A New Pooling Strategy based on Local Feature Distribution: A Case Study for Human Action Classification

Raquel Almeida, Zenilton Kleber Gonçalves do Patrocínio Jr., Silvio Jamil F. Guimarães
Audio-Visual Information Processing Laboratory (VIPLAB)
Graduate Program in Informatics – Computer Science Department
Pontifical Catholic University of Minas Gerais (PUC Minas)
raquel.almeida.685026@sga.pucminas.br
{zenilton, sjamil}@pucminas.br

***Abstract*—Mid-level representations are used to map sets of local features into one global representation for a given media descriptor. In visual pattern recognition tasks, Bag-of-Words (BoW) is one popular strategy, among many methods available in literature, due mainly by the simplicity in concept and implementation. Despite the overall good results achieved by BoW in many tasks, the method is unstable in high dimensional feature space and quantization errors are usually ignored in the final representation. To cope with these problems, we propose a new pooling function based on feature points distribution around codewords. We propose to use the standard deviation associated with each codeword to measure attribution discrepancy and weight the impact that feature points will assume in the final representation. The main contribution of this article is the study of more discriminative representations, which amplify values of feature points close to codewords border regions. Experiments were conducted in human action classification task and results demonstrated that our pooling strategy has improved the classification rates in 25.6% for UCF Sports dataset and 21.4% for UCF 11 dataset, with respect to the original pooling function used in BoW.**

I. Introduction

Data representation is critical in many computational tasks, and usually aims to incorporate desirable properties, such as accessibility, discrimination and compactness. When dealing with visual data, such as images and videos, the data representation plays a decisive protagonism, since there is still a big gap between what we are able to see and interpret and what we are able to process computationally. Advances in visual data processing were made by the use of low-level descriptors, which identify regions of interest in images or videos. Thanks to this identification, it is possible to extract useful features to be used in tasks such as indexation and classification. Concerning classification tasks, one very popular approach is to map the set of local descriptors into a single set, so-called mid-level representation.

Among the methods for creating mid-level representations, one should point out Bag-of-Words (BoW), Spatial Pyramids (SP) and Convolutional Neural Networks (CNN) for their notable results. As stated by [1], mid-level representations have three steps in common: (i) coding; (ii) pooling; and (iii) concatenation. Coding stands for the local transformation applied to features vectors, creating codebook words and extracting distribution characteristics. Pooling, in turn, explores the spatial relation between these characteristics; and concatenation constructs the final representation.

Despite the overall good results achieved by these representations, the creation of a compact and discriminative mid-level representation is a challenging problem due to: (i) noise information captured during the feature extraction; (ii) possible selection of irrelevant data; (iii) degradation of semantic relation between elements; (iv) manipulation and interpretation of high-dimensional elements; and (v) quantization errors ignored in the final representation. All these aspects can lead to improper representation and miss classification.

In [2] the authors highlighted the importance of pooling feature vectors over spatially local neighborhoods as a way to achieve invariance and robustness in mid-level representations. According to them, local coding makes pooling more tight to regions of the multidimensional feature space and the restrictions to certain regions boost the performance.

Following these directives, in this work we argue that the distribution of feature points around codewords, specifically in the frontier between codewords, can provide clues of quantization errors produced during feature extraction and codification. This hypotheses is grounded on content-based approaches which incorporate local information during pooling operation and by an empirical observation of feature points distribution inside codewords regions. The main contribution of this work is the study of discriminative representations of feature points, taking into account the spatial distribution for the pooling function, thanks to the amplification of small differences between points close to the border of the regions in a high-dimensional space.

This paper is organized as follows. Section II provides a review of local pooling strategies found in literature. Section III describes methodology to create a mid-level representation using the BoW model. Is also presented in Section III the mathematical formulation for BoW strategies and for the new error based pooling function. In Section IV,

we present the framework used to validate the proposed strategy, the experimental setup and results. Finally, some conclusions are drawn in Section V.

