Homography based Visual Bag of Word Model for Scene Matching in Indoor Environments

Nabeel Younus Khan
Computer Science Department
Otago University, NZ
Email: nabeel@cs.otago.ac.nz

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

Geoff Wyvill
Computer Science Department
Otago University, NZ
Email:geoff@cs.otago.ac.nz

***Abstract*—This paper proposes a data driven approach to perform scene localization in indoor environments. The proposed algorithm named p-BoW is designed to cope with self-repetitive and confusing patterns in indoor environments of any type. The algorithm uses the Visual Bag of Words (BoW) model along with proposed voting scheme to perform scene localization from a database of captured images. In the first phase, a small subset of images closer to the query image is found via standard BoW. In the second phase, verification is performed (if required) to identify the best matched image from this subset against the query image.**

ntf weighting scheme has been found to outperform $ntfidf$ scheme in matching precision. Proposed algorithm makes use of visual BoW based on SIFT features in the first phase and perspective transformation during the verification phase for image matching. The resulting proposed system has been able to perform scene matching efficiently in our indoor environment (having about 35 indoor locations) with an accuracy of more than 91% on different cluster sizes.

I. Introduction

Localization in an indoor environment without any global positioning information (GPS) is a challenging problem. Indoor localization is a basic requirement for navigation of robots and blind people. This area has been widely studied by robotics and computer vision researchers. In robotics, a system able to autonomously build a map while estimating its position in the indoor environment is the optimal solution and is referred as Simultaneous Localization and Mapping (SLAM) [1]. A key part of most of the SLAM implementations is the detection of previously visited locations by the robot [2]. The target application for this work is a vision-based navigation system for blind people in indoor environments using common hardware (preferably smart phones). The first stage in this process is location recognition. We envisage that a user will take a photo using their smart phone, the phone may perform some initial processing then send a representation of the image to a server for location recognition which is then communicated back to the phone. In this paper, we focus purely on the location recognition problem and assume that a suitable database of labeled images has already been collected for the building in question. Such a system would be particularly suitable for office buildings.

A. Related Work

Visual Bag of Word Model (BoW) has recently been used for recognition of scenes and video event analysis due to its robustness and good accuracy. There is a substantial amount of research work in the area of image retrieval but most of it focuses on outdoor images.

Few researchers have worked on indoor environments [3]–[5]. In [3], SIFT, hue and texture features are used for visual BoW followed by a voting scheme to perform scene localization. Although it has been tested for a small scale indoor environment it has not been shown to work in office buildings which have similar color/texture schemes in many places. In [4], the idf weighting scheme based on global and local statistics is used to perform scene localization in large scale indoor office environments. Their algorithm is efficient and robust. The algorithm utilizes the trained images distances information to identify the best match from top 8 image matches.

More recently, object detection and probabilistic semantics have been used in a small scale indoor environment to identify the place type [5]. Their work should be applicable to indoor environments of any type/scale but will perform slower as objects segmentation, objects classification and then use of semantics is usually slow.

The closest work to that described in this paper is proposed in [6]. In that 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.

In some ways, our work can be viewed as an extension of that of [6]. On the one hand, we update their approach using SIFT features and the BoW algorithm. We also note that plane homographies can be used in many environments and not just building facades as with [6]. In our case, only coarse localisation is necessary and therefore we do not require camera calibration. Finally, we show that the normalised term frequency (ntf) weighting scheme is superior for this dataset than the more popular normalised term frequency - inverse document frequence ($ntfidf$) scheme.

B. Approach

We propose a planar homography based Bag of Word Model (p-BoW) in this paper. p-BoW can perform scene localization efficiently in an indoor environment with good precision. The environment is represented as an image based topological map [7], [8]. We have developed a database of indoor images of our building for evaluation purposes [9] . It is a standard office-type building with some classroom size computer laboratories — many different locations within the building look very similar. Figure 1 shows an example of our system in which a captured indoor image is matched against the stored images for localization. Our p-BoW works requires only database of labeled images and works efficiently. Section 2 of paper discusses the first phase which uses traditional BoW based on two weighting schemes. Then in Section 3, we discuss the validation techniques for verifying the potential candidates. Results are presented in Section 4 followed by conclusions in Section 5.

