EXPLORING INTRA-BOW STATISTICS FOR IMPROVING VISUAL CATEGORIZATION

Usman Niaz, Bernard Merialdo

EURECOM, Campus SophiaTech,
450 Route des Chappes, 06410 Biot FRANCE

ABSTRACT

Research in video retrieval systems is mainly inspired by the state of the art text retrieval where high dimensional descriptors are quantized to visual words making a Bag Of Words (BOW) histogram for an image. For a small BOW model potentially different descriptors could get assigned to the same visual word. Recently however refinements have been proposed to recover some of this *representation loss* for this simplistic model of visual description by studying the distribution of descriptors within the visual words [1, 2, 3]. Following the same foot-steps we enhance the BOW by encoding the position of each of the descriptor inside the quantized cell according to its centroid. Embedding this information to represent images increases precision of video concept detection. We compare our method to a BOW based baseline on TRECVID 2007 and TRECVID 2010 [4] datasets and show that adding the refinement proposed always improves the semantic indexing task. We also compare our method to that of [3] and show that it outperforms the Hamming Embedding Similarity based classification on the TRECVID 2007 dataset and illustrates comparable performance on the TRECVID 2010 set.

1. INTRODUCTION

A popular method to represent video frames is the Bag Of Words (BOW) model based on quantization that is inspired from text retrieval [5]. It is a histogram based representation of scene description obtained through vector quantizing thousands of visual descriptors into a discrete visual dictionary. The visual descriptors are usually patch based features containing rich visual information on key interest points in video frames. The descriptor vectors are quantized using some unsupervised clustering process, like k-means, to divide the visual space into adjacent Voronoi cells. The centroids of the Voronoi cells in the visual space correspond to bins of the final histogram that counts the number of features assigned to that cell (bin) for an image. The number of clusters thus size of the histogram is fixed at the time of clustering. Each image is then represented by a fixed dimensional histogram which can be directly fed into a discriminative classifier like Support Vector machines (SVM) to build a model for each category.

For video concept detection typical size of BOW vector or visual dictionary varies from 200 to several thousands of words. This size is directly related to the categorization performance as well as retrieval efficiency. Precision of categorization increases with the size as there are more cells in the same clustering space but this affects the generalization ability of the model as noisy descriptors are miss-assigned. Contrarily if BOW size is kept small the discrimination is low as patches belonging to significantly different parts from the images are assigned to the same cell. There is thus a need to find a compromise between the dictionary size, its discrimination ability and its generalization capability. Although BOW is a very sparse representation, as with the increase in size the sparsity increases nevertheless the training and prediction efficiency are also affected with this increase. Moreover the time taken to construct the dictionary is prohibitive for a large number of centers. We build small dictionaries with an added refinement to overcome the coarseness of the BOW model.

In the rest of the paper section 2 presents a brief review of the state of the art and our inspiration for the proposed method followed by presentation of difference BOW in section 3. In section 4 we detail experimentation and present the results with the improvements. Finally section 5 concludes the paper.

2. RELATED WORK

In the recent years many researchers have improved the BOW framework. Perronnin et al. [6] represent each image with a bipartite histogram by building universal and class specific dictionaries. Authors in [7, 8] use soft assignment to assign a descriptor to r closest cluster centroids, instead of a single centroid. Nevertheless all the descriptors assigned to the same visual words are considered similar. Contrary to those methods we try to find the difference between descriptors assigned to the same visual word.

