Supplementary Material for Paper 1184

Rongrong Ji[1], Xing Xie[2], Hongxun Yao[1], Wei-Ying Ma[2]
[1]Harbin Institute of Technology
No.92, West Dazhi Street, Harbin, 150001, Heilongjiang, P. R. China
{rrji,yhx}@hit.edu.cn
[2]Microsoft Research Asia
No.49, Zhichun Road, Haidian Distinct,100190 Beijing, P. R. China
{xxie, wyma}@Microsoft.com

Abstract

This supplementary material offers: (1). Basic explanation about TF-IDF functionality in BoW model of patch-based recognition; (2). Some fusion results of experimental figures so that reviewers could understand the experimental results more easily.

1. TF-IDF Functionality

Since visual vocabulary analogizes images to text document, the contribution of "visual words" to recognition can be evaluated via TF-IDF term weighting. In retrieval process, the query image is represented as *Bag-of-Visual-Word* vector, based on which the similarity between two images can be easily evaluated.

Given a visual vocabulary, a query image q or a database image d can be represented as an N dimensional vector of visual words. Each word has a weight associated with it. N is the number of words in the vocabulary. Similar to text retrieval, the relevance between q and d can be calculated as the cosine of the angle between the two word vectors. That is,

$$r(d,q) = \frac{\sum_{i=1}^{N} w_{di} w_{qi}}{|d||q|} \quad (1)$$

where w_{di} is the weight of the i^{th} word in document d, w_{qi} is the weight for the i^{th} word in query q.

The weight of each word usually takes two factors into consideration: term frequency (*TF*) and inverse document frequency (*IDF*). Term frequency means the normalized frequency of a word in a document. In our case, large term frequency means the word has appeared multiple times in the same image, which shows that the feature is more robust. *TF* is calculated as:

$$TF(t_i, d) = \frac{n_i}{\sum_{k=1}^{n_d} n_k} \quad (2)$$

where n_i is the number of occurrences of term t_i in document d, N_d is the number of words in document d. The motivation for using inverse document frequency is that terms which appear in many documents are not very useful for distinguishing a relevant document from a non-relevant one. In our case, they may be those noisy features. *IDF* can be calculated as:

$$IDF(t_i) = log \frac{|D|}{|\{d|t_i \in d\}|} \quad (3)$$

where D is the total number of documents in the database, $\{d|t_i \in d\}$ is the number of documents where t_i appears. In text retrieval, if a word appears in too many documents, i.e. *IDF* is small, it will be ignored since it contributes little while brings many noises. Such words are called "stop words". By deleting stop words from the index, both memory cost and retrieval time are reduced.

Finally the weight for word t_i in document d is defined as the multiplication of *TF* and *IDF*:

$$w_{di} = TF(t_i, d) IDF(t_i) \quad (4)$$

2. Quantization Error Validation

Based on hierarchical feature space quantization, the nearest neighbor search in each visual word of visual vocabulary is more inaccurate comparing with the same process in global feature space, which is because of the hierarchical quantization. For validation, we compare the matching ratio of nearest neighbor search results both inside leaves and among overall feature space.

For visual vocabulary, the nearest neighbor search in visual word is inaccurate due to its local nature in hierarchical structure of vocabulary generation. For validation, we compare the matching ratio of Nearest Neighbor (NN) search results both inside leaves and among overall feature space.

Table 1: Hierarchical Quantization Error Test

NN\GNP	1	3	5	10	15	20
50	41.46%	73.34%	85.00%	94.53%	97.11%	98.18%
200	57.46%	66.21%	79.00%	92.02%	95.00%	97.48%
1000	11.54%	38.27%	51.57%	67.48%	85.16%	94.91%
2000	6.38%	25.68%	40.59%	58.54%	79.21%	92.42%

Tab.1 presents the investigation of quantization errors in a 3-branch, 5-level vocabulary tree. We select 3K images from our urban street scene database to form the vocabulary, with 0.5M features (average 2K features in each visual word). We compare the matching ratio between global-scale NN and leaf-scale NN, in which global-scale is in overall feature space while leaf-scale is inside leaf nodes. We extend leaf-scale to include more local neighbors using

Greedy N-best Path (GNP) [5]. The quantization error is evaluated by this Matching Ratio to see to what extend the quantization would cause feature point mismatching. In Tab1, NN means the nearest neighbor search scope and GNP 1-5 means the number of branches we parallel in GNP search extension. From Tab1 the match ratios between inside-leaf and global-scale search results are extremely low when GNP number is small.

