A new bag of visual words encoding method for human action recognition

Xavier Cortés, Donatello Conte, Hubert Cardot

A new bag of visual words encoding method for human action recognition

Xavier Cortés
xavier.cortes@univ-tours.fr

Donatello Conte
donatello.conte@univ-tours.fr

Hubert Cardot
hubert.cardot@univ-tours.fr

LiFAT (EA 6300)
University of Tours
64, Avenue Jean Portalis
37000, Tours
France

***Abstract*—Human action recognition in videos is one of the key problems in computer vision. Inspired by image classification models, techniques based on bags of visual words have become one of the most effective approaches to solve this problem. The most usual way to engage an interest point in a bag of words is by means of the closest word found in a previously trained codebook. However, the quality of representation decreases when interest points have different visual words at similar distances or when we map noisy interest points. The aim of this paper is to present a new encoding procedure to engage the interest points in a bag of visual words improving the quality of the representation. The encoding that we propose tries to map only the relevant interest points detected in the scene. We experimentally show that using the new encoding method we can significantly improve the classification ratio.**

Fig. 1. Human action frame in a video sequence. Green points: relevant interest points. Red points: irrelevant interest points.

I. Introduction

Human action recognition is the task of giving a label to an image sequence, corresponding to the activity of a human or a group of humans, or even an interaction between humans and objects. An action, in fact, can be viewed at different levels of complexity, starting from one person making an action like running, walking, etc. to complexes interactions between people and objects in order to recognize events. Recognizing human actions in real-world environment is a necessary task in several applications such as intelligent video surveillance, urban traffic management, and so on. Accurate recognition of actions is a very challenging task, due to different problems like cluttered backgrounds, luminosity and viewpoint variations, and others. One more challenge is to design robust algorithms that are efficients varying datasets, environments, etc. In this paper we focus on designing a new bag of visual words encoding method to classify human actions in videos when only one person is present. This scenario is the base for building more complex recognition algorithms, but still this context remains challenging, as demonstrated by the numerous papers still recently published in the major computer vision conferences. In this context, our paper shows the effectiveness of the proposed approach. The paper is organized as follows: in Section II Related Works are presented together with the description of original contribution of our approach with respect to the state-of-the-art. Section III shows the general framework in which our algorithm is applied. Details of the approach are described in Section IV while Section V is consecrated to show Experimental Phase and Results. Finally some conclusions and a discussion on future works are presented in Section VI.

II. Related Works and Motivation

Bag of Visual Words (BoVW) is a pattern representation model that has been successfully used in image classification for several years [1], [2] and [3]. Originally inspired by the classical bag-of-words models used in documents analysis and natural language processing [4], the key idea of BoVW is to represent an object by means of a histogram counting the number of occurrences of a set visual words belonging to a trained codebook that can be detected in the original object. The last years, some works have shown how BoVW can also be successfully applied in video classification [5], [6], [7], [8], [9].

In this framework it is crucial the process of mapping the interest points extracted form the videos in the BoVW. The aim of the mapping process, also known as encoding, is to assign different values to one or more bins of the BoVW depending of the features of each interest point detected in a video. Different ways to encode a BoVW have been

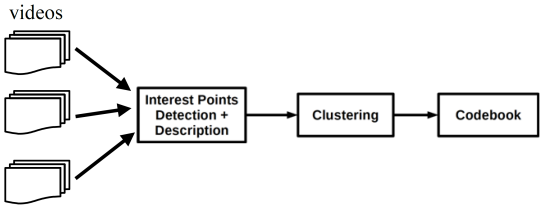

Fig. 2. Human action codebook generation scheme.

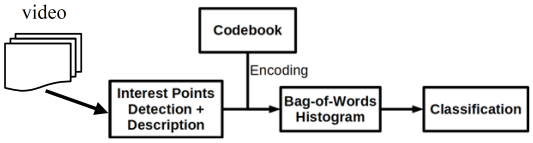

Fig. 3. Human action classification based on BoVW scheme.

explored in image processing such as *Vector quantization* [1], *Soft-assignment* [10], [11], *Sparse encoding* [2], *Locality-constrained Linear encoding* [3] and *Fisher Kernel* [12].