In [1] authors have localized each descriptor inside a Voronoi cell based on a Hamming Embedding (HE) mechanism to increase the precision of the BOW histogram. They perform image classification in a similarity space adapted from the HE based precision mechanism [3]. Jegou et al. [2] perform large scale image search using a signature obtained from calculating statistics on a small dictionary. The signature, called Vector of Locally Aggregated Descriptors (VLAD), is itself large and affects the problem efficiency. Their work is a non-probabilistic approximation of the Fisher

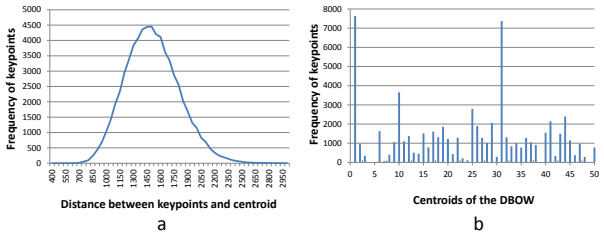

Fig. 1. Distribution of Difference vectors in a Voronoi cell: (a) magnitude of Difference vectors, (b) quantized over a 50-centroid global difference dictionary

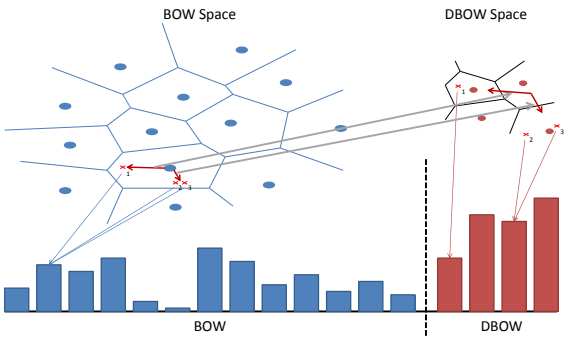

Fig. 2. Representation of BOW and DBOW

vectors [9] represented by the gradient vector of the log likelihood depicting the direction in which parameters of the generative model of data should be modified to best fit the data.

We find a compromise between retrieval efficiency and discriminative power of the BOW model. We add *important* but *little* information to the coarse dictionary by exploring intra-BOW statistics rendering it more discriminative. The codebook is enhanced by distinguishing the descriptors assigned to the same visual word, following [1]. We calculate the difference between descriptors and centroids as in [2] but instead of summing along each dimension we quantize them to build a secondary signature, much shorter in length than that of [2], to be used along with the primary BOW feature. Our method is directly adaptable in BOW framework as the added signature is computed directly from quantization information as opposed to conversion into a similarity space [3].

The proposed approach significantly improves video indexing performance for TRECVID 2007 and 2010 datasets. Results also indicate that adding the very small refinement signature to a smaller dictionary outperforms the performance of a much larger dictionary. We also show that the indexing performance is superior to Hamming Embedding [3].

3. DIFFERENCE BOW

In the BOW framework all the image descriptors assigned to the same visual word are considered identical, irrespective of their position in the high dimensional Voronoi cell. To make the visual dictionary more discriminative its size should be increased, however this affects the application efficiency along with the generalization capability of the dictionary. Furthermore the construction time of the dictionary increases with size. On the other hand a small dictionary generalizes well to noisy descriptors but does not give good classification results. Using a coarse dictionary for classification precision is lost as potentially different image patches are assigned to the same visual word due to their somewhat similar appearance in an image. Consider as an example a visual word depicting *tire* in the visual space. Now this word may contain tires from cars, motor bikes or even bicycles. We intend to further subdivide the space inside a Voronoi cell so that *tires* that are visually very close to each other group together.

We partition the high dimensional clustering space to measure the similarity between SIFT descriptors assigned to the same visual word. In [1] authors localize each descriptor inside the clustering cell and then generate a low dimensional binary signature capturing this localization. Following the same inspiration we employ a simple *global* mechanism to localize a descriptor inside a Voronoi cell focusing on the notion that two *distinct* descriptors inside the same cell should have a sub-signature different from each other. A trade-off between training / retrieval efficiency and discriminative power of the dictionary is also found as the information we encode is small in size compared to the BOW model but is meaningful.

