Action Recognition by Pairwise Proximity Function Support Vector Machines with Dynamic Time Warping Kernels

Mohammad Ali Bagheri[1,2(✉)], Qigang Gao[1], and Sergio Escalera[3,4]

[1] Faculty of Computer Science, Dalhousie University, Halifax, Canada
bagheri@cs.dal.ca, ma.bagheri@gmail.com
[2] Faculty of Engineering, University of Larestan, Lar, Iran
[3] Computer Vision Center, Campus UAB, Edifici O, 08193 Bellaterra, Spain
[4] Dept. Matemtica Aplicada i Anlisi, Universitat de Barcelona,
Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain

Abstract. In the context of human action recognition using skeleton data, the 3D trajectories of joint points may be considered as multi-dimensional time series. The traditional recognition technique in the literature is based on time series dis(similarity) measures (such as Dynamic Time Warping). For these general dis(similarity) measures, k-nearest neighbor algorithms are a natural choice. However, k-NN classifiers are known to be sensitive to noise and outliers. In this paper, a new class of Support Vector Machine that is applicable to trajectory classification, such as action recognition, is developed by incorporating an efficient time-series distances measure into the kernel function. More specifically, the derivative of Dynamic Time Warping (DTW) distance measure is employed as the SVM kernel. In addition, the pairwise proximity learning strategy is utilized in order to make use of non-positive semi-definite (PSD) kernels in the SVM formulation. The recognition results of the proposed technique on two action recognition datasets demonstrates the ourperformance of our methodology compared to the state-of-the-art methods. Remarkably, we obtained 89 % accuracy on the well-known MSRAction3D dataset using only 3D trajectories of body joints obtained by Kinect.

1 Introduction

Support Vector Machine (SVM) is one of the leading pattern classification techniques used in various vision application tasks, such as image and video recognition [29]. Given labeled training data of the form $\{(x_i, y_i)\}_{i=1}^{m}$, with $y_i \in \{-1, +1\}$[1], the standard form of SVM finds a hyperplane which best separates the data by minimizing a constrained optimization problem:

$$\tau(w, \xi) = \frac{1}{2}||w||^2 + C\sum_{i=1}^{m} \xi_i \tag{1}$$

[1] In our formulation, the input samples, x_i, are not restricted to be a subset of R^n and can be any set, e.g. set of images or videos.

R. Khoury and C. Drummond (Eds.): Canadian AI 2016, LNAI 9673, pp. 3–14, 2016.
DOI: 10.1007/978-3-319-34111-8_1

$$\text{subject to: } y_i((w.x_i) + b) + \xi_i \geq 1$$
$$\xi_i \geq 0$$

where ξ_i are slack variables and $C > 0$ is the tradeoff between a large margin and a small error penalty.

The cornerstone of SVM is that non-linear decision boundaries can be learnt using the so called 'kernel trick'. A *Kernel* is a function $\mathcal{K} : \mathcal{X} \times \mathcal{X} \mapsto \mathcal{R}$, such that for all x_i, $i = 1, \dots, m$ yields to a symmetric positive semi-definite (PSD) matrix K, where $K_{ij} = \kappa(x_i, x_j)$. Indeed, the kernel function implicitly maps their inputs into high-dimensional *feature spaces*, $x \mapsto \Phi(x)$. Two common kernel functions are the Gaussian Kernel and linear kernel.

In the dual formulation, the SVM algorithm maximizes:

$$W(a) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{ij} \alpha_i \alpha_j y_i y_j \kappa(x_i, x_j) \tag{2}$$
$$\text{subject to: } 0 \leq \alpha_i \leq C \text{ and } \sum_{\alpha_i y_i} = 0$$

The decision function is given by:

$$f(x) = sign\left(\sum_{i=1}^{m} y_i \alpha_i \kappa(x, x_i) + b \right) \tag{3}$$

where the threshold b is defined as:

$$b = y_i - \sum_{i=1}^{m} y_i \alpha_i \kappa(x_i, x_j) \tag{4}$$

In this paper, we aim to classify human actions by employing spatio-temporal information of skeleton joint points, i.e. the real positions of body joints over the time. More specifically, we use the 3D trajectories of dominant body joints obtained by the Kinect camera. These trajectories encode significant discriminative information and is sufficient for human beings to recognize different actions [7]. In addition, according to an influential computational model of human visual attention theory [21], visual attention leads to visual salient entities, which provide selective visual information to make human visual perception efficient and effective. Trajectories of skeleton joints are visual salient points of human body, and their movements in 4D space reflect motion semantics.