At this point, it is important to remark that the process of detecting interest points can be affected by distortions such as shadows, background textures or irrelevant objects that appear in the scene (Figure 1). This means that many of the detected interest points are non-informative for action classification. Mapping non-informative interest points into the BoVW adds noise in the representation and can reduce the performance of the model. Interest points detection algorithms try to avoid the noise when detecting the relevant points, however, still remain non-informative interest points that are impossible to remove.

The aim of this paper, is to propose a new encoding method that discriminates the irrelevant points from the other ones when doing the mapping. To this end we propose a method to deduce a covering region for each visual word when we build the codebook, defining a limited area on the domain of the points descriptors within which a visual word maps interest points that are inside.

III. General Framework

We can divide the model in two different parts. The first part referred to the training process (Figure 2) and the second part referred to the classification (Figure 3). Each part can be described by a sequence of different steps. In this section, we explain the different steps.

A. Interest points detection

The first step in both training and classification is to detect the interest points appearing in the scene. Several methods can be used, based on spatial information [13], [14], [15] or based on temporal and spatial information [16]. STIP detector [16] can be seen as an extension of Harris corners [13] including the temporal dimension. The model consists in to apply the Harris corners equations with horizontal, vertical and temporal dimensions as in [16].

B. Interest points description

To describe the interest points detected in Section III-A we propose to use, Histograms of Oriented Gradients (HOG) [17] based on counting in localized portions of a frame the occurrences of oriented gradients or Histograms of Optical Flows (HOF) [18] based on counting the occurrences of oriented optical flows. These descriptors contain a set of attributes that define the interest points and are represented as vectors in which each cell corresponds to a different attribute. The Euclidean distance is used to determine the closeness between two different descriptors.

C. Visual codebook

The codebook is the dictionary of visual words used to construct the BoVW (explained in detail in section III-D). A visual word carries a set of descriptive characteristics (section III-B). During the training process, the codebook is built by clustering the interest points detected from different training videos (Figure 2). One of the most popular clustering algorithms used in order to build a codebook is the *K-means* due its good balance in terms of classification accuracy and computational complexity [19]. The visual words represented in the codebook correspond to the centroids found by the *K-means* algorithm. Generating a representative codebook is a fundamental point in this framework, an adequate number of interest points (inputs) and the total number of visual words (outputs) will determine its performance. In [20], to limit the computational complexity achieving reasonable results, the authors propose to select a reduced sampling of interest points randomly generated from the training videos to generate the codebook. More recently, in [7], the authors propose a multi-level approach to generate the codebook. They apply three clustering levels: clustering per videos, clustering per classes and final clustering. Using this strategy they reduce the computational complexity of the problem and they get a more representative codebook as they do not need to arbitrarily reduce the sampling of interest points for computational reasons. The centroids found by the clustering algorithm will be the visual words represented in the codebook.

D. Bag of visual words

BoVW is a structure that describes a video by means of a frequency histogram counting the number of occurrences of visual words included in a codebook. The process of building a BoVW is based on mapping the interest points detected in the video to the BoVW by means of an encoding method. One of the most popular methods of encoding is the *Vector quantization* encoding.

The *Vector quantization* encoding also know as *Hard assignment* [1] consists in to increase the value of the bin corresponding to the nearest word given an interest point as follows:

$$B_j^i = \begin{cases} B_j^i + 1, & \text{if } j = \underset{j \in \{1,\dots,J\}}{\arg\min} \|x_n^i - w_j\| \\ B_j^i, & \text{otherwise} \end{cases} \quad \forall n \in \{1,\dots,N\} \tag{1}$$

where B^i is the BoVW that represents video i, initialized to $B^i = \overrightarrow{0}$, j is the bin number out of J in the BoVW, X^i is the set of D-dimensional interest points detected in video i, *i.e.* $X^i = [x_1^i, x_2^i, \ldots, x_N^i] \in \mathbb{R}^{D \times N}$ and W is the codebook of visual words, *i.e.* $W = [w_1, w_2, \ldots, w_J] \in \mathbb{R}^{D \times J}$. Note that we use a single codebook to encode all the videos. This encoding is shown visually with a toy example in Figure 4a.

