WHEN CODEWORD FREQUENCY MEETS GEOGRAPHICAL LOCATION

Rongrong Ji[⋆†] *Ling-Yu Duan*[⋆] *Jie Chen*[⋆] *Hongxun Yao*[†] *Wen Gao*[⋆†]

[⋆]Institute of Digital Media, Peking University, Beijing 100871, China
[†]Visual Intelligence Laboratory, Harbin Institute of Technology, Heilongjiang, 150001, China

ABSTRACT

When codeword frequency meets geographical location in landmark search applications, is it still discriminative for the search procedure? In this paper, we give a systematic investigation about how geographical location affects the effectiveness of codeword frequency. We explain why the standard IDF in the BoW models is less effective in location related search applications [11][12]. Consequently, we propose a "location discriminative codeword frequency " strategy to introduce the location context into the codeword discriminability measurement. This new codeword frequency is calculated in each geographical region, for which a spectral clustering scheme is proposed to partition the geographical map of each city into distinct regions. Extensive comparisons over the standard codeword frequency in state-of-the-art landmark search systems [1][1] demonstrates our approach's effectiveness.

Index Terms— landmark search, mobile search, visual vocabulary, codeword frequency, geographical clustering

1. INTRODUCTION

Coming with the popularization of mobile devices, the success in patch-based image retrieval [1][6] facilitates many landmark search applications [11][12]. Generally speaking, given a landmark database, where each photo is bound with a GPS location, we have the typical scenario: In the server-end computer(s), images are offline inverted indexed using patch quantization schemes such as Vocabulary Tree [1] and its variances [7][11][12]. In the user-end client, a landmark query is captured using camera-embedded PDA or mobile phone. A photo or its visual descriptors [2][3][5][4] is transmitted to the server, where its Bag-of-Words (BoW) feature is built to search near-duplicated landmark photos.

This paper focuses on the effectiveness of codeword frequency in the context of landmark search. Codeword frequency is a universal setting in most existing landmark search systems [11][12][13][10]. Given a BoW histogram extracted (or transmitted) from the user query, codeword frequency such as Term Frequency (TF) [14], Inverted Document Frequency (IDF) [14], and Mutual Information (MI) [10], aims to distinguish the contributions of different codewords in the subsequent similarity ranking. This is usually achieved by applying an importance weighting to the original BoW histogram distance (L2 or Cosine).

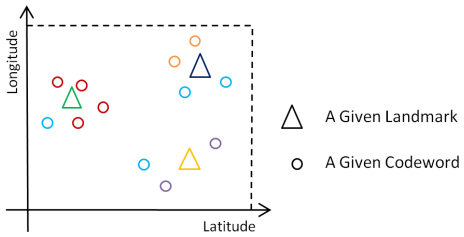

Fig. 1. A toy example of why codeword frequency is suboptimal in landmark search.

While the standard codeword frequency aims to measure the visual discriminability of a given codeword, the frequency used in landmark search should emphasize more on the landmark queries. In such context, the contribution of a given codeword not only relies on the visual discriminability of its quantized local patches, but also depends on how it can distinguish different geographical landmarks in the subsequent ranking. Figure 1 shows a toy example: To find the "Green" landmark, the "red" codeword is much more discriminative than the "blue" codeword, even both codewords have an identical frequency. As the "red" codeword concentrates more on the "Green" landmark, "red" codeword is a more discriminative feature for the image ranking. On the contrary, the "blue" codeword is uniformly distributed over all landmarks, making it less discriminative, even if it has higher frequency. Figure 1 indicates the existing codeword frequency is not specified for the landmark search or other location sensitive applications, as it fails to incorporate the geographical locations into the codeword discriminability estimation.

In this paper, we give a systematic evaluation on how location cues influence the codeword frequency effectiveness. We show that the existing codeword frequency outputs suboptimal performances in the context of landmark search. It in turn explains why previous works cannot achieve satisfactory results by adding codeword frequency into the BoW distance [11][12]. Towards optimal codeword frequency calculation, we propose a **L**ocation **D**iscriminative **C**odeword **F**requency (**LDCF**), which outperforms the standard frequency with a large margin. To enable location discriminative frequency

measurement, we further introduce a spectral clustering scheme to partition the geographical map into discrete regions. In online search, once a mobile user enters a given geographical region, the remote server identifies its current location (based on his GPS or based station signals), subsequently makes use of the corresponding LDCF in this region for an optimal landmark search.

