TOWARDS OPTIMAL VLAD FOR HUMAN ACTION RECOGNITION FROM STILL IMAGES

Lei Zhang[1]*, Xiantong Zhen*[2]*, Jiqing Han*[3]

[1] College of Information and Communication Engineering, Harbin Engineering University, Harbin, PRC

[2] The University of Western Ontario, London, ON, Canada

[3] School of Computer Science and Technology, Harbin Institute of Technology, Harbin, PRC

ABSTRACT

Human action recognition from still image has recently drawn increasing attention in human behavior analysis vision and also poses great challenges due to the huge inter ambiguity and intra variability. Vector of locally aggregated descriptors (VLAD) has achieved state-of-the-art performance in many image classification tasks based on local features. The great success of VLAD is largely due to its high descriptive ability and computational efficiency. In this paper, towards optimal VLAD representations for human action recognition from still images, we improve VLAD by tackling two important issues in VLAD including empty cavity and assignment ambiguity. The empty cavity issue severely compromises the performance of VLAD and has long been overlooked. We investigate the empty cavity and provide an effective solution to deal with it, which largely improves the performance of VLAD; we propose middle level assignments to conquer the assignment ambiguity, which are more reliable and can provide more useful information for realistic activity. We have conducted extensive experiments on two widely-used benchmarks to validate the proposed method for human action recognition from still images. Our method produces competitive performance with state-of-the-art algorithms.

***Index Terms*—** VLAD, empty cavity, ambiguity, generalized max pooling, human activity recognition

1. INTRODUCTION

Human action recognition [1, 2, 3, 4] from still images plays important role in human behavior analysis and poses even larger challenges compared with video-based action recognition [5, 6] due to the large intra-class variation and inter-class ambiguity. The bag-of-feature (BoF) model [7, 8] is one of the most effective framework in image and video representations based on local features. Recently, Fisher vector (FV) [9] and its non-probabilistic version, i.e., vector of locally aggregated descriptor (VLAD) [10] have been successfully used due to their high performances in different tasks including image retrieval, scene/object classification and human activity recognition [11].

BoF, FV and VLAD can be viewed in a unified framework which describes images with local features by two main components: codebook creation and feature encoding. Most researchers are focused on these two aspects to generate good codebooks and achieve discriminative encodings. However, the empty cavity issue always happens in these methods, which severely compromises the performance. Although [12] discussed the negative effect of empty cavity in BoF model and [13] gave the analysis on ambiguity of codeword assignment, the negative effects of the empty cavity issue and the imbalance influence of assignment that cause ambiguity in VLAD has long been neglected.

In this paper, towards optimal VLAD, we propose improving VLAD by addressing two key issues in VLAD: empty cavity and assignment ambiguity. We make the following two major contributions:

1) By exploiting the negative effect of empty cavity theoretically and experimentally, we propose an effective method to tackle the empty cavity issue, which can significantly improve the performance.

2) By investigating the distribution of the number of assignment in each codeword in one image, and rebuilding the relations of the imbalance assignment and codeword ambiguity, we propose selecting the reliable codewords to enhance the weight of the pooled vector related to those codewords in VLAD.

2. EMPTY CAVITY IN VLAD

We first provide an theoretical analysis of the empty cavity issue in VLAD from the perspective of kernels and then provide three solutions to deal with empty cavity to improve the performance of VLAD.

2.1. Negative impact of empty cavity

Empty cavity means during the assignment of local descriptors in one image, there are always some codewords with no

This work is partially sponsored by National Science Foundation of China (No. 61571147 and No. 91220301), National Science Foundation of Heilongjiang (F2015027)

local descriptor assigned to, leaving empty in the final representation. As a result, the obtained representation tends to be less discriminative and therefore compromises the recognition performance. The phenomenon of empty cavity does widely exist in both BoF and VLAD, which severely compromises the overall performance [12].

2.1.1. *Revisit of VLAD*

Let $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_K\}$ be the codebook learned by k-means approach. For one image X, each local descriptor $\mathbf{x}_t \in R^D$ is assigned to the nearest codeword as $\mathbf{u}_i = NN(\mathbf{x}_t)$. Moreover, the *pooled vector* $\mathbf{v}_i$ for codeword i is computed by :

$$\mathbf{v}_i = \sum_{\mathbf{x}_t : NN(\mathbf{x}_t) = \mathbf{u}_i} (\mathbf{x}_t - \mathbf{u}_i) \tag{1}$$

which is sum pooling strategy of the residual vectors, *i.e.*, the subtraction of local descriptor $\mathbf{x}_t$ and its belonging codeword. Finally, K pooled vectors are concatenated as a single $K \times D$ dimensional vector.

