ADAPTABLE AUTOREGRESSIVE MOVING AVERAGE FILTER TRIGGERING CONVOLUTIONAL NEURAL NETWORKS FOR CHOREOGRAPHIC MODELING

Ioannis Rallis[1], Nikolaos Bakalos[1], Nikolaos Doulamis[1], Anastasios Doulamis[1] *

[1] Photogrammetry lab, School of Rural and Survey Engineering, National Technical University of Athens

Commission II, WG II/5, II/8

KEY WORDS: Deep learning, Dynamic Scene Analysis, Intangible Cultural Heritage, Choreographic Modeling

ABSTRACT:

Choreographic modeling, that is identification of key choreographic primitives, is a significant element for Intangible Cultural Heritage (ICH) performing art modeling. Recently, deep learning architectures, such as LSTM and CNN, have been utilized for choreographic identification and modeling. However, such approaches present sensitivity to capturing errors and fail to model the dynamic characteristics of a dance, since they assume a stationarity between the input-output data. To address these limitations, in this paper, we introduce an AutoRegressive Moving Average (ARMA) filter into a conventional CNN model; this means that the classification output feeds back to the input layer, improving overall classification accuracy. In addition, an adaptive implementation algorithm is introduced, exploiting a first-order Taylor series expansion, to update network response in order to fit dance dynamic characteristics. This way, the network parameters (e.g., weights) are dynamically modified improving overall classification accuracy. Experimental results on real-life dance sequences indicate the out-performance of the proposed approach with respect to conventional deep learning mechanisms.

1. INTRODUCTION

The domain of Intangible Cultural Heritage (ICH) comprises a vast range of non-material elements, such as performing arts (e.g., folklore dances), music and oral cultural traditions (Kurin, 2004). It is clear that ICH elements are of great importance and therefore, these assets have been identified by UNESCO to ensure an efficient protection and preservation. As far as preservation of performing arts is concerned, kinesiology analysis and choreographic modeling constitute a very important aspect of folklore dance modelling. One of the most important elements of choreographic analysis is the identification of the dancer's movements and poses (i.e., dancer's postures). Recently motion capturing digitization systems are capable of providing 3D measurements of the body parts of a dancer (Rallis et al., 2018). Then, we can proceed to the identification of key primitives of a dance.

In general, deep learning models receives as inputs either raw visual signals of a choreographic sequence or transformed data, that is, 3D features, and then they generate labelled classes corresponding to dance choreographic primitives. Recently, Long Short Term Memory (LSTM) has proven especially useful in choreographic modeling (Rallis et al., 2019). The LSTM networks usually operates on 3D skeleton data of a dancer, instead of RGB content. This way the complexity of the input data is reduced, increasing choreographic classification performance. Actually, the main advantage of an LSTM network is its recurrent characteristics, implemented also in a bi-directional way (e.g., non causal modelling). Non-causality is necessary since modeling and identification of choreographic primitives depends on both backward and forward dancer's steps.

The main drawback of using 3D skeleton data sequences through an LSTM network is that the choreographic modeling performance is highly sensitive to skeleton signal errors. Missing skeleton points, as a result of errors of the motion capturing devices, significantly affect the performance of choreographic primitives classification. Another limitation is the assumption of stationarity between the input-output data. This means that the network weights of the LSTM model remains constant during choreographic modeling. However, a dance sequence presents several dynamics and dancer's attributes such as gender, age and personalized style, significantly affect the overall dance performance.

Instead, using RGB content as input to a deep learning network, we face the aforementioned skeleton error issues. Convolutional Neural Network (CNNs) have proven, recently, to be robust classifiers, especially of processing high-dimensional RGB visual data (LeCun et al., 1998), (Makantasis et al., 2017a). Therefore, CNN networks have been used for human action recognition (Varol et al., 2018), (Kamel et al., 2019).

However, issues related with the dynamic nature of a choreographic can not be addressed using conventional CNN models since model parameters (i.e., network weights) remains constant during the operation of the model. Additionally, the RGB data alone deteriorate the overall choreographic modeling performance due to the existence of enormous spatial-temporal information, confusing the classification due to the following reasons: First, the purpose of the convolutional layer of a CNN is to transform the raw RGB visual data into low-forms of representations, through the "deep convolutions". In this case, the convolutional layer transforms the whole input image frame, including the irrelevant visual background content to the choreographic modeling, into low dimensional forms of representation, which are then fed to a fully connected neural network. Second, a conventional CNN structure has not the recurrent characteristics inherently existing in a LSTM model let alone its main bi-directional capabilities. Finally, network weights are assumed to be constant throughout network operation, failing,

* Corresponding author: Nikolaos Doulamis Email: ndoulam@cs.ntua.gr

therefore, to address the dynamic characteristics of a dance.