II. Local Pooling in Bag-of-Words model

Bag-of-Words (BoW) is a notable member of mid-level representations for the simplicity in concept and implementation, and also by the achieved results in many applications. BoW is an approach, first applied to textual retrieval tasks, which represents data in terms of an histogram of codewords frequency. In visual retrieval tasks, input data are documents with a set of unordered local descriptors representing the media and a codebook is usually created by an unsupervised clustering algorithm. According to [3], BoW approaches can be grouped into three main categories based on their encoding methods: (i) voting based; (ii) reconstruction based; and (iii) super-vector based encoding.

In voting based methods, given a set of unordered local descriptors, the output is a single vector representing the codewords frequency in the input document. Vector Quantization (VQ) [4] is the main strategy of voting based methods, in which, each local descriptor votes for just one codeword (hard-assignment) or for multiple codewords (soft-assignment). In [5], some improvements were obtained in VQ by smoothing the distribution during the pooling function. This approach models two types of ambiguity between codewords: (i) codeword uncertainty; and (ii) codeword plausibility. Codeword uncertainty indicates that one descriptor may distribute probability mass to more than one codeword, while codeword plausibility signifies that one descriptor may not be close enough to guarantee representation by any codeword. The authors argue that larger vocabularies increase the probability of multiple relevant codewords to represent one feature point, increasing codeword uncertainty.

BossaNova [6] proposes an image representation in a voting based model, which keeps more information during the pooling step by using a density-based pooling strategy. In [7], the authors follow BossaNova framework by exploring quantization errors in the used representation for human action classification, called BossaNova Directly to Video (BNDTV). Also following BossaNova, was proposed in [8], the use of density local information in conjunction with orientation information between local descriptors and codewords to pool features.

Reconstruction based methods are formulated in a least square framework to attain a small reconstruction error. Main strategies in this group are Sparse Coding [9], [10], Local Coordinate Coding (LCC) [11] and Locality-Constrained Linear Coding (LLC) [12], in which the last two methods explicit reinforce locality during coding and pooling steps. Inspired by convolutional neural networks architecture in [13] the authors propose a two layer hierarchical mix-pooling in a Sparse Coding scheme. In their model, the first level pooling is performed over intermediate-size regions collecting statistics of sparse vectors, and the second level pooling is applied in the previous level representation, incorporating local statistics in the final representation.

In a different direction, super-vector based approaches, also known as aggregation methods, use high-order statistics to create a high dimensional representation locally constrained during pooling. One should mention Fisher Vector (FV), introduced by [14], and VLAD [15] as distinguished methods of super-vector encoding. In [16] the authors propose to equalize the similarity between images patches and their pooled representation in a variation of FV, while in [17] and [18] the authors used these high-order statistics to perform a Gaussian distribution pooling and second-order average pooling over regions, respectively.

III. New pooling function in Bag-of-Words mid-level representation

In this work, we propose a new pooling function, based on feature distribution around codewords. This function is applied in Bag-of-Words model with Vector Quantization (VQ) encoding method and soft-probabilistic distribution. In the following, an overview of the method and mathematical formalization.

Let $\mathbb{X} = \left\{\mathbf{x}_j \in \mathbb{R}^d\right\}_{j=1}^{N}$ be an unordered set of d-dimensional descriptors $\mathbf{x}_j$ extracted from the data. Let also $\mathbb{C} = \{\mathbf{c}_m \in \mathbb{R}^d\}_{m=1}^{M}$ be the codebook learned by an unsupervised clustering algorithm, composed by a set of M codewords, also called prototypes or representatives. Consider $\mathbb{Z} \in \mathbb{R}^M$ as the final vector representation. As formalized in [1], the mapping from $\mathbb{X}$ to $\mathbb{Z}$ can be decomposed into three successive steps: (i) coding; (ii) pooling; and (iii) concatenation, as follows:

$$\alpha_j = f(\mathbf{x}_j), j \in [1, N] \quad \text{(coding)} \quad (1)$$

$$h_m = g(\alpha_m = \{\alpha_{m,j}\}_{j=1}^{N}), m \in [1, M] \quad \text{(pooling)} \quad (2)$$

$$z = [h_1^T, \ldots, h_M^T] \quad \text{(concatenation)} \quad (3)$$

In VQ [4], the coding function f aims to minimize the distance to codewords and pooling function g leverages these distances, as follows:

$$\alpha_{m,j} = 1 \text{ iff } j = \underset{1 \leq m \leq M}{\arg\min}\, D(\mathbf{c}_m, \mathbf{x}_j) \quad (4)$$

$$h_m = \frac{1}{N} \sum_{j=1}^{N} \alpha_{m,j} \quad (5)$$

in which $D(\mathbf{c}_m, \mathbf{x}_j)$ is the Euclidean distance between j-th descriptor and m-th codeword. A soft version of this approach, so-called soft-assignment, attributes $\mathbf{x}_j$ to the n nearest codewords, and usually presents better results than the hard version. In the soft-assignment version $n \in [1..N]$.

A soft probabilistic approach for the $\alpha_{m,j}$ function, proposed in [5], smooths the distribution over codeword regions by the following function:

$$\alpha_{m,j} = \frac{\exp(-\beta D(\mathbf{c}_m, \mathbf{x}_j))}{\sum_{k=1}^{M} \exp(-\beta D(\mathbf{c}_k, \mathbf{x}_j))} \quad (6)$$

in which β is a fixed parameter that controls the smoothness of the assignment. Despite the improvement achieved by this probabilistic distribution, there is no information about the distance-to-codeword in the final representation.

BossaNova [6] overcomes this issue by using a probability density-based local pooling strategy. BossaNova first computes the distance between $\mathbf{c}_m$ and $\mathbf{x}_j$, and divides codeword neighborhood in B bins. It also restricts the range of each region using α_m^{min} and α_m^{max} bounds, during computation of histogram z_m, as follows:

$$z_{m,b} = \text{card}\left(\mathbf{x}_j \mid \alpha_{m,j} \in \left[\frac{b}{B}; \frac{b+1}{B}\right]\right) \tag{7}$$

$$\frac{b}{B} \geq \alpha_m^{min} \quad \text{and} \quad \frac{b+1}{B} \leq \alpha_m^{max} \tag{8}$$

BossaNova implements the soft probabilistic distribution proposed by [5], but instead of a fixed parameter β, it uses a cluster-related parameter β_m based on each codeword standard deviation σ_m, more precisely $\beta_m = \sigma_m^{-2}$.

A. Proposed pooling strategy

In clustering methods that divide feature space into regions to create a codebook, data are represented by the regions centers and the standard deviation σ_m associated with each codeword $\mathbf{c}_m$ carries an important information about data distribution during codebook creation. Due to the curse of dimensionality and through empirical observations, it was notice that most of feature points are close to σ_m with a small difference between the points. We propose to amplify these differences adopting a new pooling function $\alpha'_{m,j}$, as follows:

$$\alpha'_{m,j} = \exp\left(\gamma \frac{E(\mathbf{x}_j)}{W(\mathbf{c}_m, \mathbf{x}_j)}\right) \tag{9}$$

in which, γ is an expansion controlling factor and

$$E(\mathbf{x}_j) = \sum_{k=1}^{M} (D(\mathbf{c}_k, \mathbf{x}_j))^2 \tag{10}$$

$$W(\mathbf{c}_m, \mathbf{x}_j) = \left(D(\mathbf{c}_m, \mathbf{x}_j)^2 - \sigma_m\right)^2 \tag{11}$$

In the proposed strategy, the term E represents the total assignment error between the descriptor $\mathbf{x}_j$ and the codebook $\mathbb{C}$, while the W term weights the error of descriptor $\mathbf{x}_j$ to the codeword $\mathbf{c}_m$ in relation to the standard deviation σ_m.

