International Journal of Neural Systems, Vol. 33, No. 1 (2023) 2350002 (20 pages)
© World Scientific Publishing Company
DOI: 10.1142/S0129065723500028

A Multimodal Fusion Approach for Human Activity Recognition

Dimitrios Koutrintzes
Institute of Informatics and Telecommunications
National Center for Scientific Research — "Demokritos"
Athens, Greece
dkoutrintzes@gmail.com

Evaggelos Spyrou
Department of Informatics and Telecommunication
University of Thessaly, Lamia, Greece
espyrou@uth.gr

Eirini Mathe[*] and Phivos Mylonas[†]
Department of Informatics
Ionian University
Corfu, Greece
[*]*c17math@ionio.gr*
[†]*fmylonas@ionio.gr*

Accepted 26 September 2022
Published Online 27 December 2022

The problem of human activity recognition (HAR) has been increasingly attracting the efforts of the research community, having several applications. It consists of recognizing human motion and/or behavior within a given image or a video sequence, using as input raw sensor measurements. In this paper, a multimodal approach addressing the task of video-based HAR is proposed. It is based on 3D visual data that are collected using an RGB + depth camera, resulting to both raw video and 3D skeletal sequences. These data are transformed into six different 2D image representations; four of them are in the spectral domain, another is a pseudo-colored image. The aforementioned representations are based on skeletal data. The last representation is a "dynamic" image which is actually an artificially created image that summarizes RGB data of the whole video sequence, in a visually comprehensible way. In order to classify a given activity video, first, all the aforementioned 2D images are extracted and then six trained convolutional neural networks are used so as to extract visual features. The latter are fused so as to form a single feature vector and are fed into a support vector machine for classification into human activities. For evaluation purposes, a challenging motion activity recognition dataset is used, while single-view, cross-view and cross-subject experiments are performed. Moreover, the proposed approach is compared to three other state-of-the-art methods, demonstrating superior performance in most experiments.

Keywords: Human activity recognition; multimodal fusion; deep convolutional neural networks.

[†]Corresponding author.

1. Introduction

Human activity recognition (HAR) from visual data is one of the of the most challenging computer vision tasks,[87] gaining an increasing amount of attention within the research community.[89] It may be defined as the recognition of some human motion and/or behavior within an image or a video sequence. Moreover, an activity (or "action") may be defined as a type of motion performed by a single human, taking place within a relatively short (however, not instant) time period and involving multiple body parts.[89] The aforementioned informal definition differentiates activities from gestures; the latter are typically instant and involve at most a couple of body parts, while an activity may even involve the whole body. Similarly, interactions may involve either a human and an object or two humans. Finally, group activities involve more than one humans. Another categorization of HAR tasks is the following[89]: (a) *segmented* HAR, wherein the input video sequence depicts exactly one activity example, which should be recognized, i.e. as in a typical classification problem. This means that all frames before and after the activity have been trimmed; and (b) *continuous/online* HAR, wherein the input video may contain some actions (this means that it may not contain any or it may contain one, or even more). Activity temporal boundaries are not provided and should be detected within the approach. The main domains of application of HAR include video surveillance, human–computer/robot interaction, augmented reality (AR), ambient assisted environments, health monitoring, intelligent driving, gaming and immersion, animation, etc.,[67,89] however the number of possible HAR applications is ever-growing.

There exist several HAR approaches that are based on either wearable sensors or sensors installed within the subject's environment. A plethora of such sensors has emerged during the last few years. In the former case, the most popular ones include smartwatches, hand/body worn sensors, smartphones, etc. Moreover, in the latter case, typical sensors include video/thermal cameras, microphones, infrared, pressure, magnetic, Radio-Frequency IDentification (RFID) sensors,[12] etc. However, it has been shown that wearable sensors are not preferred by the users, while their usability is below average.[36,58] Moreover, overloading the users' environment with a plethora of sensors may be an expensive task, requiring in some cases many interventions in home furniture and/or appliances, e.g. in case of a home environment. Therefore, several low-cost solutions tend to be based solely on one or more cameras, detecting activities using solely the subjects' captured motion.

Common HAR approaches that are based on cameras use as input some raw sensor data from the performed activity, while their output is the classification result, i.e. the determination of the aforementioned activity. In between lie the processing and reasoning steps. More specifically, a given HAR approach may consist of (some of) the following steps: (a) raw data are pre-processed e.g. for noise and redundancy removal; (b) within the video sequence, temporal segmentation takes place, aiming to extract video segments that contain exactly one action to be recognized. Approaches that contain this step perform "segmented" recognition; (c) feature extraction, aiming to extract important temporal, spatial or visual features from human motion; and (d) dimensionality reduction to increase the quality and to decrease the size of features.

Note that the feature extraction step requires knowledge and expertise regarding the specific domain of application, in order to provide features that will be able to discriminate between activities. However, a common problem of these features is that even though they may demonstrate satisfactory performance within the given domain of application, they may fail when applied to a similar domain. A solution to the aforementioned problem is omitting the feature extraction step and instead using a classification approach that is based on deep learning. In that case, features are not engineered; they are *learnt* from the specific training data. Common deep learning architectures that have been successfully applied in the problem of HAR are the convolutional neural networks (CNNs)[43] and the recurrent neural networks (RNNs),[20] outperforming the majority of traditional machine learning approaches. In case of the segmented HAR tasks, both architectures may be applied, while in case of continuous HAR tasks, RNNs are the most common approach. Although a vast amount of research has been conducted on improving recognition performance, several principal challenges, such as the representation and the analysis of actions, still remain unresolved.

In the early years of HAR, the first publicly available datasets consisted of a relatively small number of very simple activities and consisted of either still images or low-quality videos.[21] For example, the KTH dataset[68] was limited to a small number of simple actions such as *walking*, *running*, *hand clapping*, etc. Some years later, the next "generation" of datasets targeted more "realistic" activities. Moreover, the Hollywood dataset[42] contained classes such as *answer phone*, *get out of car*, *hand shake*, etc. However, this increased activity complexity was not accompanied by an increase to the number of classes; most datasets were still limited to approximately 10–15 classes or less. Then, several challenging datasets comprising a large number of more complex activities arrived. Notable examples include the UCF101 dataset[73] with 101 activities, the HMDB dataset[40] with 51 categories, etc. These datasets contained large numbers of more complex actions, including interactions with objects such as *playing cello*, *horse riding*, *swing baseball bat*, *fencing*, etc.

With the advent of cost-effective sensors such as Microsoft Kinect, depth data have become widely available. This way, several challenging human activity datasets now provide 3D multimodal raw data, i.e. consisting of RGB video and depth information. The latter has also allowed for the extraction of a third modality, i.e. skeleton sequences that consist of 3D coordinates of human joints as the subject moves in space, over time. Moreover, it is well known that when working with deep learning approaches, a large-scale multi-class dataset may be the key to effectiveness and robustness. Therefore, contemporary datasets provide a large number of training videos. Notable examples include PKU-MMD[52] and NTU-RGB+D[56] datasets, which provide RGB, depth and skeleton data, for 51 and 120 classes, respectively. Note that the depth modality, unlike the conventional RGB, is invariant to illumination changes and also reliable for the estimation of body silhouettes. Nevertheless, RGB information contains color and texture which are significant for discriminating several actions involving e.g. human–object interactions. Different modalities offer different perspectives of actions, thus, intuitively, a fusion of their complementary correlations should be meaningful. Furthermore, the existence of skeletal information can be very helpful for accurately capturing the human body posture. However, in scenarios where the source of motion features is limited to sequence data, the challenge of CNN-based methods is to find efficient encoding techniques for representing skeleton sequences, while capturing spatio-temporal activity features.

In this paper, a novel approach for HAR that fuses multiple modalities, incorporating RGB, depth and visual representations of 3D skeletal information is presented. Inspired by previous work on fusion of different representations,[77] and early experiments using representations of 3D skeletal data that are based on image transforms[61] or on artificially created pseudo-colored images,[85] in this paper a late fusion methodology that uses as input RGB and depth video sequences is proposed. The former are "summarized" into a single "dynamic" image. Both data representations are used to create 3D skeletal sequences of the subjects. These are then used to create (a) four visual representations of skeletal data upon applying well-known image transforms to the spectral space; and (b) a pseudo-colored image representation of skeletal data. For each of the aforementioned modalities, a CNN is trained and is used as feature extractor. Extracted features are fused, scaled and used as input to an Support Vector Machine (SVM) classifier. The proposed method is evaluated on three publicly available human motion datasets, namely, PKU-MMD dataset,[52] SYSU 3D Human–Object Interaction (HOI) dataset[27] and UTKinect-Action3D dataset.[90]

The novelties of the proposed approach are as follows: (a) it is based on the extraction of deep features and early fusion of different image transforms that correspond to skeletal motion; (b) it applies a novel viewpoint augmentation scheme prior to create both image transforms and pseudo-colored image representations of skeletal motion; (c) it incorporates dynamic images to exploit the raw RGB information of motion; and (d) it proposes the fusion of skeletal motion and raw RGB features using only CNNs.

This paper is arranged as follows. Section 2 presents recent research works closely related to the proposed approach. Section 3 initially provides an overview on the visual data that are used within this work, the camera setup that has been assumed and the data augmentation strategy that has been adopted. Then, it describes the whole classification methodology, i.e. the creation of image representations, network architectures, training process, data fusion and classification. The datasets used, the experimental setup, evaluation and results are presented in Sec. 4. Finally conclusions are drawn in Sec. 5, wherein plans for future work are also presented.

2. Related Work

In this section, a brief review of recent scientific literature on HAR using deep learning is presented. Since the problem of HAR is very wide and since the novelties of this work lie in the areas of image representations of 3D skeletal data sequences and multimodal fusion of human motion modalities, focus is given on two major research categories, (a) methods for extracting skeletal information by image-based representations; and (b) approaches that utilize information from multiple modalities.

2.1. *Visual skeletal representations*

When a CNN is used and the only available motion features are skeletal data, an intermediate visual representation of skeletal sequences is required. This representation should capture both spatial and temporal information regarding the motion of joints and reflected to its color and/or texture properties.

