Enlarging the Discriminability of Bag-of-Words Representations with Deep Convolutional Features

Daniel Manger and Dieter Willersinn
Fraunhofer IOSB, Karlsruhe, Germany
e-mail: firstname.lastname@iosb.fraunhofer.de

***Abstract*— In this work, we propose an extension of established image retrieval models which are based on the bag-of-words representation, i.e. on models which quantize local features such as SIFT to leverage an inverted file indexing scheme for speedup. Since the quantization of local features impairs their discriminability, the ability to retrieve those database images which show the same object or scene to a given query image is decreasing with the growing number of images in the database. We address this issue by extending a quantized local feature with information from its local spatial neighborhood incorporating a representation based on pooling features from deep convolutional neural network layer outputs. Using four public datasets, we evaluate both the discriminability of the representation and its overall performance in a large-scale image retrieval setup.**

***Keywords*— Content-based Image Retrieval, Bag-of-Words, Spatial Context of local Features, CNN features, 2D index**

I. Introduction

During the last decade, content-based image retrieval (CBIR) has not only been a lively field of computer vision research, it also became apparent in many successful applications including countless apps, which recognize items based on a snapshot taken with a mobile device. CBIR aims at preparing a large database of images such that all database images can be efficiently searched for candidates showing similar scenes or objects to a given arbitrary query image. The seminal paper of Sivic et al. [1] introduced the quantization of local features such as SIFT [2] for image retrieval proposing the bag-of-words-model (BoW) which allows to apply text retrieval methods to images. To that end, the local features are quantized with a so-called visual codebook, i.e. by assigning every feature to one element of the codebook which in turn is a large, but limited set of feature representatives termed visual words. Thus, the task of matching image content can be translated to analyzing the co-occurring visual words of images.

Although enabling computational speed-up and memory benefits, the quantization of features in the BoW model impairs the discriminative power of the underlying local features. This limits the retrieval accuracy in large-scale datasets beyond a few million images. Research therefore turned towards global image representations, where local features are aggregated with strategies such as VLAD [3] or Fisher Vectors [4] into a high-dimensional embedding followed by a compression step – mostly PCA and whitening - to encode images into compact codes. Searching the image database for similar images is hence typically performed by calculating the Euclidean distances between the compact code of the query image and the codes from all database images. With typical code sizes of 64 to 512 dimensions, this exhaustive search is still fast for databases of moderate sizes. Finally, the recent advances of convolutional neural networks (CNNs) have attracted attention to image retrieval researchers. Using deep-learned CNN features out-of-the-box [5] or by pooling responses from fully connected [6], [7] or convolutional layers [8–11], competitive and better results than the aggregated local feature approaches have been obtained for low dimensional image codes.

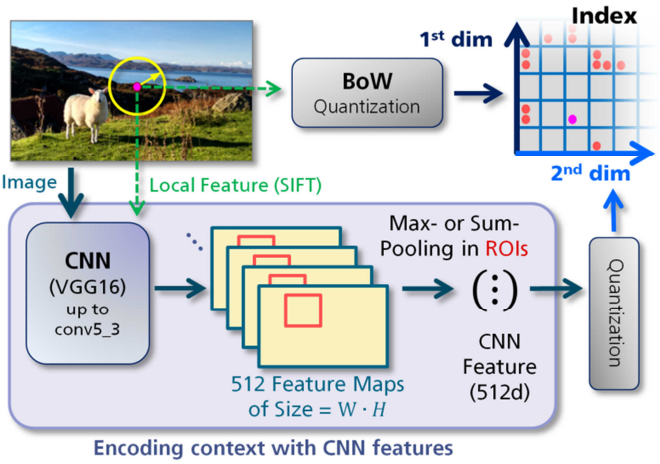

Figure 1. Summary of the proposed system for extending the information content of a local feature (magenta point) using features from pooling CNN layer activations. In the end, the BoW quantization of a feature is enriched with a quantized value encoding its neighboring context and inducing a second dimension in the index.

However, if, for instance, all of the image content is encoded (via CNN features or aggregation of local features) into just 64 floats, it becomes obvious that retrieving very small objects surrounded by plenty of heavily cluttered background becomes difficult. In other words, to date, neither CNN features nor aggregation of local features into global image codes were able to preserve the locality of the initial BoW-model at high retrieval performance to large-scale datasets.

