applied sciences

MDPI

Article

Unsupervised 3D Motion Summarization Using Stacked Auto-Encoders

Eftychios Protopapadakis [1,*,†], Ioannis Rallis [1,†], Anastasios Doulamis [1,†], Nikolaos Doulamis [1,†] and Athanasios Voulodimos [2,†]

[1] School of Rural and Surveying Engineering, National Technical University of Athens, 9th, Heroon Polytechniou Str., 15773 Athens, Greece; irallis@central.ntua.gr (I.R.); adoulam@cs.ntua.gr (A.D.); ndoulam@cs.ntua.gr (N.D.)

[2] Department of Informatics and Computer Engineering, University of West Attica, Egaleo, 12243 Athens, Greece; avoulod@uniwa.gr

* Correspondence: eftprot@mail.ntua.gr; Tel.: +30-210-772-2664

† These authors contributed equally to this work.

Received: 19 October 2020; Accepted: 17 November 2020; Published: 20 November 2020

Abstract: In this paper, a deep stacked auto-encoder (SAE) scheme followed by a hierarchical Sparse Modeling for Representative Selection (SMRS) algorithm is proposed to summarize dance video sequences, recorded using the VICON Motion capturing system. SAE's main task is to reduce the redundant information embedding in the raw data and, thus, to improve summarization performance. This becomes apparent when two dancers are performing simultaneously and severe errors are encountered in the humans' point joints, due to dancers' occlusions in the 3D space. Four summarization algorithms are applied to extract the key frames; density based, Kennard Stone, conventional SMRS and its hierarchical scheme called H-SMRS. Experimental results have been carried out on real-life dance sequences of Greek traditional dances while the results have been compared against ground truth data selected by dance experts. The results indicate that H-SMRS being applied after the SAE information reduction module extracts key frames which are deviated in time less than 0.3 s to the ones selected by the experts and with a standard deviation of 0.18 s. Thus, the proposed scheme can effectively represent the content of the dance sequence.

Keywords: video summarization; deep learning; motion capturing

1. Introduction

One interesting procedure for video visual analysis is video content summarization, a technique which has received wide research interest in recent years due to its wide application spectrum. The scope of a video summarization algorithm is to find out a set of the most representative key-frames of a video sequence, taking into consideration salient events and actions on video content so as to form a short but meaningful synopsis [1]. The existing video summarization techniques abstract the input data using three different approaches [2]. The first is the so-called *representative key-frame selection* that creates video summaries through a collection of representative key frames [3]. *The key subshot-oriented approach* selects the representative subshots of key-frames to form the video synopsis [4]. Finally, *the key object detection method* decomposes the whole video sequence into several single frames, each revealing representative objects in a given video sequence [5].

In the context of performing arts, such as dance sequences, variations of human body signals and gestures are essential elements describing a storyline or choreography in a symbolic way [6]. One important aspect in the analysis is the extraction of the choreographic motifs since these elements provide a fine summarization of the semantic information encoded the overall storyline [7–9].

Automatic summarization of choreographic sequences is an important issue in computer graphics research due to the following reasons. First, labelling procedures are time-consuming and occasionally require feedback from experts since motion capturing data are often unlabelled. Second, spatio-temporal analysis demands the reduction of 3D motion data and thus the automatic definition of all important features in a dance sequence. Third, implementation of advanced classification algorithms, based for example on deep learning neural network structures [10] require a large amount of labelled training data. Therefore, unsupervised summarization methods are necessary of producing representative training samples especially when large amount of video content is available.

The recent achievements of deep machine learning [10] have been proven to be very effective for visual recognition especially in the context of motion primitive identification or for object detection and recognition on benchmarked datasets [11]. The main advance of deep learning compared to traditional shallow learning approaches is that the former can automatically extract a set of optimal features for classification (pre-training) by deeply process raw visual content and analyse it on a discriminatory basis. Instead, the traditional shallow learning methods exploit hand-crafted image descriptors in their analysis which is application sensitive.

However, few works can be found dealing with the identification of 3D moving subjects and extracting motion primitives from dance sequences, creating a summarized representation of a choreography. In general, video summarization within motion content exploits methods that receive as inputs 3D skeleton data, captured by motion capturing systems (i.e., Kinect, OptiTrak, VICON) representing choreographic primitives of a dancer's performance. In particular, the capturing system extracts 3D coordinates of salient humans' joints measured them in a global coordination system and then video summarization is carried out by processing these (x, y, z) data instead of the raw image pixels. Usually, representational models have been applied for performing the summarization of a dance such as the Sparse Modeling for Representative Selection (SMRS) algorithm [12] or its hierarchical implementation [6]. However, since there is a great redundancy both in space and time (many frames represent similar characteristics), these methods fail to effectively represent dance video sequences, especially when multiple actors (dancers) are performing.

To address the aforementioned difficulties, we introduce a novel unsupervised-driven summarization scheme for dance sequences. Our method first exploits a stacked auto-encoder (SAE) mechanism followed by representational algorithms for key frame extraction. The purpose of SAE is to compress the raw captured inputs (containing a significant amount of redundant information both in space and time) in a way that an optimal reconstruction is achieved from the compressed data. That is, the encoded data (e.g., the compressed ones) are reconstructed in a way to optimally represent the raw input signals [13]. Data compression can be achieved using other approaches, apart from SAE. The wavelet transform is one of these approaches [14]. It can be applied to identify the salient features and reduce the redundancy/irrelevancy in a deterministic process using a time-frequency decomposition. This yields sufficient results, depending on the selection of the mother wavelet. However, highly non-linear schemes, like neural networks can be more effective especially when the statistical properties of the signal are dynamically changed [15,16]. Yet, SAEs is a deep example of a highly non-linear compression scheme which, through an unsupervised training phase, can learn all important properties of the dance, handling efficiently variations in spatial and temporal redundancy.

The 3D skeletal coordinates are used for data sequence representation obtained using the VICON motion capturing interface. The 3D motion coordinates are propagated into a stacked encoder with the main purpose to produce a compressed input signal of low redundancy that can optimally characterize the dance sequence. Then, representational algorithms, such as the hierarchical SMRS, are implemented to perform the final summarization. This way, the performance is maximized since summaries are extracted on a compressed input signal instead of the redundant high-dimension input signal data.

Previous works [6,8,17] implemented summarization techniques to extract the synopsis of choreographic sequences. Our work exploits the reduction of the redundant raw input-data to

create a fine-grained representation. This is achieved by refining the input data using SAEs, so that any redundant information is discarded. Such an approach is very important especially when multiple dancers are present in the dance sequences, unlike to the previous works, which focus on the performance of a single dancer. The presence of multiple dancers make the analysis much more complicated due to (i) humans' joint occlusions (some joints of one dancer are not visible since they are occluded by the other dancers in the 3D space) and (ii) merging of some joints of the dancers together. Although, the VICON motion capturing system can extract the labels of the passive markers with respect to the dancers, in our setup, we have not considered these labels, making the problem more challenging.

Figure 1 shows an example of the geometric challenges that the presence of two dancers causes to our analysis. Looking these two dancers, the right hand of the left dancer is overlapped with the left hand of the right dancer. Another example is depicted in Figure 2. By looking at the fifth and sixth frame of the sequence, one can notice that only one dense body (dancer) executes the choreography (fourth row) while as it can be observed from the RGB content the dancers are two (third row). Thus, the application of conventional video summarization algorithms will yield to a failure. All these bottlenecks, that is, (i) *overlapping of the skeletal joints* and (ii) *redundancy of the raw input data* are addressed in this paper through the use of a combined SAE scheme followed by a hierarchical implementation of a SMRS.

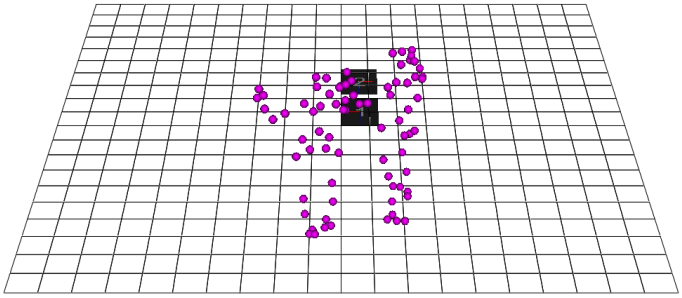

Figure 1. An example of geometric challenges due to the presence of multiple dancers.

