Human Action Classification Using an Extended BoW Formalism

Raquel Almeida[1(✉)], Benjamin Bustos[2], Zenilton Kleber G. do Patrocínio Jr.[1], and Silvio Jamil F. Guimarães[1]

[1] Audio-Visual Information Laboratory (VIPLAB), Department of Computer Science, Pontifical Catholic University of Minas Gerais (PUC Minas), Belo Horizonte, Brazil
raquel.almeida.685026@sga.pucminas.br, {zenilton,sjamil}@pucminas.br
[2] Department of Computer Science, University of Chile, Santiago, Chile
bebustos@dcc.uchile.cl

Abstract. In human action classification task, a video must be classified into a pre-determined class. To cope with this problem, we propose a mid-level representation which extends the Bag-of-Words formalism in order to better described the low-level features, exploring distance-to-codeword histograms. The main contribution of this article is the assembly of low-level features by a mid-level representation enriched with information about distances between descriptors and codewords. The proposed representation takes into account volumes of hyper-regions obtained from hyperspheres centered at codewords. Experimental results demonstrated that our strategy either has improved the classification rates more than 6% with respect to the compared mid-level representation for UCF Sports, or it is a competitive one, for KTH and UCF-11.

Keywords: Human action classification · Mid-level representation · Pooling strategy

1 Introduction

Human Action Classification is a pattern recognition task in which the main goal is to identify the action displayed in a media content. Regardless of the media source, such as images, sequential selection of video frames, raw video data or annotations, this task gained a lot of attention in the past few years. Here the focus is on video data, and one could define the main task as: given a video, one needs to classify the action displayed into one of a predetermined set of actions using only the content presented in the video.

To achieve this goal we explore ways to better represent the video content, transforming the original input information into a more suitable representation for the classifier. Usually researchers address this problem using two stages [9]:

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.

S. Battiato et al. (Eds.): ICIAP 2017, Part I, LNCS 10484, pp. 185–196, 2017.
https://doi.org/10.1007/978-3-319-68560-1_17

(i) feature extraction; and (ii) action classification. In a typical approach, feature extraction is performed directly on the raw data, here called low-level description, trying to avoid noise or irrelevant information. Action classification involves learning statistical models from the extracted features, and using those models to classify new feature observations.

The most discriminative low-level descriptors available in literature today rely on identifying regions of interest. Once these regions are identified, desirable features are then extracted around these regions. The output created by this process is a set of features, which are related to regions, representing the media. Facing this scenario, one popular approach is to map the set of local descriptors into one vector used as a global representation, so-called *mid-level representation*.

Among the methods for creating a mid-level representation, standout Bag-of-Words (BoW), Spatial Pyramids and Convolutional Networks for their notable results. As stated by [2], mid-level representations have three steps in common: (i) coding; (ii) pooling; and (iii) concatenation. Coding stands for the transformation locally applied into features vectors, extracting distribution characteristics. Pooling, in turn, explores the spatial relation between these characteristics; and concatenation constructs the final vector representation.

Here we explore a new strategy for the pooling step based on a volumetric partition of an hypersphere centered at codewords. The goal is to maintain the same probability of assignment to a given hyper-region. We argue that this kind of pooling could decrease the quantization error created during codification.

This paper is organized as follows. In Sect. 2, some related works involving mid-level representations and human action classification are described. While in Sect. 3, a formalization of traditional BoW is presented, in Sect. 4, the new mid-level representation is given. Experiments for human action recognition taking into account three well-known datasets are presented in Sect. 5, and finally, some conclusions are drawn in Sect. 6.

2 Related Work

Human action recognition is a popular topic in video processing, but it still an open problem due to the difficult in creating a representation able to capture and describe action motions in different scenarios. Among the most pronounced action descriptors, local spatio-temporal features, as proposed in [6,15], have been successfully used in several applications. In [6] the Space-Time Interest Point (STIP) descriptor are proposed. In STIP, interest points are detected in multiple scales and associated to a patch. Each path is described using Histogram of Oriented Gradients (HOG) and Histogram of Optical Flow (HOF). Regarding Dense Trajectories descriptor, proposed in [15], the trajectories are obtained by densely tracking sampled points obtained with optical flow fields. After tracking, feature point descriptors are extracted using HOG, HOF and Motion Boundary Histogram (MBH) around the trajectories.

These two description approaches represent action videos by a set of local features. Inspired by the success of mid-level representations of local features in

image processing, they rely on vector quantization in a BoW scheme to create a global video representation. Although they achieve good results, the loss of information during codification still an open problem.