In this work, we therefore analyze the combination of the BoW approach with CNN features by extending each quantized feature in the BoW-model with more context information from the respective local neighborhood. Our contributions are threefold: (1) we propose to integrate quantized CNN context information thus adding new dimensions to the index of a retrieval system, (2) we propose an evaluation framework for analyzing the discriminabiliy of the context information and (3) we give an overall large-scale evaluation which confirms that in fact more corresponding images can be found when adding the features' contexts.

The paper is structured as follows: Section 2 presents related work for integrating more context information into the BoW model, Section 3 describes our approach of encoding and subsequently quantizing context information. In Section 4, we focus on a proper framework for evaluation, followed by Section 5 for details on experiments with public datasets. The paper concludes with Section 6.

978-1-5386-1842-4/17/$31.00 ©2017 IEEE

II. Related work

Many advancements of the bag-of-words model have been proposed which aim at different aspects of the image retrieval pipeline in order to incorporate more information into the retrieval process. We neglect methods for short-list re-ranking via spatial verification [12], [13] or query expansion [14–17] since they rely on proper initial retrieval results or assume multiple corresponding database images and are too expensive to be applied to all images in the database. In contrast to that, the approaches which incorporate additional information into the inverted file indexing scheme (termed index), are applied to all images and can be separated into three strategies:

Extending the accumulator: These methods extend the accumulator, which - for every query - holds bins for the scores of all the database images by new dimensions assuming that irrelevant features will spread along multiple bins of one database image while corresponding features will accumulate in one or few of the bins of a similar image. For instance, Jegou et al. [18] use orientation and scale information of SIFT features to push database images with features having consistent differences in scale and orientation compared to the query image. Zhang et al. [19] process the position of the features introducing a larger accumulator called "offset space" in X and Y image dimension. Shen et al. [20] extends that accumulator yielding invariance to translation, rotation and scale differences resulting in 16,000 bins for each target image (16×X, 16×Y, 8×Scale, 8×Orientation). Besides the benefit that retrieved objects are localized in the target images, these extensions of the accumulator add additional storage requirements to the main memory hosting the index. Furthermore, initialization and evaluation of large accumulators takes a lot of time for large-scale datasets.

Filtering of features: Another approach is to keep the accumulator compact (one bin per database image) and to use additional information in the index to filter matches prior to casting votes into the accumulator. Zhang et al. [21] determine the four closest features in the image coordinate space and capture their appearance and relative geometry. During retrieval, each BoW-match is further examined as to how many of the four neighboring features are consistent. While keeping the accumulator space compact, this filtering of features during retrieval requires significantly more computational resources since still all entries of the feature's visual word in the index have to be processed. Furthermore, in realistic image data with challenging transformations, a larger neighborhood than just four features is relevant and within this larger neighborhood, only a fraction of features match.

2D-Index: In order to overcome the runtime, performance and storage limits of both accumulator extension and filtering of features, Zheng et al. [22] uses a multi-index. The first dimension of the index is still dedicated to the BoW vectors while the second dimension is based on the color name descriptor [23], which is an 11-dimensional descriptor mapping color values to 11 categories. Using a Color-Codebook of size 200, every feature in the index is assigned up to the 100 closest Color-Words which however obviously eliminates the advantages of the second dimension because still up to 50% of the index has to be traversed.

In this paper, we follow the latter strategy of integrating context information as a second dimension into the inverse file. However, with CNN features, we use different features as basis for the second dimension. Adding such a new dimension to the index is attractive in multiple aspects: In contrast to the filtering strategy, the runtime during retrieval can be optimized because only features which match both dimensions have to be considered for the accumulator. Depending on the parameters, this can allow storing the inverse file on a disk rather than in main memory which is a current limit in terms of scalability. Furthermore, since fewer entries will end up in the accumulator, sparse implementations become feasible which can be evaluated faster. Finally, retrieval accuracy can benefit from the second dimension because many incorrect matches of features are discarded that match w.r.t. the first dimension only (the quantized local feature descriptor) but not in the second dimension (the larger context of the feature). See Figure 2 for an illustration of the index when adding new dimensions based on context of features. Please note that, for the `OR` combination, many cells have to be processed during retrieval (see the yellow cells in Figure 2). However, still much less than in the previously mentioned *Filtering of features* approach of [21] and – most notably - without any filtering operation (only memory accesses).

