Bags of Graphs for Human Action Recognition

Xavier Cortés, Donatello Conte, Hubert Cardot

Bags of Graphs for Human Action Recognition

Xavier Cortés[1], Donatello Conte[1], Hubert Cardot[1]

[1] LiFAT, Université de Tours, Tours, France
{xavier.cortes, donatello.conte, hubert.cardot}@univ-tours.fr

Abstract. Bags of visual words are a well known approach for images classification that also has been used in human action recognition. This model proposes to represent images or videos in a structure referred to as bag of visual words before classifying. The process of representing a video in a bag of visual words is known as the encoding process and is based on mapping the interest points detected in the scene into the new structure by means of a codebook. In this paper we propose to improve the representativeness of this model including the structural relations between the interest points using graph sequences. The proposed model achieves very competitive results for human action recognition and could also be applied to solve graph sequences classification problems.

1 Introduction

Human action recognition in video sequences has become a necessary task in several applications such as human-robot interaction, autonomous driving, surveillance systems and many others. However, an accurate recognition performance of human actions is a very challenging task.

Bags of Visual Words (BoVW) used before for images classification [1-3] have been shown as a successful way to address the problem of human action recognition [4-7]. The key idea of this approach is to map the interest points detected in a human action video in a representative structure referred to as BoVW taking into account its features.

In order to improve the representativeness of the BoVW model, we propose to include in the representation the structural relations between the interest points instead of evaluating the points individually.

A typical way to represent structured objects is by means of graphs. Graphs are defined by a set of nodes (interest points in our case) and edges (connections between the nodes) and they have become very important in pattern recognition. Graphs have been successfully applied in several domains such as cheminformatics, bioinformatics and computer vision among others [8-10]. We propose to represent human actions by means of graph sequences.

It is important to remark that most of the fields in which graphs have been applied in pattern recognition, are based on single graph representations estimating graph distances [11] or classifying graphs [12]. However, dynamic or time dependent problems are very common in several pattern recognition applications. For instance signal processing, study of chemical interactions, proteins folding, evaluation of diseases behaviors on populations or the human action recognition problem addressed

in this paper can be represented by streams of graphs evolving through the temporal dimension. Due to this, another important contribution of this paper is to present a method to classify graph sequences.

The paper is organized as it follows, in section 2, we introduce the necessary definitions to understand the paper, in section 3, we present a model to transform a video in a graphs sequence, in section 4, we present a classification model for graph sequences, finally, in sections 5 and 6, we show the experimental results and the conclusions.

2 Definitions

In this section we introduce some definitions necessary to contextualize and understand the paper.

2.1 Attributed Graph

Formally, we define an attributed graph as a quadruplet $g = (\Sigma_v, \Sigma_e, \gamma_v, \gamma_e)$, where $\Sigma_v = \{v_i \mid i = 1, \dots, n\}$ is the set of attributed nodes, $\Sigma_e = \{e_{ij} | i, j \in 1, \dots, n\}$ is the set of edges connecting pairs of nodes, γ_v is a function that maps the nodes to their attributed values and γ_e maps the edges.

2.2 Graph Edit Distance

The Graph Edit Distance (GED) [13, 14] defines a distance model between two attributed graphs g^p and g^q through the minimum amount of distortion required to transform g^p into g^q. To do this, a set of edit operations of insertion, deletion, and substitution of nodes and edges are required. Edit cost functions are typically used to evaluate the level of distortion of each edit operation. A sequence of edit operations that completely transform g^p into g^q is referred to as *editpath* between g^p and g^q. The total cost of the edit operations included in an *editpath* could be considered as a distance between g^p and g^q. Note, that there are several *editpaths* between two graphs depending on the edit operations we use to do the transformation. Formally, GED is defined as the minimum cost under all possible editpaths T.

$$\mathrm{GED}(g^p, g^q) = \min_{\Upsilon \in T} \mathrm{EditCost}(g^p, g^q, \Upsilon) \quad (1)$$

2.3 Sub-optimal Graph Edit Distance Computation

Optimal algorithms for computing the GED are based on complex search procedures. These procedures typically explore a wide range of possible *editpaths* between g^p

and g^q selecting the smaller in terms of total cost. The main drawback of these methods is that they are very complex in terms of computational cost.

