# Deep Learning for Digitizing, Analysing and Modelling Choreographic 3D Signal Sequences of Intangible Cultural Heritage

**Ioannis Rallis**

School of Rural, Surveying and Geoinformatics Engineering
National Technical University of Athens

This dissertation is submitted for the degree of
*Doctor of Philosophy*

NTUA

September 2021

# Βαθιά Μηχανική Μάθηση για την Ψηφιοποίηση, Ανάλυση και Μοντελοποίηση Τρισδιάστατων Χορογραφικών Ακολουθιών της Άυλης Πολιτισμικής Κληρονομιάς

**Ιωάννης Ράλλης**

Σχολή Αγρονόμων & Τοπογράφων Μηχανικών - Μηχανικών Γεωπληροφορικής
Εθνικό Μετσόβιο Πολυτεχνείο

Αυτή η διατριβή υποβάλλεται για το πτυχίο του
*Διδάκτορα Φιλοσοφίας*

Ε.Μ.Π Σεπτέμβριος 2021

# Research Committee

***Advisory Committee***

1. Nikolaos Doulamis, Associate Professor · School of Rural Surveying Engineering · National Technical University of Athens, Supervisor

2. Andreas Georgopoulos, Professor · School of Rural Surveying Engineering · National Technical University of Athens

3. Nikolaos Grammalidis, Senior Researcher · The Centre for Research and Technology Hellas · Information Technologies Institute (ITI)

***Examination Committee***

1. Nikolaos Doulamis, Associate Professor · School of Rural and Surveying Engineering · National Technical University National Technical University of Athens, Supervisor
2. Andreas Georgopoulos, Professor · School of Rural and Surveying Engineering · National Technical University National Technical University of Athens
3. Nikolaos Grammalidis, Senior Researcher · The Centre for Research and Technology Hellas · Information Technologies Institute (ITI)
4. Athanasios Voulodimos, Assistant Professor · Department of Informatics and Computer Engineering · University of West Attica
5. Vassilios Vescoukis, Associate Professor · School of Rural and Surveying Engineering · National Technical University National Technical University of Athens
6. Charalabos Ioannides, Professor · School of Rural and Surveying Engineering · National Technical University of Athens
7. Andreas-Georgios Stafylopatis, Professor · School of Electrical and Computer Engineering · National Technical University of Athens

I would like to dedicate this thesis to Melina

# Declaration

I hereby declare that except where specific reference is made to the work of others, the contents of this dissertation are original and have not been submitted in whole or in part for consideration for any other degree or qualification in this, or any other university.

Ioannis Rallis  
September 2021

# Abstract

In performing arts, such as choreography, dance and theatrical kinesiology, movements of human body signals and gestures are essential elements used to describe a storyline in an aesthetic and symbolic way. Although, we, as humans, can inherently perceive and decipher such human body signals in a natural way, this process is challenging for a computer system. One important aspect in the analysis of a performing dance is the automatic extraction of the choreographic patterns/elements since these elements provide an abstract and compact representation of the semantic information encoded in the overall dance storyline. One salient issue in the analysis of a performing dance is to automatic extract its choreographic patterns since these elements provide an abstract representation of the semantics of the dance and encodes the overall dance storytelling. However, application of conventional video summarization algorithms on dance sequences cannot appropriately retrieve their choreographic patterns since a dance is composed of an ordered set of sequential elements which are repeated in time. Additionally, the 3D geometry of a dance is too complicated to be described using only the RGB color information.

This thesis is distinguished into three parts. Part I describes the theoretical background regarding ICH and the principles with respect to the mathematical modelling of folklore choreographic sequences. In Chapters 1, 2, 3 the recent trends on choreographic representation in terms of machine learning, video summarization, pose identification and dance annotation are described. Part II presents the adopted techniques for content-based sampling of the selected folklore choreographic sequences. This part is oriented on the semantic compression and the video summarization taking into consideration the complexity of the spatio-temporal sequences. In particular, Chapter 4 exploited a hierarchical scheme that implements spatio-temporal variations of the dance features. Chapter 5 describes an abstract representation of the semantic details of choreographic sequences taking into consideration a key-frame selection algorithm. Chapter 6 compares the summarization performances taking into account four sampling algorithms all implemented under a SAE scheme's projected data. Specifically, a SAE framework followed by a hierarchical SMRS algorithm implemented to summarize choreographic sequences. Part III (Chapters 7, 8, 9) focused on modelling and analysis of folklore choreographic sequences. Chapter 7 explored the feasibility of pattern matching between heterogeneous motion capturing systems. In this chapter, a trajectory interpretation in folklore sequences is described. The conducted experiments indicate that if significant levels of precision are ensured during initial data collection, design, development and fine-tuning of the system, then low-cost and widely popular motion capturing sensors suffice to provide a smooth and integrated experience on the user end, which would allow for relevant educational or entertainment applications to be adopted at scale. Chapter 8 focuses on the enhancement of the learning experience of folklore dances by introducing machine learning tools with the capability of providing a scalable quantifiable assessment of a choreography at different level of

hierarchies; yielding a from coarse to fine evaluation. Chapter 9 describes an adaptable autoregressive and moving average layer (R-ARMA) into a conventional CNN filter to model the dynamic behavior of a choreography. In addition, to face the choreography dynamics, we introduced 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 indicated the efficiency of the proposed model against conventional deep machine learning filters. Chapter 10 summarizes the thesis by representing the overall contribution and the future works.

# Εκτεταμένη Περίληψη

Πολλές προσπάθειες έχουν μέχρι σήμερα γίνει προκειμένου να καταγραφεί και να διασωθεί η υλική πολιτιστική κληρονομία. Αντιθέτως η άυλη κληρονομιά λόγω της μη απτής φύσης της επιφέρει δυσκολίες τόσο στην επεξεργασία όσο και στην καταγραφή της. Παρόλη την τεράστια πρόοδο που έχει επιτευχθεί στην τεχνολογία της ψηφιοποίησης, κυρίως όσον άφορα στα απτά πολιτιστικά αγαθά στο επίπεδο της τρισδιάστατης απεικόνισης (3D-COFORM, EPOCH, CARARE, IMPACT, PRESTOSPACE, V-CITY), η ηλεκτρονική τεκμηρίωση των άυλων πολιτιστικών αγαθών δεν είναι πλήρως φανερή, ειδικότερα στις λαϊκές μορφές τέχνης. Αυτό οφείλεται κυρίως στο σύμπλεγμα διεπιστημονικότητας των φολκλόρ παραστάσεων που παρουσιάζουν μια σειρά από προκλήσεις που περιλαμβάνουν τη χορογραφία, την ψηφιοποίηση, την επεξεργασία, την χορογραφική ανάλυση/σημειογραφία, την μηχανική μάθηση και την υπολογιστική όραση. Είναι σημαντικό να αναφέρουμε ότι αυτή είναι η πρώτη φορά που υλοποιείται τέτοιο καινοτόμο πεδίο έρευνας, το οποίο έχει ως στόχο να λειτουργήσει ως ένας πρωτοποριακός μηχανισμός για την ενοποίηση του περιεχομένου της Άυλης Πολιτιστικής Κληρονομιάς (ΑΠΚ) με ήδη υπάρχον ψηφιοποιημένο περιεχόμενο από ψηφιακές βιβλιοθήκες (π.χ. Europeana), την σύνδεση της ΑΠΚ με το πεδίο της μηχανικής μάθησης οδηγώντας σε προηγμένες επιστημονικές δημοσιεύσεις. Στόχος της επιδιωκόμενης ερευνητικής μελέτης της διδακτορικής διατριβής είναι η ψηφιοποίηση της ΑΠΚ, δηλαδή των χορευτών και των χορευτικών τους κινήσεων, χορευτικών εκφράσεων, καθώς και η αρχειοθέτηση των σχετικών δεδομένων/μεταδεδομένων σε κατάλληλη ψηφιακή βιβλιοθήκη, προκειμένου να διατηρηθεί τμήμα της ΑΠΚ. Επιπλέον, τίθεται η ανάγκη να μειωθεί η πολυπλοκότητα της ψηφιοποίησης που διέπει την καταγραφή, την απεικόνιση, τη μοντελοποίηση και την εικονική αναπαράσταση. Οι παραστατικές τέχνες, όπως η χορογραφία, ο χορός και η θεατρική κινησιολογία, οι κινήσεις ανθρώπινων σωμάτων και χειρονομιών είναι βασικά στοιχεία που χρησιμοποιούνται για να περιγράψουν μια ιστορία με αισθητικό και συμβολικό τρόπο. Παρόλο που εμείς, ως άνθρωποι, μπορούμε εγγενώς να αντιληφθούμε και να αποκρυπτογραφήσουμε τέτοια σήματα ανθρώπινου σώματος με φυσικό τρόπο, αυτή η διαδικασία είναι δύσκολη για ένα σύστημα υπολογιστή. Μια σημαντική πτυχή στην ανάλυση ενός παραστατικού χορού είναι η αυτόματη εξαγωγή των χορογραφικών προτύπων/στοιχείων δεδομένου ότι αυτά τα στοιχεία παρέχουν μια αφηρημένη αναπαράσταση των σημασιολογικών πληροφοριών που κωδικοποιούνται στη ακολουθία του χορού. Επιπλέον ένα σημαντικό ζήτημα στην ανάλυση εκτέλεσης ενός χορού είναι η αυτόματη εξαγωγή χορογραφικών προτύπων της, δεδομένου ότι αυτά τα στοιχεία παρέχουν μια αφηρημένη αναπαράσταση της σημασιολογίας του χορού και κωδικοποιούν τη συνολική αφήγηση του συγκεκριμένου χορού. Ωστόσο, η εφαρμογή συμβατικών αλγορίθμων συνοπτικών βίντεο ακολουθιών δεν μπορεί να ανακτήσει κατάλληλα τα χορογραφικά τους μοτίβα αφού ένας χορός αποτελείται από ένα διατεταγμένο σύνολο διαδοχικών στοιχείων τα οποία επαναλαμβάνονται

στο χρόνο. Τέλος, η τρισδιάστατη γεωμετρία του χορού είναι πολύ περίπλοκη για να περιγραφεί χρησιμοποιώντας μόνο τις πληροφορίες χρώματος RGB.

Η υπό εξέταση διατριβή αναπτύσσεται αναπτύσσεται σε 10 κεφάλαια:

Το Κεφάλαιο 1 εισάγει τις βασικές έννοιες της ΆΠΚ, αναλύει τους ερευνητικούς στόχους, την πρωτοτυπία και την καινοτομία της προτεινόμενης διατριβής.

Το Κεφάλαιο 2 παρουσιάζει τις πρόσφατες τάσεις στη χορογραφική αναπαράσταση όσον αφορά τη μοντελοποίηση, τη βιντεοπερίληψη, την σημειογραφία και την αναγνώριση χορογραφικών μοτίβων. Επιπλέον, δημιουργήθηκαν δύο χορογραφικά σύνολα δεδομένων. Τα χορογραφικά σύνολα περιλαμβάνουν τριάντα παραδοσιακές ψηφιοποιημένες ακολουθίες χορού που καταγράφηκαν σε συνεργασία με το Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης στο πλαίσιο του έργου TERPSICHORE. Αυτά τα σύνολα δεδομένων εμπεριέχουν περισσότερες από 83663 εικόνες RGB και 7362 εγγραφές νεφών σημείων (μορφή .c3d) συμβατές με διάφορες βάσεις δεδομένων (π.χ. Europeana, βάση δεδομένων CMU, AMASS).

Το Κεφάλαιο 3 περιγράφει τη διαδικασία ψηφιοποίησης χορογραφικών δεδομένων, την περιγραφή του συνόλου δεδομένων, την επεξεργασία χορογραφικών μοτίβων και την κινησιολογική μοντελοποίηση. Επιπλέον, περιγράφονται τα συστήματα καταγραφής κίνησης. Συγκεκριμένα, σε αυτό το Κεφάλαιο περιγράφονται τα χαρακτηριστικά βήματα των υπό εξέταση ελληνικών χορών που έχουν καταγραφεί προκειμένου να εξαχθούν τα χορογραφικά μοτίβα και οι αντιπροσωπευτικές στάσεις/κινήσεις.

Το Κεφάλαιο 4 πρότεινε ένα νέο σχήμα σύνοψης χορού (βιντεοπερίληψης) αναφορικά με τα δεδομένα που καταγράφηκαν χρησιμοποιώντας το σύστημα καταγραφής κίνησης VICON. Η προτεινόμενη μέθοδος εξαγωγής αντιπροσωπευτικών στιγμιοτύπων εφαρμόζει ένα ιεραρχικό σχήμα κατάτμησης που εκμεταλλεύεται τις χωροχρονικές παραλλαγές των κινησιολογικών μεταβολών των χορευτών. Αρχικά, οι ολιστικοί περιγραφείς εξάγονται για να εντοπίσουν τα βασικά βήματα ενός χορού (μια χονδροειδής αναπαράσταση). Στη συνέχεια, κάθε τμήμα αποσυντίθεται περαιτέρω σε λεπτομερή τμήματα για τη βελτίωση της αντιπροσωπευτικότητας του χορού (λεπτή αναπαράσταση). Το σχήμα ιεραρχικής κατάτμησης χορού τροποποιεί τον αλγόριθμο SMRS κατάλληλα προκειμένου να επιτραπεί η χωροχρονική μοντελοποίηση σύνθετων χορευτικών ακολουθιών. Η προσέγγισή αξιολογήθηκε σε τριάντα φολκλορικές χορευτικές ακολουθίες που καταγράφηκαν στο Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης στο πλαίσιο του έργου TERPSICHORE που αντιπροσωπεύει πέντε διαφορετικές χορογραφίες και σε σύνολα δεδομένων από το Πανεπιστήμιο Carnegie Mellon, ελεύθερα διαθέσιμα, που απεικονίζουν παραστάσεις στη θεατρική κινησιολογία.

Το Κεφάλαιο 5 εισάγει δύο τεχνικές: μια μέθοδο «ανεξάρτητη από το χρόνο» που βασίζεται σε αλγόριθμο ομαδοποίησης k-means++ για την εξαγωγή αντιπροσωπευτικών στιγμιοτύπων του χορού και μια τεχνική που βασίζεται στη ερμηνεία των φυσικών χαρακτηριστικών της κινησιολογίας δημιουργώντας χρονικές περιλήψεις σε διαφορετικά επίπεδα λεπτομέρειας. Οι προτεινόμενες μέθοδοι αξιολογήθηκαν σε δύο σύνολα δεδομένων κίνησης χορού.

Το Κεφάλαιο 6 εισάγει ένα μη επιβλεπόμενο πλαίσιο βαθιάς στοίβας αυτόματου κωδικοποιητή (SAE) ακολουθούμενο από έναν αλγόριθμο ιεραρχικής κατάτμησης για να συνοψίσει τις χορογραφικές ακολουθίες. Στόχος του SAE είναι ο περιορισμός των περιττών θορύβων στα μη επεξεργασμένα δεδομένα και συνεπώς η βελτίωση της απόδοσης της χορογραφικής περίληψης. Αυτό γίνεται εμφανές όταν δύο χορευτές καταγράφονται ταυτόχρονα. Αλγόριθμοι βίντεο-περίληψης εφαρμόζονται για την εξαγωγή των αντιπροσωπευτικών χορογραφικών στάσεων χρησιμοποιώντας τον Kennard-Stone, το συμβατικό SMRS και το ιεραρχικό του σχήμα που ονομάζεται H-SMRS. Τα πειραματικά αποτελέσματα αξιολογήθηκαν σε πραγματικές χορευτικές ακολουθίες ελληνικών παραδοσιακών χορών, ενώ τα αποτελέσματα συγκρίθηκαν με χορογραφικά δεδομένα που επέλεξαν ειδικοί χορού. Τα αποτελέσματα δείχνουν ότι το H-SMRS που εφαρμόζεται μετά τη μείωσης του θορύβου υπό την εφαρμογή του SAE εξάγει βασικά καρέ που αποκλίνουν σε χρόνο μικρότερο από 0,3 δευτερόλεπτα από αυτά που επιλέγονται από τους ειδικούς και με τυπική απόκλιση 0,18 δευτ. Έτσι, το προτεινόμενο σχήμα μπορεί να εξάγει αντιπροσωπευτικά στιγμιότυπα αποτελεσματικά.