The conjunction of these terms enhances values of feature points closer to the frontier of the region associated to the codeword standard deviation and penalize those points that are more distant from the assigned codeword. This strategy creates representations more tight to feature points distribution, and experiments demonstrated that it is more accurate and discriminative.

Taking into account the local feature distributions around codewords with the new pooling function, we could rewrite the original pooling function for Vector Quantization (VQ) and BossaNova (BN) as follows:

$$h'_m = \sum_{j=1}^{N} \alpha'_{m,j} \quad \text{(VQ)} \tag{12}$$

$$z'_{m,b} = \text{sum}\left(\alpha'_{m,j} \mid \mathbf{x}_j \in \left[\frac{b}{B}; \frac{b+1}{B}\right]\right) \quad \text{(BN)} \tag{13}$$

IV. Experiments

In order to validate the proposed strategy, we developed video mid-level representations to be applied in human action classification task. We assessed the proposed pooling in two mid-level representations: (i) VQ with soft-assignment, here called traditional BoW for simplicity; and (ii) BossaNova extended formalism applied directly to video as in [7], called BNDTV.

In this section, we describe the human action classification framework, the datasets used for classification and the experimental setup. Moreover, a quantitative analysis, in terms of classification rates, comparing the results obtained with the new pooling strategy in both representations and also a comparison with state-of-the-art approaches for the datasets.

A. Human Action Classification Framework

Human action classification task can be defined as: given a video, we need to classify the action displayed into one of the pre-determined number of actions. An overview of the framework to achieve this, using mid-level representations, is presented in Figure 1, as in [7].

For the low-level cuboid and feature extraction we chose Dense Trajectories (DT) descriptor [19], a dense and high-dimensional descriptor, which achieved good results in action classification tasks. In DT, the cuboids are in the form of trajectories, obtained by densely tracking sampled points gathered with optical flow fields. After tracking, feature characteristics are extracted using Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF) and Motion Boundary Histogram (MBH) around the trajectories.

Next, a visual codebook must be created before the encoding, and it should be able to provide information about data distribution. Thus, we applied the k-means clustering algorithm in sampled feature points, storing the standard deviation associated with each cluster creation and using the cluster centers as codewords. The penalty error is applied in both Bow and BNDTV pooling functions.

Finally, once we obtained the final representation, we performed the action classification using a non-linear SVM with RBF kernel, which is a popular classifier that is used throughout different works for human action classification [19]. Since this classifier is vastly used in human action classification, it is interesting to use it to make fair comparisons between different approaches.

A growing trend of feature learning-based methods to create action representations can be observed in literature today, as in [20] and [21]. Even not directly been related to our proposal, it is necessary to provide a quick overview of these methods once, for the best of our knowledge, they