In order to deal with quantification errors, a wide range of methodologies for creating mid-level representation has been proposed. Most of these representations, when applied in action recognition task, follow the vector quantization based on BoW model, but try to preserve spatial temporal relations during the coding process taking into account multiple weighted representation.

In [18], it was proposed the combination of local histograms with body regions histograms in order to preserve spatial temporal relations between interest points. In [8], it was proposed a sparse coding with max pooling framework applied in multiple contexts, which are defined according to the spatial scale and nearest neighbors of local features. Moreover, it is constructed one vocabulary and one histogram for each defined context. At the end, these data are concatenated for the final representation.

In [17], it was proposed a method based on multiple hierarchical levels for creating histograms. The first level is constructed using the descriptors extracted from video cuboids. The other levels are created by applying a neighboring function regarding previous level description and by creating a new codebook and new histogram for the current level. This scheme is called hierarchical BoW. In [13] an hierarchical BoW is constructed by recursive computing partitions of depth maps sequence in temporal domain, called Temporal Bag-of-Words.

In order to explore a representation driven by the histogram information, in [20], it was defined a contextual domain surrounding a spatial temporal area. After that, a contextual distance is calculated by adding a penalty value proportional to the probability density function computed from the local descriptors and codewords of the contextual domain. The contextual information is also used in [19], however it is obtained by histogram intersections, using both spatial and temporal distances as weighted controlling factors. In [16], Term Frequency-Inverse Document Frequency (TF-IDF) of visual words is used to create histograms representing video segments. These histograms are applied in a continuous framework using a data stream algorithm to update the system knowledge based on the classification score obtained by the histogram. In [3] contextual information incorporate depth camera data and global frame descriptors in a BoW framework.

In image processing domain, enriched BoW representations with extra knowledge from the set of local descriptors have been explored on several approaches [11,21]. However, those works use parametric models leading to a very high-dimensional representation. On other hand, BossaNova model [1], which follows BoW formalism (coding/pooling), keeps more information than the traditional BoW during the pooling step. It estimates a probability density function by computing a histogram of distances between local descriptors and codewords. In addition to the pooling strategy, in [1], it also proposed a localized soft-assignment coding that considers only the k-nearest codewords for coding a local descriptor.

3 Traditional Bag-of-Words

In the traditional Bag-of-Words (BoW) model for mid-level representation, the input is a set of unordered local descriptors, representing the whole data. The BoW model first requires a dictionary learned from the feature points. The most common approach to create the dictionary is by an unsupervised clustering algorithm (*e.g.*, K-means algorithm). The dictionary is composed by a set of M codewords. More precisely, 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 and let $\mathbb{C} = \{\mathbf{c}_m \in \mathbb{R}^d\}_{m=1}^{M}$ and $\mathbb{Z} \in \mathbb{R}^M$ be the dictionary learned and the final vector representation, respectively. As formalized in [2], the mapping from $\mathbb{X}$ to $\mathbb{Z}$ can be decomposed into three sucessive steps: (i) coding; (ii) pooling; and (iii) concatenation, as follows:

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

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

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

In the traditional BoW framework [14], the coding function f minimizes the distance to a codebook, and the pooling function g computes the sum over the pooling region. As illustrated in Fig. 1, the coding and pooling functions can be visualized in terms of the matrix $\mathbf{H}$ with N column and M rows, in this example, the coding function f for a given descriptor $\mathbf{x}_j$ corresponds to information obtained from the j^{st} column. Moreover, the pooling function g for a given visual word $\mathbf{c}_m$ corresponds to the m^{st} row of the $\mathbf{H}$ matrix. Both functions could be, more precisely, defined as follows:

$$\alpha_j \in \{0,1\}^M = \alpha_{m,j} = 1 \text{ iff } j = \underset{m \leq M}{\mathbf{arg\,min}} \parallel \mathbf{x}_j - \mathbf{c}_m \parallel_2^2 \tag{4}$$

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

in which $\mathbf{c}_m$ denotes the m-th codeword.

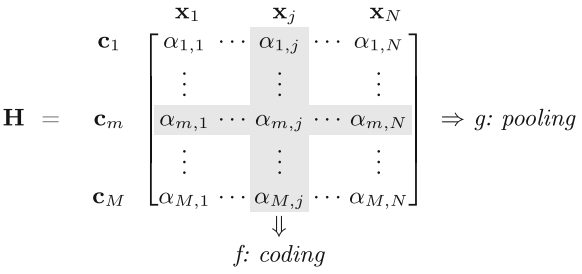

Fig. 1. Matrix $\mathbf{H}$ of BoW model in which the rows and columns are related to coding and pooling functions, respectively, as presented in [1].