In order to reduce its computational complexity, the problem of minimizing the GED has been sub-optimally reformulated in [15, 16]. However in these works the problem still has a considerable computational complexity. More recently, in [17], the authors propose a quadratic time approximation of GED based on the Hausdorff matching algorithm. For a better understanding of the details of this algorithm we encourage to read the original paper [17].

2.4 Graph Sequences

We define a graphs sequence $G = \{g_1, \dots, g_n, \dots, g_N\}$ as a stream of graphs representing the evolution through N different states, represented by graphs, of a single object.

2.5 Bags of Graphs

Bags of Words (BoW) are a kind pattern representation model that has been used for several years in language processing [18] and more recently as BoVW in image [1-3] and video classification [4-7]. A BoW is a global object descriptor consisting of different bins counting the mappings between the components of the represented object and the words of a codebook. We can distinguish three fundamental parts in this model, the first one is the codebook generation, the second one is the encoding procedure to embed the objects in a BoW and the last one is the classification algorithm.

In [19], the authors introduced Bags-of-Graphs (BoG), a particular type of BoW to encode digital objects into histograms based on local structures defined by graphs. The authors propose to use the BoG to encode single graph representations as proteins, letters or images.

Inspired by [19], in this paper, we propose to use a BoG to encode and classify graph sequences.

3 Representing Human Actions by means of Graph Sequences

We propose to represent each video by means of a graphs sequence. The original video is divided into splits of a predefined number of consecutive frames and each split is represented by a graph. The process consists of the following steps. First, we extract the interest points that appear in the frames of the original video. To do this, we propose to use a Spatio-Temporal Interest Point detector (STIP) [20] that can be seen as an extension of the Harris detector [21] but taking into account the temporal dimension. Next, we divide the original video into splits of consecutive frames and we group the interest points within the split where they have been detected. We build one graph per split. To do this we find the Convex Hull [22] on the spatial coordinates where the interest points have been detected to find which points are the vertexes of

the smallest polygon enveloping all the points detected in the same split. Applying this method, we filter the interest points using only the vertexes and consequently we limit the cardinality and the density of the graph representations reducing also the computational complexity of the problem. Moreover, we assume that for human action recognition tasks, the peripheral interest points are more informative than the internal interest points. To feature the nodes we propose to use the Histogram of Optical Flows (HOF) [23] of the corresponding interest points as attributes. Finally, to represent the structure, we use the sides of the Convex Hull polygon. If two nodes belong to the ends of the same side, we connect them by an edge. Fig. 1 shows the process described in this section.

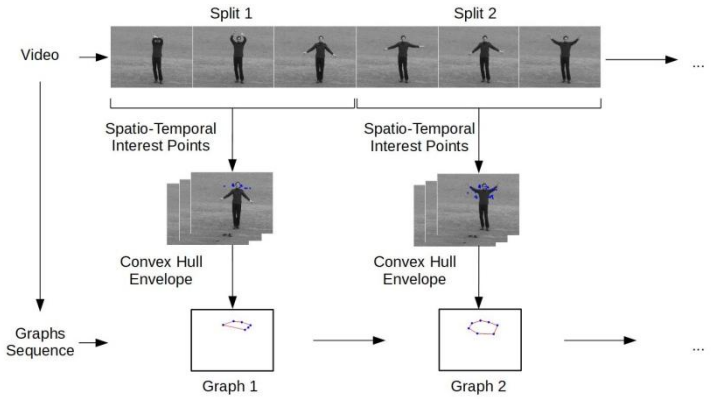

Fig. 1. Representing human action videos by means of graph sequences.

4 Graph sequences Classification using Bags of Graphs

We propose to use BoGs representations (introduced in section 2.5) to encode graph sequences into histograms, mapping the graphs of the sequence to the graphs represented in a codebook.

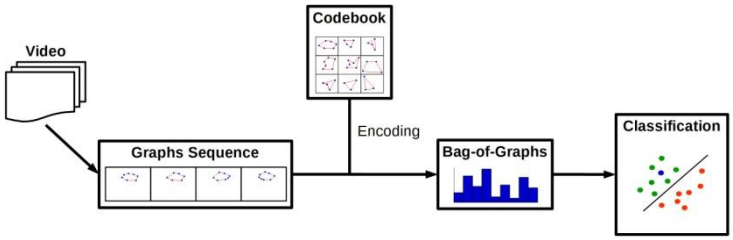

Fig. 2. Human action classification based on BoG scheme.

