inventions

MDPI

Article

Skeleton-Based Human Action Recognition through Third-Order Tensor Representation and Spatio-Temporal Analysis

Panagiotis Barmpoutis [1,*], Tania Stathaki [1] and Stephanos Camarinopoulos [2]

[1] Department of Electrical and Electronic Engineering, Faculty of Engineering, Imperial College London, London SW7 2AZ, UK; t.stathaki@imperial.ac.uk

[2] RISA Sicherheitsanalysen, GmbH, 10707 Berlin, Germany; s.camarinopoulos@risa.de

* Correspondence: p.barmpoutis@imperial.ac.uk

Received: 17 December 2018; Accepted: 3 February 2019; Published: 8 February 2019

Abstract: Given the broad range of applications from video surveillance to human–computer interaction, human action learning and recognition analysis based on 3D skeleton data are currently a popular area of research. In this paper, we propose a method for action recognition using depth sensors and representing the skeleton time series sequences as higher-order sparse structure tensors to exploit the dependencies among skeleton joints and to overcome the limitations of methods that use joint coordinates as input signals. To this end, we estimate their decompositions based on randomized subspace iteration that enables the computation of singular values and vectors of large sparse matrices with high accuracy. Specifically, we attempt to extract different feature representations containing spatio-temporal complementary information and extracting the mode-n singular values with regards to the correlations of skeleton joints. Then, the extracted features are combined using discriminant correlation analysis, and a neural network is used to recognize the action patterns. The experimental results presented use three widely used action datasets and confirm the great potential of the proposed action learning and recognition method.

Keywords: human action recognition; higher-order decomposition; discriminant component analysis; pattern recognition

1. Introduction

Human action recognition has been an active research topic due to its wide range of applications, including surveillance, healthcare, safety, transportation, human–computer interactions and response prediction [1,2]. Furthermore, with the continuous development of cost-effective RGB (Red–Green–Blue) [3] and depth cameras [4], inertial sensors [5], and algorithms for real-time pose estimation, human action recognition receives growing attention nowadays. Comparing these types of capturing sensors, RGB cameras provide rich texture information but are sensitive to illumination changes. Otherwise, depth sensors provide 3D structural information of the scene but are sensitive to materials with different reflection properties while inertial sensors can work in an unconfined environment but are sensitive to the sensor location on the body [6]. Although there are many different benefits from the use of all these sensors in numerous applications, access to 3D information and skeleton data brings unique advantages including robustness in action and gesture recognition. Specifically, through 3D skeletons, the set of connected body-joints that evolve in time can effectively be used for the representation and analysis of human behaviors.

To date (we present approaches that are most related to ours however, a comprehensive review of skeleton-based action recognition methodologies can be found in [1]), most of the skeleton-based literature approaches [1] consider human action recognition as a time series problem in which the

input observations are the 3D locations of the major body joints at each frame. Thus, characteristics of body postures and their dynamics over time are extracted to represent a human action. One of the most used approaches for modeling of time-evolving data and specifically of human actions is the Hidden Markov Model (HMM) [7–9], a graphical oriented method to characterize real-world observations in terms of state models. Single or multiple, they are often employed either for hand gesture recognition [10] or human action recognition [11]. Furthermore, Kosmopoulos et al. [12] employed a Bayesian filter supported by hidden Markov models and used user's feedback in the learning process to achieve online recognition.

Moreover, to model the temporal dynamics for action recognition, Xia et al. [13] extracted histograms of 3D joint locations and used discrete HMMs. Another widely used technique is the Conditional Random Field (CRF) model [14], which is an undirected graphical method that allows the dependencies between observations and the use of incomplete information about the probability distribution of a certain observable. However, these methods are incapable of identifying the representative patterns or modeling the structure of the data, thus, are lacking in discriminative power [1]. Furthermore, CRF is a highly computationally complex model at the training stage of the algorithm, making it difficult for researchers to re-train the model.

To overcome the above limitations and since not all poses in video sequences of an action are informative for the classification of that action, researchers focus on the identification of key poses and localization of the action in an unsegmented stream of frames. Thus, Zhou et al. [15] proposed the extraction of discriminative key poses represented by normalized joint locations, velocities and accelerations of skeleton joints, while Sharaf et al. [16] extracted features based on the probability distribution of skeleton data and employed a pyramid of covariance matrices and mean vectors to encode the relationship between these features. To facilitate the recognition task, Meshry et al. [17] encoded the position and kinematic information of skeleton joints proposing gesturelets, while Patrona et al. [18] extended gesturelets by adding automatic feature weighting at frame level and employing kinetic energy to identify the most representative action poses. However, these methods often lead to the selection of unneeded key poses or the omission of salient points containing important action information.

Recently, deep learning networks have been employed in the automated classification of human actions, aiming to overcome the extraction of hand-crafted features and the discrimination limitations of previous methods through effective deep architectures. For action recognition, Du et al. [19] proposed a hierarchical recurrent neural network for skeleton-based action recognition. They divided the human skeleton into five parts according to the human physical structure, and then separately fed them to five subnets. Furthermore, Hou et al. [20], encoded the spatio-temporal information of a skeleton sequence into color texture images and employed convolutional neural networks to learn dynamic features for action recognition. Bilen et al. [21] introduced dynamic images in combination with convolution networks and Chen et al. [22] combined deep convolution neural networks with CRFs, to achieve action recognition and improved image segmentation.

Additionally, Recurrent Neural Networks (RNN) and Long Short-Term Memory (LSTM) have been proposed to model temporal relationships among frames [23–25]. However, RNNs are incapable of capturing long-term dependencies between frames and skeletal joints in the spatial domain [26]. Liu et al. [26] used a spatio-temporal long short-term memory (ST-LSTM) network, to model the dynamics and dependency relations in both temporal and spatial domains. More recently, Konstantinidis et al. [27] proposed a four stream LSTM neural network based on two types of spatial skeletal features and their corresponding temporal representations extracted by the Grassmannian Pyramid Descriptor. However, the training of these complex deep learning networks requires the creation of large datasets for the accurate definition of their parameters. Thus, although high accuracies have been achieved, these methods need large labeled datasets and incur high time cost for training [28].

Significantly fewer works focus on action and gesture recognition adopting third-order tensor representations and modeling. Kim et al. [29] formed third-order tensors and applied tensor canonical correlation analysis to perform action and gesture classification in videos while Vasilescu

and Terzopoulos [30] used higher-order tensors and their decompositions for face recognition. Koniusz, et al. [31] defined kernels on 3D joint sequences, which are linearized to form kernel descriptors. Then, tensor representations were formed from these kernels and were used for action recognition. More recently, Dimitropoulos et al. [32,33] took advantage of the correlation between the different channels of data and proposed a stabilized higher order linear dynamical system to extract appearance information and dynamics of multidimensional patches. However, the last method requires high computational cost making real-time action recognition nearly impossible. Furthermore, one of the challenges in these methods is how data could be effectively represented and fed to the developed classification systems.