Most works rely on 3D skeletal sequences, comprising a set of moving joints in the 3D space. Typical simple features extracted from this motion are inter-joint distances and orientations, either between joints of the same or consecutive frames.[28,34,45,65] Other features include joints' motion direction[88] and magnitude.[34,88] Several approaches also consider duration of activities.[55,57] A skeleton is typically treated as a single set, though in some cases, its subsets that correspond to body parts may be treated independently.[34,77] Many works opt to extract features from projected 3D skeletons into the 3 2D planes.[10,34,45,84] The extracted simple features may be stacked so as to create 2D pseudo-colored images.[28,65] In several cases, images are created by joints' trajectories[26,82,88] or heatmaps.[14] In the following, several recent intermediate representations of skeletal data are briefly presented.

The "pose-transition feature to image" technique proposed by Huynh-The *et al.*,[28] extracts inter-joint distances and orientations within the same and consecutive frames and uses them to create a feature vector per frame. All feature vectors are stacked and their values are encoded to create an RGB image. The same distances and orientations were also used by Pham *et al.*[65] to create "enhanced Skeleton Posture-Motion Feature" (SPMF) while they also used a color enhancement method, to increase contrast and highlight texture and edges the representation. A similar approach was presented by Ke *et al.*,[34] wherein body parts are represented by subsets of joints, while distances and magnitudes are calculated between parts instead of joints.

A similar idea was proposed by Silva *et al.*[69] wherein, the position of the joints in the final representation is clustered into groups and by Yang *et al.*[91] who proposed a "tree structure skeleton image" (TSSI), based on the idea that spatially related joints in original skeletons have direct graph links between them. This way, spatial correlations between joints are better preserved. Moreover, to aggregate more temporal dynamics to the representation, Caetano *et al.*[6] used several temporal scales. Another approach to encode joint motion that does not calculate joint distances is the one of Duan *et al.*[14] who used joint heatmaps which are composed upon posing a set of Gaussian maps centered at each joint.

In "joint trajectory maps" (JTM) proposed by Wang *et al.*,[88] skeleton data sequences are represented by three 2D images wherein motion direction and magnitude are reflected as the hue, saturation and brightness of the colored image. Also, different body parts are represented by multiple color maps. Similarly, in "joint distance maps" (JDM), proposed by Li *et al.*,[45] three maps correspond to inter-joint distances in the orthogonal planes, while the fourth in the 3D space, all encoded by hue. Extending the aforementioned works,[65,88] Li *et al.*[46] used both CNN and Long Short Term Memory (LSTM) networks to classify spatial and temporal motion properties, while a late fusion approach was adopted. The idea of projecting 3D joints to planes was also used in "Temporal Pyramid Skeleton Motion Maps", proposed by Chen *et al.*,[10] who created hierarchical structures using different types of joint visualization, calculation of inter-joint distances in consecutive frames and pseudo-color coding and by Verma *et al.*[84] who created skeleton intensity images, for top, front and side views of skeletons.

The durations of activities have been incorporated in the representation of "Skepxel", proposed by Liu *et al.*[57] along with the coordinate values of joints. A group of Skepxels are generated for a single skeleton frame while the representation is constructed upon concatenation of groups of Skepxels in a column-wise manner. Similarly, Liu *et al.*[55] used joint coordinates, labels and the corresponding timestamps. This 5D representation was projected to a 2D image using labels and timestamps, and the remaining dimensions were used as R, G, B, channels to form pseudo-colored images. A variation of this idea was adopted in "joint skeleton spectra", proposed by Hou *et al.*,[26] wherein joint distribution maps are projected onto three Cartesian planes, reflecting the temporal variations of joints to hue values. The idea of changing colors of joints as time is passing was proposed by Tasnim *et al.*[82] and was used to create spatio-temporal images.

Although most of the aforementioned approaches demonstrate more than satisfactory performance in ideal conditions, they do not deal with two main problems that occur in real-life situations, i.e. viewpoint changes and occlusion. In this work, we deal with the first problem and propose a data augmentation technique, applied in skeletal data.

2.2. *Multimodal fusion methods*

Moreover, several approaches dealing with the fusion of more than one data modalities have been proposed. A notable early work is the one of Simonyan and Zisserman,[71] who trained a CNN for capturing spatial features from raw video frames and another for capturing motion features from dense optical flow. A late fusion approach was adopted, using an SVM. Another early work is the one of Chaaraoui *et al.*,[8] who combined body pose estimation and 2D shapes, to obtain skeletal and silhouette-based features, which were then combined by early fusion.

The majority of more recent works is based on the use of both RGB and depth data.[8,15,39,86] Other modalities that have been used as input in multimodal fusion approaches are thermal data,[23] inertial measurements,[15,29] audio data[47] and RFID data.[47] In those works, each data modality is independently processed, in order to extract features. In many cases, features are handcrafted,[15] yet in the majority of works, deep features are extracted, aiming to capture spatial[71] and/or temporal[71] properties of human motion. To this goal, most approaches rely only in CNN architectures[8,39,54,71,86] that are trained as feature extractors and aim to extract features from either from raw video frames or from intermediate 2D feature representations. Although deep features have almost completely dominated the research area of HAR, it has been shown that handcrafted features may capture complementary motion properties, thus, their fusion may boost performance.[39]

In several cases, especially when dealing with small datasets, pre-trained networks in the same or in a similar domain or dataset[86] or transfer learning approaches[54] are used. More recent works may additionally use LSTM networks so as to extract temporal features,[23,29,47,96] while in few cases only LSTM networks are used.[78]

Typically, fusion is implemented as concatenation of these feature vectors (i.e. early fusion) and recognition involves a traditional machine learning classifier, such as a support vector machine,[15,39,71,86] or a clustering approach such as k-means.[8,15] Yet, late fusion, i.e. fusion of classifier decisions is also used.[63,71]

However, the majority of works relies on a single representation per data modality. Specifically, it has not been investigated whether different representations of skeletal joint motion may carry complementary information, thus their fusion could provide a performance boost on recognition approaches. Moreover, raw RGB data are often ignored or used only for skeleton extraction, yet, in this work we demonstrate that they may significantly assist recognition in several cases.

3. Proposed Methodology

In this section, the proposed methodology for HAR is presented. In brief, it is based on the fusion of raw RGB and 3D skeletal motion data sequences. The former are used to create a condensed representation of the activity, consisting a single image. The latter are used to create four representations of the spectral content of the skeletal motion, which are based on well-known image transforms and also a pseudo-colored representation of skeletal motion. Each image is then fed to an appropriately trained CNN. The second from the end dense layer of each network is used as a feature vector. All feature vectors are then concatenated and upon applying principal component analysis (PCA), are classified using a support vector machine. A visual overview of the proposed approach is illustrated in Fig. 1.

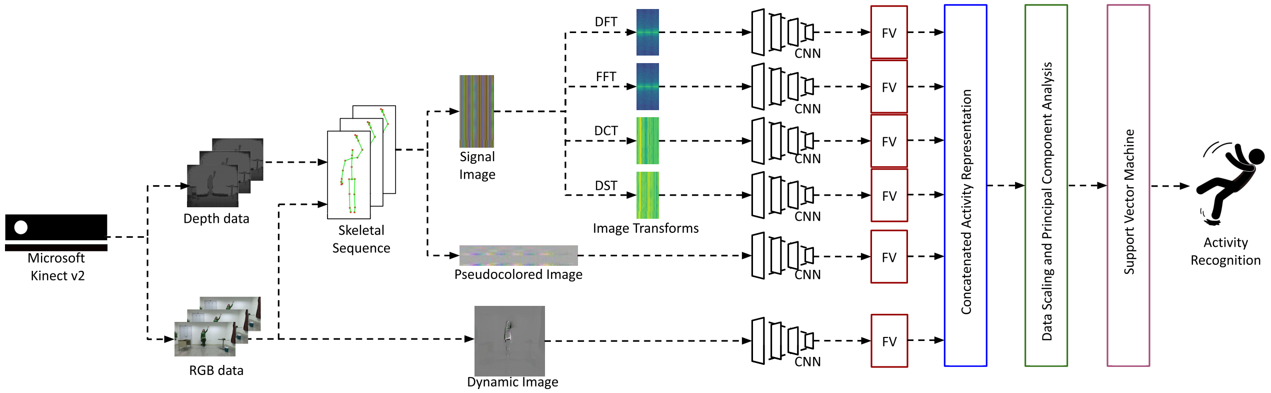

Fig. 1. A visual overview of the proposed approach.

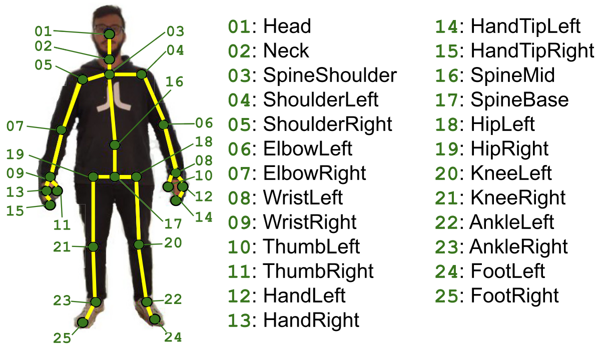

Fig. 2. The 25 skeletal joints extracted by Microsoft Kinect v2.

3.1. *Visual data*

The proposed approach relies on visual data modalities derived from the 3D motion of humans, since in typical 3D HAR problems, subjects perform actions in space and over time. Herein, both raw RGB video and 3D motion of human skeletons are considered. The latter are represented as structured sets of 3D skeleton joints moving in space. More specifically, in the context of this work, RGB and skeleton data that have been captured using the Microsoft Kinect RGB/depth camera[a] are used. Within captured sequences, a human skeleton comprises 25 3D joints, which are organized as a graph; each node corresponds to a body part such as hands, feet, head, neck, etc., while edges follow the body structure, connecting pairs of joints. In Fig. 2, a skeleton extracted using the Kinect camera is illustrated. For the sake of explanation, a visual example of an activity is illustrated in Fig. 3. Note that the proposed approach is not tied to the use of the Kinect v2 camera. The only requirement is to have as input both video sequences and 3D skeleton sequences. In case of 2D skeleton sequences, such as the ones provided by e.g. PoseNet,[9] OpenPose[7] or Movenet[b] a performance drop should be expected.

3.2. *Camera setup and data augmentation*

Generally speaking, data augmentation is a process that aims to expand the size and/or the diversity of some data set. This is typically achieved upon creating "artificial" data samples, which however are quite similar to the original sample, yet not identical to any of them. In the context of computer vision problems, usually data are images, thus data augmentation aims to construct synthetic images, based on the properties and limitations of the given problem. Data augmentation approaches in HAR problems have employed local averaging and sampling,[76] transformations such as rotation, scaling, jittering, etc.[32] and data warping.[25] Little work has been demonstrated when working with skeletal data. In that case, augmentation approaches include direct application of geometric transformations[55] or RNNs[92] in raw skeletal sequences.

