Multi-stage image retrieval based on feature augmentation with truncated polynomial weight

Keundong Lee
SW & Contents Research Laboratory
ETRI
Daejeon, Rep. of Korea
zacurr@etri.re.kr

Seungjae Lee
SW & Contents Research Laboratory
ETRI
Daejeon, Rep. of Korea
seungjlee@etri.re.kr

Wonyoung Yoo
SW & Contents Research Laboratory
ETRI
Daejeon, Rep. of Korea
zero2@etri.re.kr

***Abstract*—In this paper, we propose an effective image retrieval method. Based on a conventional global image representation, multi-stage image retrieval pipeline with feature augmentation is constructed to improve retrieval accuracy. To suppress irrelevant images while boosting relevant images, a novel weighting scheme for feature augmentation is introduced. In addition, the relationship between database images is leveraged to update or re-rank shortlist of the retrieved images. The proposed method was evaluated on Google-landmarks dataset, and the experimental results validate the effectiveness of the proposed method.**

Keywords—landmark retrieval, Google landmark challenge, Contents-based image retrieval

I. INTRODUCTION

Content-based image retrieval is a task of searching images of the same object, depicted in a query image, within a large collection of images. It has a wide range of applications, such as product recognition, image-based localization and management of personal photo collections.

Conventional methods for image retrieval are based on bag-of-visual-words (BoVW) [1,2] representation of hand-crafted local features, such as SIFT [3], with inverted-index structure. To further improve retrieval accuracy, initial results are re-ranked by local feature matching and geometric verification at the cost of expensive computation.

More recent works have introduced global image representations based on local feature aggregation, such as VLAD [4] or Fisher Vector (FV) [5].

After the success of Krizhevsky et al. [6] in image classification, global [7-12] or local [13,14] image representations for image retrieval, based on Convolutional Neural Networks (CNN) features have been introduced.

Regional maximum activations of convolutions (R-MAC) descriptor [8] is a global image representation based on CNN, which is one of the state-of-the-art methods. It uses a pre-trained, fully convolutional network to extract feature map, and regional features are max-pooled across a fixed layout of spatial regions. R-MAC is finally obtained by sum-aggregating regional features into a fixed-length vector.

Deep Image Retrieval (DIR) [10, 11] further improved R-MAC by fine-tuning CNN with ranking triplet loss for image retrieval task and adopting region proposal network to predict region of interest.

Image retrieval methods are usually evaluated on landmark datasets [1, 15, 16] for several reasons: (1) from huge amounts of landmark images on the web, it is easy to collect a large set of images for each landmark, (2) landmark datasets are challenging due to partial occlusion, background clutter and extreme changes in viewpoint and illumination.

However, several standard landmark datasets [1, 15, 16] have a limited variety of landmarks, and hence are small in scale. Oxfords5K [1] and Paris6K [15] has 5,062 and 6,392 landmark images, respectively and each dataset contains 55 query images of 11 landmarks.

To represent the diversity of real-world landmarks, Google-landmarks dataset was first introduced in [17] and recently made publicly available in [18] as part of the two Google Landmark Challenges: landmark recognition [19] and landmark retrieval [20].

The proposed method was developed based on DIR while taking part in the Google Landmark Retrieval challenge.

The rest of the paper is organized as follows: Section II explains Google-Landmarks dataset and challenges. In Section III, the proposed method is presented. Experimental results are given in Section IV, and we conclude in Section V.

II. GOOGLE-LANDMARKS DATASET AND CHALLENGES

Google-landmarks dataset contains 1.2M training images of 15K unique landmarks for the recognition challenge, and 1.1M index images of another 15K unique landmarks for the retrieval challenge, and 117K query images including 100K non-landmark images shared between challenges.

For the recognition challenge, participants were asked to build classification models to predict the correct label of the landmark depicted in each query image, while the goal of retrieval challenge is to retrieve all relevant index images that contain the same landmark depicted in each query image.

The proposed method is our submission to the Google Landmark Retrieval challenge, where we won the 8th place among 209 teams.

III. PROPOSED METHOD

This section introduces our method for large-scale image retrieval, particularly focused on Google-landmarks dataset.