Το Κεφάλαιο 7 παρουσιάζει ένα νέο πλαίσιο βασισμένο στη μοντελοποίηση της κινησιολογίας με βάση τα φυσικά χαρακτηριστικά (ταχύτητα, επιτάχυνση) για την εξαγωγή αντιπροσωπευτικών στιγμιοτύπων χρησιμοποιώντας νέφη σημείων. Προτείνονται δύο προσεγγίσεις: (i) μια μέθοδος βασισμένη σε αυτόματη ομαδοποίηση για την επιλογή των βασικών πρωτευόντων μιας χορογραφίας και (ii) μια προσέγγιση βασισμένη στα κινηματικά χαρακτηριστικά των χορογραφιών. Το πλαίσιο περίληψης χορού έχει επικυρωθεί επιτυχώς σε σύνολα χορογραφικών δεδομένων με τη συμμετοχή επαγγελματιών χορού και εμπειρογνωμόνων.

Το Κεφάλαιο 8 περιγράφει ένα δίκτυο Μακράς και Βραχείας μνήμης (LSTM) με ικανότητα ανάλυσης χορογραφικών στάσεων λαμβάνοντας υπόψιν νέφη σημείων προερχόμενα από την ψηφιοποίηση των χορευτών. Αυτή η διαδικασία ταυτοποίησης θέσης είναι ικανή να παρέχει μια λεπτομερή αξιολόγηση του χορογραφικού μοτίβου. Επιπλέον, προτείνεται μια αρχιτεκτονική χορογραφικής περίληψης που βασίζεται στην εφαρμογή της ιεραρχικής κατάτμησης προκειμένου να εξάγει τα χορογραφικών μοτίβα. Τέλος αναπτύχθηκε μια πλατφόρμα σοβαρού παιχνιδιού υποστηρίζοντας την οπτικοποίηση της χορογραφίας χρησιμοποιώντας Laban σημειογραφία, προκειμένου προσδιορίσει την απόδοση της προτεινόμενης προσέγγισης με επίσημη τεκμηρίωση.

Το Κεφάλαιο 9 περιγράφει ένα αυτοπαλίνδρομο κινητού μέσου όρου (ARMA) φίλτρο που εφαρμόζεται σε ένα συμβατικό Συνελικτικό Νευρωνικό Δίκτυο (CNN). Αυτό σημαίνει ότι η έξοδος ταξινόμησης επιστρέφει στο επίπεδο εισόδου, βελτιώνοντας τη συνολική ακρίβεια ταξινόμησης. Επιπλέον, εισάγεται ένας προσαρμοστικός αλγόριθμος, εκμεταλλευόμενος την επέκταση της σειράς Taylor πρώτης τάξης. Με αυτόν τον τρόπο, οι παράμετροι του δικτύου (π.χ. βάρη) τροποποιούνται δυναμικά βελτιώνοντας τη συνολική ακρίβεια ταξινόμησης. Τα πειραματικά αποτελέσματα σε πραγματικές χορευτικές ακολουθίες δείχνουν την απόδοση της προτεινόμενης προσέγγισης σε σχέση με τους συμβατικούς μηχανισμούς βαθιάς μάθησης.

Το Κεφάλαιο 10 ολοκληρώνει τη διατριβή παραθέτοντας τη συνολική συμβολή της και τα μελλοντικά βήματα.

# Acknowledgements

I would like to express my gratitude to the people who have supported me during the elaboration of this PhD thesis. First and foremost to three advisors ; Nikolaos Doulamis, for his outstanding ethos and for believing in me and encouraging me since my early steps on research on 2017 with his relentless optimism, inspiration, friendship and collaboration. Andreas Georgopoulos except of being an important pillar of this PhD and my research career so far, shared with me his principles and ethics in research, shaping my research and personal identity; Nikolaos Grammalidis for his outstanding contribution to the digitization of the Intangible Cultural Heritage, for the valuable discussions on the problems and the solutions found during this dissertation, for introducing me to choreographic datasets and motion capturing systems; Anastasios Doulamis for introducing me to the fundamental principles of Machine Learning and Intangible Cultural Heritage domains during our collaboration in the context of the H2020 projects. Anastasios gave me endless research freedom and believe within the setting of those projects, offer assistance, comments and proposals and was effectively included in most of the parts of this dissertation. At the same time, I would like to thank Athanasios Voulodimos for the moral, psychological, professional support, he generously showed me throughout this path. I will never forget that Athanasios Voulodimos was the man who urged me to pursue doctoral studies. Big Thanks to Eftychios Protopapadakis, Nikolaos Bakalos, Maria Kaselimi, George Kopsiaftis for the fruitful collaboration and support. The excellent participation between all the individuals of the counseling committee served as a steady inspiration for advancement, setting a exceptional establishment for my professional career. I would like to thank the rest of the individuals of the Photogrammetry Lab. Moreover, I would like to thank Ioannis Papadonikolakis for his valuable support, his optimism as well as his valuable advice on multiple levels. Last but not least, I would like to thank my family for supporting me all these years.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations and Nomenclature

## Abbreviations

| | |
|---|---|
| ARMA-CNN | Autoregressive Moving Average Convolutional Neural Network |
| BOBi-LSTM | Bayesian Optimized Bi-directional Long Short Term Memory |
| CH | Cultural Heritage |
| CNN | Convolutional Neural Network |
| DBI | Davies-Bouldin index |
| DTW | Dynamic Time Warping |
| EU | European Union |
| FIS | Fuzzy Inference System |
| GAN | Generative Adversarial Networks |
| HMM | Hidden Markov models |
| ICH | Intangible Cultural Heritage |
| LMA | Laban Movement Analysis |
| LPF | LowPass Filter |
| LSTM | Long-Short Term Memory |
| MA | Moving Average |
| ML | Machine Learning |
| NARMA | Non-linear Auto Regressive Moving Average |
| PCA | Principal Component Analysis |
| RELU | Rectified Linear Unit |
| SAE | Stacked Auto-Encoder |

| | |
|---|---|
| SVM | Support Vector Machine |
| TDL | Tapped Delay Line |

# Part I

# Introduction and Fundamentals

# Chapter 1

# Intangible Cultural Heritage

## 1.1 Introduction

The ICH content encompasses *"the practices, representations, expressions, knowledge, skills – as well as the instruments, objects, artefacts and cultural spaces associated therewith"* [11]. Although, the ICH content and especially the traditional folklore performing arts, are considered to be worthy of preservation by UNESCO (Convention for the Safeguarding of ICH) and the EU committee, most of the current research efforts are focused on tangible cultural assets [12], while the ICH content seems to be underestimated. The primary disadvantage emerges from the complex structure of ICH, its dynamic composition, the interaction among the objects and the environment, as well as from a variety of emotional elements (e.g., the way of expression and dancers' style) [13], [14]. This thesis focuses on analysing, designing, researching, training and validating a novel framework, that implements machine learning algorithms, for digitization, modelling, archiving and e-preserving ICH content related to folk dances.

European's CH is considered to be one of the greatest diversities around the world. The fusion of these multiple cultural diversities leads to a common place, that draws millions of visitors every year to cultural heritage sites, as well as to theaters, concert halls, folklore festivals and to other festivities. This common cultural European underlay merges the accumulation of past artistic achievements with the dynamical expressions of tradition and creativity, during the 4-th decade of the industrial revolution. The cultural outcomes are considered as economic triggers that boost activities and job opportunities, reinforcing the social and political cohesion of the EU [15]. Culture is playing an emblematic role into supporting the European integration process, attempting to bring people regardless of their different habits, traditions and languages. Towards this aim, the prosperity and the adequacy of the EU is in its ability to pay respects to each Member States' identity and inter-related history and cultures, while forging mutual understanding and policies that have ensured peace, stability, prosperity and solidarity [16] for decades. The significance of the European Cultural Heritage is prominent via very important international decisions, declarations and agreements, such as (a) the adoption of the Commission Communications to the Council, (b) the EU Lisbon Treaty (article 3) [17], (c) the European Parliament 2006/2040(INI) [18], (d) the different resolutions of the European Council (such as the 2006/C297/01), (e) the Communication by the European Commission in 2007 as the famous European agenda for culture, that was later also endorsed by the Council of Ministers in November 2007. This agenda is structured around three main pillars, in order to foster (i) cultural diversity and intercultural dialogue, (ii) culture, as a catalyst for promoting creativity within the framework of the Lisbon Strategy for EU growth and (iii) to sharpen new competencies concerning culture as a vital element in the EU foreign affairs.

Furthermore, the European Council Resolution stated in 2008 and in 2010 (8843/10), that *"Digitization and online accessibility of cultural material are essential to highlight cultural and scientific heritage, to inspire the creation of new content and to encourage new online services to emerge. They assist to help democratize access to culture and knowledge and to develop the information society and the knowledge-based economy"*. Finally, the Commission Recommendations to the Member States (2011/711/EU[1]) and the report of the "Comite de Sages" on Bringing Europe's Cultural Heritage Online are the most important EU policy documents about the e-documentation and e-preservation of the Cultural Heritage Commons [19].

On the other hand, the UNESCO Convention for the Safeguarding of Intangible Cultural Heritage, defines that the tangible, along with the intangible cultural assets, determine intellectual, materialistic and emotional features, that defines a society or a social group. Hence, ICH is considered to be an important factor in maintaining cultural diversity, in the content of globalization. Its emblematic role is not the cultural manifestation itself, but rather the wealth of knowledge and skills that are imparted to the next generations [20].

Cultural expression, in any form, includes fragile intangible live expressions and elements. Such expressions are built upon certain knowledge, skills and craftsmanship. These manifestations of human intelligence and creativeness constitute our ICH, a basic factor of local cultural identity and a guaranty for sustainable development [21],[22]. UNESCO refers that ICH assets (e.g., music, dance, craft) are of equal importance to the tangible ones. Folk dances are important parts to ICH; they are directly connected to local culture and identity [23], [24]. Recently, research approaches have been carried out for digitization, modelling [25], choreographic analysis [26], posture classification [27], documentation [9], [10] and representation of folklore choreographies [28]. In this context, research projects have been funded, such as i-TREASURES [11], TERPSICHORE [2] [29], Wholodance [12], WebDANCE [13], AniAge and projects with the purpose of capturing and modelling ICH. Beyond the political support, the significant EU investment in the area of cultural heritage, which is more than 1B Euro in the last decade, has a number of past and currently important active projects[3] of this scope.

The study of dance from a computational point of view has been enabled by the development of heterogeneous sensors, including visual cameras and motion capture devices, on the one hand, and the advancements in motion analysis fueled by the progress made in machine learning, as well as signal and image processing [16]. Regarding the part of motion acquisition, characteristic examples of Motion Capture Systems are Kinect [17], Vicon, and OptiTrak [18], which can be seen as one of the most accurate motion schemes used to digitize humans' movements [19], [20], [16]. Now, these systems are being rapidly incorporated as a critical component to many applications like gaming, 3D animation, education,

[1] https://ec.europa.eu/digital-single-market/en/news/european-commission-report-cultural-heritage-digitisation-online-accessibility-and-digital

[2] http://terpsichore-project.eu/

[3] e.g. AGAMEMNON, CALIMERA, DELOS, MEMORIES, MICHAEL, MICHAEL+, MINERVA, MINERVA+, 3D-COFORM, PRESTOSPACE, IMPACT, V-CITY, EPOCH, CULTURA, V-NET, DC-NET, INDICATE, ATHINA, ATHINA+, DC-NET, INDICATE, EUROPEANALOCAL, APARSEN, AXES, CHESS, PATHS, PrestoPRIME, 4DCH-WORLD, ITN-DCH

engineering, rehabilitation and sports industry [9], [16]. In addition, the visualization of the human body through joint identification and extraction of the dance movement based on motion capture and Labanotation [30], [31], [32], expand new horizons in several fields such as kinesiology, neuroscience and computer graphics research.

A choreographic sequence is a time-varying 3D process (4D modeling), which contains dynamic co-interactions among different actors, emotional and style attributes, and supplementary elements, such as music tempo, and costumes. Dance analysis is an important research field in the cultural sector since it constitutes one of the components of ICH. Nowadays, research focuses on the utilization of motion acquisition sensors, in an attempt to handle kinesiology issues. The extraction of skeleton data, in real-time, contains a significant amount of information (data and metadata), allowing for various choreography-based analytics. Analyzing choreographic sequences is a highly complicated task as it involves the inclusion and processing of many factors such as the dancer's emotions [14], motion capturing systems calibration issues, the dancer's expressions [15] and kinesiology differences. Moreover, folklore choreographies are very important not only for preserving ethnological aspects but is a different area in the kinesiology field encompassing the rhythm, the expression, specific postures and the folklore music.

## 1.2 Main Research Objectives of the Thesis

ML learning techniques have progressed dramatically over the past decades, from researched curiosity to a practical technology, in many applications such as Computer Vision, Natural Language Processing, Bioinformatics, etc., succeeding to provide solutions to difficult research problems, while also leading to a wide range of exciting applications. In the domain of ICH content, and particularly dance, ML provides many opportunities for analysis, classification, semantic annotation and emotional understanding of human choreographic movement. In this thesis, we will present a brief survey of the main approaches that have been proposed in the literature exploiting ML techniques, to analyze choreographic time series (see Chapter 2). We focused on three main pillars: (a) *the extraction of the key choreographic postures, taking into consideration time series analysis*, i.e. video summarization, (b) *the identification of key posture in dance movement*, i.e. dance pose recognition of choreographic content, and (c) *the semantic representation and notation of dance movements through Laban Movement Analysis*. The way of developing the research plan and the methodology that will be followed for the investigation of the above mentioned topic, is a complex process and is presented briefly below.

*Objective 1*: Review and evaluation of the state-of-the-art digitization technologies for decreasing the capturing complexity, by introducing low cost devices, able to acquire high quality depth information in real-time and compare such smart sensors with imaging technologies (enhanced through the use of computer vision tools, data processing and 3D modeling) in order to result in accurate virtual reconstructions of moving, complex (non-rigid) objects which dynamically interact with each other and with the environment.

*Objective 2*: Review, analysis and evaluation of digitization technologies to identify the one that will provide significantly reduced complexity in recording using motion capturing systems, so that they will be able to obtain high quality information in real time.

*Objective 3*: Comparison among motion capturing sensors utilizing computer vision tools, machine learning algorithms, time-series analysis and 3D modeling [14], in order to define the most accurate semantic representations of movements and complex choreographic performances.

*Objective 4*: Research on data processing, mathematical modelling machine learning and computer vision techniques, which serve in the automation of 3D/4D digitization. These are tools that allow the connection of multimodal data, to a common work environment, reducing the complexity of digitization and enhancing the ability to automate it.

*Objective 5*: Investigating the ways of extracting semantic and ontological signatures, with the assistance of 3D computer vision and computational methodologies for modeling human movements and for measuring human expression. At the same time, advanced methods in the field of video summarization will be explored, which will be exploited with ML techniques of the recorded choreographic data.

