Mini-Batch VLAD for Visual Place Retrieval

Reem Aljuaidi, Jing Su, Rozenn Dahyot
School of Computer Science and Statistics
Trinity College Dublin
Dublin, Ireland
aljuaidr@tcd.ie, Jing.Su@tcd.ie, Rozenn.Dahyot@tcd.ie

***Abstract*—This study investigates the visual place retrieval of an image query using a geotagged image dataset. Vector of Locally Aggregated Descriptors (VLAD) is one of the local features that can be used for image place recognition. VLAD describes an image by the difference of its local feature descriptors from an already computed codebook. Generally, a visual codebook is generated from k-means clustering of the descriptors. However, the dimensionality of visual features is not trivial and the computational load of sample distances in a large image dataset is challenging. In order to design an accurate image retrieval method with affordable computation expenses, we propose to use the mini-batch k-means clustering to compute VLAD descriptor(MB-VLAD). The proposed MB-VLAD technique shows advantage in retrieval accuracy in comparison with the state of the art techniques.**

***Index Terms*—feature extraction; content-based image retrieval; image processing.**

I. Introduction

The availability of images with their geolocations coupled with additional information (e.g. text, time stamp, etc.), has led to many applications such as object geo-localization [1] and flood monitoring [2]. However, many images are lack of accurate (or any) GPS information - for instance tweets often include GPS of where people tweet but pictures posted in a tweet may have no GPS tag. In order to recover lost GPS information, Bulbul et al. proposed to query Google street view image database [3]. In general, the pipeline of recognizing a certain place using a single visual query has three successive steps. This pipeline can be applied to a large dataset such as a city scale. First, regions are located in the query image. Second, descriptors are generated over these selected regions in order to provide an accurate representation of the query image. Finally, this representation is matched over geotagged images in the reference dataset and the GPS information of the retrieved reference is then retrieved to the query image.

Local scale invariant feature transform (SIFT) [4] has been used as powerful features for describing informative regions of images. SIFT is also robust to photometric and geometric changes [5], [6]. Therefore, SIFT has an important role for image retrieval. The Vector of Locally Aggregated Descriptors (VLAD) [7] have been shown to be powerful local features for image geo-localization and retrieval. VLAD represents an image by a single fixed-size vector using K-means clustering. The issue however with VLAD is the dimensionality of visual features and the computational load of sample distances in a large image dataset. We propose instead to learn VLAD by using mini batch k-means clustering [8] (MB-VLAD). This paper is organized in the following way: Section II introduces related works. MB-VLAD is presented in Section III, and it is assessed experimentally over state of the art techniques in Section IV-D.

II. Related work

Image geo-localization can be defined as predicting the GPS coordinates of a query image using a geotagged image dataset [9]. Recently, the presence of large scale geo-tagged image collections enables image retrieval approaches of transferring geotagged data from a reference dataset to the query image. Instances of these applications include adding and refining geotags in image collections [10], [11], navigations [12], photo editor [13] and 3D reconstruction [14].

With the task of visual place retrieval we aim to find an estimated location of a query image by selecting the geo-tagged images capturing the same visual scene in a reference dataset [10]. This challenge can be approached as an image retrieval task from a large scale image repository.

Content-based image retrieval (CBIR) differs from text-based image retrieval (TBIR) in that CBIR uses visual information as evidence of matching between a query image and an image repository [15]. In this paper, we study the problem of using a robust and effective local feature descriptor against geometric changes for visual place retrieval.

Typically, many of the local features can be extracted from an input image. However, a main challenge is the dimensionality of visual features. Computation load of feature vector matching is overwhelming with raw image feature vectors. To get a compact representation, high dimensional local features are first translated into visual words using a pre-trained visual codebook. Based on the results of this quantization, local features of an image can be projected to a fixed-length vector [16], based on Bag-of-Visual-Words

978-1-7281-2800-9/19/$31.00 (c) 2019 IEEE

model [17], VLAD [7], [18], [19], or Fisher Vector [18], [20]. Visual codebook is generated by clustering visual features. For example, the cluster centers of K-means clustering can be used as visual words.

