Discriminative Adaptation of Codebooks for Large-Scale Video Retrieval with Varying Content Queries[*]

Sangmin Oh, Amitha Perera, Anthony Hoogs
{sangmin.oh, amitha.perera, anthony.hoogs}@kitware.com
Kitware Inc. 28 Corporate Dr., Clifton Park, NY

The approach of using bag-of-words (BoW) or variants is ubiquitous in computer vision and related fields. The approach is very simple, nonetheless, works extremely well for diverse classes of problems. For retrieval problems in computer vision, each data (e.g., images or videos) is usually decomposed into spatially and temporally local regions where local descriptor vectors are extracted from each. Then, clustering is conducted to quantize these descriptor vectors into discrete words where each data is eventually represented as a BoW histograms. The formed BoW representations are used to train classifiers from given exemplars, then, also used as basis to find proper matches during retrieval.

In this work, we propose a scalable BoW-based indexing framework which can adapt to varying content query sets on the fly to maximize discriminative power during retrieval. Formally, our problem assumes a large archive of visual datasets from which matches will be retrieved. The archive data is assumed to be indexed apriori, where a strong constraint is imposed to prohibit indexing from being altered afterwards. In our problem, the retrieval process is initiated with a set of exemplars from which a classifier, e.g., SVMs or nearest neighbors, is trained and used to retrieve proper matches. For example, for video retrieval problems, there may be millions of Internet consumer videos, e.g., YouTube, and tens of exemplars may be given as varying content query sets for classes to be searched, such as 'wedding' and 'soccer game'.

Traditionally, a fixed codebook dictionary of a certain size is 'blindly' generated and used to index the archive data with BoW representations. Usually, the codebook is selected by evaluating the retrieval accuracy against a number of development queries from different classes. However, this approach has a limitation, because the size and centroids of dictionary may not be optimal for diverse future query classes. To alleviate this issue, schemes such as term-frequency-inverse-document-frequency (tf-idf) are frequently used to highlight the effect of rare words [2]. However, existing literature indicates that such solution does not always improve the search and retrieval results unfortunately [1]. Indeed, an optimal codebook which is versatile for any test class may not exist. Actually, it is apparent that the size and centroids of dictionary influence the retrieval results significantly, and the notion of optimality for a codebook will depend heavily on data and given content queries. Unfortunately, such conclusion simply confirms that retrieval performance with blindly constructed codewords is likely to be only sub-optimal.

A related machine learning approach which sheds some light on this optimality issue for codeword generation is the venue of 'discriminative' codebook clustering . Overall, discriminative clustering aims to produce an optimized codebook by incorporating auxiliary information on data labels. Most discriminative clustering process is designed to decrease intra-class distance and increase inter-class distance. Unfortunately, while such solution may be optimal for a particular class, the process incurs needs to re-index archive, which is often practically not feasible due to the sheer size of the archive dataset. Additionally, the user satisfaction on retrieval heavily depends on speed as well, accordingly, such discriminative clustering approach is not suitable for time-sensitive tasks which involve a large archive data.

Our approach in this work aims to address the above-mentioned shortfalls of both the blind or discriminative clustering for archive indexing. The closest work to ours is [4]. In detail, an overly segmented dictionary is generated in the first place, using approaches such as hierarchical topdown clustering where sufficiently large dictionary (e.g., 100K codewords) is generated. Such fine-clustering will leave us with dictionary with redundant levels of granularity. Then, these codewords are used to index entire archive data. Later, given a set of query data, we can select a subset of the codewords (e.g., 2K codewords), which are optimal for the query, from which we can learn a classifier using the labeled data. During retrieval, it is straightforward to map the original indexing to a new sparser dictionary.

[*] Topic: visual processing and pattern recognition. Preference: poster.

This way, class-specific dictionary can be generated or selected on the fly at minimal costs, without incurring needs to alter archive indexing anymore.

Our evaluation results show benefits of the proposed approach where we conducted preliminary experiments on both synthetic and large archive of real-world Internet consumer videos. We also compare difference in performance using multiple detailed components, such as hard vs soft assignments and recently popularized difference coding schemes, e.g., [3, 5].

Acknowledgements

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20069. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

[1] Yu-Gang Jiang, Chong-Wah Ngo, and Jun Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In *Proceedings of ACM International Conference on Image and Video Retrieval*, 2007.

[2] David Nister and Henrik Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006.

[3] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, 2010.

[4] J. Winn, A. Criminisi, and T. Minka. Object categorization by learned universal visual dictionary. In *ICCV*, 2005.

[5] X. Zhou, K. Yu, T. Zhang, and T. S. Huang. Image classification using super-vector coding of local image descriptors. In *ECCV*, 2010.