## 1.3 Originality and contributions of this Thesis

The recent advances in digitization technology as regards tangible cultural assets and especially in the area of 3D virtual reconstruction and rehabilitation, the e-documentation of ICH assets is not yet evident, especially of folklore performing arts. This is mainly due to the complex multi-disciplinarity of the folklore performances which presents a series of challenges ranging from the choreography, the folk music, the –uniforms, -music and from the digitization and computer vision to spatio-temporal (4D) dynamic modeling and virtual scene generation as discussed above. 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 the objectives of this thesis, the main contributions to the research community are summarized follows:

- A key frame extraction framework that implements a hierarchical scheme exploiting spatio-temporal variations of the dance features is introduced (see Section 4). Initially global holistic descriptors are extracted to localize the key choreographic steps of a dance (a coarse representation). Then, each segment is further decomposed into finer sub-segments to improve dance representativity (fine representation). Dance abstraction scheme exploits the concepts of a Sparse Modeling Representative Selection (SMRS) appropriately modified to enable spatio-temporal modelling of the dance sequences through a hierarchical decomposition algorithm.

- A machine learning method exploiting deep learning paradigms is proposed (see Section 8). In particular, we introduced a LSTM memory network with the main capability of analyzing 3D captured skeleton feature joints of a dancer into predefined choreographic postures. This pose identification procedure is capable of providing a detailed (fine) evaluation score of a performing dance. In addition, this proposed framework proposes a choreographic summarization architecture based on SMRS in order to abstractly represent the performing choreography through a set of key choreographic primitives. We have modified the SMRS algorithm in a way to extract hierarchies of key representatives. Choreographic summarization provides an efficient tool for a coarse quantitative evaluation of a dance. Moreover, hierarchical representation scheme allows for a scalable assessment of a choreography. The serious game platform supports advanced visualization toolkits using Labanotation in order to deliver the performing sequence in a formal documentation.

- Development of a method to address dynamic limitations of sequences (e.g., stationarity). We introduced 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.

- Development of a deep stacked auto-encoder (SAE) scheme followed by an algorithm 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 (see Section 9).

- A method that matches trajectories' patterns, existing in a choreographic database, to new ones originating from different sensor types such as VICON and Kinect II. Then, a Dynamic Time Warping (DTW) algorithm proposed to find out similarities/dissimilarities among the choreographic trajectories. The goal of this method is to evaluate the performance of the low-cost Kinect II sensor for dance choreography compared to the accurate but of high-cost VICON-based choreographies. Experimental results on real-life dances are carried out to show the effectiveness of the proposed DTW methodology and the ability of Kinect II to localize dances in 3D space (see Section 7).

- Development of two choreographic datasets (see Sections 3.3.1, 3.3.2, 3.5). Our approach encompasses thirty folkloric dance sequences recorded at the Aristotle University of Thessaloniki under the framework of TERPSICHORE project representing five different choreographies. These datasets encompass more than 83663 RGB images and more than 7362 point clouds records (.c3d format) compatible with various databases (e.g., Europeana, CMU database, AMASS [4]).

## 1.4 Outline of the Thesis

The structure of this thesis is divided into ten chapters. The second chapter presents an overview of the previous works relevant to the choreographic domain describing the state-of-the-art approaches. The

[4] https://amass.is.tue.mpg.de/en

third part encompasses the choreographic datasets created within this research taking into consideration the benchmarked motion databases. The fourth part analyses the fundamentals regarding kinessiological modelling and the pre-processing state of the training data. The fifth part describes our approach to analyse choreographic patterns from heterogeneous motion capture systems using DTW algorithm. The sixth part exploits an hierarchical summmarization schema to decompose the choreographic sequences taking into consideration the spatio-temporal dependencies.

(i) **Chapter 2** presents recent trends in choreographic representation in terms of modelling, summarization and choreographic pose recognition. We survey recent approaches employed for the extraction of representative primitives of choreographic sequences, the recognition of choreographic pose and dance movements, as well as for the analysis and semantic representation of choreographic patterns.

(ii) **Chapter 3** represents the dataset creation and the components integration and includes the state-of-the-art solutions, the adopted acquisition process, the TERPSICHORE dataset description, the data processing and the kinessiological modelling. Moreover, it is described the adopted motion capturing systems and the adopted acquisition process. In addition, it is introduced the adopted folklore dances and the description of the adopted choreographies. Specifically, in this Chapter the adopted folklore Greek dances are annotated in order to extract the choreography patterns and the most representative key postures.

(iii) **Chapter 4** proposed a new dance summarization scheme on data being recorded using the Vicon motion capturing system. This way, skeleton information of the 3D joints of a dancer is available. The proposed key frame extraction method implements a hierarchical scheme that exploits spatio-temporal variations of the dance features. Initially global holistic descriptors are extracted to localize the key choreographic steps of a dance (a coarse representation). Then, each segment is further decomposed into finer sub-segments to improve dance representativity (fine representation). Dance abstraction scheme exploits the concepts of a Sparse Modeling Representative Selection (SMRS) appropriately modified to enable spatio-temporal modelling of the dance sequences through a hierarchical decomposition algorithm. Our approach is evaluated over thirty folkloric dance sequences recorded at the Aristotle University of Thessaloniki under the framework of TERPSICHORE project representing five different choreographies and on datasets from the Carnegie Mellon University, freely available, that depict performances on theatrical kinesiology.

(iv) **Chapter 5** provides an abstract and compact representation of the semantic information of choreographic sequences using a key-frame selection algorithm. In this chapter two techniques are introduced: a "time-independent" method based on k-means++ clustering algorithm for the extraction of prominent representative instances of a dance, and a physics-based technique that creates temporal summaries of the sequence at different levels of detail. The proposed methods are evaluated on two dance motion datasets.

(v) **Chapter 6** introduces a deep stacked auto-encoder (SAE) scheme followed by a hierarchical Sparse Modeling for Representative Selection (SMRS) algorithm in order 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.

(vi) **Chapter 7** presents a novel framework based on physical modeling for the extraction of salient 3D human motion data from real-world choreographic sequences. Two approaches are proposed:(i) a clustering-based method for the selection of the basic primitives of a choreography, and (ii) a kinematics-based method that generates meaningful summaries at hierarchical levels of granularity. The dance summarization framework has been successfully validated and evaluated with two real-world datasets and with the participation of dance professionals and domain experts.

(vii) **Chapter 8** describes a Long-Short Term Memory (LSTM) network with the main capability of analyzing 3D captured skeleton feature joint of a dancer into predefined choreographic postures. This pose identification procedure is capable of providing a detailed (fine) evaluation score of a performing dance. In addition, the paper proposes a choreographic summarization architecture based on Sparse Modelling Representative Selection (SMRS) in order to abstractly represent the performing choreography through a set of key choreographic primitives. We have modified the SMRS algorithm in a way to extract some hierarchies of key representatives. Choreographic summarization provides a efficient tool for a coarse quantitative evaluation of a dance. Moreover, hierarchical representation scheme allows for a scalable assessment of a choreography. The serious game platform supports advanced visualization toolkits using Labanotation in order to deliver the performing sequence in a formal documentation

(viii) **Chapter 9** describes an Auto-Regressive 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.

(ix) **Chapter 10** concludes the thesis by representing the overall contribution and the future works.

# Chapter 2

# Related Works

## 2.1 Introduction

Performing arts and in particular dance is one of the most important domains of Intangible Cultural Heritage [14]. However, preserving, documenting, analyzing and visually understanding choreographic patterns is a challenging task due to technical difficulties it involves. A choreography is a time-varying 3D process (4D) including dynamic co-interactions among different actors (dancers), emotional and style attributes, as well as supplementary ICH elements such as the music tempo, the rhythm, traditional costumes etc. Recent technological advancements have unleashed tremendous possibilities in capturing, documenting and storing Intangible CH content, which can now be generated at a greater volume and quality than ever before. The massive amounts of RGB-D and 3D skeleton data produced by video and motion capture devices. The huge number of different types of existing dances and variations dictate the need for organizing, archiving and analyzing dance-related cultural content in a tractable fashion and with lower computational and storage resource requirements. Motion capturing devices extract humans' skeleton data in terms of 3D points each corresponding to a human joint. This information can be combined with computer graphics software toolkits for modelling, classification and summarization purposes. In this chapter, we present recent trends in choreographic representation in terms of modelling, summarization and choreographic pose recognition. We survey recent approaches employed for the extraction of representative primitives of choreographic sequences, the recognition of choreographic pose and dance movements, as well as for the analysis and semantic representation of choreographic patterns [14].

Works focusing on choreographic acquisition and modelling can be distinguished into those that deal with 3D digitization and capturing and those that mainly focus on the analysis and processing of dances. Regarding 3D digitization, the work of [33] is considered as one of the first approaches in the field. In particular, this work introduces a 3D archive system for Japanese traditional performing arts. The graph-cuts algorithm is used to reconstruct the 3D model of the scene from multi-view videos. In the same context, the [34] digitizes Cypriot dances using the Phasespace Impulse X2 motion capture system. In the same work, a video game is developed for making the teaching of Cypriot dances more attractive. In [26], the capturing architecture of the i-Treasure European Union funded project is described, mainly focusing on 3D digitization and analysis of rare European folkloric choreographies. A digitization framework suitable for tele-immersive applications of a dance is proposed in [35]. The purpose of this research is to design a creativity framework for dance choreography based on LMA [30]. Advanced motion captured architectures for digitizing folklore performing arts presented in [36]. In this work, motion analysis algorithms are investigated with the main aim to transform the captured motion trajectories of the dancers into meaningful and semantically enriched LMA features.

Although 3D digitization technologies provide an efficient framework for documentation and preservation of the ICH artifacts of folklore dances, it has the limitation that the delivered 3D data are too large for processing, storing and archiving. For this reason, skeletonization is first performed, which is a process that emphasizes the geometrical and topological properties of the motion trajectories, extracting the medial axis. In this context, Kinect depth senors [37], Phasespace capturing [36] or Vicon [38] motion interface has been exploited.

Regarding choreographic analysis approaches, classification algorithms have been proposed on data expressing the human body movements. In this context, the work of [39] proposes a real-time classification system in detecting choreographed gesture classes. The input data have been acquired using the Kinect depth sensor [40], extracting a 3D wireframe skeleton of the dancers. Another dance classification approach is proposed in [41] using again data capturing from the Kinect sensor. In particular, the authors of [41] combine a PCA , acting as a feature selection process, with two classifiers; a Gaussian mixture and a hidden Markov model. A combination of principal component and Fisher's linear discriminant analysis, which is called fisherdance, is proposed in [42], for classifying Korean pop dances. The inputs are again from the Kinect sensor.

A dance recognition system is introduced in [43]. The platform compares an unknown move with a specified start and stop against known dance moves. The recognition method consists of a classification algorithm and a template matching using a database of model moves. Similarly, in the works of [44], [37] a markerless tracking system, exploiting the principles of the Kinect sensor, is presented for motion trajectory interpretation and folklore dance pattern recognition.

Recently, video summarization algorithms have been proposed for choreographic motion trajectories [8]. This scheme exploits input data from a Vicon motion capturing interface and then applies a k-means classification algorithm to find out key frame representatives that abstractly model the choreography. In the broad research area of dance summarization, algorithms focusing on extracting key frames of human actions can be also considered. More specifically, the works of [45] and [46] introduce a classification framework for retrieving representative human actions, while the work of [47] proposes a hierarchical union of sub-spaces for human activity abstraction under a semi-supervised framework. In addition, the work of [48] proposes Histograms of Grassmannian Points for classifying multidimensional time-evolving data in dynamic scenes. A stylistic analysis of the variations of dance movements has been recently proposed in [49]. In addition, in the works of [50] and [51] emotional analysis and characterization of dance sequences are discussed.

## 2.2 Previous Work

### 2.2.1 Choreographic Summarization

Content summarization is very useful application domain in the multimedia research community in general. Focusing on choreographic sequences, the automatic extraction of the choreographic elements is of significant interest, since such elements provide an abstract and compact representation of the semantic information encoded in the overall dance storyline. A large number of sensors capture the kinesiology of the dancers around the clock producing huge video sequences. Processing these videos is a time, energy, hardware and man power consuming progress. Due to the aforementioned parameters video summarization has an important role in this field enhancing the storage, browsing and retrieval of large collection of video data without losing important details of the captured subject. One of the first approaches for extracting the most representative key frames from video programs introduced in [52]. After that, many approaches used kinesiological features for extraction the most representative frames. The approach in [53] focuses on the decomposition of the dance movements into elementary motions. Placing this problem into a probabilistic framework, we propose to exploit Gaussian processes to accurately model the different components of the decomposition [54]. The proposed framework relies on Gaussian processes allowing for a flexible representation, from extremely coarse to detailed, capturing the periodicities of the dance movement.

In [1], the authors focus on segmentation and classification algorithms using depth images and videos of folkloric dances in order to identify key movements and gestures, compare them against database instances and determine the dance genres they represent, as well as to provide helpful metadata. A set of six traditional Greek dances consists the investigated data. A two-step process was adopted. At first, the most descriptive skeleton data were selected using a combination of density based and sparse modelling algorithms. Then, the representative data served as training set for a variety of classifiers. In [55], a segmentation method that can separate cyclic activities and their transitions for a number of data modalities is presented. This approach tackles the segmentation problem on a general level in terms of the choice of crucial parameters, e.g. the search radius and the feature offsets for stacking. The proposed feature bundling is a novel contribution and proves to be especially helpful for processing noisy data modalities such as EMG, accelerometer and Kinect motion capture. The authors used a five-point derivation to estimate the direction of movement in the bundling, but when faced with severe noise, one will need more robust methods. This will further reduce variance in the feature space, with few implications, as long as one does not try to synthesize new sequences from the feature space.

Furthermore, the spatio-temporal summarization algorithm proposed in [56] considers 3D motion captured data, instead of RGB information, represented by 3D joints that model human skeleton is introduced. In particular, the proposed approach, 3D joints are derived from the Vicon motion capture system. The advantage of directly handling 3D human skeleton points instead of raw depth data is that few data samples are involved in the processing of the dance sequences, making summarization far more efficient. The authors describe an hierarchical framework taking into consideration the Sparse Modeling Representative Selection algorithm [57]. The basic idea behind this approach is that every image frame of the choreographic sequence can be expressed as a linear combination of one or more representative samples. A dynamic hierarchical layered structure to represent human anatomy is the core of the method proposed in [58], which uses low-level motion parameters to characterize motion in the various layers of this hierarchy, which correspond to different segments of the human body. This characterization is used with a naive Bayesian classifier to derive choreographer profiles from empirical data that are used to predict how particular choreographers segment gestures in other motion sequences. In contrast, the works of [59], [60] propose two summarization approaches: a "time-independent" method based on k-means++ clustering algorithm for the extraction of prominent representative instances of a dance, and a physics-based technique that creates temporal summaries of the sequence at different levels of detail are presented. The main scope of the proposed framework is to extract the most representative instances of the dance, its key postures, or, differently put, its basic primitives, regardless of their order in the sequence. The authors define the selection of the most representative frames as an unsupervised clustering problem. Since a feature vector is assigned for each frame of a dance frame sequence, the vectors of all frames form a trajectory in a high dimensional feature space, which expresses their temporal variation. In the pro-posed work, the authors denote the magnitude of the second derivative of feature vectors for all frames within a sequence with respect to time as a curvature measure. The second derivative expresses the degree of acceleration or deceleration of an object that traces out the feature trajectory.