Fig. 2 shows the general scheme of this classification model. First, we find the corresponding graphs sequence from a human action video, this procedure is described in section 3, next, we encode the graphs sequence in a BoG using a graph codebook and finally we perform the classification. In section 4.1 we propose a method to build a graph codebook of representative graphs from a training set while

in section 4.2 we explain how to encode a graphs sequence in a BoG given a graph codebook. Finally, in section 4.3 we detail how to classify BoGs.

4.1 Generation of Graph Codebooks by means of Graph Clustering

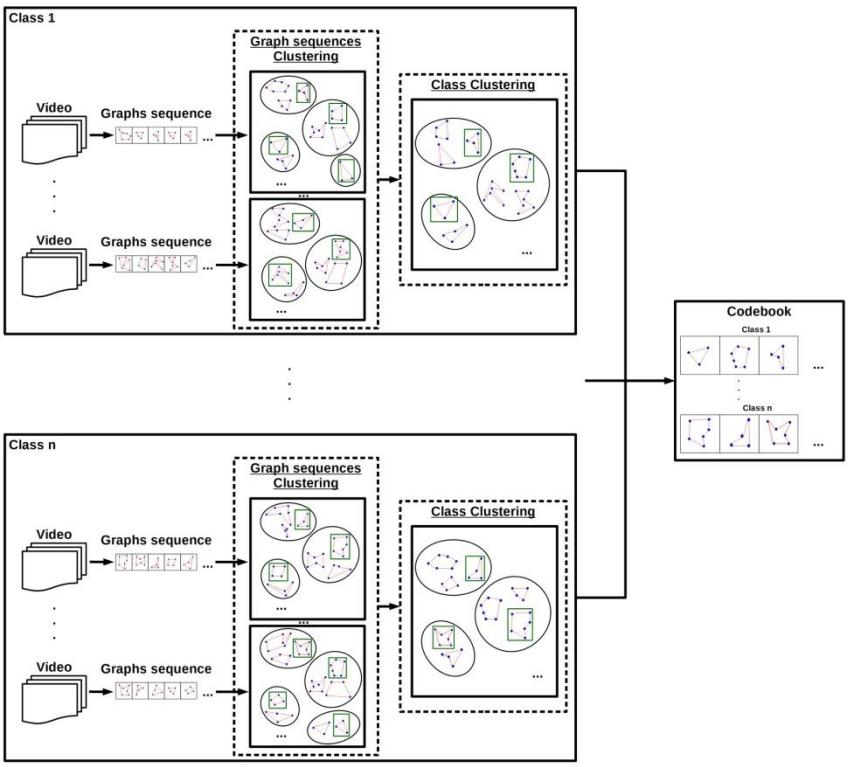

Fig. 3. Graph codebook generation scheme.

Graph codebooks are graph collections used to encode graph sequences in BoGs. A representative selection of graphs in the codebook is crucial for the performance of the model. To build a graph codebook we propose to follow a multi-level clustering approach based on the k-means algorithm [24] similar to the one presented in [7]. This approach proposes to build the codebook by means of clustering the interest points extracted from a set of training videos. The clustering is performed at different levels in order to reduce the computational complexity of the process and to be more robust to the noise. In our model we propose to cluster graphs instead of interest points. In the first level we cluster the graphs of the sequences extracted from the training videos (section 3) in order to select a subset of representative graphs per sequence while in the second level we cluster the output graphs of the first level to select the action representatives. The codebook is finally built attaching the output graphs of the second level in a single structure. In Fig. 3, we show a general scheme of the codebook generation process.

The graph clustering problem has been addressed by several authors in the literature as in [25, 26] because is not trivial given the computational complexity of the GED. We followed a similar approach to the one presented in [26] to perform the graph clustering. The authors propose to embed the graphs before applying the k-means clustering algorithm. The embedding problem aims to convert graphs into another structure to make more manageable the operations. There are different methods to solve the graph embedding problem as in [26-27]. In our model, we propose to embed graphs in n-dimensional vector spaces. The values of the embedded vector are filled by taking the GED between the graphs we are embedding to each one of the graphs in the set we are clustering. Once all the graphs have been embedded the k-means algorithm is applied on the embedded representations. The outputs of the k-means algorithm are k-centroids in the vector space corresponding to k-clusters. Finally, as clusters representatives, we select the graphs whose embedded representations are the closest to the centroids found by the k-means algorithm.

