Exploring Feature Distribution to Create Mid-level Representations: A Case Study in Human Action Recognition

Raquel Almeida, Zenilton K. G. 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)
raquel1908@gmail.com, {zenilton,sjamil}@pucminas.br

***Abstract*—Data representation is a critical task in many areas of computational studies, particularly in the case of visual data representation, in which subtleties can undermine the perception and interpretation of the visual content. In this study, it is proposed strategies to exploit visual mid-level representations, aiming to transform the detailed description extracted directly from the visual media into a simplified and discriminative representation. More specifically, the proposed strategies are delineated in Bag-of-Words mid-level representation model and are used to aggregate distribution information within partitions and regions of interest on feature space. Experiments on three well-known public datasets, namely, KTH, UCF Sports and UCF 11, demonstrated that feature points spatial distribution information is useful to create more discriminative representations. All three proposed representations were published and outperform, in terms of recognition rate, conventional strategies on BoW model and are, in many cases, superior or comparable with the state-of-the-art.**[1]

I. Introduction

Data representation is a critical task in many areas of computational studies. Through an appropriate representation it is possible to incorporate desirable characteristics of interest for an application. In the case of visual data representation, a careful conception is required. The visual information is usually subtle and can undermine the perception and interpretation of the visual content. In this study, it is proposed strategies to create visual representations by aggregating information of feature points distribution within regions of interest on the multidimensional feature space.

The proposed strategies are delineated in the Bag-of-Words (BoW) mid-level representation model. This model is a popular method to map a set of local descriptors, extracted directly from the media, into a global representation. The output is a single feature vector, representing histograms of feature frequency distribution over learned data prototypes, called codewords. The mid-level approach provided great advances in visual data representation, supported by the very discriminant local descriptors and the output compact form, which is more suitable for classifier use.

[1]Work presented as a MSc dissertation at PUC-Minas

BoW is a notorious method, for the simplicity in concept and implementation, and also by the achieved results in many applications. Nevertheless, limitations are imposed by the discrete analysis of visual data and it is commonly associated with some issues which could lead to improper representation, such as: (i) presence of noise information captured during the feature extraction; (ii) possible selection of irrelevant data; (iii) semantic relation degradation between elements; (iv) manipulation and interpretation of high-dimensional elements; (v) and quantization errors usually ignored during the codification.

From the premise that functions which combine features over spatial neighborhoods could achieve invariance and robustness in representations [1], it is argued that spatial distribution can provide useful information for representations. Three hypothesis are raised, regarding feature distribution: (i) data feature distribution can provide clues to create a concise visual data representation; (ii) spatial restrictions can make representations more discriminative; and (iii) the frontier of bounded regions can provide additional knowledge for creating mid-level representations. The main goal is to create more discriminative representations, reasoning that, with the feature distribution knowledge, it is possible to filter out some noise information, impute relevance into distinct features and establish some order relation.

II. Human action recognition task in literature

Among the many pattern recognition tasks which could benefit from a discriminative mid-level representation, the proposed strategies are applied into the Human Action Recognition task. Human action recognition is an open problem, which gained a lot of attention recently due many possible real life applications. To create a comprehensive overview of the task, in this section it is presented a brief review, focusing on publications that address the task of human action recognition using the BoW model.

To delineate the core of the BoW model, one should point out Vector Quantization (VQ) [2] as the main strategy and the improvements obtained by a soft-probabilistic approach called Soft-Assignment (SA) [3]. The most pronounced action

descriptors use local spatio-temporal features aggregated using VQ or SA [4]–[6].

Methods applied in action recognition, which explore spatial distribution in BoW model, could be grouped in the following categories:

- **Weighted methods**: which aim to preserve spatial temporal relations taking into account multiple weighted representation [7], [8];
- **Hierarchical methods**: which create histograms based on multiple hierarchical levels [9]–[11];
- **Combination methods**: which combine local histograms to create the video representation [12], [13];
- **Contextual methods**: which create a contextual spatial temporal domain driven by the histogram information [14], [15];
- **High-order methods**: which use spatial-temporal statistics to create a spatial context [16]–[19].