1.1 Related Works

Kinesiology modelling are distinguished into methods that exploit supervised learning and those algorithms of using an unsupervised paradigm. In the literature, the works proposed cover human activity indexing (Ben-Arie et al., 2002), pose identification (Chéron et al., 2015), action prediction (Hadfield, Bowden, 2013), emotion recognition (Fan et al., 2016) and background subtraction (Piccardi, 2004). In (Milbich et al., 2017), an unsupervised approach is proposed for modelling human activities, while in (Rallis et al., 2018), summarization of folklore dances have been introduced using an hierarchical SMRS algorithm. In this context, the work of (Wang et al., 2011) has introduced an action recognition framework exploiting dense trajectories. Finally, in (Kolekar, Dash, 2016) hidden Markov models (HMM) has proposed for human activity recognition.

Recently deep machine learning methods have been introduced for analysis of folklore sequences. A brief review of deep learning for computer vision applications one can be found at (Voulodimos et al., 2018). In (Zeng et al., 2014), a CNN neural network model have been introduced for human activity analysis, while the work of (Khaire et al., 2018) uses RGB-D and skeleton data for activity analysis. In (Simonyan, Zisserman, 2014), the authors introduce a two-stream convolutional neural network structure for action recognition in videos. In this context, the work of (Wang et al., 2017) introduces a three-stream CNN for action recognition modelling, while the work of (Kamel et al., 2018) proposes CNNs structures on depth maps and postures for human action recognition. Finally, Makantasis el al. (Makantasis et al., 2016) introduces a behavioural understanding approach for industrial environments, while in (Gan et al., 2015), the authors introduces a flexible Deep CNN for detecting spatio-temporal relationships in videos.

Another area of research related with this paper is background modeling and consequently foreground extraction. Towards this direction salient maps have been proposed in (Makantasis et al., 2013) exploiting concepts of visual attention algorithms. In this context, the work of (Babaee et al., 2018) introduces a background modeling algorithm using CNN structures. Similarly, in (Varadarajan et al., 2015), the authors introduce methods of Mixture of Gaussians to face background dynamics. In (Bianchi et al., 2019), the authors proposed a neural network implementation of the ARMA filter with a recursive and distributed formulation, obtaining a convolutional layer that is efficient to train, localized in the node space, and can be transferred to new graphs unseen during training. In (Defferrard et al., 2016) the authors are interested in generalizing CNN from low-dimensional regular grids to high-dimensional irregular domains, such as social networks, brain connectomes or words' embedding, represented by graphs.

1.2 Paper contribution

To face the aforementioned limitations, in this paper, we introduce a novel CNN model with Autoregressive Moving Average (ARMA) capabilities. In addition, we introduce adaptive capabilities into the proposed non-linear ARMA model in a way that the network weights are dynamically adapted to face the current choreographic dynamics. We call this model adaptable ARMA-based CNN filer due to its adaptive and Autoregressive-Moving Average capabilities.

In particular, the proposed network filter feeds back its classification output to the input layer, implementing an autoregressive triggering mechanism; the output variable depends on its own previous values. In addition, we introduce a Tapped Delay Line (TDL) input to the CNN model in order to capture the temporal dependencies of a choreography. The TDL filter implements a moving average (Doulamis et al., 2003).

Finally, we introduce a computationally efficient and adaptive algorithm for dynamically modifying the network weights of the fully connected layer of the CNN model to fit the dynamic nature of a choreography. The proposed way of adaptation allows to the new ARMA-enriched CNN to automatically adapt its behavior to the current conditions while simultaneously respecting the already accumulated knowledge as much as possible. This way, the new model is able to capture the non-stationary behaviors of a choreography.

In addition, to face the first limitation of using a conventional CNN model for choreographic modeling, we prior to the classification stage. In this context, the irrelevant to the choreographic modeling background content is isolated, creating an RGB mask of dancers' postures. In this way, the hierarchies of convolutions of the CNN transforms the RGB dancers' postures into low forms of representations, e.g., kinesiology dancers' features, which are then used for choreographic modeling. Therefore, the proposed approach faces the skeleton error sensitive issues of the current LSTM filters and simultaneously addresses the previous discussed limitations of using conventional CNN models on the raw RGB data (that is dynamic training and adaptive since the output of a dance pose estimator should affect its own previous value). This paper is organized as follows: Section 1.1 describes previous works. The new proposed ARMA-enriched CNN model is discussed in Section 2. In this section, the adaptive behavior of the model is also given along with the proposed optimization process to maximize its efficiency and the variational inference-based background subtraction method. Experimental results on real-life dances are presented in Section 3. Finally, Section 4 draws the conclusions.

2. AN ARMA-ENRICHED CNN FOR CHOREOGRAPHY MODELING

Fig. 1 indicates our proposed overall architecture for choreographic modeling. As is observed, our proposed framework encompasses the following components. The first is responsible for the data acquisition (the motion capturing sensors) that is used to obtain the RGB images of a choreographic sequence as well as the skeleton data. The second component is related with the background subtraction for reducing the irrelevant to choreographic modeling content. This information is fed as input to the proposed *adaptive ARMA-enriched CNN model* (the third component). The adaptive ARMA-enriched CNN filter is a conventional CNN enriched with an ARMA Filter as well as with adaptive network weight strategies for dynamically adjust model response to fit dance dynamics. The MA component is responsible for delaying the input signals into several taps. In addition, the AR filter is responsible to feed back the classification output to the input in a way that the current choreographic modeling is related with its own previous values. Finally, the adaptive algorithm is responsible for dynamically modifying the weights of the fully connected layer of the CNN to face the dynamic nature of a choreography.