From the classification point of view, these trajectories may be considered as multi-dimensional time series. The traditional recognition technique in the literature is based on time series dis(similarity) measures (such as Dynamic Time Warping). For these general dis(similarity) measures, k-nearest neighbor algorithms are a natural choice. In practice, given two actions represented by two multi dimensional time series, a time series distance measure calculates the distance between two actions. To classify an unlabeled test action (sample), its distance to all training samples is calculated. Consequently, the nearest neighbor algorithm is employed for classification. Given a test action, we calculate its distance to all training actions, e.x. by using DTW, and the target of the closest sample is predicted as the target class.

In general, the k-NN classification algorithms work reasonably well; but are known to be sensitive to noise and outliers. Since SVMs often outperform k-NNs on many practical classification problems where a natural choice of PSD kernels exists, it is desirable to extend the applicability of kernel SVMs.

In our action classification problem, however, time series distances measures are generally non-PSD kernels and basic SVM formulations are not directly applicable. To include non-PSD kernels in SVM, several ad-hoc strategies have been proposed. The straightforward strategy is to simply overlook the fact that the kernel should be non-PSD. In this case, the existence of a Reproducing Kernel Hilbert Space is not guaranteed [18] and it is no longer clear what is going to be optimized.

Another strategy, which has been applied in our work, is based on *pairwise proximity function* SVM(ppfSVM) [5]. This strategy involves the construction of a set of inputs such that each sample is represented with its dis(similarity) to all other samples in the dataset. The basic SVM is then applied to the transformed data in the usual way. As a consequence, sparsity of the solution may be lost. The ppfSVM is related to the arbitrary kernel SVM, a special case of the generalized Support Vector Machines [13]. The name is due to the fact that no restrictions such as positive semi-definiteness, differentiability or continuity are put on the kernel function.

In this paper, we investigate the effectiveness of this strategy for human action classification when the pairwise similarities are based on time-series distances measures. More specifically, we demonstrate the effectiveness of the derivative of Dynamic Time Warping (DTW), as SVM kernel function. The experimental results on two benchmark datasets prove the outperformance of the proposed method compared to the state-of-the-art techniques.

The contributions of our work are as follows: (1) we propose a new class of Support Vector Machine (SVM) that is applicable to trajectory classification, such as action recognition; (2) we introduce the derivatives of Dynamic Time Warping distance measures as pairwise similarity measures for SVM kernel; (3) we demonstrate the validity of the proposed methodology for action/gesture classification.

The rest of the paper is organized as follows: Sect. 2 reviews the related work on action recognition, and briefly introduces Dynamic Time Warping. Section 3 presents our methodology for action recognition. Section 4 evaluates the proposed method and Sect. 5 concludes the paper.

2 Related Work

2.1 Action Recognition

The fast and reliable recognition of human actions from captured videos has been a goal of computer vision for decades. Robust action recognition has diverse applications including gaming, sign language interpretation, human-computer interaction (HCI), surveillance, and health care. Understanding gestures/actions from a real-time visual stream is a challenging task for current computer vision algorithms. Over the last decade, spatial-temporal (ST) volume-based holistic approaches and local ST feature representations have been reportedly achieved good performance on some action datasets, but they are still far from being able to express the effective visual information for efficient high-level interpretation.

Various representational methodologies have been proposed to recognize human actions/gestures. Based on extracted salient points or regions [9] from ST volume, several local ST descriptor methods, such as HOG/HOF [10] and extended SURF [3] have been widely used for human action recognition from RGB data. Inspired from the text mining area, the intermediate level feature descriptor for RGB videos, Bag-of-Word (BoW) [12,23], has been developed due to its semantic representation and robustness to noise. Recently, BoW-based methods have been extended to depth data.

Development of low-cost depth sensors with acceptable accuracy has greatly simplified the task of action recognition [19]. Most importantly, the recent release of the Microsoft Kinect camera and its evolving skeleton joints detection technique in late 2011 led to a substantial revolutionary effect in the field of Computer Vision and created a wide range of opportunities for demanding applications. Shotton et al. [19] proposed one of the greatest advances in the extraction of the human body pose from depth data, which is provided as a part of the Kinect platform. Their work enables us to recover 3D positions of skeleton joints in real time and with reasonable accuracy.