In [4], some improvements were obtained by smoothing the distribution during the pooling function. This approach, called soft-assignment, models an ambiguity concept in the attribution, creating more expressive models for classification. They indicate that large vocabulary increases the probability of multiple relevant visual words to represent one feature point. This is called visual word uncertainty and can be formulated as follows:

$$\alpha_{m,j} = \frac{\exp(-\beta \parallel \mathbf{x}_j - \mathbf{c}_m \parallel_2)}{\sum_{k=1}^{M} \exp(-\beta \parallel \mathbf{x}_j - \mathbf{c}_k \parallel_2)} \tag{6}$$

where β is a parameter that controls the softness of the soft assignment (hard assignment is the limit when $\beta \to \infty$).

4 An Extended BoW Formalism

In the traditional BoW framework [14], the function g for pooling computes the number of descriptors over the pooling region, thus the mid-level representation could be defined by a concatenation of all values related to the codewords. Unfortunately, this pooling strategy is quite poor in terms of information inside each pooling region, mainly related to spatial distribution of the descriptors. To cope with this lack of information, we propose a new mid-level representation, so-called BOH (**B**ag **O**f local distribution of descriptors on concentric **H**yperspheres), which explores the descriptor position inside the largest hypersphere centred at each codeword for computing the pooling. For that, we propose to divide this hypersphere into equally probable hyper-regions in which the descriptors inside one hyper-region have similar distances to the codeword.

Let S_i and S_j be two hyperspheres centered at codeword $\mathbf{c}_m$ with radius r_i and r_j, respectively, in which $r_i < r_j$. We define the hyper-region $R_{i,j}$ between the hyperspheres S_i and S_j as the hyper-region computed by the difference of S_i and S_j. More precisely, a d-dimensional descriptor belongs to the $R_{i,j}$ if the distance to the codeword $\mathbf{c}_m$ is higher than r_i and smaller than or equal to r_j.

Two hyper-regions $R_{i,j}$ and $R_{i',j'}$ are considered equally probables if they have the same volume, *i.e.*, $V(R_{i,j}) = V(R_{i',j'})$. Let E be the number of equally probable hyper-regions related to the codeword $\mathbf{c}_m$. Without loss of generality, let S_E and S_1 be two hyperspheres with radius r_E and r_1 centered at $\mathbf{c}_m$, $V(R_{E-1,E}) = V(R_{0,1})$ iff $V(S_E) = E \times V(S_1)$. From this definition, it is easy to show that $r_e = r_1 \times \sqrt[n]{e}$, $\forall\ e \in [1, E]$.

Considering these E equally probable hyper-regions, the proposed pooling strategy is the histogram of distances between the local descriptors and the codewords taking into account the radius of the largest n-dimensional hypersphere over the pooling region. Let $\mathbb{X} = \{\mathbf{x}_j\}$ be an unordered set of d-dimensional descriptors $\mathbf{x}_j$ extracted from a video, such that $j \in [1, N]$. The proposed strategy for pooling is defined by:

$$h_{m,e} = \text{card}\left(\mathbf{x}_j \mid \alpha_{m,j} \in \left[r_E^{\mathbf{c}_m} \sqrt[N]{\frac{e}{E}}, r_E^{\mathbf{c}_m} \sqrt[N]{\frac{e+1}{E}}\right]\right), e \in [0, E-1] \tag{7}$$

in which $r_E^{\mathbf{c}_m}$ is the radius of largest n-dimensional hypersphere centered at codeword $\mathbf{c}_m$. The final representation $\mathbf{z}$ is given by:

$$\mathbf{z} = [h_{m,e}]^{\mathrm{T}}, \ (m,e) \in \{1,...,M\} \times \{1,...,E\} \tag{8}$$

where $\mathbf{z}$ is a vector of size $M \times E$.

When the number of equally probable hyper-regions is equal to 1, our pooling strategy is similar to the traditional BoW. As the number of hyper-regions and codewords increase, the vector $\mathbf{z}$ is more sparse but it approximates better the actual distribution of distances. Thus there is a trade-off between the sparsity and this size.