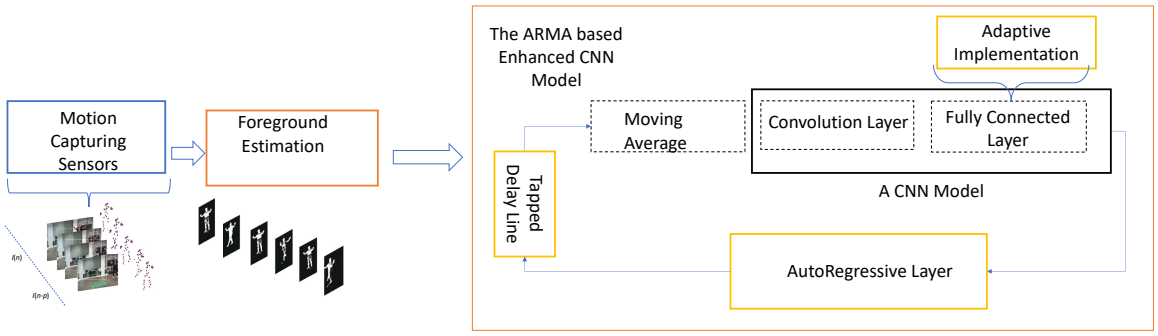

Figure 1. The overall proposed architecture adopted in this paper for choreographic modeling.

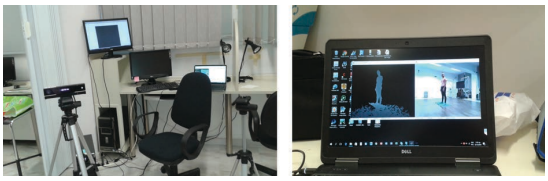

Figure 2. The architecture of the Kinect-II motion capturing interface used for digitising dance choreographic performances

2.1 The Kinect-based Acquisition Component

The acquisition module adopted for modeling the dancer's motion trajectories in 3D space exploits the Kinect-II motion capturing System. It should be mentioned that the Kinect motion capturing system also extracts the respective RGB visual data. Fig. 2 shows a snapshot of the proposed Kinect-II architecture used for motion capturing of the dance sequences.

The recorded data from the Kinect system is to extract a) the RGB visual content of the choreography and b) the respective 3D skeleton joints. In this paper, we use only the RGB information as sensorial input to identify the choreographic primitives, since skeleton sensorial data are sensitive to errors, especially in case of using low-cost motion capturing systems such as the Kinect.

2.2 The Autoregressive Moving Average Convolutional Neural Network

In the following we assume a non-linear relationship, denoted as $g(\cdot)$. This relationship relates the output of the neural network model $y(n)$ with input sensorial signals $x(n)$ at a time instance n. Actually, the purpose of $g(\cdot)$) is to transform the raw RGB input signals $x(n)$ into labeled choreographic primitives classes. Therefore, we have that

$$y(n) = g(x(n), x(n-1), \cdots, x(n-q), \\ y(n-1), \cdots, y(n-p)) + e(n) \quad (1)$$

where q expresses a time window of previous observations affecting the choreographic classification of the current image frame n, while p the order of the previous classification outputs affecting the choreographic modeling. Error $e(n)$ is an independent and identically distributed (i.i.d) process.

In order to approximate the non-linear function of $g(\cdot)$, we use machine learning methods. The machine learning algorithms minimizes the error $e(n)$ through training. In particular, it has been proven that a Tapped Delay Line (TDL) input filter can approximate the non-linear function of (1) with any degree of accuracy (Doulamis et al., 2003).

The main limitation of using a simple fully connected neural network (e.g., a feedforward one) is the training procedure are unstable especially in cases where large amount of multi-dimensional data are used as input signals, such as series of RGB image content. To face these difficulties, CNN models have been proposed as an alternative classification mechanism for processing RGB input signals compared to conventional feedforward structures (LeCun et al., 1998). A CNN model includes a pre-training layer, the convolutional layer, with the purpose of transforming the high-dimensional RGB data into low forms of representations. This means that the convolutional layer extracts from the raw visual inputs appropriate features for maximizing the overall classification performance. A CNN model have been shown very promising results in effective feature selection in a high dimensional space for choreographic modeling (Bakalos et al., 2019).

However, conventional CNN structures have not designed to approximate a non-linear ARMA filter as the one of Eq. (1). For this reason, in this paper, we extent the conventional CNN models to have ARMA characteristics

2.2.1 The Moving Average behavior: A folklore video sequence depends on several previous frames. Therefore, choreographic modeling is not relationship of only a single folklore input frame. Instead, several dance sequence frames contribute to the video modeling. For this reason, a moving average operator is adopted to model this temporal relationship.