Summarization can also be useful in the context of fast searching of content in large motion databases, and for efficient motion analysis and synthesis. In [61], the authors demonstrate that identifying locally similar regions in human motion data can be practical even for huge databases, if medium-dimensional feature sets are used for kd-tree-based nearest-neighbor searches. Moreover, efficient approaches for local and global motion matching, which are applicable even to huge databases, have been presented. Moreover, the authors of [62] present a framework that encompasses a connected set of avatar behaviors that can be created from extended, free form sequences of motion, automatically organized for efficient search, and exploited for real-time avatar control using a variety of interface techniques. The motion is pre-processed to add variety and flexibility by creating connecting transitions where good matches in poses, velocities, and contact state of the character exist. An approach for performance animation that employs video cameras and a small set of retro-reflective markers to create a low-cost, easy-to-use system that might someday be practical for home use is introduced in [63]. The low-dimensional control signals from the user's performance are supplemented by a database of pre-recorded human motion. The system automatically learns a series of local models from a set of motion capture examples that are a close match to the marker locations captured by the cameras. A framework for synthesizing dance performance matched to input music, based on the emotional aspects of dance performance is proposed in [64]. This framework consists of a motion analysis, a music analysis, and a motion synthesis component based on the extracted features. In the analysis steps, motion and music feature vectors are acquired. Motion vectors are derived from motion rhythm and in-tensity, while music vectors are derived from musical rhythm, structure, and intensity. On a different note, the work of [65] focuses on the use of game design elements for the transmission of ICH knowledge and, especially, for the learning of traditional dances. More specifically, the authors present a 3D game environment that employs an enjoyable natural human computer interface, which is based on the fusion of multiple depth sensors data in order to capture the body movements of the user/learner. Moreover, the proposed framework automatically assesses the users' performance by using a combination of DTW with FIS approach providing feedback in a form of a score as well as instructions from a virtual tutor in order to promote self-learning. Finally, the authors of

[66] propose ways of comparing two similar dance performances, using the DTW algorithm. The DTW method is validated for use with dance performance motion tracking data by comparing its results with 'ground truth' results obtained from a comparison between videos of two motion tracked performances. The technique was extended to investigate two processes that affect movement timing-scaling (a fixed ratio alteration) and lapsing (caused by insertion or deletion of movement material). The authors applied the method to a comparison of dances performed with a musical soundtrack and without a musical soundtrack.

### 2.2.2 Pose Recognition and Dance Movement Classification

The particularities of dance motion make the already challenging computer vision problems of pose and action recognition even more interesting when explored in a choreographic context. In [27], the authors scrutinized the effectiveness of a series of well-known classifiers (k Nearest Neighbors, Naïve Bayes, Discriminant Analysis, Classification Trees and Support Vector Machines) in dance recognition from skeleton data. In particular, the goal was to identify poses which are characteristic for each dance performed, based on information on body joints, acquired by a Kinect sensor. The datasets used include sequences from six folk dances and their variations. Multiple pose identification schemes are applied using temporal constraints, spatial information, and feature space distributions for the creation of an adequate training dataset. A similar approach for defining choreographic postures from data sequences is introduced in [67]. The selected classifiers are either probabilistic, linear or non-linear kernels.

A framework for body motion analysis in dance using multiple Kinect sensors is presented in [68]. The proposed method applies fusion to combine the skeletal tracking data of multiple sensors in order to solve occlusion and self-occlusion tracking problems and increase the robustness of skeletal tracking. Finally, body part postures are combined into body posture sequences and Hidden Conditional Random Fields (HCRF) classifier is used to recognize motion patterns. Furthermore, a Convolutional Neural Network-based approach for 3D human body pose estimation from single RGB images is presented in [69], addressing the issue of limited generalizability of models trained solely on the starkly limited publicly available 3D pose data is proposed. Using only the existing 3D pose data and 2D pose data, the authors show state-of-the-art performance on established benchmarks through transfer of learned features, while also generalizing to in-the-wild scenes.

A combined approach, involving 3D spatial datasets, noise removal prepossessing and deep learning regression is presented in [70] aiming at the estimation of rough skeleton data. The application scenario involved data sequences from Greek traditional dances. In particular, a visualization application interface was developed allowing the user to load the C3D sequences, edit the data and remove possible noise. The 3D points are selected on the use of a Convolutional Neural Network (CNN) model. Experimental results on real-life dances being captured by the Vicon motion capturing system are presented to show the great performance of the proposed scheme.

In [71], the authors introduce a deep machine learning framework that exploits CNN representational capabilities to identify choreographic postures captured through the RGB channel of a Kinect II capturing device. To increase the performance, a background subtraction algorithm is utilized for pre-processing, so as to minimize the captured noise and only consider the motion data. To enhance the classification performance, a background subtraction framework was utilized, while the CNN architecture was adapted to simulate a moving average behavior. The overall system can be used as an AI module for assessing the performance of users in a serious game for learning traditional dance choreographies. The main scope of the proposed architecture is to develop a pose identification tool for choreographic educational purposes in order to define automatically the appropriate dance postures from a video sequence.

A method for classifying 3D dance motions especially selected from Korean POP (K-POP) dance performance is proposed in [72]. Compared to actions addressed in daily life and existing games, K-POP dance motions are much more dynamic and vary substantially according to the performers. To cope with the variation of the amplitude of pose, a practical pose descriptor based on relative rotations between two body joints in the spherical coordinate system is presented. As a method to measure similarity between two incomplete motion sequences, subsequence DTW algorithm is explored that supports partial matches.

On a different note, the authors of [73] present an algorithm for real-time body motion analysis for dance pattern recognition using a dynamic stereo vision sensor. Dynamic stereo vision sensors asynchronously generate events upon scene dynamics, so that motion activities are on-chip segmented by the sensor. Using this sensor body motion analysis and tracking can be efficiently performed. For dance pattern recognition, a machine learning method based on the Hidden Markov Model is used. On the other hand, in [74], a music-oriented dance choreography synthesis method using a long short-term memory (LSTM)-autoencoder model to extract a mapping between acoustic and motion features is proposed. Moreover, the authors improve the proposed model with temporal indexes and a masking method to achieve better performance.

A novel Spatio-Temporal Laban Feature descriptor (STLF) for dance style recognition based on Laban theory is proposed in [73]. A novel feature descriptor for dance style recognition and test it on Indian Classical Dance (ICD) is presented. Using inspirations from Laban theory, the authors formulate its major entities and model seemingly trivial biological and psychological kinematics of body-motion into features. At another level, the authors of [75] introduce a Bayesian Optimized Bi-directional Long Short Term Memory (LSTM) model, called BOBi-LSTM, that automatically estimates dancers' poses through 3D skeleton data processing. Bi-directionality models non-causal relationships occurred in a dance performance, in the sense that future dancer's steps depend on previous/current steps. Additionally, long-range dependence correlates choreographic primitives on a long time (memory) window. To model the aforementioned principles, the authors modify the conventional LSTM networks under a Bayesian Optimized framework in order to define the best network structure.

Chor-RNN [72] is a recurrent neural network that is trained using a corpus of motion captured contemporary dance. The system can produce novel choreographic sequences in the choreographic style represented in the corpus. Using a deep recur-rent neural network, it is capable of understanding and generating choreography style, syntax and to some extent semantics. Although it is currently limited to generating choreographies for a solo dancer there are a number of interesting paths to explore for future work. This includes the possibility of tracking multiple dancers and experimenting with variational autoencoders that would allow the automatic construction of a symbolic language for movement that goes beyond simple syntax. A multimodal approach to recognize isolated complex human body movements, i.e. Salsa dance steps is proposed in [76]. The proposed framework exploits motion features extracted from 3D sub-trajectories of dancers' body-joints (deduced from Kinect depth-map sequences) using principal component analysis (PCA). These sub-trajectories are obtained thanks to a footstep impact detection module (from recordings of piezoelectric sensors installed on the dance floor). Two alternative classifiers are tested with the resulting PCA features, namely Gaussian mixture models and hidden Markov models (HMM).

Another interesting application is the transfer of motion between human subjects in different dance videos [53]. Given a video of a source person and another of a target person, the main goal of this work is to generate a new video of the target person enacting the same choreography as the source. To address this task, the authors divide the proposed framework into three stages – pose detection, global pose normalization, and mapping from normalized pose stick figures to the target subject. In the pose detection stage the authors use a pretrained state of the art pose detector to create pose stick figures given frames from the source video. The global pose normalization stage accounts for differences between the source and target body shapes and locations within frame. Finally, the authors design a system to learn the mapping from the normalized pose stick figures to images of the target person with adversarial training. In order to extract pose keypoints for the body the authors adopt Open-Pose [71]. For the image translation stage, a framework proposed in the pix2pixHD [75] is provided. Additionally the authors adopt a single 70x70 Patch Generative Adversarial Networks (GAN) for the face discriminator [77].

The work of [78] presents a method for action recognition using depth sensors and representing the skeleton time series sequences as higher-order sparse structure tensors to exploit the dependencies among skeleton joints and to overcome the limitations of methods that use joint coordinates as input signals. Moreover, the authors estimate their decompositions based on randomized subspace iteration that enables the computation of singular values and vectors of large sparse matrices with high accuracy. Specifically, the authors attempt to extract different feature representations containing spatio-temporal complementary information and extracting the mode-n singular values with regards to the correlations of skeleton joints. Then, the extracted features are combined using discriminant correlation analysis, and a neural network is used to recognize the action patterns. The experimental results presented use three widely used action datasets and confirm the great potential of the proposed action learning and recognition method.

### 2.2.3 Laban Movement Analysis

Human movement analysis and recognition is an important field in computer vision area, and is of particular interest in the choreographic domain. Due to the fact that choreographic performances use complex kinesiology movements is necessary to define the notation of the body joints variations. Laban Movement Analysis (LMA) or Kinetography [32] encodes the choreographic sequences of the body joints into dance notations. The Labanotation system encompasses symbols in order to recognize and to encode the human body movements defining a dance score as a music score respectively. Dance notation includes a set of scores, symbols and rules for encoding dance (or movement in general), in a similar way that music notation records music. Labanotation is recognized as one of the most widely used and accurate notation systems for recording dance highlights.

In [51], the authors present a framework based on the principles of LMA that aims to identify style qualities in dance motions. The pro-posed algorithm uses a feature space that aims to capture the four LMA components (Body, Effort, Shape, Space), and can be subsequently used for motion comparison and evaluation. The proposed framework is designed and implemented using a virtual reality simulator for teaching folk dances in which users can preview dance segments performed by a 3D avatar and repeat them.

A mathematical framework that can automatically extract motion qualities, in terms of LMA entities, is presented in [79]. The aforementioned approach aims to distinguish motions with different emotional states. The authors aim to appraise the significance of the proposed features in motion classification using PCA, where the weight of each feature in separating the performer's feeling is presented. A new classification space is introduced based, not only on the basic description of motion such as the posture, but on the motion qualitative and quantitative characteristics. PCA has been also used for dimensionality reduction, resulting in a less complex system; the reduced segments (principal components) are used as input to a SVM classifier, which decides about the segment with respect to emotion.

Moreover, in [80], LabanDance, a serious game for Labanotation is presented. The LabanDance is a real-time game using the Kinect sensor. The user is asked to perform a sequence of moves at a specific time as they are recorded in a score displayed on the screen. The game has two modes of operation. The first is addressed to users with little familiarity with Labanotation and is accompanied by a virtual trainer. In the second, the user is only required to perform the moves based on the score. The game includes four levels with hand, foot, jump, and a level with a combination of all moves. A different aspect of the use of LMA is presented in [81], where the authors describe a framework in order to extract characteristic poses as well as high-light parts from data of dancing movement obtained by motion capturing technique. For this, the theory of LMA has been applied, and the physical feature values corresponding to the LMA components are defined. By observing the change over time of these feature values, body movements corresponding to the LMA components are extracted. In this approach, the authors focus on effort and shape components of LMA.

The similarities between various emotional states with regards to the arousal and valence of the Russell's circumplex model have also been investigated [51]. A variety of features that encode, in addition to the raw geometry, stylistic characteristics of motion based on LMA is presented. Motion capture data from acted dance performances were used for training and classification purposes. The experimental results show that the proposed features can partially extract the LMA components, providing a representative space for indexing and classification of dance movements with regards to the emotion. In [82], an automatic motion capture segmentation method based on movement qualities derived from LMA is presented. LMA provides a good compromise between high-level semantic features, which are difficult to extract for general motions, and low-level kinematic features which, often yield unsophisticated segmentations. The LMA features are computed using a collection of neural networks trained with temporal variance in order to create a classifier that is more robust with regard to input boundaries.

Another work [83] proposes a set of body motion features, based on the Effort component of LMA, that are used to provide sets of classifiers for emotion recognition in a game scenario for four emotional states: concentration, meditation, excitement and frustration. Experimental results show that, the system is capable of successfully recognizing the four different emotional states at a very high rate. From the results achieved the authors conclude that Laban Movement Analysis is a valid and promising approach for emotion recognition from body movements due to the abstract level of Laban technique. Specifically this framework describes that two of Effort's component motion factors, Time and Space can result to high emotion recognition rates.

In the context of educational frameworks, a proposal for analysis and visualization of dance kinesiology based on Labanotation and embodied learning concepts is presented in [84]. The low-cost Kinect sensor is employed to extract skeletal data which are then processed and transformed geometrically. In the sequel, they are analyzed based on the Labanotation system to characterize the posture of the human limbs. Two modules have been developed. The first module serves for recording, analyzing and visualizing body movements. The second module is an application in which the user is required to perform with his upper limbs, a sequence of gestures given by the system in the form of Labanotation symbols. Dance notation consists of a set of symbols and rules for recording dance (or movement in general), in a similar way that music notation records music.

Lastly, a motion analysis framework based on LMA is described in [85], which also accounts for stylistic variations of the movement is presented. Implemented in the context of Motion Graphs, it is used to eliminate potentially problematic transitions and synthesize style-coherent animation, without requiring prior labeling of the data. The effectiveness of the proposed method is demonstrated by synthesizing contemporary dance performances that include a variety of different emotional states. The constructed LMA-based Motion Graph (MG) by default satisfies posture correlation; in the proposed implementation, the authors select the transition with the highest LMA correlation. Although the MG algorithm may encourage transition to frames of other motions where body posture is highly similar, in contrast LMA MG selects those transitions that motion style is more coherent, despite body posture being less similar.

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 [86], pose identification [87], action prediction [88], emotion recognition [89] and background subtraction [90]. In [91], an unsupervised approach is proposed for modelling human activities, while in [7], summarization of folklore dances have been introduced using an hierarchical SMRS algorithm. In this context, the work of [92] has introduced an action recognition framework exploiting dense trajectories. Finally, in [93] 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 [94]. In [95], a CNN neural network model have been introduced for human activity analysis, while the work of [96] uses RGB-D and skeleton data for activity analysis. In [97], the authors introduce a two-stream convolutional neural network structure for action recognition in videos. In this context, the work of [98] introduces a three-stream CNN for action recognition modelling, while the work of [99] proposes CNNs structures on depth maps and postures for human action recognition. Finally, Makantasis el al. [100] introduces a behavioural understanding approach for industrial environments, while in [101], 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 [102] exploiting concepts of visual attention algorithms. In this context, the work of [103] introduces a background modeling algorithm using CNN structures. Similarly, in [104], the authors introduce methods of Mixture of Gaussians to face background dynamics. In [105], 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 [106] the authors are interested in generalizing CNN from low-dimensional regular grids to high-dimensional irregular domains, such as social networks, brain connections or words' embedding, represented by graphs.