Many methods have been proposed to solve the problem of local feature descriptors in addition to VLAD. Delhumeau et al. [21] proposed a method to solve the problem of the accumulative residuals of VLAD. This method helped to improve retrieval performance by contributing all descriptors equally to the aggregated residual vector. Arandjelovic et al. [18] addressed the problem of vocabulary sensitivity. In some cases, the cluster centers used for VLAD are not consistent with the dataset when new images are added to the dataset after initial vocabulary learning. To solve the sensitivity to the choice of vocabulary, vocabulary adaptation method was proposed. This method can update the centroids when images from a different dataset are processed without requiring re-clustering on the current dataset. In this work, cluster centroids are updated first to keep consistency and then all VLAD vectors are re-computed based on new centroids. Results of this method show significant improvement over the original VLAD approach. On the other hand, VLAD re-computation has high computation cost when the collectionâs size increases. In addition, Eggert et al. [19] proposed hierarchical VLAD (HVLAD) method, which is the representation of applying cluster-wise PCA on aggregated residual vectors before concatenation.

Among the works of improving local feature descriptors for image geo-localization, Kim et al. [22] propose PBVLAD as a novel method to locally integrate SIFT features detected with a MSER blob. The descriptor is called Per-bundle vector aggregated locally vector (PB-VLAD). The purpose behind this descriptor is to find a robust local feature descriptor against geometric and photometric changes. The idea is describing each maximally stable region (MSER) by a vector of locally aggregated descriptors (VLAD) on multiple features detected in the region. As an accumulation of variance between the descriptors that are allocated to the visual word and the centroid, a sub-vector of the per-bundle VLAD was proposed. In fact, this processing takes place on each image patch, which is time consuming when computing VLAD descriptors for large datasets.

The basic clustering algorithm for VLAD is k-means clustering algorithm. The idea behind this algorithm is to select k (the final clusters number) beforehand. The initial k centroids are selected randomly. Distance is calculated between each sample and the centroids. Sample points are assigned to the most similar class (closest centroid), and then the centroid is recalculated for each class. After iterations, centroids keeps stable and each sample is allocated to the closest cluster centroid [23].

The mini batch k-means algorithm [24] is an alternative algorithm of k-means, which divides the data logically into multiple small batch data subsets. The algorithm here randomly extracts subsets of data at each training iteration [23]. In another word, the strategy of the mini batch k-means is using per-centre learning rates and a stochastic gradient descent algorithm. It takes one mini batch as an input, which are random subsets of the whole dataset. Then, the samples in this mini batch are associated with the nearest centroid: for every single sample in the mini batch, the allocated centroid is updated by taking the current mean of the sample and all former samples allocated to that centroid. Algorithm 1 explains mini-batch optimization for k-means clustering [25].

In our work, we try to improve local feature descriptors for visual place retrieval by using alternative clustering algorithm. We adopt mini batch k-means algorithm that was proposed by Sculley [8], to compute VLAD [7]. Mini batch k-means algorithm works better on large scale datasets than the original k-means algorithm by reducing the computation cost and time for data [23].

Feizolla et al [26] evaluate the performance of k-means and mini batch k-means clustering for malware detection purposes. Results of this comparative study of k-means and mini batch k-means clustering show that the performance with the mini batch is actually better [26] .

Algorithm 1 Mini batch k-means clustering

Given k, mini-batch size m,iterations t, dataset X
Initialize each $c \in C$ with an x picked randomly from X $v \leftarrow 0$
for $i = 1 \rightarrow t$ **do**
  $S \leftarrow b$
  for $x \leftarrow S$ **do**
    $d[x] \rightarrow f(C, x)$ //Cache the center nearest to x
  end
  for $x \in M$ **do**
    $d[x] \leftarrow c$ // Get center for this x temporary
    $v[c] \leftarrow v[c] + 1$ // Update per-center counts
    $\eta \leftarrow \frac{1}{v[c]}$ // Get per-center learning rate
    $c \leftarrow (1 - \eta)c + \eta x$ // Take gradient step
  end
end

III. Methodology

In this section we introduce the methodologies to apply mini batch k-means clustering in feature based visual place recognition. The pipeline of feature-based visual place recognition entails three stages: First, selecting a single query image with unknown location as an input. Second, extracting features using SIFT keypoints. The final stage is to compute a vector representation using the k centroids to match the database. Figure 1 depicts the proposed pipeline used in this paper. The following sections discuss each pipeline stage in detail.

