Image Matching with Distinctive Visual Vocabulary

Hongwen Kang Martial Hebert Takeo Kanade
School of Computer Science
Carnegie Mellon University
{hongwenk, hebert, tk}@cs.cmu.edu

Abstract

In this paper we propose an image indexing and matching algorithm that relies on selecting distinctive high dimensional features. In contrast with conventional techniques that treated all features equally, we claim that one can benefit significantly from focusing on distinctive features. We propose a bag-of-words algorithm that combines the feature distinctiveness in visual vocabulary generation. Our approach compares favorably with the state of the art in image matching tasks on the University of Kentucky Recognition Benchmark dataset and on an indoor localization dataset. We also show that our approach scales up more gracefully on a large scale Flickr dataset.

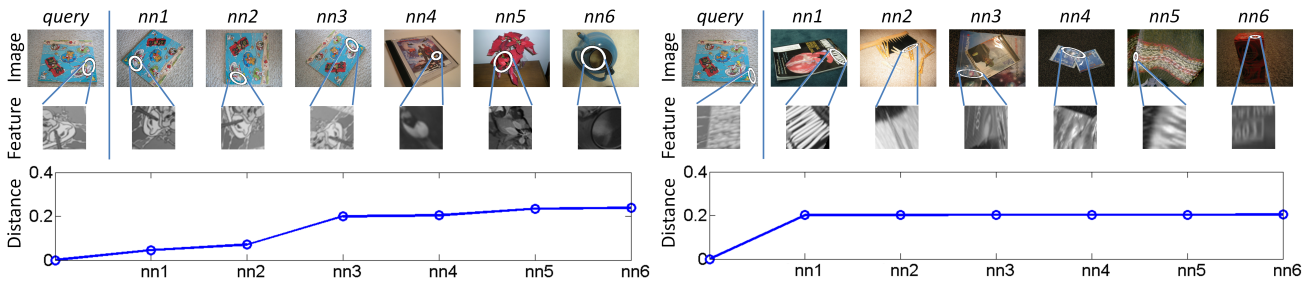
1. Introduction

Much progress has been achieved in image matching in recent years [25, 26, 27]. Despite considerable advances, achieving good performance for applications involving large image databases remains challenging [13, 34]. There are obvious computational issues because, as the database size increases, finding exact nearest neighbors becomes inefficient, especially for high-dimensional image features. More importantly, as the database size increases, the number of similar feature points in a unit distance interval increases exponentially [3, 36], i.e., the curse of dimensionality starts to affect performance (Figure 1). Therefore, distinguishing distinctive features from indistinctive features is important for both efficiency and, more importantly, matching accuracy.

Much of the literature has focused on the computational issues. For example, [33] proposed a framework based on the bag of words (BOW) model from the text retrieval domain. To further improve the efficiency, a tree data structure was adopted in [29]. The limitations of the vector quantization approaches have been addressed and the proposed improvements include adding local soft distance [15, 30], improving the symmetry among nearest neighbors [16], min-Hash and spatial verification [4, 31], fast search with metric learning [14], and other variations. Advances in approximate nearest neighbor search also contributed to improving efficiency [1, 28].

In this paper we explore the feasibility of addressing the second issue, i.e., distinguishing distinctive features from indistinctive features. Intuitively, we note that the concept of feature distinctiveness only makes sense when we consider the feature vector in the context of its nearest neighbors in the feature space (Figure 1). When all of its nearest neighbors are similarly close, a feature is indistinctive and has little or even negative contribution to image matching. From the literature, we found these heuristics can be summarized well by a formal definition of feature distinctiveness based on statistics and information theory [11], which has been successfully applied in content-based image retrieval as a retrieval quality criteria, *after the retrieval is accomplished* [19]. In this work, we explore its effectiveness in feature selection and indexing large image databases. We show that feature distinctiveness is strongly correlated with image matching performance. This observation is especially important, given the tendency to use larger and larger number of high dimensional features in the image matching problem [29, 31, 16]. We propose a bag-of-words algorithm that combines the feature distinctiveness in visual vocabulary generation.