Recently, there is a growing trend of feature learning-based methods, and they usually present the state-of-the-art on the task. The growing trend of learning-based methods is followed by the use of larger datasets, commonly required for the learning step. In contrast, feature distribution analysis regards only the regions of interest of a determined data. Consequently, most of strategies which use a feature learning framework could not be directly compared with the proposed methods. For completeness and comparison, it is here presented four strategies assessed in the same datasets used in this work, namely: (i) ISA [20], which performs an independent subspace analysis to learn spatial-temporal features from unlabeled data; (ii) Action bank [21], which creates a high-level action representation using a bank of action detectors; (iii) TMAR [22], which describes actions using optical flow motion features clustered by Gaussian mixture model; and (iv) Stream learning [23], which uses a deep learning architecture to create feature models in a streamed learning framework.

III. Proposed strategies using local feature distribution

Mid-level representations have three steps in common: (i) coding; (ii) pooling; and (iii) concatenation. In Bag-of-Words (BoW) mid-level representation model at each step it is used an auxiliary structure learned from the data, called codebook. The coding step uses a function to relate the features within the regions of interest with the codewords composing the codebook. The pooling step uses a function to aggregated the codified values as histograms at each codeword. And the final step, the concatenation, insert the histograms side by side to create the final representation.

Formally, let $\mathbb{X}$ be an unordered set of local descriptors and $\mathbb{Z}$ the final representation. The steps to create the mid-level representation are defined as follows:

$$\mathbb{X} \in \mathbb{R}^N \mapsto \mathbb{Z} \in \mathbb{R}^M$$

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

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

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

in which:

- $\mathbf{x}_j$ is a d-dimensional descriptor extracted from the data.
- N is the number of identified regions.
- M is the number of codewords.

The functions commonly used in traditional BoW methods, tend to average out locality information, particularly during the pooling step.

From the premise that spatial distribution can provide useful information to create a concise visual data representation, it is proposed strategies to design regions of interest in feature space and study the feature points distribution considering these regions. It aims to incorporate locality information in the final representation. It is proposed three strategies, outlined as follows:

A) **Pooling over linear local distance distribution of feature space:** For this approach, it is proposed to study distance-to-codeword histograms based on equally linear subdivisions of codewords neighboring space. The goal is to establish a relation between locality pooling constrains and recognition rates.

B) **Pooling over volumetric distribution and partition of feature space:** For this approach, it is proposed to study distance-to-codeword histograms based on equally volumetric partitions of codewords neighboring space. It is designed to maintain the same probability of assignment to a given hyper-region creating volumetric partitions of a hypersphere centered at each codeword.

C) **Weighed distribution based on bounding regions frontiers:** For this approach it is proposed to use the frontier of codewords as a reference value to measure attribution discrepancy and weight the impact that feature points will assume in the final representation. This representation takes into account the spatial distribution to amplify the contribution of points close to the border of the regions in a high-dimensional space.

In the following it is presented the main concepts and formulation for each proposed representation.

A. Linear feature space partition and distance distribution

The pooling function proposed for this approach creates histograms based on equally linear subdivisions of codewords neighboring region. The proposed video descriptor is called **B**ossa **N**ova **D**irectly **T**o **V**ideo (BNDTV), due the extension of the BossaNova [24] image representation scheme to represent video features. Although BossaNova was originally portrait as an image mid-level descriptor, it could be easily derived as a video descriptor, due the many similarities in the process of describing images and videos. Therefore, BossaNova and BNDTV are similar in form, but differ in nature.

BNDTV uses a density-based pooling strategy, which determine the pooling region distance radius by a factor of the codeword standard deviation. This radius is used to restrict the distances range and bound the codeword neighboring region by a certain number of bins. The number of bins is used to