4.2 Bags of Graphs Encoding

The encoding is the procedure to represent a graphs sequence in a BoG. The BoG is a histogram divided in different bins. Each bin corresponds to one of the graphs in the codebook. We propose to follow a soft-approach [28] updating each bin according the GED between the graph of the sequence that we are mapping and the corresponding graph in the codebook. Formally:

A $\mathrm{BoG} \in \mathbb{R}^{J}$ is defined as a vector of J bins representing a graphs sequence where N is the number of graphs in a sequence $G = \{g_1, \dots, g_n, \dots, g_N\}$ and J is the number of representative graphs in a codebook $W = [w_1, \dots, w_j, \dots, w_J]$.

We encode the graphs sequence G into each bin BoG_j of the BoG using the graph codebook W as follows:

$$\mathrm{BoG}_j = \sum_{n=1}^{N} u(g_n, w_j) \tag{2}$$

Where:

$$u(g_n, w_j) = \frac{e^{(-\beta \cdot \mathrm{GED}(g_n, w_j))}}{\sum_{k=1}^{J} e^{(-\beta \cdot \mathrm{GED}(g_n, w_k))}} \tag{3}$$

Where βa parameter to control the softness of the assignment and GED is the distance function between two graphs.

4.3 Bags of Graphs Classification

To perform the classification we propose to train one linear SVM [29] per class targeting the BoGs to the corresponding classes of the training videos. The trained

SVMs are used to identify if the BoGs representing the videos that we want to classify belong to a class or not.

5 Experiments

The aim of our experiments is to empirically evaluate the performance of the model classifying videos of humans performing different actions. We tested the experiments on the KTH [30] dataset, which is commonly used in the human action recognition domain to compare results.

The dataset consist of 599 videos corresponding to 6 different action classes. The actions are performed by 25 actors in 4 different scenarios. The testing set consists of the videos performed by the first 9 actors and the training set by the videos performed by the next 16 actors.

We build the codebook using the graph sequences generated from the training videos following a multilevel clustering approach as we describe in section 4.1. In the first level, we select a sample of the 10% of graphs that appear in the original sequence and in the second level we select 50 graphs for each human action. Finally, given 6 actions and 50 graphs per action we build a graph codebook of 300 graphs. To build the graphs sequence as we described in section 3 we divide the original video into splits of 50 frames. The parameter β of the encoder (section 4.2) is fixed to 0.75. Due to its good balance in terms of computational complexity and classification accuracy, we have used the Hausdorff-GED (section 2.3) as the GED measure and the Clique centrality [31] as the costs function penalizing the structural dissimilarities.

Table 1. Accuracy results of our method and other state-of-the-art models following a similar experimental configuration.

Method	Accuracy
Elshourbagy et al. [7]	97.7
Bilinski et al. [32]	96.3
Bregonzio et al. [33]	94.3
Wang et al. [6]	92.1
Klaser et al. [34]	91.4
Laptev et al. [5]	91.8
Zhang et al. [35]	91.3
Dollr et al. [36]	81.2
Our method	96.5

In Table 1 we show a comparison between our method and other recently presented results following a similar experimental configuration. The values correspond to the average classification accuracy percentage achieved on each human action using a linear SVM classifier per class. Our method is the second best with respect to the state-of-the-art presented in the table, so proving the competitiveness of our solution.

Fig. 4 shows some sample graphs appearing in the original sequence and the corresponding BoG belonging to different action classes. We observe how BoG representing videos of the same action class tend to be more similar.

