SIFT correspondences against query.
2) At least 20% inliers; consider it a best match.

IV. Results

The data sets and performance measures are briefly described in Section IV-A followed by the experimental results. All experiments are performed on a single core of 3.6 GHz Intel Core 2 Duo machine.

A. Data Sets and Performance Measures

We have used three data sets and have selected subsets of images from these data sets to get a reasonable number of trained features for a fair comparison of the three weighting schemes.

1) **David Nister (DN):** This data set contains 4 different images of 2500 objects/scenes [2]. We have used the first 4000 images with 3000 for training and 1000 for testing. About 0.5 million trained features are extracted.
2) **Hongwen Dataset (IE):** This data set contains 8000 images of an office environment taken over some period of time [4]. We have used the first 3000 images for training and 1000 images for testing. About 1.4 million trained features are extracted.
3) **Indoor Dataset(CS):** This dataset contains 700 indoor images of an office building with offices and some classroom sized computer laboratories [6] — many different locations within the building look very similar. 70 images are used for testing and 630 for training. About 0.17 million trained features are extracted. Since it is a smaller data set, 15-fold cross-validation with different test and training sets chosen randomly is performed for each experiment.

The following definitions are used to define the performance metrics:

Q_t	Total number of query images.
V_t	Total number of images passed to the verification method.
N_v	Number of images correctly matched by the voting scheme.
M_v	Number of images correctly matched by the verification method: $M_v \leq V_t$.
N_d	Number of images for which no decision is made.

The following evaluation metrics are used in our work:

C_a is the correct acceptance rate. Higher is better for this metric:

$$C_a = (N_v + M_v)/Q_t \tag{4}$$

W_c is the consistency of the weighting scheme:

$$W_c = (Q_t - V_t)/Q_t \tag{5}$$

Higher is better for this metric, but must be used in conjunction with N_c.

N_c is the matching accuracy of images not passed to the verification method. Higher is better for this metric and indicates the success of weighting scheme.

$$N_c = N_v/(Q_t - V_t). \tag{6}$$

W_m is the wrong match rate. Lower is better for this metric.

$$W_m = 1.0 - \frac{(N_v + M_v) + N_d}{Q_t} \tag{7}$$

R_{nd} is no-decision rate. Lower is better for this metric however it must be used in conjunction with W_m.

$$R_{nd} = N_d/V_t. \tag{8}$$

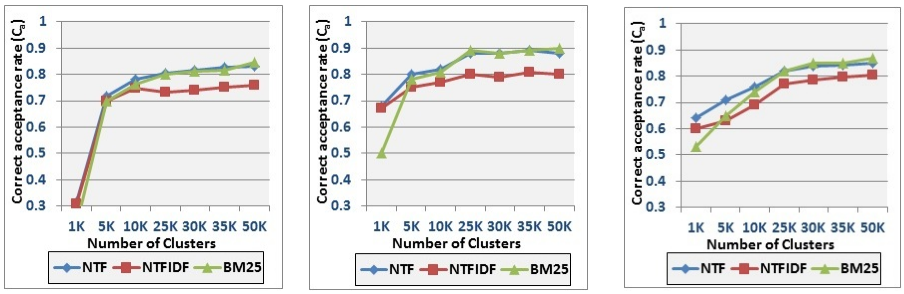
B. Comparison of Weighting Schemes

We have evaluated the performance of standard visual BoW with all weighting schemes mentioned in Section II-A across all data sets and results are shown in Figure 2. For this experiment, the top ranked image is considered the best match for the query image. The main purpose of this evaluation is to identify the best weighting scheme for such applications.

A good weighting scheme should generate good image rankings and should result in fewer calls to the verification methods. Results in Table I show that $ntfidf$ scheme performs worst in regards to consistency but that the matching accuracy is equally good for all schemes. Table I shows that overall ntf and $BM25$ schemes perform equally well. However, since ntf is more efficient it is used in the following experiments.

C. fm-BoW vs p-BoW Analysis

Results of the comparison between *p-BoW* and all *fm-BoW* variants are shown in Figure 3 (with the ntf scheme). *fm1-BoW* is found to give the best performance among its variants and is comparable with *p-BoW* as shown in Table II. Results show that *p-BoW* and *fm1-BoW* both provide reasonable C_a with a very low W_m. There is not much to differentiate between *p-BoW* and *fm1-BoW* as both can be configured to give almost the same classification performance.



(a) DN Dataset. (b) CS Dataset. (c) IE Dataset.

Fig. 2: The correct acceptance rates C_a for all weighting schemes across all data sets using standard visual BoW.

TABLE I: Weighting schemes analysis.