The proposed method has several key features to further improve retrieval accuracy based on DIR: multi-stage retrieval with iterative query expansion and database-side feature augmentation, a novel weighted averaging scheme for feature augmentation, re-ranking or updating shortlist using similarity matrix of index database, and a hybrid approach using global CNN feature and hand-crafted local feature.

In the following subsections, baseline and key features of proposed method are presented.

A. Baseline

As shown in Fig. 1, our model is built upon DIR, query expansion [21-23], and database-side feature augmentation [23, 24]. The released source code and fine-tuned ResNet-101 model with ranking triplet loss and regular grid [25] were used

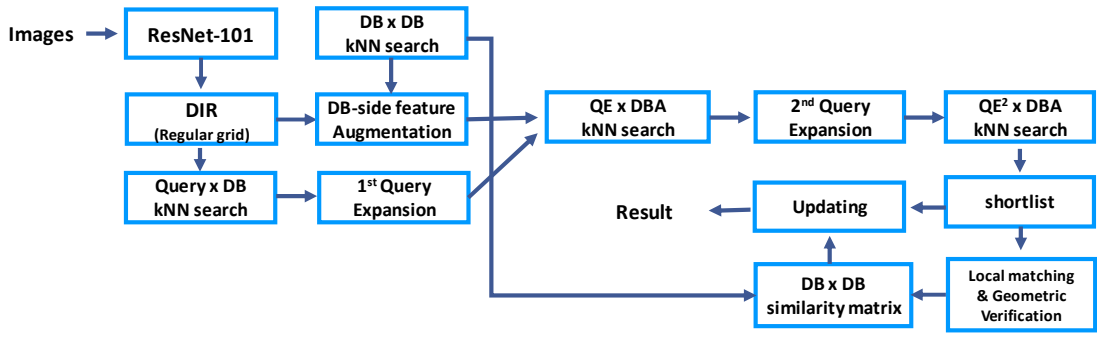

Fig. 1. Proposed image retrieval pipeline

to extract DIR descriptor and perform QE and DBA. In retrieval stage, exhaustive search is performed to retrieve top-k nearest neighbors.

B. Multi-stage retrieval with iterative QE and DBA

In large-scale dataset, due to a wide variety of imaging conditions, images depicting the same objects are scattered in the feature space. Considering this, query expansion (QE) [21-23] and database-side feature augmentation (DBA) [23, 24] were introduced.

In QE, a query feature is expanded by the features of the retrieved images. A standard scheme is average QE (AQE). In AQE, a new query feature is obtained by averaging the features of a query and top-k ranked images in initial search, and retrieval is performed once again with this feature. This can be understood that a query is moved close to a cluster of relevant images, and another adjacent cluster of relevant images is found in the feature space. Because of significant improvement in accuracy and its efficiency, QE has become a standard technique in image retrieval.

DBA has an analogy to QE, since each database image is queried against the database to retrieve relevant images, and its new representation is obtained by weighted averaging features of itself and its neighbors. Even though DBA involves heavy computation in the offline stage, it brings a significant boost in accuracy, and even greater improvement can be achieved, combined with QE [11].

In the proposed method, multi-stage retrieval pipeline employing DBA and iterative QE is built. In the offline stage, original features of database are augmented (denoted by DBA in Fig. 1), and similarity matrix of database, which will be explained later, is constructed. A query is first expanded by original features of database, and secondly by augmented features of database. The original and augmented features of database, first and second expanded query are denoted by DB, DBA, QE and QE^2, respectively in Fig. 1.

To augment features, AQE is used in the first QE, and a novel weighting scheme is applied in the second QE (QE^2) and DBA. This will be introduced in the following subsection.

Finally, by querying QE^2 against DBA, a shortlist of the retrieved images is obtained, which will be re-ranked by the methods described later in this section.

C. Truncated polynomial weight for feature augmentation

When QE and DBA are combined with geometric verification, only relevant images are augmented and a further improvement can be achieved. However, geometric verification requires local feature matching, which is a high computational burden in DBA and not feasible for the methods employing only global feature. Instead, weighting schemes for feature augmentation [11, 12] to effectively suppress irrelevant images were introduced.