Since 2011, hundreds of studies are devoted to action analysis using depth information. In [28], visual features for activity recognition are computed based on the spatial and temporal differences among detected joints. This feature set contains information about static posture, motion, and offset. Then, Naive Bayes Nearest Neighbor method was applied for the classification task. Alternatively, a histogram of 3-D joint locations (HOJ3-D) for body posture representation is proposed in [27]. In this representation, the 3D space is partitioned into bins using a spherical coordinate system, and the HOJ3-D histogram is constructed by casting joints into certain bins. After applying linear discriminant analysis (LDA) for dimensionality reduction, HOJ3-D vectors are clustered into k posture visual words. The temporal behaviour of these visual words is coded by discrete HMMs. Reyes et al. [16] used 15 joints from Primesense API to represent a human model. Dynamic Time Warping (DTW) with weighted joints is used to achieve real-time action recognition. Sung et al. [20] proposed a 459-element feature vector from various body joints for each frame, and then a two-layered Maximum Entropy Markov Model (MEMM) was applied to recognize single person activities. Despite active research for action/gesture recognition, none of the previous skeleton-based approaches considers a multiple classifier system philosophy. In [6], Bag-of-Visual-and-Depth-Words defined containing a vocabulary from RGB and depth sequences. This novel representation was also used to perform multi-modal action recognition. In [1], the authors proposed an ensemble

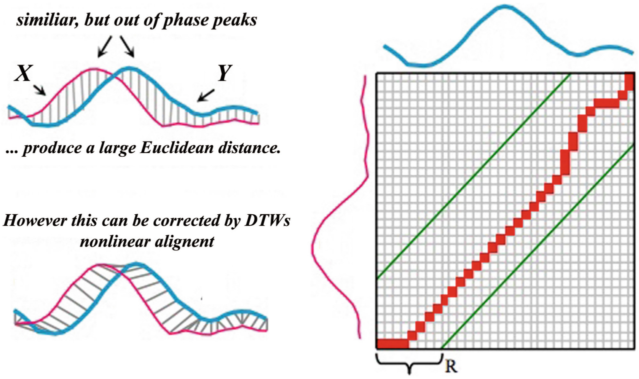
similiar, but out of phase peaks
X
Y
... produce a large Euclidean distance.
However this can be corrected by DTWs nonlinear alignent
R

Fig. 1. Top left: two time series which are similar but out of phase produce a large Euclidean distance. Bottom left: this can be corrected by DTWs nonlinear alignment. Right: to align the signals we construct a warping matrix, and search for the optimal warping path

of five action learning techniques, each performing the recognition task from a different perspective and combined the outputs of these classifiers based on the Dempster-Shafer combination theory.

2.2 Dynamic Time Warping

Dynamic Time Warping (DTW) is a well-known algorithm which aims to compare and align two temporal sequences, taking into account that sequences may vary in length (time) [16]. DTW employs the dynamic programming technique to find the minimal distance between two time series, where sequences are warped by stretching or shrinking the time dimension. Although it was originally developed for speech recognition [17], it has also been employed in many other areas like handwriting recognition, econometrics, and action recognition.

An alignment between two time series can be represented by a warping path which minimizes the cumulative distance, shown in Fig. 1. The DTW distance between time series x and y of length n and m will be recursively defined as:

$$DTW(i,j) = d(i,j) + min \begin{cases} DTW(i, j-1) \\ DTW(i-1, j) \\ DTW(i-1, j-1) \end{cases}$$

Here, $d(i,j)$ is the square Euclidean distance of x_i and y_j.

3 The Proposed Algorithm

The proposed algorithm works as follows:

1. **Feature extraction:** Given a depth image, 20 joints of the human body can be tracked by the skeleton tracker. Instead of using the positions of joints, we employ the relative position of each joint to the torso at each frame, as more discriminative and intuitive 3D joint features.
2. **Compute non-PSD kernels:** we compute the pairwise distance of each normalized 3D trajectory to other trajectories, using the derivative of DTW, as described in following subsections.
3. **Classification:** we train the ppfSVM using the computed kernel and evaluate the model on unseen test samples.