Normalize the BoVWs, with a normalization method such as *Min-max normalization* [21], allows to manage videos of different duration:

$$B_j^i = \frac{B_j^i - \min_{i \in \{1,\ldots,I\}} B_j^i}{\max_{i \in \{1,\ldots,I\}} B_j^i - \min_{i \in \{1,\ldots,I\}} B_j^i} \quad (2)$$

where I is the total number of videos.

E. Classification

The last step is the classification. Support Vector Machines (SVMs) are supervised learning algorithms successfully used for classification in several domains [22]. We propose to train the model using one SVM per action class targeting the BoVWs to the corresponding classes of the training videos. Finally, the trained SVMs are used to determine if the BoVWs representing the videos we want to classify (Figure 3) belong to an action class or not.

F. Overview

In our model we followed these design configurations. The interest point detection algorithm is the STIP. As in [20] and [7] we use HOF descriptors for the KTH dataset [23] and HOG descriptors for the Weizman dataset [24]. To generate the codebook, we use a multi-level K-means approach as in [7]. The normalization method for the BoVW is the *Min-max normalization* considering results reported in [8]. Finally, we use a linear SVM to classify the videos.

IV. Covering Region Encoding

In this section is proposed a new encoding model aiming:

- to avoid the ambiguity problems when an interest point has two or more words at similar distances and
- to remove irrelevant points by penalizing the large distances between interest points and visual words.

Note that when an interest point is irrelevant (for example, interest points detected in objects that are not involved in the human action, immobile body parts, clothing pieces...) it is not informative to represent the point in the BoVW. Adding non-informative points distorts the quality of the representation and could reduce the performance of the model in terms of classification accuracy.

The encoding we propose it follows a soft-approach taking into account the radius of the covering region corresponding to each visual word obtained during the generation of the codebook through a clustering algorithm. The covering region refers to the region surrounding each visual word in the interest points descriptors space, this region limits the area in which a visual word can map an interest point, unlike other encodings, such as *Vector quantization* encoding, which only takes into account the closest visual word independently of the distance (Figure 4a). The boundary of the covering region is defined by its radius. When the covering region radius covers a larger region it is more tolerant to distant interest points than when a covering region covers a smaller region.

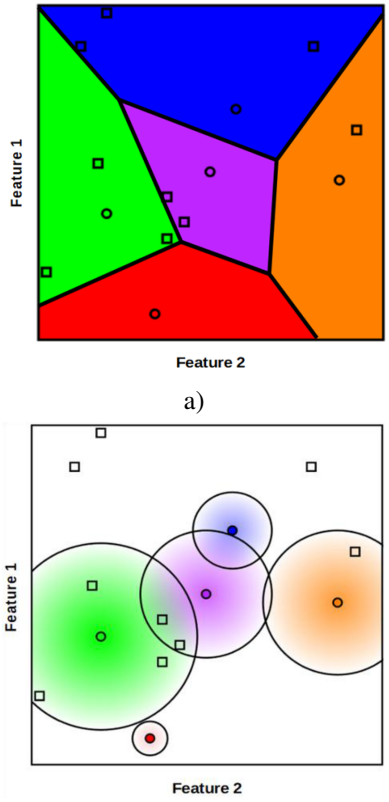

Fig. 4. Toy examples of (a) *Vector quantization* encoding and (b) *Covering region* encoding. Visual words embedded in a $2D$ featured space, are represented by small circles of different color, regions borders are represented by black lines for the *Vector quantization* encoding and big circles for the *Covering region* encoding, assigned interest points are represented by colored squares depending on the word where they are mapped and white for unassigned interest points. Regions colors represents the word that is assigned to the points that are within the region domain.