In this paper, we aim to address the problem of human action recognition through the encoding of the spatio-temporal information of a skeleton sequence and forming novel higher-order sparse tensors which describe 3D space relationship of joints. Through the decomposition of the tensor, we aim to overcome challenges, such as subject variations in performing actions and recognition in an unsegmented stream of frames, through the exploitation of the higher-order and hidden correlations between 3D coordinates of the body-joints and their temporal variations. Furthermore, we factor out the need for computationally costly operations, aiming to support daily living smart appliances and essential services offered to their end-users.

More specifically, this paper makes the following contributions: (a) We introduce a novel modeling of the skeleton time series sequences as higher-order sparse tensors to capture the dependencies among skeleton joints. Towards this end, a given sequence of the skeleton is represented by a third-order sparse tensor in which 3D coordinates corresponding to the joints are equal to one. (b) We adopt the higher order singular value decomposition of the formed tensor to exploit the correlations between body-joints. (c) We propose the extraction of a spatial and a temporal descriptor that is able to encode the global shape of action performers and motion of an action respectively. (d) We propose the fusion of the two descriptors adopting the discriminant correlation analysis (DCA). The proposed fused feature descriptor exploits the possibility of using third-order representations and their decompositions for 3D action recognition, which poses a different set of challenges. The final classification towards action recognition of the given sequence is obtained through the use of an artificial neural network. The proposed method improves the performance of action recognition, in terms of both positive detection rates (improving the average classification accuracy by 0.69%) and computational-time cost (achieving average classification in 0.42 s) making the proposed approach suitable for real-time applications. Finally, it can be combined with other local or global models; however, the goal of this paper is not to propose an ad-hoc algorithm, but an automated, fast, and robust approach through the sparse tensor modeling and its analysis for the support of the daily living action recognition applications.

The remainder of this paper is organized as follows: The next section presents the materials and methods used for the evaluation of the proposed methodology and for the automated human action recognition. Subsequently, experimental results are presented and discussed, while finally conclusions are drawn in the last section.

2. Materials and Methods

Skeleton sequences reflect the connectivity and topology of skeletal joints allowing the exploitation of both spatial time-evolving inter-correlated and intra-correlated patterns that are unique for each human action. Thus, for the modeling of different human actions, the creation of structures that will efficiently represent the spatial and temporal joints correlations play a key role. In the proposed methodology (Figure 1), to exploit the spatial time-evolving dependencies among skeleton joints, we represent the skeleton time series sequences as high-order sparse structure tensors. This allows the estimation of two descriptors with regards to spatio-temporal complementary information to exploit the intra-correlations of skeleton joints (Figure 2). For the estimation of the first descriptor, each skeleton frame is considered as a sparse binary tensor for which the elements that correspond to 3D skeletal joints' coordinates are equal to one. In the second case for each frame, the tensor elements that correspond to 3D joints' coordinates from the first frame to the examined frame

are set equal to one. The corresponding dimensions for each tensor are x, y, and h for the created tensor length, width, and height respectively. Then, to exploit the correlations between body-joints, we estimate their decompositions based on randomized subspace iteration. This enables the computation of the mode-n singular values of the sparse matrices with high accuracy. For the construction of feature vectors and to keep only the meaningful information and to reduce the complexity of data we use only the first ten singular values of each unfolded mode-n. The extracted features are combined using DCA. Finally, to recognize human action patterns a neural network is used.

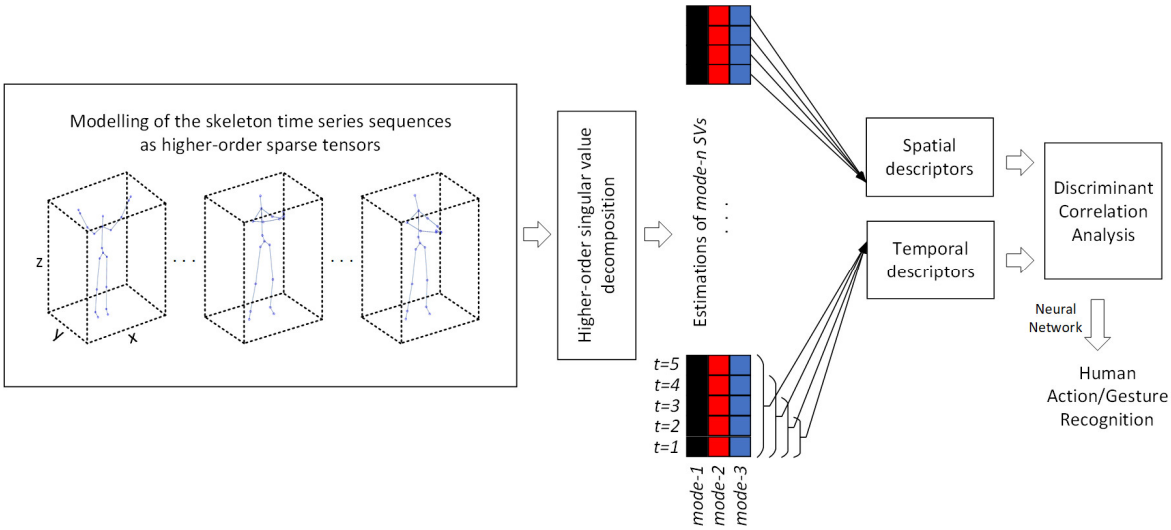


Figure 1. The proposed methodology.