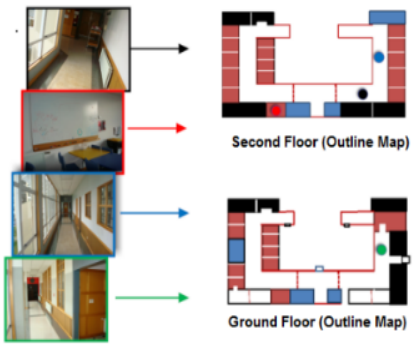

Fig. 1. Scene Localization by the system. Colored circles indicate the identified indoor places against the corresponding input images.

II. Image Matching Phase

A. Key points Extraction

Features or key points from the images can be detected either as salient patches using Harris, Laplacian, DoG or Maximal Stable Extreme Region [10]–[13] or by visual descriptors using SIFT, SURF, PCA-SIFT [12], [14], [15].

We have used SIFT descriptors in our work. The SIFT implementation is our own and we have used shorter 96 dimensional SIFT descriptors. In our proposed SIFT, the image size is kept constant and some orientation values are skipped from 4 x 4 orientation rows obtained in the "Key-point descriptor" phase resulting in shorter 96D SIFT features [16]. These shorter SIFT descriptors are found to give almost the same classification accuracy on different benchmark datasets as we get with 128D SIFT features with the benefit of being twice as fast with lower memory requirements.

B. Vocabulary Building

We build the vocabulary by applying clustering on the extracted SIFT features from the training images. Different clustering techniques have been used by researchers such as k-mediods, hierarchical clustering etc [3], [4], [17]–[20]. For large scale databases, a very large vocabulary is needed and time complexity will be high. We have used Approximate k-means clustering (AKM) and have performed nearest neighbor search by kd trees [21] to reduce the time complexity to $N \log k$ where N is the number of features and k is the number of clusters. AKM has been reported to be superior than HKM in performance [20].

Vocabulary can be built using hard and soft assignment of words. In hard assignment, features are mapped to one closest cluster during clustering while in soft, features are mapped to multiple nearest cluster centers. In practice hard assignment may lead to errors because of variability in the feature descriptor such as image noise, varying scene illumination etc. This may result in the same surface patch being assigned to different visual words in different images.

We have used both hard and soft assignment of visual words in our experiments. People have used different vocabulary sizes ranging from 0.8K to 1M [3], [18]–[20], [22], [23] but trade off between discrimination and generalization motivates the use of appropriate dictionary size [24]. We have tested 7 vocabulary sizes ranging from small to large ones i.e. 1K to 50K.

C. Keywords Weighting Scheme

We obtain word distributions or histograms for every training image via a weighting scheme. We have used two weighting schemes: normalized term frequency (ntf); normalized term frequency-inverse document frequency $(ntfidf)$ [24].

In term frequency, each histogram bin refers to the actual count of visual words in an image d. With vocabulary size of K and n_d as the total number of visual words in the image d, we use a K-bin histogram $T_d = [t_{d1}, t_{d2}..t_{dK}]$ where each histogram bin refers to a normalized frequency count i.e. $t_{di} = n_{di}/n_d$. Normalization eliminates the difference between short and long documents.

In $ntfidf$, we penalize visual words which appear in many images and give more weight to those words which appear in few images. For a vocabulary of size K, normalized $ntfidf$ can be computed as follows:

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

where N is the total number of images and n_i is the number of images having visual word i.

D. Classification via Inverted Indexing

To classify a query image, we are using the inverted indexing scheme to quickly retrieve 200 trained images which could be similar in appearance to the query image. We then compare the histograms of the query image against the obtained images using either ntf or $ntfidf$ to form a ranked list of potential candidate images. Image with highest rank can be considered to be the best match. Histograms are compared using the χ^2 distance. χ^2 can be computed between two histograms, $(H1)$ and $(H2)$, as follows [3]:

$$||H_1 - H_2||^2 = \sum \frac{(H_{1,i} - H_{2,i})^2}{H_{1,i} + H_{2,i}} \tag{2}$$

E. Weighting Schemes Analysis

Most people have preferred $ntfidf$ weighting scheme in their works [3], [4], [17]–[20]. We analyzed from experiments that ntf scheme performs better on our indoor dataset. We decided to use other indoor and outdoor datasets to verify our hypothesis. We have evaluated our simple BoW with both weighting schemes on the following benchmark datasets:

1) **UK Benchmark (U.K.B):** Contains outdoor images and is used as a standard for classification tasks [18]. There are 4 different images of 2500 objects i.e.10,000 images in total. We have used the first 4000 images i.e. 3000 for training and 1000 for testing. The 1st image is used for testing and remaining object images are used for training. 3000 trained images have been used to ensure there are a large number of trained SIFT features i.e. 0.99M for reasonable BoW analysis.
2) **Indoor Environment (I.E):** Contains images of indoor environments (i.e. from one floor of a building having official setup) taken over some period of time [4], [25]. There are 8000 images for training and 100 images for testing. We have used 3000 out of 8000 images for training and all 100 images for testing in our experiments. A reasonable number of SIFT features are extracted from trained data i.e. 14M.
3) **CS Indoor (CS):** Contains indoor images of our building. Details are stated in Section IV.

Figure 2 shows that traditional BoW with ntf scheme has performed better then $ntfidf$ on all datasets thus making ntf scheme more feasible.

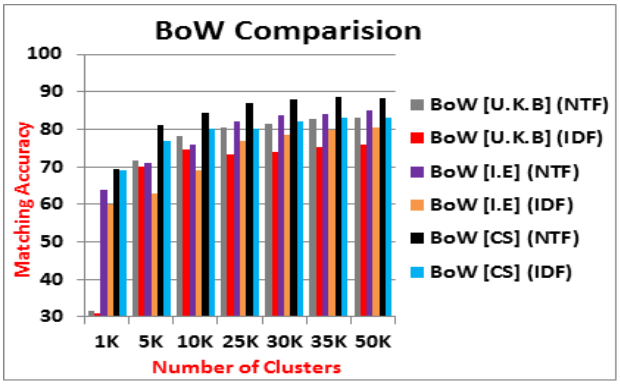

Fig. 2. Traditional BoW Matching Performance on different datasets.

III. Verification Phase based on Homography

In the simple BoW model there is no verification phase. The 1st candidate image in the top 200 images (subset) is considered to be the best match for the query image. However, the simple BoW model does not take into account the spatial configuration of features or other image attributes (such as colour) and this often leads to spurious matches. Nevertheless, the correct matching image is often in the top few candidate matches, and incorporating a verification phase should significantly improve performance.

The proposed verification algorithm works as follows:

1) Use Inverted Indexing to retrieve top 200 images.
2) Calculate ntf or $ntfidf$ histograms of 200 images.
3) Compute image rankings using χ^2 measure.
4) Take top 50 ranked images in increasing order.
5) If top 3 ranked images refer/vote to same location:
   a) **Best Case:** Then return that location.
6) Else **[Worst Case Scenario]**
   a) Perform validation on top 50 images one by one.
   b) If any image matches with query; return location.
   c) if no match found in 50 images; this means "no decision" i.e. no match found.

Inverted indexing helps in efficient retrieval of images similar to the query image. In our proposed scheme, if any of top 3 images disagrees, we perform validation on top 50 images. Validation is performed on every image one by one (if required). We compute best SIFT correspondences between query and top ranked candidate image. We use RANSAC to pick four of these correspondences randomly and compute a plane homography. We then transform the candidate features to new locations. If both images are the same, then most of the transformed features will be in the field of view and will be approximately at the same location as corresponding features are. But this may be the case sometimes when two images are different. So we identify such transformed features and check the feature similarity with corresponding query features to correctly identify the perspective correspondences. If we find sufficient number of perspective correspondences with different homographies then we select the current candidate image as the best match. Otherwise, we pick next candidate image and repeat whole process. Details of the algorithm are given in Section III-C.

Validation can be performed for every query image but it is expensive operation. Incorporation of voting results in efficiency gains. As in many cases, we get localization decision via votes. Any number of top images rather then 3 can be picked to check the votes. We experimented different number of images for voting (i.e. 5, 7 and 9) and identified that top 3 images should be picked because this configuration gives best results i.e. more then 99% accuracy on average. Other configurations yield more wrong matches thus making them less feasible. So proposed verification algorithm is a reasonable trade-off between accuracy and efficiency.

