SELECT INFORMATIVE FEATURES FOR RECOGNITION

Zixuan Wang[1]*, Qi Zhao*[2]*, David Chu*[3]*, Feng Zhao*[4] *and Leonidas J. Guibas*[5]

[1]Department of Electrical Engineering, Stanford University, Stanford, CA, USA
[2]Department of Technology & Information Management, University of California Santa Cruz, Santa Cruz, CA, USA
[3]Microsoft Research, Redmond, WA, USA
[4]Microsoft Research Asia, Beijing, China
[5]Department of Computer Science, Stanford University, Stanford, CA, USA

ABSTRACT

The state of the art rigid object recognition algorithms are based on the bag of words model, which represents each image in the database as a sparse vector of visual words. We propose a new algorithm to select informative features from images in the database. which can save the memory cost when the database is large and reduce the length of the inverted index so it can improve the recognition speed. Experiments show that only using the informative features selected by our algorithm has better recognition performance than the previous methods.

Index Terms— Image recognition, Bag of words, Informative features, Inverted index

1. INTRODUCTION

The task of image recognition refers to identifying images containing the same object in the query image from the image database. Image recognition becomes a challenging problem when the number of images in the database scales up. The challenge lies in 1) maintaining the recognition accuracy and 2) returning the results efficiently.

So far, a common approach for image recognition is adopting the bag of words (BoW) model which was first introduced by Sivic et al. [1]. In this model, the local image feature descriptors are quantized into visual words and further an image is represented by a sparse vector via employing term frequency-inverse document frequency (tf-idf) weighting scheme. Image candidates are ranked by sorting the cosine similarity of two sparse vectors. Ranking becomes computationally expensive due to a large number of candidates resulted from long inverted index. However, ranking is a vital phase of image recognition and can become the bottleneck if the performance is inefficient.

Given a fixed database size, the average length of the inverted index depends on the number of features per image. We are motivated to shorten the inverted index by using less features per image. The key of the feature reduction process is selecting a subset of informative features for each image. The selected features should be robust against geometry distortion caused by viewpoint change and informative so that they capture the uniqueness of an object. The benefit of using such features include 1) shorter inverted list thus memory efficient, 2) higher relevance for the result candidate set and 3) faster recognition by accelerated ranking.

In this paper, we propose a new algorithm to perform image recognition efficiently by designing criterion to evaluate the importance of image features and preserving those features of large importance for retrieval. Our work is based on the work of Heath et al. [2] by first building a feature graph across the image database. Features corresponding to the same point of an object are linked by edges. Unlike previous approach [3], which only use matched features per image, we augment each feature with all features in the same connected component in the graph and rank them by their scores. The scores of features are defined according to their abilities to distinguish different objects. Only top ranked features are selected to quantize to visual words for recognition.

We evaluate the performance of our method on Oxford dataset and Paris dataset and compared to the state of the art approaches. The experimental results demonstrated that using the informative features selected by our algorithm has better recognition performance than the previous methods.

Section 2 outlines the previous work in the content-based image retrieval and recognition. Our feature selection algorithm is presented in Section 3 and Section 4. Experiments are shown in Section 5. Section 6 concludes our paper.

2. PREVIOUS WORK

In recent years, most image retrieval and recognition algorithms are based on the bag of words model [1] [4] [5]. Based on this model, researchers propose novel ingredients to improve the recognition performance from various aspects. In [6], query expansion is adopted to retrieve the relevant documents which are missed due to the variation of particular visual words present in different images of the same object. In [5], the idea of the soft assignment is introduced such that

each image feature of the query image is quantized to multiple visual words instead of just the nearest one. Doing so allows the inclusion of the features which get lost in the quantization stage of previous retrieval systems and achieves better retrieval performance at the cost of increased storage for the index. In [7], binary signatures are used to refine visual word matching and a weak geometry constraint is incorporated to validate match pairs efficiently. Wu et al. [8] propose the semantics-preserving vocabulary and its corresponding model to incorporate semantic information in the BoW model. In [9], a novel similarity measure is learned from the vocabulary to mitigate the quantization error.

When the number of images in the database is large, the inverted index is inefficient. Hashing [10] has been proposed to solve this problem. But the low recall ratio of hashing keeps it from being widely used in image recognition.

All these methods use all image features to create the inverted index and perform recognition for the query image. In [11], a subset of features is identified via manifold regularization. Turcot et al. [3] propose to boost the recognition performance by selecting a small number of features which are robust and distinctive across the image dataset. Knopp et al. [12] present a method to avoid confusing features in location recognition. Our algorithm differs from theirs in two aspects: First, the feature graph associated with the database is used to augment a single feature. Second, a subset of features are further selected per image, which are informative to distinguish different objects.

3. BUILDING THE FEATURE GRAPH

Given a collection of images in the database for recognition, the feature graph is first built to capture the relations of local image features across images. For each image, Hessian affine covariant detectors [13] are used to detect interest points and the SIFT descriptor [14] is extracted on each interest point. We find image feature matchings across images. To find feature correspondences efficiently, we use the algorithm similar to Heath's work [2] that used content-based image retrieval [4] to find related images and selected top k candidate images. RANSAC [15] was used to geometrically verify feature correspondences between images.