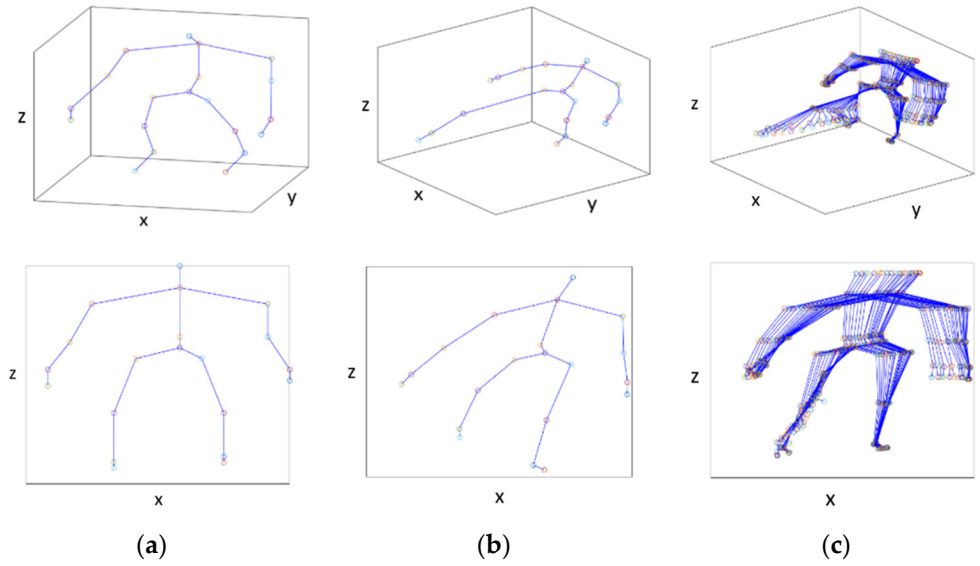


Figure 2. Tensor representation of skeleton and skeletal joints in 3D and 2D (xz-projection) graphs respectively for the extraction of spatial correlations (**a**), (**b**) and for the extraction of time-evolving correlations (**c**). Representation of skeletal joints (a) in the initial position, (**b**) in the position of right kick and (**c**) for the period from the initial position to right kick.

More specifically, we represent each frame as a sparse third-order tensor $Y \in \mathbb{R}^{x \times y \times h}$ (Table 1 contains the basic notations and their definitions). To compress the data, minimizing data loss, and significantly reduce its size we apply higher order singular value decomposition (Higher-Order SVD-HOSVD or Multi-Linear Singular Value Decomposition-MLSVD) using sequential truncation and a randomized SVD algorithm based on randomized subspace iteration [34]. Furthermore, this allows

the computation of singular values and vectors of large sparse matrices with high accuracy. Summarily, for each mode-n unfolding $Y_{(n)}$ of the sparse third-order tensor Y, we estimate a Gaussian random matrix Ω and we compute the $I_{(n)} = Y_{(n)}\Omega$ and its QR factorization so that it is $I_{(n)} = QR$. Then, we form the matrix $B_{(n)} = Q^T I_{(n)}$ and compute the corresponding singular value decomposition: $B_{(n)} = \widetilde{U}\Sigma V^T$. Thus, the orthonormal matrices of the higher-order singular value decomposition are described by the Equation $U_{(\mathrm{n})} = Q\widetilde{U}$. Therefore, the tensor Y is written as the multilinear tensor-matrix product of S and $U_{(\mathrm{n})}$.

$$Y = S \times_1 U_{(1)} \times_2 U_{(2)} \times_3 U_{(3)} \quad (1)$$

where, $S \in \mathbb{R}^{x \times \mathrm{y} \times h}$ is the core tensor, while $U_{(1)} \in \mathbb{R}^{x \times \mathrm{x}}$, $U_{(2)} \in \mathbb{R}^{y \times \mathrm{y}}$, and $U_{(3)} \in \mathbb{R}^{\mathrm{h} \times \mathrm{h}}$ are orthogonal matrices containing the orthonormal vectors spanning the column space of the matrix and with the operator $\times_j$ denoting the j-mode product between a tensor and a matrix.

Table 1. Basic notations and definitions.

Symbol	Definition
Y	observed frame data
S	core tensor
$Y_{(n)}$	mode-n unfolding
U	orthogonal matrix
$\sigma^{(n)}$	mode-n singular values
D_s	spatial representation of three mode MSVs
D_t	temporal representation of three mode MSVs
f	number of spatial and temporal descriptors for each human action sequence
D_S	feature vector (term frequency histogram) for spatial analysis
D_T	feature vector (term frequency histogram) for temporal analysis
ds_{ij}	spatial feature vector corresponding to the i^{th} class and in j^{th} sample
$\overline{ds_i}$	means of ds_{ij} vectors in the i^{th} class
$\overline{ds}$	means of the whole feature set
S_{bds}	between-class scatter matrix
P	matrix of orthogonal eigenvectors
$\widehat{\Lambda}$	diagonal matrix of real and non-negative eigenvalues
W_{bds}	transformation that unitizes S_{bds}
D_S'	feature vector (reduced dimensionality) for spatial analysis
D_T'	feature vector (reduced dimensionality) for temporal analysis
W_{DS}	transformation matrices for the spatial feature vectors
W_{DT}	transformation matrices for the temporal feature vectors
$\acute{D}_S$	transformed spatial feature
$\acute{D}_T$	transformed temporal feature

2.1. Modeling of Human Actions through Mode-n Singular Values

After obtaining the core tensor and the set of U matrices, we evaluated two modeling approaches for the extraction of a spatio-temporal features set. The key of a successful modeling is the automatic extraction of discriminative features. Thus, we calculate the mode-n singular values (MSVs) of core tensor unfoldings extracting the topological and the time-evolving properties of tensors as shown in Figure 3. These features can also be considered as artificial characteristics providing crucial measures [35,36]. Generally, the singular values are different for mode-n unfoldings but not completely independent.

Specifically, to model the spatial and the temporal inter-correlations and intra-correlations of skeletal joints in each frame, and by the assumption that spatio-temporal changes and action patterns are reflected by the sums of squared SVs, MSVs are computed using the core tensor S of each segment. Thus, the MSVs spatial descriptors, denoted as $\sigma^{(n)}$, are given by:

$$\sigma_k^{(1)} = \sqrt{\sum_{i=1}^{x}\sum_{j=1}^{y} s_{ijk}^2} \in \mathbb{R}^h \quad k = 1, 2, \ldots, h$$

$$\sigma_j^{(2)} = \sqrt{\sum_{i=1}^{x}\sum_{k=1}^{h} s_{ijk}^2} \in \mathbb{R}^x \quad j = 1, 2, \ldots, x \tag{2}$$

$$\sigma_i^{(3)} = \sqrt{\sum_{j=1}^{y}\sum_{k=1}^{h} s_{ijk}^2} \in \mathbb{R}^y \quad i = 1, 2, \ldots, y$$

where x, y, and h are the sizes of n-mode dimensions and s_{ijt} are the elements of the core tensor. As a mode-n SV descriptor corresponds to each sequence frame, a total number of f spatial and f temporal descriptors are produced for each human action sequence. In our experiments, to keep only the meaningful information and to reduce the complexity of data, for the construction of feature vectors we used the first ten singular values of each unfolded submatrix. Then, we concatenated the reduced three mode MSVs into vectors for spatial and temporal representation for each frame respectively, as follows:

$$D_s = \left[\sigma_1^{(1)}, \ldots, \sigma_k^{(1)}, \sigma_1^{(2)}, \ldots, \sigma_j^{(2)}, \ldots, \sigma_1^{(3)}, \ldots, \sigma_i^{(3)}\right] \in \mathcal{R}^{mxf} \tag{3}$$

$$D_t = \left[\sigma_1^{(1)}, \ldots, \sigma_k^{(1)}, \sigma_1^{(2)}, \ldots, \sigma_j^{(2)}, \ldots, \sigma_1^{(3)}, \ldots, \sigma_i^{(3)}\right] \in \mathcal{R}^{mxf} \tag{4}$$

where m is the size of the descriptors after concatenation and following the use of the first ten singular values of each unfolded submatrix. Following the concatenation, we need to create a spatial and a temporal feature that will represent each human action sequence. To this end, we estimate two histogram representations by defining two different codebooks. Thus, we apply the bag of systems approach and use the k-means clustering method for the collection of f spatial and f temporal descriptors. Each codebook consists of f codewords corresponding to the f representative sequence frames. Hence, the set of codewords encode all kinds of spatial and temporal patterns. Then, using Euclidean distance and the representative codewords, each human action is represented as a term frequency histogram (D_S and D_T) of the predefined codeword of the MSVs. These vectors may be considered to represent the distinctive classes of the human actions. Comparing the results of Figure 3 which show the variations of the SVs of each mode for the different actions, it is obvious that the SVs can be used as discriminating features for actions differentiation. Furthermore, we observe that singular values for mode-n unfoldings are not completely independent, but there are variations between neighboring singular values.

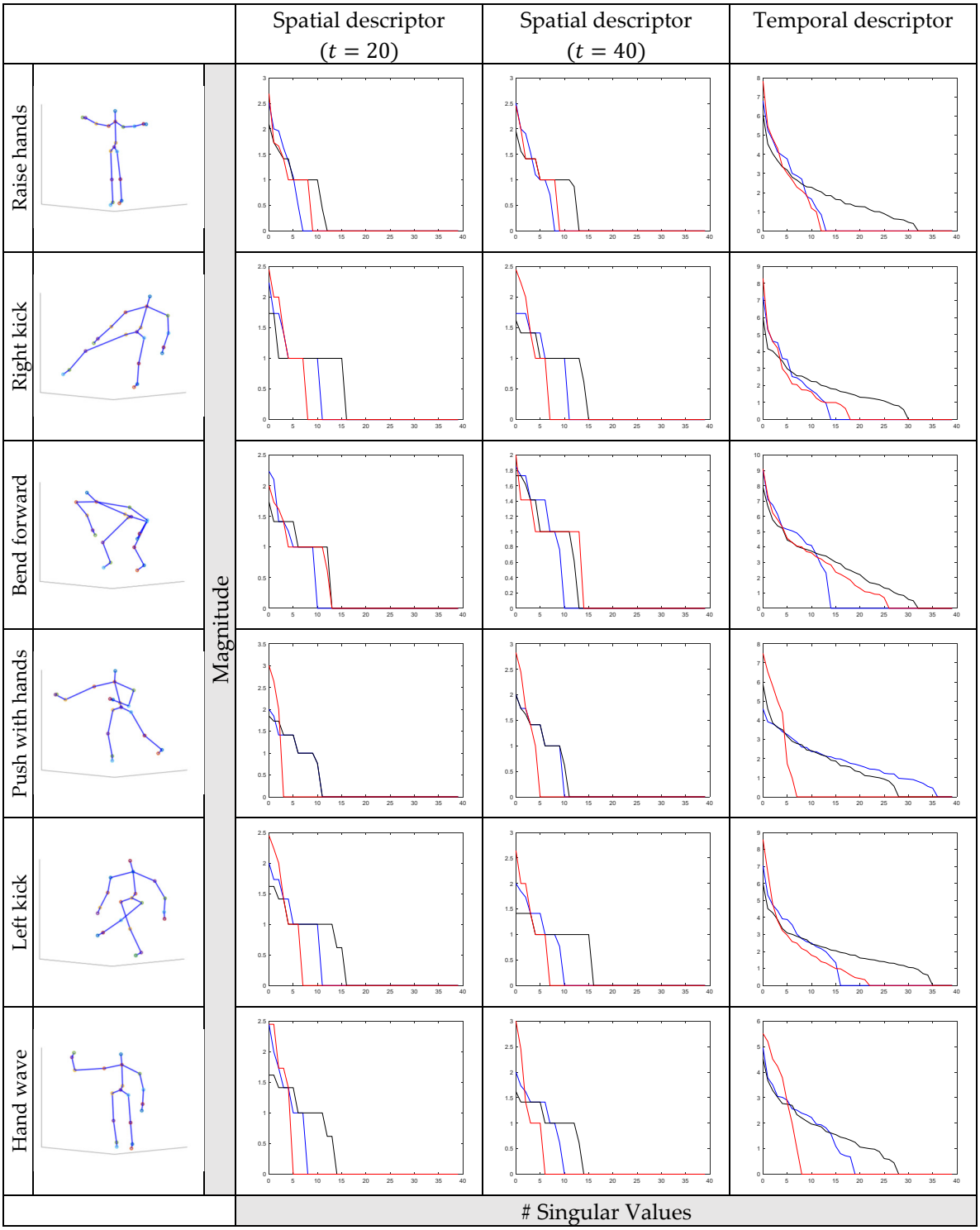


Figure 3. Distributions of three mode singular values (SVs) for the different human actions of CERTH [32] database. In the two first columns, the proposed spatial descriptors for $t = 20$ and $t = 40$ are shown. In the third column t,he proposed temporal descriptor is shown. The blue line corresponds to the $mode - 1$ SVs, the black line to the $mode - 2$ SVs and the red line to the $mode - 3$ SVs.

2.2. Feature Fusion through Discriminant Correlation Analysis and Classification

For the fusion of the extracted features, we adopted the DCA [37,38], a level fusion technique, that incorporates the class associations into the correlation analysis of the feature sets. Thus, we aim to maximize the pairwise correlations across the extracted spatial and temporal feature set. Simultaneously, we aim to eliminate the between-class correlations and to restrict the correlations to

be within the different human actions. The DCA is a low computational complexity method, and it can be employed in real-time applications meeting the requirements of daily living applications.