2. SUBOPTIMAL CODEWORD FREQUENCY IN LANDMARK SEARCH

Towards efficient landmark search in a million scale database, the Scalable Vocabulary Tree (SVT) [1] is well exploited in previous works [2][3][10][11]. SVT uses hierarchical k-means to partition local descriptors into quantized codewords. A H-depth SVT with B-branch produces $M = B^H$ codewords, and the scalable search typically settles $H = 6$ and $B = 10$ [1]. Given a query photo $\mathbf{I}_q$ with local descriptors $\mathbf{S}_q = [S_1^q, S_2^q, ..., S_J^q]$, SVT quantizes $\mathbf{S}_q$ by traversing in the vocabulary hierarchy to find the nearest codeword, which converts $\mathbf{S}_q$ to a BoW signature $\mathbf{V}_q = [V_1^q, V_2^q, ..., V_M^q]$. In search, the ranking aims to minimize the following ranking lost with respect to the ranking position $R(x)$ of each photo $\mathbf{I}_x$ in a n-photo database:

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

where TF-IDF weighting is calculated similar to its original form [14] in the document retrieval as:

$$\mathbf{W}_x = [\frac{n_1^x}{n^x} \times log(\frac{N}{N_{V_1}}), ..., \frac{n_M^x}{n^x} \times log(\frac{N}{N_{V_M}})] \tag{2}$$

n^x denotes the number of local descriptors in $\mathbf{I}_x$; $n_{V_i}^x$ denotes the number of local descriptors in $\mathbf{I}_x$ quantized into V_i; N denotes the total number of images in the database; N_{V_i} denotes the number of images containing V_i; $\frac{n_i^x}{n^x}$ serves as the term frequency of V_i in $\mathbf{I}_x$; and $log(\frac{N}{N_{V_i}})$ serves as the inverted document frequency of V_i in the database.

One important issue lies in: IDF in Equation 2 should operate at the landmark level: If we can identify which photo belongs to which landmark, we should calculate codeword frequency for each V_i as:

$$IDF_{Landmark} = log(\frac{N}{N'_{V_i}}) \tag{3}$$

where N'_{V_i} refers how many landmarks (not photos) containing the codeword V_i; However, tagging every photo for the entire database is extremely difficult if not impossible in a large-scale database.

3. LOCATION DISCRIMINATIVE FREQUENCY

We propose a refined"**L**ocation **D**iscriminative **C**odeword **F**requency" (**LDCF**), which largely improves the landmark ranking precision in online search. We'd like to replace the IDF weighting in $\mathbf{W}_{TF-IDF}$ to truly distinguish contributions of codewords in locating landmarks: A codeword that is geo-scatted over the entire region is less discriminative, comparing with a codeword that is geo-concentrated in a given landmark location, even with an identical IDF. Our LDCF incorporates such geographical distribution concentration measurement to re-estimate the codeword discriminability.

For a given codeword V_i, we incorporate the geo-distances among images containing V_i to re-estimate its discriminability. The original IDF for V_i is calculated by:

$$IDF^i_{Original} = \frac{N_{Region}}{N_i} \tag{4}$$

N_{Region} denotes the number of photos in the current region R; and N_i denotes the number of photos in the current region and contain V_i. Our $LDCF^i$ refines Equation 4 by:

$$LDCF^i = \frac{\sum_{\mathbf{I}_m \in R} \sum_{\mathbf{I}_n \in R} Dis_{Geo}(\mathbf{I}_m, \mathbf{I}_n)}{\sum_{\mathbf{I}_m \in R,\ \mathbf{V}_i \in \mathbf{I}_m} \sum_{\mathbf{I}_n \in R,\ \mathbf{V}_i \in \mathbf{I}_n} Dis_{Geo}(\mathbf{I}_m, \mathbf{I}_n)} \tag{5}$$

where $Dis_{Geo}(\mathbf{I}_m, \mathbf{I}_n)$ denotes their geographical distance, measured by L2 distance of their corresponding geographical location; $\mathbf{I}_m \in R$ denotes image $\mathbf{I}_m$ falling into the current geographical region; $\mathbf{V}_i \in \mathbf{I}_m$ denotes image $\mathbf{I}_m$ containing V_i.

Therefore, a codeword that is distributed in a more concentrated geographical location is more likely to produce a higher LDCF, and vice versa.