Each vertex in the graph is a local image feature. The edge connecting two vertices represents two features are correspondences across images. Each connected component in the feature graph is a track of features usually from different views of the same physical point in the 3D world. Because of the image noise and different views, most connected components have small size ranging from two to hundreds.

Singleton features, which are not matched to other features, are discarded when building the feature graph. This is reasonable because most singleton feature come from transient objects like people or cars. Features from transient objects are not helpful for the retrieval and recognition.

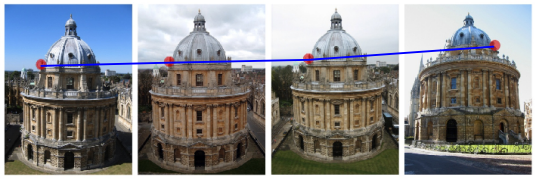

Fig. 1. Features in one connected component in the feature graph.

4. SELECTING INFORMATIVE FEATURES

4.1. Remove confusing features

(a) Correct matches

(b) Confusing matches

Fig. 2. Examples of correct matches and confusing matches.

4.2. Augment features

The image is augmented with features in the same connected component in the feature graph. The spirit behind the feature augmentation is the query expansion. We use feature correspondences from other view points to augment one feature. For image I_i in the database, we represent the image not only with useful features obtained from the previous step , but also include features in the same connected component in the feature graph. Turcot et al. [3] showed that using feature augmentation can improve the recognition performance. Our method is different from their work in that they used the image adjacency to augment features in one image. We augment features in the same connected component, which is shown to have better performance in the evaluation section.

If one image in the database are not matched to other images, it contains no useful feature. This happens when the database only contains a single view of the object. Sometime the isolated single view of an object is important for the ap-

plication. So we add all original features of that image back and select a subset features in the following step.

4.3. Rank features

Features are ranked in one image according to the score of their corresponding visual word. Because each image in the database has the label information, we use the label information when computing the score of visual words. The score of each visual word is defined using an idea similar to the tf-idf scoring:

N_j is the number of features quantized to visual word W_j. N_j^i is the number of features in image with the label i, $i = 1, 2 \ldots k$. We have $\sum_{i=1}^{k} N_j^i = N_j$.

We define the term frequency for each label. The term frequency of W_j for label i is defined as

$$tf(W_j, i) = N_j^i / N_j \tag{1}$$

The inverse document frequency is defined as

$$idf(W_j) = \log \frac{k}{|\{i : W_j \in I_l, label(I_l) = i\}|} \tag{2}$$

where k is the number of labels in the database, and the denominator is the number of labels containing W_j.

The inverse document frequency of each visual word is fixed, but its term frequency varies for each label.

For one feature f in a given image, if the label of the image is i, and its corresponding visual word is W_j. The score of f is defined as:

$$s(f) = tf(W_j, i) \cdot idf(W_j) \tag{3}$$

We rank augmented features in each image by their scores and select at most top m features to convert to the bag of words representation for recognition.

5. PERFORMANCE EVALUATION

The Oxford building dataset and Paris dataset [4] are used to evaluate our algorithm. Both datasets have 11 different buildings. For other unlabeled images, a dummy label is assigned to them. So the datasets contain 12 labels in all. The ground truth of the dataset tells the quality of images: *Good*, *Okay* and *Junk*. We use K-fold validation, with number of folds set to 5. We divide *Good* and *Okay* images into 5 groups. In each validation, we use 4 groups and remaining images as the database and use 1 group as query images.

To compute the feature graph, the visual vocabulary of size one million is used, which is computed using Approximate K-means (AKM) [4] from 10k images. The top 30 candidate images are selected for geometrical verification.

After selecting informative features, a different vocabulary of size 200k is used to test the recognition performance. The recognition accuracy was evaluated using the mean average precision (mAP) the same as [4], which is the mean of the area under the precision-recall curve for all queries.

Fig. 3 shows the performance as we select different number of features per image.

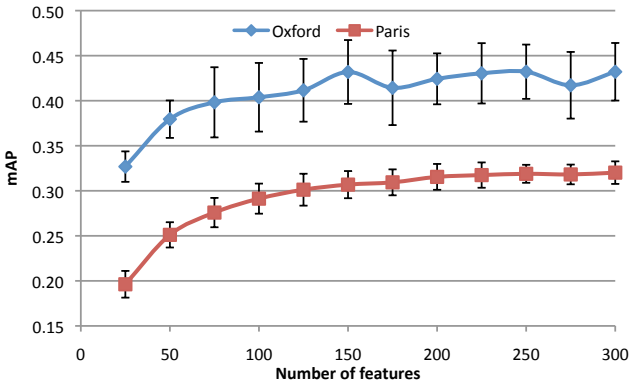

Fig. 3. The mAP performance when selecting different number of features. Error bars represent standard deviation across cross validation folds.