By deeply analyzing the pooled vector $\mathbf{v}_i$, it can be seen that the distance of local descriptors to codewords are encoded. Thus the similarities relative to the codewords during matching are all incorporated which increases the accuracy of measuring the relationship between local descriptors. While for BoF, only the number of pairs of local descriptors assigned to the same codeword in two images are counted, with no consideration about the similarity of local descriptors. This is the main reason for the better performance of VLAD.

2.1.2. *Negative effect of empty cavity from the kernel perspective*

Traditional methods including the standard VLAD do not provide any way to handle empty cavity phenomenon, which means the *pooled vector* $\mathbf{v}_i$ is zero for the empty cavity codeword. We will demonstrate the negative effect of this zero vector from the kernel aspect.

In both classification and retrieval tasks, the essential part is to compute the similarity between two images X and Y. Kernel is one way to fulfill this aim.

Let $\mathbf{X} = \{\overbrace{x_1, ...x_D}^{\mathbf{X}_1}, ..., \overbrace{x_{D\times(K-1)+1}, ...x_{D\times K}}^{\mathbf{X}_K}\}$ and be $\mathbf{Y} = \{\overbrace{y_1, ...y_D}^{\mathbf{Y}_1}, ..., \overbrace{y_{D\times(K-1)+1}, ...y_{D\times K}}^{\mathbf{Y}_K}\}$ the VLAD representations for image X and Y, respectively. Then we can kernelize the match between $\mathbf{X}$ and $\mathbf{Y}$ as:

$$K(\mathbf{X}, \mathbf{Y}) = \sum_i K_i(\mathbf{X}_i, \mathbf{Y}_i) \tag{2}$$

where $K_i(\mathbf{X}_i, \mathbf{Y}_i)$ represents the *partial kernel* in codeword i. The essence of Eq. (2) is to decompose the whole kernel function into several independent elements, and each element only focuses on the part generated by one cluster during the construction of VLAD.

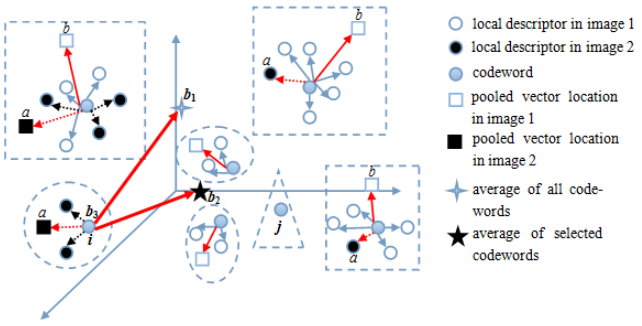

Fig. 1. Illustration of different solutions to empty cavity.

Without loss of generality, we consider a linear kernel widely used for recognition, and therefore we have

$$K(\mathbf{X}, \mathbf{Y}) = \mathbf{X}^T\mathbf{Y} = \sum_i \mathbf{X}_i^T\mathbf{Y}_i \tag{3}$$

Where $\mathbf{X}_i$ is part representation build on codeword i, namely, the *pooled vector* above, and $K_i(\mathbf{X}_i, \mathbf{Y}_i) = \mathbf{X}_i^T\mathbf{Y}_i$.

From Eq. (3), it is obvious that if $\mathbf{X}_i$ happens to be a zero vector caused by empty cavity in codeword i, and no matter what value $\mathbf{Y}_i$ is, the *partial kernel* K_i is the same. This kind of matching has misleading result under two conditions: 1) When the codeword is mutually missed in two images, *i.e.*, both $\mathbf{X}_i$ and $\mathbf{Y}_i$ are zero vectors, the codeword that can give more information will receive higher weights in the similarity measurement. 2) When the codeword is missed in only one image, that is, either $\mathbf{X}_i$ or $\mathbf{Y}_i$ is a zero vector, the final similarity measurement should vary with non-zero vector left, rather than fixed zero.