To model a MA property into a CNN filter, we include a Tapped Delay Line (TDL) layer to the network. This is illustrated in Fig. 3. The TDL layer is responsible for delaying the input signal for q discrete time instances. Therefore, it is responsible for implementing the $x(n), x(n-1), \cdots, x(n-q)$ relationship of (1). MA behavior means that identification of a choreographic primitive at a time instance n should not limited to a single image frame, but rather to a set of q frames. That is, vector $y(n)$ depends on q previous samples $x(n-j),\ j = 0, \cdots, q-1$.

2.2.2 The AutoRegressive behavior: On the other hand, the output of the pose estimator should not only depend on external, even cumulative, input but also on its classification output history, so as to eliminate abrupt spikes in the recognition output. Therefore, including an additional time window of previous classification outputs in the input of the model can effect the consideration of previous identification behavior and ensure smoother output. This is also illustrated in Fig. 3, where the classification output feeds back to the input layer. Actually, the

AR behavior implements the second part of (1), that is the non linear function of $y(n)$ is related with its own previous values $y(n-1), \cdots, y(n-p)$.

2.2.3 The Convolutional Layer: The purpose of this layer is extract descriptors from the sensorial input signals with a latent way. In the following, the outputs of the convolutional layer of the CNN is denoted as $f_1, f_2, \cdots, f_L$. These outputs are fed as inputs to the classification layer which is resposible for choreographic modeling. The structure of the convolutions layer adopted in this paper are the following: It consists of convolutions and RELU, max pooling filters. The first layer of convolutions consists of 32 filters of a size of 5x5x3. ON the other hand, the second layer composes of 64 convolutional filters of a size of 5x5x32. The classification layer uses the descriptors of the convolutional layer, that is the $f_1, f_2, \ldots, f_L$, to provide the final choreographic modeling. Fig.3 depicts the structure of the proposed deep learning model for choreographic modeling.

Therefore, our proposed ARMA-enriched CNN architecture supports both input- and output memory to the model, thus approximating a Non-linear NARMA filter, functioned with the power of a CNN. We call this model Autoregressive Moving Average Convolutional Neural Network, named in short **ARMA-CNN model**. Fig. 3 presents the proposed ARMA-CNN) architecture adopted for choreographic modeling.

2.3 The Adaptive Behavior of the ARMA-Enriched CNN

The main limitation of the aforementioned architecture is that it is assume a stationary input-output relationship. However, this is not valid in a choreographic modeling since many dynamics are involved. Therefore, adaptable strategies are required to update the model response in a highly dynamic way.

Let us now denote as w_b the parameters of the fully connected neural layer, that is the network weights, before the network adaptation. Let us also assume that w_a is the network weights are the adaptation. We assume that these weights are related as follows

$$w_a = w_b + dw \tag{2}$$

In Eq.(2) dw refers to a small perturbation of the network weights. Eq. (2) means that we only need to compute the small perturbation of the network weights dw in order to estimate the new network weights (that is after the adaptation) from the previous ones, w_b. Usually, a choreography consists of a constant main choreographic pattern. For example, the main choreographic pattern of two different choreographies are depicted in Fig. 4. A frequency domain approach is adopted for estimating the main choreographic pattern as in (Baihua Li, Holstein, 2002). Let us denote that using the method of (Baihua Li, Holstein, 2002), the main choreographic pattern have been estimated as

$$\gamma = \{c_1(n_s), \cdots, c_L(n_e)\} \tag{3}$$

In Eq. 3 $c_i(t)$ expresses the choreographic primitive that the image frame at time instance t belongs to. This means that n_s and n_e refers to the start and end time instance of the main choreographic pattern. In case that a misclassification occurs within the a choreographic pattern group, network weight adaptation is needed. Therefore, the new network weights are estimated in a way that the network response, after the weight adaption, approximates the main choreographic pattern group sequence.

$$y_{w_a}(n) \approx c_i(n) \; \forall c_i(n) \in \gamma \tag{4}$$

In Eq. (4), $y_{w_a}(n)$ denotes the response of the network at the time instance n of using the new adapted weights w_a. Eq. (4) means that the network response should respect the main choreographic pattern sequence.

Using the assumption of Eq. (2), one can apply first-order Taylor series expansion for estimating the small weight perturbation dw. In this way, a system of linear equations are derived as follows

$$e_i(n) = A_i \cdot dw \tag{5}$$

In Eq. (5) matrix A_i expresses a matrix that it is derived from the previous network weights, that is w_b, while $e_i(n)$ is a scalar expresses the difference of the network response before and after the adaptation. Therefore,

$$e_i(n) = y_{w_a}(n) - y_{w_b}(n) \tag{6}$$

Solving Eq. (5) one can estimate the the small weight perturbation dw and thus the new weights w_a. The new ways are estimated in a way that the previous behavior of the network is optimized (see Eq. (4)).