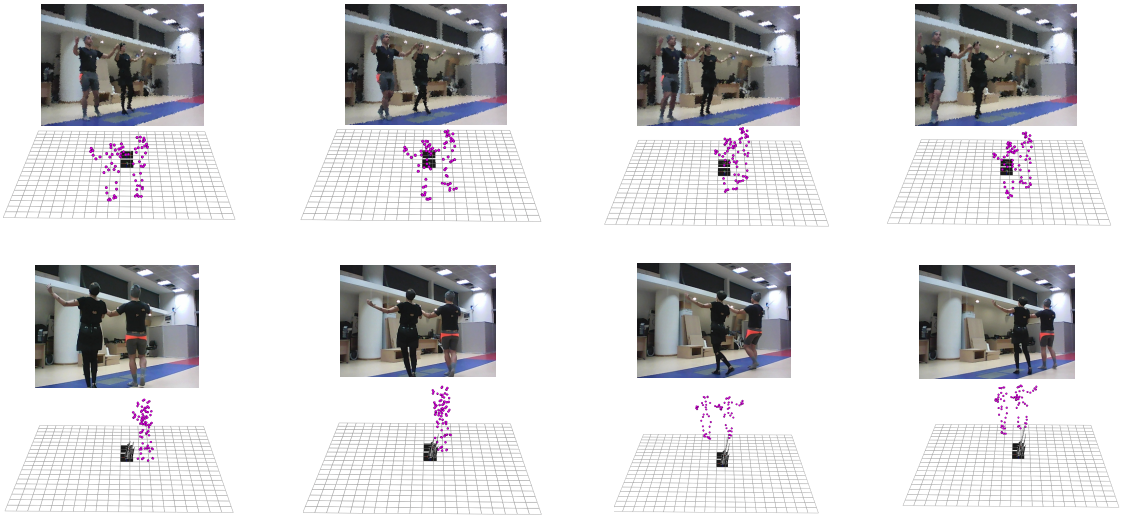

Figure 2. This visual sequences depict the motion capturing process. 3D skeletal data are obtained by the VICON motion capturing system (**second** and **fourth row**) and the respective RGB content (**first** and **third row**). This figure refers to Makedonikos dance sequence, executed by two dancers simultaneously.

This article is organized as follows: In Section 2, a description of the current state-of-the-art is given along with the proposed contribution of this paper (see Section 2.1). Section 3 gives an overview of the proposed summarization workflow which combines an SAE scheme with sampling algorithms.

The adopted SAE structure is discussed in Section 4. Section 5 presents the hierarchical sparse modelling representative selection algorithm, called H-SMRS. In Section 6, experiments are carried out using real-life dance sequences and objective criteria are proposed for a comparative evaluation for different summarization methods. Finally, Section 7 draws the conclusions of this paper.

2. Related Works

In general, video summarization techniques are distinguished into the following main categories [2]: (a) representative frame-based selection, (b) key-frame subshot detection and (c) key object detection algorithms. The representative frame-based selection focuses on identifying a series of discontinuous frames to comprise a synopsis that represents the whole video content as much as possible. In this context, References [18,19] propose different key frame extraction methods based on visual descriptors. On the other hand, Reference [3] performs the key frame extraction using only the temporal variations of a video sequence.

In the context of key-subshot detection, Reference [4,20] can be considered. These approaches extract short-time segments of a video as meaningful representations of its visual content. In Reference [21], an unsupervised video summarization algorithm is introduced that uses title-based image search results to find out shots of visual similarity. In Reference [22], the authors introduce a video summarization technique that decomposes the whole sequence into key objects. This representative selection problem is formulated as a sparse dictionary selection problem. Finally, Reference [23] identifies key-frames as a set of local interest points description and repeatability graph clustering. The selection of key frames is performed using graph clustering by approaching modularity principle.

In choreographic context, video summarization techniques use motion variations of spatio-temporal data in order to define the most representative key frames of the dance sequence. An example of this category is Reference [6] that applies the SMRS algorithm under a hierarchical modification for video dance summary. This method captures the variations or movements of each human action in different subspaces, which allow them to be represented as sequences of transitions from one subspace to another. This work is valid only for a single dancer while its performance when multiple dancers are present severely deteriorates. In Reference [24], the problem of learning motion primitives as one of temporal clustering is addressed deriving an unsupervised hierarchical bottom-up framework called hierarchical aligned cluster analysis (HACA).

HACA defines a partition of a given multidimensional time series into disjoint segments such that each segment belongs to one of clusters. HACA combines kernel k-means. In Reference [25], a robust method for detection and tracking human poses in videos is presented by matching video trajectories to a 3D motion capture model. The main novelty of this work resides in computing the correspondences between video and motion capture data. Reference [26] detects local minima in the temporal variation of the motion speed. The analysis is obtained by applying a low-pass filter to a one-dimensional motion speed data stream. In Reference [17], the authors have incorporated unsupervised clustering method for extracting key frames from choreographies. Toward this direction, classification of motion primitives of a dance using Long-Short Term Memory (LSTM) structures has been proposed in Reference [27].

In Reference [7], motion motifs and motion signatures are represented as a succinct but descriptive representation of motion sequences. Firstly, the motion sequences decomposed to short-term movements called motion words, and then the words are clustered in a high-dimensional feature space to find motion patterns. To this end, a deep learning architecture is exploited to embed the motion words into features. In Reference [28], an exploratory search system for large data collections of motion capture data is presented. The system provides an overview of human poses in a hierarchical dendrogram visualization that represents the result of a clustering procedure. A node-link diagram enables the user to analyze human poses as nodes, where each node shows a collection of similar pose instances.

2.1. Our Contribution

As we have previously stated, the main limitation of the aforementioned methods is that they apply the representational algorithms for dance summarization directly on the raw captured data, containing a significant amount of redundancy. Therefore, their performance is deteriorated, especially for long-dance video sequences. The redundancy problem is even more evident when multiple dancers are presence in a choreography, since the interactions among them may lead to a high confusion, as far as, the extracted key-frames are concerned. To address these issues, we introduce an SAE scheme prior to the representational sampling algorithms to reduce redundancy and, therefore, increase the dance summarization performance.

The paper compares the summarization performance using four sampling algorithms all applied over the SAE scheme's projected data. The results on real-world dance sequences, captured using two dancers performing, indicate that the proposed SAE-based redundancy reduction scheme can yield an effective repsentation of the dances sequences which on average deviates less than 0.30 s from the key-frames selected by dance experts (ground truth data) and with a standard deviation of about 0.18 s.

3. The Proposed Dance Summarization Workflow

Figure 3 presents the main architecture of the proposed unsupervised approach for dance summarization. Initially, from each (x, y, z) coordinates of a skeletal dancer's joint, kinematics attributes are extracted such as velocity and acceleration [8]. Then, the enhanced 3D motion primitives are forwarded into a stacked auto-encoder with the main purpose of compressing (encoding) the raw motion captured attributes into low dimensional representations. Encoding is performed in a way that the decoder is able to optimally reconstruct the raw input signals from the compressed ones, significantly reducing spatio-temporal redundancy [10,13]. The final module of the proposed architecture is the unsupervised representational algorithm for extracting the most importance key-frames of the dance sequence. The representational algorithm receives the low dimensional compressed data as inputs instead of the high redundant (both in space and time) raw signals, improving the overall summarization performance.

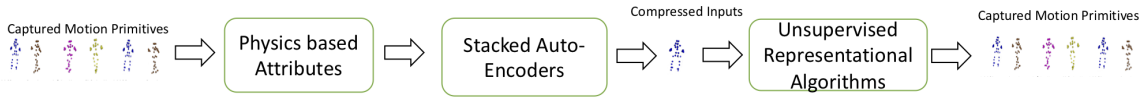

Figure 3. The proposed architecture for video dance summarization using stacked auto-encoders and representative algorithms.

3.1. Physics-Based Attributes of 3D Motion Primitives

In the following, let us denote as $\vec{J}_k^G(t) = (x_k^G(t), y_k^G(t), z_k^G(t))$ the k-th joint out of the M extracted by the Vicon architecture for each dancer for the t-th frame of the dance sequence. In our case $M = 40$, that is, 40 joints are extracted per human dancer. Variables $x_k^G(t)$, $y_k^G(t)$ and $z_k^G(t)$ indicate the coordinates of the k-th joint with respect to a reference point setting by the VICON architecture (in our case the center of the square surface) for the t-th frame. These joints have been obtained after the application of a density-based filtering on all the detected joints to remove noise from the acquisition process (see the third paragraph of Section 6). This noise becomes apparent when multiple dancers are performing in the choreography.

The main problem in directly processing the extracted joints $\vec{J}_k^G(t)$ is that they refer to the VICON coordination system which do not reflect the dancer's position in 3D space. For this reason, we first compute the center of the mass for each dancer and then the coordinates of $\vec{J}_k^G(t)$ is transformed to a local coordinate system, the origin of which coincides with the center of mass of a dancer, that is $\vec{J}_k^L(t) = \vec{J}_k^G(t) - \vec{C}_{cm}(t)$, where $\vec{C}_{cm}(t)$ denotes the center of mass of a dancer. As far as the

kinematics attributes is concerned, the velocity and the acceleration are taken into account. In particular, the velocity is given as $\vec{u}_k(t) = d\vec{J}_k^L(t)/dt$, while the acceleration as $\vec{\gamma}_k(t) = d\vec{u}_k(t)/dt$ for each detected human joint. Since velocity and acceleration are given through a derivative formula, their calculation is independent from local/global coordination system and thus they are independent of a global translation. Alternative, we could use global dancers' velocity along with small local velocities of the joints to improve the feature analysis. But in this paper, we prefer to concentrate on simpler features. Gathering all these features together a vector is constructed as $(\vec{J}_k^L(t), \vec{u}_k(t), \vec{\gamma}_k(t))$. In the aforementioned notation, we focus only on one dancer and thus we omit indices describing the dancers for clarity purposes.