In order to exemplify the traditional BoW, the BossaNova and the proposed method pooling strategies, it is illustrated in Fig. 2 how the regions related to two codewords are divided. In this example, the coding is done by a hard-assignment in which the d-dimensional descriptor is associated with just one codeword. In the traditional BoW, as illustrated in Fig. 2(a), the codewords are represented by the number of descriptors which are assigned to them. For BossaNova and BOH each codeword is represented by a histogram of descriptors which are quantized according to their distance to the codeword. While the quantization used by BossaNova is based on linear function in terms of the distance-to-codeword, as shown in Fig. 2(b), the quantization used by BOH is based on the volumes of the hyper-regions obtained by hyperspheres centered at codewords, as illustrated in Fig. 2(c).

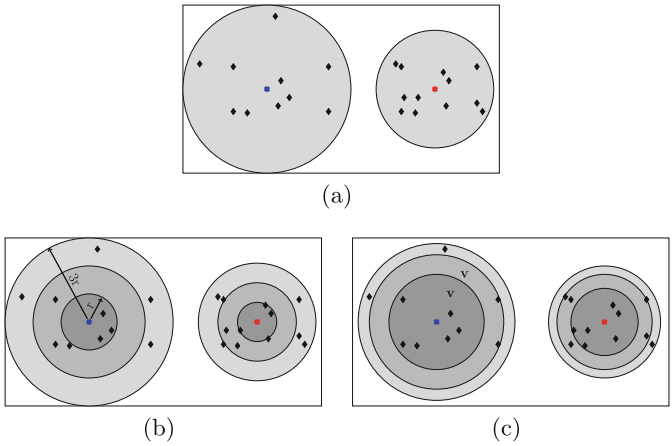

Fig. 2. Example of pooling strategy for d-dimensional descriptors taking into account BoW, BossaNova and BOH. For BossaNova the number of hyper-regions is equal to 3, and for BOH, the number of equally probable hyper-regions related to each codeword is also equal to 3.

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

Fig. 3. Example of hand waving with same subject in different scenarios

5 Experimental Analysis

In this section, we describe the three used datasets, the protocols for classification and the experimental setup. Moreover, a quantitative analysis, in terms of classification rates, comparing our method with the state-of-the-art approaches is given.

5.1 Datasets and Protocols

In order to validate the proposed method we tested our approach in three well-known action recognition datasets: (i) KTH [12]; (ii) UCF Sports [10]; and (iii) UCF 11 [7]. The datasets choices were made due their distinctive characteristics, such as video duration, intraclass variability and noise scene elements.

The KTH dataset [12] 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. 3. There are totally 600 video clips with 160×120 pixels size and different video durations. We adopt the same experimental setup as in [12,15], 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).

The UCF sports dataset [10] 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. 4. Contrary to what has been done in many works that apply their methods on this dataset, we do not extended the dataset with a flipped version of the videos, trying to prevent the classifier from learning the background instead of the actions. We adopt a split set dividing the dataset into 103 training and 47 test samples as in [5].

The UCF11 dataset [7] contains 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. This dataset is challenging due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background, and illumination conditions. Some examples are illustrated in Fig. 5. The dataset contains

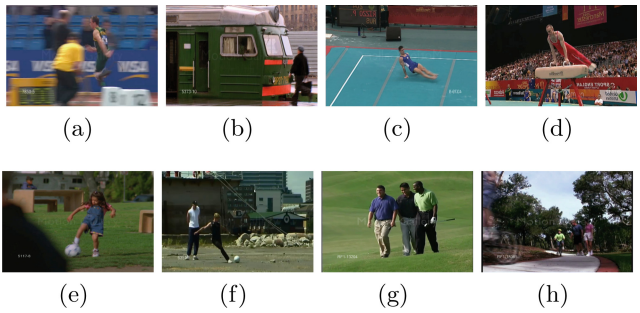

Fig. 4. Example of intra-class variability in UCF Sports dataset. (a) and (b) are both examples from running class, (c) and (d) from swinging, (e) and (f) from kicking; while (g) and (h) from walking.

Fig. 5. Example of UCF11 challenges, such as object appearence in (a) and (b), viewpoint in (c) and (d), cluttered background in (e) and (f), and illumination conditions in (g) and (h).

a total of 1646 videos. We adopt the original setup [7] using the leave-one-out cross-validation for a pre-defined set of 25 folds.