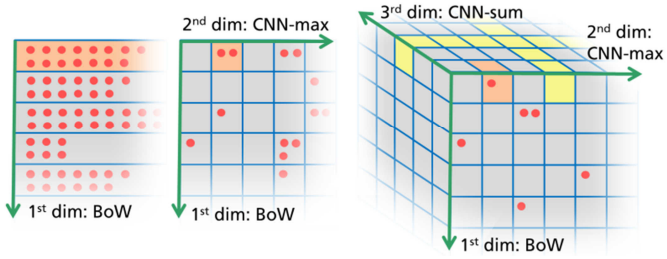

Figure 2. Visualization of an inverted file-based index for image retrieval which is extended with additional dimensions arising from the proposed quantized context features. Red points depict local features of images in the index and orange cells contain the subset of features which have to be processed for one query image feature during the retrieval. Please note that, after introducing new dimensions, fewer features have to be processed as all features are spread along all dimensions. The yellow cells indicate a setup where two context feature dimensions are combined by `OR` instead of `AND` (which is used in the orange cells).

III. Encoding context information

The aim of considering the larger context of a local feature is to include information about the larger neighborhood which may have certain similarities in other images of the same object or scene. When trying to represent the environment it is essential that the representation supports the existing invariances of the features (translation, scale and rotation for the SIFT features used in this work) to keep these invariances in a final CBIR system. We propose a representation based on pooling features from deep convolutional neural network layer outputs.

Since in 2012 Deep Convolutional Neural Networks (CNNs) yielded stunning results in classification tasks [24], they turned out to effortlessly improve the state-of-the-art in many other computer vision domains [5]. For image retrieval, the use of convolutional layer activations [8–11] showed to result in better performance than the fully-connected layers [6], [7]. More specifically, images are typically represented by a global descriptor which is obtained by sum or max-pooling over the feature channel maps of the last convolutional layer of the neural network. We adopt this for representing one local feature's context information using the 512 feature maps of the conv5_3 layer of the pre-trained VGG16 network [25]. For taking into account the different aspect ratios, we actually use three different networks for landscape, portrait and quadratic images respectively which results in feature maps of three different sizes. Images are rescaled to 800 pixels for the longest side before being processed by the networks. Since we are interested in a local representation, we

perform pooling only in a spatial quadratic region with its position and size being determined by the local feature's position and scale. Please note that this rougly preserves the invariances w.r.t. scale and orientation of the underlying SIFT features. Finally, the spatially pooled features are l2 normalized. For quantizing the context features, we use k-means clustering. Figure 1 summarizes the proposed feature for encoding context information.

IV. Evaluation framework

When designing a feature for encoding the larger neighborhood of a local feature in order to enhance its bag-of-words representation, the best evaluation would naturally be to measure the benefits with respect to the retrieval accuracy of an overall image retrieval system. However, often, indexing large image datasets with all context features cannot be carried out each time for balancing every parameter. We therefore propose a way to evade this by compiling a dataset framework which only considers the relevant parts, i.e. which models the retrieval system's view to the features. More precisely, we consider the two possibilities every BoW-match can be looked upon: either it is a correct match arising from a real object correspondence or it is an incorrect match originating from the quantization loss or random background clutter etc.

Given the datasets, we thus compile these two sets of feature pairs (correct and incorrect BoW matches) and evaluate the involved quantized context numbers accordingly, i.e. the feature pairs of a correct BoW match should also agree w.r.t. their quantized context number whereas for incorrect BoW matches, we want the context numbers to be different. We therefore measure the *False Negative Rate* (FNR, the number of correct BoW matching pairs that are not quantized to the same value) and the *False Positive Rate* (FPR, the number of incorrect BoW matches that are unluckily quantized to the same value). Ideally, both FNR and FPR are low to not loose any recall and to skip all incorrect matches during retrieval, respectively. We additionally perform experiments by combining different context features using `AND` and `OR` combinations. In these cases, FNR and FPR are adapted accordingly, e.g. for `AND`, a False Negative occurs if for a correct BoW match none or only one context feature yields identical quantized values.