Some of the published approaches used other strategies for selecting features. For example, in [33], a visual vocabulary was generated using only features that could be stably tracked through consecutive video frames, which eliminates a large number of features that are not repeatable. However, this tracking based process is not applicable to static image datasets that are not extracted from the same video stream. Also, in [25], an empirical method was used to prune the candidate feature matches by requiring that a good match must have distance less than 0.6 times that of the second best match. Although this criteria is intuitive, it is difficult to formulate the underlying theory and this approach is not applicable to large scale image matching problems where



(a) Distinctive feature (b) Indistinctive feature

Figure 1. An example comparing *distinctive* features (left) and *indistinctive* features (right). Each row shows: 1) the query and retrieved images; 2) the feature that was used as the query and its nearest neighbors; and, 3) their corresponding distances. When all of its nearest neighbors are similarly close, a feature is indistinctive and has little or even negative contribution to image matching.

the same and or similar objects can appear in multiple images.

Our approach is related to the *supervised* feature selection approaches in the literature. For example, [35] and [23] use information theory and conditional entropy. [24] also showed an interesting idea on reducing the number of features used in location representation through optimizing the weighting of features to maximize the posterior probability of a location. [32] is most similar to our approach in that they generate a visual vocabulary by maximizing the information gain in distinguishing different locations. The major difference between these approaches and ours is that our distinctive feature selection process is fully *unsupervised*, based on the non-uniformity of high dimensional feature space and feature distinctiveness. Beyond the obvious scalability advantage of unsupervised techniques, our approach is also more appropriate for the image matching problem itself. For example, in Figure 2, our approach discovers the distinctive "eye" type of features that are important for image matching. However, since this feature also appears on several *different* kinds of objects, they would be downweighed by the supervised approaches due to the confusion for classification.

The use of local nearest neighbor statistics in our approach has additional connections to estimating distance functions for recognition and retrieval tasks [9, 10, 37]. Although our primary goal is for specific instance matching, our general approach is also similar to approaches in object categorization that try to identify important features/locations for category recognition [7, 22, 38]. Also note that our distinctiveness measure is based on local nearest neighbor information, which is fundamentally differently from the dimensionality reduction literature such as [8, 20]. These are *global* approaches that seek to lower the dimensionality of the whole feature space, without taking advantage of the non-uniformity in the data distribution [6, 17].

Our approach is generally applicable to any high dimensional features used in image matching. In this paper, we use the Scale Invariant Feature Transformations (SIFT) [25]. We use the standard benchmark of [29] in order to validate the effectiveness of our approach. Since our approach is particularly well-suited for situations in which the database images are very confusing, we also evaluate the performance on a public dataset from an indoor localization task [18]. In this task, the images differ in relatively small details which cannot be captured without accounting explicitly for the distinctiveness of features. Finally, we demonstrate that our approach scales up more gracefully on a large scale Flickr dataset.

2. Distinctiveness of high dimensional features

2.1. Definition of feature distinctiveness

We first introduce the measure of feature distinctiveness and we show its connection with the concept of intrinsic dimensionality [19]. In a high dimensional space, assuming uniform distribution, the number of feature points in a unit ball increases exponentially with respect to the dimensionality (the "curse of dimensionality" [3].) This is indeed what makes information retrieval applications, such as image search, extremely challenging. As shown in [11, 19], in a high dimensional space, the expected ratio of the distance between a query feature point to the $(K+1)^{th}$ nearest neighbor and its distance to the K^{th} nearest neighbor is:

$$\frac{E\{d_{(K+1)NN}\}}{E\{d_{KNN}\}} \approx 1 + \frac{1}{Kn}, \tag{1}$$

where n is the dimension of the feature space. The ratio decreases monotonically as the dimensionality increases. For a given n, the maximum ratio is achieved between the nearest and second nearest neighbors of the query feature point, which is:

$$\frac{E\{d_{2NN}\}}{E\{d_{1NN}\}} \approx 1 + \frac{1}{n}. \tag{2}$$

For example, for a 128 dimensional feature, this expected distance ratio is merely 1.008. Features with such high