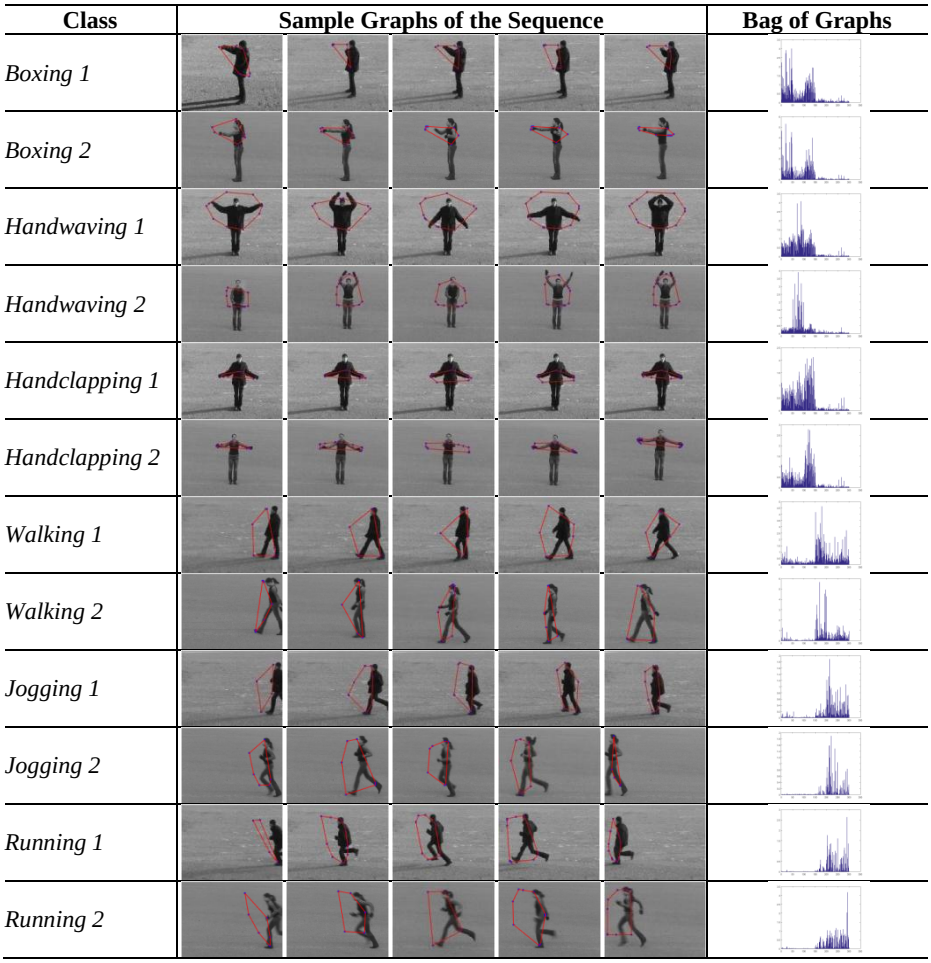

Fig. 4. Sample Graphs and BoG of different human actions in the KTH dataset.

6 Conclusions

The main purpose of the paper is to present a method for human action recognition based on BoG. To perform this task, we propose a model consisting of two main parts. The first part consists of transforming of the human action video in a sequence of graphs. The second part is to encode the sequence of graphs in a BoG before classifying. We experimentally prove that our method is competitive compared with some of the best state-of-the-art results. Another relevant contribution of our paper is the idea to use the BoG model to classify graph sequences. For future works we consider to evaluate the performance of our model using different GED measures and

to address new problems represented by graph sequences using our classification model.