In previous works,[62,74] the effect of data augmentation to the classification performance of a HAR methodology has been assessed. We showed that incorporating artificially rotated skeletons to the

[a] https://developer.microsoft.com/en-us/windows/kinect.

[b] https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html.

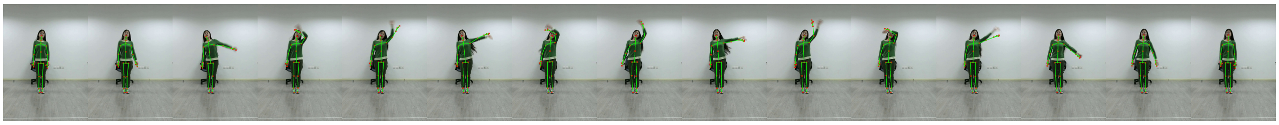

Fig. 3. A sequence of an actor performing the activity *handwaving*. Extracted human skeleton 3D joints using the Kinect SDK have been overlaid. Frames have been taken from the PKU-MMD dataset[52] and have been trimmed for illustration purposes.

training dataset may significantly assist to boost the accuracy of deep approaches in multi-camera setups. Note that due to the way image representations of skeletal data are constructed in the context of this work, popular data augmentation strategies such as rotations and random crops may not be applied; such methods could severely affect the spectral properties of activity images, by removing or distorting important information. Thus, rotation of activity samples of skeletal motion upon geometric processing them which led to the creation of artificial, valid examples has been instead adopted.

Therefore, the aforementioned augmentation methodology[62] has been incorporated to the training process of this work. In brief, given a camera setup composed by three cameras, capturing the subject under different viewpoints and assuming that they are all placed at the same distance to the subject (i.e. at the perimeter of an imaginary circle), a given camera may be "aligned" to any of the remaining two upon imposing a simple rotation transformation. Let θ the desired rotation angle. Then, the corresponding transformation $\mathbf{R}_y$ is given by[83]

$$\mathbf{R}_y(\theta) = \begin{bmatrix} \cos\theta & 0 & \sin\theta \\ 0 & 1 & 0 \\ -\sin\theta & 0 & \cos\theta \end{bmatrix}, \tag{1}$$

where in that case y denotes the axis of the rotation.

Within the considered 3-camera setup, one camera is directly facing the subject, while the remaining two are placed on its left and right, at angles θ_L and θ_R, respectively, as illustrated in Fig. 4. Therefore, given a training sample, artificial samples are created upon rotating the skeleton by an angle θ, about the y-axis, complying to the Cartesian 3D coordinate system used by the Kinect camera. Specifically, for a given sample a rotation transformation is applied with $\theta \in \{-90°, -45°, +45°, +90°\}$. An example of the data augmentation process is illustrated in Fig. 5, wherein a skeleton corresponding to a video frame of an activity example of class hand waving is illustrated along with four artificial ones, that have been created using the aforementioned process.

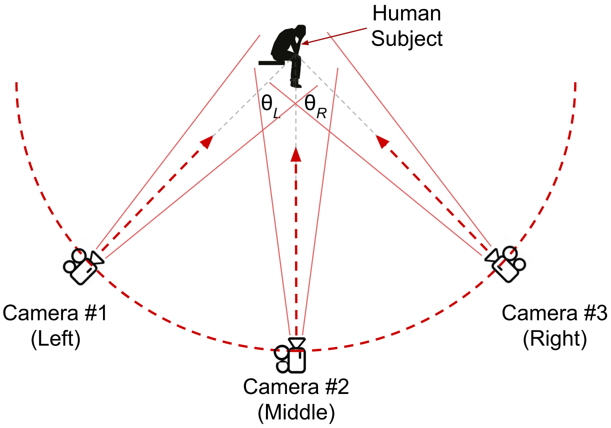

Fig. 4. The 3-camera setup used in this work. Camera #2 is directly facing the subject while she/he is performing an activity. Cameras #1 and #3 have been placed on an imaginary circle, forming angles equal to θ_L and θ_R, respectively, with Camera #2.

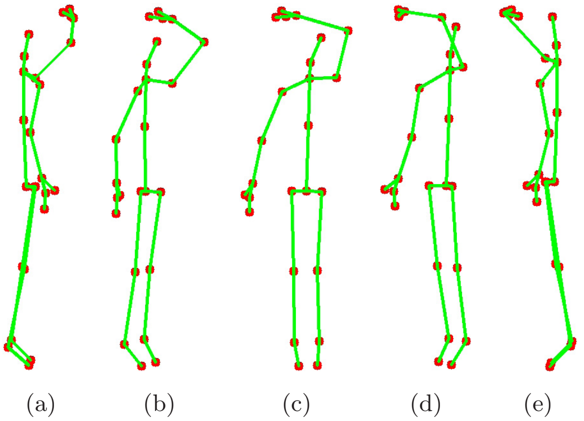

Fig. 5. Example of rotated skeletons that have been used for data augmentation. A skeleton rotated by angle θ: (a) $\theta = 90°$; (b) $\theta = 45°$; (c) $\theta = 0°$; (d) $\theta = -45°$ and (e) $\theta = -90°$. The original skeleton is denoted by $\theta = 0°$, while the other four skeletons are the augmented samples. For illustrative purposes, depth information, i.e. z-coordinate has been discarded.

3.3. *Signal images*

Inspired by the work of Jiang and Yin[31] who worked using raw sensor measurements from inertial sensors, as a first step "signal" images are created, by concatenating the signals that are produced by skeletal motion. Specifically, the motion of each skeletal joint in the 3D space, over time is treated as three independent 1D signals. Each of them corresponds to a coordinate. Therefore, for a given joint j, let $S_{j,x}(n), S_{j,y}(n), S_{j,z}(n)$ denote the 3 1D signals that correspond to its 3D motion and for the coordinates x, y, z, respectively. Thus, in the signal image, $S_{j,x}(n)$ corresponds to row $3\times j-2$. Accordingly, $S_{j,y}(n)$ and $S_{j,z}(n)$ correspond to rows $3\times j-1$ and row $3\times j$. This way, the signal image $\mathbf{S}$ for a given activity and for N joints, is created upon concatenation of the $3\times N$ signals, thus its dimension is $3\times N\times T_s$, where T_s is the duration of this activity. However, in real problems, since different subjects may perform the same activity with different duration and also different activities require different duration, it should be obvious that T_s is variable. In order to address the problem of temporal variability between subjects and between activities, so as to allow for signal concatenation, as discussed, a linear interpolation step is imposed. This way the length of all activities T_a is fixed. Note that in order to set value of T_a the process begins with the selection of a value close to the mean of all activities, yet the exact value that is finally used is chosen upon experimentation and fine-tuning. Thus, in the context of this work, the length is set to $T_a = 159$. Moreover, given $N = 25$, the dimension of $\mathbf{S}$ is equal to 75×159. An example signal image is illustrated in Fig. 6.

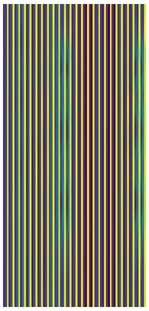

Fig. 6. The signal image that is created as a first step for the representation of skeleton data, using the methodology described in Sec. 3.3. This example corresponds to the activity handwaving, depicted in Fig. 3.

3.4. *Activity images*

Let $\mathbf{S}$ denote a signal image with dimensions $W\times H$ and $\mathbf{S}_{nm}$ the pixel at coordinates (m, n). From each signal image an "activity" image $\mathbf{A}$ is created. This may be done by applying one of the following image transforms[19,30] on a given signal image $\mathbf{S}$: (a) the 2D Discrete Fourier Transform (DFT), which is defined as

$$\begin{aligned}\mathbf{A}(u,v) = &\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}\mathbf{S}(x,y)e\\ &-j2\pi(ux/W+vy/H),\end{aligned}\tag{2}$$

where $x\in[0,W-1], y\in[0,H-1], u\in[0,W-1], v\in[0,H-1]$; (b) the 2D Fast Fourier Transform (FFT), which is a fast implementation of DFT; (c) the 2D Discrete Cosine Transform (DCT), which is defined as

$$\begin{aligned}\mathbf{A}(u,v) = a_u a_v &\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}\mathbf{S}(x,y)\\ &\times\left(\cos\frac{\pi(2m+1)u}{2W}\cos\frac{\pi(2n+1)v}{2H}\right),\end{aligned}\tag{3}$$

where $x\in[0,W-1], y\in[0,H-1], u\in[0,W-1], v\in[0,H-1]$ and also

$$a_u = \begin{cases}\dfrac{1}{\sqrt{W}}, & u = 0,\\ \sqrt{\dfrac{2}{W}}, & 1\le u\le W-1\end{cases}\tag{4}$$

and

$$a_v = \begin{cases}\dfrac{1}{\sqrt{H}}, & v = 0,\\ \sqrt{\dfrac{2}{H}}, & 1\le v\le H-1;\end{cases}\tag{5}$$

(d) the 2D DST, which is defined as

$$\begin{aligned}\mathbf{A}(u,v) = a_u a_v &\sum_{x=0}^{W-1}\sum_{y=0}^{H-1}\mathbf{S}(x,y)\\ &\times\left(\sin\frac{\pi(2m+1)(u+1)}{2W}\right.\\ &\left.\times\sin\frac{\pi(2n+1)(v+1)}{2H}\right),\end{aligned}\tag{6}$$

where $x\in[0,W-1], y\in[0,H-1], u\in[0,W-1], v\in[0,H-1]$ and a_u, a_v are given by Eqs. (4) and (5),

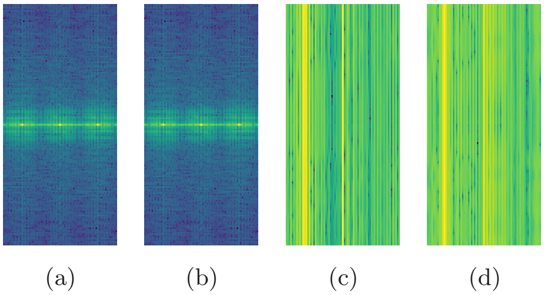

Fig. 7. Activity images resulting upon (a) DFT; (b) FFT; (c) DCT; (d) DST; from the signal image of Fig. 6 and for the activity handwaving. Note that DFT and FFT images have been processed with log transformation for visualization purposes. Figure best viewed in color.

respectively. Note that from each transform only the magnitude is preserved, while the phase is discarded. Also, note that DST and DCT are further processed by normalizing using the orthonorm. Obviously, in all cases the result is a 2D image, with the same dimension as the signal image **S**. Moreover, although FFT is a fast implementation of DFT, as it has been mentioned, it was amongst the goals of this work to assess if, due to the expected differentiation between them, any differences in performance would show in practice. In Fig. 7, the four corresponding activity images, created upon applying the aforementioned transforms on the signal image **S** of Fig. 6 are illustrated.