2.4 The Optimization Procedure

The main problem of solving Eq.(5) is that we have only one equation whereas the number of weights are many. This means that dw is a multi-dimensional vector of size equal to the number of network weights of the fully connected layer of the network (see Fig.3). Therefore, there is no a unique solution of solving Eq. (5).

To address this limitation, an additional constraint is introduced in this paper. Particularly, we select among all possible solutions that satisfy Eq. (5), the one that yields a minimum modification of the small perturbations dw. This means that we have the following constraint optimisation framework

$$\begin{gathered} min \; \|dw\| \\ \text{subject to} \\ c_i(n+1) = A_i \cdot dw \end{gathered} \tag{7}$$

Solving Eq. (7), we can estimate the small perturbation of dw. An alternative framework is not to modify the weights in a way to have the minimum possible norm of dw subject to constraint of (5). Instead, the previous network knowledge should be modified as discusses in (Doulamis et al., 2003).

2.5 Variational Inference of Gaussian Modeling for Background Subtraction

As far as background modeling is concerned, a a variational inference approach of Gaussian Mixtures is adopted (Makantasis et al., 2017b). The advantages of this algorithm compared to the usage of traditional mixture of Gaussians schemes is that it substitute scalar parameters with probability distributions. Therefore, more accurate background modeling is performed. In addition, this approach is less computationally complex compared to traditional mixture of Gaussians schemes which is an important aspect for folklore analysis. Initially, every pixel is divided

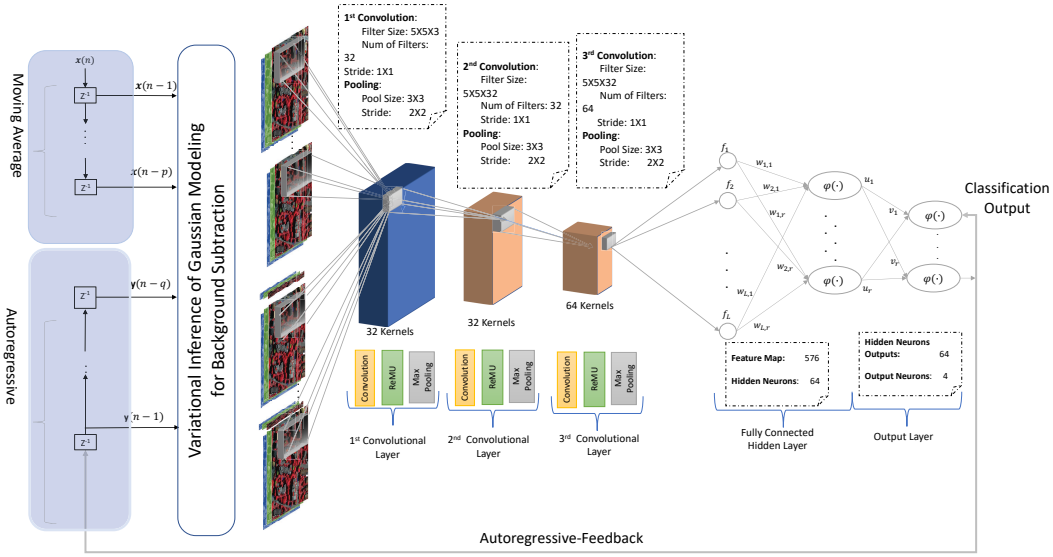

Figure 3. The architecture of the proposed Autoregressive Moving Average Convolutional Neural Network (ARMA-CNN) used for choreographic modeling in this paper

by its intensity in RGB colour space. Each pixel is computed expressing its probability whether it is included in the Foreground or Background with the following equation:

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \quad (8)$$

Actually, in a variational inference approach, variable $\omega_{i,t}$ is a probability density function, say $P(X_t|\omega)$, instead of a scalar value as in a conventional Gaussian Mixture Model. However, in Eq. (8), we have denoted as scalar for simplicity purposes (More information can be found at (Makantasis et al., 2017b)). In addition, in Eq. (8), X_t expresses the current pixel in frame t and K the number of the distributions of the mixture. The weight of the i-th distribution in frame t is expressed as $\omega_{i,t}$. Additionally, the mean of the i-th distribution in frame t is expressed as $\mu_{i,t}$ and the standard deviation of the i-th distribution in frame t is expressed as $\Sigma_{i,t}$. Moreover, the $\eta(X_t, \mu_{i,t}, \Sigma_{i,t})$ declares the probability density function and is defined as following as a Gaussian distribution.

The difference between a Gaussian mixture and a variational inference is that the weights $\omega_{i,t}$ of Eq. (8) are probability distributions instead of scalar. Therefore, better function approximations are achieved, improving background/foreground separation performance as it is discussed in (Makantasis et al., 2017b).