We have compared our homography based validation techniques with two other proposed validation techniques i.e. sift-distance based (sd-BoW) and selective-hue matching (sh-BoW) for performance comparison. Algorithms for all validation techniques are as follows:

A. sift-distance Validation (s-BoW)

1) For each candidate image:
   a) Compute SIFT correspondences with query image.
   b) Use 150 threshold to compute features similarity.
2) If correspondences $>= 3$; return location.
3) Else pick next top ranked image and repeat steps (1-2).

4) If no match found in 50 images; refers to 'no decision'.

B. selective-hue Validation (sh-BoW)

1) We use spatial information. The candidate image will be considered best match only:
   a) IF its SIFT Correspondences are $>= 3$ against query image.
   b) AND also if its Hue Histogram $=$ Query Hue Histogram by at least 50%.
2) If no match found, use next image and repeat steps (1-2).
3) If no match found in 50 images; refers to 'no decision'.
4) For Hue Computation:
   a) We have used 5 x 5 regions around key-points.
   b) To check histograms similarity, we use χ^2 measure.

C. Homography Validation (p-BoW)

1) For each candidate image:
   a) Find 10 best SIFT correspondences against query.
2) Declare $numPerspective = 0$.
3) Use RANSAC for random picking of 4 SIFT correspondences 15 times.
4) For every set of 4 SIFT correspondences:
   a) Compute transformation matrix.
   b) Transform all candidate features to new locations.
   c) Note the transformed features coming with in 3 x 3 window of corresponding query features.
   d) Check all such features similarity with query features (150 threshold) and record the "COUNT".
   e) If COUNT$>= 3$;
      i) return $numPerspective++$; else return 0.
5) If $numPerspective >= 3$;
   a) return candidate location.
   b) Otherwise pick next image and do steps (1-5).
6) If no match found in 50 images; refers to 'no decision'.
7) **Note:** [The 4 points in both images which are used to compute homography are excluded in step 4 (c-d)].

IV. Experimental Results

A. Dataset

Our indoor dataset contains about 700 images taken from 35 places over three floors of the building [9]. 70 images are used for testing and 630 for training. Test images contain images of every place. 15-fold cross-validation with different test and training sets is performed to compute average performance. About 0.17M SIFT features are extracted on average. The dataset is quite challenging because of the similarity of many locations and is therefore a good test set for many localization problems.

B. Proposed BoW Analysis

We have evaluated the performance of our proposed validation techniques against the simple BoW with ntf and $ntfidf$ schemes to identify the best among them. For evaluation, we have used seven cluster sizes in experiments. and have run simple BoW and all proposed BoW 15 times on our dataset. The average matching precision and standard deviations are shown in Figures 3 and 4. Results indicate that proposed p-BoW performs best in terms of matching precision and stability (due to its low standard deviation) as compared to simple BoW and proposed s-BoW, sh-BoW. sh-BoW performs slightly better then s-BoW and appears to be more stable due to incorporation of color information as shown in Figure 4.

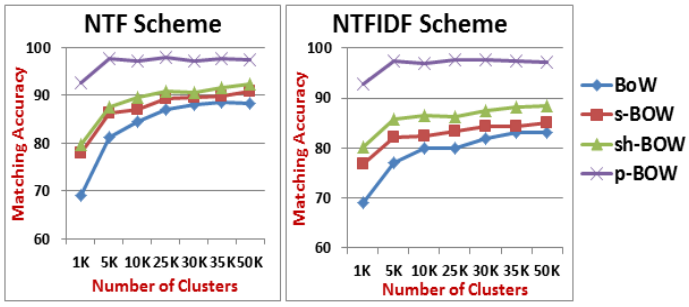

Fig. 3. Simple and proposed BoW Models Performance Evaluation with seven clusters on our indoor dataset

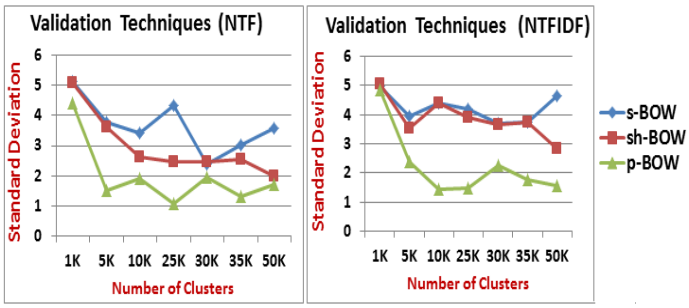

Fig. 4. Proposed BoW Models Standard Deviation for ntf and $ntfidf$

C. Weighting Schemes Analysis

We analyzed that for ntf scheme, worst case scenario i.e. frequency of invoking of validation technique (in %) is less then $ntfidf$. For very small cluster sizes, worst case occurrence is similar for both however with increase in cluster sizes ntf scheme holds and performs better then $ntfidf$ as shown in Table I. This also supports our hypothesis in Section II that ntf scheme performs better then $ntfidf$.