In our problem, D_S and D_T denote the feature vector of each action for spatial and temporal analysis and are collected from c human action classes. Assuming that $ds_{ij} \in D_S$ denote the spatial feature vector corresponding to the i^{th} class and in j^{th} sample, then, the $\overline{ds}_i$ and $\overline{ds}$ denote the means of ds_{ij} vectors in the i^{th} class and the whole feature set, respectively. Thus, the between-class scatter matrix is defined as:

$$S_{bds_{(m\times m)}} = \sum_{i=1}^{c} n_i\left(\overline{ds}_i - \overline{ds}\right)\left(\overline{ds}_i - \overline{ds}\right)^T = \Phi_{bds}\Phi_{bds}^T \tag{5}$$

where

$$\Phi_{bds_{(m\times c)}} = \left[\sqrt{n_1}\left(\overline{ds}_1 - \overline{ds}\right), \sqrt{n_2}\left(\overline{ds}_2 - \overline{ds}\right), \ldots, \sqrt{n_c}\left(\overline{ds}_c - \overline{ds}\right)\right] \tag{6}$$

Then, to estimate the most significant eigenvectors of the covariance matrix $\Phi_{bds}\Phi_{bds}^T$ or $\Phi_{bds}^T\Phi_{bds}$ (if the number of features is higher that the number of classes) we calculate the transformations that diagonalize it.

$$P^T(\Phi_{bds}^T\Phi_{bds})P = \widehat{\Lambda} \tag{7}$$

where the P is the matrix of orthogonal eigenvectors and $\widehat{\Lambda}$ is the diagonal matrix of real and non-negative eigenvalues sorted in decreasing order [38]. Thus, if Z consists of the r most significant eigenvectors then, the r most significant eigenvectors of S_{bds} can be obtained with the mapping: $Z \to \Phi_{bds}Z$.

$$(\Phi_{bds}Z)^T S_{bds}(\Phi_{bds}Z) = \Lambda_{(r\times r)} \tag{8}$$

Thus, the $W_{bds} = \Phi_{bds}Z\Lambda^{-1/2}$ is the transformation that unitizes S_{bds} and reduces the dimensionality of the data matrix, S_{bds}, from m to r. That is

$$W_{bds}^T S_{bds} W_{bds} = \mathrm{I} \tag{9}$$

The new feature vector D_S' is resulting from

$$D_S' = W_{bds}^T D_S \tag{10}$$

Similarly, to the above, we estimate also the D_T' for temporal analysis features D_T. Then, to make the features of the spatial feature set to have a nonzero correlation with their corresponding temporal feature sets, we diagonalize the between-set covariance matrix. To achieve this, we diagonalize the $S_{DSDT}' = D_S' D_T'^T$ applying singular value decomposition and estimating the Σ.

$$S_{DSDT}' = U\Sigma V^T \text{ and } U^T S_{DSDT}' V = \Sigma \tag{11}$$

Furthermore, to unitize the between-set covariance matrix, we set $W_{cDS} = U\Sigma^{-1/2}$ and $W_{cDT} = V\Sigma^{-1/2}$.

$$\left(U\Sigma^{-1/2}\right)^T S_{DSDT}'\left(V\Sigma^{-1/2}\right) = I \tag{12}$$

Thus, the features are transformed as follow:

$$\acute{D}_S = \mathrm{W}_{cDS}^T D_S' = \mathrm{W}_{cDS}^T \mathrm{W}_{bDS}^T D_S = W_{DS} D_S \tag{13}$$

$$\acute{D}_T = \mathrm{W}_{cDT}^T D_S' = \mathrm{W}_{cDT}^T \mathrm{W}_{bDT}^T D_T = W_{DT} D_T \tag{14}$$

where $W_{DS} = \mathrm{W}_{cDS}^T \mathrm{W}_{bDS}^T$ and $W_{DT} = \mathrm{W}_{cDT}^T \mathrm{W}_{bDT}^T$ are the transformation matrices for the extracted spatio-temporal feature vectors D_S and D_T, respectively.

Finally, for the classification of human actions, we used neural networks that have been proved [39] a useful tool for various applications which require extensive classification. Specifically, we used MATLAB's Neural Network Toolbox 11 and a two-layer feedforward network, with sigmoid transfer functions in both the hidden layer and the output layer. The number of hidden neurons was set to 10.

3. Results

In this section, we present a detailed experimental evaluation of the proposed methodology using three datasets. The goal of this experimental evaluation is three-fold: (a) Initially, we aim to define the number of MSVs that will be used for the evaluation, then (b) we want to show that the fusion of proposed descriptors improves the classification accuracy significantly, and finally (c) we intend to demonstrate the superiority of the proposed algorithm in human action recognition against a number of current state-of-the-art approaches.

To evaluate the proposed method we initially used the Centre for Research and Technology Hellas (CERTH) dataset [32], which contains a relative small number of actions, i.e., six different actions (bend forward, left kick, right kick, raise hands, hand wave, and push with hands), performed by six subjects, each repeated ten times (i.e., 360 actions in total). With regards to the CERTH dataset, we used 6 instances of each action per subject for training and 4 instances for testing. Furthermore, we used the well-known G3D dataset [40], containing a large range of gaming actions, i.e., 20 gaming actions (punch right, punch left, kick right, kick left, defend, golf swing, tennis swing forehand, tennis swing backhand, tennis serve, throw bowling ball, aim and fire gun, walk, run, jump, climb, crouch, steer a car, wave, flap, and clap), repeated three times by ten subjects (i.e., 600 actions in total). With regards to the G3D dataset, we used 1 instance of each action per subject used for training and 2 instances of each action for testing. Finally, we used the Microsoft Research Cambridge (MSRC)-12 Kinect Gesture Dataset [41] which consists of 12 gestures (kick, beat both, change weapon, had enough, throw, bow, shoot, wind it up, googles, punch right, duck and start system) performed by 30 people. In all experimental results, we split the datasets into training and testing subsets based on [32] and on [27]. Specifically, to have a fair comparison in CERTH and G3D datasets the training and testing sets based on [32] were used (i.e., for CERTH dataset, we used six instances of each action per subject for training and four instances for testing and for G3D dataset, one instance of each action per subject used for training and two instances of each action for testing). For the evaluation of MSRC-12 dataset, we followed two different protocols: initially, 37% of the dataset was used for training and 63% for testing based on [32] (Table 2—MSRC-12 [1]) and secondly a modality-based "leave-persons out" protocol was used as proposed in [27] (Table 2—MSRC-12 [2]). In the second case, for each of the instruction modalities of MSRC-12 dataset, the 'leave-persons-out' protocol was adopted, keeping the minimum subject subset containing all the gestures as the test set and employing all the others for training.

Table 2. Experimental comparison results for CERTH, G3D, and MSRC-12 datasets.