### 2.2.4 Discussion

The rapid developments in machine learning and computer vision technologies have enabled a variety of interesting applications in a vast range of domains, including human motion understanding. In this context, several steps have been made by the research community towards a multifaceted analysis of dance. The use of appropriately designed and fine-tuned machine learning models on data acquired by both visual sensors and motion capture devices has led to significant progress in the fields of choreography summarization, dance pose recognition, as well as further analysis of style and emotion, often using Laban Movement Analysis notation (a concise list of important milestones attained is given in Table 2.1). Despite the significant steps already made and, further research is needed towards a deeper understanding and analysis of dance and related elements, such as style, tradition and affect. The advancements of deep learning as well as the increasing accuracy and cost-effectiveness of visual and motion capture sensors are bound to play an important role to this direction in the following years

Table 2.1 Important milestones in the history of the choreographic analysis

| Milestones/Contribution | Motion Capturing Systems | Contributor, Year |
|---|---|---|
| Graphical editor for dance notation | Kinect | [31], 2002 |
| Real-time control of three-dimensional avatars | VICON | [62], 2002 |
| Performance Animation | Pulnix video cameras | [63], 2005 |
| Dancing-to-Music Character Animation | VICON | [64], 2006 |
| Real-time body motion analysis for dance pattern recognition (Hidden Markov Model) | Kinect | [73], 2012 |
| Analysis of dance movements using gaussian processes | Kinect | [107], 2012 |
| Multimodal classification of dance movements using motion trajectories and sound | Kinect | [76], 2013 |
| Hierarchical aligned cluster analysis (HACA) for temporal segmentation | VICON | [84], 2013 |
| Motion indexing of different emotional states using LMA components | n/a | [79], 2013 |
| Dance analysis using multiple Kinect sensors | Kinect | [108], 2014 |
| Dynamic dance warping | VICON | [66], 2014 |
| Emotion Analysis and Classification | VICON | [13], 2015 |
| Classification of Dance Motions with Depth Cameras Using Subsequence Dynamic Time Warping | Kinect | [109], 2015 |
| A Game-like Application for Dance Learning Using a Natural Human Computer Interface | Kinect | [65], 2015 |
| Folk Dance Evaluation Using Laban Movement Analysis | VICON | [13], 2015 |
| Unsupervised Temporal Segmentation of Motion Data | Kinect | [55], 2017 |
| Key postures identification | VICON | [110], 2017 |
| CNN-based approach for 3D human body pose estimation | Monocular camera | [69], 2017 |
| Hierarchical Sparse Modeling Representative Selection | VICON | [111], 2018 |
| Physics-based keyframe selection for human motion summarization | VICON | [112], 2018 |
| Style-based motion analysis for dance composition | VICON | [85], 2018 |
| An LSTM-autoencoder Approach to Music-oriented Dance Synthesis | VICON | [74], 2018 |
| Spatio-Temporal Laban Feature descriptor (STLF) for dance style recognition | n/a | [113], 2018 |
| Everybody Dance Now | n/a | [53], 2018 |
| Human action recognition through third-order tensor representation and spatio-temporal analysis | n/a | [78], 2019 |
| Learning to Generate Diverse Dance Motions with Transformer | n/a | [114], 2020 |
| AI Choreographer: Music Conditioned 3D Dance Generation with AIST++ | n/a | [115], 2021 |

# Chapter 3

# Motion Digitization and Kinesiology Modelling

## 3.1 Introduction

The exploration of the digitization technology , regarding folklore performances, constitutes a significant aim at an European level. On the one hand, the multi-cultural intangible/tangible heritage of Europe gets documented, preserved, and accessible. During the 20th century there have been several attempts to model human creativity in performing arts. Rudolf Laban developed a system of movement notation, that eventually evolved into modern-day Laban Movement Analysis (LMA) [30], which provides a language for describing, visualizing, interpreting, and documenting all varieties of human movement, in an attempt to preserve classic choreographies. More specifically, LMA has been extensively used for analyzing dance performances and creating digital archives of dancing, in the area of education and research. Currently, digital technology has been widely adopted, which greatly accelerates efforts and efficiency of CH preservation and protection. At the same time, it enhances the assimilation of the ICH in the digital era, creating enriched virtual representations. Although, the aforementioned significant achievements for improving the digitization technology towards a more cost-effective automated and semantically enriched representation, protection, presentation and re-use of the CH via the European Digital Library EUROPEANA [1], very few efforts exist in creating breakthrough digitization technology (i.e. audio, visual and stereoscopic recordings). The core subject of the approaches above, focuses on improving the e-documentation (3D modelling enriched with multimedia metadata and ontologies), the e-preservation (standards) and the reuse of ICH traditional artefacts. These efforts are limited to the following well documented technologies:

- Visual Stereoscopic Recordings have been usually utilized for digitizing choreographic performances. It is imperative to declare that the digitization technology through AV recordings is not spatio-temporaly defined adequately, in the sense that there is no possibility for important symbolic characteristics, representing human creativity to be extracted. For this reason, it is difficult for the way (styling) of a dance, the way of expression and the human emotions [116] to be preserved sufficiently. In addition, 2D AV recordings do not allow the implementation of 3D/4D modelling

[1] https://www.europeana.eu

and rendering technologies that result in enriched virtual environments, which enhance physical objects with virtual ones. This enhancement is critical for the preservation of the intangible cultural content and its integration with additional information. The tremendous advance in hardware engineering, has boosted stereoscopic digitization technologies allowing stereo video data in real-time to be captured [29].

- Digitization and preservation of folklore performances using RGB/point clouds sequences. The developed datasets aim to define and exploit new interoperable and compatible to Europeana and UNESCO MoW Library metadata formats, that permit repositoring, archiving and harvesting ICH assets to support new forms of representations. This is a very critical perspective for the protection, preservation and re-use of the CH metadata, since it formulates the framework for archiving the digitized folklore performances.
- The re-use of the e-Folklore Digital libraries contain large amounts of tangible CH content. Nevertheless, the corresponding amounts of digitized ICH content lags significantly. This thesis aims to fill this gap, by providing an open-access and integrated digital CH repository which includes intangible CH content, in arrange to extend the research impact, the re-use of data and boost sustainable industrial growth. This holistic digital repository can be a useful tool in the hands of stakeholders that utilize ML techniques in the domain of ICH.

## 3.2 The Adopted Motion Capturing Systems

Section 3.2 presents the sensors network that is adopted to obtain the choreographic data and metadata. Within our research, two of the most popular motion capture systems; Kinect II and Vicon are adopted. Furthermore, this chapter describes the motion capture workstations, the cameras and each component for the capturing and the calibration process.

### 3.2.1 VICON motion capturing system

Motion capturing systems are widely used in biomechanics sports, computer graphics and computer animation. The effectiveness of motion capturing systems depending on their system setup and are sensitive against variations. Marker properties, optical projections, video-digital conversion, camera configuration, lens distortion, calibration procedures. A set of spherical reflective markers are attached to the research object, in our case is the dancer. The reflective markers are tracked by a number of grayscale cameras which are placed around the research area and via the Vicon software is calculating and calibrating the 3D position for each reflective marker. The Vicon system consists both hardware and software components. The hardware includes 10 high precision and sampling camcorders (Fig 3.1) to

Table 3.1 Comparison of the VICON and Kinect motion capturing systems

| Motion Capture System | Cost | Accuracy | Calibration | Camera Resolution |
|---|---|---|---|---|
| Kinect | Low | Low | Simple | Low |
| VICON | High | High | Difficult | High |

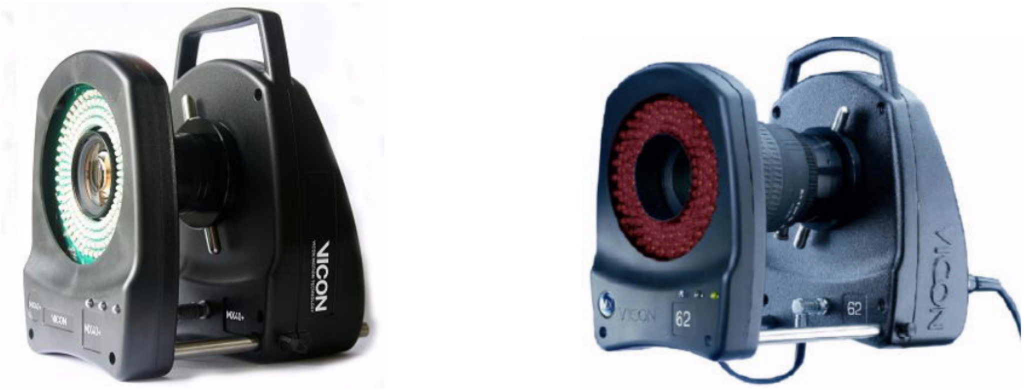


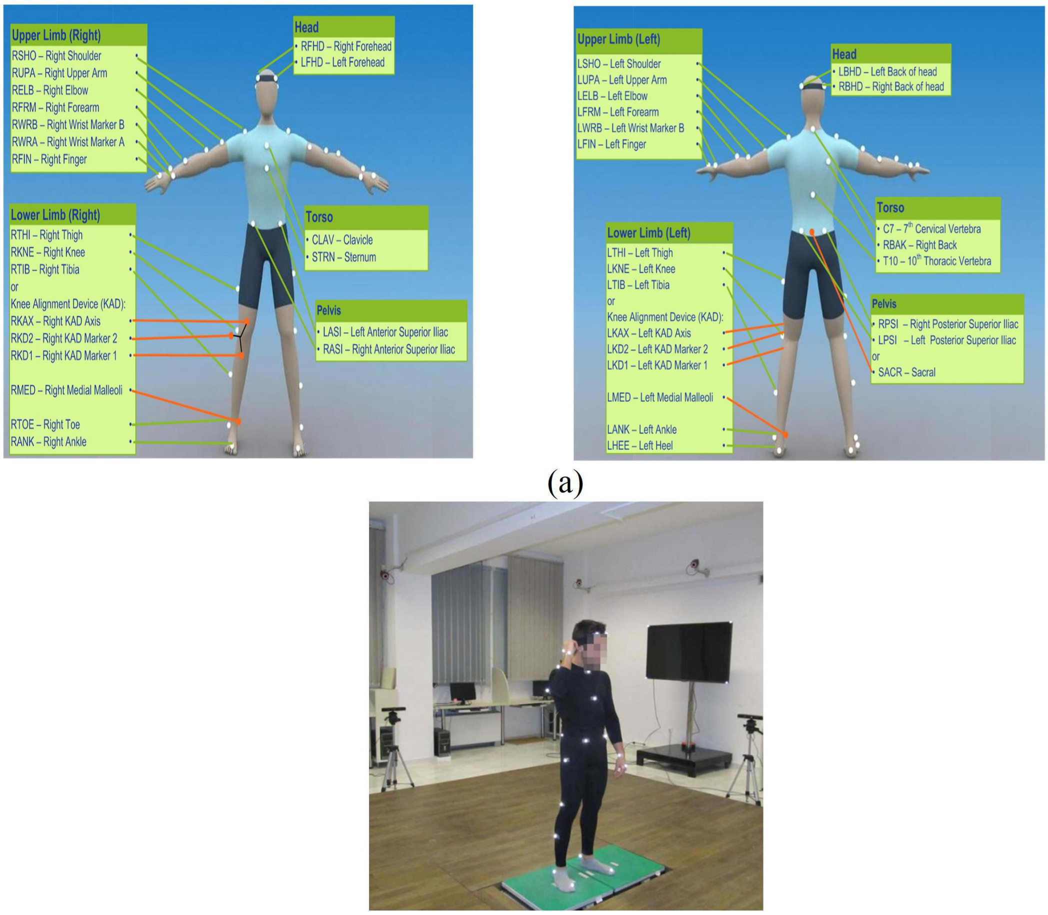
Figure 3.1 Vicon cameras

record human motion, and an analog data acquisition module. The software includes Vicon Workstation that collects and processes the motion and analog data.

A Vicon motion capture area consists of a space surrounded by high resolution Vicon cameras. Each Vicon camera has a ring of LED strobe lights around the Vicon camera lens. The recorded subject has a number of reflective markers to their body. The cameras are recording the subject as it moves in the motion capture space. The Vicon Datastation controls the Vicon cameras and collects the signals, passing them to the host computer on which the Vicon software suite is installed. Vicon workstation is the main application of the Vicon software used to collect, filter and process the raw data. This module processes the 2D data from each Vicon camera, consolidate them reconstruct the 3D motion. This process is depicted in Vicon Workstation as a virtual 3D motion subject. After the aforementioned process the extracted data can be passed to other Vicon applications for further analysis. Our Vicon System consists of the Vicon Datastation, Vicon Workstation, the Vicon Cameras and the Vicon Software Suite. Specifically, the Vicon cameras and the strobes are collecting the light from the reflective markers. The strobe send out light at the same time with the Vicon camera, illuminating the reflective markers. Then, the reflective markers send straight back the light to the Vicon camera. Furthermore, the Vicon Datastation controls the Vicon cameras and every device is used to capture data. In addition, the Vicon Workstation is an application software for controlling the Vicon Data station and the motion capturing process. The heart of the acquisition component adopted for modelling the dancer motion trajectories in 3D space is based on the VICON Motion System [2], which is a motion capturing framework used in several application domains, ranging from gaming, film production, clinical research and entertainment. In our implementation, ten Bonita B3 cameras are included, running the Nexus1.8.5.61009h software. The movement area is a 6.75 meters square. The origin of the VICON coordinate system is the centre of the square surface. A calibration wand with markers is used to calibrate the ten cameras. The user's body is measured by attaching 35 markers at fixed positions on the body. After sticking all the markers, the height, the weight and other specific anthropometric characteristics of the user are measured.

The motion capture area of VICON is surrounded by a number of high resolution cameras with LED strobe light rings (see Fig. 3.1). A setup of VICON workstation is illustrated in Fig.3.2. Reflective markers facilitate the recording of the moving subject by the cameras, while signal collection is controlled

[2] https://www.vicon.com/



(a)

(b)

Figure 3.2 (a) VICON body joints capturing capabilities. (b) Placement of the passive markers to the dancers' body.

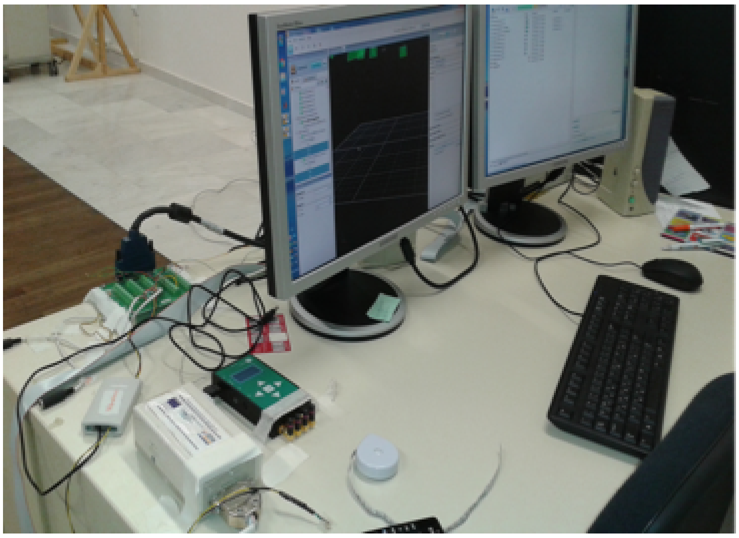


Figure 3.3 Vicon workstation

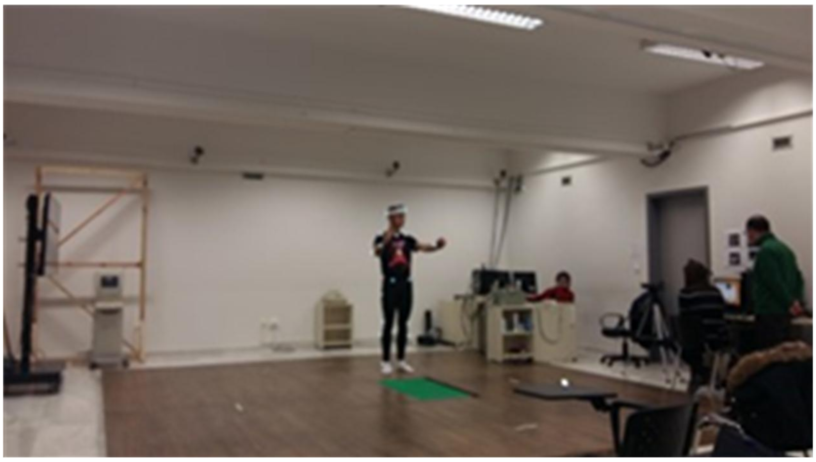
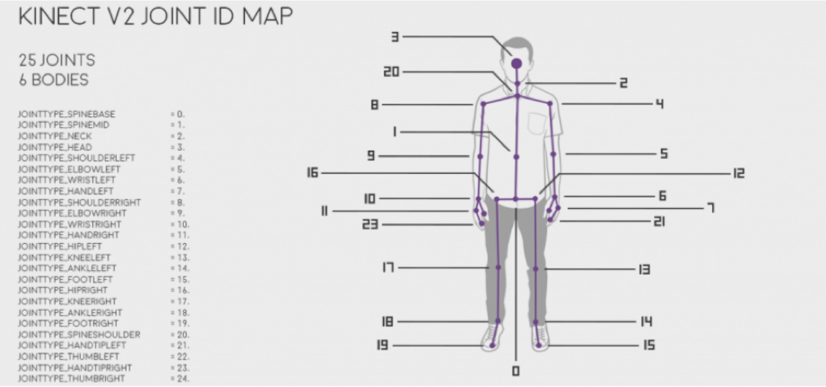


Figure 3.4 A snapshot from the experiments conducted at the Aristotle University of Thessaloniki for capturing the folklore dances.

by Data station controls (see Fig.3.3). Signals are then passed to the VICON workstation, equipped with a specialized software for collection, filtering and processing of raw data. Two-dimensional data from cameras are processed and combined in order for the three-dimensional motion to be reconstructed. Fig 3.2 presents the topology of the markers used for capturing the motion properties of the dancer.In this figure, we depict a dancer, participating in the experiment. The markers are exploited by the VICON component for modelling the 3D dancer attributes and to extract the joints.