TABLE I
WORST CASE OCCURRENCE FOR BOTH ntf AND $ntfidf$ SCHEMES

	1K	5K	10K	25K	30K	35K	50K
NTFIDF	76	60	57	57	52	51	53
NTF	76	58	48	43	42	41	40

D. Validation Techniques Analysis

Some query images contain less discriminatory information and are difficult to recognize even with human eye. Such images are called 'confusing' and algorithms mostly find wrong matches against such images. p-BoW also performs some wrong matches. Sometimes the central hall of 3rd floor

is matched with the central hall of 2nd floor, rooms may be wrongly matched, corridors of different floors may be mismatched etc. Some examples of such wrong matches and confusing query images are shown in Figure 5.

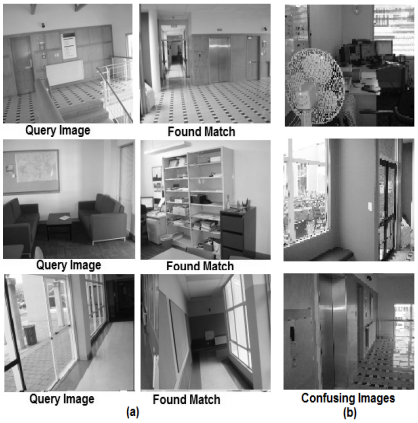

Fig. 5. p-BoW (a) Wrong matches (b) Confusing Images

For validation techniques analysis, we have considered only those query images for which worst case happens and validation techniques are invoked to find the best match. Proposed validation techniques have been evaluated in terms of:

1) **No Decision Rate (R.R):** Number of query images against which validation cannot find best match.
2) **Correct acceptance rate (C.A):** Ratio of correct matches i.e. average matching accuracy.

We evaluated average performance of all techniques w.r.t above parameters. Figure 6 clearly shows that the perspective transform validation technique outperforms others tested here for both weighting schemes. For 'confusing query images', the top 50 images may not contain the desired match against the query image. s-BoW and sh-BoW do not handle such cases well and normally finds the wrong match resulting in lower accuracy. For sh-BoW, the R.R is a bit higher then s-BoW. The incorporation of hue and SIFT features make this validation technique better then s-BoW. On the other hand, in p-BoW use of homography ensures avoidance of most of the wrong matches. Resulting R.R is higher for p-BoW which prevents many wrong matches against confusing query images resulting in better accuracy.

The R.R of homography validation technique can be reduced at the cost of computation by applying validation on more than 50 images to find the desired match.

E. Scene Confusion Matrix

For a blind person, it is important that they know the type of place in which they are present even if specific location information is not available. A scene confusion matrix has been developed for 25K cluster configuration for both p-BoW ntf and $ntfidf$ as shown in Table II that groups place type as opposed to simply location. As can be seen, the type of place is recognized extremely well. These results are quite good and are comparable with results shown in [5].

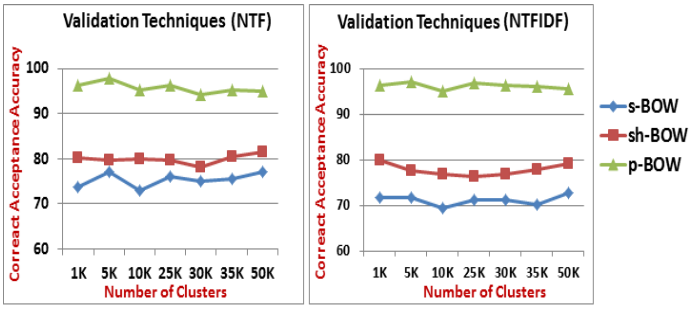

Fig. 6. Validation Techniques (s-BoW, sh-BoW and p-BoW) Average Correct Acceptance Rate for ntf and $ntfidf$ schemes