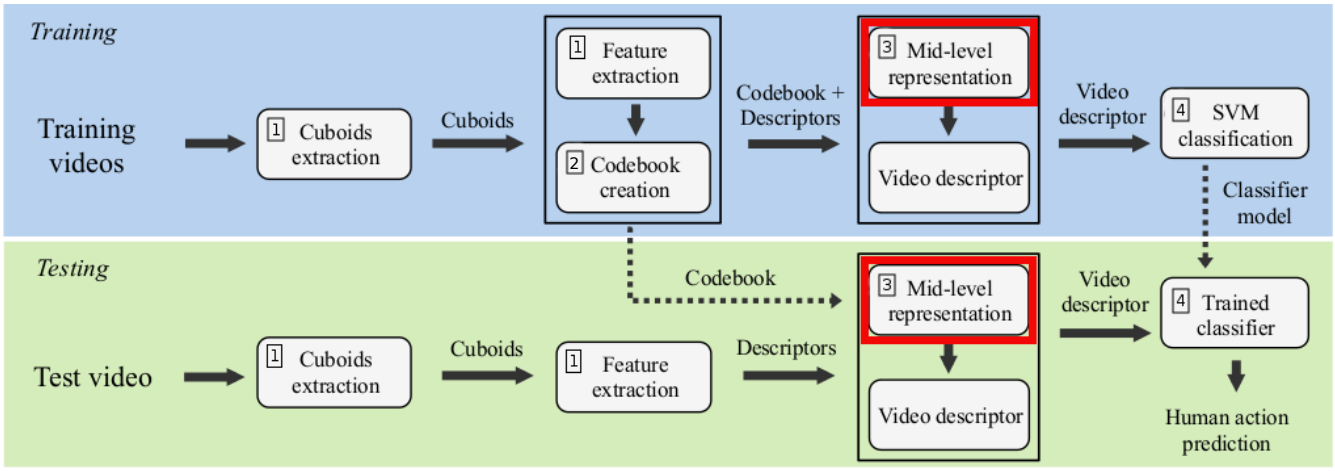


Fig. 1. Illustration of generic human action classification framework using mid-level representations, according to [7]. 1 In our application, cuboid and feature extraction is performed by the same algorithm, which generates the low-level descriptors. 2 Codebook creation can be performed by any clustering algorithm which can provides information about data distribution. 3 The proposed strategy operates during the mid-level representation step to create the global video descriptor, as detailed in Section III. 4 The classification final step is performed using the created mid-level video descriptor as input.

represent the state-of-the-art in human action classification task. Le et al. [20] proposed an independent sub-space analysis, called ISA, to learn spatial-temporal features from unlabeled data. Vrigkas et al. [21] proposed a method, called TMAR, which uses optical flow motion features to describe an action and a Gaussian mixture model to cluster then, in a learning-based framework.

B. Datasets and classification protocols

We tested in three well-known datasets for human action classification task: (i) KTH [22]; (ii) UCF Sports [23]; and (iii) UCF 11 [24]. We chose these datasets due their distinctive characteristics, such as dataset size, colorspace, video duration, intraclass variability and noise scene elements, as detailed in the following.

The KTH dataset [22] contains six types of human actions: walking, jogging, running, boxing, hand waving and hand clapping. These actions are performed by 25 different subjects in four scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and inside. Some examples are illustrated in Fig. 2. There are totally 600 video clips with 160×120 pixels size. We adopted the same experimental setup as in [22], so-called split, where the videos are divided into a training set (eight subjects), a validation set (eight subjects) and a test set (nine subjects).

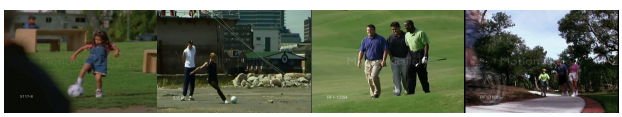
(a) Outdoors (b) Scaled (c) Clothing (d) Inside

Fig. 2. Example of hand waving with same subject in different scenarios in KTH dataset [22].

The UCF sports dataset [23] contains ten different types of sports actions: swinging, diving, kicking, weight-lifting, horse-riding, running, skateboarding, swinging at the high bar, golf swinging and walking. The dataset consists of 150 real videos with a large intra-class variability. Each action class is performed in different ways, and the frequencies of various actions also differ considerably, as can be seen in Fig. 3. Contrary to what has been done in many works that apply their methods on this dataset, we did not extend the dataset with flipped versions of the videos. We adopted a split set dividing the dataset into 103 training and 47 test samples as in [25].

The UCF11 dataset [24] contains 1646 videos in 11 action categories: biking/cycling, diving, golf swinging, soccer juggling, trampoline jumping, horse riding, basketball shooting, volleyball spiking, swinging, tennis swinging, and walking with a dog. Contains large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, and illumination conditions, some illustrated in Fig. 4. We adopted the original setup [24] using the leave-one-out cross-validation for a pre-defined set of 25 folds.

(a) (b) (c) (d)



(e) (f) (g) (h)

Fig. 3. Example of intra-class variability in UCF Sports [23]. Figures 3a and 3b are both examples from running class, 3c and 3d from swinging, 3e and 3f from kicking; while 3g and 3h from walking.