3.1 Kernels from Pairwise Data

According to [5], it is assumed that instead of a standard kernel function, all that is available is a proximity function, $P : \mathcal{X} \times \mathcal{X} \mapsto R$. No restrictions are placed on the function P, not symmetry nor even continuity. The mapping $\varPhi(x)$ is defined by:

$$\varPhi(x) : x \mapsto (P(x, x_1), P(x, x_2), \ldots, P(x, x_m))^T \tag{5}$$

where $x_i, i = 1, \ldots, m$ are the examples in dataset. Here, we represent each sample x_i by $x_i = \varPhi_m(x_i)$ i.e. an m-dimensional vector containing proximities to all other samples in the dataset. Let P denote the $m \times m$ matrix with entries $P(x_i, x_j), i, j = 1, \ldots, m$. Using the linear kernel on this data representation, the resulting kernel matrix becomes $K = PP^T$. In this case the decision rule (3) simplifies to

$$f(x) = sign\left(\sum_{i=1}^{m} y_i \alpha_i P \varPhi_m(x) + b\right) \tag{6}$$

All elements of $\varPhi_m(x_i)$ must be computed when classifying a point x.

3.2 Kernel Using Derivative of Dynamic Time Warping Distance

Despite the success of time series dis(similarity) measures they may fail in some situations. For example, since the DTW algorithm aims to explain variability in the Y-axis by warping the X-axis, it may results in unintuitive alignments where a single point on one sequence maps onto a large subsection of the other sequence; which is referred to as *"singularity"* in the related literature [8]. Also, they may fail to find obvious, natural alignments of two time series simply because a feature (i.e. peak, valley, inflection point, plateau etc.) in one series is slightly higher or lower than its corresponding feature in the other time series.

To deal with such problems, the derivatives version of DTW are employed in this work in order to consider the higher level features. This modified version is named *Derivative DTW (DDTW)* as defined as following:

$$DDTW(x, y) = DTW(\nabla x, \nabla y) \tag{7}$$

where ∇x and ∇y are the first discrete derivatives of x and y.

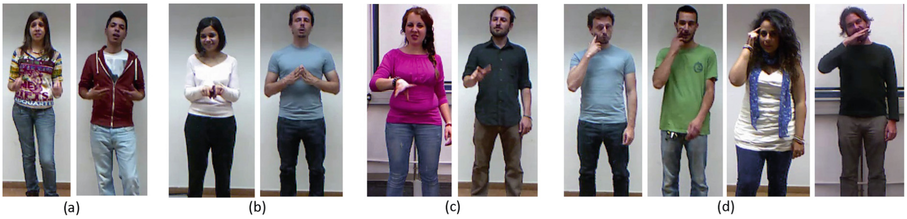

Fig. 2. Some example gestures in the Chaleran dataset are very easy to be confused, even from human visual perception. (a) *Che vuoi* vs. *Che due palle*. For the *Che vuoi* gesture, both hands are in front of the chest area; where for *Che due palle* gesture they are near the waist region. (b) *Vanno d'accordo* vs. *Cos hai combinato*: both hand positions are very close and with the same motion directions; (c) both gestures, *Si sono messid'accordo* and *non ce ne piu*, require hand rotations; (d) four gestures, *Furbo, seipazzo, buonissimo*, and *cosatifarei* are required with the finger pointing to the head area, which cannot be easily determined, even with human eyes.

4 Experimental Evaluation

Here, we present the experimental details of evaluation, including the datasets used, settings of the experiments, as well as the obtained results. The codes were implemented in C/C++ with an interface in Matlab and is available upon request.

4.1 Datasets

We evaluated our framework on two publicly available datasets: the Multi-modal Gesture Recognition Challenge 2013 (Chalearn) and MSR Action3D.

Chalearn Dataset: This dataset is a newly released large video database of 13,858 gestures from a lexicon of 20 Italian gesture categories recorded with a Kinect camera, including audio, skeletal model, user mask, RGB and depth images [4]. It contains image sequences capturing 27 subjects performing natural communicative gestures and speaking in fluent Italian, and is divided into development, validation and test parts. We conducted our experiments on the depth images of development and validation samples which contains 11,116 gestures across over 680 depth sequences. Each sequence lasts between 1 and 2 min and contains between 8 and 20 gesture samples, around 1,800 frames. Some examples of RGB images are shown in Fig. 2.