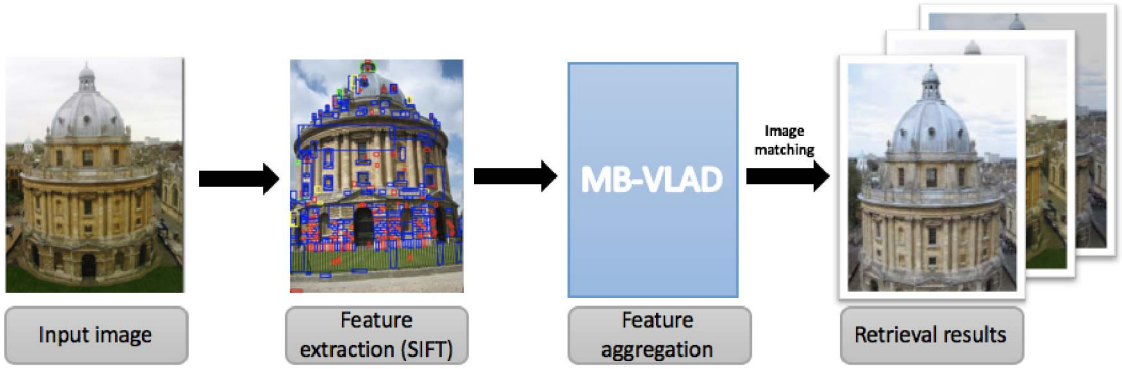

Fig. 1: Our proposed pipeline to process a input (query) image.

A. Mini-batch VLAD

Our goal is to retrieve image using parts of an input image for geo-localization. One challenge of feature learning is the scale of dataset. In case of learning from a city dataset, the computational load is very high. Here we propose Mini Batch Vector of Locally Aggregated Descriptors (MB-VLAD). The key idea is to aggregate features using a vector with a fixed-size, and to learn the vocabulary word using mini batch k-means clustering algorithm. In this way we reduce computational load and it is still convenient to use standard distance measures to retrieve relevant images.

The concept of a mini batch k-means algorithm was proposed by Sculley [8], in two iterated steps. The first step is to form a mini-batch by taking samples randomly from the dataset, and then each sample in the mini batch is assigned to its nearest centroid. In the second step, the centroids are updated as the mean over its associated samples in the mini batch. These two steps are then repeated for several iterations. Please note that, the quality of clustering is affected by the chosen number of iterations: the more iterations used in the mini batch k-means algorithm, the better the clustering result quality is [27].

The original VLAD approach [7] builds a codebook dictionary $C = \{c_1, c_2, \ldots, c_k\}$ from $m \geqslant k$ feature vectors in the reference dataset. To generate the dictionary, a k-means clustering algorithm is used. For an image having m descriptors $I = \{x_1, \cdots, x_m\}$, the VLAD coefficient V_i is computed by accumulation over these descriptors in cluster c_i:

$$V_i = \sum_{x \in I / q(x) = c_i} x - c_i \qquad (1)$$

where $q(x)$ is the cluster associated with x.

The final VLAD representation is a concatenation $v = \{v_1, \ldots, v_i, \ldots, v_k\}$ followed by L_2 normalization $v : v/\|v\| \rightarrow \tilde{v}$. Thereafter, VLAD encodes feature by computing the residuals [18], and the residuals are stacked together as vector v. In this study, we propose to replace the set of centroids inferred by k-means c_i algorithm [8] with cm_i the cluster centers from mini batch k-means. When generating the dictionary, a mini batch k-means algorithm is applied. Clustering input data are unnormalized SIFT descriptors before adding them into mini batches. When the dictionary is generated, two normalizations are applied to compute the final VLAD. First, the power law normalization V_u is applied: for $u = \{1, \ldots, Kd\}$, $V_u = sign(V_u)|V_u|^{\alpha}$ [19]. Then L_2 normalization is used.