There are two possibilities to define the covering region radius (during the training phase). The first one is to consider the distance between the farthest point belonging to the cluster with respect to the visual word (note that each visual word is the centroid of its cluster) and the second one is to consider the mean distance between all the points of the cluster with respect to the visual word. We tested both possibilities and we have experimentally observed that the performance is better considering the arithmetic mean. Then, we formally define the covering region radius as follows:

$$R_j = \frac{1}{C} \sum_{c=1}^{C} \|p_c^j - w_j\| \quad (3)$$

where R_j is the radius of the covering region that represents the codebook word w_j and p_c^j is an interest point c out of C that belongs to the cluster corresponding to w_j.

To determine the degree of membership of an interest point with respect to a visual word, we compute the relation between the distance of the interest point with respect to the visual word and the radius times a parameter α that gauges a threshold of acceptability as follows:

$$B_j^i = B_j^i + \hat{u}(x_n^i, w_j) \; \forall n \in \{1, \ldots, N\} \tag{4}$$

where:

$$\hat{u}(x_n^i, w_j) = \begin{cases} 1 - \frac{e^{\beta \cdot \|x_n^i - w_j\|}}{e^{\beta \cdot \alpha \cdot R_j}}, & \text{if } \|x_n^i - w_j\| \leq \alpha \cdot R_j \\ 0, & \text{otherwise} \end{cases} \tag{5}$$

and β is a parameter for controlling the softness of the assignment.

After mapping all interest points that appear in a video and before normalizing by the bins as we show in Eq. (2), we propose to normalize each BoVW as it follows:

$$B_j^i = \frac{B_j^i - \min_{j \in \{1,\ldots,J\}} B_j^i}{\max_{j \in \{1,\ldots,J\}} B_j^i - \min_{j \in \{1,\ldots,J\}} B_j^i} \tag{6}$$

Figure 4 shows a toy example to illustrate the differences between the classical *Vector quantization* encoding (Fig. 4a) and the *Covering region* encoding (Fig. 4b) proposed in this section. If an interest point is in two covering regions (represented by big circles) at the same time, is mapped to both visual words weighed according to its distance from the two centroids, similarly to the encoding proposed in *Soft assignment* [10]. Moreover, defining a covering region, allows us to discard the interest points that are located outside. White boxes in Fig. 4b represent those unassigned interest points. Figure 5 shows the corresponding histograms for both encodings. Note that interest points that were assigned to the blue word with the *Vector quantization* encoding are unassigned with the *Covering region* encoding, changing the resulting histogram.

V. Experiments

We have used the KTH [23] and Weizman [24] datasets in our experiments, which are commonly used in the context of human actions recognition to compare results. This section presents the databases characteristics, the experiments configuration and the achieved experimental results.

A. Databases

The KTH dataset [23] consists of 599 videos corresponding to 6 different action classes (Figure 6). The actions are performed by 25 actors in 4 different scenarios. The testing set is composed by the videos performed by the first 9 actors and the training set by the next 16 actors.

The Weizman dataset [24] consists of 90 videos corresponding to 10 different actions performed by 9 actors in a single scenario (Figure 7). The experimental configuration in the case of Weizman dataset is based on leave-one-person. Then, for each actor, there are 10 videos that correspond to the testing set and 80 to the training set.

The accuracy corresponds to the average of the recognition rates for each human action class, in both databases.

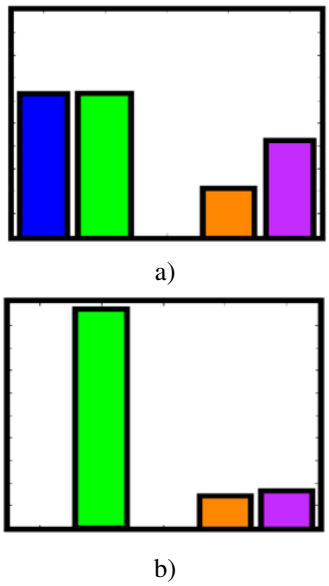

Fig. 5. Histograms representing the BoVW corresponding to the toy examples: (a) *Vector quantization* encoding, (b) *Covering region* encoding.

Fig. 6. Different frames of the KTH dataset corresponding to different actions.