In these methods, a new feature, F_n is obtained by weighted averaging an original feature F_o and top-k ranked images as in (1). A weight for an original feature (query for QE, and each queried database for DBA) is set to 1.

$$F_n = \left(F_o + \sum_{r=1}^{k} w_r F_r\right) \Big/ \left(1 + \sum_{r=1}^{k} w_r\right) \quad (1)$$

The feature and weight of the *r*-th ranked image is denoted by F_r and w_r, respectively.

In [11], a rank weight is used for w_r. The weight of the image is determined by its rank, r and the number of images considered in feature augmentation, k.

$$\mathrm{w_r} = \frac{\mathrm{k-r}}{\mathrm{k}} \quad (2)$$

This is effective under the condition that high-ranked images are relevant. Otherwise, however, a new obtained feature is significantly affected by irrelevant images, which results in performance degradation.

This problem can be effectively resolved by α-weighted query expansion (αQE), which was proposed in [12]. Considering, in general, irrelevant image has low similarity to a query, the weight is defined as follows.

$$\mathrm{w_r} = s^{\alpha} \quad (3)$$

The similarity between a query and a retrieve image is denoted by *s*. The exponent α is a pre-defined constant. For DIR, a cosine similarity is used, whose value ranges from 0 to 1 (identical). In our experiments on Google-landmarks dataset, the similarity of relevant images to a query is mostly over 0.6. Fig. 2 shows the various weight functions versus the cosine similarity. As α is increased, weights corresponding to the low similarity are decreased more. α was set to 3 in [12]

This scheme can suppress irrelevant images in feature agmentation. However, small weights are assigned to relevant images, which weakens the effect of augmentation.

We propose a novel truncated polynomial weight (TP weight) for feature augmentation, which can effectively suppress irrelevant images while highly leveraging relevant images. The proposed weight is computed by (4).

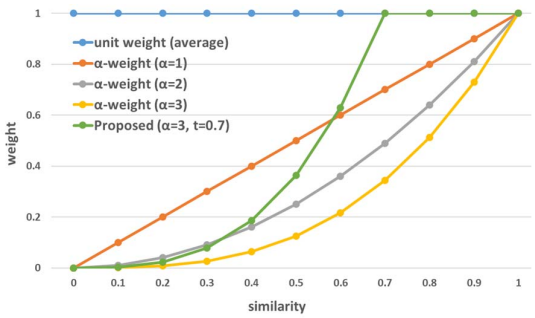

Fig. 2. Comparison of weights for feature augmentation

$$w_r = \begin{cases} (s/t)^{\alpha}, & s < t \\ 1, & s \geq t \end{cases} \quad (4)$$

If a similarity, s is smaller then threshold, t, similarity is divided by t, and raised to the α-th power. Otherwise, a weight is truncated to 1, not to give higher weight than an original feature. We set t=0.7 for QE and t=0.8 for DBA. In both cases, α is set to 3.

As shown in Fig. 2, the proposed method assigns larger weights for neighbors with high similarity compared to αQE.

In DBA and QE^2, the proposed weight is used.

D. Updating shortlist based on similarity matrix

The shortlist of top 500 ranked images is obtained by querying QE^2 against DBA. To further improve retrieval accuracy, this list is re-ranked or updated based on geometric verification and similarity matrix of index database.

The similarity matrix of index database is constructed during DBA, where each databasde image is queried against the database. It contains top 100 neighbors of each database image and their similarities.

For the computational efficiency, only top 30 images in the shortlist, rather than total 500 images are geometrically verified using local SIFT features and RANSAC [26].

If the relevance of a neighbor to a query is verified, neighbors of the neighbor in database are also likely to be relevant with a query. Considering this, for each verified image in the shortlist, its nearest neighbors, whose cosine similarity is over 0.6, are looked up in the similarity matrix. Using the similarity scores of these neighbors, the shortlist is re-ranked or updated. If they are already included in the shortlist, their similarity scores are updated by (5).

$$s_n = s_q + s_d \quad (5)$$

For each neighbor of the verified image, a new similarity score, s_n is obtained by adding the original similarity of the image to a query, s_q and the similarity of the image to a verified image s_d.