5.2 Experimental Setup

Regarding the feature descriptor, we have chosen to use an approach with a dense descriptor (dense trajectories [15]) because it is simple and achieved good results. After the feature extraction step, BoW, BossaNova and BOH are used to organize the low-level features to represent each video using the mid-level representation. Here, we used the following parameter values for computing the BossaNova: $\lambda_{min} = 0.4$, $\lambda_{max} = 2$, $knn = 10$ (semi-soft assignment), $B = \{2, 4\}$ and $M = \{512, 2048\}$ (number of visual codewords). For the proposed pooling strategy BOH, we used the following parameter values: $knn = 10$ (semi-soft assignment), $E = \{2, 4\}$ and $M = \{512, 2048\}$ (number of visual codewords).

For classification we used non-linear SVM with an RBF kernel which is a popular classifier that is used throughout different works for human action

classification [15]. Since this classifier is vastly used in human action classification, it is interesting to use it to make fair comparisons between different approaches.

5.3 Comparison with the State-of-the-art

In order to compare the proposed method to some of the state-of-the-art approaches we adopted the classification rate (also called recognition rate). Usually, in literature, there is a bit confusion between the use of classification rate and accuracy. For the sake of clarification, in this work, the classification rate is the number of correct video classification by the number of videos. In Table 1, a comparison, in terms of classification rate, is presented. Except for BossaNova, the rates of the compared methods were obtained from the original paper. As one can note, ours give competitive rates for KTH and UCF 11, and much better results for UCF Sports. When compared to BossaNova, which uses a similar pooling strategy, our results are better in UCF Sports and UCF 11.

Table 1. The classification rates for the compared approaches.

Approach	Parameters	KTH	UCF sports	UCF 11
Dense trajectories [15]	-	94.2%	-	84.1%
BoW	M = 2048	85.7%	55.3%	53.9%
BossaNova [1]	M = 512, B = 2	94.9%	66.0%	78.4%
	M = 512, B = 4	96.3%	66.0%	81.0%
	M = 2048, B = 2	97.7%	70.2%	78.0%
	M = 2048, B = 4	97.7%	70.2%	75.7%
Ours	M = 512, E = 2	94.9%	72.3%	75.3%
	M = 512, E = 4	94.0%	72.3%	76.0%
	M = 2048, E = 2	96.8%	74.5%	79.3%
	M = 2048, E = 4	96.3%	76.6%	81.4%

The performances, in terms of classification rates, of the compared approaches applied to the KTH, UCF Sports and UCF 11 are illustrated in Fig. 6. As we can see, the rate for BOH increases when the number of codewords and hyper-regions increase. This behavior does not occurs for BossaNova since there is no a monotonic increasing neither for number of codewords nor for the number of hyper-regions. Furthermore, both methods are better than traditional BoW. In terms of time performance, there is no significant difference between BOH, BossaNova and BoW, once the main time consuming operation rely on calculating the distances between feature points and codewords during coding and are the same for all methods.

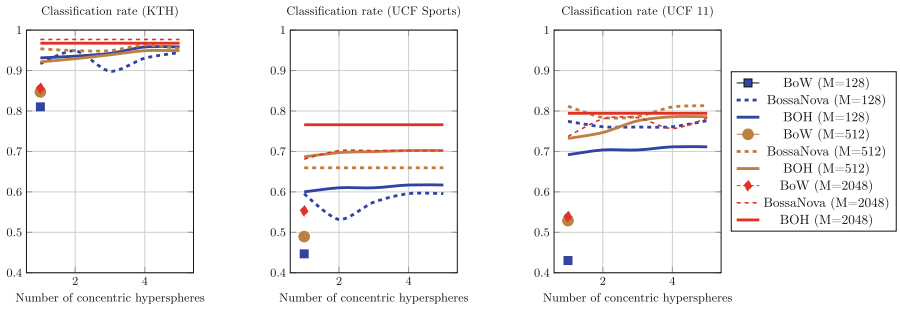

Fig. 6. A comparison between the proposed method and BossaNova concerning the classification rate according to the number of hyperpheres from 1 to 5. The classification rate for BoW is also illustrated (when the number of hypersphere is equal to 1).

6 Conclusions and Further Works

In this work, we addressed the task of human action classification using only the information present in the content of the video. Also, we focused on an intermediate stage between feature extraction and classification by using an extended BoW formalism, so-called BOH to generate a new mid-level video representation which is obtained directly from densely sampled features extracted around trajectories.

The idea is to increase the classification rate by careful use of a well disseminate motion descriptor. Here we explored a new strategy for the pooling step based on a volumetric partition of the hypersphere centered at codewords in order to maintain the same probability of assignment to a given hyper-region. The results indicates that this kind of pooling could decrease the quantization error of the descriptors.