Fig. 7. Different frames of the Weizman dataset corresponding to different actions.

B. Results

The aim of our experiments is to show the improvement in terms of classification accuracy when we use the *Covering region* encoding.

Figure 8 shows the accuracy results of the first experiment using the KTH dataset [23]. The codebooks have been generated using the training videos with a three-levels clustering approach as in [7]. The first level size corresponding to the videos clustering, is the 20% of interest points detected in the scene, the second level size corresponds to the actions clustering and it is $k = 500$ for each action, the third level clustering is the final codebook size and corresponds to the horizontal axis of the figure. In order to compare the classical *Vector quantization* and the *Covering region* encodings on equal terms, a single codebook for each size configuration it is generated. The vertical axis corresponds to the accuracies. The computational time required to generate the codebooks

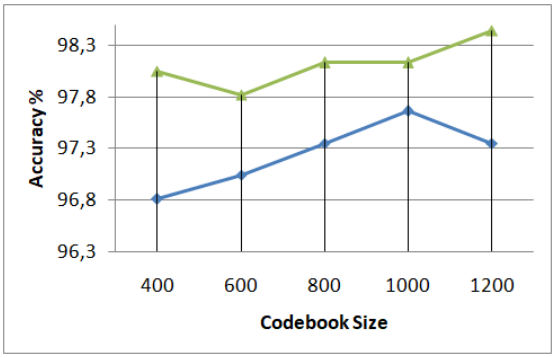

Fig. 8. Classification accuracy using different encoding methods with the KTH dataset. Blue: *Vector quantization*. Green: *Covering region*.

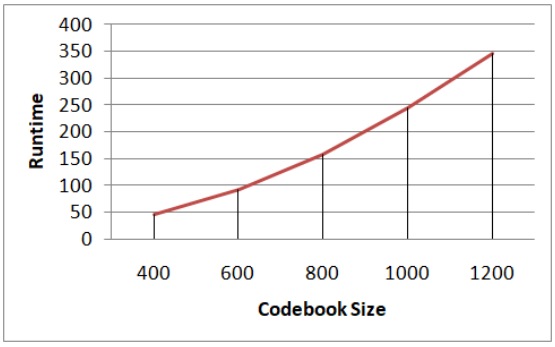

Fig. 9. Computational cost in seconds for codebook generation with the KTH dataset.

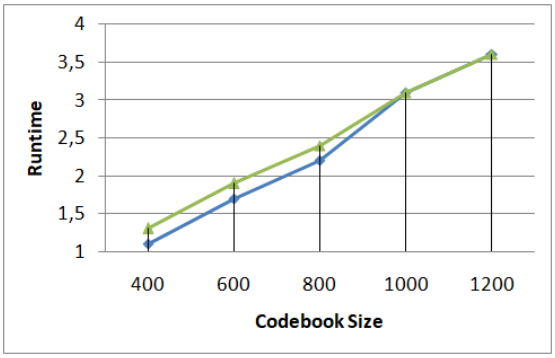

Fig. 10. Computational cost in seconds of the encoding process depending on the size of the codebook with the KTH dataset. Blue: *Vector quantization*. Green: *Covering region*.

using the KTH dataset is shown in Figure 9. Figure 10 shows the computational time required to perform a single video classification depending on the codebook size. Note that as bigger is a codebook, more visual words have to be compared to map each interest point in the BoVW and the computational time increases.

The parameters configuration for the *Covering region* encoding in this experiment has been $[\beta, \alpha] = [0.01, 3]$.

The second experiment is performed using the Weizman dataset [24], in this case we have generated a codebook with a single-level clustering approach to study the performance

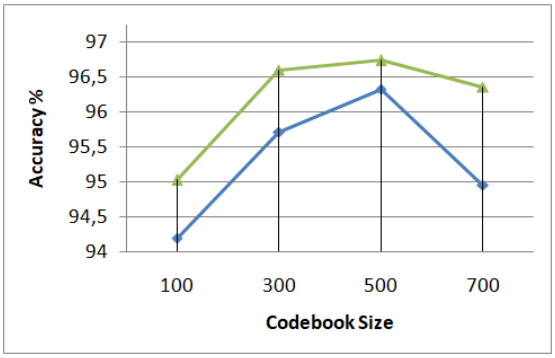

Fig. 11. Classification accuracy using different encoding methods with the Weizman dataset. Blue: *Vector quantization*. Green: *Covering region*.