V. Experiments

For experiments, we use four public datasets often used in CBIR:

- **Oxford5k** [13] containing 5,062 images including 5 query images and several corresponding images for each of 11 different buildings.
- **Paris6k** [26] containing 6,392 images including 5 query images and several corresponding images for each of 11 different buildings.
- **Holidays** [18] containing 1,491 images of 500 different scenes.
- **Landmarks** [6] originally containing 213,678 images of 672 different landmarks. We use the "clean" subset and due to broken links and after removing landmarks without corresponding images, we obtained 35,224 images of 586 landmarks.

From all images - rescaled to 800 pixels for the longest side - we extract SIFT features [2] at Difference-of-Gaussians Keypoints and apply the RootSift normalization[14]. Using the features from Oxford dataset, we generate a visual Codebook of size 100,000 by hierarchical k-means clustering which is used for bag-of-words quantization of local features in all our experiments.

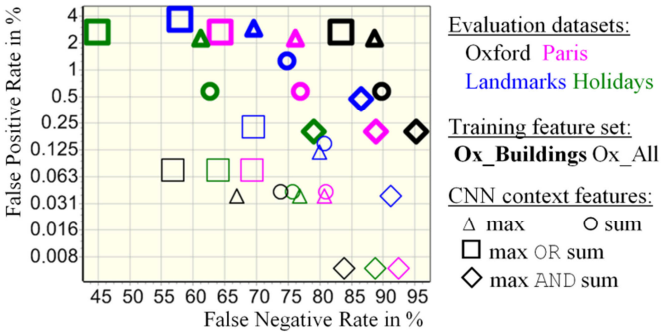

Figure 3. Evaluation of the discriminability of the proposed context features and their combinations for four public datasets (models trained on Oxford). Please note the logarithmic scale of the False Positive Rate.

A. Discriminability of CNN-based context features

Using the evaluation framework described in section 4, we evaluate the discriminability of the context features as follows:

1) Identifying feature pairs for correct BoW-matches: Given the annotations of the datasets, we identify all pairs of images showing the same object or scene. Since the Landmarks dataset has some groups with very many images, we limit the number of randomly chosen pairs for each landmark to 1,000. For each pair of corresponding images, we then extract a set of 'correct matches' (see Section 4) by taking the bag-of-words matches as starting point and subsequently keeping only those matches which are proved to be real feature correspondences. To ensure this, we also match the raw (not quantized) local features of the two images using the well-known ratio check (nearest neighbor vs. second nearest neighbor in descriptor space) [2] and additionally perform spatial verification, see [27] for details on the geometric consistency checks of neighboring features we use. As the datasets contain some near-duplicates (taken at the same time with the same camera etc. and thus having a lot of matches) we limit (by random selection) the number of extracted correct matches per image pair to 100. The main reason for that is that the matches will subsequently be processed with respect to their local image context in the neighborhoods of the respective features, and if we allow too many matches per image, the respective contexts will often overlap which limits their informative content for the experiments. Furthermore, for Paris6k, we limit the number of image pairs by random selection in order to obtain a similar number of matching features pairs compared to Oxford5k and Landmarks datasets.

2) Identifying feature pairs for incorrect BoW-matches: For evaluating the ability of the two different context features with respect to filtering out incorrect Bow-matches (i.e. for calculating the false negative rates, see Section 4) we report results based on two sets of feature pairs:

OxPa_iBoW: we randomly take pairs of images (each time one image from Oxford5k and one from Paris6k), calculate the BoW matches and randomly keep 30% of them. Given the fact that the images from Oxford5k are taken in Oxford and those from Paris6k in Paris, virtually all of those should be incorrect BoW matches.