Figure 2 show the humans' joints extracted both on RGB content (the first and the third row of Figure 2) and on a plane depicting the movement of the dancers in the space (second and fourth row of Figure 2). Since we have two dancers executing the choreography, it is clear that severe occlusions and merges are encountered, mainly due to the 3D geometry of the dancers. This is the case, for example, of the fifth and sixth frame of Figure 2 where one can notice, by observing the frame content, that only one dancer appears to perform.

3.2. The Proposed Stacked Auto-Encoder (SAE) Module for Dimensionality Reduction

The core idea of our SAE representation is to capture a meaningful content of the main patterns of the raw data inputs by discarding any redundant information, that is, any outlier in data samples which will not be justified well using that representation. The learning process is described simply as minimizing a loss function over a training set. But since no desired outputs are required, the whole process is unsupervised. That is, the desired outputs are the same with the inputs. The final results will be a representation of low dimensionality of the input data. Thus, an SAE works similar to a Principal Component Analysis (PCA) but under a non-linear framework. Figure 4 depicts the proposed SAE approach for input data dimensionality reduction. In the following Section 4, we analyze with more details the SAE structure adopted in this article.

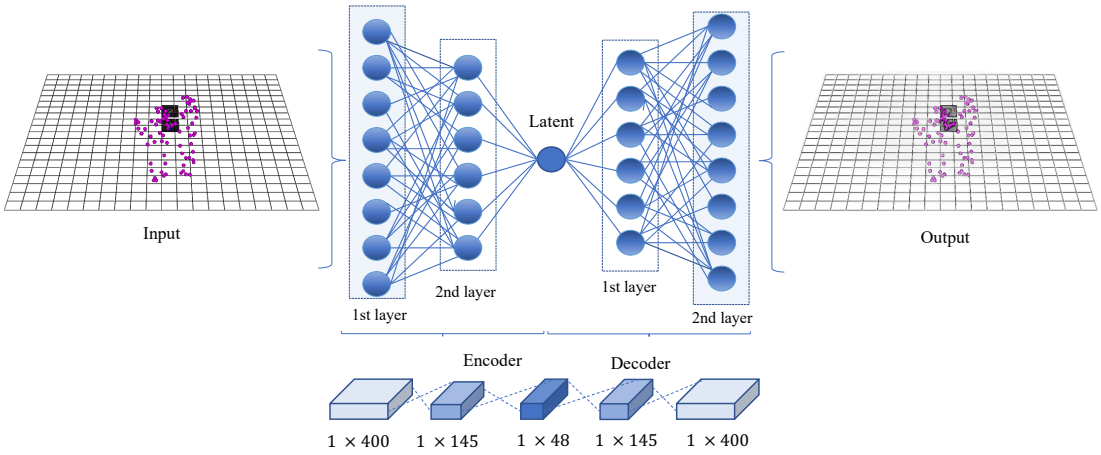

Figure 4. The structure of the proposed auto-encoder used for dimensionality reduction of the raw input signals.

3.3. Unsupervised Representational Sampling Algorithms

The last step of the proposed unsupervised video summarization algorithm employs traditional representational methods, such as the hierarchical SMRS [6], SMRS [12], K-OPTICS and Kennard Stone [29] for performing the final dance sequence summarization. K-OPTICS combines density-based and centroid based approaches [17,30]. The idea is implemented in a two step process. Start by clustering the available data using a centroid based approach, for example, k-means. Then, in each cluster run a density based approach, that is, OPTICS. The Kennard Stone (KenStone)

algorithm applied in order to generate a training set when no standard experimental design can be implemented. All samples are considered as candidates for the training set. The selected candidates are chosen sequentially.

Sparse Modelling for Representative Selection (SMRS) estimates correlations among different frames to extract the key ones. The principle of this scheme is to make the coefficient matrix as sparse as possible so as to achieve reconstruction of the whole dance sequence only from few data samples, that is the representative ones. In our recent work [6], a hierarchical implementation of the SMRS, called H-SMRS has been introduced. This hierarchical approach extracts a set of representative frames using the compressed input data under a hierarchical manner to take into account dance content complexity and fluctuations.

4. The Proposed Sae Scheme for Dance Sequence Summarization

The structure of the proposed SAE is depicted in Figure 4. As is observed, an SAE includes two modes of operations; the *encoding* and *decoding* mode. The goal of training is to minimize a loss function, say $L(\cdot)$ over a mean square error criterion. In particular, if x are the input data, then the loss function is expressed as $L(x, g(\beta(x)))$. In this notation $\beta(\cdot)$ is the overall non-linear function of the SAE encoder, whereas $g(\cdot)$ denotes the non-linear function of the decoder. Therefore the relationship $g(\beta(x)))$ denotes the operation of the encoding followed by the decoding.

In our particular implementation, three hidden layers are used for encoding phase. As we are moving deeper and deeper in the encoding hidden layers, the number of neurons that a hidden layer consists of is reduced. This forces the encoder to compress the input signals into a lower transformed versions of them. The input signal $\vec{x}_k \in R^n$ of the encoder are the kinematic driven attributes of 3D skeletal human's joint points (see Section 3.1). Variable n denotes the dimension of the input signal, that is, it is equal to the number of frames of the dance sequence N, by the number of joints per dance M, by the number of dancers D. That is, $n = N * M * D$. In our case, we focus on two dancers and on 40 humans' joints thus, $M = 40$ and $D = 2$. In addition, number N depends on the length of the dance sequence. In the current notation, we have omitted the dependence of the feature vector $\vec{x}_k$ on time t just for simplicity purposes.

The $\vec{x}_k$ triggers the first hidden layer to generate a transformed version of it of lower dimension. In particular, the output of the first hidden layer $\vec{h}_k^{(1)} \in R^{m^{(1)}}$ is given by

$$\vec{h}_k^1 = f(W_1^T * \vec{x}_k + \vec{b}_1), \tag{1}$$

where W_1 is the encoding weight matrix, $\vec{b}_1$ is the corresponding bias vector and $f(\cdot)$ the sigmoid vector-valued function. Variable $m^{(1)}$ denotes the dimension of the first hidden layer output signal. It is held that $m^{(1)} << n$ in order to yield a compressed version of the input signal $\vec{x}_k$.

In a similar way, the output of the second hidden layer transforms the hidden signals of the first layer (that is the $\vec{h}_k^{(1)} \in R^{m^{(1)}}$) into a further dimensionality reduced representation $\vec{h}_k^2 \in R^{m^{(2)}}$. Then, the ne woutput will be given as $\vec{h}_k^2 = f(W_2^T * \vec{h}_k^{(1)} + \vec{b}_2)$, where W_2 is the respective weight matrix of the second hidden layer, $\vec{b}_2$ the respective bias and again $f(\cdot)$ the sigmoid vector-valued function. It is held that $m^{(2)} << m^{(1)}$, so that a further compression is carried out. With the same way, the output of the second hidden layer $\vec{h}_k^2$ is propagated to the third hidden layer to produce a new reduced version $\vec{h}_k^3 \in R^{m^{(3)}}$ of the input signal with a much lower dimension $m^{(3)} << m^{(2)}$.

The parameters of the SAE, that is, the matrices W_i^T as well as the bias $\vec{b}_i$, are given through a training procedure minimizing a least square loss function $L(\cdot)$. The unsupervised operation of SAE is to generate as outputs, signals which are as close as possible to the input signals $\vec{x}_k$. This is achieved through minimization of the following loss function.

$$min \sum_{i=1}^{Q} L(\vec{x}_k, \hat{\vec{x}}_k), \tag{2}$$

where $\hat{\vec{x}}_k$ denotes the approximate version of the input signal $\vec{x}_i$ as generated by the encoder-decoder. This means that $\hat{\vec{x}}_k = g(\beta(\vec{x}_k))$. Training is performed over a set of Q samples of the same form of $\vec{x}_k$.

Dropout is used to reduce overfitting in the training process of neural networks. The overfitting problem is faced when the training dataset is small, which would result in a low accuracy on the test dataset. Dropout can randomly affect the neurons of the hidden layer to lose power in the training process. Technically, dropout is able to be achieved by setting the output date of some hidden neurons to 0 and then these neurons cannot be related to the forward-propagation algorithm.

5. The Hierarchical-Sparse Modelling Representative Selection

A hierarchical implementation of the Sparse Modelling Representative Selection (SMRS) algorithm, say H-SMRS [6], is adopted in this paper for key-frame extraction. The H-SMRS is applied on the compressed transformed signals, $\vec{h}_k^{(n)}$ of the encoding mode of SAEs instead of our previous works where this algorithm has been applied directly on the 3D attributes. This way, we discard redundant information existing in the data samples, a process which is very important especially in case where multiple humans are dancing in a sequence.

The proposed hierarchical scheme is based on the Sparse Modelling for Representative Selection (SMRS) algorithm [12] which reconstructs the N total frames of the dance sequence from K representatives. The optimization of the algorithm is achieved using the Alternative Direction Method of Multipliers (ADMM) [31]. Actually, this method comprises of iterative steps, taking into consideration the Lagrange multipliers.

