Extending the Bag-of-Words Representation with Neighboring Local Features and Deep Convolutional Features

Daniel Manger and Dieter Willersinn

Fraunhofer IOSB, Karlsruhe, Germany

Abstract

In this work, we propose and compare two methods to extend the bag-of-words representation which is still widely used in the domain of content-based image retrieval where a query image is used to search for those images in a large image database that show the same object or scene. To this end, typically, local features such as SIFT are quantized and treated independently to leverage an inverted file indexing scheme for speedup. As the quantization of local features impairs their discriminability, the ability to retrieve the relevant database images is decreasing in larger databases. We address this issue by extending every quantized local feature with information from its local spatial neighborhood. More precisely, we make use of two approaches widely used for global image features: the Fisher Vector representation aggregating the neighboring local features and a representation based on pooling features from deep convolutional neural network layer outputs. Using four public datasets, we evaluate the representations in terms of their performance after quantization.

Keywords: Content-based Image Retrieval, Bag-of-Words, Spatial Context of Local Features

1 Introduction

While becoming apparent in many successful applications, e.g. apps for recognizing items based on snapshots of mobile devices, content-based image retrieval (CBIR) with local features is still limited to small databases compared to today's web-scale flood of pictures. The main reason for that is, that the methods to find common local content in a pair of images based on local features such as SIFT do not scale for searching in databases with billions of images. Nevertheless, since the seminal work of [Sivic and Zisserman, 2003] introducing the codebook-based quantization of features into visual words in order to manage the image retrieval problems with text retrieval methods, many work has been done to enrich the discriminative power of their bag-of-words (BoW) model. Alternatively, various *global* image representations based on local features have been proposed for image retrieval. The Fisher Vector encoding [Perronnin et al., 2010], for instance, aggregates local features into a high-dimensional embedding followed by a compression - typically PCA and whitening - to encode an image into a compact fix-sized code. Finally, the recent advances of convolutional neural networks (CNN) have led to many approaches using deep-learned features - either out-of-the-box or by pooling responses from fully-connected or convolutional layers. Since both CNN- and Fisher Vector-based representations are targeted for very compact codes of e.g. just 64 floats for one image, it becomes obvious that retrieving very small objects surrounded by plenty of heavily cluttered background becomes difficult for large databases.

In this work, we therefore analyze specific combinations of the three approaches by extending each quantized local feature in the BoW model with more context information from the respective local neighborhood encoded with a Fisher Vector- or CNN-based representation.

2 State of the Art

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. In this work, we neglect methods refining the shortlist (re-ranking) and focus on those approaches that incorporate additional information into the inverted file indexing scheme (termed index) so that *all* images can benefit from. These can be separated into three strategies:

Extending the accumulator which 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., 2008] 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.

Filtering of features: This keeps the accumulator compact (still one bin per database image) and adds additional information into the index to filter matches prior to casting votes into the accumulator. [Zhang et al., 2013] integrates information about the four closest features in the image coordinate space and during retrieval, each BoW match is further examined as to how many of the four neighboring features are consistent.

2D-Index: In order to overcome the runtime, performance and storage limits of both accumulator extension and filtering of features, [Zheng et al., 2014] 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 [Khan et al., 2012], 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 target the latter strategy of integrating context information as a second dimension into the inverse file. However, with Fisher Vector encodings of local features on the one hand and CNN features on the other hand, 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 leading to fewer memory accesses. Furthermore, 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).

3 Encoding context information

When encoding the larger context of a local 'central' feature into a context descriptor it is essential to not loose the existing invariances of the features (translation, scale and rotation for the SIFT features used in this work) for the final CBIR system. As first option, we encode local features which lie both in the spatial neighborhood and in nearby scales (yielding typically 10 to 100 features) using the Fisher Vector (FV) encoding. More specifically, we use their descriptors - reduced to 64D by PCA - and concatenate their spatial configuration relative to the central feature spending another four dimensions (scale-normalized distance, polar angle, feature scale ratio and difference of descriptor orientations). Using a GMM model with 32 mixtures, the 68D feature vectors are condensed into a fixed-size Fisher Vector of $62 \times 32 \times 2 = 4352$ dimensions.

As a second method to generate context descriptors, we pool the activations in the 512 channels of the last convolutional layer of the VGG16 network [Simonyan and Zisserman, 2014] in a rectangular region defined by the feature's position and scale. We use sum- and max-pooling since they proved successful for global CNN-based image retrieval [Babenko and Lempitsky, 2015, Tolias et al., 2015].

After reducing the Fisher Vectors to 512D with PCA and L2-normalizing the CNN features, we quantize both context features with separate Codebooks of sizes $10,000$ in order to obtain quantized context numbers.