The similarity grouping is performed by distance measurement. We use Euclidean distance as sample similarity metric. When searching for the closest VLAD vector, the one with the lowest Euclidean distance is selected.

The dimension of MBVLAD can be reduced when searching the nearest neighbor by using principal component analysis (PCA). We fit PCA in offline mode, and then transform all MBVLAD feature vectors in the dataset. PCA is calculated on subvectors v_i that are generated from each visual word c_i. We generate a coarse vocabulary of 128 visual words(16,384-dimensional MBVLADs raw). Thereafter, some majors components are used to reduce dimensional VLAD vectors size using 128 visual words.

IV. Experimental Results

Performance is evaluated using Mean average precision (mAP), which is the mean of the average precision scores for each query. We evaluate proposed descriptor MB-VLAD on Oxford building datasets (Sec. IV-A) and obtain good mAPs compared to state of the art for uncompressed descriptors (Sec. IV-B), compressed descriptors (Sec. IV-C). The robustness of our approach to the choice of centroids is evaluated in paragraph IV-D.

A. Dataset

For standalone MB-VLAD descriptor evaluation, the Oxford building dataset [17] was used. It is usually called as Oxford 5k. It consists of several image subsets, assembled together by an image quality measure (mainly the percentage of represented object visibility) and with each image set partitions labelled as $\{good, ok, ugly, bad\}$. Additional images placed in *query* subset served as query test set. However, we used entire dataset to compute the visual dictionary to get best possible distribution for the calculation of the cluster centroids while clustering.

B. On Uncompressed VLAD

In our experiment, we set the maximum number of iterations over the complete dataset to 100, and the size of the mini batches m to 500000.

Table I shows the image retrieval performance of uncompressed VLAD [19] and PB-VLAD [22], HVLAD [19] and MB-VLAD (our). The VLAD vector size is typically kxd-dimensional. We use d= 128 in all experiments. The adapted method has a significant improvement comparing with other methods in term of improving local features descriptors. Comparing with PB-VLAD our method increases the retrieval performance for image geo-localization by 11%. Figure 3 shows examples of our successful retrieval images with high accuracy (true-retrievals / total-images).

TABLE I: Comparison of the mean Average Precision (mAP) performance of several (uncompressed) VLAD signatures evaluated on the Oxford dataset.

Descriptor	# Vocabulary	mAP
VLAD [19]	128	0.33
PB-VLAD [22]	128	0.36
HVLAD [19]	128	0.40
MB-VLAD(Our)	128	0.47

Figure 2 is showing the mAP in a graphical form. The greater the area under the curve, the higher is the reported mAP metric. The curve is generated by applying step-by-step thresholding of the prediction scores. The overall mAP score is 0.47 when we calculate Precision-Recall curve. The mini batches algorithm is more robust to the noise introduced by random selection of initial centroids, and retrieval performance is not affected.

C. On PCA compressed VLAD

Table II shows retrieval performance of MB-VLAD on the Oxford5k dataset, before and after the dimensionality reduction using PCA (k= 128, 64, 32, 16, 8). MB-VLAD achieves 0.47, which outperforms other feature selection approaches in literature.

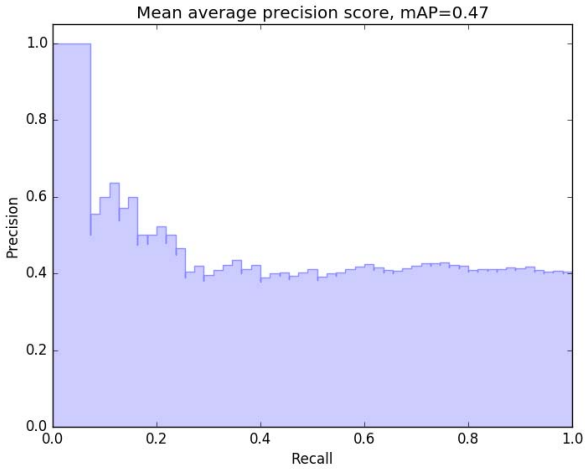

Fig. 2: Precision-Recall curve for uncompressed VLAD, with mAP of 0.47.