If the neighbor of the verified image is not in the shortlist, it is appended to the shortlist. Since s_q is not available, its score is set to s_d. This can be explained that, even though images are not close to a query in feature space, they are re-evaluated by the similarity to a verified sample.

Note that, the score can be updated repeatedly since the geometrically verified images are close to each other in the feature space, and their neighbors are largely overlapped. A common neighbor of the verified images will have much larger score in this scheme. In contrast, an isolated irrelevant image will not have a chance to update its score.

This can be understood that the relationship of neighbors is leveraged in this scheme while the features of neighbors are used for feature augmentation in a more direct manner.

A final retrieved list is obtained by sorting the updated shortlist, where images are appended and scores are updated.

IV. Experimental Results

The proposed method is evaluated on Google-landmarks dataset. The implementation details, evaluation measures, and experimental results are presented in the following.

A. Implementation details

The numbers of neighbors considered in feature augmentation for DBA, QE and QE^2 are 20, 1 and 5, respectively. Since an initial retrieval result is not reliable, only top 1 ranked image is augmented with a query in first QE.

For each image, 300 SIFT features are extracted to be used in local matching. Effective feature selection and extraction method [27] is ued to select informative features and reduce computational complexity.

An image is considered to be geometrically verified when the number of inlier, found by RANSAC, is more than 10.

B. Evaluation

We follow the standard evaluation protocol of Google landmark retrieval challenge. In this challenge, official evaluation measure is mean average precision (mAP) at top 100 ranked images.

The query set of 117K images including 100K non-landmark images are split into the public and private set.

The public set contains one third of query images for retrieval task, and the private set includes the rest of them. During the challenge, mAP of the submitted predictions on the public set is open to the participants. After the challenge is over, mAP scores on the private set are revealed and the final ranking is determined based on it. In experimental results, we presents mAP scores on both public and private set.

C. Experiments

To examine the various components of the proposed method, a set of experiments were conducted. The results are summarized in Table I.

In the baseline method, DIR is combined with AQE and DBA with rank weight (denoted by rDBA in Table I).

We evaluate the multi-stage retrieval structure with iterative query expansion. Performance was improved by appending the second AQE (denoted by AQE^2 in Table I. (c)) to the baseline.

Based on this result, the proposed TP weight for feature augmentation (denoted by tp-X in Table I) are compared with rank weight and αQE. Comparing (d), (e) and (f), the proposed weight outperforms the others. A further improvement was achieved by replacing the rank weight with TP weight for DBA in (g).

TABLE I. RETRIEVAL ACCURACY COMPARISON ON GOOGLE-LANDMARKS DATASET

Method		mAP @ 100	
		Public	*Private*
(a) DIR [10]		0.437	0.423
(b) Baseline: DIR + AQE + rDBA [11]		0.540	0.548
Iterative QE (Ours)	(c) DIR + AQE + rDBA + AQE^2	0.542	0.550
	(d) DIR + AQE + rDBA + rQE^2	0.545	0.552
	(e) DIR + AQE + rDBA + αQE^2	0.544	0.552
	(f) DIR + AQE + rDBA + $tpQE^2$	0.548	0.555
	(g) DIR + AQE + tpDBA + $tpQE^2$ (Ours: TP weight)	0.549	0.560
(h): (g) + updating shortlist based on similarity matrix (Ours)		0.551	0.559

Comparing (b) and (g), the combination of iterative QE and TP weight gives a substantial boost in performance.

The effect of shortlist updating based on similarity matrix and geometric verification was different across the two set (see (h) in Table I). It may result from various reasons such as inlier threshold in geometric verification, the number of extracted SIFT features, the spatial scale of landmark in image.

Our final submission to the challenge was (h) in Table I, which achieved the highest score among our methods on the public set, and we won the 8th place among 209 teams.

V. CONCLUSION

In this paper, an effective image retrieval method was presented. The proposed method has three key features. First, multi-stage image retrieval pipeline was built on the combination of iterative query expansion and database-side feature augmentation. Second, we proposed a truncated polynomial weight to augment the features of neighbors more effectively in QE and DBA. Third, the relationship of neighbors was leveraged based on geometric verification and similarity matrix of database to obtain the final result.