MSRAction3D Dataset: This dataset [11] is a well-known benchmark dataset for 3D action recognition. This dataset contains 20 actions, including *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pick up & throw*. Each action was performed 2 or 3 times by each subject. Skeleton joint data of each frame is available having a variety of motions related to arms, legs, torso, and their combinations. In total, there are 567 depth map sequences with a resolution of 320 × 240.

4.2 Classification Results

For Chalearn dataset, the classification performance is obtained by means of stratified 5-fold cross-validation. For MSR Action3D dataset, many studies follow the experimental setting of Li et al. [11], such that they first divide the 20 actions into three subsets, each having 8 actions. For each subset, they perform three tests. In test one and two, 1/3 and 2/3 of the samples were used as training samples and the rest as testing samples. In the third test, half of the subjects are used as training and the rest subjects as testing. The experimental results on the first two tests are generally very promising, mainly more than 90 % accuracy. On the third test, however, the recognition performance dramatically decreases. It shows that many of these methods do not have good generalization ability when a different subject is performing the action, even in the same environmental settings. In order to have more reliable results, we followed the same experimental setup of [15,25]. In this setting, actors 1,3,5,7, and 9 are used for training and the rest for testing.

The summaries of the results are reported in Table 1 for Chalearn and MSRAction3D datasets. In these tables, accuracies of traditional k-NN-based techniques using DDTW distance measures along with the corresponding accuracies using ppfSVMs are reported. It is important to note the outperformance of the results in comparison with the traditional kNN-based classifiers. The result are quite promising, considering the facts that the skeleton tracker sometimes fails and the tracked joint positions are quite noisy. In addition, the confusion

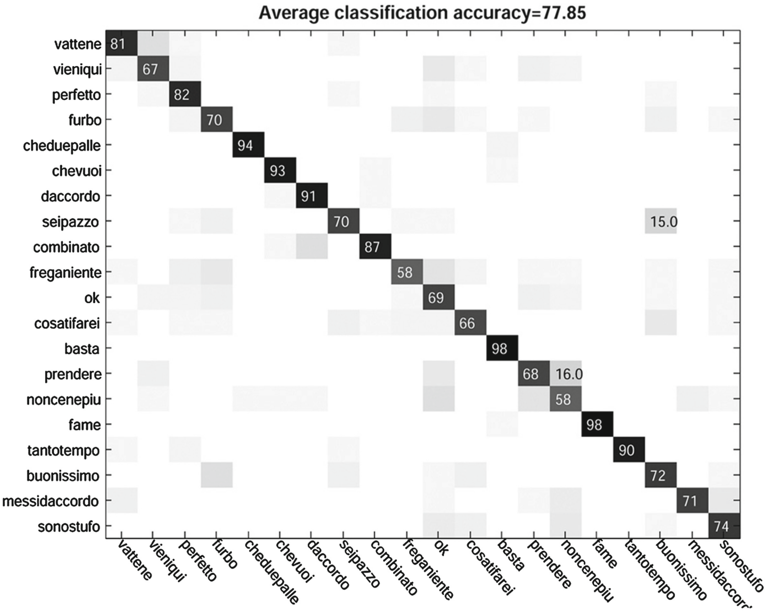

Fig. 3. Confusion matrices of the proposed technique on the Chalearn dataset.