TABLE II: Retrieval performance of (our) on the Oxford 5k dataset [17], before and after the dimensionality reduction using PCA (128 vocabulary size). The performance is measured by the mean Average Precision (mAP)

	Full	PCA			
	16384	8192	4096	2048	1024
PB-VLAD [22]	0.36	0.36	0.33	0.26	0.21
MB-VLAD (our)	0.47	0.44	0.40	0.43	0.39

D. Sensitivity to initial centroids

When the visual dictionary is generated again, the initial centroids from mini batch k-means algorithm can be different every time. Table III shows the mAP for five runs, leading to an average mAP of 0.444 with standard deviation 0.0152. Significant improvement is observed in comparison with the state of the art (see Tab. I for comparison).

TABLE III: Retrieval performance of (our) after generating the visual dictionary for five times

Run Times	# Vocabulary	D	mAP
1^{st}	128	16384	0.47
2^{nd}	128	16384	0.44
3^{rd}	128	16384	0.43
4^{th}	128	16384	0.44
5^{th}	128	16384	0.44

V. Conclusion

In this paper we address the problem of finding visual places over city area using a query image. We propose mini batch VLAD descriptor with the goal of improving

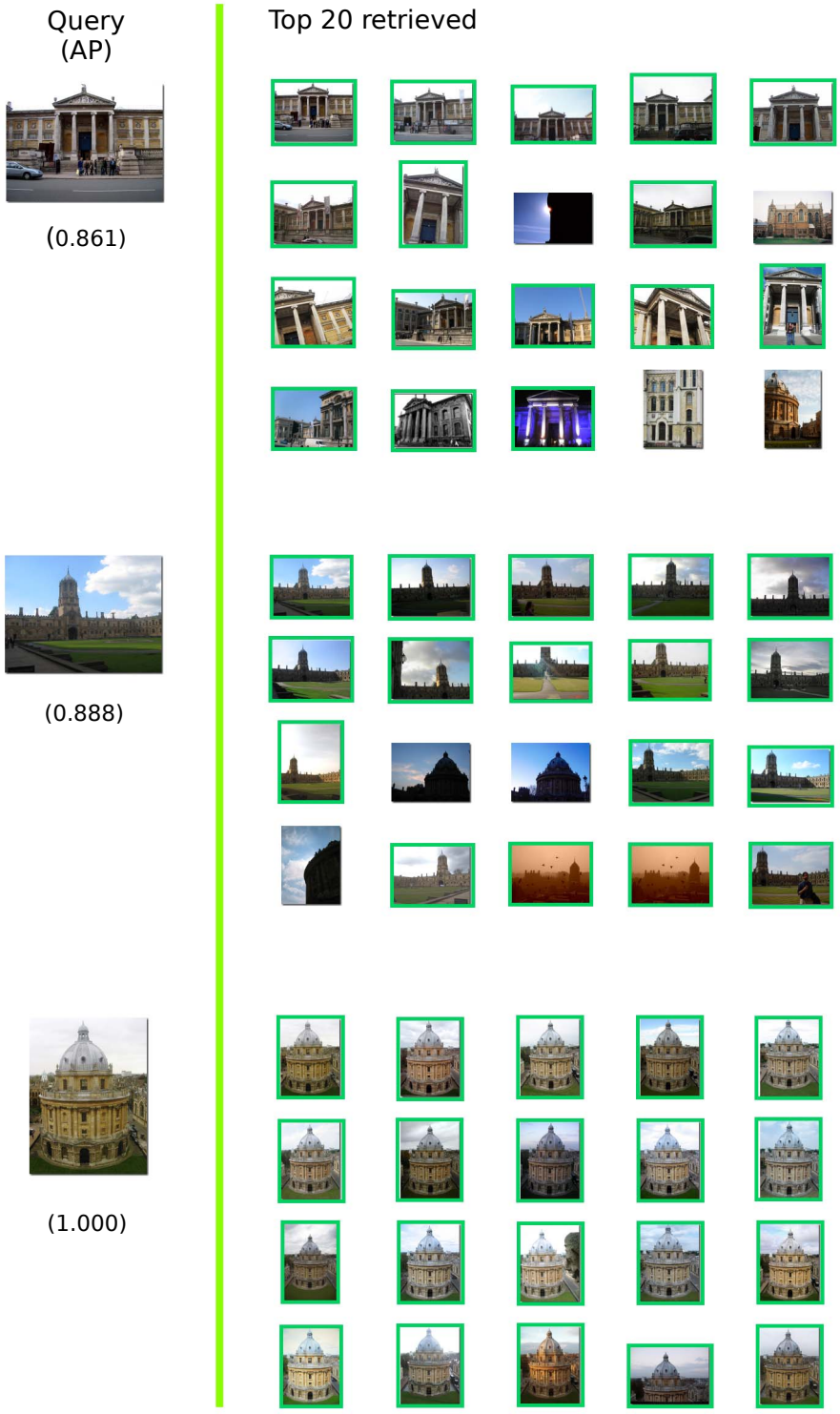

Fig. 3: Example results (uncompressed MB-VLAD): Query images (left) with different sizes, (right) Top 20 retrieved images using our proposed MB-VLAD

the performance of a visual place recognition system, under challenges of geometric changes. Comparing with the original k-means clustering approach, MB-VLAD has significant accuracy improvement on image retrieval. A key challenge of k-means algorithm is that clustering output is not deterministic and it is influenced by choices of initial centroids. From experiments we find that the mini batch version is more robust to the randomness of initial cluster centroids as well as significantly reduce computational load.

ACKNOWLEDGMENTS