3. Fuse Experimental Figures Together

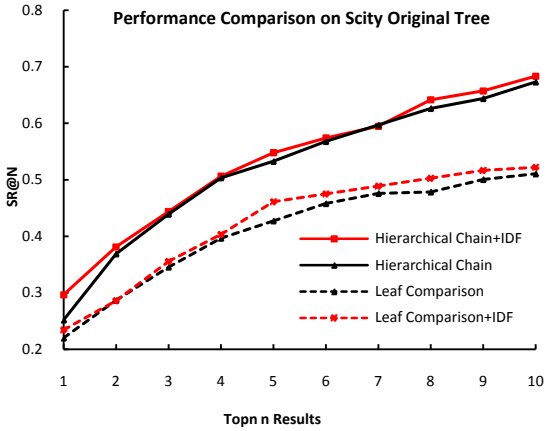

Figure 1: Scity Original Tree Test

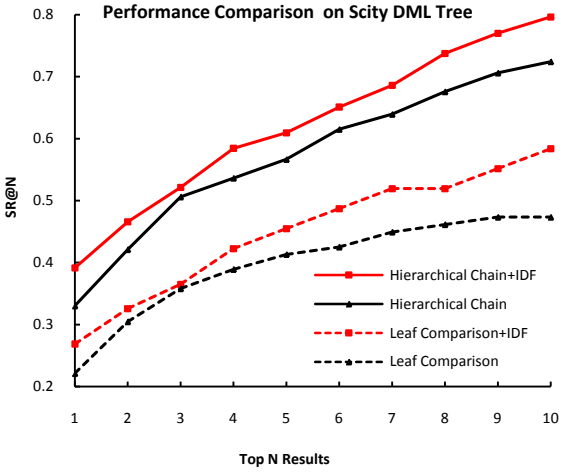

Figure 2: Scity DML-based Tree Test

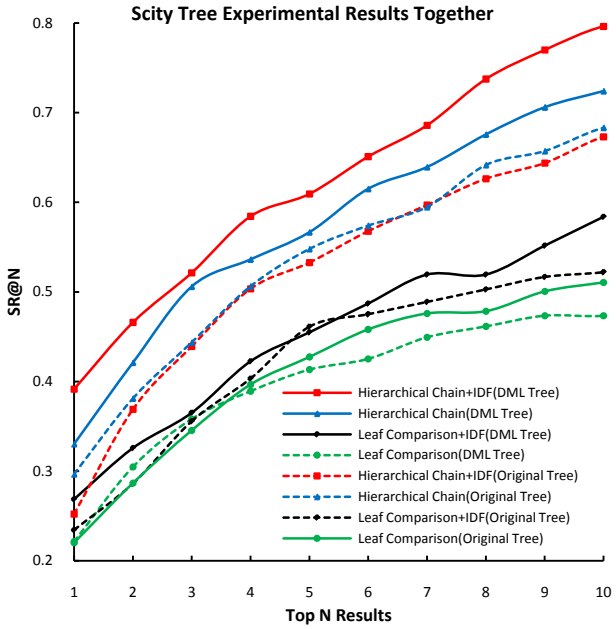

Figure 3: Fuse above Two Figures Together

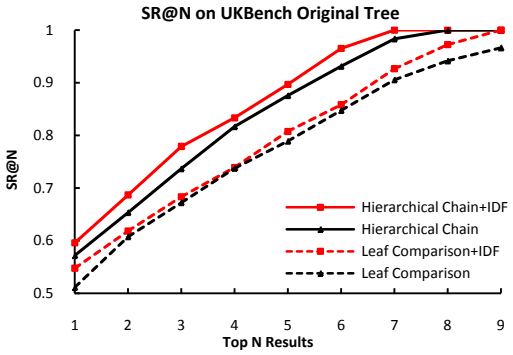

Figure 4: UKBench Original Tree Test

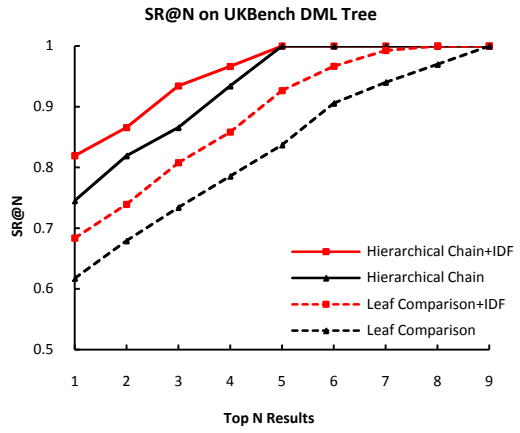

Figure 5: UKBench DML-based Tree Test

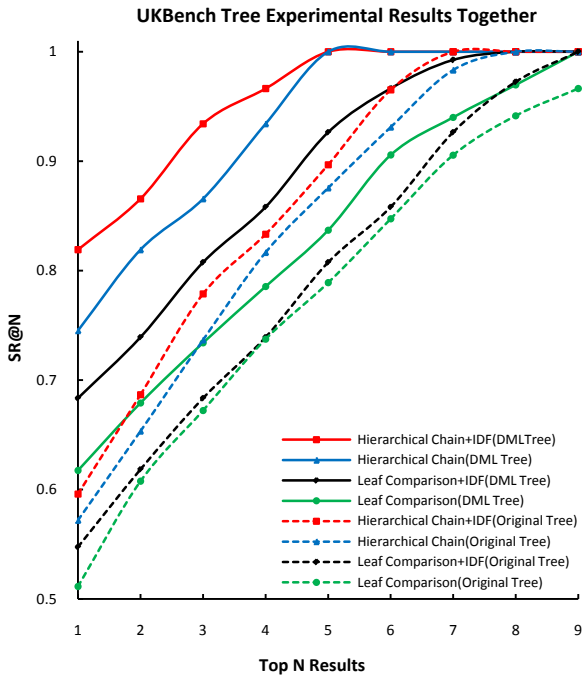

Figure 6: Fuse above Two Figures Together