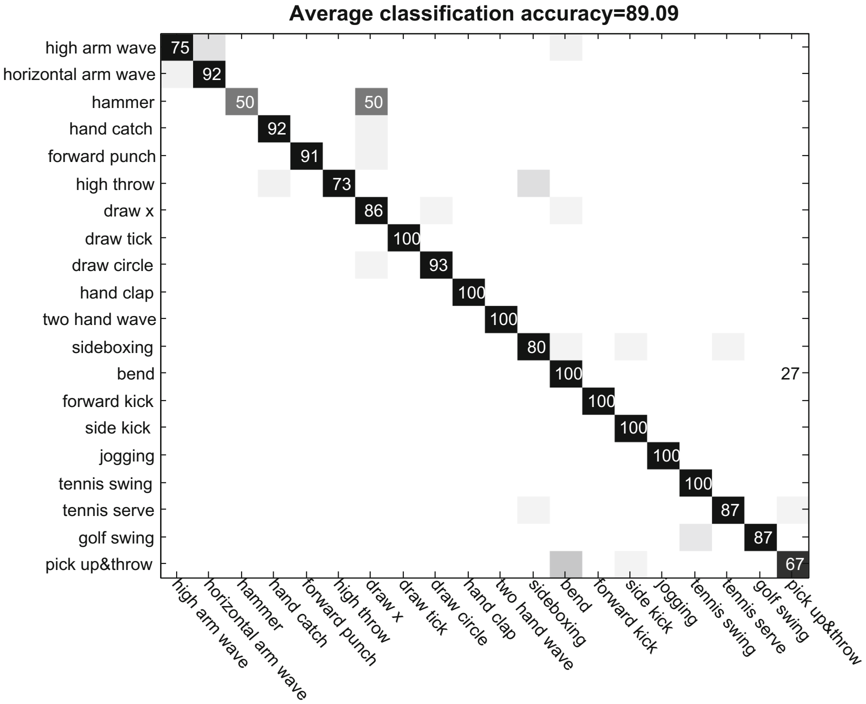

Fig. 4. Confusion matrices of the proposed technique on the MSRAction3D dataset.

Table 1. Classification accuracy of different learners on the Chalearn and MSRAction3D datasets.

	MSRAction3D dataset	Chalearn dataset
K-NN	80.12	68.90
ppfSVM	89.09	77.85

matrix of the proposed classification algorithm for these datasets are demonstrated in Figs. 3 and 4. It is important to note the outperformance of the results in comparison with the traditional kNN-based classifiers. The result are quite promising, considering the facts that the skeleton tracker sometimes fails and the tracked joint positions are quite noisy.

We then compare our classification results on MSRAction3D dataset with state-of-the-art methods. Table 2 shows the accuracy of our method and the rival methods on this dataset based on the cross-subject test setting [11]. Most studies use depth data in addition to skeleton joint information. However, processing sequences of depth images is much more computationally intensive.

The results provided in Table 2 along with the confusion matrix depicted in Figs. 3 and 4 demonstrate the superiority of the proposed methodology.

Table 2. Comparing classification accuracy of our method with the state-of-the-art methods on the MSRAction3D dataset.

Method	Accuracy
Studies employed depth data	
Action Graph [11]	74.70
HON4D [15]	85.85
Vieira et al. [22]	78.20
Random Occupancy Patterns [24]	86.50
DMM-LBP-FF [2]	87.90
Studies employed only skeleton data	
Actionlet Ensemble [26]	88.20
Histogram of 3D Joint [27]	78.97
GB-RBM & HMM [14]	80.20
Ensemble classification [1]	84.85
Proposed method	89.09

By only considering the skeleton data, our results achieved the accuracies of many works based on depth data Considering the fact that we have only employed the skeleton data, not depth sequences, the results are promising. The confusion matrix also reveals that almost all classes, except the "Hammer" class have been classified very well. This is due to fact that skeleton tracker sometimes fails and the tracked joint positions are quite noisy.

5 Conclusion

In this paper, we tackled the problem of human action classification using the 3D trajectories of body joint positions over the time. To do that, we utilized the derivatives of two time series distance measures, including Dynamic Time Warping and Longest Common subsequences. However, instead of employing these general measures as a distance measure for k-NN, we transformed these measures using the pairwise proximity function in order to be used for powerful SVM classification algorithm. Comparing the recognition results of the proposed methods with state-of-the-art techniques on two action recognition datasets, showed significant performance improvements. Remarkably, we obtained 89 % accuracy on the well-known MSRAction3D dataset using only 3D trajectories of body joints obtained by Kinect.