This work is partly funded by Prince Sattam bin Abdalaziz University Scholarship Program from Saudi Arabian Government, and the ADAPT Centre for Digital Content Technology www.adaptcentre.ie (funded under the SFI Research Centres Programme Grant 13/RC/2106 and co-funded under the European Regional Development Fund).

REFERENCES

[1] V. A. Krylov, E. Kenny, and R. Dahyot, "Automatic discovery and geotagging of objects from street view imagery," *Remote Sensing*, vol. 10, no. 5, 2018. [Online]. Available: http://www.mdpi.com/2072-4292/10/5/661

[2] K. Ahmad, K. Pogorelov, M. Riegler, O. Ostroukhova, P. Halvorsen, N. Conci, and R. Dahyot, "Automatic detection of passable roads after floods in remote sensed and social media data," *Signal Processing: Image Communication*, vol. 74, pp. 110 – 118, 2019. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0923596518311536

[3] A. Bulbul and R. Dahyot, "Social media based 3d visual popularity," *Computers Graphics*, vol. 63, pp. 28 – 36, 2017. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0097849317300146

[4] D. G. Lowe, "Object recognition from local scale-invariant features," in *Proceedings of the International Conference on Computer Vision-Volume 2 - Volume 2*, ser. ICCV '99. Washington, DC, USA: IEEE Computer Society, 1999, pp. 1150–. [Online]. Available: http://dl.acm.org/citation.cfm?id=850924.851523

[5] Y. Li, N. Snavely, and D. P. Huttenlocher, "Location recognition using prioritized feature matching," in *Proceedings of the 11th European Conference on Computer Vision: Part II*, ser. ECCV'10. Berlin, Heidelberg: Springer-Verlag, 2010, pp. 791–804. [Online]. Available: http://dl.acm.org/citation.cfm?id=1888028.1888088

[6] A. R. Zamir and M. Shah, "Accurate image localization based on google maps street view," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

[7] H. JÃ©gou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3304–3311.