TABLE I
COMPARATIVE TABLE BETWEEN OTHER WORKS USING DIFFERENT ENCODINGS BASED ON BOVW, AND OUR BEST RESULT ON THE KTH DATASET.
(VQ)*Vector quantization*, (SA) *Soft assignment*, (LSA) *Localized soft assignment*, (LCL) *Locality-constrained Linear*, (SP) *Sparse*, (FK) *Fisher Kernel*, (CR) *Covering region* ENCODING

Method	*Encoding*	*Accuracy*
Elshourbagy et al. [7]	VQ	97.7
Laptev et al. [18]	VQ	91.8
Wang et al. [25]	VQ	86.1
Wang et al. [25]	SA	89.8
Wang et al. [25]	LSA	88.9
Wang et al. [25]	LCL	89.8
Wang et al. [25]	SP	90.7
Wang et al. [25]	FK	92.1
Our best results	***CR***	**98.4**

of the encoding methods in this context. The results of this experiment are shown in Figure 11. In this experiment we used the following configuration $[\beta, \alpha] = [0.5, 1.5]$ for the *Covering region* encoding.

Finally, a comparison with previously published results following a similar experimental configuration is presented in Table I and Table II. Table I, presents previous works based on BoVW that use different methods of encoding on the KTH dataset. Table II shows a comparative study with the best state-of-the-art results that are not necessarily based on BoVW. We can see that we improve the results for the KTH dataset and we can also achieve good results for the Weizman dataset.

TABLE II
COMPARATIVE TABLE WITH THE STATE-OF-THE ART AND OUR BEST RESULTS.

Method	*Year*	*KTH*	*Weizman*
Elshourbagy et al. [7]	2016	97.7	98.9
Bilinski et al. [26]	2012	96.3	–
Bregonzio et al. [27]	2012	94.3	96.7
Bregonzio et al. [28]	2009	93.2	96.7
Niebles et al. [5]	2008	83.3	90
Klaser et al. [29]	2008	91.4	84.3
Zhang et al. [30]	2008	91.3	92.9
Dollr et al. [31]	2005	81.2	85.2
Our best results	**2018**	**98.4**	96.6

VI. Conclusions and Future Work

This paper proposes a new encoding method (*Covering region*) to map the interest points in a BoVW. The main advantage of our method is that it excludes the irrelevant interest points during the mapping process by taking into account the radius of the covering region; moreover, it uses a soft assignment approach that allows to map a single interest point to different visual words, when an interest point has two or more words at a similar distance, to manage ambiguity problems.

The experimental results in figures 8 and 11 show that the presented method improves the results in terms of classification accuracy with respect to the classical approach *Vector quantization*. The computational cost still remains comparable with other approaches.

Results show that using our proposed encoding we achieve, with respect of the state of the art approaches, the best results on the KTH dataset and very competitive results on the Weizman dataset.

As future work, we consider that it will be interesting to define the covering regions through multidimensional radius $R \in \mathbb{R}^n$. Moreover it will be interesting to propose a method to automatically learn the parameters of the encoders. This encoding could also be used for image classification.

Acknowledgment

This work is part of the LUMINEUX project supported by a Region Centre-Val de Loire (France). We gratefully acknowledge Region Centre-Val de Loire for its support.

References

[1] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *Workshop on statistical learning in computer vision, ECCV*, vol. 1, no. 1-22. Prague, 2004, pp. 1–2.

[2] J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1794–1801.

[3] 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.

[4] Z. S. Harris, "Distributional structure," *Word*, vol. 10, no. 2-3, pp. 146–162, 1954.