References

1. Bagheri, M.A., Hu, G., Gao, Q., Escalera, S.: A framework of multi-classifier fusion for human action recognition. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 1260–1265. IEEE (2014)
2. Chen, C., Jafari, R., Kehtarnavaz, N.: Action recognition from depth sequences using depth motion maps-based local binary patterns. In: WACV, pp. 1092–1099 (2015)
3. Dollar, P., Rabaud, V., Cottrell, G., Belongie, S.: Behavior recognition via sparse spatio-temporal features. In: IEEE Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 65–72. IEEE (2005)
4. Escalera, S., Gonzlez, J., Bar X., Reyes, M., Lopes, O., Guyon, I., Athistos, V., Escalante, H.: Multi-modal gesture recognition challenge 2013: dataset and results. In: ICMI (2013)
5. Graepel, T., Herbrich, R., Bollmann-Sdorra, P., Obermayer, K.: Classification on pairwise proximity data. In: Advances in Neural Information Processing Systems, pp. 438–444 (1999)
6. Hernndez-Vela, A., Bautista, M.A., Perez-Sala, X., Ponce, V., Bar X., Pujol, O., Angulo, C., Escalera, S.: BoVDW: bag-of-visual-and-depth-words for gesture recognition. In: ICPR, pp. 449–452. IEEE (2012)
7. Johansson, G.: Visual perception of biological motion and a model for its analysis. Percept. Psychophy. **14**(2), 201–211 (1973)
8. Keogh, E.J., Pazzani, M.J.: Derivative dynamic time warping. SIAM (2001)
9. Laptev, I.: On space-time interest points. IJCV **64**(2–3), 107–123 (2005)
10. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2008)
11. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3D points. In: CVPR Workshop (CVPRW), pp. 9–14. IEEE (2010)
12. Liu, J., Kuipers, B., Savarese, S.: Recognizing human actions by attributes. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3337–3344. IEEE (2011)
13. Mangasarian, O.L.: Generalized support vector machines. In: Advances in Neural Information Processing Systems, pp. 135–146 (1999)
14. Nie, S., Ji, Q.: Capturing global and local dynamics for human action recognition. In: 2014 22nd International Conference on Pattern Recognition (ICPR), pp. 1946–1951. IEEE (2014)
15. Oreifej, O., Liu, Z., Redmond, W.: HON4D:: histogram of oriented 4D normals for activity recognition from depth sequences. In: IEEE Conference on Computer Vision and Pattern Recognition (2013)
16. Reyes, M., Dominguez, G., Escalera, S.: Feature weighting in dynamic timewarping for gesture recognition in depth data. In: CVPR Workshops (CVPRW), pp. 1182–1188. IEEE (2011)
17. Sakoe, H., Chiba, S.: Dynamic programming algorithm optimization for spoken word recognition. IEEE Trans. Acoust. Speech Sig. Process. **26**(1), 43–49 (1978)
18. Schölkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2002)
19. Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., Blake, A.: Real-time human pose recognition in parts from single depth images. In: IEEE Conference on Computer Vision and Pattern Recognition (2011)

20. Sung, J., Ponce, C., Selman, B., Saxena, A.: Unstructured human activity detection from RGBD images. In: ICRA, pp. 842–849. IEEE (2012)
21. Treisman, A., Schmidt, H.: Illusory conjunctions in the perception of objects. Cogn. Psychol. **14**(1), 107–141 (1982)
22. Vieira, A.W., Nascimento, E.R., Oliveira, G.L., Liu, Z., Campos, M.F.M.: STOP: space-time occupancy patterns for 3D action recognition from depth map sequences. In: Alvarez, L., Mejail, M., Gomez, L., Jacobo, J. (eds.) CIARP 2012. LNCS, vol. 7441, pp. 252–259. Springer, Heidelberg (2012)
23. Wang, H., Klaser, A., Schmid, C., Liu, C.L.: Action recognition by dense trajectories. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 3169–3176. IEEE (2011)
24. Wang, J., Liu, Z., Chorowski, J., Chen, Z., Wu, Y.: Robust 3D action recognition with random occupancy patterns. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part II. LNCS, vol. 7573, pp. 872–885. Springer, Heidelberg (2012)
25. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Mining actionlet ensemble for action recognition with depth cameras. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 1290–1297. IEEE (2012)
26. Wang, J., Liu, Z., Wu, Y., Yuan, J.: Learning actionlet ensemble for 3D human action recognition. PAMI **36**(5), 914–927 (2014)
27. Xia, L., Chen, C.C., Aggarwal, J.: View invariant human action recognition using histograms of 3D joints. In: CVPR Workshops (CVPRW), pp. 20–27. IEEE (2012)
28. Yang, X., Tian, Y.: Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In: CVPR Workshops (CVPRW), pp. 14–19. IEEE (2012)
29. Yang, X., Tian, Y.L.: Action recognition using super sparse coding vector with spatio-temporal awareness. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part II. LNCS, vol. 8690, pp. 727–741. Springer, Heidelberg (2014)