3.5. *Pseudo-colored image representation of skeletal data*

As it has already been mentioned in Sec. 2, another approach to use the 3D skeletal information as input to a CNN, is to create image representations of skeletal data. However, when creating such representations, one should consider capturing and preserving spatio-temporal properties of skeleton trajectories, so that the resulting representations could be able to discriminate among different activities. Moreover, they should also comply to the graph structure skeleton representation. In the context of this work, the representation initially presented in previous work[85] and whose early results indicated that it could be successfully applied to the problem of HAR is adopted.

Specifically, this representation aims to capture inter-joint distances as they vary through the duration activity. These are then used to create pseudo-colors within an artificial RGB image. Note that it is based on the 3D trajectories of skeletal joints. Let $x(n)$, $y(n)$ and $z(n)$ denote the sequences of coordinates of each of the N available joints, at the nth frame F_n of the video sequence depicting the activity. Through the duration of any activity, a set of $3 \times N$ signals is collected for a given video sequence. To address the problem of temporal variability between actions and between users, as it has been described in Sec. 3.3, a linear interpolation step in the exact same way as in the case of signal images is imposed. Then, from each of the aforementioned sequences difference between consecutive frames is calculated. To create pseudo-colored images, x, y, z coordinates are assigned to R, G, B color channels of the pseudo-colored image, respectively.

Particularly, the process of creating a pseudo-colored image $\mathbf{I}$ is as follows: Let $x_i(n), i = 1, \ldots, N$ denote the x-position of the ith joint in the nth frame. Let r, g, b denote the red, green and blue channel of $\mathbf{I}$, respectively. Pixel values of $r(i,n)$, $g(i,n)$, $b(i,n)$ are calculated as

$$\begin{aligned} r(i,n) &= x_i(n+1) - x_i(n), \\ g(i,n) &= y_i(n+1) - y_i(n), \\ b(i,n) &= z_i(n+1) - z_i(n), \end{aligned} \tag{7}$$

where $n = 1, \ldots, T_a$ and $i = 1, \ldots, N$. As it is exhibited, the way these pseudo-colored images are formed, leads to preserving both the temporal and the spatial properties of the skeleton trajectories. Obviously, the dimension of the pseudo-colored image is $N \times T_a \times 3$, which in our case is $25 \times 159 \times 3$. In Fig. 8, a pseudo-colored image that corresponds to the activity depicted in Fig. 3 is illustrated.

3.6. *Dynamic images for activity representation*

A dynamic image is a typical RGB image, which by construction aims to summarize the appearance and dynamics of a given video sequence.[5] Specifically, the idea that lies behind dynamic images[16]

Fig. 8. The pseudo-colored image for the activity *handwaving* that is illustrated in Fig. 3.

is known as "rank pooling" and aims to represent a video sequence as a ranking function $S(\bullet)$ of its frames $F_1, \ldots, F_{T_a}$. By using $S(\bullet)$ a feature vector $\psi(F_i)$ from each frame F_i is extracted. Let $V_n = \frac{1}{n}\sum_{i=1}^{t}\psi(F_n)$ denote the time average of the aforementioned features ranging from F_1 to F_n. The ranking function associates each moment n with a score $S(n)$, given a parameter set $\mathbf{d}$ and assigning larger scores to later frames. Scores are learned by solving a convex optimization problem which defines a function $\rho(\bullet)$ mapping a given video sequence comprising T frames to a vector $\mathbf{d}^*$, used as feature descriptor. Using the RankSVM formulation,[72] $\mathbf{d}^*$ may be estimated as follows:

$$\mathbf{d}^* = \rho(F_1, \ldots, F_n; \psi) = \arg\min_{\mathbf{d}} E(\mathbf{d}), \tag{8}$$

where

$$\begin{aligned} E(\mathbf{d}) = {} & \frac{\lambda}{2}\|\mathbf{d}\|^2 + \frac{2}{T(T-1)} \\ & \times \sum_{q>t}^{\max}\{0, 1 - S(q \mid \mathbf{d}) + S(t \mid \mathbf{d})\}. \end{aligned} \tag{9}$$

Although $\psi(\bullet)$ may be any feature extractor, Bilen *et al.*[5] opted for simply using raw RGB pixel values and reported remarkable results. However, the most important aspects of such an approach are (a) $\mathbf{d}^*$ may be interpreted as an RGB image as it has the exact same number of elements; and (b) this image is obtained by rank pooling, thus it may be regarded as a summary of the whole video sequence. Note that the pixels in the produced dynamic images tend to focus on salient information rather than the background; this is considered the reason for them being appropriate for the problem of HAR, wherein typically the background is static and a subject consists a nonstatic part within the video sequence. This is clearly depicted in Fig. 9 where a dynamic image that corresponds to the activity of Fig. 3 is illustrated. In this image, it is clear that information regarding the background has been discarded, while emphasis has been given to the subject and her moving arm.

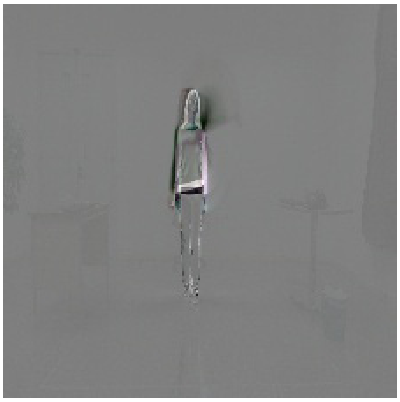

Fig. 9. The dynamic image for the activity *handwaving* that is illustrated in Fig. 3.

3.7. *Multimodal fusion and classification*

The architecture of the CNN that has been used throughout our experiments with DFT, FFT, DCT and DST images has been experimentally defined and has been initially used in previous work,[61] while it is illustrated in detail in Fig. 10(a). It consists of three convolutional layers with 32, 64 and 128 kernels of size 3 × 3. Each is followed by a pooling layer using "max-pooling" to perform 2 × 2 subsampling. A flatten layer transforms the output image to a vector, used as input to a dense layer of size 128, using dropout[75] and a second dense layer produces the output of the network. In case of the pseudo-colored image representations, the CNN architecture was based on the aforementioned one and has been experimentally modified to better fit this representation. It is illustrated in detail in Fig. 10(b). It consists of a convolutional layer with 16 kernels of size 3 × 3, followed by a pooling layer. Then, two convolutional layers with 32 kernels of size 3 × 3 and a pooling layer follow, succeeded by a convolutional layer with 64 kernels of size 3 × 3 followed by a pooling layer. Each pooling layer uses "max-pooling" to perform 2×2 subsampling. Then, once again a flatten and two dense layers follow as in the aforementioned network.

Finally, in case of dynamic images, the CNN architecture that was adopted is based on the well-known VGG16[70] and is illustrated in detail in Fig. 10(c). First, two convolutional layers filter the input image with 64 kernels of size 3 × 3. A pooling layer follows. Then, two convolutional layers filter the input image with 128 of size 3×3, followed by another pooling layer. Then three convolutional layers filter the resulting image with 256 kernels of size 3 × 3 and are again followed by a pooling layer. Then, two triplets convolutional layer follow, filtering the resulting image with 512 kernels of size 3×3, each followed by a pooling layer. Then, a flatten layer transforms the output image of size 7×7 of the fifth pooling into

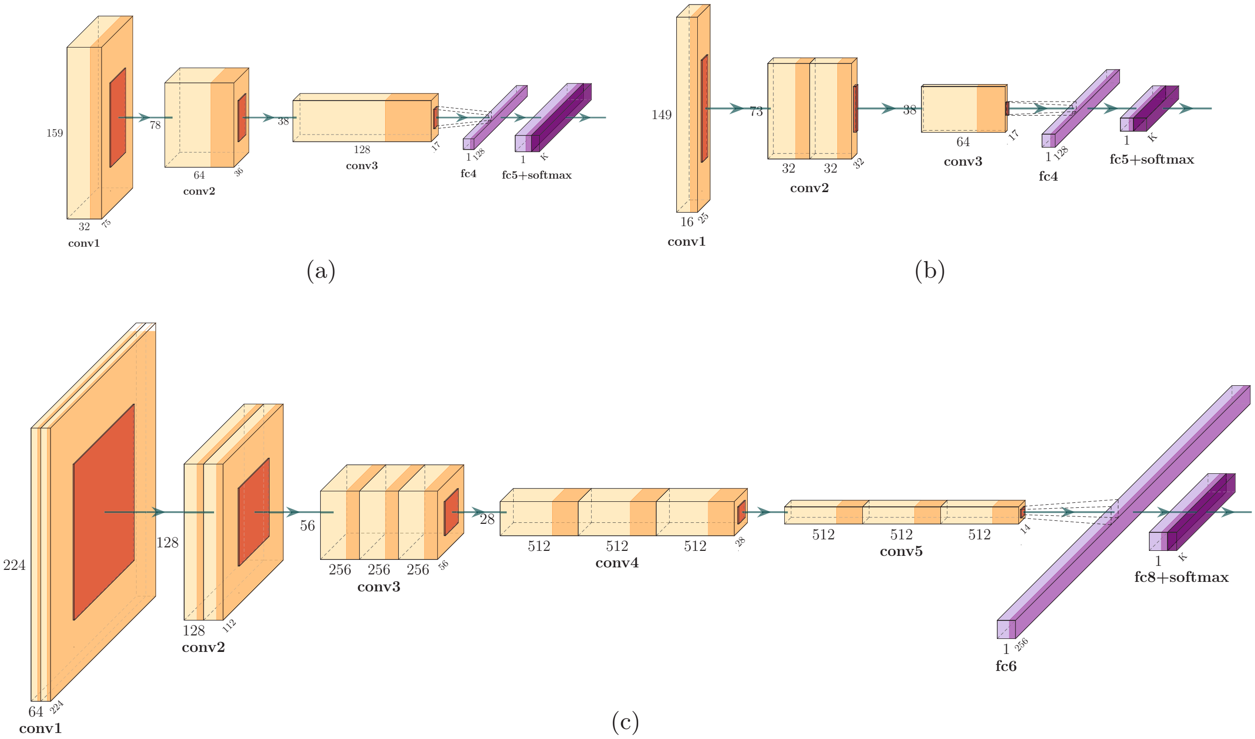

Fig. 10. The three deep convolutional network architectures that have been used in this work and for the cases of (a) DFT, FFT, DCT, DST activity images; (b) pseudo-colored images; and (c) dynamic images.

a vector, with is then used as input to a dense layer of size 256 using dropout. Finally, a second dense layer produces the output of the network. All pooling layers use "max-pooling" to perform 2×2 subsampling.