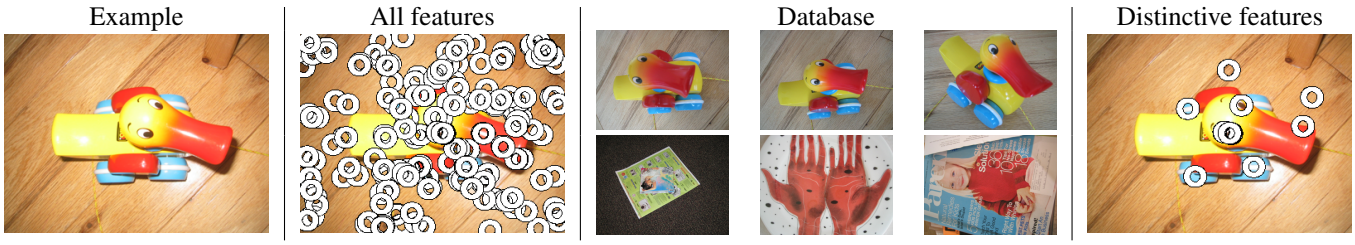


Figure 2. Given an image and the features detected, it is insufficient to determine which feature is more informative by itself; however, after putting this image in the context of the given database, our algorithm is able to select the distinctive features based on the statistics of their nearest neighbors. (Features shown here only illustrate the locations of the original MSER regions.)

dimensionality are therefore very unreliable for similarity search, because any small disturbance to the feature space, e.g., lighting or viewpoint, could change the nearest neighbor ordering. Fortunately, in real-world applications, 1) feature spaces are rarely uniform, and 2), locally, the feature intrinsic dimensionality is much lower than the actual dimensionality n of the parent feature space.

Based on these observations, we can draw a direct connection between feature distinctiveness and its intrinsic dimensionality. Then our goal is to select the features that have low intrinsic dimensionality (*distinctive*) and filter out the ones that have high intrinsic dimensionality (*indistinctive*). To this end, [19] suggested a generative model for estimating the data likelihood for a given intrinsic dimensionality, by simply counting the number of nearest neighbors appearing within some range from the feature point of interest. It was shown that given an intrinsic dimensionality n', if the nearest neighbor of a feature x is at distance d_{NN}, the likelihood of observing more than N_c data points in the distance range of $R_p \times d_{NN}$ ($R_p \geq 1.0$) is:

$$P(N_c|n', R_p) = (1 - \frac{1}{(R_p)^{n'}})^{N_c}, \qquad (3)$$

note that $P(N_c|n', R_p)$ is independent of the absolute value of d_{NN}. In this paper, we use this likelihood definition as our measure of the feature distinctiveness. Intuitively, the more features (larger N_c) observed in this distance range the less distinctive x is. For example, the query feature of Figure 1(b) is less likely to have a low intrinsic dimensionality than that of Figure 1(a), because many of the other neighbors have almost the same distance as its nearest neighbor, therefore N_c will be quite large. In practice, we will choose a desired intrinsic dimensionality n', we will estimate N_c for each feature and select the features for which $P(N_c..)$ is larger than certain threshold. In our experiments, we calculate the values of R_p for each n' in the same way exactly as in [11, 19].

2.2. Correlation between feature distinctiveness and image matching performance: a case study

From a public image dataset [29] that we used in our experiments, we generate some empirical statistics of feature distinctiveness (Figure 3). We set $n' = 6$ and $R_p = 2.77$ for this empirical study. To see the strong correlation between feature distinctiveness and image matching performance, we categorize query images based on their matching performance, using a baseline algorithm that does not use feature distinctiveness [29]. We notice that the more distinctive features the query image has, the better it will get matched to the correct database images.

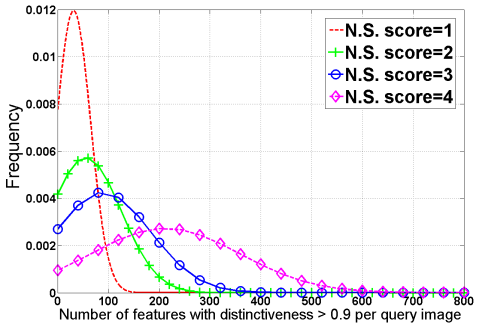


Figure 3. Statistics of feature distinctiveness estimated from the dataset of [29]. A 128 dimensional MSER-SIFT feature is used; we assume that the intrinsic dimensionality is 6. In each graph, statistics are drawn from images categorized based on their matching scores ($1-4$) resulted from a baseline algorithm (N.S. [29]), which does not use distinctiveness. The graphs show a strong correlation between the number of distinctive features and the retrieval performance, i.e., the more distinctive features exist in the query image, the better it will get matched to the correct database images.