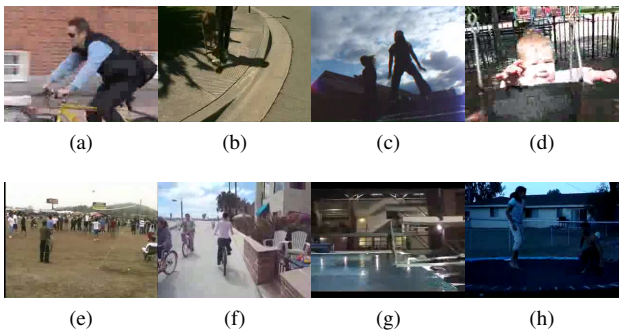


Fig. 4. Example of UCF11 [24] challenges, such as object appearance in 4a and 4b, viewpoint in 4c and 4d, cluttered background in 4e and 4f, and illumination conditions in 4g and 4h.

C. Experimental setup

Regarding the low-level descriptor, we used the same setup described in [19], thus each trajectory has a fixed length of 15 frames, with pixel sampling step of 5 pixels, a neighborhood size of 32 pixels and cell grid structure of 2×2 pixels. For low-level descriptors we have used HOG, HOF, MBHx and MBHy combined. In terms of dimensionality, the low-level descriptor have multiple sets of 426 dimensions.

The codebook creation is here performed by a k-means clustering algorithm, with Euclidean distance over one million randomly sampled descriptors and with number of clusters $k \in \{128, 512, 2048\}$.

For computing traditional BoW, we set just two parameters, the input codebook size, $M \in \{512, 2048\}$, and the number of codewords the feature points were assigned, $n = 10$. With regard to BNDTV parameters: $M \in \{512, 2048\}$, $n = 10$, the region bounds α_{min} and α_{max} were equal to 0.4 and 2.0 respectively, and the number of bins $B \in \{1, 2, 4\}$. The expansion controlling factor for both is $\gamma = 0.5$.

For classification we used a non-linear SVM with an RBF kernel [26]. For adjusting RBF parameters, C and γ, we performed a grid search using the validation set for KTH, one video per class, randomly selected and removed from the training set of UCF Sports and five videos per class, randomly selected and removed from the training set of UCF 11. The reported results are the average of ten executions of the classification, using the split protocol proposed for each dataset.

D. Experimental analysis

The performances, in terms of classification rates, for the traditional BoW and BNDTV are presented in Table I and in Table II, respectively. In both cases, the mid-level representation are applied to three datasets and by using the traditional pooling function, labeled as α, and the proposed pooling function, labeled as α'. As we can see, in boldface, the proposed pooling function increases recognition in almost all parameters variation, for UCF Sports and UCF 11. For KTH, there is no gain with the proposed approach, but the recognition rate remains very close.

TABLE I
THE CLASSIFICATION RATES FOR TRADITIONAL BOW.

Parameters	Approach	KTH	UCF Sports	UCF 11
M=512	BoW with α	84,7%	48,9%	52,9%
	BoW with α'	**89,8%**	**68,1%**	**69,5%**
M=2048	BoW with α	85.7%	55.3%	53,9%
	BoW with α'	**95,8%**	**80,9%**	**75,3%**

TABLE II
THE CLASSIFICATION RATES FOR BNDTV [7].

Approach	Parameters	KTH	UCF Sports	UCF 11
M=512 B=1	BNDTV with α	**95.4%**	66.0%	81.2%
	BNDTV with α'	89,8%	**68,1%**	69,5%
M=512 B=2	BNDTV with α	**94.9%**	66.0%	78.4%
	BNDTV with α'	90.3%	**70.2%**	**84.4%**
M=512 B=4	BNDTV with α	**96.3%**	**66.0%**	81.0%
	BNDTV with α'	91.2%	59.6%	**81.3%**
M=2048 B=1	BNDTV with α	**97.7%**	68.1%	73.7%
	BNDTV with α'	95,8%	**80,9%**	**75,3%**
M=2048 B=2	BNDTV with α	**97.7%**	70.2%	78.0%
	BNDTV with α'	96.3%	**80.9%**	**86.2%**
M=2048 B=4	BNDTV with α	**97.7%**	70.2%	75.7%
	BNDTV with α'	96.3%	**80.9%**	**85.7%**