2.2. Solutions to empty cavity

In order to solve the empty cavity problem, a non-zero reference vector should be found for pooled vector $\mathbf{v}_i$ by intuition. To obtain a better insight into reference vector selection, Fig. 1 illustrates the procedure of VLAD and shows the meaning of a reference vector in our approach. In Fig.1, the real blue arrows and black dashed arrows represent the residual vector as $\mathbf{x}_t - \mathbf{u}_i$ in VLAD from two images. Moreover, the real red arrow and the dashed red arrows represent the pooled vectors from different images. For a certain codeword which local descriptors from both images are assigned to, as the dashed rectangles parts, the *partial kernel* is to compute the inner product between two vectors, the origin of which is the associated codeword.

However, there are also two other conditions for codewords as shown by dashed ellipses and dashed triangle in Fig.1. The dashed ellipse case is to show the mismatched empty cavity condition between two images (for example, for

codeword i, where only two local descriptors in image 2 are assigned and no points in image 1 is in this cavity). The other case is for co-missing codeword j in both two images, which is shown by dashed triangle part. The standard VLAD is to neglect the effect of this codeword under these two conditions. No matter how many local descriptors in image 2 are assigned to the codeword i, the partial kernel has no difference, which is zero. Similar, for the dashed triangle case, the partial kernel is also zero. To fix this problem, we aim to find a point as the reference to keep the pooled vector nonzero. For empty cavity from distinct image, this reference point should keep unchanged. b_1, b_2 and b_3 are three strategies of reference point which is independent to different image itself. They are type I to type III cases as follows.

Type I: We treat the mean of all codewords (blue four-points star) as the reference point in Fig. 1. Then the pooled vector for image 1 is the vector starting from codeword i and ending to this blue four-points star. The main idea for this strategy is to find the point with highest probability as the reference point in R^D space, since the average of all codewords is the real center of training data points. The corresponding pooled vector $\hat{v}_i$ is obtained by:

$$\hat{\mathbf{v}}_i = \frac{1}{K}\sum_{k=1}^{K}\mathbf{u}_k - \mathbf{u}_i \tag{4}$$

Type II: The nearby codewords with smaller assignments are similar to the codeword with no assignment. This is the case of the red arrow to five-points star in Fig. 1. The corresponding pooled vector $\hat{\mathbf{v}}_i$ is obtained by:

$$\hat{\mathbf{v}}_i = \frac{1}{\mathrm{card}(\mathscr{S})}\sum_{k\in\mathscr{S}}\mathbf{u}_k - \mathbf{u}_i \tag{5}$$

where $\mathscr{S}$ is the set that codeword is close to codeword i and at the same time, the number of points assigned to this codeword is smaller than a threshold.

Type III: For this case, we just treat the codeword itself as the pooled vector if empty cavity happens. The corresponding pooled vector $\hat{\mathbf{v}}_i$ is obtained by: $\hat{\mathbf{v}}_i = \mathbf{u}_i$. This is a simple but sometime effective way to handle the empty cavity problem.

3. AMBIGUITY IN ASSIGNMENT

The ambiguity in codeword assignment can be reflected in two-field. The first one is the misrepresentation situation, that is, the codeword could not represent the characteristic of local descriptors. The second one corresponds to the codewords with higher assignments, which however are from backgrounds carrying indiscriminate information for different images.

3.1. Analysis of ambiguity

Fig. 2 gives the location information of local descriptors with different number of assignments to one codeword. The blue

Fig. 2. The locations of local descriptors in images.

circles in left sub-figure represent the locations of local descriptors in the cavity with highest assignments, and each red stars are the locations for those cavities with only one assignment. Moreover, stars of four different colors (pink, red, yellow and green) in right in Fig. 2(b) show the locations of local descriptors in the cavity with middle level of assignments.

It is clear to see that the locations in Fig. 2 (a) could not grasp the essence of the image. In fact, the blue circles in Fig. 2 (a) expresses the burstness phenomenon, that is, a given visual element appears much more times in an image than a statistically independent model would predict [14]. Since the visual word with burstness mostly provides the background information, it should be attenuated during later processing. These peak or near peak assignments correspond to the ambiguity situation similar to that *function words* in the document vector space.

However, the red stars carry similar indiscriminate information for recognition which can be found in Fig. 2 (a), which is related to the misrepresentation of the ambiguity. It is interesting to find that this kind of one assignment appearance does not occasionally happen and may exist many times in one image (e.g., over 20 times as shown in Fig. 2 (a)). It also should be scaled down during recognition.