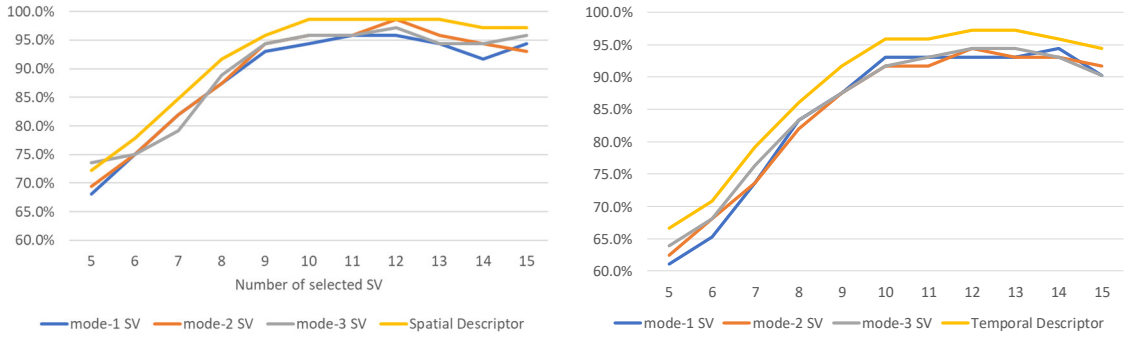
Method	CERTH	G3D	MSRC-12 [1]	MSRC-12 [2]
Dynamic Time Warping	87.5%	57%	48.12%	-
Hidden Markov Model	96.25%	77.4%	76.2%	-
Restricted Boltzmann Machine	97.1%	84%	79.8%	-
Conditional Random Fields	97.91%	69.25%	67.95%	-
Histograms of Grassmannian Points	98.61%	90.75%	80.15%	-
Multi-Scale Action Detection	-	-	-	63.9%
Bags of Gesturelets	-	-	-	87.1%
Extended Gesturelets	-	-	-	91.2%
LSTM and Grassmannian Pyramids	-	92.38%	-	94.6%
Proposed	100%	92.6%	83.1%	92.8%

[1] 37% of the MSRC-12 dataset was used for training and 63% for testing [32]. [2] Training and testing using "Leave-persons out" protocol [27].

3.1. Defining the Number of MSVs

The selection of the number of MSVs is based on the contribution of each mode-n singular value to the recognition results for the proposed spatial and temporal descriptors. Specifically, we compared the proposed method results using the first five to fifteen SVs. We used the CERTH dataset

to run the experiments. As one can easily see in Figure 4 using more than 10 SVs does not lead to any significant improvement in the recognition results. Furthermore, as was expected from observing Figure 3, the use of a low number of SVs results in low recognition rates. This is related to the first values of SVs that are similar to each other. Significant differences in values are observed between the eighth and the twelfth SVs where better results are achieved.



(**a**) (**b**)

Figure 4. Action recognition rates of (**a**) Spatial Descriptor and (**b**) Temporal Descriptor using the first five to fifteen SVs and applying the proposed method to the CERTH dataset.

3.2. Contribution of Different Feature Representations and Fusion Results

In this subsection, we elaborate a more detailed analysis to evaluate the contribution of different feature representations to the human action recognition. Specifically, we analyze the contribution of the two different descriptors, and we aim to show that the fusion of these features can improve the recognition rates. For the classification of the human actions, in the first case, we used the proposed spatial descriptor that achieves 98.6% and 90.5% for CERTH, G3D, and MSRC-12 datasets respectively. In the second case, we used the temporal descriptor that achieves lower accuracy rates than the spatial descriptor and achieves 95.8% and 89.4% for the two datasets. The performance of the proposed DCA method and the explanation of how the different descriptors contribute to classification rates are shown in Figure 5. It is clear that the fusion approach using the proposed spatial and temporal descriptors and a neural network for classification achieves 100%, 92.6%, and 83.1% classification rates for CERTH, G3D, and MSRC-12 datasets, respectively. The accuracy rates make evident that the individual feature representations contain complementary information and, therefore, the detection accuracy after fusion is increased.

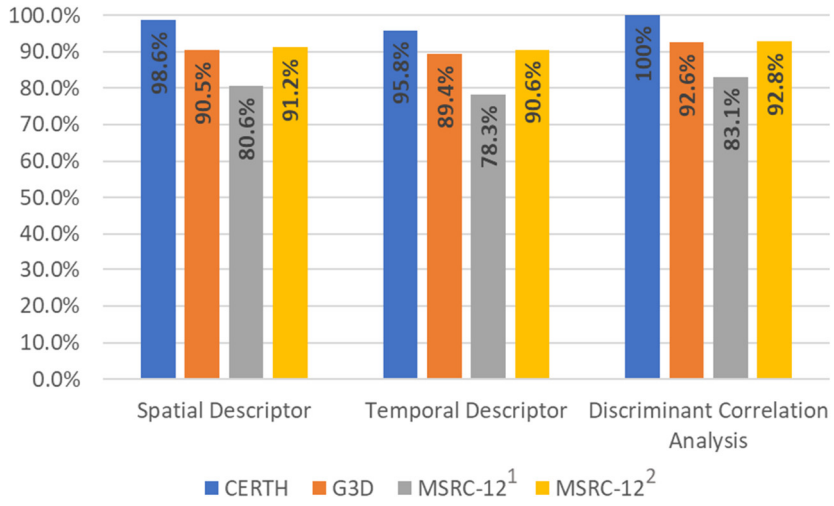


Figure 5. Contribution of different feature representations and fusion results through discriminant correlation analysis for CERTH, G3D, and MSRC-12 databases. [1] 37% of the MSRC-12 dataset was used for training and 63% for testing. [2] "Leave-persons out" protocol was used.

3.3. Comparison with State-of-the-Art Approaches

Finally, we evaluated the performance of the proposed methodology against other state-of-the-art techniques, and the comparison is shown in Table 2. Table 2 presents the classification rates of the proposed method against nine other state-of-the-art algorithms, i.e., dynamic time warping [42], HMMs [8], restricted Boltzmann machine [43], CRFs [14], Histograms of Grassmannian Points [33], pyramid of covariance matrices and mean vectors [16], Bags of Gesturelets [17], Extended Gesturelets [18], and LSTM and Grassmannian Pyramids [27], on the three datasets. To have a fair comparison, we used the joints coordinates as an input signal for all algorithms. Obviously then, for the CERTH and G3D datasets, our proposed new method outperforms previously presented action recognition approaches. More specifically, our methodology improves the state-of-the-art results by 1.39% and 0.22% for these datasets respectively. Moreover, from Table 2, it can be observed that our methodology outperforms the state-of-the-art methodologies in MSRC-12 dataset when the protocol based on [32] was followed. However, employing the modality-based "leave-persons out" protocol, the proposed approach shows slightly lower recognition rates from the LSTM and Grassmannian Pyramids method. It can be explained by the fact that in this method, a meta-learner step which takes advantage of the meta-knowledge is applied. However, as it is observed from the recognition results in G3D dataset, it does not work efficiently in the cases that training datasets are small. In contrast, the proposed method works better in these cases aiming to support efficiently daily living action recognition applications that use small training sets (e.g., safety and security in home appliances). Furthermore, the proposed method achieves action recognition in far less than half a second (the average time—in a Core i5, 8 GB RAM, 2 GB Internal Graphics Card—for the extraction of feature vector and classification was estimated in 0.42 s) making it suitable for real-time applications, which could not be achieved using deep learning techniques.