4. GEOGRAPHICAL CITY PARTITION

LDCF calculation is location-sensitive, so we should estimate LDCF within each geographical region respectively. Therefore, we partition the geo-tagged image database into discrete regions in each city. Note that we should avoid the incorrect division of an identical landmark into different regions to improve the effectiveness of LDCF: We assume that photos containing an identical landmark are geographically nearby and visually similar. Therefore, a *Visual Aware Spectral Clustering* is proposed to partition each city into visually coherent geographical regions.

Suppose there are in total N photos in a given city, we first construct a N-node fully connected graph $\mathbf{G}$. Each node g_i represents a photo, and each link l_{ij} denotes the (visual + geographical) distance between g_i and g_j. Then, our goal is to partition $\mathbf{G}$ into L subgraphs $\{\mathbf{G}'_l\}$ where $l \in [1, L]$. As graph partition retains NP hard, we resort to a spectral clustering to handle partition, which is proven to be equivalent to the normalized graph cut in [15].

We first build a geographical similarity matrix $\mathbf{A}$, where $\mathbf{A}_{i,j}$ measures the geographical distance between g_i and g_j. To

Algorithm 1: Visual Aware Spectral Clustering to Partition Geographical Regions

Input: Geographical Distance Matrix $\mathbf{A}_{N\times N}$.
Output: Spectrum Clustering Graph $\mathbf{S}_{N\times L}$.
ε-Ball Operation on $\mathbf{A}_{N\times N}$ using Equation 6;
Build Laplasians Graph $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$;
Spectral Graph Construction by SVD over $\mathbf{L}$ into $\mathbf{S}_{N\times L}$, with eigenvectors $[e_1, e_2, ..., e_L]$;
Clustering N rows in $\mathbf{S}_{N\times L}$ into L clusters with $Sim(\mathbf{S}_i, \mathbf{S}_j) = \mathbf{W}_{LDCF}\|BoW_i, BoW_j\|_{Cosine}\|\mathbf{S}_i, \mathbf{S}_j\|_2$;
Return Spectrum Clustering Graph $\mathbf{S}_{N\times L}$;

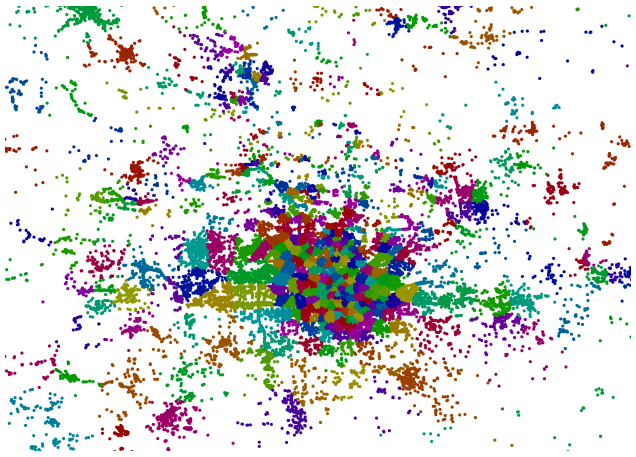

Fig. 2. The visual aware spectral clustering partitions Beijing into geographical regions.

ensure only nearby photos are grouped into the identical region, we use a ε-ball to disconnect far-away photos in $\mathbf{A}_{N\times N}$:

$$\mathbf{A}_{N\times N} = \begin{cases} \mathbf{A}_{i,j}, & \mathbf{A}_{i,j} < \varepsilon; \\ \infty, & \mathbf{A}_{i,j} \geq \varepsilon. \end{cases} \quad (6)$$

Then, we build a diagonal matrix $\mathbf{D}$ whose (i,i)-element is the sum of $\mathbf{A}$'s ith row ($d_k = \sum_{n=1}^{N} \mathbf{A}_{k,n}$), based on which a Laplasians matrix $\mathbf{L}$ is built via $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$. Subsequently, we extract the L largest eigenvectors $[e_1, e_2, ..., e_L]$ from $\mathbf{L}_{N\times N}$. They transform $\mathbf{L}_{N\times N}$ into a spectral matrix $\mathbf{S}_{N\times L}$, in which each row $\mathbf{S}_i$ is a L-dimensional eigenvector in $\mathbb{R}_L$ (normalized with a unit length).

We then incorporate the visual similarity into the clustering of each row $\mathbf{S}_i$ and $\mathbf{S}_j$ in $\mathbf{S}_{N\times L}$ (L-dimensional):

$$Sim(\mathbf{S}_i, \mathbf{S}_j) = \mathbf{W}_{LDCF}\|BoW_i, BoW_j\|_{Cosine}\|\mathbf{S}_i, \mathbf{S}_j\|_{Cosine} \quad (7)$$

Algorithm 1 shows the overall clustering process, and exemplar partitions in Beijing are given in Figure 2.