Interesting to notice that the traditional BoW with α' results are identical to BNDTV with α' at B=1, indicating that all locality factors in BNDTV pooling function, the bounding values and bins, are subdue in the final representation, given rise to the relevance of standard deviation neighborhood information for error quantification.

According to the Table III, it is clear the significant improvement achieved by the proposed pooling, regarding the top recognition rate of the compared methods. Specially in UCF Sports dataset, in which the recognition rate is 25,6% higher for traditional BoW with α' and 10.7% for BNDTV with α'. Nonetheless, for UCF 11, the improvement are 21.4% and 5%, in traditional BoW with α' and BNDTV with α', respectively, when compared to the pooling function proposed in [5].

Furthermore, in Table III, it is presented a comparison with state-of-the-art approaches. Our results can be directly compared to DT-MB [27] and ISA [20], regarding classification protocols. It is also possible to directly compare our results to Dense Trajectories [19], except for UCF Sports, since they used flipped versions of the videos as an extension of the dataset and a different classification protocol. TMAR, for the best of our knowledge, holds the state-of-the-art recognition rate for all three datasets. Their results are reported using leave-one-out cross validation classification protocol. In this work we avoided this protocol, since there is many visually similar videos within the same class in UCF datasets.

V. CONCLUSION

In this work, we addressed the study of more discriminative mid-level representations in Bag-of-Words model applied to human action classification task. For that, we proposed a new pooling function, which amplifies values of feature points close to codewords border regions and weights the impact that feature points will assume in the final representation. Results indicate the relevance

TABLE III
COMPARISION WITH THE STATE-OF-THE-ART.

Framework	Approach	KTH	UCF Sports	UCF 11
Learning-based	TMAR [21]	**98.0%**	**95.1%**	**93.2%**
	ISA [20]	93.9%	-	75.8%
No learning-based	Dense trajectories [19]	94.2%	**88.2%**	84.1%
	DT-MB [27]	95.6%	-	**86.6%**
	BoW with α	85.7%	55.3%	53,9%
	BoW with α'	95.8%	**80.9%**	75.3%
	BNDTV with α	**97.7%**	70.2%	81.2%
	BNDTV with α'	**96.3%**	**80.9%**	**86.2%**

of codeword standard deviation neighborhood in a mid-level representation. We evaluated our approach by using two mid-level representation, vector quantization with soft-assignment, here called traditional BoW for simplicity and BNDTV. The representations are applied in human action classification task, using three well-known datasets, KTH, UCF Sports, and UCF 11. In both mid-level representations, we showed that thanks to the new pooling function, the classification rates have increased in almost all cases for UCF Sports and UCF 11. For KTH, the results for traditional BoW using the proposed pooling function are better than the original one, but the inverse occurs for BNDTV. In terms of values, our strategy has improved the classification rates more than 25.6% with respect to the original pooling function used in traditional BoW and 10.7% in BNDTV for UCF Sports. For UCF 11, the improvement are 21.4% and 5%, in traditional BoW and BNDTV by using the proposed pooling function. For further works, we will explore the new pooling function in other applications and with other mid-level representations.

ACKNOWLEDGMENT

The authors are grateful to FAPEMIG (PPM-00006-16), CNPq (Grant 421521/2016-3), PUC Minas and CAPES for the financial support to this work.

REFERENCES

[1] Y. L. Boureau, F. Bach, Y. LeCun, and J. Ponce, "Learning mid-level features for recognition," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 2010, pp. 2559–2566.

[2] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: multi-way local pooling for image recognition," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2651–2658.