The traditional SMRS algorithm is sensitive to temporal redundancies. Therefore, it fails to model the temporal dependencies of a choreography. To overcome this difficulty, we have introduced in Reference [6] a hierarchical decomposition scheme of the SMRS algorithm which first detects time intervals on which further decomposition takes place so as to create hierarchies of the key frame representatives. Thus, hierarchical SMRS segments the initial feature space into suitable sub-spaces that better model the choreography. The proposed H-SMRS is able to efficiently describe more complicated choreographic patterns, since the feature fluctuation within a sub-time interval (sub-space) is less than the fluctuation of the entire feature space of the sequence. Figure 5 presents an example of the hierarchical decomposition framework (H-SMRS). At the first layer, three representatives are extracted to model the whole video sequence (marked in green). Therefore, the initial video sequence is decomposed into four further sub-sequences (intervals), since the first and the last frame are also considered as representatives. Then, we assume that the third out of the fourth video sub-sequences. that is the interval $\Delta\tau(1,2)$, is further decomposed. $\Delta\tau(1,2)$ expresses the first layer at the second sub-sequence (interval). For this reason, the SMRS algorithm is applied within the interval $\Delta\tau(1,2)$ for extracting representatives that best fit the frames of this interval. In this example, two representatives are identified, again marked in blue color at layer 1. Therefore, the video segment of $\Delta\tau(1,2)$ is further decomposed into three more sub-segments. This procedure is iteratively applied until the decomposition criterion identifies that no further decomposition is required.

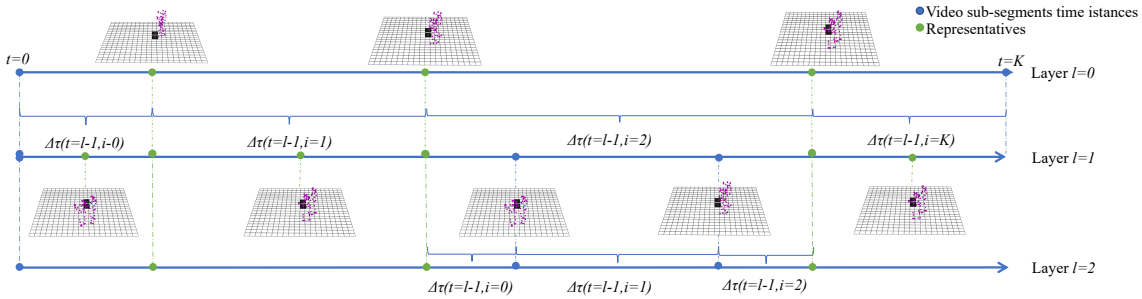

Figure 5. The architecture of the H-SMRS algorithm [6].

6. Experimental Results

In this section, we present several experiments to demonstrate the performance of the proposed unsupervised 3D motion summarization framework based on a stacked auto-encoder used to reduce the redundant information. The proposed stacked auto-encoder scheme is evaluated over three different dance sequences (see Section 6.2). Each choreographic sequence is executed by two humans, dancing simultaneously. We present several experiments to demonstrate (i) *the encoding capabilities* and (ii) *the similarity of the automatically selected frames against the ground-truth*.

As input data we use the ones presented in Section 3.1. That is, we extract for each human joint the relative coordinates and its kinematics, that is 5 elements (3 for the joint coordinates and two for the velocity and acceleration). We recall that we have 40 joints per human dancer. Thus, the total feature space is of dimension 400 (40 joints by 2 dancers by 3 coordinates per joint plus velocity and acceleration).

Due to the presence of two dancers in the sequences, a severe noise exists. To remove it, we first pre-process the data to exclude some frames which seem to be noisily represented. This is accomplished by just thresholding the differences of the joint coordinates among few consecutive frames. If this difference is greater than a threshold, this implies that a severe difference is noticed among the successive frames revealing an erroneous performance in 3D data encoding. A dancer (and thus his/her joint coordinates) cannot be moved long within the grid space during a choreographic performance. Having refined the captured data from potential noisy inputs, then we feed the features into the proposed SAE scheme to get a compressed input signal where all redundant information will be discarded.

Once, the stacked auto-encoder (see Equation (4)) is trained, we maintain the encoder part and project the feature values onto a latent space of lower dimension. In our experiments, we keep only 48, out of 400, feature element dimensions. This number has been selected after several experiments since it gives an acceptable performance while retaining the dimension as low as possible. A set of summarization approaches are applied, including the adopted unsupervised representational algorithms, along with other prominent methods such as k-OPTICS and Kennard Stone [29]. The last step of the analysis involves the calculation of similarity scores and the time divergence between the summarized frames and a set of selected key-frames by expert users in traditional dances (ground truth data sets). The former is calculated by the correlation scores between each frame of the original dance sequence to all the frames, provided by the sampling method. A higher score indicates a better match. Time divergence is simply calculated by the difference in frames, which is the same as the difference in times (seconds). In this case, the lower the difference is, the better the sumarization performs.

6.1. The Acquisition Module

The heart of the acquisition module adopted for modelling the dancers' motion trajectories in 3D space is based on the VICON Motion Capturing System. In our implementation, ten Bonita B3 cameras are used, running the Nexus 1.8.5.61009 h software. The movement area is 6.75 m^2. The origin of the VICON coordinate system is the centre of the square surface. A wand with markers is used to

calibrate the ten cameras. User body is measured by attaching passive markers on it at fixed positions for each dancer. After sticking all the markers, the height, weight and other specific anthropometric characteristics of the users are measured (see Figure 6). The data sets contain three recordings from Greek folklore dances, performed simultaneously by two professionals. We chose male and female expert-dancers since for those particular dances, the choreographic performance between men and women is slightly different. Specifically, men dance proud and imperious, while women modest and humble. On the contrary, dance style differences among professionals of the same gender are slight and mainly due to the personality of the dancer and how she/he executes the predefined choreographic performance.

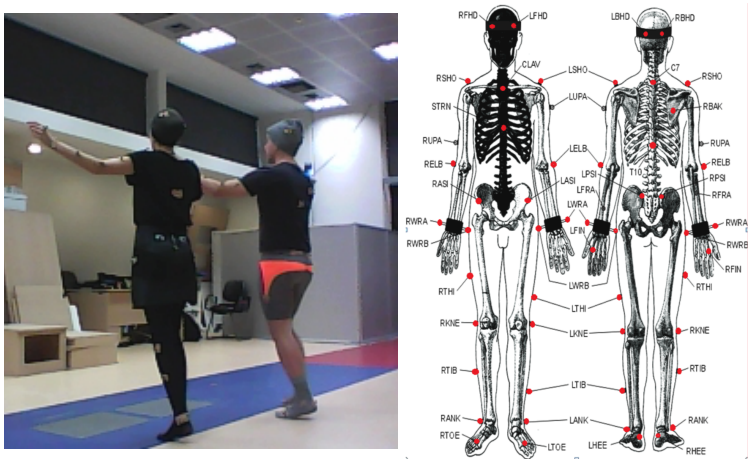

Figure 6. The motion capturing process takes into consideration the appropriate topology of the passive markers according to the VICON manual. 40 passive markers have been attached to the body of each dancer.

6.2. Dataset Description

Three dance sequences have been recorded using the VICON motion capturing platform [32]. These dance sequences refer to three different performances (dances), each executed simultaneously by two dancers (one male and one female). The recording process took place at the School of Physical Education and Sport Science of the University of Thessaly in Trikala Greece in January 2019. All sequences are Greek traditional folkloric dances, the selection of which was made by dance experts of traditional dances of the Schools of Sport Science of the Universities of Thessaloniki and Thessaly in Greece. The selection fulfils (i) different types of complexities in the dance main patterns, (ii) circular performances of the dance, (iii) different styles and (iv) different rhythmical tempos. All dancers are professional actors and each dance was executed twice per actor so as to record different paths of the same choreography. Figure 6 presents a photo of the environment used for the acquisition of the dance sequences using the VICON motion interface.

In Table 1, we also present a brief description of the dances along with their main steps. These steps have been defined by the dance experts who have designed the whole choreography and refer to the main variations of the dance as acquired through the VICON capturing module. Thus, the main steps of the dance in Table 1 do not refer to the steps of the choreography as being taught to a dancer trainee but to the main "activities" of the dance as being captured by the digitization unit. For instance, Sirtos (3-beat), consists, in its digital space, of six main choreographic units; (1) Initial Posture (IP)—the dancer faces a forward position; (2) Cross Leg (CL)—the dancer crosses the legs as she/he is moving, the left leg is in front of the right; (3) Initial Posture (IP)—again the dancer faces a forward position; (4) Left Leg Up (LLU)—the dancer rises her/his left leg up; (5) Initial Posture (IP)—after lowering her/his leg, the dancer is again in an in front position; (6) Right Leg Up (RLU)—the dancer rises her/his right leg up (see, Figure 7). Then, the main patterns of the dance stop and the choreography starts from scratch.

Table 1. A brief description of the recorded dances.