5. IMPLEMENTATIONS AND RESULTS

Data Collection: We collected over 10 million geo-tagged photos from photo sharing websites of Flickr and Panoramio.

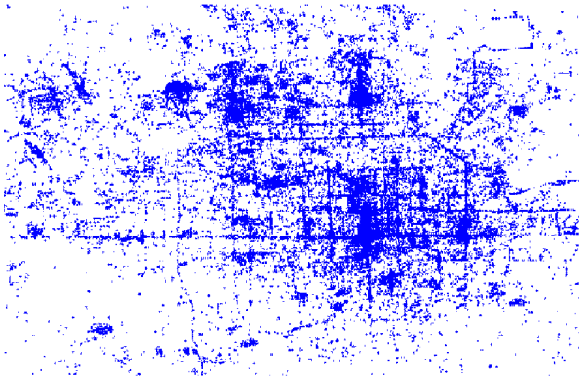

Fig. 3. The geographical distribution of community contributed photos with geographical tagging in Beijing.

Our dataset covers typical areas including Beijing, New York City, Lhasa, Singapore, and Florence. Figure 3 shows the typical photo distribution in the geographical map of Beijing. Typically speaking, such data can naturally outline the human photographing activity (such as user traveling manners on the roads) from their photographing viewpoint.

Labeling Query Ground Truth: From the geographical map of each city, we selected the top 30 most dense regions and 30 random regions. Since manual identifying all related photos of an identical landmark is intensive, for each of these 60 regions, our volunteers manually identify one or more dominant landmark views. All near-duplicated landmark photos are labeled in its current and nearby regions. Then, we sample 5 images from each region to form the ground truth. It generates 300 queries with ground truthes in each city.

Parameter and Evaluation: For the landmark photo collection in Beijing, we extract both SIFT [9] and CHoG [8] features from each photo. Then, we build a Scalable Vocabulary Tree [1] to generate the initial Vocabulary $\mathbf{V}$, which generates a bag-of-words signature $\mathbf{V}_i$ for each database photo $\mathbf{I}_i$. We use the vocabulary generated in Beijing to search all five cities, for each of which the boosting is carried out to build the $\mathbf{M}$ transformation. We denote the hierarchical level as H and the branching factor as B. In a typical settlement, we have $H = 6$ and $B = 10$, producing approximate 100,000 codewords. We use Mean Average Precision at N ($MAP@N$) to evaluate our system performance, which reveals its position-sensitive ranking precision in the top N positions.

Spectral Clustering Tuning: Our geographical partition involves two factors: (1) the visual discriminability embedding; (2) the number of eigenvectors in our clustering. Figure 4 shows that, with visual embedding, we achieve better search performance by maintaining identical amount of highest IDF and LDCF codewords. The influence of different eigenvector selections in Beijing is also shown. In each city, we leverage the similar scheme of Figure 4 to determine the best region partition number. Smaller regions typically give higher retrieval performance, but meanwhile, there is a

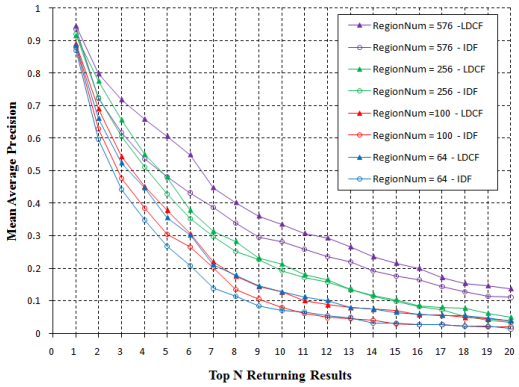

Fig. 4. The influence of both visual discriminability embedding and (different eigenvector volumes (region numbers)) in Visual Aware Spectral Clustering, measured by MAP performance using our ground truth query set.