We show distribution of Euclidean distances between the cluster centers and the assigned descriptors in the figure 1-(a) for a cluster (Voronoi cell) with maximum number of descriptors, from a 500 words dictionary on TRECVID 2007 training data. The distance is also the magnitude of the difference vector. Although the distance is normally distributed the range of values for distance is very large and could be a good candidate to differentiate descriptors within the cell. But doing so does not include any location which can be recovered by considering difference vector as a whole instead of its magnitude. Figure 1-(b) shows the histogram of difference vectors quantized over 50 bins for the same cell. We conclude that difference vector is a good candidate to further classify each descriptor inside a cell into a small number of classes.

We build a k-words visual dictionary $C = \{c_1 \dots c_k\}$ and to find the location of each descriptor x inside the Voronoi cell we follow [2] by calculating its difference from the nearest cluster center c_i as: $\overline{x} = x - c_i$. These difference vectors are quantized to generate a new dictionary $D = \{d_1 \dots d_l\}$ of size smaller than the original visual dictionary i.e. $(l << k)$. This leads to the formation of the Difference BOW (DBOW) feature that are extra dimensions added to the original vector to increase descriptor precision. The refinement proposed is shown in the figure 2 where keypoints belonging to the same visual word are assigned to different *difference words* based on their position inside the Voronoi cell. DBOW is

a global model which is calculated quantizing the difference vectors from all the clusters of BOW dictionary. Figure 1-(b) shows distribution of difference vectors from the most populous BOW visual word assigned using this global model. Both BOW and DBOW are used together to represent images.

3.1. Weighted DBOW

All words of the DBOW are given equal importance. Since each cluster has different number of descriptors but since the DBOW clustering is global each bin should have a separate weighting for difference vectors. Also DBOW vectors should be given image specific weights as for an example image certain bins might dominate the others in BOW.

We use frequency of keypoints in BOW to weight DBOW bins. Each difference vector belonging to a cluster is assigned a weight equal to the number of descriptors belonging to that cluster, or the size of that BOW bin: $w_{\overline{x}} = |x \in c_i|$. The weight of a DBOW word w_{d_j} is then calculated by adding the weights of all the difference vectors that are quantized to d_j

$$w_{d_j} = \sum_{\overline{x} \in d_j} w_{\overline{x}}$$

and normalized by the sum of weights for all DBOW bins. These weights are calculated for each image separately. We actually use square root of these weights in the experiments.

4. EXPERIMENTS

4.1. Experimental Setup

We have used TRECVID 2007 Sound and Vision database comprising 200 hours of 219 long videos [4] and demonstrate results on semantic indexing using the 20 concepts from TRECVID 2009. Also used is the TRECVID 2010 IACC [4] dataset containing 11644 internet videos of 400 hours with 50 concepts from TRECVID 2011 Light Semantic Indexing task. Datasets are divided equally into development and test sets according to TRECVID guidelines [4]. SIFT features are extracted from keypoints detected through LOG detector [10] from each video frame. These features are used to build visual dictionaries and also the HE signatures. We have used 1-vs-all SVM classifiers for each concept; with non-linear classifiers from LIBSVM [11] for 2007 dataset and linear SVM using homogeneous kernel map [12] for 2010 dataset. Classifier parameters are optimized on the development set. The baseline experiments are carried out on the simple BOW models of 500, 1000 and 2000 words. DBOWs of 10, 50 and 100 difference words are computed for each of these making three versions for each base dictionary. Version 1 is *baseline* + 10DBOW, version 2 is *baseline* + 50DBOW and so on.

HE Similarity Feature: We implemented the Hamming Embedding (HE) Similarity based image classification framework [3] following the similar steps with the number of hyperplanes $m = 64$ and a fixed threshold of $h_t = 22$. We have used SIFT descriptors from the development sets of TRECVID 2007 and TRECVID 2010 datasets separately to generate the HE hyperplanes (the median values) for each cluster center. This is also done separately for each of the three dictionary sizes from 500 to 2000 as they embed different clustering spaces. After cross validation we retain 0.3 and 0.5 as values of α for power normalization of the HE signature. For each base dictionary there are three versions of HE similarity: the un-normalized (version 1) and with α as 0.3 (version 2) and 0.5 (version 3).