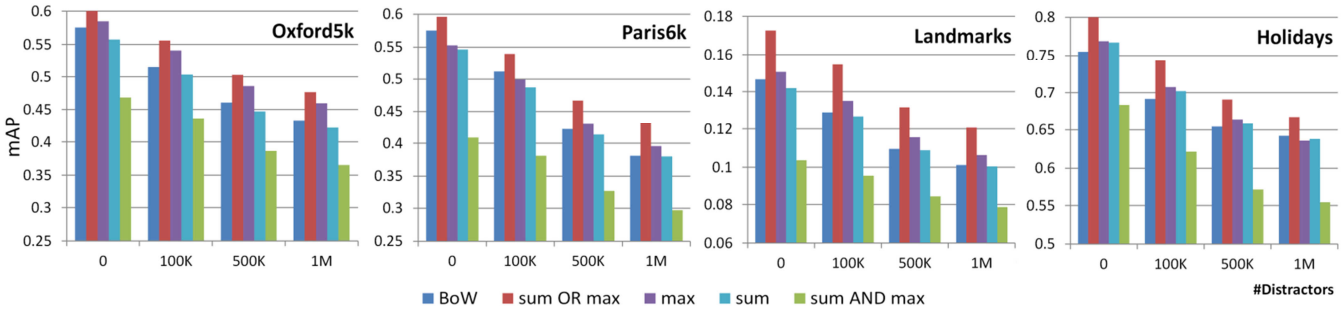

Figure 4. Evaluation of different variants of context features within a large-scale image retrieval setup: compared to the baseline bag-of-words model (BoW), integrating the OR-combination of max- and sum-pooled CNN features into the inverse file improves results for all databases of all sizes. The zeros on the abscissa indicate that only the images within respective datasets have been used (e.g. 5,062 for Oxford5k) and no further images from the MIRFlickr1M-dataset have been added as distractors. Please note the different intervals of the ordinate which shows mean average precision.

La_iBoW: we randomly take pairs of images from different landmarks of the Landmarks dataset, calculate the BoW matches and randomly keep 30% of them.

3) Feature set for learning the quantizers. We train all our models, i.e. the Codebook for BoW-quantization of features and the quantizers with data from Oxford5k only and use two different sets of context features for learning the quantizers:

Ox_All, consisting of about 2 million context features randomly calculated from all features in all Oxford5k images and

Ox_Buildings, which contains all features of the correct BoW matches from Oxford5k dataset, i.e. the latter set contains only features from buildings. Both feature sets are k-means clustered into 10,000 context words to obtain the quantizers.

See Table 1 for details on numbers of images in the datasets, identified pairs of corresponding images, used pairs of images and extracted pairs of features, respectively.

Figure 3 condenses the results for the CNN-based context features. As can be seen, the quantizers trained with *Ox_Buildings* yield fewer FNRs at the price of dramatically increased FPRs. In other words, it is essential to train the quantizers with background clutter in order to be able to filter out incorrect BoW matches. Furthermore, OR-combining sum- and max-pooled CNN features interestingly boosts FNR without sacrificing too much FPR. For example, the holidays dataset (green square) yields a FNR of 63.83% and a FPR of 0.0748%, which means that 36 out of 100 correct BoW-matches are saved while only one incorrect BoW-match remains out of 1,336.

B. Large-scale Image Retrieval evaluation

After modelling the image retrieval system's view to the context features to evaluate their discriminability (FNR vs. FPR), the question arises which trade-off to take in order to find as much relevant images as possible in large-scale datasets. To this end, we also measure the overall benefit of adding the features' contexts into a complete retrieval system. We focus on the quantizers build with *Ox_All* and evaluate the retrieval setup on the datasets with the common mean average precision (mAP) values. For the large-scale setups, we incrementally add up to one million images from the public MIRFlickr1M dataset[28].

The results in Figure 4 show that integrating both max- and sum-pooled CNN features as a second and third dimension of the index using the OR combination leads to a significant increase of mAP scores across all database and sizes. Interestingly, integrating just one context feature performs as good as the BoW baseline in most cases although much less features have to be processed in the accumulator. Another order of magnitude fewer features are processed using the AND combination where both the max-pooled and the sum-pooled feature quantizations (and the BoW of course) have to be identical to cast votes for a database image. Given this strong filtering mechanism, still more than 75% of the corresponding images are found among the 1 million distractors compared to the BoW baseline.