TABLE I
SUMMARIZED DEFINITIONS OF BNDTV AND BOH REPRESENTATIONS

Aspects	BNDTV	BOH
Regions of interest	Codeword neighboring region bounded by B bins	E equally probable hyper-regions in codeword neighboring region
Bounding	Factor λ of normal distribution: $\alpha_m^{lim} = \lambda^{lim}.\ \sigma_m$	Radius of largest hypersphere: $r_E^{\mathbf{c}_m} = \sigma_m$
Pooling regions	Linear partition $r_b, \forall b \in [1, B]$ for $\frac{b}{B} \geq \alpha_m^{min}$ and $\frac{b+1}{B} \leq \alpha_m^{max}$	Volumetric partition $r_e, \forall e \in [1, E]$ for $r_e = r_1 \times \sqrt[d]{e}$
Range of distances	$\mathbb{B}_{range} \doteq \left[\frac{b}{B}; \frac{b+1}{B}\right]$	$\mathbb{E}_{range} \doteq \left[r_E^{\mathbf{c}_m} \sqrt[d]{\frac{e}{E}}, r_E^{\mathbf{c}_m} \sqrt[d]{\frac{e+1}{E}}\right]$
Pooling function	Probability density function: $h_{m,b} = \mathrm{card}(\mathbf{x}_j \mid \alpha_{m,j} \in \mathbb{B}_{range})$	Sum of values inside hyper-region: $h_{m,e} = \mathrm{sum}\left(\alpha_{m,j} \mid \mathrm{D}(\mathbf{x}_j, \mathbf{c}_m) \in \mathbb{E}_{range}\right)$
Final form	$\mathbf{z}_{bndtv} = \left[\left[z_{m,b}\right], t_m\right]^{\mathrm{T}}$ $(m, b) \in \{1, ..., M\} \times \{1, ..., B\}$	$\mathbf{z}_{boh} = [h_{m,e}]^{\mathrm{T}}$ $(m, e) \in \{1, ..., M\} \times \{1, ..., E\}$
Final size	$M \times (B + 1)$	$M \times E$
Representation form	Normal distribution	Sparse

Consider the following notations: $\mathbf{c}_m$ a codeword, σ_m a codeword standard deviation and $\mathrm{D}(\cdot)$ Euclidean distance

equally divide the radius and to quantify the low-level feature points according to their distance-to-codeword and the bins they fall. BNDTV deals with the plausibility of a codeword [3], meaning that, if a codeword is among the closest codewords of a feature point, but is not close enough to be relevant, it will be thrown out by the pooling region restriction.

B. Volumetric distribution and feature space partition

The volumeric strategy is called **B**ag **O**f local distribution of descriptors on concentric **H**yperspheres (BOH). The BOH pooling function is designed to create equally probable hyper-regions, aiming to maintain the same probability of assignment to a given hyper-region. Within the context of data distribution, two hyper-regions are considered equally probable if they have the same volume. The pooling regions are determined by placing concentric hyperspheres at each codeword, enclosing hyper-regions with the same volume. The radius of the largest hypersphere is the codeword mean distribution value and the histogram of distances takes into account the feature point position amongst the hyper-regions. In this way, feature points attributed to one hyper-region have similar distances to the codeword, which includes a locality context during pooling.

The BOH representation is more sparse, as the number of hyper-regions or codewords increases, but it approximates better the actual distribution of distances. Due the high-dimensional feature space, the representation codifies more feature points close to the largest hypersphere limit, leading to most points falling in the same hyper-region. This creates a representation more tight to the boundary region than to the number of internal partitions of the pooling region.

In Table I it is presented the formal concepts for both BNDTV and BOH representations and their main characteristics in terms of representation form.

C. Weighed distribution based on bounding regions frontiers

For this strategy, it is proposed to use the frontier of codewords as a reference value to measure attribution discrepancy and weight the impact a feature point will assume in the final representation. This approach takes into account the spatial distribution to amplify the contribution of points close to the border of regions in a high-dimensional space. As stated before, codewords are data prototypes, usually learned by clustering methods that divide feature space into regions to determine the clusters centers. In these clustering methods, the standard deviation associated with each cluster center, carries an important information about data distribution.