3. Image matching with distinctive features

We propose an integration of the feature distinctiveness with the bag-of-visual-words framework [33]. The major steps of our approach are: distinctive visual vocabulary generation, distinctive feature selection for database image indexing and query image representation, vector quantization and retrieving, as shown in Figure 4.

3.1. Distinctive visual vocabulary generation

The purpose of visual vocabulary generation is to cluster a large set of high dimensional features to a finite set that are representative for the visual patterns in the database.

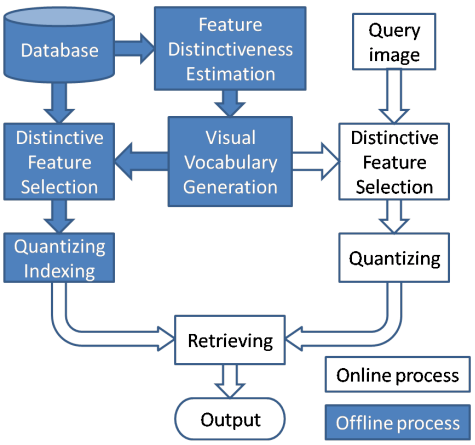


Figure 4. Diagrams of the distinctive bag-of-words (Distinctive_BOW) approach.

This visual vocabulary essentially compresses the dataset because it is usually much smaller than the original feature set [12, 33]. The K-means algorithm is frequently used for this purpose. The original K-means algorithm can be formulated into the following minimization problem,

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{K} \sum_{x_j \in \mathbf{C}_i} \|x_j - c_i\|^2, \tag{4}$$

where $\mathbf{C} = \cup \mathbf{C}_i$. $\mathbf{C}_i$ is a cluster consists of a set of data points with their averaged center at c_i.

This clustering approach, however, is sensitive to the non-uniformity or bias in the dataset [21]. This is especially problematic for the K-means algorithm, since the calculation of a cluster center can be skewed by a few data points that are far away from the actual cluster centers [11, 12]. In this paper, we adopt a weighted clustering algorithm that weighs feature points with their distinctiveness. The weighing strategy has two advantages. First, it leads to cluster centers that are closer to regions where features with higher distinctiveness are located at; and second, the skewing effect of indistinctive features far away from the cluster centers will be alleviated by their low distinctiveness. We define the clustering problem as the following,

$$\arg\min_{\mathbf{C}} \sum_{i=1}^{K} \sum_{x_j \in \mathbf{C}_i} w_j * \|x_j - c_i\|^2, \tag{5}$$

where w_j is the weight for each data point x_j. In our experiments, we set w_j to the distinctiveness measure of x_j. The output of the distinctiveness weighted clustering process is a set of cluster centers. This set is called distinctive visual vocabulary, and a member of it is called a distinctive visual word.

3.2. Database image indexing and retrieving with bag-of-distinctive-words

After the visual vocabulary is constructed, we assign a discrete digit to represent a high dimensional feature vector. This digit represents the visual word that is the closest to the feature vector, under the Euclidean distance. This process is normally known as vector quantization [12]. Conventionally, any given feature vector is forced to be assigned to at least one of the clusters, regardless of the specific distance configuration between this feature vector and all the cluster centers. In this paper, however, instead of quantizing every feature vector in the image, we propose to select only a subset of features that are most distinctive, then quantize and use them for indexing and searching.

The distinctiveness of an image feature is calculated in the same way as described in Section 2. But this time, the reference feature set that an image feature vector compare with is the distinctive visual vocabulary generated from the previous step, instead of using the original features from the database. Because the distinctive visual vocabulary is relative small, we can efficiently calculate the distinctiveness of each feature vector. Also, since one needs to find the nearest neighbor for the purpose of vector quantization any way, using the visual vocabulary for distinctiveness calculation adds negligible computational cost.

After the distinctiveness of the query features is calculated, we apply a threshold (0.9) to select the distinctive ones and vector quantization is used to transform the original feature vectors to discrete visual words indexes, we note the generated image representation by bag-of-distinctive-words. Based on these bag-of-distinctive-words, each image is now represented by a term frequency inverse document frequency vector (TF-IDF). More details of TF-IDF formulation and its application in image search can be found in [2, 33]. Distances between a query image and the database images are then calculated by the L_1 distances of these TF-IDF vectors [29].