### 3.2.2 Kinect-II motion capturing system

The Microsoft KinectTM SDK sensor has a great potential to be adopted for motion capturing system as a low-cost motion analysis tool. It allows to capture 3D objects and human movements and export them to disk for use in 3D packages.In particular, the method applies fusion to combine the skeletal tracking data of multiple sensors in order to solve occlusion and self-occlusion tracking problems and increase the robustness of skeletal tracking. The Microsoft KinectTM SDK sensor contains: (a) a depth sensor, (b) a color camera and (c) a four-microphone array that provide a full body 3D motion capture [117]. More specifically, the depth sensor consists of the infrared projector with the infrared camera. Moreover, the



KINECT V2 JOINT ID MAP

25 JOINTS
6 BODIES

| | |
|---|---|
| JOINTTYPE_SPINEBASE | = 0. |
| JOINTTYPE_SPINEMID | = 1. |
| JOINTTYPE_NECK | = 2. |
| JOINTTYPE_HEAD | = 3. |
| JOINTTYPE_SHOULDERLEFT | = 4. |
| JOINTTYPE_ELBOWLEFT | = 5. |
| JOINTTYPE_WRISTLEFT | = 6. |
| JOINTTYPE_HANDLEFT | = 7. |
| JOINTTYPE_SHOULDERRIGHT | = 8. |
| JOINTTYPE_ELBOWRIGHT | = 9. |
| JOINTTYPE_WRISTRIGHT | = 10. |
| JOINTTYPE_HANDRIGHT | = 11. |
| JOINTTYPE_HIPLEFT | = 12. |
| JOINTTYPE_KNEELEFT | = 13. |
| JOINTTYPE_ANKLELEFT | = 14. |
| JOINTTYPE_FOOTLEFT | = 15. |
| JOINTTYPE_HIPRIGHT | = 16. |
| JOINTTYPE_KNEERIGHT | = 17. |
| JOINTTYPE_ANKLERIGHT | = 18. |
| JOINTTYPE_FOOTRIGHT | = 19. |
| JOINTTYPE_SPINESHOULDER | = 20. |
| JOINTTYPE_HANDTIPLEFT | = 21. |
| JOINTTYPE_THUMBLEFT | = 22. |
| JOINTTYPE_HANDTIPRIGHT | = 23. |
| JOINTTYPE_THUMBRIGHT | = 24. |

Figure 3.5 A list of body joints captured by Kinect. For each joint, position and rotation values are stored in XML format (source: https://vvvv.org/documentation/kinect).

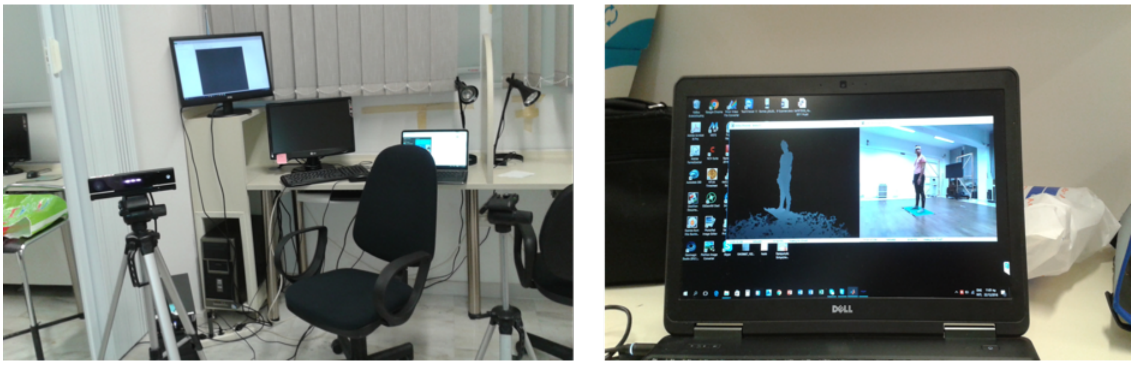


Figure 3.6 Kinect-II workstation

infrared projector is an infrared laser that passes through a diffraction grating and set of infrared dots. Kinect sensor has been discontinued and replaced by Azure Kinect DK with AI capabilities.

In our research, a Microsoft Kinect sensor II is deployed for 3D skeletonization of a dancer. Microsoft Kinect sensor II is a markerless motion capturing framework of low-cost [40]. The extracted skeleton is consisted of twenty joints, each including the 3D coordinates, the rotation parameters and a tracking state property. The topology of the 3D skeleton joints is depicted in Fig. 3.5. The skeleton tracking exploits the human variations, generated by the Kinect sensors.

The Kinect innovation depends on the advantages in human tracking. The skeletal tracking is defined by a number of human joints i.e., head, neck, shoulders and arms. Each identified skeletal joint is represented by a 3D coordinate system. Kinect determines all the 3D joint variations in real-time allowing the interactivity between the tracked subject and the Kinect software [118] [37]. Fig. 3.6 shows a snapshot of the proposed Kinect-II architecture. The tracked skeleton distinguish into twenty five joints with each one to include the 3D position coordinates, its rotation and a tracking state property: “Tracked”, “Inferred”, and “UnTracked” [14]. Furthermore, the sensor work in dark and bright environments and the capture frame rate is 30fps. In parallel, there are some limitations that should be considered: it is designed to track the front side of the user and as a result the front and back side of the user cannot be discerned, and that the movement area is limited (approximately 0.7–6 m)[118].

## 3.3 3D Captured Datasets

### 3.3.1 Single Dancer Dataset

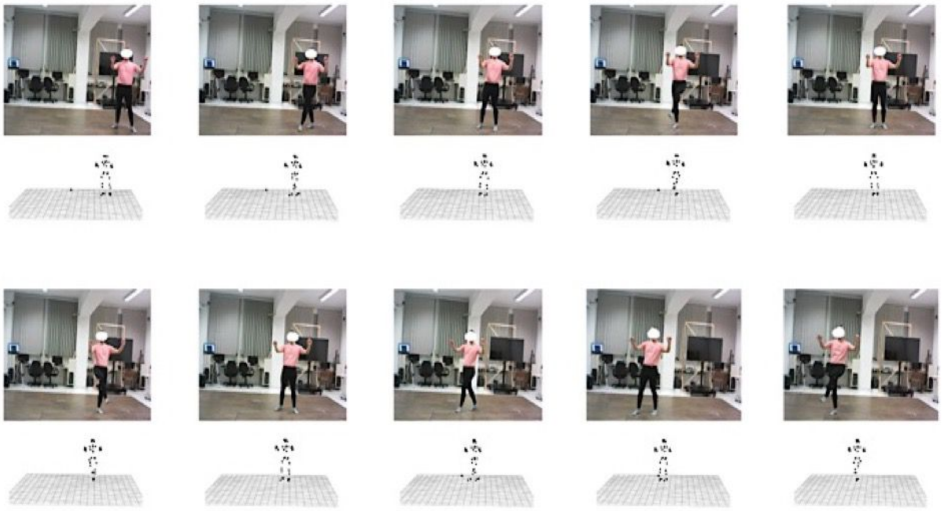


Figure 3.7 An indicative sequence of 10 image frames of “Syrtos” dance by Dancer C, along with the respective skeleton data.

Figure 3.8 Enteka dance performed by dancer I (female) [3].

The first choreographic data set is distinguished into six dances, additionally their execution was in straight and in circle way. Table 3.5, depicts the investigated dances with a description of the cultural information and the main choreographic steps. Each dance is described by a set of RGB images. Every frame ($I_i = 1,...,n$), has a corresponding extensible mark-up language (.XML), (.C3D) and (.CSV) files with positions, rotations and confidence scores for N joints on the body, in addition to timestamps (see Chapter 3.7). The dances in Kinect are described by a matrix, $D_i$, of size b×m×n, where b is the number of body joints (i.e. 25), m is the number of feature vectors (i.e. 3 coordinates and 4 rotations, plus 2 more binary indicators, explaining if values are measured or estimated), and n is the duration of the dance. On the other hand, Vicon capturing described by a matrix with 35 passive markers extracted by the Vicon architecture. $J_i$=1,...,$k$. Subsequently, it is described the six dances with the most representative

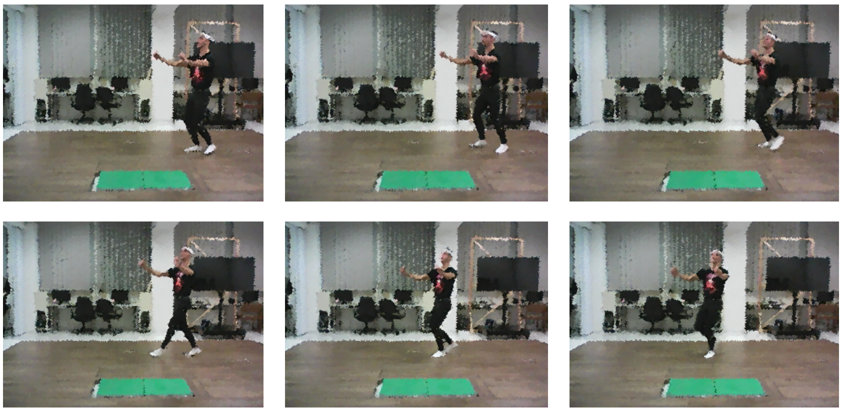


Figure 3.9 Illustration of Syrtos at 2 beat dance [3].

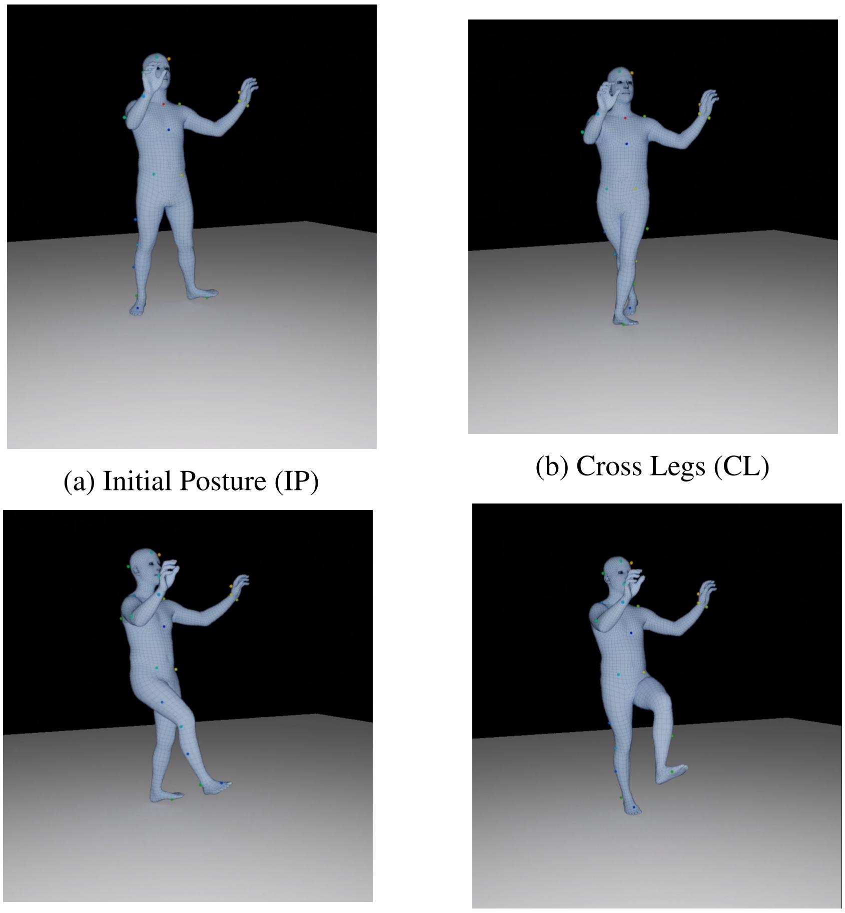
key frames summarizing the choreography pattern. For the teaching/understanding of Greek traditional dances is important to exercise in special preparatory rhythmical steps which characterize a variety of dances such as: (a) Gait in three, (b) Gait in both, (c) Crosswise, (d) Hops, (e) Simple-complex steps. The first folklore dataset encompass the Greek traditional dances performed by 3 dancers. Therefore, six dances containing these steps were selected for this research, (these steps are included in many Greek traditional dances). The dances which are proposed to this research are the following: (i) Enteka (liftings-hops), (ii) Syrtos at two beat (see Fig. 3.9), (iii) Syrtos at three beat, (iv) Makedonikos, (v) Kalamatianos, (vi) Trehatos (Simple-complex steps / liftings-hops). Fig. 3.8 depicts Enteka folklore dance performed by dancer I. Table 3.2 presents the list of the redorded dances and their variations as well as their duration recorded from the Kinectt-II sensor.

Table 3.2 The folklore dances recorded from the Kinect-II sensor [1]

| Dance | Variation | Duration (frames) | | |
|---|---|---|---|---|
| | | Dancer1 | Dancer2 | Dancer3 |
| Enteka | Straight | 749 | 807 | 858 |
| Kalamatianos | Circular | 655 | 593 | 561 |
| | Straight | 304 | 378 | 455 |
| Makedonikos | Circular | 424 | 582 | 409 |
| | Straight | 283 | 367 | 418 |
| Syrtos 2 beat | Circular | 608 | 543 | 352 |
| | Straight | 623 | 639 | 334 |
| Syrtos 3 beat | Circular | 608 | 964 | 947 |
| | Straight | 1366 | 678 | 511 |
| Trehatos | Circular | 991 | 723 | 443 |
| | Straight | 315 | 295 | 355 |

Table 3.3 The folklore dances recorded from the Vicon motion capturing system. Those recordings refer to the first Terpischore dataset.

| Dance | Variation | Duration (frames) | | |
|---|---|---|---|---|
| | | Dancer1 | Dancer2 | Dancer3 |
| Enteka | Straight | 3457 | 4116 | 1897 |
| Kalamatianos | Circular | 1423 | 2449 | 1943 |
| | Straight | 844 | 1256 | 1542 |
| Makedonikos | Circular | 2160 | 1980 | 1529 |
| | Straight | 856 | 1458 | 1789 |
| Syrtos 2 beat | Circular | 2045 | 1727 | 1701 |
| | Straight | 1458 | 1481 | 1495 |
| Syrtos 3 beat | Circular | 4856 | 5754 | 3449 |
| | Straight | 2241 | 2812 | 1698 |
| Trehatos | Circular | 1972 | 2788 | 1832 |
| | Straight | 1329 | 1542 | 1052 |



(a) Initial Posture (IP)

(b) Cross Legs (CL)

(c) Right Leg Up (RLU)

(d) Left Leg Up (LLU)

Figure 3.10 These representation illustrate Syrtos at 3 beat visualized using SMPL agloritthm [4].