Type of Dance	Description	Main Choreographic Steps
Sirtos (3-Beat)	A Greek folklore dance in a slow rhythm performed by both women and men.	(1) Initial Posture (IP); (2) Cross Leg (CL); (3) Initial Posture (IP); (4) Left Leg Up (LLU); (5) Initial Posture (IP); (6) Right Leg Up (RLU)
Sirtos (2-Beat)	A Greek folkloric circular dance performed by both women and men, with a 7/8 musical beat.	(1) Initial Posture (IP); (2) Left Leg Back (LLB); (3) Cross Legs (CL); (4) Cross Legs (CL); (5) Cross Legs (CL); (6) Initial Posture (IP); (7) Right Leg Back (RLB);
Makedonikos	A Greek folkloric circular dance performed by both women and men, with a 9/8 musical beat.	(1) Initial Posture (IP); (2) Cross Legs Backwards (CLB); (3) Cross Legs (CL); (4) Left Leg Front (LLF); (5) Right Leg Back (RLB)

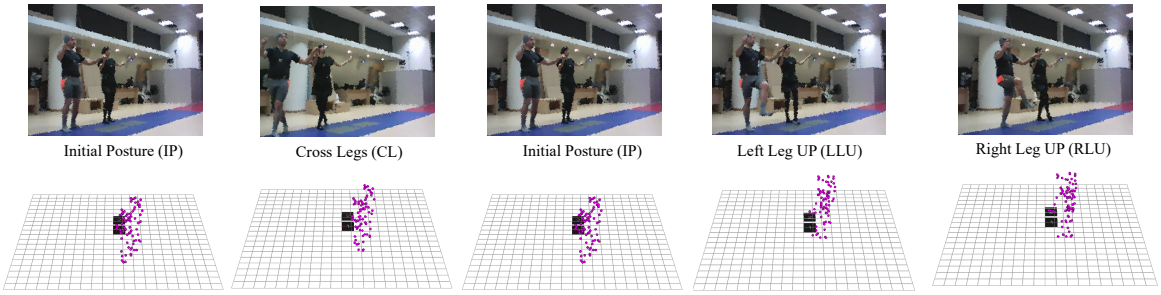

Figure 7. The main choreographic steps of Syrtos (3-beat) dance.

6.3. Evaluation Metrics

As we have stated above, ground truth data have been created by experts of Greek traditional dances. These experts are affiliated with the schools of sport science of the University of Thessaloniki and University of Thessaly in Greece. The ground truth data include a set of desired key frames, as being specified by the experts. Let us denote as $\vec{g}_l$ the selected key frames by the experts, with $l = 1, 2, ..., L$ where L is the number of representative frames as being indicated by the experts. We also symbolize as G the set containing all these selected frames, that is, $G = \{\vec{g}_1, \cdots, \vec{g}_L\}$. Let us also denote as $\vec{r}_k, k = 1, 2, .., K$ the extracted representative frames by any summarization algorithm and as $R = \{\vec{r}_1, \cdots, \vec{r}_K\}$ the respective set containing all K representatives extracted. Indices l, k are actually the frame instances of the ground truth key frames and the ones extracted by a summarization algorithm respectively. Thus, one objective criterion for evaluating the performance of a summarization scheme is to find, for each of the K extracted frames by an algorithm, the time instance (i.e., the frame index) of the experts' selected frame which is closest to the first one and then take the frame index difference of the ideal (experts' selected frame) and the extracted one. In other words,

$$\hat{l}(k) = \arg \min_{for\ all\ l \in G} |l - k| \quad \forall \vec{r}_k \in R, \tag{3}$$

where $\hat{l}(k)$ is the optimal frame index returned over all L selected frames in G for an examined extracted frame in R, say the k-th. We should notice that different extracted key frames $\vec{r}_k 1, \vec{r}_k 2$ with $k_1 \neq k_2$ may yield the same selected frame $\vec{g}_{\hat{l}(k)}$ meaning that some of the L selected frames may not correspond to any of the K extracted key frames. Then, the absolute difference $|\hat{l}(k) - k|$ describes how close is the k-th representative frame (by a summarization algorithm) to the closest ground truth one. In particular,

$$\mu = \frac{1}{K}\sum_{k=1}^{k=K} |\hat{l}(k) - k|$$
$$\mu_{max} = \max_{\forall k \in R} |\hat{l}(k) - k|, \tag{4}$$

where μ is the average time instance deviation among all K extracted representatives and μ_{max} the maximum deviation (worst case) among all K extracted frames.

Another criterion is to estimate how well all frames of a dance sequence can be reconstructed (represented) by the key frames. This is performed in our case by calculating the correlation coefficient of the feature vector for each frame of the dance sequence $\vec{x}_i, i = 1, ..., N$ against all representative frames $\vec{r}_k, k = 1, ..., K$.

$$\rho_i^{max} = \max_{\forall \vec{r}_k \in R} \rho(\vec{x}_i, \vec{r}_k) \;\; \forall \vec{x}_i, \tag{5}$$

where $\rho(\cdot)$ refers to the correlation coefficient of two vectors. The maximum the value ρ is the better the matching of that particular feature to a key frame. Thus, by taking the maximum value over all representative frames $\vec{r}_k$ as being set by a summarization algorithm, we estimate the best relation of any frame of the dance sequence to the extracted representatives. If this correlation is high, then the extracted key frames can well represent all frame sequences. Instead a small maximum correlation for some frames means that these cannot be reliably reconstructed by the key representatives.

6.4. Dance Summarization Experiments

In this sub-section, we present some results of different summarization algorithms on the above-mentioned dance sequences. In particular, Figure 8 demonstrates the results obtained on Syrtos (2 beat) dance sequence, consisting of more than 5000 frames, using as summarization algorithm the K-OPTICS. More specifically, we extract 32 key-representatives using the K-OPTICS algorithm and then we calculate the maximum correlation score ρ_i^{max} for each frame of Syrtos (2 beat) dance sequence against the 32 key frames extracted [see Equation (5)]. As shown in Figure 8, the average ρ_i^{max} for all 5000 frames (that is for all $i \in N$) is 0.5 with a variance of 0.25, which is a relatively low score. However, as we have stated previously, some frames of the dance sequence have been erroneously encoded mainly due to the simultaneous presence of two dancers in the choreography and the dense occlusions this causes. Thus, if we refine the frames of the dance sequence by excluding the ones whose the joint coordinates between two consecutive frames present high differences, greater than a threshold (in our case the threshold is set to 20% rate of change in joint's coordinates, for more than 20% of joints), then the correlation score is significantly improved. In particular, in this case the average ρ_i^{max} for all 5000 frames becomes more than 0.6, indicating a good summarization ability. Additionally, the majority of excluded frames, shown as purple crosses in Figure 8 can be found bellow the average similarity score. Such an outcome suggests that the applied rules for corrupted frames removal are adequate for the problem at hand.

Figure 9 illustrates the summarization performance when the Kennard Stone sampling algorithm is applied over Syrtos (3 beat) dance sequence. Again, as in Figure 8, the non-corrupted frames achieve a high average similarity score, close to 0.67, indicating that the summarized sequence can adequate describe (correlate) most of the originally captured frames. The fluctuations are also limited, and appear around frame 1500.

Table 2 summarizes the maximum correlation coefficients scores before and after the exclusion of the corrupted frames for all the three dances and the four examined sampling algorithms. It can be seen that the correlation scores obtained is about 0.6 revealing a satisfactory performance of the key frames as representatives of the whole dance sequence variation. In this table, we have presented as bold the highest correlation values.

Figure 10 demonstrates the average differences in frames (time instances) between a frame selected using a specific sampling approach (i.e., a summarization algorithm) and the experts' selected frames

(ground truth), for a particular dance. That is, the criterion μ of Section 6.3. Since the the frame rate of the system is 120 fps, a value of 50 indicate that the sampling approach generates frames less than half-a-second earlier/latter compared to the experts' selection. The impact of using raw against encoded data is, also, assessed. Results indicate that SMRS based approaches perform better to the other summarization schemes, for both raw and encoded data, when we have a single dancer sequence.

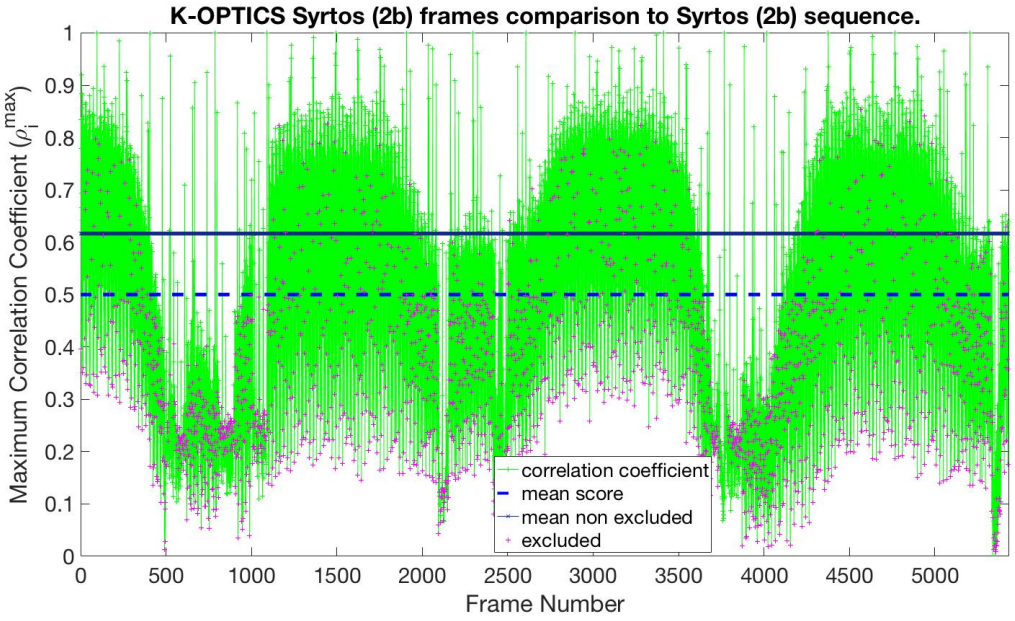

Figure 8. The maximum correlation scores ρ_i^{max} for each frame of the original Syrtos at 2 beat dance sequence compared to the summarized one using K-OPTICS.