Regarding classification protocols, we experimented the training and testing classification (here called split) for KTH and UCF Sports, and the leave-one-group-out cross-validation for UCF 11. Experimental results demonstrated that our strategy either has improved the recognition rates with respect to the BossaNova, expect for KTH.

For further endeavors, we will study different ways encoding quantization errors into video descriptors. Another interesting research path is to investigate the quality of video data used during (and filter it out before) training time for the classification step and its relationship with the support vectors needed to produce better accuracy results in human action classification.

References

1. Avila, S., Thome, N., Cord, M., Valle, E., Araújo, A.D.A.: Pooling in image representation: the visual codeword point of view. Comput. Vis. Image Underst. (CVIU) **117**(5), 453–465 (2013)

2. Boureau, Y.L., Bach, F., LeCun, Y., Ponce, J.: Learning mid-level features for recognition. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 2559–2566, June 2010
3. Foggia, P., Percannella, G., Saggese, A., Vento, M.: Recognizing human actions by a bag of visual words. In: 2013 IEEE International Conference on Systems, Man, and Cybernetics (SMC), pp. 2910–2915. IEEE (2013)
4. van Gemert, J.C., Veenman, C.J., Smeulders, A.W.M., Geusebroek, J.M.: Visual word ambiguity. IEEE Trans. Pattern Anal. Mach. Intell. **32**(7), 1271–1283 (2010)
5. Lan, T., Wang, Y., Mori, G.: Discriminative figure-centric models for joint action localization and recognition. In: International Conference on Computer Vision (ICCV) (2011)
6. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8 (2008)
7. Liu, J., Luo, J., Shah, M.: Recognizing realistic actions from videos. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2009, 20–25 June 2009, Miami, Florida, USA, pp. 1996–2003 (2009)
8. Luo, H., Lu, H.: Multi-level sparse coding for human action recognition. In: 2016 8th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), vol. 1, pp. 460–463. IEEE (2016)
9. Poppe, R.: A survey on vision-based human action recognition. Image Vis. Comput. **28**(6), 976–990 (2010)
10. Rodriguez, M.D., Ahmed, J., Shah, M.: Action mach a spatio-temporal maximum average correlation height filter for action recognition. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2008, pp. 1–8, June 2008
11. Sánchez, J., Perronnin, F., Mensink, T., Verbeek, J.: Image classification with the fisher vector: theory and practice. Int. J. Comput. Vis. (IJCV) **105**(3), 222–245 (2013)
12. Schuldt, C., Laptev, I., Caputo, B.: Recognizing human actions: a local SVM approach. In: 17th International Conference on Proceedings of the Pattern Recognition, ICPR 2004, Washington, DC, vol. 3, pp. 32–36. IEEE Computer Society (2004)
13. Shukla, P., Biswas, K.K., Kalra, P.K.: Action recognition using temporal bag-of-words from depth maps. In: MVA, pp. 41–44 (2013)
14. Sivic, J., Zisserman, A.: Video Google: a text retrieval approach to object matching in videos. In: Proceedings of the Ninth IEEE International Conference on Computer Vision, ICCV 2003, Washington, DC, vol. 2, p. 1470. IEEE Computer Society (2003)
15. Wang, H., Kläser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3169–3176. IEEE (2011)
16. Wiliem, A., Madasu, V., Boles, W., Yarlagadda, P.: Adaptive unsupervised learning of human actions. In: 3rd International Conference on Imaging for Crime Detection and Prevention, ICDP 2009, pp. 1–6, December 2009
17. Wu, J., Zhou, D., Xiao, G.: A hierarchical bag-of-words model based on local space-time features for human action recognition. In: 2013 International Conference on IT Convergence and Security (ICITCS), pp. 1–4. IEEE (2013)
18. Yan, X., Luo, Y.: Making full use of spatial-temporal interest points: an adaboost approach for action recognition. In: 2010 17th IEEE International Conference on Image Processing (ICIP), pp. 4677–4680. IEEE (2010)

19. Yi, T., Qiuqi, R.: Weight and context method for action recognition using histogram intersection. In: 5th IET International Conference on Wireless, Mobile and Multimedia Networks, ICWMMN 2013, pp. 229–233, November 2013
20. Zhang, Z., Wang, C., Xiao, B., Zhou, W., Liu, S.: Action recognition using context-constrained linear coding. IEEE Sig. Process. Lett. **19**(7), 439–442 (2012)
21. Zhou, X., Yu, K., Zhang, T., Huang, T.S.: Image classification using super-vector coding of local image descriptors. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6315, pp. 141–154. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15555-0_11