### 3.3.2 Two Dancers Dataset

The second 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 second dataset includes the aforementioned folklore dances performed by two dancers simultaneously. Figure 3.11 and 3.12 illustrate an example of the geometric challenges that the presence of two dancers causes to our analysis (see Section 6.2.2). As we observe, the passive markers of the dancers are very close. In this second dataset VICON and Kinect II motion capturing systems are utilized. In this dataset we face the occlusion limitation as explained in Section 6.2.2. The right hand of the left dancer is overlapped with the left hand of the right dancer.

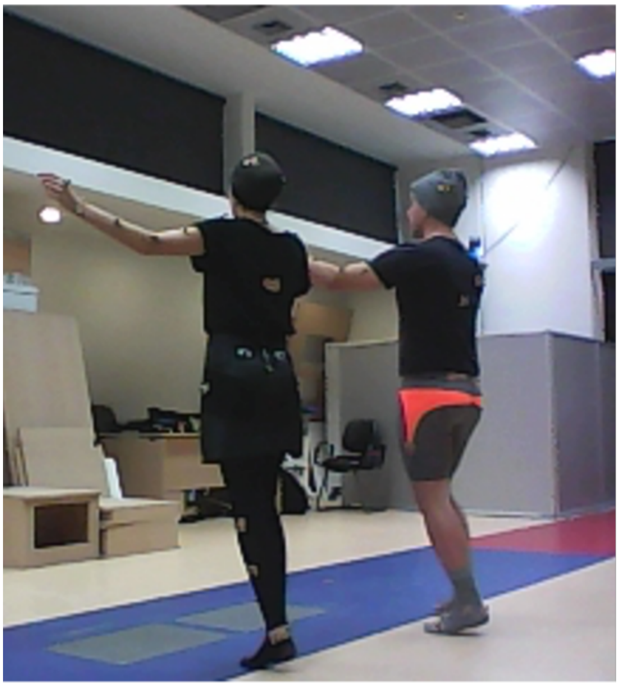


Figure 3.11 The motion capturing process takes into account the variations of the dancers' joints simultaneously [5].

Table 3.4 The folklore dances recorded from the Vicon motion capturing system. Those recordings refer to the second Terpischore dataset.

| Dance | Variation | Duration (frames) |
|---|---|---|
| Makedonikos | Circular | 5430 |
| Syrtos 2 beat | Circular | 3466 |
| Syrtos 3 beat | Circular | 4835 |

Three dance sequences have been recorded using the VICON motion capturing platform [119]. These dance sequences refer to three different performances (dances), each executed simultaneously by two dancers (one male and one female). 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

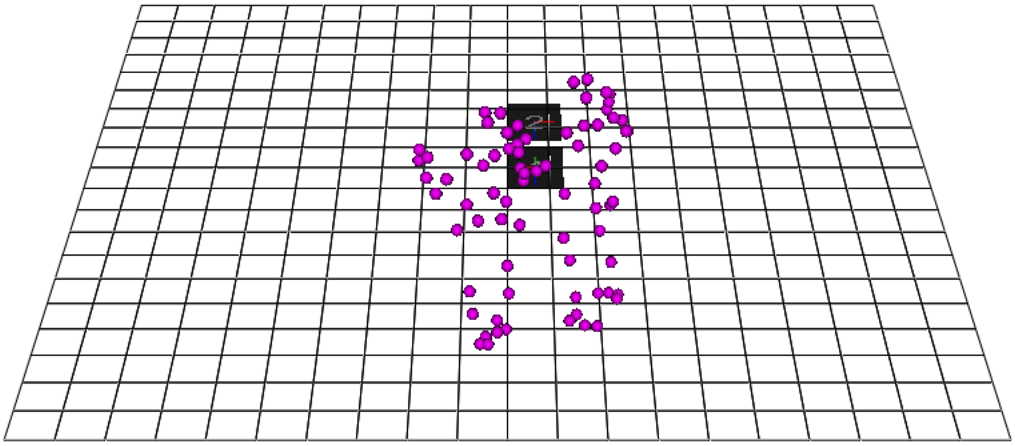


Figure 3.12 Syrtos at 3 beat performed by two dancers simultaneously

(a)

(b)

Figure 3.13 Syrtos at 2 beat performed by two dancers.

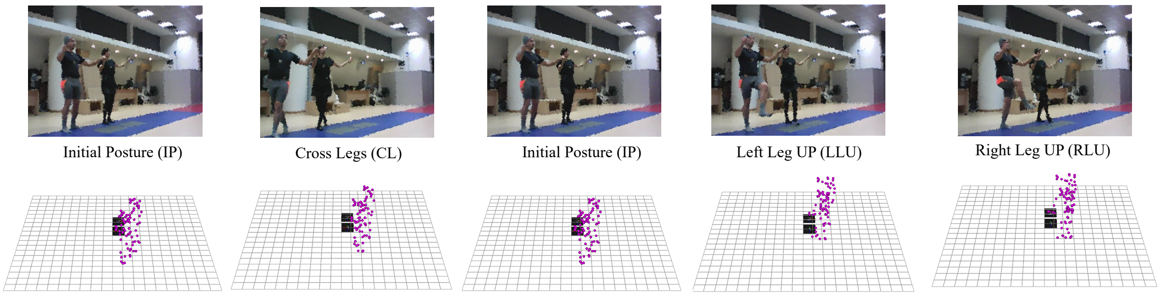


Figure 3.14 The main choreographic steps of Syrtos (3-beat) dance.

tempos. All dancers are professional actors and each dance was executed twice per actor so as to record different paths of the same choreography.

## 3.4 Annotation of the Datasets

The recorded dance sequences refer to five different choreographic sequences (dances), each executed twice by three dancers (two male and one female). The recording processes took place at the School of Physical Education and Sport Science of the Aristotle University of Thessaloniki and at the School of Physical Education and Sport Science of the University of Thessaly in Trikala. All sequences are

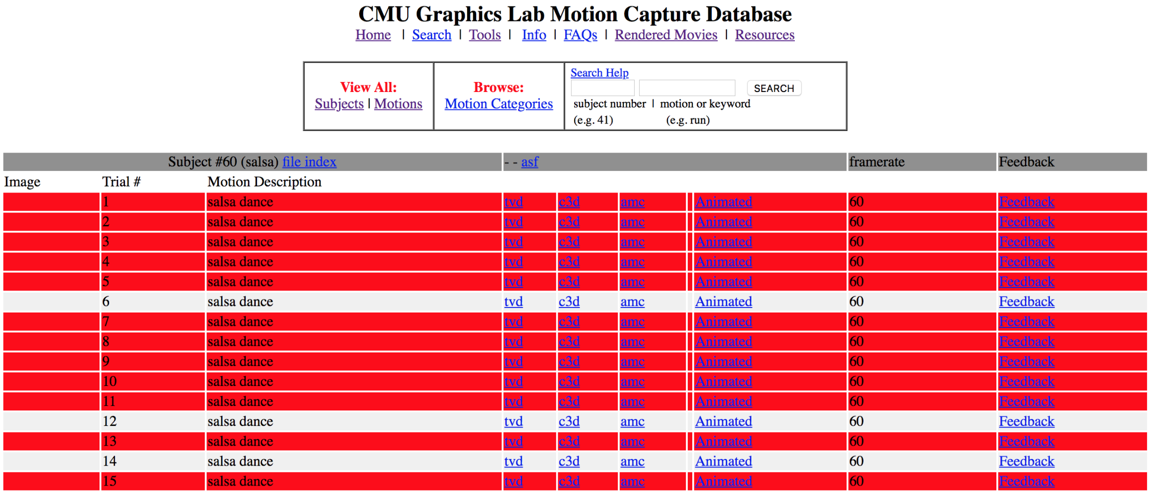
Table 3.5 A brief description of the dances recorded from Vicon.

| **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 folkdance 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) |
| 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) |

Greek traditional folkloric dances, the selection of which was made by dance experts from the Aristotle University of Thessaloniki to fulfill (i) different types of complexities in the dance main patterns, (ii) linear and circular performances of the dance, (iii) different styles the choreography and (iv) different rhythmical tempos. The selection of different human sexes is due to the fact that the main steps of the dances are slightly different between men and women. For men, the dancing style is proud and imperious while for women modest and humble. All dancers are professional actors and each dance was executed twice per actor so as to record different paths of the same choreography. Fig. 3.7 presents a photo of the environment used for the acquisition of the dance sequences using the Vicon motion interface. In Table 3.5, we show the five choregraphies recorded. In this table, we also present a brief description of the dance along with its 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 and Kinect-II capturing modules. Thus, the main steps of the dance (see Table 3.5) 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, the first recorded dance, Sirtos at 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. Then, the main patterns of the dance stop and the choreography starts from scratch. Different steps are recorded for the other types of dances. For example, in Sirtos at 5-beat, except for the initial posture (IP) and cross legs (CL) patterns we also have some leg movements backwards, named as Left Leg Back (LLB). In addition, the cross legs patterns (CL) are sequentially repeated three times. Thus, each of the three CL pattern should be considered as a different choreographic element. Similarly, in Kalamatianos and Trehatos dance, there exists CL patterns repeated sequentially along with cross legs backwards movements. These two dances have totally different rhythmical tempos. Finally, the Enteka dance includes dancer's about-face positions; the dancer is turning around and facing the other position. The proposed methods have been validated in the context of Terpsichore [29], a European research project that aims to create affordable tools for the digitization, modeling, analysis, archiving and promotion of ICH content and, in particular, European folk dances. Fig. 3.7 depicts an example (10 indicative frames along with the respective joints) of a sequence choreography captured for the Syrtos dance.

## 3.5 Choreographic Benchmarked Datasets

Over the past decades, researchers in the domain of Computer Graphics, Computer Vision, Robotics and ICH have worked with large collection of motion sequences to encode human motion applications. The study and analysis of human body movements and gestures is a core issue in various domains including sports and performing arts. Although humans can inherently perceive and decipher such human body signals in an intuitive way, this is a challenging process for artificial computer-based systems. Focusing on the domain of dance, an important aspect is the automatic extraction of the choreographic patterns, which can provide a compact, "bird's eye" representation of the semantic information encoded in the overall



**CMU Graphics Lab Motion Capture Database**

Home | Search | Tools | Info | FAQs | Rendered Movies | Resources

| | | |
|---|---|---|
| **View All:**<br>Subjects \| Motions | **Browse:**<br>Motion Categories | Search Help<br>SEARCH<br>subject number \| motion or keyword<br>(e.g. 41) (e.g. run) |

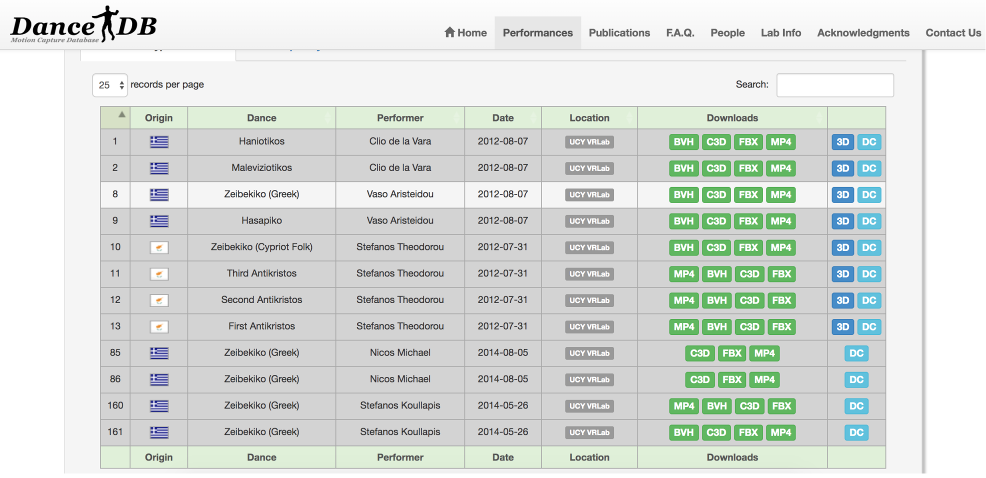
| Subject #60 (salsa) file index | | | - - asf | | | | framerate | Feedback |
|---|---|---|---|---|---|---|---|---|
| Image | Trial # | Motion Description | | | | | | |
| | 1 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 2 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 3 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 4 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 5 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 6 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 7 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 8 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 9 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 10 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 11 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 12 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 13 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 14 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |
| | 15 | salsa dance | tvd | c3d | amc | Animated | 60 | Feedback |

Figure 3.15 Carnegie Mellon Motion Capture Database

dance storyline [120]. Such a compact content representation may be useful in a variety of applications ranging from multimedia systems (e.g., indexing, browsing, content-based search and retrieval) [121] to education (e.g., teaching/learning of a dance choreography) [122], as well as documentation and preservation of the Intangible Cultural Heritage (ICH) assets [123], [124], among which dance holds a prominent spot.

