PRUNING TREE-STRUCTURED VECTOR QUANTIZER TOWARDS LOW BIT RATE MOBILE VISUAL SEARCH

Jie Chen, Ling-Yu Duan, Rongrong Ji, and Wen Gao

The Institute of Digital Media, School of CS & EE, Peking University, Beijing 100871, China
e-mail:{cjie,lingyu,rrji,wgao}@pku.edu.cn

ABSTRACT

Coming with the proliferation of mobile devices, mobile visual search emerges. One fundamental issue here is the query transmission latency, especially in a bandwidth constraint wireless link. Towards low bit rate retrieval, recent works have proposed to extract compact visual descriptors directly on the mobile end, where the vocabulary tree based bag-of-words representation has shown superior performance in producing compact descriptors [2][9]. However, the corresponding tree-structure vector quantizer is extremely large against a mobile end implementation. In this paper, we propose two alternatives to prune this tree structure based on the subtree discriminability analysis, where either *information gain based* or *ranking based* pruning are investigated. Furthermore, we have unveiled that the tree structure can be even discarded while retaining only the discriminative leaves together with their radii in practice. We evaluate our tree pruning on Android HTC Desire G7, with application to low bit rate mobile landmark search in a 10-million landmark photo collection, where over 10 scale memory reduction with almost identical search accuracy is reported.

Index Terms— Mobile Visual Search, Compact Visual Descriptors, Visual Vocabulary, Tree Pruning

1. INTRODUCTION

Handheld mobile devices, such as smart phones, have great potential for emerging mobile visual search applications. In this scenario, the reference images are typically stored on the remote server. Therefore, online query should be transmitted from the mobile device to the remote server and the query image has to be transmitted over a relatively slow wireless link, which therefore heavily degenerate the user experience.

With the ever growing computational power on the mobile device, recent works have proposed to directly extract compact yet discriminative visual descriptors [2][4][6][9] to be subsequently transmitted in a low bit rate. Such compact visual descriptors should be also efficient in terms of the online extraction complexity and cast light memory footprint due to the memory limitation of most mobile devices.

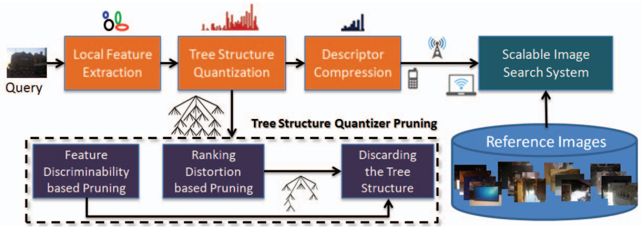
Query
Local Feature Extraction
Tree Structure Quantization
Descriptor Compression
Scalable Image Search System
Tree Structure Quantizer Pruning
Feature Discriminability based Pruning
Ranking Distortion based Pruning
Discarding the Tree Structure
Reference Images

Fig. 1. Tree-structure quantizer pruning towards efficient descriptor extraction on the mobile end.

Recent works on low bit rate descriptor for mobile visual search can be categorized in two directions: (1) *local descriptor based*, where both CHoG [4] and SIFT compression [6] schemes are well exploited in the literature. (2) *Bag-of-Words based*, which is a more compact solution that further quantizes local descriptors into a Bag-of-Words (BoW) histogram followed by lossless or lossy histogram compression at the mobile end [2][9]. The latter works typically reply on the vocabulary tree [1] based tree-structure quantizer in the mobile-end feature extraction.

One practical and fundamental issue in BoW based compression comes from the size of this tree-structure quantizer. For instance, as shown in [1][2][9], tuning the best search accuracy typically involves 0.1-1 million visual words. As a result, this tree-structure vector quantizer (TSVQ) would cost over 60MB main memory of a mobile device while running. However, many smart phones have memory limit per process or application, *e.g.*, the limit is 16Mbyte for 1st generation Android phones and 24Mbyte for 2nd generation phones.