4. Discussion and Conclusions

In this paper, we presented a novel methodology for human action recognition. The main advantage of the proposed approach is that it exploits the spatio-temporal inter-correlations between skeleton joints in different actions by extracting feature representations that contain complementary information. More specifically, to exploit the spatial and time-evolving information as well as to better model human motion correlations, we use a third-order tensor structure, and then we extract different feature representations containing complementary information with regards to the spatial and temporal correlations of the signals. Subsequently, we extracted feature representations based on higher-order singular value decompositions and mode-n singular values. Furthermore, the experimental results in Figure 5 show that the combination of descriptors using DCA and a neural network significantly increase the detection rates of individual feature representations. We notice that the fusion approach provides excellent results in both CERTH, G3D, and MSRC-12 datasets.

The proposed method can significantly enhance the accuracy of human action recognition and support modern daily living through the action recognition services offered to end-users. Even when the number of classes increases, the discrimination ability of the method remains higher than other methods. This is mainly because we use intra-correlation information associated with the spatial distribution of joints, while we aim to combine it with the time-evolving information of the actions encoded in the descriptors.

However, in the MSRC-12 dataset, the discrimination ability of the extracted features is lower than the other datasets. This can be explained by the fact that there is a high intra-class variation of the dataset. To overcome this, a weighted approach for automatic adjustment of selected mode-n SVs could be employed. Furthermore, a meta-learner that would take the advantage of the meta knowledge would achieve better results. Furthermore, some misclassifications in the G3D dataset are explained by the fact that there are actions during which most of their human skeleton joints remain almost in the same position. Thus, the extracted features lose the high ability to distinguish these actions. To overcome this limitation in the future, we aim to create subsets of joints and to apply the proposed algorithm in the different subsets weighting the extracted features.

In the future, we aim to evaluate the proposed methodology in more databases (e.g., JHMDB [44] and in Human3.6M [45]) to assess the effectiveness of the proposed methodology. Based on the extracted results and the number of actions (21 and 17 for the two datasets respectively) we believe that the extracted features will have the ability to accurately discriminate the actions that are included in the datasets.

Author Contributions: Conceptualization, P.B., T.S. and S.C.; Methodology, P.B., T.S. and S.C.; Validation, P.B.; Formal Analysis, P.B.; Investigation, P.B., T.S. and S.C.; Writing—Original Draft Preparation, P.B.; Writing—Review & Editing, P.B., T.S. and S.C.; Visualization, P.B.

Funding: This research was funded by the EU H2020 TERPSICHORE project "Transforming Intangible Folkloric Performing Arts into Tangible Choreographic Digital Objects" under the grant agreement 691218.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Han, F.; Reily, B.; Hoff, W.; Zhang, H. Space-time representation of people based on 3D skeletal data: A review. *Comput. Vis. Image Underst.* **2017**, *158*, 85–105.
2. Lokare, N.; Zhong, B.; Lobaton, E. Activity-Aware Physiological Response Prediction Using Wearable Sensors. *Inventions* **2017**, *2*, 32.
3. Ramanathan, M.; Yau, W.Y.; Teoh, E.K. Human action recognition with video data: Research and evaluation challenges. *IEEE Trans. Hum. Mach. Syst.* **2014**, *44*, 650–663, doi:10.1109/THMS.2014.2325871.
4. Han, J.; Shao, L.; Xu, D.; Shotton, J. Enhanced computer vision with microsoft kinect sensor: A review. *IEEE Trans. Cybern.* **2013**, *43*, 1318–1334, doi:10.1109/TCYB.2013.2265378.
5. Ngo, T.T.; Makihara, Y.; Nagahara, H.; Mukaigawa, Y.; Yagi, Y. Similar gait action recognition using an inertial sensor. *Pattern Recognit.* **2015**, *48*, 1289–1301, doi:10.1016/j.patcog.2014.10.012.
6. Chen, C.; Jafari, R.; Kehtarnavaz, N. A survey of depth and inertial sensor fusion for human action recognition. *Multimed. Tools Appl.* **2017**, *76*, 4405–4425.
7. Kim, E.; Helal, S.; Cook, D. Human activity recognition and pattern discovery. *IEEE Pervasive Comput. IEEE Comput. Soc.* **2010**, *9*, 48.
8. Rabiner, L.R.; Juang, B.H. An introduction to hidden Markov models. *IEEE ASSP Mag.* **1986**, *3*, 4–16.
9. Hidden Markov Model. Available online: https://www.cs.ubc.ca/~murphyk/Software/HMM/hmm.html (accessed on 30 May 2016).
10. Liu, K.; Chen, C.; Jafari, R.; Kehtarnavaz, N. Fusion of inertial and depth sensor data for robust hand gesture recognition. *IEEE Sens. J.* **2014**, *14*, 898–1903, doi:10.1109/JSEN.2014.2306094.
11. Ofli, F.; Chaudhry, R.; Kurillo, G.; Vidal, R.; Bajcsy, R. Berkeley mhad: A comprehensive multimodal human action database. In Proceedings of the 2013 IEEE Workshop on Applications of Computer Vision (WACV), Tampa, FL, USA, 15–17 January 2013; pp. 53–60.
12. Kosmopoulos, D.I.; Doulamis, N.D.; Voulodimos, A.S. Bayesian filter-based behavior recognition in workflows allowing for user feedback. *Comput. Vis. Image Underst.* **2012**, *116*, 422–434.
13. Xia, L.; Chen, C.C.; Aggarwal, J.K. View invariant human action recognition using histograms of 3D joints. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, 16–21 June 2012; pp. 20–27.
14. Conditional Random Field. Available online: https://www.cs.ubc.ca/~murphyk/Software/CRF/crf.html (accessed on 30 May 2016).
15. Zhou, L.; Li, W.; Zhang, Y.; Ogunbona, P.; Nguyen, D.T.; Zhang, H. Discriminative key pose extraction using extended lc-ksvd for action recognition. In Proceedings of the 2014 International Conference on Digital Lmage Computing: Techniques and Applications (DlCTA), Wollongong, Australia, 25–27 November 2014; pp. 1–8.
16. Sharaf, A.; Torki, M.; Hussein; M. E.; El-Saban, M. Real-time multi-scale action detection from 3D skeleton data. In Proceedings of the 2015 IEEE Winter Conference on Applications of Computer Vision (WACV), Waikoloa, HI, USA, 5–9 January 2015; pp. 998–1005.