Acknowledgments. 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". Workshop on statistical learning in computer vision, ECCV, vol. 1, no. 1-22. Prague, pp. 1–2, 2004.
2. J. Yang, K. Yu, Y. Gong, and T. Huang, "Linear spatial pyramid matching using sparse coding for image classification". Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on. IEEE, pp. 1794–1801, 2009.
3. J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, "Locality constrained linear coding for image classification". Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on. IEEE, 2010, pp. 3360–3367.
4. 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.
5. 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, pp. 1–8, 2008.
6. 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, pp. 572–585, 2013.
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. P. Mahé, J.-P. Vert. "Graph kernels based on tree patterns for molecules". Machine Learning 75(1):3-35, 2009.
9. X. Qi, Q. Wu, Y. Zhang, E. Fuller, C.-Q. Zhang. "A novel model for DNA sequence similarity analysis based on graph theory". Evolutionary Bioinformatics (7), pp. 149-15, 2011.
10. D. Conte, P. Foggia, C. Sansone, M. Vento. "Thirty Years Of Graph Matching In Pattern Recognition". International Journal of Pattern Recognition and Artificial Intelligence, Vol. 18, No. 3 pp: 265-298, 2004.
11. T. Li, H. Dong, Y. Shi, M. Dehmer. "A comparative analysis of new graph distance measures and graph edit distance". Information Sciences. Volumes 403–404, Pages 15-21, 2017.
12. A. Solé-Ribalta, X. Cortés, F. Serratosa. "A Comparison between Structural and Embedding Methods for Graph Classification". SSPR/SPR 2012: 234-242. 2012.
13. A. Sanfeliu, K. Fu. "A distance measure between attributed relational graphs for pattern recognition". IEEE Trans. on Sys., Man and Cybern., 13, pp. 353-362, 1983.
14. H. Bunke, G. Allermann. "Inexact graph matching for structural pattern recognition". Pattern Recognition Letters, 1(4): p. 245-253, 1983.
15. K. Riesen, H. Bunke. "Approximate graph edit distance computation by means of bipartite graph matching". Image and Vision Computing, vol. 27, no. 4, pp. 950–959, 2009.

16. F. Serratosa. “Speeding up Fast Bipartite Graph Matching trough a new cost matrix”. International Journal of Pattern Recognition and Artificial Intelligence, 29 (2), 2015.
17. A. Fischer, C. Y. Suen, V. Frinken, K. Riesen, H. Bunke. “Approximation of graph edit distance based on Hausdorff matching”. Pattern Recognition 48(2): 331-343, 2015.
18. Z. S. Harris, “Distributional structure”. Word, vol. 10, no. 2-3, pp. 146– 162, 1954.
19. Fernanda B. Silva, Rafael de Oliveira Werneck, Siome Goldenstein, Salvatore Tabbone, Ricardo da Silva Torres: “Graph-based bag-of-words for classification”. Pattern Recognition Volume 74, Pages 266-285, 2018.
20. I. Laptev, “On space-time interest points”. International journal of computer vision, vol. 64, no. 2-3, pp. 107–123, 2005.
21. C. Harris and M. Stephens, “A combined corner and edge detector”. in Alvey vision conference, vol. 15, no. 50. Manchester, UK, pp. 10–5244, 1988.
22. Andrew, A. M., "Another efficient algorithm for convex hulls in two dimensions". Information Processing Letters, 9 (5): 216–219, 1979.
23. J. Pers, V. Sulic, M. Kristan, M. Perse, K. Polanec, S. Kovacic. “Histograms of optical flow for efficient representation of body motion”. Pattern Recognition Letters 31(11): 1369-1376. 2010.
24. Hartigan, J. A.; Wong, M. A. "Algorithm AS 136: A K-Means Clustering Algorithm". Journal of the Royal Statistical Society. Series C (Applied Statistics). 28 (1): 100–108. 1979.
25. L. Galluccio, Olivier J. J. Michel, P. Comon, A. O. Hero III: “Graph based k-means clustering”, Signal Processing 92(9): 1970-1984, 2012.
26. M. Ferrer, E. Valveny, F. Serratosa, I. Bardají, H. Bunke: “Graph-Based k-Means Clustering: A Comparison of the Set Median versus the Generalized Median Graph”. CAIP: 342-350, 2009.
27. H. Bunke, K. Riesen: “Improving vector space embedding of graphs through feature selection algorithms”, Pattern Recognition 44(9): 1928-1940, 2011.
28. L. Liu, L. Wang, and X. Liu, “In defense of soft-assignment coding”. in Computer Vision (ICCV), 2011 IEEE International Conference on. IEEE, pp. 2486–2493. 2011.
29. 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.
30. C. Schuldt, I. Laptev, and B. Caputo, “Recognizing human actions: a local svm approach”. in Pattern Recognition. ICPR 2004. Proceedings of the 17th International Conference on, vol. 3. IEEE, 2004, pp. 32–36. 2004.
31. Francesc Serratosa, Xavier Cortés. “Graph Edit Distance: Moving from global to local structure to solve the graph-matching problem”. Pattern Recognition Letters 65: 204-210, 2015.
32. 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. 2012.
33. 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.
34. 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. 2008.
35. 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.
36. P. Doll´ar, 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. 2005.