To shrink memory footprint, we have proposed to prune the tree-structure quantizer without any serious loss in the subsequent compact descriptor extraction and search accuracy. As shown in Figure 1, our tree pruning paradigm are three-fold: First, we propose to prune the subtree in the quantizer via analyzing the subtree discriminability using an information gain principle. Second, we look at the ranking distortion of removing each subtree, based on which trying to remove subtrees with minimal effects in search accuracy degeneration. Finally, we discard the tree-structure quantizer and maintain the remaining words instead, together with their corresponding radii in the local feature space. Quantitative evaluations on a 10-million landmark image collection have shown our pruning effectiveness as well as a comparable

978-1-4673-0046-9/12/$26.00 ©2012 IEEE

search accuracy over alternatives.

The rest of this paper is organized as follows: Section 2 introduces the visual vocabulary based search preliminary. Section 3 presents our proposed tree-structure pruning paradigm. We give quantitative comparisons in Section 4 and conclude this paper in Section 5

2. VISUAL VOCABULARY PRELIMINARY

TSVQ Quantizer for Scalable Visual Search: Towards efficient mobile visual search in a low bit rate scenario, recent works [2][3][8] have reported superior compression rate by using a vector quantizer based on a vocabulary tree model [1]. Vocabulary tree adopts hierarchical k-means to partition local descriptors into quantized codewords. An H-depth vocabulary tree $\mathbf{V}$ with B-branch produces $M = B^H$ words as in [1]. Given a query photo $\mathbf{I}_q$, we first extract local descriptors $\mathbf{S}_q = [S_1^q, S_2^q, ..., S_J^q]$ in the mobile end. Then, we adopt $\mathbf{V}$ to quantize $\mathbf{S}_q$ by traversing in the tree-structure to find the nearest word (also in the mobile end), which converts $\mathbf{S}_q$ into a high-dimensional BoW signature $\mathbf{V}_q = [V_1^q, V_2^q, ..., V_M^q]$. The $\mathbf{V}$ is then lossless or lossy compressed using some state-of-the-art techniques [2][9]. And finally, the resulting compact histogram is transmitted to the remote server.

In the server end, a desirable ranking tries to minimize the following ranking loss with respect to the ranking position $R(x)$ of each photo $\mathbf{I}_x$ in an N-photo database:

$$Loss_{Rank} = \sum_{x=1}^{N} R(x)\mathbf{W}_x \|\mathbf{V}_q, \mathbf{V}_x\|_{Cosine} \tag{1}$$

where TF-IDF weighting vector $\mathbf{W}_x$ can be also incorporated to refine the score of each reference image $\mathbf{I}_x$.

3. PRUNING TSVQ TREE

Subtree Feature Discriminability Based Pruning: Our first solution is based on the feature discriminability of each subtree, which is define as the ensemble of all its containing leaf nodes (visual words defined). Taking mobile location search for instance, for a give word W_j, we define its discriminability for location L_i as:

$$MI(L_i|W_j) = H(L_i) - H(L_i|W_j) \tag{2}$$

where $MI(L_i|W_j)$ is the information gain of recognizing location L_i by using word W_j. $H(L_i)$ is the entropy of location L_i, set as proportional to the number of reference images in L_i. $H(L_i|W_j)$ is the conditional entropy of L_i given W_j as:

$$H(L_i|W_j) = -log(\frac{N_{W_j,L_i}}{N_{W_j,L_i} + N_{W_j,\bar{L}_i}}) \tag{3}$$

where N_{W_j,L_i} is the times of word W_j happening in the reference images of location L_i, while $N_{W_j,\bar{L}_i}$ is the times of word W_j happening in other locations rather than L_i. As a result, the overall discriminability of a given $Subtree(k)$ is defined as:

$$MI_{Subtree(k)} = \sum_{W_j \in Subtree(k)} \sum_{i \in N} p(L_i) MI(L_i|W_j) \tag{4}$$

That is, for a given $Subtree(k)$ we scan all its words $W_j \in Subtree(k)$. For each word we measure its information gain for all reference locations (in total N, $p(L_i)$ is the proportion of reference images at location i over the database). As a result, if $MI_{Subtree(k)}$ is less than an empirical threshold τ we prune this $Subtree(k)$.