According to Fig. 3, the top 150 features are selected in the Oxford dataset and the top 200 features are selected in the Paris dataset when comparing with using all original features and features from Turcot et al. algorithm [3].

We reproduce their algorithm so the results are different from their original publication. Fig. 4 shows the results in Oxford dataset. Fig. 5 shows the results in Paris dataset. In most cases, our algorithm outperforms both using original features and Turcot's algorithm.

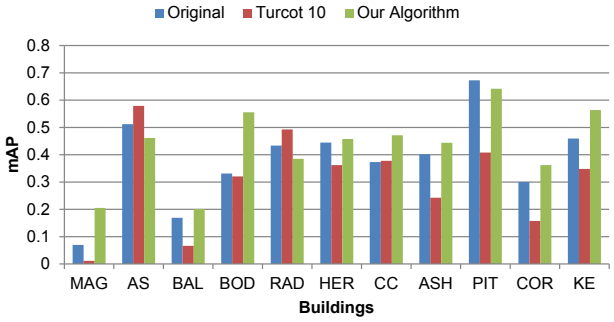

Fig. 4. The mAP on the Oxford dataset.

Table 1 shows the summary of the evaluations. We can only use a small fraction of features to represent all images in database and have the improvement in mAP. The average inverted index length for each visual word is shorter than using all features, which means we need to compare fewer images when finding related images.

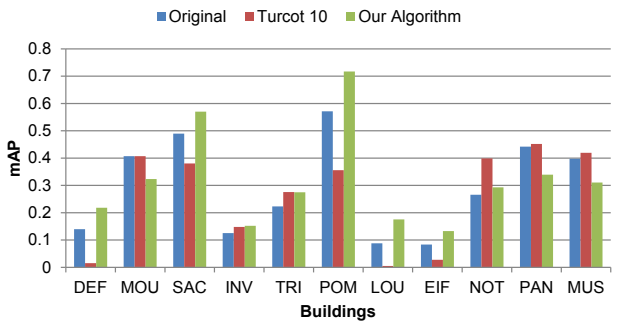

Fig. 5. The mAP on the Paris dataset.

Table 1. Image database summary of the dataset

Dataset: Oxford (4955 images)		
	Original	Informative features
Number of features	14.8M	0.73M (4.9%)
Inverted Index Length	62.1	3.2
mAP	0.37	0.43
Dataset: Paris (6059 images)		
	Original	Informative Features
Number of features	17.7M	1.2M (6.7%)
Inverted Index Length	75.2	4.7
mAP	0.29	0.32

6. CONCLUSIONS

We propose an algorithm for selecting informative features per image in the database for recognition. The feature graph associated with the database is first built and filters out features on transient objects. Confusing features are removed in the consequent step using the label information. Each feature is augmented with features in the same connected component to make it robust to different view points. Finally, top ranked features are selected to represent the image. Experimental results show that by discarding a large number of original features, the average inverted index length is substantially reduced and the recognition performance is improved.

Acknowledgement

NSF grants NSF grants CSE-0832820 and IIS-1016324, as well as ONR grant N0001470710747.

7. REFERENCES

[1] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, Oct. 2003, vol. 2, pp. 1470–1477.

[2] K. Heath, N. Gelfand, M. Ovsjanikov, M. Aanjaneya, and L.J. Guibas, "Image webs: Computing and exploiting connectivity in image collections," in *CVPR*, 2010.

[3] P. Turcot and D.G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," in *ICCV Workshops*, 2010, pp. 2109–2116.

[4] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *CVPR*, 2007.

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

[6] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *ICCV*, 2007, pp. 1–8.

[7] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," *ECCV*, pp. 304–317, 2008.

[8] L. Wu, S.C.H. Hoi, and N. Yu, "Semantics-preserving bag-of-words models and applications," *IEEE Transactions on Image Processing*, vol. 19, no. 7, pp. 1908–1920, 2010.

[9] A. Mikulık, M. Perdoch, O. Chum, and J. Matas, "Learning a Fine Vocabulary," in *ECCV*. Springer-Verlag New York Inc, 2010, p. 1.

[10] O. Chum and J. Matas, "Large-scale discovery of spatially related images," *PAMI*, vol. 32, no. 2, pp. 371–377, 2009.

[11] Z. Xu, I. King, M.R.T. Lyu, and R. Jin, "Discriminative semi-supervised feature selection via manifold regularization," *IEEE Transactions on Neural Networks*, vol. 21, no. 7, pp. 1033–1047, 2010.

[12] J. Knopp, J. Sivic, and T. Pajdla, "Avoiding confusing features in place recognition," *ECCV*, pp. 748–761, 2010.

[13] K. Mikolajczyk and C. Schmid, "Scale & affine invariant interest point detectors," *IJCV*, vol. 60, no. 1, pp. 63–86, 2004.

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

[15] M.A. Fischler and R.C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," *Communications of the ACM*, vol. 24, no. 6, pp. 381–395.