Note that the second dense layer is omitted when the CNN is used for feature extraction. Also, the first and the second convolutional networks are trained using the available dataset, while the third convolutional network is pre-trained with ImageNet.[13] The convolutional layers are "freezed" and training of the dense layers using the dynamic images continues.

Therefore, upon using the aforementioned networks as feature extractors, a feature vector per representation is formed. The early fusion step of this work consists of concatenation of these six feature vectors. This way, a combined representation of size 896 has been created. Upon scaling and PCA, only the components that correspond to 95% of total variance are kept, resulting to a feature vector of size approximately 350, which was fed to the SVM. Note that due the way PCA works, the size depends on the dataset, therefore, as expected it varies, per experiment. Finally, for classification, an SVM with an RBF kernel has been used.

4. Experiments and Results

4.1. *Datasets*

In order to experimentally evaluate the proposed multimodal HAR approach, a large-scale, public and open benchmark datasets, namely, PKU-MMD[52] has been selected. It focuses on human activity understanding and it contains approximately 20K action instances from 51 activity categories, spanning into 5.4 M video frames and performed by 66 human subjects. A multi-camera setup has been used throughout the recording sessions. More specifically, data from 3 Microsoft Kinect v2 cameras have been collected and the following visual data modalities are provided: (a) raw RGB video sequences, each depicting one or more actors while performing an action/interaction under a given viewpoint; (b) depth sequences, that is, the z-dimension corresponding to the scene depth at each pixel of an RGB sequence; (c) infrared radiation sequences, that is, modulated infrared light captured simultaneously to the RGB sequences; and (d) positions in the 3D space of the extracted human skeleton joints, varying over time.

Recordings from three camera views are available; each activity was simultaneously captured by all cameras. All subjects were asked to perform activities within a pre-determined area, in order to have an as fixed as possible distance to all cameras. Also, they were asked to face directly one of the cameras; i.e. each activity sample was recorded under three viewpoints. Throughout the experimental evaluation and as we have already illustrated in Fig. 4, the following naming convention for each camera (i.e. for each viewpoint) will be used: L (left), M (middle) and R (right). As illustrated in Fig. 4, the following fixed angles are used for their positioning: $\theta_L = -45°$ and $\theta_R = +45°$. Also, the cameras have been placed on the same height level, which remained fixed and equal to 120 cm for all activities, Moreover, since videos contain several sequential actions, inter-video temporal boundaries are available. The number of recordings (examples) is 6918, 6928 and 6934 for viewpoints L, R and M, respectively. Finally, in all examples, all skeleton joints are visible, no matter the viewpoint.

Moreover, in order to further evaluate our approach, 2 small-scale, single-camera human activity datasets are also used. The second dataset, namely, SYSU 3D HOI[27] also consists of 3D human motion. However, contrary to PKU-MMD, which contained several types of activities, SYSU 3D HOI focuses on interactions between humans and objects. It is not as large as PKU-MMD, yet it consists of 480 activity instances from 12 different activities which involve interaction of 40 subjects with one of the following objects: phone, chair, bag, wallet, mop and besom. For each activity, RGB, depth and 3D skeleton data are available. Note that there are several activities which appear highly similar, e.g. mopping and sweeping. The third and final dataset, namely, UTKinect-Action3D dataset[90] consists of 10 simple activities that have been performed by 10 subjects. Each subject performs all activities twice, thus 200 activity instances are provided. As in SYSU 3D HOI, for each activity, RGB, depth and 3D skeleton data are available.

4.2. *Experimental setup and implementation details*

Experiments were performed on a personal workstation with an AMD Ryzen$^{\text{TM}}$ 5 1600 6-core processor on 3.20 GHz and 16GB RAM, using NVIDIA$^{\text{TM}}$ Geforce GTX 1060 GPU with 6 GB GDDR5 VRAM and Ubuntu 20.04 (64 bit). The deep architecture has been implemented in Python, using Keras 2.4.3[11] with the Tensorflow 2.5[1] backend. All data preprocessing and processing steps have been implemented in Python 3.9 using NumPy,[c] SciPy[d] and OpenCV.[e] For training the CNN, the ReLU activation function has been used. Moreover, the batch size has been to 8 and the Adam/SGD optimizers (Adam in skeleton models, SGD in Dynamic model) has been used. Also, the dropout was set to 0.5, the learning rate was set to 0.001 and the network was trained for 150 epochs, using the loss of the validation set calculated via cross-entropy as an early stopping method, in order to avert overfitting. The parameters of the SVM were $C = 100$ and $\gamma = 0.001$ and have been selected upon grid search.

The four CNNs that are used to extract features from the DFT, FFT, DCT and DCT images comprise $2,164,339$ trainable parameters; training in case of the most demanding PKU-MMD dataset requires on average approximately 35 min., while extraction of features requires approximately 0.002 s. The CNN that is used to extract features from the pseudo-colored images comprises $170,611$ trainable parameters; training in case of the most demanding PKU-MMD dataset requires on average approximately 25 min., while extraction of features requires approximately 0.001 s. The VGG16 network that is used to extract features from the dynamic images comprises $21,150,579$ trainable parameters; training in case of the most demanding PKU-MMD dataset requires on average approximately 50 min., while extraction of features requires approximately 0.004 s. Finally, the SVM that is used for classification requires approximately 420 s for training and approximately 0.008 s for classification of a given fused feature vector. Note that the aforementioned times may vary in every repetition of the experiment due to the early stopping.

4.3. *Evaluation protocol*

In case of PKU-MMD, experiments are divided into three parts: (a) *Single-view* experiments, wherein the same camera viewpoint has been used to create both training and testing sets (e.g. L viewpoint was

[c] https://numpy.org/.
[d] https://scipy.org/.
[e] https://opencv.org/.

used for both training/testing); (b) *Cross-view experiments*, wherein different camera viewpoints were used for training and testing. Note that up to two viewpoints may be used for training (e.g. *L* or *L* and *R* for training, *M* for testing); (c) *Cross-subject experiments*: subjects were split in training and testing groups, i.e. each one was a member of exactly one of these groups. The goal of the single-view experiments is to test the performance of the approach wherein only one camera is available, while the subject faces this camera. Since in real-life situations this is not always possible, cross-view experiments aim to evaluate the performance wherein the camera does not face the subject directly. This is typically occurring e.g. in ambient assistive living environments. Finally, the goal of cross-subject experiments is to test the robustness of our approach into intra-class variations, i.e. when training and testing sets have been created using different subjects. This is also expected to occur during a real-life application, wherein a system has been trained e.g. in a laboratory or using public datasets and is deployed into a real environment with previously unseen subjects. For each case, the accuracy achieved is measured.

In case of PKU-MMD single-view and cross-subject experiments, 87.5% of data have been used for training, while the remaining 12.5% for testing. In the former case, the dataset was randomly split, while in the latter case, the split that has been imposed by the authors of the dataset has been used. Moreover, in case of cross-subject experiments, in all cases all available data from a given viewpoint were used. In case of SYSU 3D HOI, the authors define a typical protocol (setting 1), wherein 50% of samples are used for training and 50% for testing and a cross-subject protocol (setting 2), wherein 50% of subjects are used for training and 50% for testing, without any overlap. Finally, the evaluation protocol of UTKinect-Action3D dataset is much simpler, wherein a leave one sequence out cross validation is indicated.

4.4. *Results and discussion*

Table 1 summarizes the results achieved for the classification of the 51 activities from the PKU-MMD dataset. More specifically, it summarizes results in terms of accuracy scores that has been achieved for the following cases of input data: (a) DFT; (b) FFT; (c) DCT; (d) DST; (e) pseudo-colored images (PCI); (f) dynamic images (Dyn.); (g) all transformations (i.e. cases a–d); (h) all transformations and pseudo-colored images; and (i) all available inputs. It is evident that in cross-view setup, best accuracy is achieved when using all image transformations, fused with the pseudo-colored images. On the other hand, in cross-subject and single-view setups, all available representations are necessary in order to achieve best accuracy. Notably, and contrary to previous

Table 1. Experimental results for PKU-MMD dataset. Numbers denote accuracy. Bold numbers indicate best result, per case. “PCI” denotes pseudo-colored images, “Dyn.” denotes dynamic images, “Tr.” denotes image transformations (i.e. DFT, FFT, DCT and DST). CV, CS, SV denote cross-view, cross-subject and single-view cases, respectively.

	Viewpoint		DFT	FFT	DCT	DST	PCI	Dyn.	Tr.	Tr. + PCI	All
Experiment	Train	Test	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.	Acc.
CV	*LR*	*M*	0.75	0.76	0.85	0.84	0.80	0.59	0.92	**0.98**	**0.98**
	LM	*R*	0.70	0.69	0.70	0.79	0.75	0.50	0.86	**0.95**	0.94
	RM	*L*	0.68	0.69	0.78	0.62	0.74	0.51	0.86	**0.95**	0.94
	M	*L*	0.64	0.63	0.68	0.74	0.59	0.47	0.85	**0.94**	0.92
	M	*R*	0.63	0.62	0.76	0.64	0.60	0.50	0.85	**0.93**	0.92
	R	*L*	0.58	0.58	0.66	0.46	0.59	0.28	0.78	**0.90**	0.87
	R	*M*	0.67	0.65	0.78	0.66	0.70	0.39	0.87	**0.95**	0.94
	L	*R*	0.58	0.59	0.40	0.64	0.55	0.30	0.74	**0.89**	0.86
	L	*M*	0.66	0.66	0.67	0.73	0.67	0.47	0.86	**0.95**	0.93
CS	*LRM*	*LRM*	0.70	0.69	0.79	0.79	0.76	0.80	0.85	0.92	**0.96**
SV	*L*	*L*	0.62	0.60	0.75	0.72	0.63	0.78	0.83	0.91	**0.95**
	R	*R*	0.62	0.61	0.75	0.72	0.63	0.80	0.82	0.92	**0.97**
	M	*M*	0.65	0.66	0.79	0.75	0.66	0.86	0.85	0.93	**0.97**

works that used only a single representation of skeletal data,[61,85] it is herein demonstrated that even when the viewpoint used for training differs significantly from the one used for testing, the proposed fusion approach is able to demonstrate comparable performance to all other cases. In those cases, i.e. *R–L* and *L–R*, achieved accuracies were 0.90 and 0.89, respectively, which were slightly lower than other cases, indicating the lack of performance gap as in previous works. In less challenging cross-view cases, performance ranged between 0.93 and 0.98. As expected from previous works, high accuracy values are achieved both in single-view (i.e. 0.95–0.97) and cross-view (i.e. 0.96) cases.