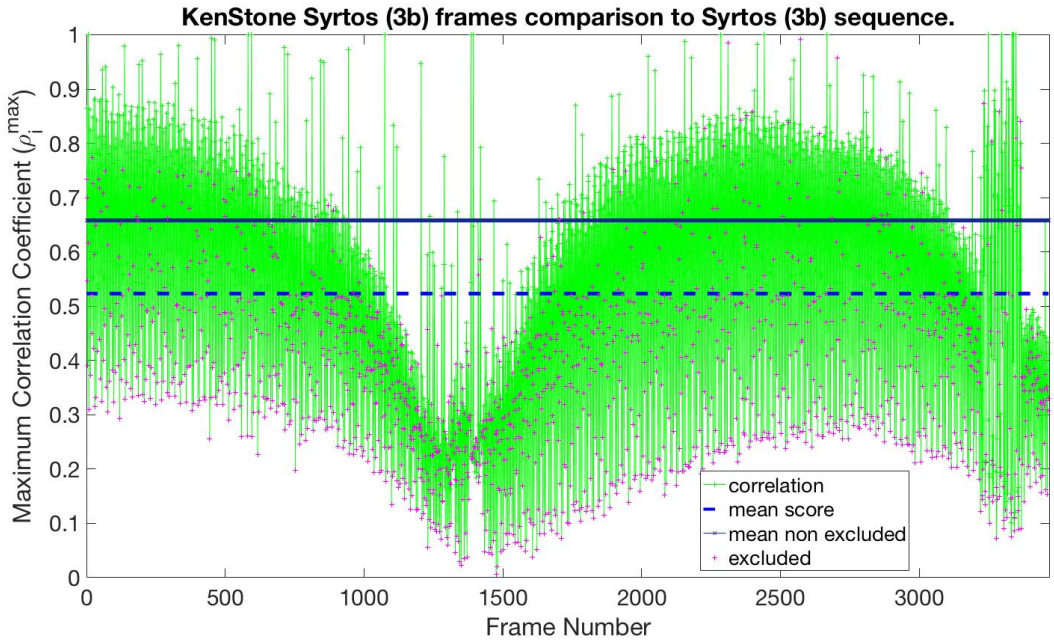

Figure 9. The maximum correlation scores ρ_i^{max} for each frame of the original Syrtos at 3 beat dance sequence compared to the summarized one using Kennard Stone.

Table 2. Maximum correlation coefficient scores (ρ_i^{max}) for different sampling algorithms and dance sequences.

Dance Sequence	Max Correlation Without Corrupted Frames	Max Correlation with Corrupted Frames	Sampling Summarization Algorithm
Makedonikos	0.64	0.52	KenStone
	0.61	0.47	K-OPTICS
	0.65	0.53	SMRS
	0.65	0.53	H-SMRS
Syrtos (2-beats)	0.30	0.29	KenStone
	0.64	0.51	K-OPTICS
	0.57	0.43	SMRS
	0.57	0.43	H-SMRS
Syrtos (3-beats)	0.63	0.50	KenStone
	0.60	0.48	K-OPTICS
	0.57	0.43	SMRS
	0.57	0.43	H-SMRS

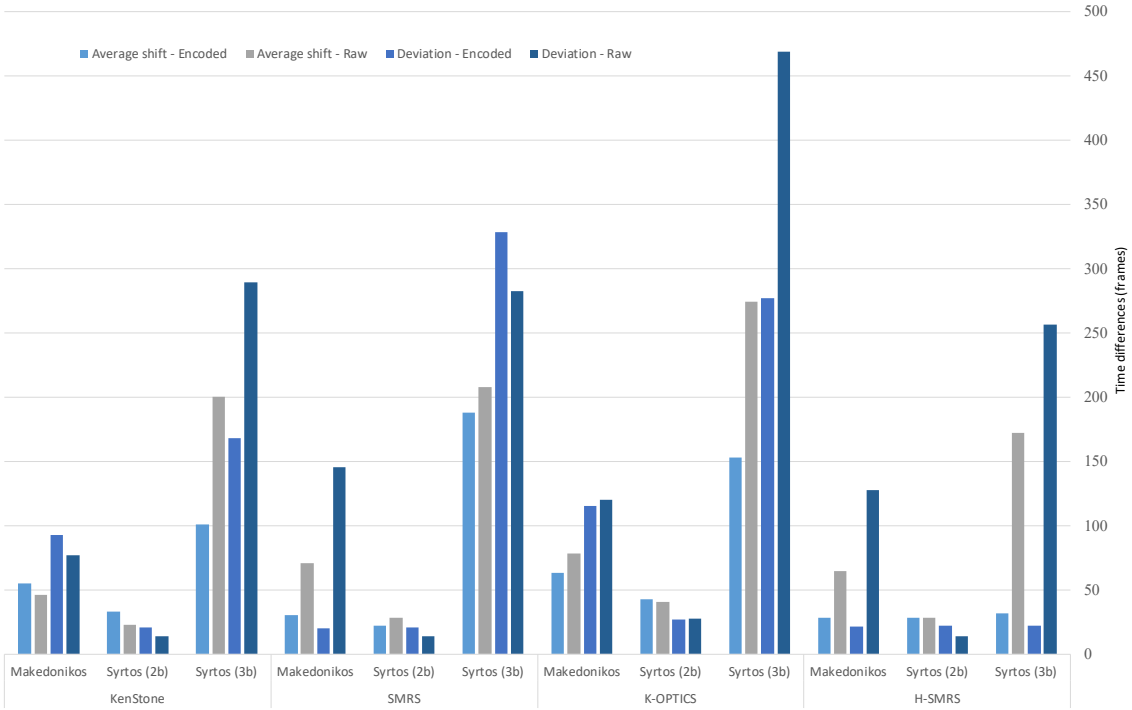

Figure 10. Data input type summarization impact when two dancers performed simultaneously for all the examined dance sequences.

In this figure, we also compare the performance derived against the four summarization methods; that is, K-OPTICS, Kennard Stone, SMRS, and the proposed hierarchical SMRS, H-SMRS. As we can observe from Figure 10, the H-SMRS gives the best performance for all dances with a deviation around 50 frames (or, approximately, 0.41 s), when encoded frames are used as inputs. The H-SMRS scheme also provides much better performance for the Syrtos(3b) dance, which seems to be more complicated than the other two dances, resulting in higher time deviations for the rest of the samplers. It is also worth mentinign the complex effect of coupling different features and samplers. For example, Syrtos(2b) input type does not affect significantly the performance for all four samplers.

Table 3 shows the average time deviation of key frames extracted by the four summarization algorithms and the ground truth data, that is, the value μ, measured, however, in seconds and not in frame index differences just for clarity. As is observed, the best performance is given for the the H-SMRS algorithm when the SAE scheme is used. In particular, the highest deviation of the H-SMRS

is achieved for the Syrtos (3b) equal to 0.26 s deviation on average which is in fact a very small deviation value. Similar performances of 0.23 and 0.24 sec deviations is also noticed for the other two dances. In the same table, we also present the standard deviation of the time shift to the ground truth data to show how these values vary. Again, H-SMRS yields the smallest standard deviation values which is about 0.18 s using the SAE, revealing its robustness against the other compared summarization algorithms.

Table 3. Average time shift among the summarization outcomes and the experts' annotations with and without the Stacked Auto-Encoder (SAE)-based data compression scheme.

Summarization Algorithm	Dance	Average Shift with SAE	Average Shift without SAE	Standard Deviation with SAE	Standard Deviation without SAE
KenStone					
	Makedonikos	0.46	0.38	0.78	0.64
	Syrtos (2b)	0.27	0.19	0.17	0.12
	Syrtos (3b)	0.84	1.67	1.4	2.41
SMRS					
	Makedonikos	0.25	0.59	0.17	1.21
	Syrtos (2b)	0.19	0.23	0.17	0.12
	Syrtos (3b)	1.03	1.73	2.74	2.35
K-OPTICS					
	Makedonikos	0.53	0.65	0.96	1
	Syrtos (2b)	0.34	0.36	0.23	0.23
	Syrtos (3b)	1.28	2.29	2.31	3.91
H-SMRS					
	Makedonikos	0.24	0.54	0.18	1.06
	Syrtos (2b)	0.23	0.24	0.18	0.12
	Syrtos (3b)	0.26	1.44	0.19	2.14

In the same table, we illustrate the results without using the SAE scheme. All summarization approaches, except KenStone algorithm, provide better results when the SAE-based compression framework is adopted. We get better scores in both average time shift and standard deviation, compared to the expert's annotated frames. For the Kenstone algorithm and only for two out of three dances, the performance remains, approximately the same, regardless of using or not the SAE.