Weighting schemes consistency W_c							
	1K	5K	10K	25K	30K	35K	50K
NTF	0.28	0.47	0.51	0.58	0.60	0.59	0.62
NTFIDF	0.27	0.42	0.43	0.44	0.47	0.47	0.47
BM25	0.17	0.39	0.48	0.57	0.58	0.57	0.58
Weighting schemes accuracy N_c							
	1K	5K	10K	25K	30K	35K	50K
NTF	0.92	0.97	0.99	1	1	0.99	0.99
NTFIDF	0.95	0.98	1	1	1	1	1
BM25	0.78	0.99	0.98	0.99	1	0.99	0.99

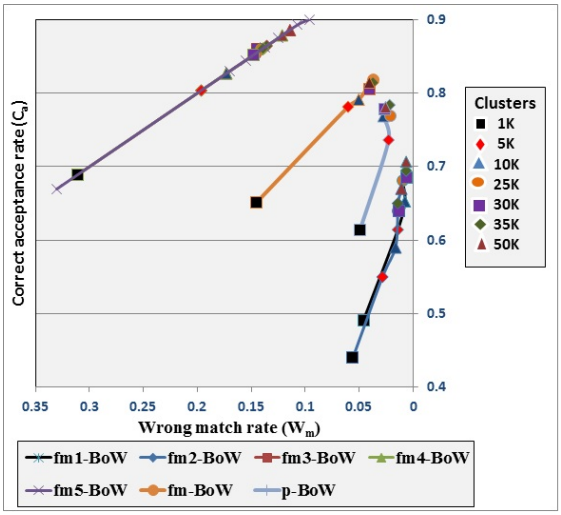


Fig. 3: *fm-BoW* variants performance against the *p-BoW*. The curves that are higher and further to the right are better.

TABLE II: Performance comparison of *fm1-BoW* and *p-BoW*

		1K	5K	10K	25K	30K	35K	50K
p-BoW	**C_a**	0.61	0.74	0.77	0.77	0.78	0.78	0.78
	W_m	0.05	0.02	0.03	0.02	0.03	0.02	0.03
fm1-BoW	**C_a**	0.65	0.78	0.79	0.82	0.81	0.81	0.81
	W_m	0.15	0.06	0.05	0.04	0.04	0.04	0.04

1) Computational Time: We have also analyzed the computational time of the fundamental matrix (FM) and homography verification methods for matching two images. We have run the methods several times and the average results are recorded in Table III. The results show that:

1) With RANSAC, *FM* verification method is not as efficient as homography method.
2) Without RANSAC option, only 8 correspondences are picked for the *FM* method which makes it efficient but then it gives poor matching performance.

TABLE III: Average time in seconds for verification methods.

	FM RANSAC	FM No RANSAC	Homography RANSAC
Time	**0.175 sec.**	**0.022 sec.**	**0.071 sec.**

D. Image matching with multiple query images

We have computed the performance of our system with multiple query images and have compared it with our system when it uses only one query image (i.e *p-BoW* and *fm1-BoW*). Both *p-BoW* and *fm1-BoW* match a query image in about 0.91 seconds on average including retrieval, voting and verification. We report the average time to match a query image of the system with 25K clusters in this section.

We designed one test and one training set. Our test set includes four query images against each location and we call them q1 (query1), q2 (query2), q3 (query3) and q4 (query). All query images differ from each other and do not exist in the training set. We have conducted the following experiments:

Using two query images (pB-BoW & fmB-BoW)

The system reports the results on a second query image if it cannot make a localisation decision on the first query image. Results in Table IV show that *pB-BoW* and *fmB-BoW* both achieve a C_a of 96% and 95% respectively because the system finds matches against the second query image most of the time. The W_m for *pB-BoW* and *fmB-BoW* is about 3% and 4% respectively. The system matches a query image in 1 second on average.

Using four query images

pC-BoW & fmC-BoW: The system reports the results on three remaining query images if it fails to make a decision on q1. A localisation decision is made only if q2, q3 and q4 all refer to the same location. Table IV shows that a C_a of 94% is obtained with both *pC-BoW* and *fmC-BoW*, with a W_m of 3%

TABLE IV: Performance with multiple query images. Note that *p-BoW* and *fm1-BoW* results are different to Table II because of different data sets.

Clusters		1K	25K	50K
One query image				
p-BoW	C_a	0.70	0.81	0.83
	W_m	0.06	0.06	0.03
	R_nd	0.24	0.13	0.14
fm1-BoW	C_a	0.77	0.86	0.89
	W_m	0.10	0.04	0.03
	R_nd	0.13	0.10	0.08

Clusters		1K	25K	50K
Two query images (with rejected images)				
pB-BoW	C_a	**0.90**	**0.90**	**0.96**
	W_m	0.07	0.09	0.03
	R_nd	0.03	0.01	0.01
fmB-BoW	C_a	**0.86**	**0.91**	**0.95**
	W_m	0.12	0.07	0.04
	R_nd	0.02	0.02	0.01

Clusters		1K	25K	50K
Four query images (with rejected images)				
pC-BoW	C_a	**0.83**	**0.87**	**0.94**
	W_m	0.06	0.09	0.03
	R_nd	0.11	0.04	0.03
fmC-BoW	C_a	**0.86**	**0.89**	**0.94**
	W_m	0.09	0.07	0.04
	R_nd	0.05	0.04	0.02