From this, it is proposed a coding function which weights the contribution of feature points in the final representation based on a relative value considering the codeword standard deviation as pooling region frontier, as follows:

$$\alpha'_{m,j} = \exp\left(\gamma \frac{T(\mathbf{x}_j)}{W(\mathbf{c}_m, \mathbf{x}_j)}\right) \tag{1}$$

in which:

- γ is an expansion controlling factor.
- $T(\mathbf{x}_j) = \sum_{k=1}^{M} (\mathrm{D}(\mathbf{c}_k, \mathbf{x}_j))^2$ is the total assignment error between the feature point and the codebook.
- $W(\mathbf{c}_m, \mathbf{x}_j) = \left(\mathrm{D}(\mathbf{c}_m, \mathbf{x}_j)^2 - \sigma_m\right)^2$ is the weight of a given feature point based on its distance from the assigned codeword and from the codeword standard deviation.

In a typical approach, codeword relevance is determined by the distribution of a probability mass [3]. In the proposed function, the relevance is modeled by weighted contributions

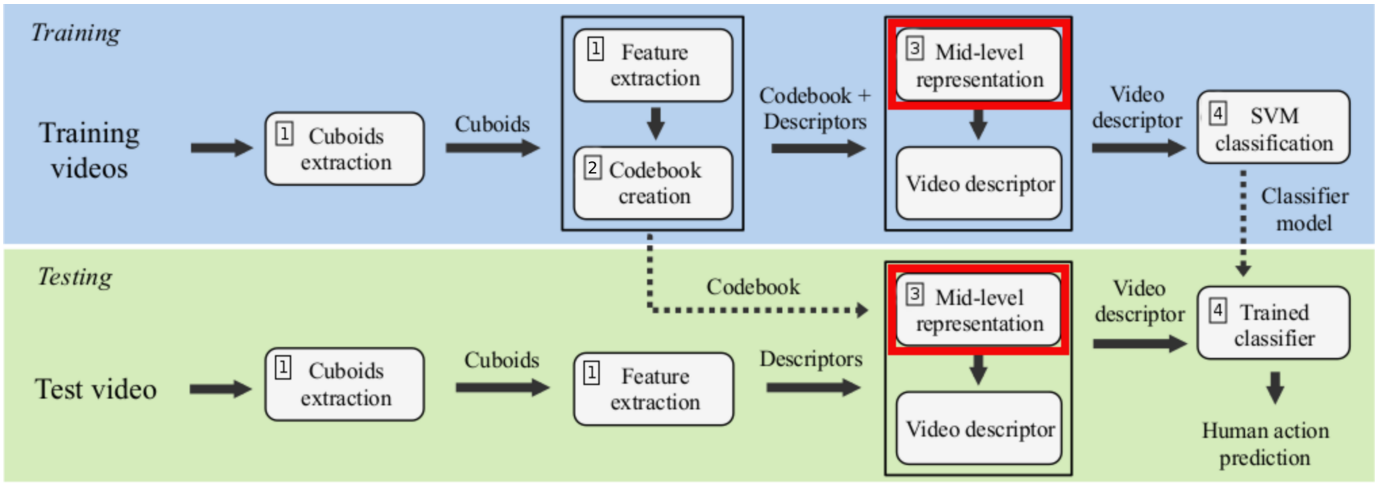

Fig. 1. Proposed Framework for Action Recognition

in the final representation, enhancing values of feature points closer to the frontier of the region and penalizing the ones more distant. Two important considerations are raised from the proposed function: (i) the value codified for each codeword is not leveraged by the number of regions, which avoid the flatten effect in the representation; and (ii) feature points which are assigned to larger regions or are outside the border, will be considered relevant for the representation if close to the border, despite their distance from the codeword.

IV. Human action recognition framework

The framework, illustrated in Figure 1 is presented in the context of human action recognition task and can be outlined in four major steps: (i) local description using the content of input videos; (ii) codebook creation using local descriptors; (iii) mid-level representation applying the proposed strategies; and (iv) classification using the final representation to learn statistical models. Each step is detailed in the following:

1. Low-level description: The video description process can be divided into cuboid and feature extraction. For this study, it is proposed to use Dense Trajectories (DT) as low-level description [5]. DT creates dense and high-dimensional descriptors, which achieve 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), histogram of oriented optical flow (HOF) and motion boundary histogram (MBH) within the trajectories.
2. Codebook creation: A visual codebook must be created before the encoding, and for the proposed strategies it should be able to provide information about data distribution. The k-means clustering algorithm is simple and yet well-suited for this purpose. The algorithm is performed using sampled trajectories descriptors. The cluster centers identified during k-means execution will be addressed as codewords for the mid-level step. The data distribution information is retrieved by storing the standard deviation associated with each cluster during the codebook creation.
3. Mid-level representation: This is target step of this study, in which each proposed strategy is applied. In BNDTV and BOH, the representations are created using their respective pooling functions to codify features based on their spatial distribution. As for the weighted function, the representations are created with an additional step to weight and enhance the features close to the codeword border that is defined by the standard deviation.
4. Classification: The final step in the framework is the classification per se, in which the recognition rate could indicate the discrimination properties of the representations. This step is proposed to be performed by non-linear support vector machine (SVM) classifiers using a Radius Basis Function (RBF) kernel, mainly to keep a fair comparison between the tests results and most of reported methods in literature. The classification can be divided in training and testing phase. In testing phase, a new test video is classified by applying the classifier model obtained during the training phase. The selection of videos for training and test, along with the classification protocol, follows the best practice according to each dataset.

V. Experiments on human action recognition task

The assessments are grouped by each representation, namely: (i) BOW (baseline), built on the main strategy in BoW model, called Soft-Assignment; (ii) BNDTV; (iii) BOH; and (iv) the feature distribution weighted function applied on each representation strategy, here referenced as BOW', BNDTV' and BOH'. For completeness, it is also presented a comparison with the state-of-the-art methods applied on the same datasets.

The results are evaluated using two classification protocols: (i) split, following the original setup proposed for each dataset, more susceptible to bias but widely used in literature; and (ii) leave-one-out cross-validation (loocv), for model generalization. The reported results are the mean value of 10 executions.

TABLE II
RECOGNITION RATE RESULTS AND COMPARISON WITH THE STATE-OF-THE-ART (%). BEST VALUE PER GROUP EMPHASIZED

	Method	KTH		UCF Sports		UCF 11	
		Split	Loocv	Split	Loocv	Split	Loocv
No-Learning	Liu, Luo and Shah [25]	–	93.8	–	–	65.4	–
	Raptis and Soatto [16]	–	94.5	–	–	–	–
	Wang et al. [5]	94.2	–	–	85.6	65.4	–
	Wang et al. [6]	95.3	–	–	**89.1**	85.4	–
	BOW (baseline)	85.7	55.3	58.8	80.1	55.0	53.8
	BOW'	95.8	–	80.9	–	75.3	–
	BNDTV	**97.7**	**97.7**	70.2	81.3	81.3	**90.0**
	BNDTV'	97.2	–	80.9	–	**86.6**	–
	BOH	97.2	97.5	**83.0**	82.7	80.8	89.1
	BOH'	96.3	–	68.0	–	79.0	–
Learning	Le et al. [20]	93.9	–	–	–	**75.8**	–
	Vrigkas et al. [22]	–	**98.3**	–	**95.1**	–	**93.2**
	Hasan and Roy-Chowdhury [23]	98.0	–	–	–	–	–
	Sadanand and Corso [21]	**98.2**	–	–	95.0	–	–
	Lan, Wang and Mori [26]	–	–	**73.1**	–	–	–

The proposed methods are tested in three well-known datasets, namely: (i) KTH [27], 600 videos and 6 classes; (ii) UCF Sports [28], 150 videos and 10 classes; and (iii) UCF 11 [25], 1646 videos and 11 classes. The dataset selection were determined due distinctive aspects that could lead to improper representation and miss classification, such as: (i) size; (ii) colorspace; (iii) video duration and resolution; (iv) intraclass variability; and (v) presence of noise scene elements.

In Table II a comparison, in terms of recognition rate, is presented for the methods and the state-of-the-art. The results are grouped into two different blocks of methods: (i) No-Learning, which there is no learning process other than the clustering and classification; and (ii) Learning, for feature learning methods.

A. Results analysis

In terms of time performance evaluation, the main time consuming operation is the distance calculation between feature points and codewords, thereby, it is the same for all methods.