	Methods	1	2	3
Base Dictionary 500 Words	**Baseline**	0.0739	-	-
	HE Similarity	0.0781	0.0741	0.0770
	DBOW	0.0764	0.0796	**0.0830**
	DBOW weighted	0.0761	0.0810	0.0821
Base Dictionary 1000 Words	**Baseline**	0.0796	-	-
	HE Similarity	0.0815	0.0777	0.0820
	DBOW	0.0804	0.0831	**0.0850**
	DBOW weighted	0.0821	0.0845	0.0813
Base Dictionary 2000 Words	**Baseline**	0.0814	-	-
	HE Similarity	0.0737	0.0768	0.0801
	DBOW	0.0824	0.0835	0.0854
	DBOW weighted	0.0848	**0.0868**	0.0829

Table 1. MAP for 20 concepts, TRECVID 2007

4.2. Results

TRECVID 2007: Table 1 shows the Mean Average Precision (MAP) scores for 20 concepts of the TRECVID 2007 dataset for the three versions of the HE similarity and DBOW based features and compares them to the baseline. Adding DBOW bins to the BOW improves results significantly even with the addition of as little as 10 difference bins. Another important result produced is the performance of the small BOW-DBOW feature compared to that of a large BOW feature. As evident from table 1 adding 100 difference words to base dictionary of 500 words outperforms the performance given by a baseline dictionary almost twice and even four times its size.

We show the increase or decrease of the performance of the methods relative to the baseline methods. For each of

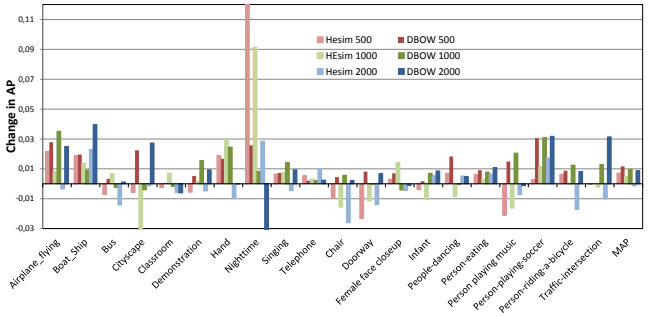

Fig. 3. Increase / decrease in AP over the baseline for the best performing HE and DBOW methods, TRECVID 2007.

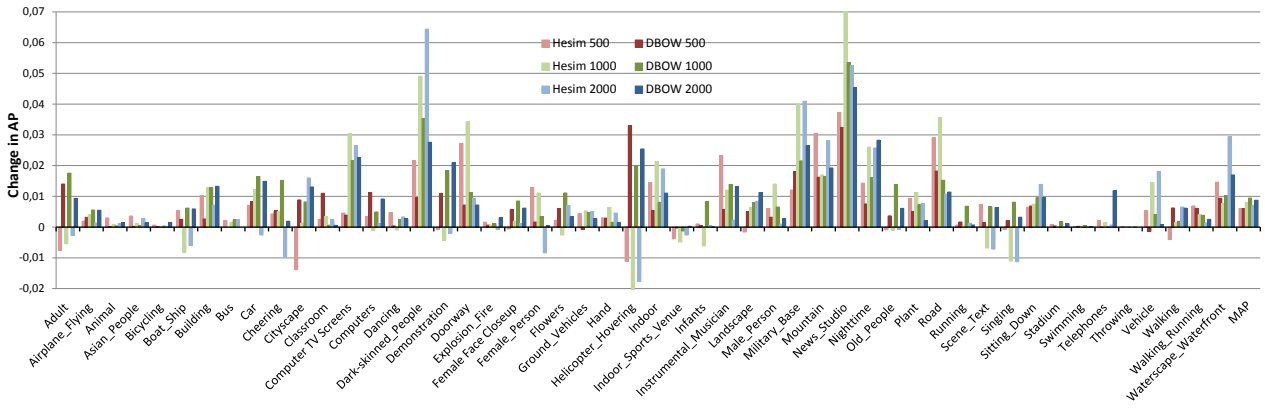

Fig. 4. Increase / decrease in AP over the baseline for the best performing HE and DBOW methods, TRECVID 2010.