C. Comparison with other work

In Table 2 we list other methods which integrate context information into the index in order to improve retrieval accurracy. To the best of our knowledge, only the coupled multi-index [29] is comparable to our strategy of adding new dimensions to the index by quantizing the context features. Unfortunately, many interesting papers only give results for small database sizes without distractors or only use near-duplicate retrieval datasets. And even reported large-scale experiments cannot be directly compared to our setup for several reasons:

- Different SIFT Keypoint Detectors (Hessian-affine vs. DoG)
- RootSIFT normalization [30] not used in early works
- Different distractor sets of 1 million images (Flickr1M, MIRFlickr1M, Panoramio1M, self-crawled)
- Some report evaluation results from models trained on the same dataset
- Feature quantization (Codebook size, hard or soft assignment to several visual words)
- Different Query images (cropped to the object vs. non-cropped including clutter; removing true corresponding images in distractors)

VI. Conclusion

In this work, we proposed ways to increase the discriminability of bag-of-words based representations of local features in the context of image retrieval. We extended a local feature with more information

Table 1. Overview of the datasets used for experiments and the derived feature sets for training and evaluation of the quantization process. Numbers with * denote that not all possible combinations have been propagated, see text for the reasons.

Dataset / *Feature set*	Imgs in dataset	Correspon-ding image pairs	Pairs used	Number of feature pairs or features (feat')
Oxford5k	5,063	33,892	18,002	612,295 pairs
Paris6k	6,393	176,440	26,441*	693,128 pairs
Landmarks	32,720	242,533*	35,447	524,014 pairs
Holidays	1,485	2,060	1,399	79,660 pairs
OxPa_iBoW	*for evaluating FPR*		14,834	990,189 pairs
La_iBoW	*for evaluating FPR*		26,694	645,053 pairs
Ox_All	*for training the quantizers*			1,974,698 feat'
Ox_Building	*for training the quantizers*			1,224,590 feat'

from its larger neighborhood comparing different combinations of a representation based on pooling features from deep convolutional neural network layer outputs. After defining an evaluation framework which models the retrieval system's view, i.e. which measures performance after quantization of the features' contexts, our experiments carried out with four public datasets contrast different combinations of sum- and max-pooled CNN features. Leveraging both the maximum and the sum of the convolutional layers maps leads to significant better retrieval results than the bag-of-words baseline while requiring much less memory access. In future work, we will evaluate on more recent CNN architectures and examine how different memory setups can make use of the proposed multi-dimensional index. For instance, during retrieval, if the AND combination of context features and BoW features requires very few memory accesses, the index could be stored on a soild state disk rather than in the expensive and limited main memory.

ACKNOWLEDGMENT

This work was supported by the German Ministry of Education and Research (BMBF) (grant number 13N14028).

REFERENCES

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

[2] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision*, vol. 60, no. 2, pp. 91–110, 2004.

[3] H. Jégou, M. Douze, C. Schmid, and P. Pérez, "Aggregating local descriptors into a compact image representation," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 3304–3311.

[4] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[5] A. Sharif Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: an astounding baseline for recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2014, pp. 806–813.

[6] A. Babenko, A. Slesarev, A. Chigorin, and V. Lempitsky, "Neural codes for image retrieval," in *European Conference on Computer Vision*, 2014, pp. 584–599.

[7] Y. Gong, L. Wang, R. Guo, and S. Lazebnik, "Multi-scale orderless pooling of deep convolutional activation features," in *European Conference on Computer Vision*, 2014, pp. 392–407.

[8] A. Babenko and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1269–1277.

[9] Y. Kalantidis, C. Mellina, and S. Osindero, "Cross-dimensional weighting for aggregated deep convolutional features," *arXiv preprint arXiv:1512.04065*, 2015.

[10] A. S. Razavian, J. Sullivan, A. Maki, and S. Carlsson, "A baseline for visual instance retrieval with deep convolutional networks," *arXiv preprint arXiv:1412.6574*, 2014.

[11] G. Tolias, R. Sicre, and H. Jégou, "Particular object retrieval with integral max-pooling of CNN activations," *arXiv preprint arXiv:1511.05879*, 2015.

[12] M. Perd'och, O. Chum, and J. Matas, "Efficient representation of local geometry for large scale object retrieval," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, 2009, pp. 9–16.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[14] R. Arandjelovi and A. Zisserman, "Three things everyone should know to improve object retrieval," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2911–2918.