Moreover, it should be clear that fusion of several representations is able to significantly boost the performance. For example, in cross-subject case, the achieved accuracy using a single representation ranged between 0.70 and 0.80, while fusion allowed for an increase of 20%. Also, note that while dynamic images exhibit poor performance in cross-view cases, thus are unable to boost performance when fused with other modalities, they exhibit strong performance in single-view and cross-subject cases, which is the best per means of a single representations. Unsurprisingly, in those case they are able to provide a strong performance boost. This was expected due to the fact that the data augmentation approach has been used (Sec. 3.2) only in cases of image transformations and pseudo-colored images. Furthermore, another reason for this, is the loss of visual information in case of viewpoint changes; in the PKU-MMD dataset skeletons always comprise 25 joints. It should also be emphasized that as demonstrated in Table 1, the fusion of all transforms leads to improved accuracy compared to the cases of using a single one, meaning that the information extracted is complementary.

For the sake of comparison using the PKU-MMD dataset five state-of-the-art works, which to the best of our knowledge exhibit highest performances in the PKU-MMD dataset have been used. More specifically, the experimental results of this work are compared to the ones of Li *et al.*,[46] Li *et al.*,[48] Li *et al.*[49] and also to the more recent works of Guo *et al.*[22] and Sun *et al.*[79]

Comparative results are depicted in Table 2. Note that in this comparison we used mAP@50 instead of accuracy. The authors of PKU-MMD propose the use of the following two tasks: (a) cross-subject, which is the same task as the one that we have already demonstrated; and (b) cross-view, wherein the case that is considered is the one that *L* and *R* viewpoints are used for training, while *M* is used for testing purposes. As it may be seen in Table 2, the proposed approach demonstrated improved performance in the cross-view case, while it shows inferior performance, yet comparable, to the results of Li *et al.*[49] which reported best results in the cross-subject case.

Table 2. Comparison of the proposed approach to state-of-the-art research works using PKU-MMD dataset. Numbers indicate mAP (%). Bold indicate best results per task. CV, CS denote cross-view and cross-subject cases, respectively.

	Methodology					
Experiment	Proposed	Li *et al.*[49]	Li *et al.*[48]	Li *et al.*[46]	Guo *et al.*[22]	Sun *et al.*[79]
CV	**95.1**	94.4	94.2	93.7	—	94.6
CS	92.1	**92.9**	92.6	90.4	87.8	93.2

Table 3. Comparison of the proposed approach to state-of-the-art research works using SYSU 3D HOI dataset. Numbers indicate accuracy (%). Bold indicate best results per task.

	Methodology					
Experiment	Proposed	Zhang *et al.*[94]	Zhang *et al.*[92]	Zhang *et al.*[93]	Ke *et al.*[35]	Hu *et al.*[27]
Setting 1	**87.3**	86.9	85.1	85.7	—	79.6
Setting 2	**88.7**	86.5	84.8	85.7	83.3	84.9

Table 4. Comparison of the proposed approach to state-of-the-art research works using UT-Kinect Action3D dataset. Numbers indicate accuracy (%).

Methodology							
Proposed	Paoletti *et al.*[60]	Koniusz *et al.*[37]	Gao *et al.*[17]	Tang *et al.*[81]	Kao *et al.*[33]	Avola *et al.*[3]	Zhang *et al.*[95]
99.6	99.5	99.2	98.5	98.5	96.0	95.7	95.6

In case of the SYSU 3D HOI dataset, we compared our work with five state-of-the-art works which exhibit strong performance in both settings, namely, to the ones of Zhang *et al.*,[94] Zhang *et al.*,[92] Zhang *et al.*,[93] Ke *et al.*[35] and Hu *et al.*[27] Comparative results are depicted in Table 3, where it may be seen that the proposed approach demonstrated best performance in both settings of the dataset. Moreover, in case of the UTKinect-Action3D dataset, we compared our work with seven state-of-the-art works which exhibit strong performance, namely, to the ones of Paoletti *et al.*,[60] Koniusz *et al.*,[37] Gao *et al.*,[17] Tang *et al.*,[81] Kao *et al.*,[33] Avola *et al.*[3] and Zhang *et al.*[95] Comparative results are depicted in Table 4, where it may be seen that the proposed approach demonstrated best performance.

5. Conclusions and Future Work

In this paper, a multimodal fusion approach which targeted the problem of HAR from video data was proposed. Specifically, it was based on moving RGB and skeletal data, both captured by a depth camera. These modalities were used to create (a) 4 2D image representations, which were based on popular spectral transformations, i.e. the DFT, the FFT, the DCT and the DST; (b) a pseudo-colored image which is formed in order to capture inter-joint differences over time; and (c) a dynamic image which is used to provide a single-frame visual "summary" of a video sequence. Experiments have been mainly performed using a dataset of human motion activities, which was recorded with a multi-camera setup and conducted a three-fold evaluation, i.e. a single-view case where the same viewpoint was used for training/testing, a cross-view case where different viewpoints were used for training/testing and cross-subject case, where different subjects were used for training/testing.

Additional experiments were performed using two smaller-scale single camera datasets. A CNN was trained for each case, which then was used for feature extraction. Thus, in order to classify a given activity sequence, all the aforementioned 2D images were first extracted and then the 6 CNNs were used so as to extract features. The latter were fused and fed into a support vector machine for classification. The experimental evaluation indicated that the proposed approach may be successfully used for HAR in all the aforementioned cases. Moreover, it has been shown that the fusion of all representations is able to boost performance in the cross-subject and single-view cases, however in case of cross-view the dynamic images are not necessary, since they cause a small, yet significant drop of performance. Also, upon comparison of our approach with state-of-the-art approaches, superior performance in the cross-view case and comparable performance in the cross-subject case have been demonstrated.

We believe that the main advantage of this work is that it is not tied to a single modality. Instead it may be used with more than one modalities. Depending on the available hardware, different types of cameras may be used, while the extraction of skeletal data may be performed with any available methodology, although it is preferred to use 3D skeletal data. Also, a different deep network is used for feature extraction for each representation, allowing for faster training and extraction times. As demonstrated, the approach could be also used with single-camera datasets, without any modification, apart from training data and with RGB datasets (i.e. without depth information), although some performance loss is expected. The main limitation is the need for a multi-camera setup, in case of cross-view experiments. Also, as the majority of contemporary research works, it does not deal with spatial and/or temporal occlusion. However, as it will be later discussed, this could be a possible future research direction.

Plans for future work include investigation on methods for creating the signal image, possibly with the use of other types of sensor measurements such as wearable accelerometers, gyroscopes, etc. and evaluation of the proposed approach on several other public datasets, and for other types of activities. Within this process, a temporal augmentation approach[41] applied on skeletal data could replace the interpolation step herein used. Moreover, a further continuation of this work could involve other types of image representations of signals such as recurrence plots,[59] which have been successfully used with deep CNNs for classification of time-series,[24] deep architectures, such as LSTMs and hybrid CNN-LSTM networks, or techniques that work with sequences, such as seq2seq[80] or time-series such as multivariate CNNs.[53]

Since handcrafted features have been shown to boost recognition performance of deep approaches,[39] it would be interesting to experiment with other methodologies for feature extraction that are based on the geometry and the motion of skeletons.[3,95] Also, the use of modern classifiers could be investigated for classification. For example, instead of using an SVM, other possible research directions could include experiments with approaches such as Neural Dynamic Classification,[66] Dynamic Ensemble Learning[2] and Finite Element Machine for fast learning.[64]

Other aspects of HAR should be also investigated, such as dealing with incomplete data due to e.g. partial spatial or temporal occlusion. It is our belief that occlusion is one of the main factors that should be investigated in the continuation of our work. Due to the absence of datasets that contain occluded samples, a possible approach would be to artificially remove structured sets of joints, i.e. corresponding to body parts and apply our already trained models to assess the effect of occlusion.[18] Furthermore, a possible approach to deal with occlusion may be the use of regression on skeletal joints.[38] It is among our immediate goals to perform an evaluation into a real-like or even real-life assistive living environment. Therein and for privacy preservation issues, pose estimation approaches that do not depend on cameras may applied. A possible approach could be based on the use of Wi-Fi signals,[4] since Wi-Fi routers are typically encountered within any home environment. Finally another possible research direction would be to extend our approach to 3D data for real-world engineering cases.[50,51]

Acknowledgment

This research was funded by the European Union and Greece (Partnership Agreement for the Development Framework 2014-2020) under the Regional Operational Programme Ionian Islands 2014-2020, project title: Indirect costs for the project TRaditional corfU Music PresErvation through digiTal innovation, project number: 5030952'".