3. EXPERIMENTAL EVALUATION

3.1 Description of the dataset used

For evaluating and comparing the proposed algorithm against SOA methods folklore video sequences are used as presented in Table 1. A Kinect-II is exploited for the capturing process. it should be mentioned that in the presented approach the skeleton data of the Kinect-II sensor have been disregarded. The motion capturing procedure carried out at the School of Physical Education and Sport Science of the Aristotle University of Thessaloniki. All video sequences are Greek traditional folkloric dances, the selection of which was made by dance experts

Type of Dance	Description	Main Choreographic Steps
Sirtos (3-Beat)	A Greek folklore dance in a slow three-beat 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 (5-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);
Kalamatianos	A very popular Greek folk-dance through Peloponnese and the Greek Islands. The tempo is at 7/8 beat.	1) Initial Posture (IP); 2) Cross Legs (CL); 3) Cross Legs (CL); 4) Cross Legs (CL); 5) Cross Legs (CL); 6) Initial Posture (IP); 7) Cross Legs Backwards (CLB)
Trehatos	A circle dance, performed by both women and men.	1) Initial Posture (IP); 2) Cross Legs (CL); 3) Cross Legs (CL); 4) Cross Legs (CL); 5) Initial Posture (IP); 6) Left Leg Up (LLU); 7) Right Leg Up (RLU); 8) Left Leg Up (LLU); 9) Cross Legs Backwards (CLB)
Enteka	A folkloric dance performed by women and men by at a line.	1)Initial Posture (IP); 2) Right Leg Up (RLU); 3) Dancer's Right Turn (DRT); 4) Initial Posture (IP) 5) Dancer's Left Turn (DLT)

Table 1. A brief description of the dances recorded.

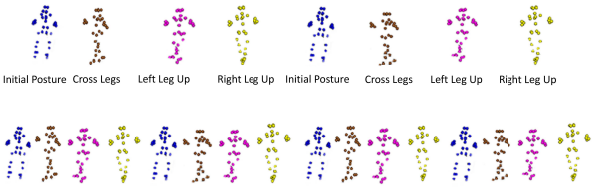

Figure 4. Choreographic primitives of two dance sequences.

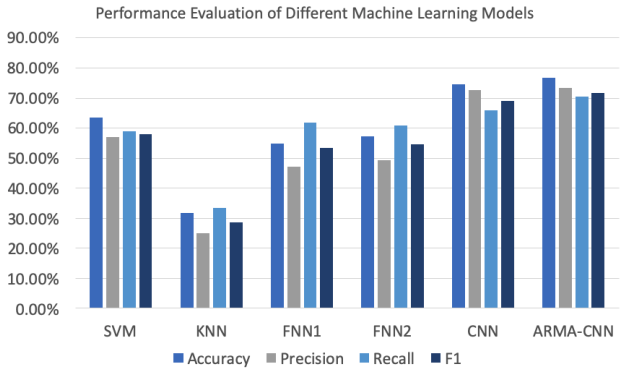

Figure 5. Performance Evaluation of different machine learning network set-ups for choreographic primitive classification

from the Aristotle University of Thessaloniki to achieve variability in terms of styling, rhythm and gender. The selection of different human sexes is due to the fact that men and women follow different style in their dance performance. Table 1 describe the folklore dance sequences used in this experiment. For every dance video sequence a small description is provided for clarification purposes. The adopted frame rate is of about 30 fps. This results in an estimate of a time window of about 15 to 30 frames, meaning of about 0.5 to 1 sec delay. In this table, we depict the main choreographic primitives of each dance. It should be mentioned that these primitives does not refer to the steps of the choreography as being taught to a dancer trainer but to the main "activities" of the dance in the digitized manner. Fig. 4 visually depicts the main choreographic primitives of two dance sequences. As is observed, the choreographic primitives same similarities with each other, imposing difficulties in the recognition process.

3.2 Choreographic Identification Performance

The proposed approach was compared with traditional adopted classifiers such as k-Nearest-Neighbor (kNN), kernel-based SVM structures, Feedforward Neural Network (FNN1) with 1 hidden layer of 10 neurons, and another FNN2 with 2 hidden layers of 10 neurons/layer. Finally, the CNN classifier was tested with a normal input layer as well as an input layer with autoregressive moving average behavior as proposed in this paper. For comparison, we include metrics from information retrieval such as precision and recall, accuracy and F1-score. During the experiments the dataset was split into a training set and a test set following an 90 to 10 ratio. Fig. 5 presents the aforementioned metrics for different machine learning configuration networks. As is observed, the proposed method, that is of using Autoregressive and Moving Average (ARMA), through an adaptive implementation, outperforms the compared machine

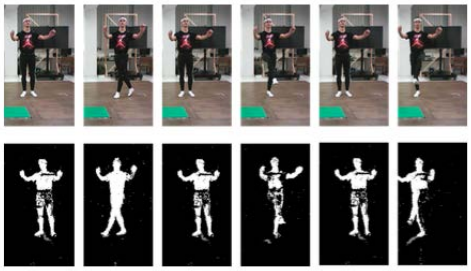

Figure 6. Simulation results regarding background/foreground estimation.