Algorithm 1: Ranking distortion based subtree pruning.

```
Input:Initial vocabulary $\mathbf{V}$; pruning threshold $\varepsilon$,
    sampled query images $\{Query(\mathbf{I'}_r)\}_{r=1}^{R}$; An empty set of
    pruned subtrees $\mathbf{P}$.
Output: The retained subtree collection $\mathbf{V} - \mathbf{C}$.
while Breadth-first-search not ended do
    Obtain $subtree(k)$ and all its leaf node words;
    Calculate $Loss_{Rank}$ by Equation 7.
    if $Loss_{Rank} < \varepsilon$ then
        Add $subtree(k)$ into $\mathbf{P}$;
    end
    $k + +$;
end
```

We visit each $Subtree(k)$ sequentially in a breadth-first-search scanning. Based on its *sum up* characteristic, the subtrees at lower level (far away from the root node) are more unlikely to bring about benefits in search accuracy. As a result, if a $subtree(k)$ is pruned, all its subtrees would be also pruned[1].

Subtree Ranking Distortion based Pruning: Our second solution is to prune each $Subtree(k)$ based on the distortion in retrieval performance due to the removal words of $Subtree(k)$ from the BoW representation.

We first randomly sample a set $[\mathbf{I}'_1, ..., \mathbf{I}'_{n_{sample}}]$ of query images from the reference database, which output the following ranking list:

$$\begin{array}{rcl} Query(\mathbf{I'}_1) & = & [\mathbf{A}_1^1, ..., \mathbf{A}_R^1] \\ ... & = & ... \\ Query(\mathbf{I'}_{n_{sample}}) & = & [\mathbf{A}_1^{n_{sample}}, ..., \mathbf{A}_R^{n_{sample}}] \end{array} \tag{5}$$

where $\mathbf{A}_i^j$ is the ith returning image of the jth query.

The above queries and returning results are used as the ground truth in the subsequent pruning. An ideal pruning $Subtree(k)$ is supposed to maintain comparable retrieval results $[\mathbf{A}_1^j, ..., \mathbf{A}_R^j]$ for each jth query using $\mathbf{V} - Subtree(k)$. Therefore, we deal with an optimal subtree pruning in an iterative procedure to determine the pruning of each subtree

[1] As a greedy search, the nodes at the lower level would probably be removed by mistake, leading to a sub-optimal pruned TSVQ tree. So that we empirical set larger τ at lower levels and smaller τ at higher levels.

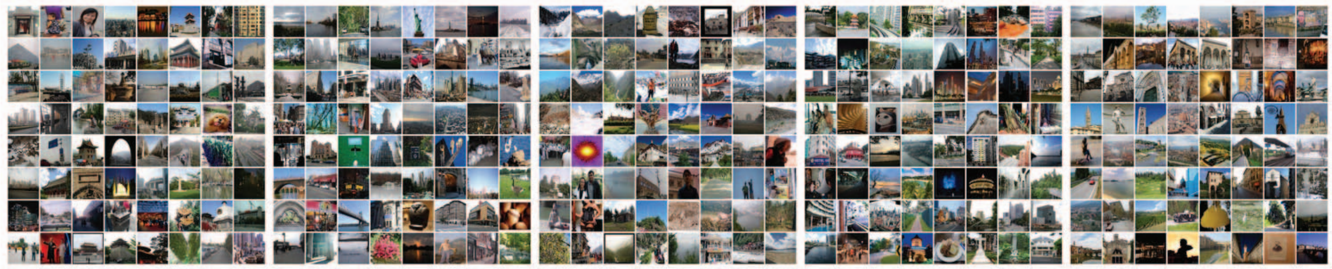

Fig. 2. Examples in our 10-million photo collection from Beijing, New York, Lhasa, Singapore, and Florence.

$Subtree(k)$ in a breadth-first-search manner, and then check the search accuracy distortion, as outlined in Algorithm 1.

We denote $\mathbf{P}$ as the set of pruned subtrees after $(t-1)$ rounds. $\mathbf{V}$ is the entire vocabulary and **V-P** is the retained vocabulary after removing the codeword union of P from V At the tth iteration, we determine whether to remove the subtree $subtree(k)$ by the ranking loss of each image:

$$Loss(\mathbf{I'}_i) = \sum_{r=1}^{R} Rank(\mathbf{A}_r^1)||\mathbf{V\text{-}P\text{-}subtree(k)}_{\mathbf{I'}_i}, \mathbf{V}_{\mathbf{A}_r^i})||_{Cosine} \quad (6)$$

where $i \in [1, n_{sample}]$; $Rank(\mathbf{A}_r^i)$ is the current position of the (originally) ith returning using query $\mathbf{I'}_i$; The $subtree(k)$ is pruned if and only if:

$$Loss_{Rank} = \sum_{i=1}^{n_{sample}} Loss(\mathbf{I'}_i) < \varepsilon \quad (7)$$

where ε is also an empirical threshold to trigger the subtree pruning operation.