References

1. M. Abadi *et al.*, TensorFlow: A system for large-scale machine learning, in 12*th USENIX Symp. Operating Systems Design and Implementation* (*OSDI* 16) (Savannah, GA, USA/The USENIX Association, 2016), pp. 265–283.
2. K. M. Alam, N. Siddique and H. Adeli, A dynamic ensemble learning algorithm for neural networks, *Neural. Comput. Appl.* **32**(12) (2020) 8675–8690.
3. D. Avola, M. Cascio, L. Cinque, G. L. Foresti, C. Massaroni and E. Rodolá, 2D skeleton-based action recognition via two-branch stacked LSTM-RNNs, *IEEE Trans. Multimedia* **22**(10) (2020) 2481–2496.
4. D. Avola, M. Cascio, L. Cinque, A. Fagioli and G. L. Foresti, Human silhouette and skeleton video synthesis through Wi-Fi signals, *Int. J. Neural Syst.* **32**(5) (2022) 2250015.
5. H. Bilen, B. Fernando, E. Gavves, A. Vedaldi and S. Gould, Dynamic image networks for action recognition, in *IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Las Vegas, NY, USA, 2016), pp. 3034–3042.
6. C. Caetano, J. Sena, F. Brémond, J. A. Dos Santos and W. R. Schwartz, Skelemotion: A new representation of skeleton joint sequences based on motion information for 3D action recognition, in *IEEE Int. Conf. Advanced Video and Signal Based Surveillance* (*AVSS*) (IEEE, Taipei, Taiwan, 2019), pp. 1–8.
7. Z. Cao, T. Simon, S. E. Wei and Y. Sheikh, Realtime multi-person 2D pose estimation using part affinity fields, in *IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Honolulu, HI, USA, 2017), pp. 7291–7299.
8. A. Chaaraoui, J. Padilla-Lopez and F. Flórez-Revuelta, Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices, in *IEEE Int. Conf. Computer Vision Workshops* (IEEE, Sydney, Australia, 2013), pp. 91–97.
9. Y. Chen, C. Shen, X. S. Wei, L. Liu and J. Yang, Adversarial PoseNet: A structure-aware convolutional network for human pose estimation, in *IEEE Int. Conf. Computer Vision*, (IEEE, Venice, Italy, 2017), pp. 1212–1221.
10. Y. Chen, L. Wang, C. Li, Y. Hou and W. Li, ConvNets-based action recognition from skeleton

motion maps, *Multimedia Tools Appl.* **79**(3) (2020) 1707–1725.

11. F. Chollet *et al.*, Keras (2015), Available at: https://github.com/fchollet/keras.
12. C. Debes, A. Merentitis, S. Sukhanov, M. Niessen, N. Frangiadakis and A. Bauer, Monitoring activities of daily living in smart homes: Understanding human behavior, *IEEE Signal Process. Mag.* **33**(2) (2016) 81–94.
13. J. Deng, W. Dong, R. Socher, L. J. Li, K. Li and L. Fei-Fei, ImageNet: A large-scale hierarchical image database, in *Proc. IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Miami, FL, USA, 2009), pp. 248–255.
14. H. Duan, Y. Zhao, K. Chen, D. Lin and B. Dai, Revisiting skeleton-based action recognition, in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition* (IEEE, New Orleans Louisiana, USA, 2022), pp. 2969–2978.
15. M. Ehatisham-Ul-Haq, A. Javed, M. A. Azam, H. M. Malik, A. Irtaza, I. H. Lee and M. T. Mahmood, Robust human activity recognition using multimodal feature-level fusion, *IEEE Access* **7** (2019) 60736–60751.
16. B. Fernando, E. Gavves, J. M. Oramas, A. Ghodrati and T. Tuytelaars, Modeling video evolution for action recognition, in *IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Boston, MA, USA, 2015), pp. 5378–5387.
17. X. Gao, W. Hu, J. Tang, J. Liu and Z. Guo, Optimized skeleton-based action recognition via sparsified graph regression, in *27th ACM Int. Conf. Multimedia* (ACM, Nice, France, 2019), pp. 601–610.
18. I. Giannakos, E. Mathe, E. Spyrou and P. Mylonas, A study on the effect of occlusion in human activity recognition, in *14th Pervasive Technologies Related to Assistive Environments Conf.* (ACM, Corfu, Greece, 2021), pp. 473–482.
19. R. C. Gonzalez and R. E. Woods, *Digital Image Processing* (Pearson Education, 2018).
20. A. Graves A. R. Mohamed and G. Hinton, Speech recognition with deep recurrent neural networks, in *IEEE Int. Conf. Acoustics, Speech and Signal Processing* (IEEE, Vancouver, Canada, 2013), pp. 6645–6649.
21. G. Guo and A. Lai, A survey on still image based human action recognition, *Pattern Recognit.* **47**(10) (2014) 3343–3361.
22. T. Guo, H. Liu, Z. Chen, M. Liu, T. Wang and R. Ding, Contrastive learning from extremely augmented skeleton sequences for self-supervised action recognition, *AAAI Conf. Artif. Intell.* **36**(1) (2022) 762–770.
23. M. A. Haque *et al.*, Deep multimodal pain recognition: A database and comparison of spatio-temporal visual modalities, in *IEEE Int'l Conf. Automatic Face & Gesture Recognition* (IEEE, Xi'an, China, 2018), pp. 250–257.
24. N. Hatami, Y. Gavet and J. Debayle, Classification of time-series images using deep convolutional neural networks, in *Int. Conf. Machine Vision (ICMV)*, Vol. 10696 (SPIE, Munich, German, 2018), pp. 242–249.
25. V. Hernandez, T. Suzuki and G. Venture, Convolutional and recurrent neural network for human activity recognition: Application on American sign language, *PLoS ONE* **15** (2020) e0228869.
26. Y. Hou, Z. Li, P. Wang and W. Li, Skeleton optical spectra-based action recognition using convolutional neural networks, *IEEE Trans. CSVT* **28**(3) (2016) 807–811.
27. J. F. Hu, W. S. Zheng, J. Lai and J. Zhang, Jointly learning heterogeneous features for RGB-D activity recognition, in *IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Boston, MA, USA, 2015), pp. 5344–5352.
28. T. Huynh-The, C. H. Hua, T. T. Ngo, T. T. and D. S. Kim, Image representation of pose-transition feature for 3D skeleton-based action recognition, *Inf. Sci.* **513** (2020) 112–126.
29. J. Imran and B. Raman, Evaluating fusion of RGB-D and inertial sensors for multimodal human action recognition, *J. Ambient Intell. Humaniz. Comput.* **11**(1) (2020) 189–208.
30. A. K. Jain, *Fundamentals of Digital Image Processing* (Prentice-Hall, 1989).
31. W. Jiang and Z. Yin, Human activity recognition using wearable sensors by deep convolutional neural networks, in *23rd ACM Int. Conf. Multimedia* (ACM, Brisbane, Australia, 2015), 1307–1310.
32. G. Kalouris, E. I. Zacharaki and V. Megalooikonomou, Improving CNN-based activity recognition by data augmentation and transfer learning, in *Int. Conf. Industrial Informatics (INDIN)* (IEEE, Helsinki-Espoo, Finland, 2019), pp. 1387–1394.
33. J. Y. Kao, A. Ortega, D. Tian, H. Mansour and A. Vetro, Graph based skeleton modeling for human activity analysis, in *IEEE Int. Conf. Image Processing (ICIP)* (IEEE, Taipei, Taiwan, 2019), pp. 2025–2029.
34. Q. Ke, S. An, M. Bennamoun, F. Sohel and F. Boussaid, SkeletonNet: Mining deep part features for 3-D action recognition, *IEEE Signal Process. Lett.* **24**(6) (2017) 731–735.
35. Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel and F. Boussaid, Learning latent global network for skeleton-based action prediction, *IEEE Trans. Image Process.* **29** (2019) 959–970.
36. A. Keogh, J. F. Dorn, L. Walsh, F. Calvo and B. Caulfield, Comparing the usability and acceptability of wearable sensors among older irish adults in a real-

world context: Observational study, *JMIR mHealth uHealth* **8**(4) (2020) e15704.

37. P. Koniusz, L. Wang and A. Cherian, Tensor representations for action recognition, *IEEE Trans. Pattern Anal. Mach. Intell.* **44**(2) (2021) 648–665.
38. I. A. Kostis, E. Mathe, E. Spyrou and P. Mylonas, Human activity recognition under partial occlusion, in *Int. Conf. Engineering Applications of Neural Networks* (Springer, Creta, Greece, 2022), pp. 297–309.
39. D. Koutrintzes, E. Mathe and E. Spyrou, Boosting the performance of deep approaches through fusion with handcrafted features, in *Int. Conf. Pattern Recognition Applications and Methods — ICPRAM* (INSTICC, 2022), pp. 370–377.
40. H. Kuehne, H. Jhuang, E. Garrote, T. Poggio and T. Serre, HMDB: A large video database for human motion recognition, in *Int. Conf. Computer Vision* (IEEE, Barcelona, Spain, 2011), pp. 2556–2563.
41. T. Kwon, B. Tekin, S. Tang and M. Pollefeys, Context-aware sequence alignment using 4D skeletal augmentation, in *IEEE/CVF Conf. Computer Vision and Pattern Recognition* (IEEE, New Orleans Louisiana, USA, 2022), pp. 8172–8182.
42. I. Laptev, M. Marszalek, C. Schmid and B. Rozenfeld, Learning realistic human actions from movies, in *IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Anchorage, Alaska, USA, 2008), pp. 1–8.
43. Y. LeCun, L. Bottou, Y. Bengio and P. Haffner, Gradient-based learning applied to document recognition, *Proc. IEEE* **86**(11) (1998) 2278–2324.
44. C. Li, Q. Zhong, D. Xie and S. Pu, Skeleton-based action recognition with convolutional neural networks, in *IEEE Int. Conf. Multimedia & Expo Workshops* (*ICMEW*) (IEEE, Hong Kong, 2017), pp. 597–600.
45. C. Li, Y. Hou, P. Wang, P. and W. Li, Joint distance maps based action recognition with convolutional neural networks, *IEEE Signal Process. Lett.* **24**(5) (2017) 624–628.
46. C. Li, P. Wang, S. Wang, Y. Hou and W. Li, Skeleton-based action recognition using LSTM and CNN, in *IEEE Int. Conf. Multimedia & Expo Workshops* (IEEE, Hong Kong, 2017), 585–590.
47. X. Li, Y. Zhang, J. Zhang, S. Chen, I. Marsic, R. A. Farneth and R. S. Burd, Concurrent activity recognition with multimodal CNN-LSTM structure (2017), arXiv:1702.01638.
48. C. Li, Q. Zhong, D. Xie and S. Pu, Co-occurrence feature learning from skeleton data for action recognition and detection with hierarchical aggregation (2018), arXiv:1804.06055.
49. T. Li, L. Fan, M. Zhao, Y. Liu and D. Katabi, Making the invisible visible: Action recognition through walls and occlusions, in *Proc. IEEE/CVF Int. Conf. Computer Vision* (IEEE, Seoul, Korea, 2019), pp. 872–881.
50. Y. Liang, F. He and X. Zeng, 3D mesh simplification with feature preservation based on whale optimization algorithm and differential evolution, *Integr. Comput.-Aided Eng.* **27**(4) (2020) 417–435.
51. Y. Liang, F. He, X. Zeng and J. Luo, An improved loop subdivision to coordinate the smoothness and the number of faces via multi-objective optimization, *Integr. Comput.-Aided Eng.* **29** (2022) 23–41.
52. C. Liu, Y. Hu, Y. Li, S. Song and J. Liu. PKU-MMD: A large scale benchmark for continuous multi-modal human action understanding (2017), arXiv:1703.07475.
53. C. L. Liu, W. H. Hsaio and Y. C. Tu, Time series classification with multivariate convolutional neural network, *IEEE Trans. Ind. Electron.* **66**(6) (2018) 4788–4797.
54. J. Liu, N. Akhtar and A. Mian, Viewpoint invariant RGB-D human action recognition, in *Int'l Conf. Digital Image Computing*: *Techniques and Applications* (IEEE, Sydney, Australia, 2017), pp. 1–8.
55. M. Liu, H. Liu and C. Chen, Enhanced skeleton visualization for view invariant human action recognition, *Pattern Recognit.* **68** (2017), pp. 346–362.
56. J. Liu, A. Shahroudy, M. Perez, G. Wang, L. Y. Duan and A. C. Kot, NTU RGB+D 120: A large-scale benchmark for 3d human activity understanding, *IEEE Trans. Pattern Anal. Mach. Intell.* **42**(10) (2019) 2684–2701.
57. J. Liu, N. Akhtar and A. Mian, Skepxels: Spatio-temporal image representation of human skeleton joints for action recognition, *CVPR Workshops* (IEEE, Long Beach California, USA, 2019), pp. 10–19.
58. S. Majumder, T. Mondal and M. J. Deen, Wearable sensors for remote health monitoring, *Sensors* **17**(1) (2017) 130.
59. N. Marwan, A historical review of recurrence plots, *Eur. Phys. J. Spec. Top.* **164**(1) (2008) 3–12.
60. G. Paoletti, J. Cavazza, C. Beyan and A. Del Bue, Subspace clustering for action recognition with covariance representations and temporal pruning, in *Int. Conf. Pattern Recognition* (*ICPR*) (Springer, 2021), pp. 6035–6042.
61. A. Papadakis, E. Mathe, I. Vernikos, A. Maniatis, E. Spyrou and P. Mylonas, Recognizing human actions using 3D skeletal information and CNNs, in *Engineering Applications of Neural Networks. EANN 2019*, Communications in Computer and Information Science, Vol. 1000 (Springer, Cham, 2019), pp. 511–521.
62. A. Papadakis, E. Mathe, E. Spyrou and P. Mylonas, A geometric approach for cross-view human action recognition using deep learning, in *IEEE Int. Symp. Image and Signal Processing and Analysis* (*ISPA*) (IEEE, Dubrovnik, Croatia, 2019), pp. 258–263.