[5] J. C. Niebles, H. Wang, and L. Fei-Fei, "Unsupervised learning of human action categories using spatial-temporal words," *International journal of computer vision*, vol. 79, no. 3, pp. 299–318, 2008.

[6] Y.-G. Jiang, Q. Dai, W. Liu, X. Xue, and C.-W. Ngo, "Human action recognition in unconstrained videos by explicit motion modeling," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 3781–3795, 2015.

[7] M. Elshourbagy, E. Hemayed, and M. Fayek, "Enhanced bag of words using multilevel k-means for human activity recognition," *Egyptian Informatics Journal*, vol. 17, no. 2, pp. 227–237, 2016.

[8] M. M. Moussa, E. Hamayed, M. B. Fayek, and H. A. El Nemr, "An enhanced method for human action recognition," *Journal of advanced research*, vol. 6, no. 2, pp. 163–169, 2015.

[9] B. Chakraborty, M. B. Holte, T. B. Moeslund, and J. Gonzàlez, "Selective spatio-temporal interest points," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 396–410, 2012.

[10] J. C. Van Gemert, J.-M. Geusebroek, C. J. Veenman, and A. W. Smeulders, "Kernel codebooks for scene categorization," in *European conference on computer vision*. Springer, 2008, pp. 696–709.

[11] L. Liu, L. Wang, and X. Liu, "In defense of soft-assignment coding," in *Computer Vision (ICCV), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2486–2493.

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

[13] C. Harris and M. Stephens, "A combined corner and edge detector." in *Alvey vision conference*, vol. 15, no. 50. Manchester, UK, 1988, pp. 10–5244.

[14] C. Tomasi and T. Kanade, "Detection and tracking of point features," 1991.

[15] W. Förstner and E. Gülch, "A fast operator for detection and precise location of distinct points, corners and centres of circular features," in *Proc. ISPRS intercommission conference on fast processing of photogrammetric data*, 1987, pp. 281–305.

[16] I. Laptev, "On space-time interest points," *International journal of computer vision*, vol. 64, no. 2-3, pp. 107–123, 2005.

[17] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, vol. 1. IEEE, 2005, pp. 886–893.

[18] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*. IEEE, 2008, pp. 1–8.

[19] J. MacQueen *et al.*, "Some methods for classification and analysis of multivariate observations," in *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14. Oakland, CA, USA., 1967, pp. 281–297.

[20] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of local spatio-temporal features for action recognition," in *BMVC 2009-British Machine Vision Conference*. BMVA Press, 2009, pp. 124–1.

[21] T. Jayalakshmi and A. Santhakumaran, "Statistical normalization and back propagationfor classification," *International Journal of Computer Theory and Engineering*, vol. 3, no. 1, p. 89, 2011.

[22] C. Campbell and Y. Ying, "Learning with support vector machines," *Synthesis lectures on artificial intelligence and machine learning*, vol. 5, no. 1, pp. 1–95, 2011.

[23] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3. IEEE, 2004, pp. 32–36.

[24] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2. IEEE, 2005, pp. 1395–1402.

[25] Q. Y. Wang X, Wang L, "A comparative study of encoding, pooling and normalization methods for action recognition." in *Lect Notes Comput Science (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics)*, vol. 7726. IEEE, 2013, pp. 572–585.

[26] P. Bilinski and F. Bremond, "Statistics of pairwise co-occurring local spatio-temporal features for human action recognition," in *Computer Vision–ECCV 2012. Workshops and Demonstrations*. Springer, 2012, pp. 311–320.

[27] M. Bregonzio, T. Xiang, and S. Gong, "Fusing appearance and distribution information of interest points for action recognition," *Pattern Recognition*, vol. 45, no. 3, pp. 1220–1234, 2012.

[28] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1948–1955.

[29] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC 2008-19th British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[30] Z. Zhang, Y. Hu, S. Chan, and L.-T. Chia, "Motion context: A new representation for human action recognition," *Computer Vision–ECCV 2008*, pp. 817–829, 2008.

[31] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance, 2005. 2nd Joint IEEE International Workshop on*. IEEE, 2005, pp. 65–72.