Table 4 shows how much the average time shift of the four examined summarization algorithms and the ground truth data is improved when the SAE-based compressed scheme is applied on the raw 3D data in case of Syrtos (3b) dance sequence. The results have been depicted for two different executions of the dance, one with a single dancer and one with two dancers. It is observed that in case of a two dancers' performance the improvement ratio is much greater than the single dancer performance execution. Moreover, the adoption of the H-SMRS combined with SAE schema exhibits great improvement which reaches 81.80%.

Table 4. The improvement ratio among the adopted summarization algorithms with and without the SAE framework for Syrtos (3b) dance sequence. Two different performances of the dance are assumed; one for a single dancer and one for two dancers.

Summarization Algorithm	Aver. Shift Without SAE (Single Dancer)	Aver. Shift With SAE (Single Dancer)	Improvement Ratio (Single Dancer)	Aver. Shift Without SAE (Two Dancers)	Aver. Shift With SAE (Two Dancers)	Improvement Ratio (Two Dancers)
KenStone	0.51	0.47	6.96%	1.67	0.84	49.79%
K-OPTICS	0.51	0.51	0.67%	1.28	2.29	79.15%
SMRS	0.47	0.41	11.41%	1.73	1.03	40.55%
H-SMRS	0.45	0.31	31.15%	1.44	0.26	81.80%

Figure 11 provides further insights on the similarity among extracted key frames, using summarization algorithms, and some user annotated (selected) key frames. This allows us to *visually* judge on the similarity between the key frames extracted by the summarization algorithms and the ground truth ones. The results demonstrate five basic postures from Makedonikos dance. Then, for each the four summarization approaches, we select the closest frame to the user annotated posture of reference. As is observed, H-SRMS selections are closer to the experts' defined key frames, compared to K-OPTICS, SMRS, and KenStone approaches.

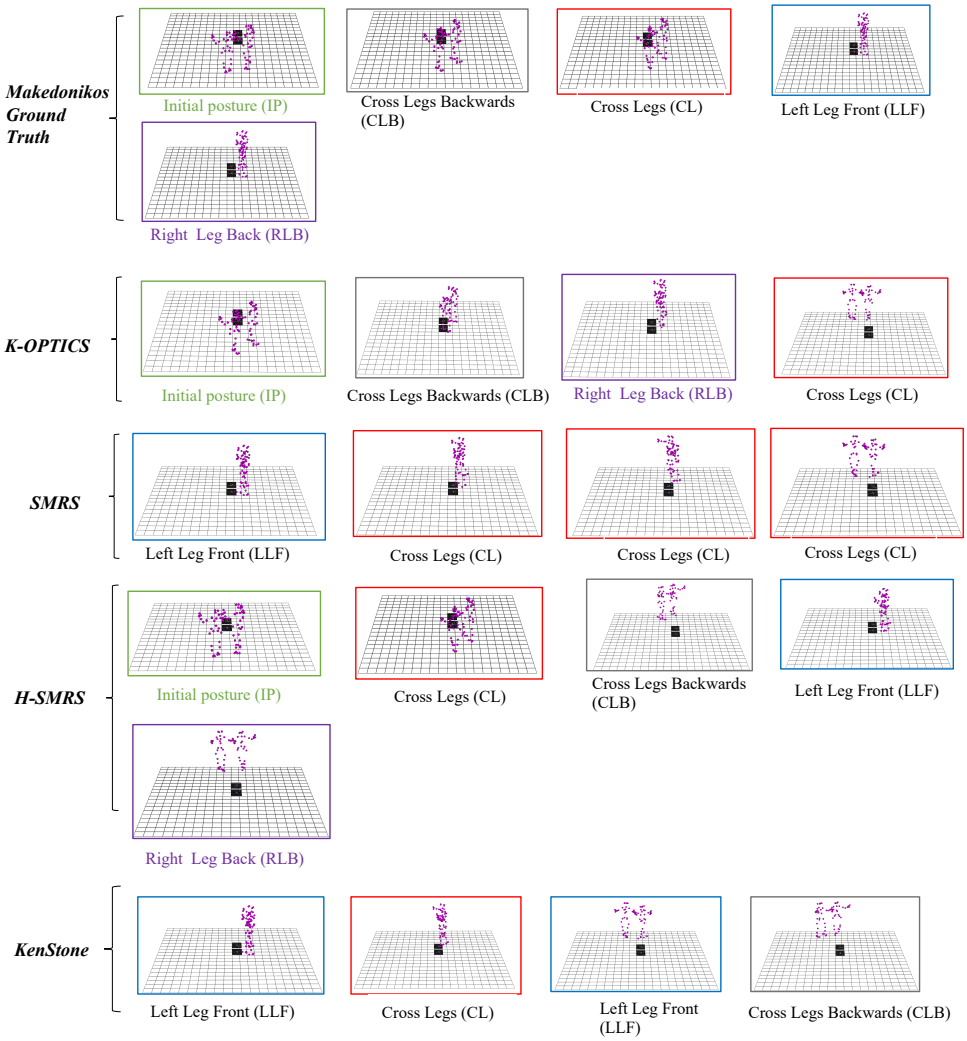

Figure 11. A visual representation of the key frames extracted by the four summarization algorithms than the ground truth ones in case of Makedonikos dance.

Figure 12 demonstrates the encoding capabilities for the adopted SAE scheme. Recall that 400 values have been reduced to 48 and then reconstructed back using SAEs. As shown, the representation of the decompressed data [see Figure 12a] are close to the original skeletal data [see Figure 12b] and maintain the two body postures and the general body form while the great compression (we retain only 48 joints than the 400 total ones). However, upper limps' joints positions have been gathered towards the body core. However, a better representation could be feasible by increasing the training epochs, which due to the limited training samples, that is, dance frames, does not affect significantly the training times.

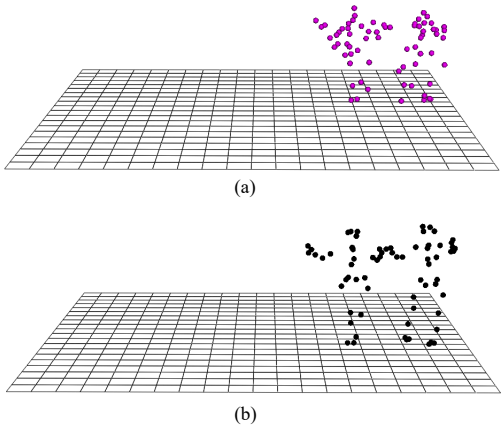

Figure 12. A representation of the decompressed data (**a**) relative to the original skeletal data (**b**), for the same time frame as for Syrtos (2b) dance sequence.

Another important criterion is how results vary (fluctuate) from the average values, as depicted in Figure 10. This is also illustrated in Table 3 where the standard deviation of the average time shift is given. But in Table 5 we also present the minimum (best) and the maximum (worst) performance [that is, μ_{max} of Equation (4)] for all the three dances. As we can see, μ_{max} reaches 0.72 s for the most difficult Makedonikos dance in case of H-SMRS. For the other two dances the worst (maximum) deviation is of about 0.5 s for the H-SMRS indicating an excellent summarization performance which is much smaller than the other summarization schemes. Regarding the minimum difference, all the summarization schemes yields excellent performance. This means that the best results obtained are very satisfactory.

Table 5. The minimum (best) and maximum (worst) time deviation (μ_{max}) among the key frames extracted using a summarization algorithm and the ground truth data. The comparison is carried out using four summarization algorithms, K-OPTICS, Kennard Stone, SMRS and H-SMRS and for the three dances.The values are in seconds.

Dance	Minimum Difference	Maximum Difference	Sampling Summarization Algorithm
Makedonikos	0.06 s	5.20 s	KenStone
	0.04 s	6.71 s	K-OPTICS
	0.04 s	6.66 s	SMRS
	0 s	0.72 s	H-SMRS
Syrtos (2-beats)	0.008 s	4.45 s	KenStone
	0.016 s	3.88 s	K-OPTICS
	0.016 s	0.5 s	SMRS
	0 s	0.74 s	H-SMRS
Syrtos (3-beats)	0.041 s	0.54 s	KenStone
	0.116 s	0.808 s	K-OPTICS
	0.033 s	0.541 s	SMRS
	0 s	0.55 s	H-SMRS

7. Conclusions

In this paper, we propose a deep stacked auto-encoder scheme followed by a hierarchical Sparse Modelling for Representative Selection (H-SMRS) summarization algorithm for performing accurate synopses of dance sequences. The sequences have been recorded through a motion capturing framework such as of VICON which produces 3D point joint of the dancers. The originality of this paper lies in the fact that our recorded dance sequences consist of two dancers performing simultaneously. This causes severe and intense errors in capturing the humans' joints in 3D coordination space. Thus, we adopt a stacked auto-encoder (SAE) scheme to reduce the redundant information of the 3D

point joints and thus improve the performance of the summarization than applying the summary algorithms directly on the raw captured data.