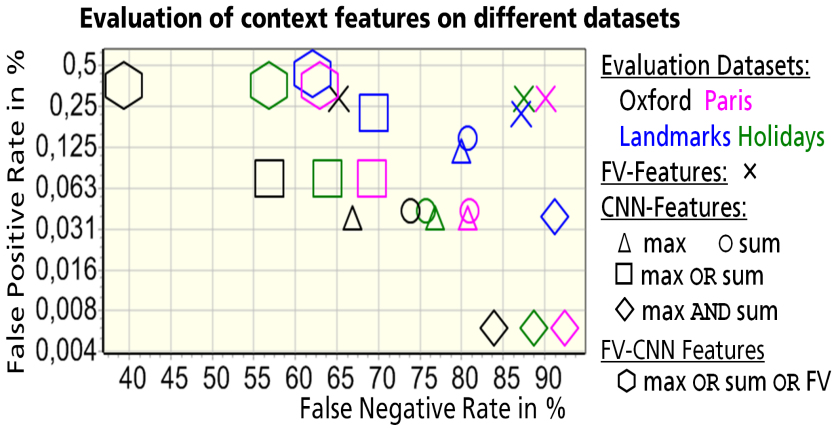

Figure 1: Overall results comparing the quantization of context features (Codebook size 10,000) based on Fisher Vectors (FV), CNNs and their combinations. Please note the logarithmic scale of the False Positive Rate (best viewed in color).

4 Evaluation framework and experiments

Instead of measuring the benefits with respect to the retrieval accuracy of an overall image retrieval system which would be very time consuming, we model 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 therefore 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 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 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.

For experiments, we use public datasets often used in CBIR: Oxford5k (5,062 images, 11 different buildings), Paris6k (6,392 images, 11 different buildings), Holidays (1,491 images, 500 different scenes) and Landmarks ("clean" subset 35,224 images due to broken links, 586 landmarks). We extract SIFT features and apply the RootSift normalization [Arandjelović and Zisserman, 2012]. The features from the Oxford dataset are used to 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. We identify feature pairs for correct BoW matches using the annotation of the datasets (specifying pairs of images that show the same object or scene) and subsequently filter matches with spatial verification. Thus, for each of Oxford5k, Paris6k and Landmarks, we obtain some 600,000 correct BoW matches and some 80,000 for Holidays. Feature pairs for incorrect BoW matches are collected by randomly taking pairs of images (each time one image from Oxford5k and one from Paris6k), calculating the BoW matches and randomly keeping 30% of them to obtain about 600,000 pairs. Given the fact that the images from Oxford5k and Paris6k are taken in different cities, virtually all of the BoW matches are incorrect BoW matches. For results on Landmarks dataset, we collect incorrect BoW matches by randomly taking pairs of images from different landmarks and keeping 30% of the BoW matches.

We train all our models, i.e. the BoW- and Context-Codebook, the GMM for the Fisher Vector representation, the PCA and the quantizers with data from Oxford5k only.

Figure 1 condenses the results for both CNN-based and FV-based context features. As can be seen, the CNN-based context features outperform FV-based features both in terms of FNR and FPR for all datasets. Interestingly, the OR-combination of sum- and max-pooled CNN features further boosts FNR without sacrificing too much FPR compared to the Fisher Vectors. For example, the holidays dataset (green square) yields a FNR of 63.83% and a FPR of 0.0748%, which means 36 out of 100 correct BoW matches are preserved while only one incorrect BoW match out of $1,336$ remains. Finally, OR-combining both sum- and max-pooled CNN and the FV-based context features offers another trade-off with a better FNR at more false positives.

5 Conclusion

In this work, we compared 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 from its larger neighborhood comparing a Fisher-Vector representation with a representation based on pooling features from deep convolutional neural network layer outputs. Representations based on CNN-features clearly outperformed the Fisher-Vectors and the combinations of different quantized features offer interesting trade-offs. The next step will be to evaluate the best representation in terms of the overall accuracy in a large-scale retrieval system with millions of images.

Acknowledgments

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

References

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

[Babenko and Lempitsky, 2015] Babenko, A. and Lempitsky, V. (2015). Aggregating local deep features for image retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1269–1277.

[Jegou et al., 2008] Jegou, H., Douze, M., and Schmid, C. (2008). Hamming embedding and weak geometric consistency for large scale image search. In *European conference on computer vision*, pages 304–317. Springer.

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

[Perronnin et al., 2010] Perronnin, F., Sánchez, J., and Mensink, T. (2010). Improving the fisher kernel for large-scale image classification. In *European Conference on Computer Vision*, pages 143–156. Springer.

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

[Sivic and Zisserman, 2003] Sivic, J. and Zisserman, A. (2003). Video google: A text retrieval approach to object matching in videos. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 1470–1477. IEEE.

[Tolias et al., 2015] Tolias, G., Sicre, R., and Jégou, H. (2015). Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*.

[Zhang et al., 2013] Zhang, S., Tian, Q., Huang, Q., Gao, W., and Rui, Y. (2013). Multi-order visual phrase for scalable image search. In *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*, pages 145–149. ACM.

[Zheng et al., 2014] Zheng, L., Wang, S., Liu, Z., and Tian, Q. (2014). Packing and padding: Coupled multi-index for accurate image retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1939–1946.