Instead of distinctive feature selection, one can use the distinctiveness of a feature as the soft weight in the TF-IDF vectors. We choose not to implement this strategy in this paper for the consideration of computational expense. Calculating the precise distinctiveness requires searching for a large number of nearest neighbors (N_c), which can be inefficient. Instead, for a distinctiveness threshold (0.9), to determine whether a feature is distinctive or not, the maximum number of nearest neighbors that one needs to retrieve is less than a hundred. This is especially useful for the scenarios where early stopping can decrease the computational expense significantly [5, 19].

4. Experiments

The University of Kentucky Recognition Benchmark [29] provides a suitable baseline to demonstrate and analyze

the effectiveness of our proposed approach. This dataset includes 2550 objects and scenes, e.g. Figure 2. Each object is captured 4 times from different viewpoints, distances and illumination conditions. For evaluation, each image is used as query and a score between 1 and 4 is calculated for the retrieval results, corresponding to the number of relevant images returned by the algorithm among the top 4 (4 is the highest achievable score, meaning that all four database images matching the query image have been found). This score divided by 4 is the value that the precision and recall measurements are equal at. In addition to this benchmark dataset, we collected a larger image dataset that consists of about 500 thousand high resolution images downloaded from the Flickr website, using the public API for the daily list of "interesting" photos[1]. In our experiments, we denote the University of Kentucky Recognition Benchmark itself by "UKBench", and the combined dataset by "UKBench+Flickr".

4.1. Object instance recognition

General object recognition has been a difficult research problem in the literature. Using image search techniques, however, one can achieve promising results in recognizing specific instances of objects [29]. The University of Kentucky Recognition Benchmark dataset has been commonly used for benchmarking the performance of various image search algorithms in this application [16, 29, 31]. In this experiment, we use the standalone UKBench dataset. To be consistent with the other baselines, here we use the original 7 million SIFT features computed from MSER regions [25, 26]. We quantitatively evaluate our approaches and compare them with the state of the art.

We vary the selections of the intrinsic dimensionality n' in the distinctive visual vocabulary generation process (n'_V) and the feature selection process for the bag-of-distinctive-words (n'_I). We compare our performance with several state-of-the-art techniques (Table 1). The first baseline algorithm is the Nistér and Stewénius algorithm [29] that uses a hierarchical K-means algorithm for vocabulary generation and a multi-level scoring strategy. The second baseline algorithm is proposed by Jegou et. al. [16] which showed that improving the nearest neighbor symmetry with a contextual similarity measure could improve the matching performance.

The third baseline algorithm [31] is most similar to our approach, which used approximate K-means algorithm for large vocabulary generation. In fact, it is exactly the same as our algorithm when $n'_V = +\infty$ and $n'_I = +\infty$, i.e., every feature has distinctiveness 1.

We pick the parameters by 10-round cross validation on a hundred samples randomly selected from the database, and choose the combination that is favored the most, i.e., $n'_V = 25$, $n'_I = 25$ and vocabulary size $1M$. Although, the performance does not vary much for other choices of n', e.g. $n'_V = 25$ and $n'_V = 30$, and vocabulary sizes, e.g., $500K$ and $1M$ (Figure 5). Considering the potential variance due to K-means initialization, for all the experiments with the $1M$ vocabulary, we used fixed initial cluster centers that are randomly initialized for once; and for each of the other experiments, we used independently randomly initialized cluster centers. We found that the initialization does not cause significant variance to the algorithm. Figure 5 shows the performance for different parameters and the comparison to other baseline algorithms is shown in Table 1.

Algorithm	Score
Nistér and Stewénius[29]	3.29
Jegou, et.al.[16]	3.38
Philbin, et.al.[31]	3.45
Distinctive_BOW	3.51

Table 1. Quantitative comparison of our approach to other baseline bag-of words approaches that do not use distinctive feature selection in vocabulary generation.