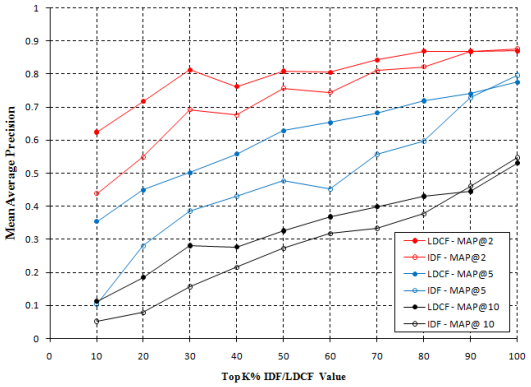

Fig. 5. The retrieval MAP lost by simply maintaining codewords with the top $k\%$ IDF or top $k\%$ LDCF values.

higher probability of incorrect matching by falsely mapping marginal photos into the nearby regions.

LDCF vs. IDF: Figure 5 investigates whether our LDCF is more discriminative than the standard IDF in finding important codewords to facilitate landmark search in each geographical region. In Figure 5, we maintain the top 0.1-0.9 codewords with either highest IDF or LDCF. Then we measure the MAP degeneration using only these codewords.

From Figure 5, it is obvious that our LDCF did better job in finding the discriminative codewords. The explanation lies in that the standard IDF ignores the geographical distributions of the visual words. That is, a frequent codeword with diverse geographical distributions is less discriminative. On the contrary, a codeword with identical or even less (original) frequency is more discriminative, once it is concentrated within a certain landmark location (In other words, it can identify this landmark; meanwhile it does not disturb the rankings of other landmarks). In fact, this is a commonsense for many popular landmarks, where geographically nearby photos tend to present near-duplicated visual appearances. Hence, the LDCF codewords are more discriminative in landmark search.

6. CONCLUSIONS

We have investigated how the geographical locations influence the visual codeword frequency in the context of landmark search. An important finding is that the standard codeword frequency can be further optimized by incorporating the pervasive geographical tags of database photos. We have introduced a **L**ocation **D**iscriminative **C**odeword **F**requency (**LDCF**) scheme for codeword discriminability measurement. Extensive experiments have shown LDCF's superior performance over the standard IDF setting . The proposed LDCF scheme can be easily implemented and plugged into most of the state-of-the-art location sensitive visual search systems that employ the standard IDF, such as works in [2][3][8][11][12].

7. ACKNOWLEDGEMENTS

This work was supported in part by grants from the Chinese National Natural Science Foundation under contract No. 60902057, in part by the National Basic Research Program of China under contract No. 2009CB320902, and in part by the CADAL Project Program.

8. REFERENCES

[1] Nister D. and Stewenius H. Scalable recognition with a vocabulary tree. *CVPR*. 2006.
[2] Chen D., Tsai S., and Chandrasekhar V. Tree histogram coding for mobile image matching. *DCC*. 2009.
[3] Chen D., Tsai S., Chandrasekhar V., Takacs G., Vedantham R., Grzeszczuk R., and Girod B. Inverted index compression for scalable image matching. *DCC*. 2010.
[4] Zheng Y., Zhao M., Song Y., and Adam H. Tour the world: building a web-scale landmark recognition engine. *CVPR*. 2009.
[5] Zamir A. and Shah M. Accurate image localization based on Google maps street view. *ECCV*. 2010.
[6] Sivic J. and Zisserman A. Video Google: a text retrieval approach to object matching in videos. *ICCV*. 2003.
[7] Philbin J., Chum O., Isard M., Sivic J., and Zisserman A. Obj. retrieval with large voc. and fast spatial matching. *CVPR*. 2007.
[8] Chandrasekhar V., Takacs G., Chen D., Tsai S., Grzeszczuk R., and Girod B. CHoG: Compressed histogram of gradients a low bit-rate feature descriptor. *CVPR*. 2009.
[9] Lowe D. G. Distinctive image features from scale-invariant keypoints. *IJCV*. 2004.
[10] Irschara A., Zach C., Frahm J., Bischof H. From s.-f.-m. point clouds to fast location recognition. *CVPR*. 2009.
[11] Schindler G. and Brown M. City-scale location recognition. *CVPR*. 2007.
[12] Ji R., Xie X., Yao H., and Ma W.-Y. Hierarchical optimization of visual vocabulary for effective and transferable retrieval. *CVPR*. 2009.
[13] Ji R., Xie X., Yao H., and Ma W.-Y. Vocabulary tree incremental indexing for scalable scene recognition. *ICME*. 2008.
[14] Salton G. and Buckley C. Term-weighting approaches in automatic text retrieval. *Info. Proc. and Management*. 1988.
[15] Ng A., Jordan M., and Weiss Y. On spectral clustering: Analysis and an algorithm. *NIPS*. 2001.