the methods in Table 1 we select the best score for each concept and show its difference from the baseline score. Thus for HE similarity the best performance is chosen from the normalized and un-normalized versions of the method and for DBOW method the choice is made between varying sizes of DBOW. These relative scores are shown in figure 3. Note that DBOW outperforms HE similarity method for most of the 20 concepts for the three base dictionary sizes.

TRECVID 2010: As table 2 shows DBOW outperforms the baseline for all dictionary sizes and performs comparable to the HE similarity based method. DBOW performs comparable to the basic BOW double its size. The overall low scores are due to the linear SVM used on the complex dataset. Figure 4 shows the relative increase or decrease in performance of the 50 concepts for the best performing HE similarity and DBOW methods compared to the BOW baselines for the three dictionary sizes. Again HE is outperformed by DBOW for most of the concepts and overall also.

5. CONCLUSIONS

The introduction of DBOW bins encoding the position of descriptors inside the clustering subspace shows a lot of promise in rendering a compact dictionary more discriminative. We compared our method to a state of the art method for image classification [3] showing that it is consistently outperformed on two video datasets for the semantic indexing task.

We plan to find an effective weighting mechanism for the DBOW as well as finding the optimal DBOW size for a given visual dictionary. Since each high dimensional Voronoi cell has different dynamics a cell based quantization would result in more precise representation but at a higher cost. Concept-wise DBOW construction can be envisaged making a DBOW histogram separately for each concept using descriptors from images labeled only with that concept.

	Methods	1	2	3
Base Dictionary 500 Words	**Baseline**	0.0336	-	-
	HE Similarity	0.0353	0.0364	0.0356
	DBOW	0.0349	0.0359	**0.0365**
	DBOW weighted	0.0356	0.0352	0.0360
Base Dictionary 1000 Words	**Baseline**	0.0368	-	-
	HE Similarity	0.0406	0.0412	0.0420
	DBOW	0.0409	0.0413	0.0420
	DBOW weighted	0.0412	0.0413	**0.0421**
Base Dictionary 2000 Words	**Baseline**	0.0403	-	-
	HE Similarity	0.0447	0.0422	0.0445
	DBOW	0.0430	0.0419	0.0415
	DBOW weighted	0.0441	**0.0459**	0.0442

Table 2. MAP for 20 concepts, TRECVID 2010

6. REFERENCES

[1] H. Jegou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, 2008.

[2] H. Jegou, M. Douze, C. Schmid, and P. Perez, "Aggregating local descriptors into a compact image representation," in *CVPR*, 2010.

[3] M. Jain, R. Benmokhtar, H. Jegou, and P. Gros, "Hamming embedding similarity-based image classification," in *ICMR*, 2012.

[4] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and trecvid," in *MIR*, 2006.

[5] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, 2003.

[6] F. Perronnin, C. R. Dance, G. Csurka, and M. Bressan, "Adapted vocabularies for generic visual categorization," in *ECCV*, 2006.

[7] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[8] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J. M. Geusebroek, "Visual word ambiguity," *IEEE TPAMI*, vol. 32, no. 7, 2010.

[9] F. Perronnin and C. R. Dance, "Fisher kernels on visual vocabularies for image categorization.," in *CVPR*, 2007.

[10] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, November 2004.

[11] C. Chang and C. Lin, "LIBSVM: A library for support vector machines," *ACM TIST*, vol. 2, 2001.

[12] A. Vedaldi and B. Fulkerson, "VLFeat: An open and portable library of computer vision algorithms," 2008.