	N.S.[29]	R.S.[18]	Distinctive_BOW
Clean set	0.996	0.999	1.0
Confusing set	0.843	0.905	0.98

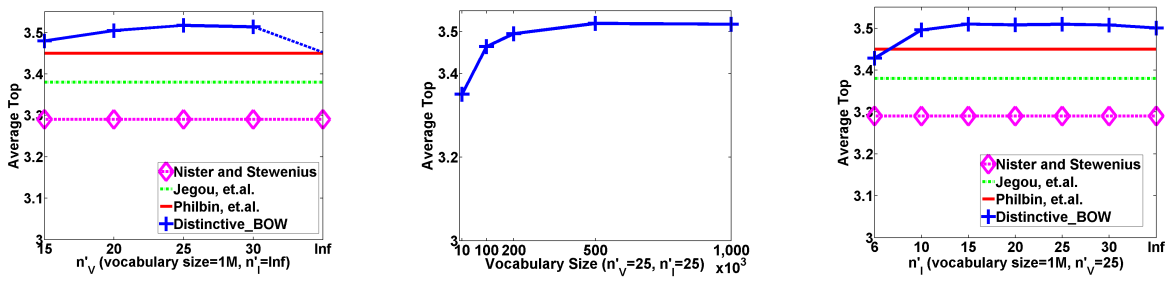
Table 2. Quantitative comparison of our approach with the N.S. algorithm [29] and the R.S. algorithm [18] on an image based indoor localization task.

Figure 6 shows a representative example comparing our result (Distinctive_BOW) and that of the baseline algorithm [29], since this is the only publicly available feature set with quantized visual words. We implemented the flat weighting scheme and we verified that our implementation of the baseline got exactly the same score as published.

4.2. Image based indoor localization

In this section, we analyze the performance of the proposed image matching approach in the context of an indoor localization application. In this application, the location of a user is estimated by matching an image taken by the user with a large database of position-tagged images. The challenge in this task stems from the high degree of repeatability in man-made environments. The resulting ambiguity complicates the matching process dramatically. For this task, [18] proposed an iterative algorithm (Re-Search, or R.S.) that refines a similarity function based on the visual words distribution in a small neighborhood. However, this approach is unstable and sensitive to quantization noises. Therefore, [18] resorted to an intermediate solution that combines the local and global similarity functions. In contrast, our approach fits very well into this scenario because

[1] http://www.flickr.com/services/api/flickr.interestingness.getList.html



(a) (b) (c)

Figure 5. The effects of different parameter choices on recognition performance, (a), different intrinsic dimensionality choices in distinctive visual vocabulary generation; (b), different sizes of distinctive visual vocabularies; (c), different intrinsic dimensionality choices in distinctive feature selection for indexing and retrieving.

Figure 6. Comparison of the performance of our approach and the N.S. algorithm [29] in some extremely challenging situations that involve severe illumination, view point changes and cluttered background.

it emphasizes the effect of feature distinctiveness for robust image matching.

For evaluation, we use the same publicly available dataset as [18]. The database consists of around 8.8 thousand images (each associated with a location label). Two sets of testing images are used, one has rich and distinctive visual structures, called the "clean set", the other "confusing set" is composed of some much more challenging images that captured more detailed parts of the scene, or objects that are common across images, such as doors, posters, etc. Most of the features in those images are ambiguous therefore techniques that do not emphasize on distinctiveness performs poorly. Both testsets are composed of 80 images with location ground-truth. For each testing image, the 8 most similar pre-captured images are retrieved. A majority vote scheme is used for location prediction and the performance is measured the same way as [18].

The provided dataset has 6.4 million SIFT features extracted with the HESAFF [27] region detector. Through cross validation on a hundred of images from the *database*, we set the intrinsic dimensionality $n' = 6$ and the distinctiveness threshold 0.9, which reduced the number of features down to 4.2M. The baseline for comparison here are the two algorithms developed in [18].

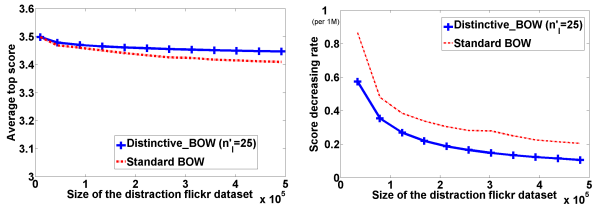
We measure the performance with precision-recall and mean average precisions (mAP)(Table 2). On the clean-set, there was no surprise that all four algorithms reached almost perfect performance. On the much more challenging confusing-set, our approach significantly outperform the Nistér and Stewénius algorithm (N.S.) [29] and the R.S. algorithm proposed by [18]. Some qualitative comparisons are illustrated in Figure 7, notice that our proposed approach performs much better than the baseline algorithms in the confusing environment and the extreme cases where very few non-ambiguous features are available.