Algorithms	Method	Metrics			
		Accuracy	Precision	Recall	F1-score
SVM	No BS	46,10%	37,85%	45,14%	41,17%
	Bs	63,51%	57,05%	58,94%	57.98%
kNN	No BS	29,38%	23,46%	32,23%	27,15%
	BS	31,76%	25,07%	33,38%	28,63%
Neural 1	No BS	51,13%	43,44%	55,81%	48,85%
	BS	54,83%	47,07%	61,94%	53,48%
Neural 2	No BS	54,28%	46,50%	59,60%	52,24%
	BS	57,27%	49,43%	60,88%	54,56%
CNN	No BS	69,99%	65,15%	65,05%	65,10%
	BS	74,47%	72,65%	65,96%	69,14%
ARMA-CNN	No BS	71,44%	66,06%	67,31%	66,68%
	BS	76,82%	73,26%	70,39%	71,80%

Table 2. Performance evaluation of the proposed model for pose identification compared with other learning methods.In this table, we have provided the effect of background subtraction as a pre-processing method

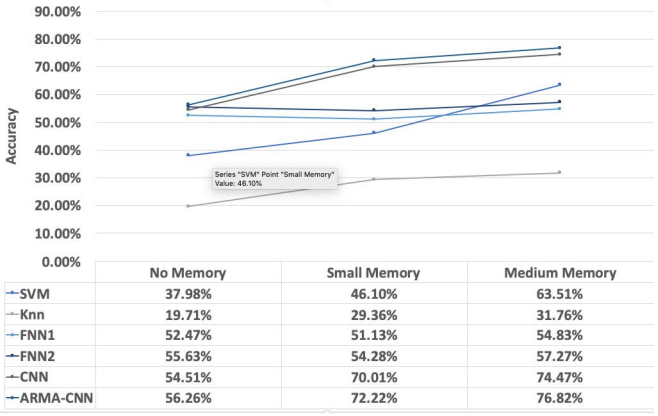

Figure 7. The effect of memory, that is the length of the tapped delay filter, on the choreographic modeling performance

learning network structures in terms of choreographic modeling. The effect of background modeling and therefore foreground separation is depicted in Table 2. It is clear that background modeling improves the overall classification performance. This is mainly due to the fact that irrelevant visual information (that is the background content) is isolated from the classification process. It should be mentioned that in Fig. 5 the results are obtained using the background separation algorithm.

The effect of the background modeling and therefore, the foreground estimation is depicted in Fig.6. Background removal is very important for choreographic modeling, since irrelevant to the choreography content is discarded. Fig. 7 indicates the effect of the size of a window (e.g., memory of window) as far as classification performance is concerned. As it is observed the implementation of the Memory Window in the classification procedure increases the total accuracy in each algorithm (SVM, kNN, FNN1, FNN2, CNN).

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-V-2-2020-467-2020 | © Authors 2020. CC BY 4.0 License.

4. CONCLUSIONS

This paper presents an adaptable autogressive and moving average layer (R-ARMA) into a conventional CNN filter to model the dynamic behavior of a choreography. The proposed architecture improves the performance of LSTM networks which is currently used for a choreography modeling, receiving as input 3D skeleton points of the dancers. The main issues of using 3D skeleton features is that the classification performance is quite sensitive to errors of the skeleton. For this reason, an alternative approach is adopted in this paper based on the capabilities of CNN models.

In particular, we use RGB input data towards choreographic modeling. RGB inputs are less sensitive to skeleton errors. However, the main drawback of this approach is that a) they can not have the recurrent characteristics of the LSTM structures, failing, therefore to handle the dynamics inherently presenting in a choreography, b) the background visual content confuses the classification accuracy since it is irrelevant to the choreography and c) they assume stationarity between the input-output data which is contradictory with the dynamic nature of a choreography. To address the aforementioned issues, we introduce, in this paper, a novel AutoRegressive, Moving Average (ARMA) filter to a CNN model in order to stimulate recurrent network characteristics. In addition, to face the choreography dynamics, we introduce an adaptation mechanisms in a way that the network weights of the fully connected hidden layer is dynamically updated to fit current environmental characteristics. Experimental results on real-life sequences illustrate the efficiency of the proposed model against conventional deep machine learning filters.

As future work, such a framework can be used in the context of educational or entertainment applications for Intangible Cultural Heritage.

ACKNOWLEDGEMENTS

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.,

REFERENCES

Babaee, M., Dinh, D. T., Rigoll, G., 2018. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, 76, 635–649.

Baihua Li, Holstein, H., 2002. Recognition of human periodic motion-a frequency domain approach. *Object recognition supported by user interaction for service robots*, 1, 311–314 vol.1.

Bakalos, N., Rallis, I., Doulamis, N., Doulamis, A., Protopapadakis, E., Voulodimos, A., 2019. Choreographic pose identification using convolutional neural networks. *2019 11th VS-Games*, IEEE, 1–7.

Ben-Arie, J., Wang, Z., Pandit, P., Rajaram, S., 2002. Human activity recognition using multidimensional indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8), 1091–1104.