17. Meshry, M.; Hussein, M.E.; Torki, M. Linear-time online action detection from 3D skeletal data using bags of gesturelets. In Proceedings of the 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, 7–9 March 2016; pp. 1–9.
18. Patrona, F.; Chatzitofis, A.; Zarpalas, D.; Daras, P. Motion analysis: Action detection, recognition and evaluation based on motion capture data. *Pattern Recognit.* **2018**, *76*, 612–622, doi:10.1016/j.patcog.2017.12.007.
19. Du, Y.; Wang, W.; Wang, L. Hierarchical recurrent neural network for skeleton based action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 1110–1118.
20. Hou, Y.; Li, Z.; Wang, P.; Li, W. Skeleton optical spectra-based action recognition using convolutional neural networks. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 807–811, doi:10.1109/TCSVT.2016.2628339.
21. Bilen, H.; Fernando, B.; Gavves, E.; Vedaldi, A. Action recognition with dynamic image networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *40*, 12, doi:10.1109/TPAMI.2017.2769085.
22. Chen, L.C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A.L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 834–848, doi:10.1109/TPAMI.2017.2699184.
23. Jain, A.; Zamir; A. R.; Savarese, S.; Saxena, A. Structural-RNN: Deep learning on spatio-temporal graphs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 5308–5317.
24. Shi, Z.; Kim, T.K. Learning and refining of privileged information-based RNNs for action recognition from depth sequences. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017.
25. Zhu, W.; Lan, C.; Xing, J.; Zeng, W.; Li, Y.; Shen, L.; Xie, X. Co-Occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, Phoenix, Arizona, 12–17 February 2016.
26. Liu, J.; Shahroudy, A.; Xu, D.; Kot, A.C.; Wang, G. Skeleton-based action recognition using spatio-temporal LSTM network with trust gates. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *40*, 3007–3021, doi:10.1109/TPAMI.2017.2771306.
27. Konstantinidis, D.; Dimitropoulos, K.; Daras, P. Skeleton-based action recognition based on deep learning and Grassmannian pyramids. In Proceedings of the 2018 26th European Signal Processing Conference, Rome, Italy, 3–7 September 2018; pp. 2045–2049, doi:10.23919/EUSIPCO.2018.8553163.
28. Ke, Q.; Bennamoun, M.; An, S.; Sohel, F.; Boussaid, F. A new representation of skeleton sequences for 3D action recognition. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 4570–4579.
29. Kim, T.K.; Wong, S.F.; Cipolla, R. Tensor canonical correlation analysis for action classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 17–22 June 2007; pp. 1–8.
30. Vasilescu, M.A.O.; Terzopoulos, D. Multilinear analysis of image ensembles: Tensorfaces. In Proceedings of the 7th European Conference on Computer Vision, Copenhagen, Denmark, 28–31 May 2002; pp. 447–460.
31. Koniusz, P.; Cherian, A.; Porikli, F. Tensor representations via kernel linearization for action recognition from 3D skeletons. In Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016; pp. 37–53.
32. Dimitropoulos, K.; Barmpoutis, P.; Kitsikidis, A.; Grammalidis, N. Classification of multidimensional time-evolving data using histograms of Grassmannian points. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *28*, 892–905.
33. Dimitropoulos, K.; Barmpoutis, P.; Kitsikidis, A.; Grammalidis, N. Extracting Dynamics from Multi-dimensional Time-evolving Data using a Bag of Higher-order Linear Dynamical Systems. In Proceedings of the International Conference on Computer Vision Theory and Applications, Rome, Italy, 27–29 February 2016; pp. 683–688.
34. Halko, N.; Martinsson, P.G.; Tropp, J.A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **2011**, *53*, 217–288, doi:10.1137/090771806.

35. Hackbusch, W.; Uschmajew, A. On the interconnection between the higher-order singular values of real tensors. *Numer. Math.* **2017**, *135*, 875–894, doi:10.1007/s00211-016-0819-9.
36. Padhy, S.; Dandapat, S. Third-order tensor based analysis of multilead ECG for classification of myocardial infarction. *Biomed. Signal Proc. Control* **2017**, *31*, 71–78, doi:10.1016/j.bspc.2016.07.007.
37. Haghighat, M.; Abdel-Mottaleb, M.; Alhalabi, W. Discriminant correlation analysis for feature level fusion with application to multimodal biometrics. In Proceedings of the 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 20–25 March 2016; pp. 1866–1870.
38. Haghighat, M.; Abdel-Mottaleb, M.; Alhalabi, W. Discriminant correlation analysis: Real-time feature level fusion for multimodal biometric recognition. *IEEE Trans. Inf. Forensics Secur.* **2016**, *11*, 1984–1996, doi:10.1109/TIFS.2016.2569061.
39. Oniga, S.; Suto, J. Human activity recognition using neural networks. In Proceedings of the 2014 15th International Carpathian Control Conference (ICCC), Velke Karlovice, Czech Republic, 28–30 May 2014; pp. 403–406.
40. Bloom, V.; Makris, D.; Argyriou, V. G3D: A gaming action dataset and real time action recognition evaluation framework. In Proceedings of the 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Providence, RI, USA, 16–21 June 2012; pp. 7–12.
41. Microsoft Research Cambridge-12 Kinect Gesture Data Set. Available online: https://www.microsoft.com/en-us/download/details.aspx?id=52283 (accessed on 14 January 2019).
42. ten Holt, G.A.; Reinders, M.J.; Hendriks, E.A. Multi-dimensional dynamic time warping for gesture recognition. In Proceedings of the Thirteenth Annual Conference of the Advanced School for Computing and Imaging, Heijen, The Netherlands, 13–15 June 2007; Volume 300, p. 1.
43. Deep Neural Network. Available online: http://www.mathworks.com/matlabcentral/fileexchange/42853-deep-neural-network (accessed on 30 May 2016).
44. Jhuang, H.; Gall, J.; Zuffi, S.; Schmid, C.; Black, M.J. Towards understanding action recognition. In Proceedings of the IEEE International Conference on Computer Vision, Sydney, Australia, 1–8 December 2013; pp. 3192–3199.
45. Ionescu, C.; Papava, D.; Olaru, V.; Sminchisescu, C. Human3.6M: Large scale datasets and predictive methods for 3D human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.* **2014**, *36*, 1325–1339.