[15] O. Chum, A. Mikulik, M. Perdoch, and J. Matas, "Total recall II: Query expansion revisited," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 889–896.

[16] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman, "Total recall: Automatic query expansion with a generative feature model for object retrieval," in *2007 IEEE 11th International Conference on Computer Vision*, 2007, pp. 1–8.

[17] G. Tolias and H. Jégou, "Visual query expansion with or without geometry: refining local descriptors by feature aggregation," *Pattern Recognition*, vol. 47, no. 10, pp. 3466–3476, 2014.

[18] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *European conference on computer vision*, 2008, pp. 304–317.

[19] Y. Zhang, Z. Jia, and T. Chen, "Image retrieval with geometry-preserving visual phrases," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, 2011, pp. 809–816.

[20] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu, "Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 3013–3020.

[21] S. Zhang, Q. Tian, Q. Huang, W. Gao, and Y. Rui, "Multi-order visual phrase for scalable image search," in *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, 2013, pp. 145–149.

[22] L. Zheng, S. Wang, Z. Liu, and Q. Tian, "Packing and padding: Coupled multi-index for accurate image retrieval," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1939–1946.

[23] F. S. Khan, R. M. Anwer, J. Van De Weijer, A. D. Bagdanov, M. Vanrell, and A. M. Lopez, "Color attributes for object detection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 3306–3313.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, 2008, pp. 1–8.

Table 2. Comparison of different methods that integrate context or the geometry of surrounding features into the index. “Strategy” refers to the three different ways for integration (adding a second Dimension or filtering or extending the accumulator during retreival; see Section 2 for details). All results are mAP scores calculated on one of the three datasets after adding 1 million distractor images. “BoW” refers to the baseline bag-of-words approach and “+Context” refers to the respective approach integrating context. See Section 5.C for more details.

Method	Strategy	Oxford BoW	Oxford +Context	Paris BoW	Paris +Context	Holidays BoW	Holidays +Context
Ours	2D-Index	43.3	47.6	38.0	43.1	64.2	66.7
Coupled Multi-index [22]	2D-Index	-	-	-	-	23.0	48.0
Spatial Bag-of-Features [31]	Filtering	40.8	55.0	27.8	39.1	-	-
Self-Contained Contextual Binary Code [32]	Filtering	-	-	20.8	27.1	31.8	48.4
Multi-order Visual Phrase [21]	Filtering	49.3	62.1	-	-	-	-
Geometry-preserving visual phrases [19]	Acc. Ext.	41.3	53.2	-	-	-	-
Visual Phraselet [33]	Acc. Ext.	44.7	55.7	-	-	-	-
Spatially constrained similarity measure [20]	Acc. Ext.	53.5	68.5	63.0	74.1	-	-
Weak geometric consistency [18]	Acc. Ext.	-	-	-	-	32.0	44.0

[27] D. Manger, M. Kubietziel, and others, “Filtering local features for logo detection and localization in sports videos,” in *2015 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2015, pp. 505–508.

[28] B. T. Mark J. Huiskes and M. S. Lew, “New Trends and Ideas in Visual Concept Detection: The MIR Flickr Retrieval Evaluation Initiative,” in *MIR ’10: Proceedings of the 2010 ACM International Conference on Multimedia Information Retrieval*, 2010, pp. 527–536.

[29] L. Zheng, S. Wang, Z. Liu, and Q. Tian, “Packing and padding: Coupled multi-index for accurate image retrieval,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1939–1946.

[30] R. Arandjelovi and A. Zisserman, “Three things everyone should know to improve object retrieval,” in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, 2012, pp. 2911–2918.

[31] Y. Cao, C. Wang, Z. Li, L. Zhang, and L. Zhang, “Spatial-bag-of-features,” in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, 2010, pp. 3352–3359.

[32] Z. Liu, H. Li, W. Zhou, R. Zhao, and Q. Tian, “Contextual hashing for large-scale image search,” *IEEE Transactions on Image Processing*, vol. 23, no. 4, pp. 1606–1614, 2014.

[33] L. Zheng and S. Wang, “Visual phraselet: Refining spatial constraints for large scale image search,” *IEEE Signal Processing Letters*, vol. 20, no. 4, pp. 391–394, 2013.