The recognition rate achieved by the baseline BOW is quite poor in UCF Sports and UCF 11, despite the protocol, although presenting slightly better values in KTH. For the proposed, in KTH comparing with: (i) No-Learning, all proposed methods overcome the highest rate; and (ii) Learning methods, both BNDTV and BOH are comparable in both protocols (less than 1% below the top). In UCF Sports, compared in: (i) split, BOH presents the best recognition rate; and (ii) loocv, all proposed methods are inferior. Unfortunately, this scenario is a consequence of the choice to not use the extended version of the dataset to avoid biases in classification. In UCF 11, compared in: (i) split, BNDTV' is slighted superior to the state-of-the-art; and (ii) loocv, both BNDTV and BOH are comparable with the state-of-the-art.

The BNDTV approach creates equally divided partitions of pooling regions. A quantitative analysis of the results places this representation comparable with the state-of-the-art for KTH dataset, and among No-Learning methods presents the top recognition rate. A qualitative analyses indicates that this representation is more susceptible to fail when subtle movements are presented in the input video.

The BOH strategy is designed to create equally probable hyper-regions. Although the top recognition rate achieved by this approach are very similar to BNDTV, the volumetric approach presented a much desirable behavior as monotonic functions in terms of codebook size and number of hyper-regions. This representation is the state-of-the-art for UCF Sports in split protocol. A qualitative analyses indicates that this representation is more susceptible to fail when the input video presents a human interaction with objects in the scene and with complex movements composed by separable multiple actions.

The last strategy studied the frontier of bounding regions as reference value to weight feature contribution in the final representation. Experiments using the proposed function for this strategy considerably increase the recognition rate of representations unaware of feature spatial distribution. As one could see from Table II, it is clear the significant improvement achieved with BOW' representations for all datasets. BNDTV' in UCF Sports, shown a considerable superior performance, in terms of recognition rate using split protocol.

VI. CONCLUSIONS

This study tackled the problem of creating discriminative mid-level representations. The proposed methods were used to create video representations and applied in human action recognition task. It was demonstrated that strategies, which use the spatial distribution of feature points can deal with some attribution errors and the relative position of the features can create more discriminative representations. In addition, it was showed that the consideration of the boundary regions of codewords as regions of interest could reduce the noise information captured and impute relevance to more significant features in the final representation.

This research opens novel opportunities for study, such as: (i) investigation of a path for generalization, searching for an optimal partition of feature space; (ii) test the representation

compact code as input of a feature learning-based system; (iii) study feature distribution using high-order statistics; and (iv) use hierarchical data structures for distribution representations.

All proposed methods were published in distinguished conferences in pattern recognition and machine learning, as listed in the following:

- **Linear feature space partition and local distance distribution (BNDTV)** Published as *Exploring quantization error to improve human action classification*, in the 2017 International Joint Conference on Neural Networks (IJCNN) [29].
- **Volumetric distribution and feature space partition (BOH)** Published as *Human action classification using an extended BOW formalism*, in the 19th Biennially International Conference on Image Analysis and Processing (ICIAP) [30]. This publication was a collaborative work with professor Benjamin Bustos of University of Chile.
- **Weighed distribution based on bounding regions frontiers** Published as *A New Pooling Strategy based on Local Feature Distribution: A Case Study for Human Action Classification*, in the 30th Conference on Graphics, Patterns and Images (SIBGRAPI) [31].

Acknowledgements

The authors are grateful to FAPEMIG (PPM 00006-16), CNPq (Universal 421521/2016-3 and PQ 307062/2016-3), CAPES (MAXIMUM STIC-AmSUD 048/14) and PUC Minas for the financial support to this work.

We would also like to show our gratitude to professor Benjamin Bustos from University of Chile for the collaborative work.

References

[1] Y.-L. Boureau, N. Le Roux, F. Bach, J. Ponce, and Y. LeCun, "Ask the locals: multi-way local pooling for image recognition," in *International Conference on Computer Vision, 13.* IEEE, 2011.

[2] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *International Conference on Computer Vision, 9.* IEEE, 2003.

[3] J. C. van Gemert, C. J. Veenman, A. W. M. Smeulders, and J.-M. Geusebroek, "Visual word ambiguity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2010.