Bianchi, F. M., Grattarola, D., Alippi, C., Livi, L., 2019. Graph neural networks with convolutional ARMA filters. *arXiv preprint arXiv:1901.01343*.

Chéron, G., Laptev, I., Schmid, C., 2015. P-cnn: Pose-based cnn features for action recognition. *Proceedings of the IEEE international conference on computer vision*, 3218–3226.

Defferrard, M., Bresson, X., Vandergheynst, P., 2016. Convolutional neural networks on graphs with fast localized spectral filtering. *Advances in neural information processing systems*, 3844–3852.

Doulamis, A. D., Doulamis, N. D., Kollias, S. D., 2003. An adaptable neural-network model for recursive nonlinear traffic prediction and modeling of MPEG video sources. *IEEE Transactions on Neural Networks*, 14(1), 150–166.

Fan, Y., Lu, X., Li, D., Liu, Y., 2016. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ACM, 445–450.

Gan, C., Wang, N., Yang, Y., Yeung, D.-Y., Hauptmann, A. G., 2015. Devnet: A deep event network for multimedia event detection and evidence recounting. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2568–2577.

Hadfield, S., Bowden, R., 2013. Hollywood 3d: Recognizing actions in 3d natural scenes. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3398–3405.

Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., Feng, D. D., 2018. Deep convolutional neural networks for human action recognition using depth maps and postures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*.

Kamel, A., Sheng, B., Yang, P., Li, P., Shen, R., Feng, D. D., 2019. Deep Convolutional Neural Networks for Human Action Recognition Using Depth Maps and Postures. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 49(9), 1806-1819.

Khaire, P., Kumar, P., Imran, J., 2018. Combining CNN streams of RGB-D and skeletal data for human activity recognition. *Pattern Recognition Letters*, 115, 107–116.

Kolekar, M. H., Dash, D. P., 2016. Hidden markov model based human activity recognition using shape and optical flow based features. *2016 IEEE Region 10 Conference (TENCON)*, IEEE, 393–397.

Kurin, R., 2004. Safeguarding Intangible Cultural Heritage in the 2003 UNESCO Convention: a critical appraisal. *Museum international*, 56(1-2), 66–77.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. et al., 1998. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.

Makantasis, K., Doulamis, A., Doulamis, N., 2013. Vision-based maritime surveillance system using fused visual attention maps and online adaptable tracker. *2013 14th international workshop on image analysis for multimedia interactive services (WIAMIS)*, IEEE, 1–4.

Makantasis, K., Doulamis, A., Doulamis, N., Nikitakis, A., 2017a. Tensor-Based Classifiers for Hyperspectral Data Analysis. *arXiv preprint arXiv:1709.08164*.

Makantasis, K., Doulamis, A., Doulamis, N., Psychas, K., 2016. Deep learning based human behavior recognition in industrial workflows. *2016 IEEE ICIP*, IEEE, 1609–1613.

Makantasis, K., Nikitakis, A., Doulamis, A. D., Doulamis, N. D., Papaefstathiou, I., 2017b. Data-driven background subtraction algorithm for in-camera acceleration in thermal imagery. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(9), 2090–2104.

Milbich, T., Bautista, M., Sutter, E., Ommer, B., 2017. Unsupervised video understanding by reconciliation of posture similarities. *Proceedings of the IEEE International Conference on Computer Vision*, 4394–4404.

Piccardi, M., 2004. Background subtraction techniques: a review. *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, 4, IEEE, 3099–3104.

Rallis, I., Bakalos, N., Doulamis, N., Voulodimos, A., Doulamis, A., Protopapadakis, E., 2019. Learning choreographic primitives through a bayesian optimized bi-directional lstm model. *2019 IEEE International Conference on Image Processing (ICIP)*, IEEE, 1940–1944.

Rallis, I., Doulamis, N., Doulamis, A., Voulodimos, A., Vescoukis, V., 2018. Spatio-temporal summarization of dance choreographies. *Computers & Graphics*, 73, 88–101.

Simonyan, K., Zisserman, A., 2014. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 568–576.

Varadarajan, S., Miller, P., Zhou, H., 2015. Region-based mixture of gaussians modelling for foreground detection in dynamic scenes. *Pattern Recognition*, 48(11), 3488–3503.

Varol, G., Laptev, I., Schmid, C., 2018. Long-Term Temporal Convolutions for Action Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6), 1510-1517.

Voulodimos, A., Doulamis, N., Doulamis, A., Protopapadakis, E., 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018.

Wang, H., Kläser, A., Schmid, C., Cheng-Lin, L., 2011. Action recognition by dense trajectories.

Wang, L., Ge, L., Li, R., Fang, Y., 2017. Three-stream CNNs for action recognition. *Pattern Recognition Letters*, 92, 33–40.

Zeng, M., Nguyen, L. T., Yu, B., Mengshoel, O. J., Zhu, J., Wu, P., Zhang, J., 2014. Convolutional neural networks for human activity recognition using mobile sensors. *6th International Conference on Mobile Computing, Applications and Services*, IEEE, 197–205.