Moreover, most stars related to the middle level assignments in Fig. 2 (b) can represent the key information which is crucial to the recognition performance. The information grasped by the middle level assignments should be further enhanced. We propose a new method called middle-level assignment to handle the ambiguity in codeword assignment.

3.2. Middle-level assignments for Ambiguity

For the unbalance assignment to each codeword, we propose weighting the codewords with the middle level assignment in VLAD instead of penalizing the codewords with the higher or lower level assignments, since the middle level assignment has less ambiguity. In order to determine the range of middle level assignment, we assume the probability distribution function of the number of assignments as Gaussian distributions. Then the **range** can be determined by the mean and standard variance: **range** $= [\max(c_1, N_\mu - k_1 N_\sigma), \min(c_2, N_\mu + k_2 N_\sigma)]$, where N_μ is the mean number of assignments in one image, and σ is corresponding standard variance. k_1, k_2 c_1

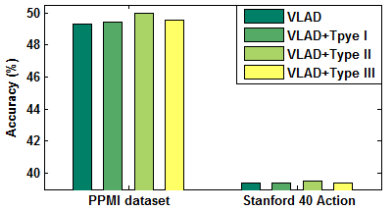

Fig. 3. Comparison of different solutions to the empty cavity problem.

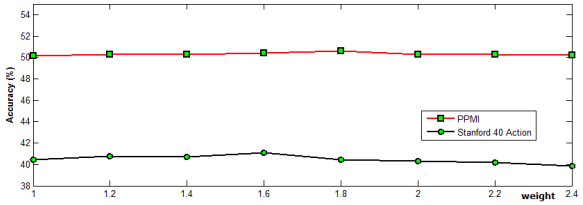

Fig. 4. Performance with different weights added to the pooled vectors in VLAD with middle level assignments.

and c_2 are constants determined experimentally. Then the pooled vector in VLAD is changed to:

$$v_i = \begin{cases} \sum\limits_{\{S_i|\mathbf{x}_t:NN(\mathbf{x}_t)=i\}} w(\mathbf{x}_t - \mathbf{u}_i) & if\ \mathrm{card}(S_i) \in \textbf{range} \\ \sum\limits_{\{S_i|\mathbf{x}_t:NN(\mathbf{x}_t)=i\}} (\mathbf{x}_t - \mathbf{u}_i) & if\ \mathrm{card}(S_i) \notin \textbf{range} \end{cases} \tag{6}$$

where $\mathrm{card}(S_i)$ means cardinal number of set S_i and w is bigger than one, and we vary this weight from 1 to 2.4 with 0.2 step increment.

4. EXPERIMENTS AND RESULTS

4.1. Experimental setting

For local descriptors, we extract dense SIFT descriptors with 3×3 grids for the better performance of dense sampling on image classification than sparse interest points [15]. The codebook is learned by the k-means clustering algorithm with randomly selecting 20% percent of training samples from each class for computational efficiency. The codebook size is selected as 512 unless specified. A linear SVM classifier [16] is adopted for final recognition and classification task to benchmark with compared algorithms. The parameters k_1, k_2, c_1 and c_2 are selected as 0.6, 1.8, 5 and 80, respectively.

PPMI [17] and Stanford datasets [18] are selected to verify the effectiveness of proposed approach.

4.2. Impact of different solutions to empty cavity

In Fig. 3, we report the comparisons of different solutions to empty cavity on the two datasets. It can be observed that all types of solutions outperform the standard VLAD. On PPMI and Stanford 40 Action datasets, the proposed solutions to the empty cavity problem have shown significant improved performance than baseline. Since the confusion between holding and playing instruments in PPMI dataset and the different background within-class in Stanford 40 dataset, the whole performances in Fig. 3 are still low.

4.3. Effect of different middle-level assignments

The performance with the variation of w from 2.4 with 0.2 is shown in Fig. 4. The trend is flattened in the upper sub-figure, while there is some oscillating in the lower sub-figure in both Fig. 4. From these results, it is clear to see that the enhancement of the codewords with middle level assignment and can improve the performance significantly.

Dataset	Accuracy		MAP	
	VLAD	Our solution	VLAD	Our solution
PPMI	47.33%	50.50% (1.8/Type II)	47.07%	48.60% (1.8/Type II)
Stanford 40 Action	37.79%	40.37% (1.4/Type II)	34.72%	36.78% (1.4/Type II)

Table 1. Comparisons on the two datasets.