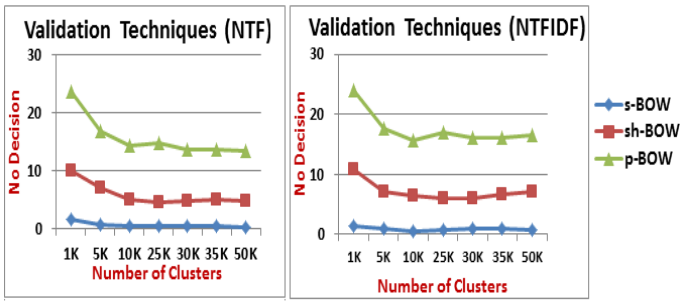

Fig. 7. Validation Techniques (s-BoW, sh-BoW and p-BoW) 'No Decision' Rate for ntf and $ntfidf$ schemes

F. Soft vs Hard Assignment

We experimented with soft assignment of visual words in simple BoW for both schemes. We mapped features to 2, 3 and 5 nearest cluster centers resulting in 3 BoW models i.e. S, S1 and S2. We compared proposed soft assignment based BoW performance with BoW based on hard assignment i.e. H-BoW. Results in Figure 8 show that soft assignment does not make a significant difference for ntf scheme as compared to hard one. However soft assignment resulted in 1-2% improvement for $ntfidf$ scheme for larger clusters. This makes soft assignment feasible for $ntfidf$ scheme. But soft assignment increases the computational cost and its difficult to determine the number of nearest cluster centers for mapping, as we find no significant pattern via choosing different number of cluster centers.

V. Conclusion

In real time applications, a blind person can be asked to take 4-5 pictures of current scene and place recognized mostly by p-BoW can be considered to be the desired location. This will further reduce the chances of wrong matches and system will be able to perform very precisely in real time.

We have presented in this paper a localization technique which can work very well in challenging indoor environments. The major findings of this work are:

1) Use of homography improves validation of candidate matches significantly.
2) p-BoW is reproducible and robust.
3) ntf weighting scheme is superior to the $ntfidf$ scheme for indoor and outdoor environments.

TABLE II

SCENE CONFUSION MATRIX FOR P-BOW FOR 25K CLUSTERS(ntf AND $ntfidf$ SCHEMES). LEGEND: AL, ALL LABS; CR, CONFERENCE ROOM; COR, COFFEE ROOM; C, CORRIDORS; H, HALLS; W, WASHROOM; O, OFFICES.

	25K NTF							25K NTFIDF						
	AL	CR	CoR	C	H	W	O	A.L	C.R.	CoR	C	H	W	O
All labs	99%	0.70%	0%	0%	0%	0%	0.3%	99%	1%	0%	0%	0%	0%	0%
Conf. Rm	0%	100%	0%	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%
Coffee Rm	0%	0%	97.8%	0%	0%	0%	2.2%	0%	0%	100%	0%	0%	0%	0%
Corridors	0%	0%	0%	99.7%	0%	0%	0.3%	0%	0%	0%	100%	0%	0%	0%
Halls	0%	0%	0%	0%	100%	0%	0%	3%	0%	0%	1%	96%	0%	0%
Washroom	0%	0%	0%	0%	0%	100%	0%	0%	0%	0%	0%	0%	100%	0%
Offices	0%	0%	0%	2%	0%	0%	98%	2%	0%	0%	0.5%	0%	0%	97.5%