Regarding summarization, this paper compares the results using four key frame extraction algorithms. The K-OPTICS scheme, the Kennard Stone, the conventional SMRS and its hierarchical representation called H-SMRS. Our approach has been evaluated over three real-world dance sequences, each executing by two dancers. The results achieved show that the H-SMRS outperforms the other three algorithms for all the examined dance sequences. More specifically, the average time deviation is less than 0.3 s compared to ground truth selected frames being annotated by dance experts. Even in its worst performance, H-SMRS yields at least 0.72 s time deviations which is an excellent result. The proposed SAE approach also reduces the time required for executing the summarization algorithms than applying the summarization schemes directly on the raw data. This way, summarization become applicable to many engineering scenarios.

Author Contributions: Conceptualization, E.P., I.R. and N.D.; Formal analysis, E.P. and A.V.; Funding acquisition, N.D. and A.D.; Investigation, I.R.; Methodology, E.P. and A.V.; Resources, N.D.; Software, I.R. and E.P.; Supervision, A.V., N.D. and A.D.; Validation, A.D. and I.R.; Visualization, I.R.; Writing—original draft, I.R. and E.P.; Writing—review & editing, N.D., A.D. and A.V. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Hellenic Foundation for Research Innovation grant No. HFRI-FM17-2972. This paper is supported by the research project: 4DBeyond: 4D Analysis Beyond the Visible Spectrum in Real-Life Engineering Applications'.

Acknowledgments: This paper is supported by the research project: 4DBeyond: 4D Analysis Beyond the Visible Spectrum in Real-Life Engineering Applications, project No. HFRI-FM17-2972 funded by the Hellenic Foundation for Research Innovation.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

SAE	Stacked Auto-Encoder
H-SMRS	Hierarchical Sparse Modelling Representative Selection
SMRS	Sparse Modelling Representative Selection
K-OPTICS	Kmeans-Ordering points to identify the clustering structure
CNNs	Convolutional Neural Networks
LSTM	Long Short-Term Memory
HACA	Hierarchical Aligned Cluster Analysis
PCA	Principal Component Analysis
KS	Kennard Stone

References

1. Avrithis, Y.S.; Doulamis, A.D.; Doulamis, N.D.; Kollias, S.D. A stochastic framework for optimal key frame extraction from MPEG video databases. *Comput. Vis. Image Underst.* **1999**, *75*, 3–24. [CrossRef]
2. Zhang, Y.; Liang, X.; Zhang, D.; Tan, M.; Xing, E.P. Unsupervised object-level video summarization with online motion auto-encoder. *Pattern Recognit. Lett.* **2020**, *130*, 376–385. [CrossRef]
3. Doulamis, A.D.; Doulamis, N.; Kollas, S. Non-sequential video content representation using temporal variation of feature vectors. *IEEE Trans. Consum. Electron.* **2000**, *46*, 758–768. [CrossRef]
4. Doulamis, A.D.; Doulamis, N.D. Optimal content-based video decomposition for interactive video navigation. *IEEE Trans. Circuits Syst. Video Technol.* **2004**, *14*, 757–775. [CrossRef]
5. Vasconcelos, N.; Lippman, A. A spatiotemporal motion model for video summarization. In Proceedings of the 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231), Santa Barbara, CA, USA, 23–25 June 1998; pp. 361–366.
6. Rallis, I.; Doulamis, N.; Doulamis, A.; Voulodimos, A.; Vescoukis, V. Spatio-temporal summarization of dance choreographies. *Comput. Graph.* **2018**, *73*, 88–101. [CrossRef]

7. Aristidou, A.; Cohen-Or, D.; Hodgins, J.K.; Chrysanthou, Y.; Shamir, A. Deep motifs and motion signatures. *ACM Trans. Graph. (TOG)* **2018**, *37*, 1–13. [CrossRef]
8. Voulodimos, A.; Rallis, I.; Doulamis, N. Physics-based keyframe selection for human motion summarization. *Multimed. Tools Appl.* **2020**, *79*, 3243–3259. [CrossRef]
9. Aristidou, A.; Lasenby, J. *Inverse Kinematics: A Review of Existing Techniques and Introduction of a New Fast Iterative Solver*; University of Cambridge: Cambridge, UK, 2009.
10. Voulodimos, A.; Doulamis, N.; Doulamis, A.; Protopapadakis, E. Deep learning for computer vision: A brief review. *Comput. Intell. Neurosci.* **2018**, *2018*, 7068349. [CrossRef]
11. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the Advances in Neural Information Processing Systems, Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
12. Elhamifar, E.; Sapiro, G.; Vidal, R. See all by looking at a few: Sparse modeling for finding representative objects. In Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 16–21 June 2012; pp. 1600–1607.
13. Protopapadakis, E.; Voulodimos, A.; Doulamis, A.; Doulamis, N.; Dres, D.; Bimpas, M. Stacked autoencoders for outlier detection in over-the-horizon radar signals. *Comput. Intell. Neurosci.* **2017**, *2017*, 5891417. [CrossRef]
14. Mallat, S. *A wavelet Tour of Signal Processing*; Elsevier: Amsterdam, The Netherlands, 1999.
15. Haykin, S. *Neural Networks: A Comprehensive Foundation*; Prentice-Hall: Upper Saddle River, NJ, USA, 2007.
16. Doulamis, A.D.; Doulamis, N.D.; Kollias, S.D. On-line retrainable neural networks: improving the performance of neural networks in image analysis problems. *IEEE Trans. Neural Netw.* **2000**, *11*, 137–155. [CrossRef]
17. Rallis, I.; Georgoulas, I.; Doulamis, N.; Voulodimos, A.; Terzopoulos, P. Extraction of key postures from 3D human motion data for choreography summarization. In Proceedings of the 2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games), Athens, Greece, 6–8 September 2017; pp. 94–101.
18. Doulamis, A.; Doulamis, N.; Kollias, S. Fuzzy video content representation for video summarization and content-based retrieval. *Signal Process.* **2000**, *80*, 1049–1067. [CrossRef]
19. Ngo, C.W.; Ma, Y.F.; Zhang, H.J. Video summarization and scene detection by graph modeling. *IEEE Trans. Circuits Syst. Video Technol.* **2005**, *15*, 296–304.
20. Medentzidou, P.; Kotropoulos, C. Video summarization based on shot boundary detection with penalized contrasts. In Proceedings of the 2015 9th International Symposium on Image and Signal Processing and Analysis (ISPA), Zagreb, Croatia, 7–9 September 2015; pp. 199–203.
21. Song, Y.; Vallmitjana, J.; Stent, A.; Jaimes, A. TVSum: Summarizing Web Videos Using Titles. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015.
22. Meng, J.; Wang, H.; Yuan, J.; Tan, Y.P. From Keyframes to Key Objects: Video Summarization by Representative Object Proposal Selection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016.
23. Gharbi, H.; Bahroun, S.; Zagrouba, E. Key frame extraction for video summarization using local description and repeatability graph clustering. *Signal Image Video Process.* **2019**, *13*, 507–515. [CrossRef]
24. Zhou, F.; De la Torre, F.; Hodgins, J.K. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Trans. Pattern Anal. Mach. Intell.* **2012**, *35*, 582–596. [CrossRef] [PubMed]
25. Zhou, F.; De la Torre, F. Spatio-Temporal Matching for Human Pose Estimation in Video. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1492–1504. [CrossRef]
26. Miura, T.; Kaiga, T.; Katsura, H.; Tajima, K.; Shibata, T.; Tamamoto, H. Adaptive keypose extraction from motion capture data. *J. Inf. Process.* **2014**, *22*, 67–75. [CrossRef]
27. Bakalos, N.; Rallis, I.; Doulamis, N.; Doulamis, A.; Voulodimos, A.; Vescoukis, V.C. Motion Primitives Classification using Deep Learning Models for Serious Game Platforms. *IEEE Comput. Graph. Appl.* **2020**, *40*, 26–38. [CrossRef]
28. Bernard, J.; Wilhelm, N.; Krüger, B.; May, T.; Schreck, T.; Kohlhammer, J. Motionexplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE Trans. Vis. Comput. Graph.* **2013**, *19*, 2257–2266. [CrossRef]

29. Saptoro, A.; Tadé, M.O.; Vuthaluru, H. A modified Kennard-Stone algorithm for optimal division of data for developing artificial neural network models. *Chem. Prod. Process. Model.* **2012**, *7*. [CrossRef]
30. Protopapadakis, E.; Voulodimos, A.; Doulamis, A.; Camarinopoulos, S.; Doulamis, N.; Miaoulis, G. Dance pose identification from motion capture data: a comparison of classifiers. *Technologies* **2018**, *6*, 31. [CrossRef]
31. Boyd, S.; Parikh, N.; Chu, E.; Peleato, B.; Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn* **2011**, *3*, 1–122. [CrossRef]
32. Doulamis, A.D.; Voulodimos, A.; Doulamis, N.D.; Soile, S.; Lampropoulos, A. Transforming Intangible Folkloric Performing Arts into Tangible Choreographic Digital Objects: The Terpsichore Approach. In Proceedings of the VISIGRAPP (5: VISAPP), Porto, Portugal, 1–27 February 2017; pp. 451–460.