The experiments on Google-landmarks dataset show that a significant boost in retrieval accuracy was achieved by the proposed method. Even though, our method was built on DIR, any existing global image representation can be plugged into our framework.

ACKNOWLEDGMENT

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIP) (No.R0132-15-1005), Content visual browsing technology in the online and offline environments

REFERENCES

[1] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," In Proc. CVPR, 2007, pp. 1–8.

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

[3] D.G. Lowe, "Distinctive Image Features from Scale-Invariant Keypoints," International Journal of Computer Vision, vol. 60, no. 2, 2004, pp. 91–110.

[4] H. Jégou, M. Douze, C. Schmidt, and P. Perez, "Aggregating Local Descriptors into a Compact Image Representation," In Proc. CVPR, , 2010, pp. 3304–3311.

[5] H. Jégou, F. Perronnin, M. Douze, J. Sanchez, P. Perez, and C. Schmid, "Aggregating Local Image Descriptors into Compact Codes," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 34, no.9, 2012, pp. 1704–1716.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," In Proc. NIPS, 2012, pp. 1097–1105.

[7] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky., "Neural Codes for Image Retrieval," In Proc. ECCV, 2014, pp. 584–599.

[8] G. Tolias, R. Sicre, and H. Jégou, "Particular Object Retrieval with Integral Max-Pooling of CNN Activations," In Proc. ICLR, 2015.

[9] F. Radenović, G. Tolias, and O. Chum, "CNN Image Retrieval Learns from BoW: Unsupervised Fine-Tuning with Hard Examples," In Proc. ECCV, 2016, pp.3–20.

[10] A. Gordo, J. Almazan, J. Revaud, and D. Larlus, "Deep Image Retrieval: Learning Global Representations for Image Search," In Proc. ECCV, 2016, pp.241–257.

[11] A. Gordo, J. Almazan, J. Revaud, and D. Larlus , "End-to-end learning of deep visual representations for image retrieval," International Journal of Computer Vision, vol. 124, no. 2, 2017, pp. 237–254.

[12] F. Radenović, G. Tolias, and O. Chum, "Fine-tuning CNN image retrieval with no human annotation," IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018. [a-QE, GeM]

[13] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua, "LIFT: Learned Invariant Feature Transform," In Proc. ECCV, 2016, pp. 467–483.

[14] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-Scale Image Retrieval with Attentive Deep Local Features," In Proc. ICCV, 2017, pp. 3456–3465.

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in Quantization: Improving Particular Object Retrieval in Large Scale Image Databases," In Proc. CVPR, 2008, pp. 1–8.

[16] H. Jégou, M. Douze, and C. Schmid, "Hamming Embedding and Weak Geometric Consistency for Large Scale Image Search," In Proc. ECCV, 2008, pp. 304–317.

[17] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han, "Large-Scale Image Retrieval with Attentive Deep Local Features," In Proc. ICCV, 2017, pp. 3456–3465.

[18] Google-Landmarks dataset, https://www.kaggle.com/google/google-landmarks-dataset

[19] Google-Landmark Recognition Challenge, https://www.kaggle.com/c/landmark-recognition-challenge

[20] Google-Landmark Retrieval Challenge, https://www.kaggle.com/c/landmark-retrieval-challenge

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

[22] O. Chum, A. Mikulik, M. Perdoch M, and J. Matas, "Total recall II: Query expansion revisited," In Proc. CVPR, 2011, pp. 889–896.

[23] R. Arandjelovic and A. Zisserman, "Three things everyone should know to improve object retrieval," In Proc. CVPR, 2012, pp.2911–2918.

[24] P. Turcot and D.G. Lowe, "Better matching with fewer features: The selection of useful features in large database recognition problems," In Proc. ICCV Workshop, 2009, pp.2109–2116.

[25] DIR source code and model, http://www.europe.naverlabs.com/Research/Computer-Vision/Learning-Visual-Representations/Deep-Image-Retrieval

[26] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," Communications of the ACM, vol. 23, no.6, 1981, pp. 381–395.

[27] K. Lee, S. Lee, and W.G. Oh, "Accelerating Local Feature Extraction Using Two Stage Feature Selection and Partial Gradient Computation," In Proc. ACCV Workshop, 2014, pp. 366–380