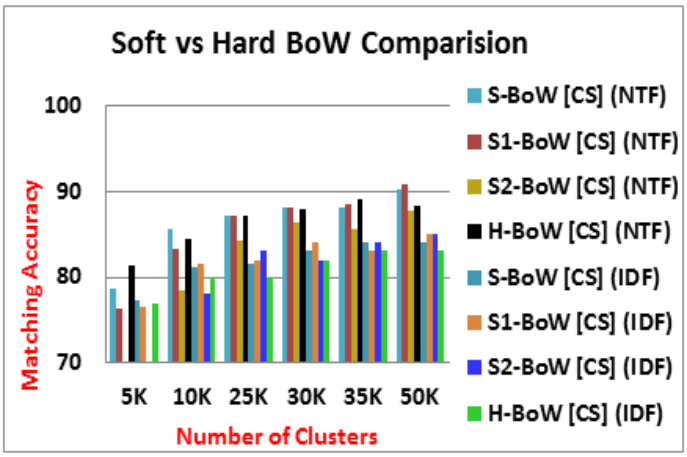

Fig. 8. Soft vs Hard Assignment in Visual BoW

4) Smaller sized SIFT descriptors can be used for BoW.
5) More clusters are generally superior to fewer clusters.
6) Hard assignment should be used for ntf scheme in indoor environments. Soft assignment may be used for $ntfidf$ but at the cost of expense and on an ad hoc basis.
7) sh-BoW validation technique is another good alternative as compared to s-BoW. Incorporation of hue information plays a good role in performance improvement.
8) SIFT features should be used in conjunction with other information for more precise matching indoors.

REFERENCES

[1] R. Smith, M. Self, and P. Cheeseman, "Estimating uncertain spatial relationships in robotics," in *Proc. Uncertainity in Artificial Intelligence, Elsevier.*, 1998, pp. 435–461.

[2] V. Ila, J. Andrade, R. Valencia, and A. Sanfeliu, "Vision-based loop closing for delayed state robot mapping," in *Proc. International Conference on Intelligent Robotics and Systems*, 2007, pp. 3892–3897.

[3] D. Filliat, "A visual bag of words method for interactive qualitative localization and mapping." in *Proc. ICRA.*, 2007, pp. 3921–3926.

[4] H. Kang, A. Efros, T. Kanade, and M. Hebert, "Image matching in large scale indoor environment." in *Proc. IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR) Workshop on Egocentric Vision*, 2009.

[5] P. Espinace, T. Kollar, A. Soto, and N. Roy, "Indoor scene recognition through object detection," in *Proc. of IEEE International Conference on Robotics and Automation (ICRA)*, 2010.

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

[7] A. Davison, G. Y Cid, and N. Kita, "Real-time 3d slam with wide-angle vision," in *Proc. IFAC Symposium on Intelligent Autonomous Vehicles*, 2004.

[8] A. Tapus and R. Siegwart, "Incremental robot mapping with fingerprints of places," in *Proc. Internationl Conference on Intelligent Robots and Systems, IROS.*, 2005, pp. 2429–2434.

[9] N. Khan, "Indoor environmental images (computer science department), otago university (nz)." available from http://www.cs.otago.ac.nz/pgdweb/nabeel/Downloads/Dataset.zip.

[10] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[11] T. Lindeberg, "Feature detection with automatic scale selection," *International Journal of Computer Vision*, vol. 30, no. 2, pp. 79–116, 1998.

[12] D. Lowe, "Distinctive image features from scale invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[13] J. Matas, O. Chum, M. Urban, and T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions." in *Proc.British Machine Vision Conference.*, 2002.

[14] H. Bay, A. Ess, T. Tuytelaars, and L. V Gool, "Robust wide baseline stereo from maximally stable extremal regions." in *Proc. of 9th European Conference on Computer Vision.*, 2006, pp. 404–417.

[15] Y. Ke and R. Sukthankar, "Pca-sift: A more distinctive representation for local image descriptors ." in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR*, vol. 2, 2004, pp. 506–513.

[16] N. Khan, B. McCane, and W. G, "Sift and surf performance evaluation against various image deformations on benchmark dataset." in *Proc. of DICTA2011 (to appear)*, 2011.

[17] T. Botterill, S. Mills, and R. Green, "Speeded-up bag-of-words algorithm for robot localisation through scene recognition." in *Proc.Image and Vision Computing New Zealand, IVCNZ.*, 2008.

[18] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree." in *Proc. CVPR.*, 2006, pp. 2161–2168.

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

[20] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

[21] A. Moore, "A tutorial on kd-trees," Extract from PhD Thesis, Tech. Rep., 1991, available from http://www.cs.cmu.edu/*sim*awm/papers.html.

[22] L. Ballan, M. Bertini, A. Bimbo, D, and G. Serra, "Video event classification using bag of words and string kernels." in *Proc.15th International Conference on Image Analysis and Processing ICIAP.*, 2009.

[23] L. S, S. C, and P. J, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories." in *Proc. of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR.*, vol. 2, 2006, pp. 2169–2178.

[24] Y. Jiang, C. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. In CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval,CIVR.*, 2007, pp. 494–501.

[25] H. Kang, "Indoor environment images." Tech. Rep., available from http://www.hwkang.com.