[3] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

[4] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *International Conference on Computer Vision (ICCV)*. Nice, France: IEEE, 2003, pp. 1470–1477.

[5] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 7, pp. 1271–1283, Jul. 2010.

[6] S. Avila, N. Thome, M. Cord, E. Valle, and A. d. A. AraúJo, "Pooling in image representation: The visual codeword point of view," *Computer Vision and Image Understanding*, vol. 117, no. 5, pp. 453–465, 2013.

[7] R. Almeida, Z. K. G. d. Patrocínio Jr., and S. J. F. Guimarães, "Exploring quantization error to improve human action classification," in *Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN 2017)*, Anchorage, Alaska, 2017, pp. 1354–1360.

[8] Q. Wang, X. Deng, P. Li, and L. Zhang, "Ask the dictionary: Soft-assignment location-orientation pooling for image classification," in *Image Processing (ICIP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4570–4574.

[9] A. M. Bruckstein, D. L. Donoho, and M. Elad, "From sparse solutions of systems of equations to sparse modeling of signals and images," *SIAM review*, vol. 51, no. 1, pp. 34–81, 2009.

[10] H. Lee, A. Battle, R. Raina, and A. Y. Ng, "Efficient sparse coding algorithms," *Advances in neural information processing systems*, vol. 19, p. 801, 2007.

[11] K. Yu, T. Zhang, and Y. Gong, "Nonlinear learning using local coordinate coding," in *Advances in neural information processing systems*, 2009, pp. 2223–2231.

[12] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality-constrained linear coding for image classification," in *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*. IEEE, 2010, pp. 3360–3367.

[13] H. Phan, P. Koch, L. Hertel, M. Maass, R. Mazur, and A. Mertins, "Cnn-lte: a class of 1-x pooling convolutional neural networks on label tree embeddings for audio scene classification," in *Proc. ICASSP*, 2017.

[14] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the fisher kernel for large-scale image classification," *Computer Vision–ECCV 2010*, pp. 143–156, 2010.

[15] H. Jegou, F. Perronnin, M. Douze, J. Sánchez, P. Perez, and C. Schmid, "Aggregating local image descriptors into compact codes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 34, no. 9, pp. 1704–1716, 2012.

[16] N. Murray and F. Perronnin, "Generalized max pooling," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2473–2480.

[17] P. Li, H. Zeng, Q. Wang, S. C. Shiu, and L. Zhang, "High-order local pooling and encoding gaussians over a dictionary of gaussians," *IEEE Transactions on Image Processing*, vol. 26, no. 7, pp. 3372–3384, 2017.

[18] J. Carreira, R. Caseiro, J. Batista, and C. Sminchisescu, "Free-form region description with second-order pooling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1177–1189, 2015.

[19] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Colorado, USA: IEEE, 2011, pp. 3169–3176.

[20] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *Computer Vision and Pattern Recognition (CVPR)*. Colorado, USA: IEEE, 2011, pp. 3361–3368.

[21] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris, "Matching mixtures of curves for human action recognition," *Computer Vision and Image Understanding*, vol. 119, pp. 27–40, 2014.

[22] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *17th International Conference on Pattern Recognition (ICPR)*, vol. 3. Washington, DC, USA: IEEE Computer Society, 2004, pp. 32–36.

[23] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action mach a spatio-temporal maximum average correlation height filter for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Anchorage, Alaska: IEEE, 2008, pp. 1–8.

[24] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos," in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*. Miami, FL, USA: IEEE, 2009, pp. 1996–2003.

[25] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in *International Conference on Computer Vision (ICCV)*. Barcelone, Spain: IEEE, 2011, pp. 2003–2010.

[26] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 21–27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[27] X. Peng, Y. Qiao, Q. Peng, and X. Qi, "Exploring motion boundary based sampling and spatial-temporal context descriptors for action recognition," in *24th British Machine Vision Conference (BMVC)*. Bristol, UK: BMVA Press, 2013, pp. 1–11.