63. A. Papadakis, I. Vernikos, E. Mathe and E. Spyrou, Skeleton geometric transformation for human action recognition using convolutional neural networks, in *ACM Int. Conf. Pervasive Technologies Related to Assistive Environments* (ACM, 2020), pp. 1–2.
64. D. R. Pereira, M. A. Piteri, A. N. Souza, J. P. Papa and H. Adeli, FEMa: A finite element machine for fast learning, *Neural. Comput. Appl.* **32**(10) (2020) 6393–6404.
65. H. H. Pham, H. Salmane, L. Khoudour, A. Crouzil, P. Zegers and S. Velastin, Spatio–temporal image representation of 3D skeletal movements for view-invariant action recognition with deep convolutional neural networks, *Sensors* **19**(8) (2019) 1932.
66. M. H. Rafiei and H. Adeli, A new neural dynamic classification algorithm, *IEEE Trans. Neural Netw. Learn. Syst.* **28**(12) (2017) 3074–3083.
67. S. Ranasinghe, F. Al Machot and H. C. Mayr, A review on applications of activity recognition systems with regard to performance and evaluation, *Int. J. Distrib. Sens. Netw.* **12**(8) (2016) 1550147716665520.
68. C. Schuldt, I. Laptev and B. Caputo, Recognizing human actions: A local SVM approach, in *Int. Conf. Pattern Recognition* (*ICPR*) (IEEE, Cambridge, UK 2004), pp. 32–36.
69. V. Silva, F. Soares, C. P. Leao, J. S. Esteves and G. Vercelli, Skeleton driven action recognition using an image-based spatial-temporal representation and convolution neural network, *Sensors* **21**(13) (2021) 4342.
70. K. Simonyan and A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014), arXiv:1409.1556.
71. K. Simonyan and A. Zisserman, Two-stream convolutional networks for action recognition in videos, in *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems*, Vol. 1 (2014) 568–576.
72. A. J. Smola and B. Schölkopf, A tutorial on support vector regression, *Stat. Comput.* **14** (2004) 199–222.
73. K. Soomro, A. R. Zamir and M. Shah, UCF101: A dataset of 101 human actions classes from videos in the wild (2012), arXiv:1212.0402.
74. E. Spyrou, E. Mathe, G. Pikramenos, K. Kechagias and P. Mylonas, Data augmentation vs. domain adaptation — A case study in human activity recognition, *Technologies* **8**(4) (2020) 55.
75. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever and R. Salakhutdinov, Dropout: A simple way to prevent neural networks from overfitting, *J. Mach. Learn. Res.* **15**(1) (2014) 1929–1958.
76. O. Steven Eyobu and D. S. Han, Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network, *Sensors* **18** (2018) 2892.
77. S. Stylianou-Nikolaidou, I. Vernikos, E. Mathe, E. Spyrou and P. Mylonas, A novel CNN-LSTM hybrid architecture for the recognition of human activities, in *Int. Conf. Engineering Applications of Neural Networks* (Springer, 2021), pp. 121–132.
78. L. Sun, K. Jia, K. Chen, D. Y. Yeung, B. E. Shi and S. Savarese, Lattice long short-term memory for human action recognition, in *Int. Conf. Computer Vision* (IEEE, Venice, Italy, 2017), pp. 2147–2156.
79. J. Sun, B. Zhou, M. J. Black and A. Chandrasekaran, LocATe: End-to-end localization of actions in 3D with transformers (2022), arXiv:2203.10719.
80. I. Sutskever, O. Vinyals and Q. V. Le, Sequence to sequence learning with neural networks, in *NIPS'14: Proceedings of the 27th International Conference on Neural Information Processing Systems* Vol. 2 (2014) 3104–3112.
81. Y. Tang, Y. Tian, J. Lu, P. Li and J. Zhou, Deep progressive reinforcement learning for skeleton-based action recognition, in *IEEE Conf. Computer Vision and Pattern Recognition* (IEEE, Salt Lake City, Utah, USA, 2018), pp. 5323–5332.
82. N. Tasnim, M. K. Islam and J. H. Baek, Deep learning based human activity recognition using spatiotemporal image formation of skeleton joints, *Appl. Sci.* **11**(6) (2021) 2675.
83. T. Theoharis, G. Papaioannou, N. Platis and N. M. Patrikalakis, *Graphics and Visualization: Principles & Algorithms* (CRC Press, 2008).
84. P. Verma, A. Sah and R. Srivastava, Deep learning-based multi-modal approach using RGB and skeleton sequences for human activity recognition, *Multimedia Syst.* **26**(6) (2020) 671–685.
85. I. Vernikos, E. Mathe, A. Papadakis, E. Spyrou and P. Mylonas, An image representation of skeletal data for action recognition using convolutional neural networks, in *ACM Int. Conf. PErvasive Technologies Related to Assistive Environments* (ACM, Rhodes, Greece, 2019), pp. 325–326.
86. I. Vernikos, E. Mathe, E. Spyrou, A. Mitsou, T. Giannakopoulos and P. Mylonas, Fusing handcrafted and contextual features for human activity recognition, in *2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)* (2019), pp. 1–6, doi:10.1109/SMAP.2019.8864848.
87. M. Vrigkas, C. Nikou, C. and I. A. Kakadiaris, A review of human activity recognition methods, *Front. Robot. AI* **2**(28) (2015), https://doi.org/10.3389/frobt.2015.00028.
88. P. Wang, Z. Li, Y. Hou and W. Li, Action recognition based on joint trajectory maps using convolutional neural networks, in *ACM Int. Conf. Multimedia* (ACM, Amsterdam, The Netherlands. 2016), pp. 102–106.

89. P. Wang, W. Li, P. Ogunbona, J. Wan and S. Escalera, RGB-D-based human motion recognition with deep learning: A survey, *Comput. Vis. Image Underst.* **171** (2018) 118–139.
90. L. Xia, C. C. Chen and J. K. Aggarwal, View invariant human action recognition using histograms of 3D joints, in 2012 *IEEE Computer Society Conf. Computer Vision and Pattern Recognition Workshops* (IEEE, Providence, Rhode Island, USA, 2012), pp. 20–27.
91. Z. Yang, Y. Li, J. Yang and J. Luo, Action recognition with spatio–temporal visual attention on skeleton image sequences, *IEEE Trans. CSVT* **29**(8) (2018) 2405–2415.
92. P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue and N. Zheng, View adaptive recurrent neural networks for high performance human action recognition from skeleton data, in *IEEE Int. Conf. Computer Vision* (IEEE, Venice, Italy, 2017), pp. 2117–2126.
93. P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao and N. Zheng, EleAtt-RNN: Adding attentiveness to neurons in recurrent neural networks, *IEEE Trans. Image Process.* **29** (2019) 1061–1073.
94. P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue and N. Zheng, Semantics-guided neural networks for efficient skeleton-based human action recognition, in *IEEE/CVF Conf. Computer Vision and Pattern Recognition* (IEEE, Seattle, WA, USA, 2020), pp. 1112–1121.
95. S. Zhang, Y. Yang, J. Xiao, X. Liu, Y. Yang, D. Xie and Y. Zhuang, Fusing geometric features for skeleton-based action recognition using multilayer LSTM networks, *IEEE Trans. Multimedia* **2**(9) (2018) 2330–2343.
96. G. Zhu, L. Zhang, P. Shen and J. Song, Multimodal gesture recognition using 3-D convolution and convolutional LSTM, *IEEE Access* **5** (2017) 4517–4524.