[4] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2008.

[5] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Action recognition by dense trajectories," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2011.

[6] H. Wang, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, 2013.

[7] T. Yi and R. Qiuqi, "Weight and context method for action recognition using histogram intersection," in *International Conference on Wireless, Mobile and Multimedia Networks, 5.* IET, 2013.

[8] Q. Wang, X. Deng, P. Li, and L. Zhang, "Ask the dictionary: Soft-assignment location-orientation pooling for image classification," in *International Conference on Image Processing, 22.* IEEE, 2015.

[9] J. Wu, D. Zhou, and G. Xiao, "A hierarchical bag-of-words model based on local space-time features for human action recognition," in *International Conference on IT Convergence and Security, 3.* IEEE, 2013.

[10] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff, "Action recognition and localization by hierarchical space-time segments," in *International Conference on Computer Vision*. IEEE, 2013.

[11] K. J. de Souza, A. de Albuquerque Araújo, Z. K. G. do Patrocínio Jr, J. Cousty, L. Najman, Y. Kenmochi, and S. J. F. Guimarães, "Decreasing the number of features for improving human action classification," in *Conference on Graphics, Patterns and Images, 29.* SBC, 2016.

[12] X. Yan and Y. Luo, "Making full use of spatial-temporal interest points: an adaboost approach for action recognition," in *International Conference on Image Processing, 17.* IEEE, 2010.

[13] Y. Guo, W. Ma, L. Duan, Q. En, and J. Chen, "Human action recognition based on discriminative supervoxels," in *International Joint Conference on Neural Networks, 29.* IEEE, 2016.

[14] Z. Zhang, C. Wang, B. Xiao, W. Zhou, and S. Liu, "Action recognition using context-constrained linear coding," *Signal Processing Letters*, 2012.

[15] H. Luo and H. Lu, "Multi-level sparse coding for human action recognition," in *International Conference on Intelligent Human-Machine Systems and Cybernetics, 8.* IEEE, 2016.

[16] M. Raptis and S. Soatto, "Tracklet descriptors for action modeling and video analysis," in *European conference on computer vision, 11.* Springer, 2010.

[17] N. Murray and F. Perronnin, "Generalized max pooling," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2014.

[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*, 2015.

[19] 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*, 2017.

[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*. IEEE, 2011.

[21] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition*. IEEE, 2012.

[22] M. Vrigkas, V. Karavasilis, C. Nikou, and I. A. Kakadiaris, "Matching mixtures of curves for human action recognition," *Computer Vision and Image Understanding*, 2014.

[23] M. Hasan and A. K. Roy-Chowdhury, "A continuous learning framework for activity recognition using deep hybrid feature models," *IEEE Transactions on Multimedia*, 2015.

[24] 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*, 2013.

[25] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos," in *Computer Society Conference on Computer Vision and Pattern Recognition*. IEEE, 2009.

[26] T. Lan, Y. Wang, and G. Mori, "Discriminative figure-centric models for joint action localization and recognition," in *International Conference on Computer Vision, 13.* IEEE, 2011.

[27] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *International Conference on Pattern Recognition, 17.* IEEE, 2004.

[28] M. D. Rodriguez, J. Ahmed, and M. Shah, "Action MACH a spatio-temporal maximum average correlation height filter for action recognition," in *Conference on Computer Vision and Pattern Recognition*. IEEE, 2008.

[29] R. Almeida, Z. K. G. d. Patrocínio, and S. J. F. Guimarães, "Exploring quantization error to improve human action classification," in *International Joint Conference on Neural Networks, 30.* IEEE, 2017.

[30] R. Almeida, B. Bustos, Z. K. G. d. Patrocínio Jr., and S. J. F. Guimarães, "Human action classification using an extended BOW formalism," in *Biennially International Conference on Image Analysis and Processing, 19.* Springer, 2017.

[31] R. Almeida, Z. K. G. d. Patrocínio, and S. J. F. Guimarães, "A new pooling strategy based on local feature distribution: A case study for human action classification," in *Conference on Graphics, Patterns and Images, 30.* Elsevier, 2017.