[8] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th International Conference on World Wide Web*, ser. WWW '10. New York, NY, USA: ACM, 2010, pp. 1177–1178. [Online]. Available: http://doi.acm.org/10.1145/1772690.1772862

[9] R. Arandjelović and A. Zisserman, "DisLocation: Scalable descriptor distinctiveness for location recognition," in *Asian Conference on Computer Vision*, 2014.

[10] J. Hays and A. A. Efros, "im2gps: estimating geographic information from a single image," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[11] A. R. Zamir, S. Ardeshir, and M. Shah, "Gps-tag refinement using random walks with an adaptive damping factor," in *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition*, ser. CVPR '14. Washington, DC, USA: IEEE Computer Society, 2014, pp. 4280–4287. [Online]. Available: https://doi.org/10.1109/CVPR.2014.545

[12] H. Lim, S. N. Sinha, M. F. Cohen, and M. Uyttendaele, "Real-time image-based 6-dof localization in large-scale environments," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, June 2012, pp. 1043–1050.

[13] C. Zhang, J. Gao, O. Wang, P. Georgel, R. Yang, J. Davis, J. Frahm, and M. Pollefeys, "Personal photograph enhancement using internet photo collections," *IEEE Transactions on Visualization Computer Graphics*, vol. 20, no. 2, pp. 262–275, Feb. 2014. [Online]. Available: doi.ieeecomputersociety.org/10.1109/TVCG.2013.77

[14] J.-M. Frahm, P. Fite-Georgel, D. Gallup, T. Johnson, R. Raguram, C. Wu, Y.-H. Jen, E. Dunn, B. Clipp, S. Lazebnik, and M. Pollefeys, "Building rome on a cloudless day," in *Computer Vision – ECCV 2010*, K. Daniilidis, P. Maragos, and N. Paragios, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 368–381.

[15] R. Datta, D. Joshi, J. Li, and J. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Computing Surveys*, vol. 40, no. 2, 4 2008.

[16] W. Zhou, H. Li, and Q. Tian, "Recent advance in content-based image retrieval: A literature survey," *CoRR*, vol. abs/1706.06064, 2017. [Online]. Available: http://arxiv.org/abs/1706.06064

[17] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

[18] R. Arandjelovic and A. Zisserman, "All about vlad," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 1578–1585.

[19] C. Eggert, S. Romberg, and R. Lienhart, "Improving VLAD: hierarchical coding and a refined local coordinate system," in *2014 IEEE International Conference on Image Processing, ICIP 2014, Paris, France, October 27-30, 2014*, 2014, pp. 3018–3022. [Online]. Available: https://doi.org/10.1109/ICIP.2014.7025610

[20] F. Perronnin, Y. Liu, J. SÃ¡nchez, and H. Poirier, "Large-scale image retrieval with compressed fisher vectors," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 3384–3391.

[21] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, "Revisiting the vlad image representation," in *Proceedings of the 21st ACM International Conference on Multimedia*, ser. MM '13. New York, NY, USA: ACM, 2013, pp. 653–656. [Online]. Available: http://doi.acm.org/10.1145/2502081.2502171

[22] H. J. Kim, E. Dunn, and J.-M. Frahm, "Predicting good features for image geo-localization using per-bundle vlad," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 1170–1178. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.139

[23] J. Newling and F. Fleuret, "Nested mini-batch k-means," in *Advances in Neural Information Processing Systems 29*, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Curran Associates, Inc., 2016, pp. 1352–1360. [Online]. Available: http://papers.nips.cc/paper/6481-nested-mini-batch-k-means.pdf

[24] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *IEEE International Conference on Computer Vision*, vol. 2, 2003, pp. 1470–1477.

[25] J. Bejar, "K-means vs mini batch k-means: a comparison," Tech. Rep., 2013.

[26] A. Feizollah, N. Anuar, R. Salleh, and F. Amalina, "Comparative study of k-means and mini batch k- means clustering algorithms in android malware detection using network traffic analysis," in *International Symposium on Biometrics and Security Technologies (ISBAST)*, 08 2014.

[27] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 881–892, Jul. 2002. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2002.1017616