4.4. Comparison with state-of-the-art

We compare our method with existing algorithms in Table 4.4. It can be seen that our method produces a much better performance than most of the state-of-the-art algorithms. Our method has great superior performance on the PPMI and produces the competitive results on the Stanford 40 Action datasets compared to other sophisticated approaches. The much better performance of our method on the two datasets has validated the effectiveness of the proposed solutions to the empty cavity problem.

5. CONCLUSION

We have presented an improved VLAD for human action recognition from still images based on local features. We have successfully tackled the empty cavity and ambiguity problems. The obtained optimal VLAD significantly improves the performance of baseline VLAD and achieves competitive and even better performance than state-of-the-art algorithms on two widely used benchmark datasets for human action recognition from still images.

PPMI		Stanford 40 Action	
method	MAP	method	MAP
VLAD [19]	47.07%	OB [20]	32.50%
SPM [21]	39.10%	SPM [21]	34.90%
LLC [22]	41.80%	LLC [22]	35.20%
Grouplets [17]	36.70%	Sparse bases [18]	45.70%
BoW	22.70%	EPM [23]	42.20%
Our method	**48.06%**	Our method	**37.14%**

Table 2. Comparison of our scheme with the state-of-the-art on the two datasets.

6. REFERENCES

[1] X. Zhen and L. Shao, "Introduction to human action recognition," *Wiley Encyclopedia of Electrical and Electronics Engineering*.

[2] X. Zhen, L. Shao, D. Tao, and X. Li, "Embedding motion and structure features for action recognition," *IEEE TCSVT*.

[3] X. Zhen, L. Shao, and X. Li, "Action recognition by spatio-temporal oriented energies," *Information Sciences*, 2014.

[4] L. Shao, X. Zhen, D. Tao, and X. Li, "Spatio-temporal laplacian pyramid coding for action recognition," *IEEE Transactions on Cybernetics*, 2013.

[5] H. Liu, M. Liu, and Q Sun, "Learning directional co-occurrence for human action classification," in *ICASSP*, 2014.

[6] S. Manel, , M. Mahmoud, and Chokri A., "Spatio-temporal pyramidal accordion representation for human action recognition," in *ICASSP*, 2014.

[7] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *CVPR*, 2003.

[8] X. Zhen and L. Shao, "A local descriptor based on laplacian pyramid coding for action recognition," *Pattern Recognition Letters*, 2012.

[9] F. Perronnin and T. Sanchez, J.and Mensink, "Improving the fisher kernel for large-scale image classification," in *ECCV*, 2010.

[10] H. Jegou, F. Perronnin, M. Douze, and J. Sanchez, "Aggregating local image descriptors into compact codes," *IEEE T-PAMI*.

[11] K. Sande, C. Snock, and A. Smeulders, "Fisher and vlad with flair," in *CVPR*, 2014.

[12] Ondrej Chum Herv Jgou, "Negative evidences and cooccurences in image retrieval: The benefit of pca and whitening," in *ECCV*, 2012.

[13] J. van Hemert, C. Veenman, A. Smeulders, and J. Geusebroek, "Visual word ambiguity," *IEEE T-PAMI*, vol. 32, no. 7, pp. 1271–1283, 2010.

[14] H. Jegou, M. Douze, and C. Schmid, "On the burstiness of visual elements," in *CVPR*, 2009.

[15] E. Nowak, F.Jurie, and B. Triggs, "Sampling strategies for bag-of-features image classification," in *ECCV*, 2006.

[16] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "Liblinear: a library for large linear classification," *JMLR*.

[17] B. Yao and L. Fei-Fei, "Grouplet: A structured image representation for recognizing human and object interactions," in *CVPR*, 2010.

[18] B. Yao, X. Jiang, A. Khosla, A. Lin, L. Guibas, and Fei-Fei. L., "Human action recognition by learning bases of action attributes and parts," in *ICCV*, Barcelona, Spain, November 2011.

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

[20] L. Li, H. Su, E. Xing, and L. Fei-Fei, "Object bank: a high-level image representation for scene classification and semantic feature sparsification," in *NIPS*, 2010.

[21] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *CVPR*, 2006.

[22] J. Wang, J.Yang, K. Yu, F. Lv, T. Hunag, and Y.Gong, "Locality-constrainted linear coding for image classification," in *CVPR*, 2010.

[23] G. Sharma, F. Jurie, and C. Schmid, "Expanded parts model for human attribute and action recognition in still images," in *CVPR*, 2013.