Clusters		1K	25K	50K
Two query images				
pD-BoW	C_a	0.64	0.76	0.77
	W_m	**0.02**	**0.01**	**0.01**
	R_nd	0.34	0.23	0.22
fmD-BoW	C_a	0.70	0.81	0.83
	W_m	**0.01**	**0**	**0**
	R_nd	0.29	0.19	0.17

Clusters		1K	25K	50K
Three query images				
pE-BoW	C_a	0.58	0.71	0.72
	W_m	**0.01**	**0.01**	**0.01**
	R_nd	0.41	0.28	0.27
fmE-BoW	C_a	0.64	0.77	0.79
	W_m	**0.01**	**0**	**0**
	R_nd	0.35	0.23	0.21

Clusters		1K	25K	50K
Four query images				
pE-BoW	C_a	0.53	0.67	0.69
	W_m	**0.01**	**0.01**	**0.01**
	R_nd	0.46	0.32	0.30
fmE-BoW	C_a	0.57	0.71	0.74
	W_m	**0**	**0**	**0**
	R_nd	0.43	0.29	0.26

and 4% for *pC-BoW* and *fmC-BoW* respectively. The system matches a query image in about 1.2 seconds on average.

pD-BoW & fmD-BoW: The system reports the results on two query images. A localisation decision is made if two queries refer to the same location, otherwise a decision is not made. The system matches the two query image in 1.8 seconds.

pE-BoW & fmE-BoW: The system reports the results on three query images in about 2.9 seconds. A localisation decision is made only if three query images refer to the same location.

pF-BoW & fmF-BoW: The system reports the results on all query images in about 3.8 seconds. A localisation decision is made only if q1,q2,q3 and q4 all refer to the same location.

V. Conclusion

In this paper we first compared three weighting schemes for visual BoW location recognition. Surprisingly, we found that the simplest scheme, ntf, was as good as the more sophisticated $BM25$ scheme. Both ntf and $BM25$ were superior to the $ntfidf$ scheme. These results are surprising as the $BM25$ scheme is the standard baseline comparison for document-retrieval experiments, and the $ntfidf$ scheme is the default weighting scheme used for visual BoW. It is unclear why ntf performs so well in these experiments, but we suspect it is related to the specific nature of the localisation problem with many images for a relatively small number of locations (numbering in the tens rather than thousands).

Secondly, we introduced a verification method based on computation of the fundamental matrix between query and candidate images and compared it with previous work based on homography. We hypotheised that the fundamental matrix method would be more accurate but less efficient than the planar homography method. We found that it was indeed less efficient, but surprisingly not much more accurate. However, the extra processing cost is minimal compared to the whole retrieval system when used in a tiered fashion. Although the fundamental matrix method is more physically plausible, for indoor localisation, there are many planar surfaces, and we strongly suspect that this allows the simpler planar homography method to do as well as the fundamental matrix method. We believe that these results would transfer to many built environments, but probably not to less structured ones. Thirdly, we tested our system performance for indoor image matching with multiple query images. Unsurprisingly, using multiple query images improved accuracy significantly, but there was a trade-off between correct acceptance rate (83% to 96%), wrong match rate (3% to 0%), and no decision rate (30% to 1%) for 50K clusters. Which configuration to choose depends entirely on the application. For blind users, minimising the wrong match rate is key, which means reducing the correct acceptance rate and increasing the no decision rate.

References

[1] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos.," in *International Conference on Computer Vision (ICCV)*, pp. 1470–1477, 2003.

[2] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. 2161–2168, 2006.

[3] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping.," in *International Conference on Robotics and Automation*, pp. 3921–3926, 2007.

[4] M. H. H. Kang, A. Efros and T. Kanade, "Image matching in large scale indoor environment," in *Proc. CVPR*, pp. 33–40, 2009.

[5] D. Robertson and R. Cipolla, "An image-based system for urban navigation," in *British Machine Vision Conference*, pp. 819–828, 2004.

[6] N. Khan, B. McCane, and G. Wyvill, "Homography based visual bag of word model for scene matching in indoor environments.," in *International Conference on Image and Vision Computing (IVCNZ)*, pp. 70–75, 2011.

[7] N. Khan, B. McCane, and G. Wyvill, "Sift and surf performance evaluation against various image deformations on benchmark dataset.," in *Proc. Digital Image Computing and Techniques*, pp. 501–506, 2011.

[8] Y. Jiang, C. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *International conference on Image and video retrieval.*, pp. 494–501, 2007.

[9] K. Jones, S. Walker, and S. Robertson, "A probabilistic model of information retrieval: development and comparative experiments: Part 2," *Information Processing and Management*, pp. 809–840, 2000.

[10] C. D. Manning, P. Raghavan, and H. Schtze, *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008.

[11] S. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, and M. Gatford, "Okapi at trec-3," pp. 109–126, 1996.

[12] G. Wollberg, *Digital Image Warping*. IEEE Computer Society, 1994.

[13] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, second ed., 2004.