4.3. Scaling up image search with distinctive features

In addition to the performance on the small scale benchmark dataset, one important performance measurement of an image search algorithm is its scalability to large scale datasets. We evaluate the performance of our approach on the UKBench+Flickr dataset. In this evaluation, images in the original UKBench dataset are used as query. The search score is measured the same way using the top 4 returns as mentioned earlier. A retrieval is correct if the retrieved image is from the original UKBench dataset and contains the same object as in the query image.

In this experiment, we use a fixed visual vocabulary that has been generated using the approach proposed in this paper, and we focus on the scalability of our approach and evaluate the added benefit of the distinctive feature selection approach in large scale applications. Two algorithms are under comparison, one is the standard bag-of-words algorithm that uses all the features for indexing and retrieving (Standard BOW); the other uses the proposed distinctive feature selection to select and quantize only a part of the features (Distinctive_BOW).

Figure 7. Some qualitative examples comparing our approach (Distinctive_BOW) and the R.S. algorithm [18] on the localization task. Our approach significantly outperforms these baseline algorithms in extremely confusing environment, where very few visual features are available. (Red boxes indicate incorrectly matched images.)



(a) (b)

Figure 8. Performance evaluation on scaling up the object instance recognition to large scale dataset (UKBench+Flickr). In the experiment, we vary the size of the distraction Flickr dataset and monitor the changes in the recognition performance when the database scales up. The two measurements are, (a), the absolute value of recognition accuracy; and (b), the recognition score decreasing rate with respect to sizes of the database.

For this experiment, we extract the HESAFF SIFT features from the UKBench+Flickr dataset using a publicly available package [27]. Using the default parameters, the HESAFF SIFT feature extractor generates on average 2000 features per image, and in total about $20M$ for the UKBench dataset. We randomly sampled $10M$ of them for the distinctive visual vocabulary generation. We use the same parameters as we used in the previous experiment, i.e., $n'_V = 25$, $n'_I = 25$ and vocabulary size $1M$.

In addition to the absolute value of the object instance recognition score (Figure 8(a)), a better measurement for scalability is the changing rate of the recognition scores with respect to the database size. We measure this as the decrease of the recognition score for every unit number ($1M$) of distraction images that are added, i.e., the slope of the curve in Figure 8(a). Figure 8(b) shows the score decreasing rate when more and more distraction images are added to the database.

The changing rate is high at the beginning when new distraction images are added and then it starts to decrease. This phenomenon is due to the way that we measure the recognition rate, i.e., all the images from the Flickr distraction dataset are considered equally. Therefore, newly added distraction images do not add a significant distraction effect to the ones that are already in the database.

The ratio between the score decreasing rates of the two approaches under comparison is $0.7 \sim 0.9$, i.e., applying feature selection is $10\% \sim 30\%$ better in scalability compared to the standard bag-of-words approach, which is benefited directly from the selection of distinctive features for indexing and retrieving since the same distinctive visual vocabulary is used in both approaches.

5. Conclusions and future work

In this paper, we explored an approach for image matching that builds on the distinctiveness of high dimensional features, reflected in their relationship with their nearest neighbors. This approach compares favorably with the state of the art in image matching tasks such as the University of Kentucky Benchmark dataset and an indoor localization dataset, also our approach scales up more gracefully on a large scale Flickr dataset

There are several directions that remain to be explored. First, the distinctiveness we rely on right now assumes a single intrinsic dimensionality across the dataset, without fully taking advantage of the non-uniformity property of high dimensional space. Second, we would also like to evaluate the generalizability of our visual vocabulary to other datasets.

6. Acknowledgements

The authors would like to thank anonymous reviewers for helpful and constructive suggestions. This material is based upon work partially supported by the National Science Foundation under Grant No. EEC-0540865.