Discarding The Entire Tree Structure: Our final simplification is to discard the tree structure, by just maintaining the leave nodes in all retained subtrees $\mathbf{V} - \mathbf{P}$ together with their radii. The radius $Radius_i$ of each W_i is a scaler as:

$$Radius_i = \min_d \max \forall_{W_j} ||W_i, W_j||_2 < d \quad (8)$$

As a result, the memory footprint includes |**V-P**| visual words (*e.g.* 128-dim SIFT descriptor or 60-dim CHoG descriptor [4]) together with |**V-P**| scalar values of their radii.

4. EVALUATIONS

Data Collection: We choose mobile location search application to evaluate the functionality of our TSVQ pruning solutions. We collect over 10 million geo-tagged photos from photo sharing websites of Flickr and Panoramio. This dataset covers cities including Beijing, New York City, Lhasa, Singapore, and Florence. Figure 2 shows some exemplar photos in our 10 million photo collection.

Ground Truth Labeling: From the geographical map of each city, we choose 30 densest regions and 30 random regions. Since manually identifying all the related photos of each landmark is intensive, for each of these 60 regions, we ask volunteers to manually identify one or more dominant views. Then, all near-duplicate landmark photos are labeled in its current and nearby regions. Finally, we sample 5 images from each region to form the ground truth, which result to totally 300 user query logs for each city.

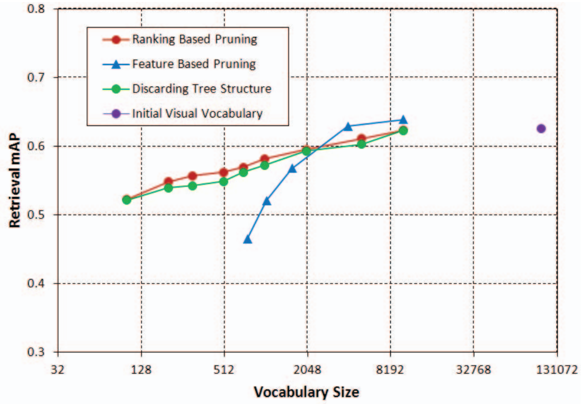

Fig. 3. Vocabulary size vs. search accuracy distortion using different pruning methods as well as the baseline from the initial large vocabulary.

Parameters and Evaluations: From our 10-million landmark photo collection, we extract SIFT [7] features from each photo. Then, we adopt the Vocabulary Tree [1] to build the initial tree-structure quantizer $\mathbf{V}$, which generates a bag-of-words signature $\mathbf{V}_i$ for each database photo $\mathbf{I}_i$. As mentioned before, we denote the hierarchical level as H and the branching factor as B. In a typical settlement, we have $H = 6$ and $B = 5$, which produces approximate 0.1 million words. We use **m**ean **A**verage **P**recision (**mAP**) to evaluate our search performance, which reveals its position-sensitive ranking precision in all ground truth reference image for a give query.

Baselines: (1). *Initial Visual Vocabulary*, which maintains the entire vocabulary tree structure with the memory footprint upper bound. upper bound. (2). *Feature based Pruning*, which is the first alternative based on information gain without considering ranking loss in pruning decision. (3). *Ranking based Pruning*, which is the second alternative by evaluating whether or not the pruning would generate the search accuracy. (4). *Discarding Tree Structure*, which is the final solution to maintain the most discriminative words with their corresponding radii in the local feature space.

For all baselines, we adopt the Tree Histogram Coding [2]

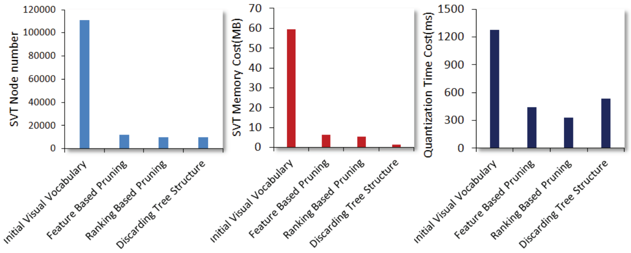

Fig. 4. Memory and time cost comparison. Left to right: SVT node number, memory cost and online quantization time.

to achieve the bag-of-words level compression to output the final histogram to be transmitted through the wireless link, without any loss in the search accuracy.