References

[1] S. Arya and D. M. Mount. Approximate nearest neighbor queries in fixed dimensions. In *SODA '93*, 1993. 1

[2] R. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval*. Addison Wesley, May 1999. 4

[3] R. Bellman. *Adaptive Control Processes: A Guided Tour*. Princeton University Press, 1961. 1, 2

[4] O. Chum, J. Philbin, and A. Zisserman. Near duplicate image detection: min-hash and tf-idf weighting. In *BMVC*, 2008. 1

[5] C. Faloutsos. *Searching Multimedia Databases by Content*. 1996. 4

[6] C. Faloutsos and I. Kamel. Beyond uniformity and independence: Analysis of r-trees using the concept of fractal dimension. In *PODS*. ACM Press, 1994. 2

[7] R. Fergus, P. Perona, and A. Zisserman. Weakly supervised scale-invariant learning of models for visual recognition. *Int. J. Comput. Vision*, 2007. 2

[8] I. K. Fodor. A survey of dimension reduction techniques. *LLNL technical report*, 2002. 2

[9] A. Frome, Y. Singer, and J. Malik. Image retrieval and classification using local distance functions. In *NIPS*. 2007. 2

[10] A. Frome, Y. Singer, F. Sha, and J. Malik. Learning globally-consistent local distance functions for shape-based image retrieval and classification. In *ICCV*, 2007. 2

[11] K. Fukunaga. *Introduction to statistical pattern recognition (2nd ed.)*. 1990. 1, 2, 3, 4

[12] A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, Norwell, MA, USA, 1991. 4

[13] J. Hays and A. A. Efros. Scene completion using millions of photographs. *SIGGRAPH*, 2007. 1

[14] P. Jain, B. Kulis, and K. Grauman. Fast image search for learned metrics. In *CVPR*, 2008. 1

[15] H. Jegou, M. Douze, and C. Schmid. Hamming embedding and weak geometric consistency for large scale image search. In *European Conference on Computer Vision*, 2008. 1

[16] H. Jégou, C. Schmid, H. Harzallah, and J. Verbeek. Accurate image search using the contextual dissimilarity measure. *PAMI*, 2009. 1, 5

[17] F. Jurie and B. Triggs. Creating efficient codebooks for visual recognition. In *ICCV '05*, 2005. 2

[18] H. Kang, A. A. Efros, M. Hebert, and T. Kanade. Image matching in large scale indoor environment. In *CVPR: Workshop on Egocentric Vision*, 2009. 2, 5, 6, 7

[19] N. Katayama and S. Satoh. Distinctiveness-sensitive nearest neighbor search for efficient similarity retrieval of multimedia information. In *ICDE*, 2001. 1, 2, 3, 4

[20] Y. Ke and R. Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. *CVPR*, 2004. 2

[21] K. Kerdprasop, N. Kerdprasop, and P. Sattayatham. Weighted k-means for density-biased clustering. In *In Data Warehousing and Knowledge Discovery*, pages 488–497, 2005. 4

[22] G. Kim, C. Faloutsos, , and M. Hebert. Unsupervised modeling of object categories using link analysis techniques. In *CVPR*, 2008. 2

[23] S. Lazebnik and M. Raginsky. Supervised learning of quantizer codebooks by information loss minimization. *IEEE Trans. PAMI*, 2009. 2

[24] F. Li. Probabilistic location recognition using reduced feature set. In *ICRA*, 2006. 2

[25] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004. 1, 2, 5

[26] J. Matas, O. Chum, M. Urban, and T. Pajdla. Robust wide-baseline stereo from maximally stable extremal regions. *Image and Vision Computing*, 2004. 1, 5

[27] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, (1), 04. 1, 6, 7

[28] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *VISSAPP (1)*, 2009. 1

[29] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006. 1, 2, 3, 4, 5, 6

[30] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR08*. 1

[31] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 1, 5

[32] G. Schindler, M. Brown, and R. Szeliski. City-scale location recognition. *CVPR*, 2007. 2

[33] J. Sivic and A. Zisserman. Video Google: A text retrieval approach to object matching in videos. In *ICCV*, 2003. 1, 3, 4

[34] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *PAMI*, 2008. 1

[35] M. Vidal-Naquet and S. Ullman. Object recognition with informative features and linear classification. In *ICCV '03*, 2003. 2

[36] Z. Wu, Q. Ke, M. Isard, and J. Sun. Bundling features for large scale partial-duplicate web image search. 2009. 1

[37] H. Zhang, A. C. Berg, M. Maire, and J. Malik. Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In *CVPR*, 2006. 2

[38] S. Zhang, Q. Tian, G. Hua, Q. Huang, and S. Li. Descriptive visual words and visual phrases for image applications. In *MM '09*. ACM, 2009. 2