Quantitative Results: Similar to the comparisons in many video coding systems, we give the *rate distortion analysis* in comparisons to (1) maintaining the original tree-structure without pruning and (2) the state-of-the-art works of [5] in terms of the memory requirement to maintain the feature extraction component. Here, "*rate*" changes is tuned by varying τ (feature discriminability based pruning) or ε (ranking based pruning) in our pruning. And the "*distortion*" is measured in terms of the mAP based search accuracy degeneration. Figure 3 shows that by both feature based and ranking based pruning, our quantizer only requires minimal memory cost to run feature extraction in the mobile end, while maintaining an almost identical search accuracy.

In addition, by further replacing the tree-structure (as our final pruning simplification) by the retained words and their corresponding radii, we can even achieve 10 scale compression rate with less than 10% loss in our search accuracy as in Figure 4. It turns out to be a very practical solution in many scenarios where the light memory cost is more preferable[2]. As in Figure 4, the time costs of online quantization using the original tree quantizer is also largely reduced (over 50%) by using our pruned trees or retaining words. Finally, Figure 5 shows the snapshot of low bit rate mobile visual search system deployed in HTC Desire G7 using the pruned tree-structure quantizer, where less than 10 MB memory are required to maintain a million scale codebook based on our pruning simplification.

5. CONCLUSION

We have presented a novel tree pruning paradigm towards memory light bag-of-words extraction in the mobile end, which extremely suits the emerging scenario of low bit rate mobile visual search *i.e.* compact descriptors are directly extracted at the mobile end to reduce the upstream query delivery delay. Our main idea is to prune the vocabulary tree [1] based quantizer that is well used in the state-of-the-art compact descriptor extractions [2][9]. We adopt a breadth-first-search over the subtrees in the quantizer, where the subtree discriminability is evaluated based on the discriminability of its contained words. We propose two alternatives towards subtree discriminability evaluation, including (1) *feature based approach*: information gain based discriminative subtree selection and (2) *ranking based approach*: simulated pruning subtrees to evaluate the ranking order distortions of sampled queries. We further discard the original tree structure by maintaining only the retained words together with their radii. Quantitative results on mobile visual search over 10 million landmark image collection show that, comparing to maintaining the entire tree-structure quantizer, our proposed scheme has achieved almost identical search accuracy with a memory reduction over a scale of 10.

Fig. 5. The snapshot of the implementation scenario for mobile visual search scenario using Android smart phone.

6. ACKNOWLEDGEMENT

This work was supported by National Basic Research Program of China (2009CB320902), in part by Chinese National Nature Science Foundation (60902057) and CADAL Project Program.

7. REFERENCES

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

[2] D. Chen, S. Tsai, and V. Chandrasekhar, *et al.* Tree histogram coding for mobile image matching. *DCC*. 2009. 1, 2, 3, 4

[3] D. Chen, S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod. Inverted index compression for scalable image matching. *DCC*. 2010. 2

[4] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, and B. Girod. CHoG: Compressed histogram of gradients a low bit-rate feature descriptor. *CVPR*. 2009. 1, 3

[5] H. Jegou, M. Douze, C. Schmid, P. Perez. Aggregating local descriptors into a compact image representation. *CVPR*. 2010. 4

[6] V. Chandrasekhar, G. Takacs, D. Chen, *et al.* Transform coding of image feature descriptors. *VCIP*. 2009. 1

[7] D. G. Lowe, Distinctive image features from scale-invariant keypoints. *IJCV*. 2004. 3

[8] G. Schindler and M. Brown. City-scale location recognition. *CVPR*. 2007. 2

[9] R. Ji, L.-Y. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, and W. Gao. Location discriminative vocabulary coding for mobile landmark search. *IJCV*. 2011. 1, 2, 4

[2] Note that the tree quantizer is the major part of the memory footprint in the mobile-end compact feature extraction.