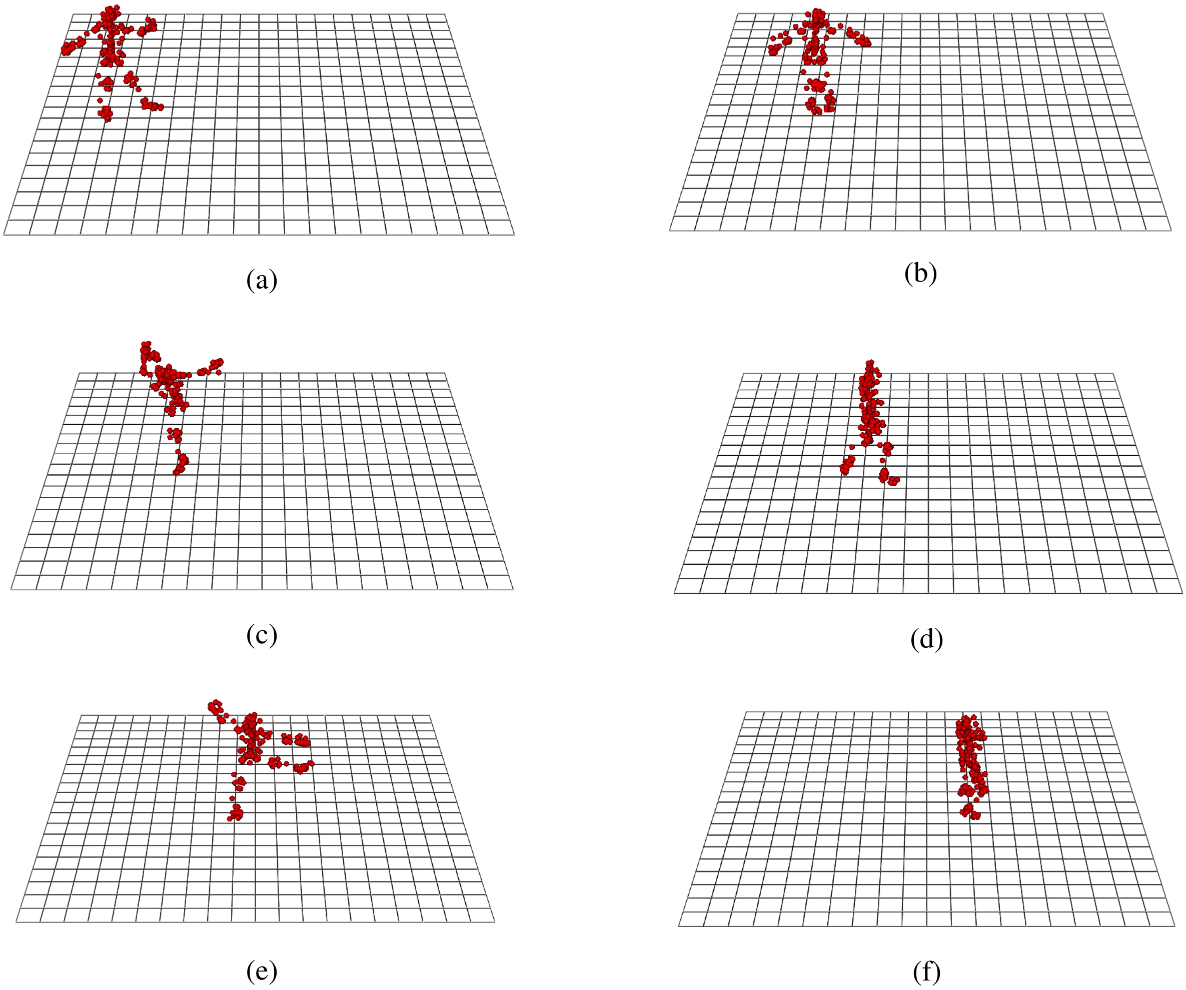
Most of the benchmarked datasets are created taking into consideration specific real-life applications. Some databases such as NTU RGB+D, Berkley MHAD and KIT [3] [4] created for kinessiological analysis of the human movements focus on every-day activities. Databases such as Dance Motion Capture Database of University of Cyprus, Carnegie Mellon Motion (CMU), ACCAD [5], Let's Dance[6] [125] and HDM05 [7] encompass motion capture data in the context of theatro/choreographic expressions. Table 3.6 encompasses the majority of the benchmarked motion databases. Most of the databases provide the raw marker data in C3D, BVH, Autodesk FBX and AMC format. Multi-modal datasets (e.g., IEMOCAP [8]) provide also video, audio, RGB images, labels and physiological recordings describing the choreographic sequences.

[3] https://motion-database.humanoids.kit.edu/list/motions/
[4] https://tele-immersion.citris-uc.org/berkeley$_m$*had*
[5] https://accad.osu.edu/research/motion-lab
[6] https://www.cc.gatech.edu/cpl/projects/dance/
[7] http://resources.mpi-inf.mpg.de/HDM05/
[8] https://sail.usc.edu/iemocap/



| ▲ | Origin | Dance | Performer | Date | Location | Downloads | |
|---|---|---|---|---|---|---|---|
| 1 | | Haniotikos | Clio de la Vara | 2012-08-07 | UCY VRLab | BVH C3D FBX MP4 | 3D DC |
| 2 | | Maleviziotikos | Clio de la Vara | 2012-08-07 | UCY VRLab | BVH C3D FBX MP4 | 3D DC |
| 8 | | Zeibekiko (Greek) | Vaso Aristeidou | 2012-08-07 | UCY VRLab | BVH C3D FBX MP4 | 3D DC |
| 9 | | Hasapiko | Vaso Aristeidou | 2012-08-07 | UCY VRLab | BVH C3D FBX MP4 | 3D DC |
| 10 | | Zeibekiko (Cypriot Folk) | Stefanos Theodorou | 2012-07-31 | UCY VRLab | BVH C3D FBX MP4 | 3D DC |
| 11 | | Third Antikristos | Stefanos Theodorou | 2012-07-31 | UCY VRLab | MP4 BVH C3D FBX | 3D DC |
| 12 | | Second Antikristos | Stefanos Theodorou | 2012-07-31 | UCY VRLab | MP4 BVH C3D FBX | 3D DC |
| 13 | | First Antikristos | Stefanos Theodorou | 2012-07-31 | UCY VRLab | MP4 BVH C3D FBX | 3D DC |
| 85 | | Zeibekiko (Greek) | Nicos Michael | 2014-08-05 | UCY VRLab | C3D FBX MP4 | DC |
| 86 | | Zeibekiko (Greek) | Nicos Michael | 2014-08-05 | UCY VRLab | C3D FBX MP4 | DC |
| 160 | | Zeibekiko (Greek) | Stefanos Koullapis | 2014-05-26 | UCY VRLab | MP4 BVH C3D FBX | DC |
| 161 | | Zeibekiko (Greek) | Stefanos Koullapis | 2014-05-26 | UCY VRLab | BVH C3D FBX MP4 | DC |
| | Origin | Dance | Performer | Date | Location | Downloads | |

Figure 3.16 The Dance Motion Capture Database was created by the Graphics and Virtual RealityLab of Cyprus [6]. http://www.dancedb.eu/


(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.17 This figure illustrates ballet dance stored in CMU database. This choreographic sequence consists of three representative postures; (i) quasi-cou-de-pied, (ii) raised leg above hip-height and (iii) jete en tourant.

Table 3.6 A brief review of the motion capturing databases. This work is an extended version of the [2]

| Databases | Scope and Content | Brief Description | Capturing System |
|---|---|---|---|
| Ohio Dataset (ACCAD) [126] | Video Games and Animation<br>Motion Capture Production<br>Intermedia | 300 sequences<br>-Locomotion<br>-Martial arts | Vicon |
| AffectMe [127]<br>(UCL Interaction Centre) | Study of body<br>posture as an<br>indicator of human affective | Collection of datasets:<br>-Acted emotions<br>-Non-acted affective states | n/a |
| Carnegie Mellon Dataset [128] | General Research<br>Human real-life activities | -Human Interactions<br>-Interaction with Environment<br>-Locomotion<br>-Physical Activities<br>-Situations and Scenarios<br>-Test Motions | Vicon |
| MoCap Database<br>of TH Köln [129] | Unspecified | Arm Gestures<br>Locomotion | Markers |
| Dance Motion Capture<br>Database (CY)<br>[34] | Digital Repository<br>of Folklore Dances | Digitization of<br>Cypriot Folk Dances | Phasespace<br>Impulse X2<br>with 38 markers |
| HDM05 [130] | General Research | 3 hours of motion captures<br>70 different classes | Vicon |

**Table 3.6 continued from previous page**

| Databases | Scope and Content | Brief Description | Capturing System |
|---|---|---|---|
| | | - Locomotion<br>- Grabbing and Depositing<br>- Sports<br>- Sitting and Lying Down<br>- Dance | |
| HumanEva-I<br>[131] | Human movement and pose estimation from video data | - Walking<br>- Jogging<br>- Throw/Catch<br>- Combinations of the above | - ViconPeak |
| IEMOCAP [132] | -Recognition and Analysis of Emotional Expression<br>-Analysis of Human Dyadic Interactions<br>-Sensitive Human Computer Interfaces and Virtual Agents | - Facial expression<br>- Head and hand movements<br>Audio recordings of the conversations | Vicon |
| National University of Singapore (NUS) Capture Database [133] | Unspecified | - Locomotion<br>- Interaction with Obstacles<br>- Martial Arts | Vicon |

**Table 3.6 continued from previous page**

| Databases | Scope and Content | Brief Description | Capturing System |
|---|---|---|---|
| | | - Dance<br>- Yoga | |
| UPenn Database [134] | - Multi-Actor behaviours<br>- Diverse personalities<br>- The effects of posture and dynamics on the perception of emotion<br>- Study human fatigue | Collection of multimodal datasets<br>- Walking<br>- Emotional Actions<br>- Emotional Body Language | Unspecified |
| NTU RGB+D [135], [136] | RGB+D human action recognition | 60 action classes<br>within daily actions, health-related actions, and inter-personal actions | Kinect V2 devices |
| Berkeley Multimodal Human Action Database (MHAD) | RGB+D human action recognition | 11 actions<br>All the subjects performed 5 repetitions of each action, yielding about 660 action sequences | - Impulse<br>- Video: 12 Dragonfly2 cameras<br>-Depth: 2 Microsoft Kinect V2<br>- Acceleration: 6 three-axis wireless accelerometers |

Table 3.6 continued from previous page

| Databases | Scope and Content | Brief Description | Capturing System |
|---|---|---|---|
| KIT [137],[138] | Human pose estimation | 15 actions within upper body movement,<br>full-body upright variations, walking variations, sitting on the floor, and miscellaneous movements | - Mocap: Vicom T40<br>- TOF: Mesa SR4000<br>- Video: Basler piA1000<br>- Body Scan: Vitus Smart LC3 |
| Perception Action and Cognition (PACO) [139] | Human behaviour and brain activity across a variety of research domains | -Ballet<br>-Indian dances | n/a |
| Terpsichore Project [29] | Digitization of Intangible Cultural Heritage | -Greek Folklore Dances<br>Multiple Subjects | Vicon and Kinect |
| AMASS [4] | 4D Scan<br>converting mocap data into realistic 3D human meshes | Transformation from 4D to Mocap coordinates<br>40 hours of motion data,<br>300 subjects,<br>more than 11000 motions | OptiTrak |

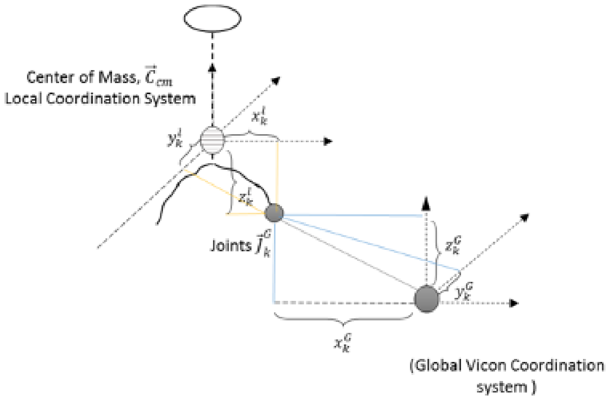

Center of Mass, $\vec{C}_{cm}$
Local Coordination System
$x_k^l$
$y_k^l$
$z_k^l$
Joints $\vec{J}_k^G$
$z_k^G$
$y_k^G$
$x_k^G$
(Global Vicon Coordination system )

Figure 3.18 Transformation of the VICON global coordination system to a local one, the center of which coincides with the center of mass of the dancer. This is an important aspect of analysing the captured moving trajectory of the dance, since dancer spatial positioning is compensated.

## 3.6 Human Body Modelling

In the following, let us denote as $\vec{J_k^G} = (x_k^G, y_k^G, z_k^G)$ the $k$-th joint out of the $M$ extracted by the aforementioned motion capturing systems. Variables $x_i^G$, $y_i^G$ and $z_i^G$ indicate the coordinates of the respective $i$-th joint with respect to a reference point setting by the motion capturing architecture (in our case the center of the square surface). 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.

The main problem in directly processing the extracted joints $\vec{J_k^G}$, $k$=1,2,...,$M$ is that they refer to the Vicon/Kinect coordination system which do not reflect the dancer's position in 3D space and thus the actual choreography. Thus, we need to transform the $\vec{J_k^G} = (x_k^G, y_k^G, z_k^G)$ from the Vicon/Kinect coordinate system to a local coordinate system, the center of which coincides with the center of mass of the dancer. This is obtained through the application of Eq. (7.5) on the joints coordinates $\vec{J_k^G}$,

$$\vec{J_k^L} = \vec{J_k^G} - \vec{C_{cm}} \tag{3.1}$$

where $\vec{C_{cm}}$ denotes the center of mass of the dancer with respect to the Vicon/Kinect coordination system expressed as

$$\vec{C_{cm}} = \sum_{k=1}^{M} \frac{\vec{J_k^L}}{M} \tag{3.2}$$

and we recall that $M$ refers to the total number of joints extracted by the Vicon.

Fig. 3.18 presents the approach adopted in this paper to transform the global Vicon coordinates of the joints $\vec{J_k^G}$ into a local coordinate system ( $\vec{J_k^L}$). The adopted local coordinate system coincides with the center of mass of the dancer. Therefore, the captured skeleton coordinates is transformed with respect to the dancer movement, making them independent from the spatial location of the dancer. It should be mentioned that the local coordinate system is dynamically updated as the dancer is moving in the space throughout the capturing experiment.

## 3.7 Kinematic Representation

In order to model the motion of a dancer, we exploit principles from the theory of rigid body dynamics [140]. In particular, let us denote as $\vec{J}_k^L(t)$ the coordinates of the $k$-th joint at time $t$. Then, the function $\vec{u}_k(t) = d\vec{J}_k^L(t)/dt$ expresses the velocity of the $k$-th joint at time $t$. It is clear that the velocity $\vec{u}_k(t)$ is a vector of three elements, $\vec{u}_k(t) = (u_k^x, u_k^y, u_k^z)$, where variables $u_k^x, u_k^y, u_k^z$ refer to the $x$, $y$, $z$ coordinates of the velocity of the $k$-th joint.

Similarly, the derivative of the velocity expresses the acceleration of a dancer's joint. Therefore, we have that $\vec{\gamma}_k(t) = d\vec{u}_k(t)/dt$. Again, the acceleration is a vector of three elements, expressing the $x$, $y$, $z$ coordinates of the $\vec{\gamma}_k(t)$. The acceleration $\vec{\gamma}_k(t)$ actually models the force imposed on the $k$-th joint at time $t$. In particular, assuming that each joint has a mass $m$=1, the force $\vec{F}_k(t)$ acting on it at time $t$ equals the acceleration $\vec{\gamma}_k(t)$. That is, we have that $\vec{F}_k(t) = \vec{\gamma}_k(t)$. In this way, a state vector $\boldsymbol{S}_k(t)$ is constructed modelling both the joint's position, velocity and acceleration (i.e., the force) concerning the $k$-th joint. Therefore, we have that

$$\boldsymbol{S}_k(t) = \begin{pmatrix} \vec{j}_k^L(t) \\ \vec{u}_k^L(t) \\ \vec{\gamma}_k^L(t) \end{pmatrix} = \begin{pmatrix} x_k^L(t) & y_k^L(t) & z_k^L(t) \\ u_k^x(t) & u_k^y(t) & u_k^z(t) \\ \gamma_k^x(t) & \gamma_k^y(t) & \gamma_k^z(t) \end{pmatrix} \tag{3.3}$$

where again variables $\gamma_k^x, \gamma_k^y, \gamma_k^z$ express the $x$, $y$, $z$ coordinates of the acceleration of the $k$-th joint. It is clear that $S_k(t)$ is a matrix of 3x3 elements.

In order to represent the kinematics of the whole dancer, we take the contribution of all the $M$ available joints. Therefore, a $3 \cdot M \times 3$ state matrix is constructed, expressing the kinematics of the dancer at time $t$.

$$\boldsymbol{S}(t) = \begin{pmatrix} \boldsymbol{S}_1(t) \\ \vdots \\ \boldsymbol{S}_M(t) \end{pmatrix} \tag{3.4}$$

### 3.7.1 Training/Test and Validation DataSet Construction

The study and development of algorithms that can learn from and make predictions on data is a popular task in machine learning [94]. The data used to create the final model usually comes from multiple datasets. More specifically, the model is first fitted on a training dataset, which is a subset of data used to suit the model's parameters (for example, the weights of connections between neurons in artificial neural networks). A supervised learning approach is used to train the model (e.g., a neural net or a naive Bayes classifier) on the training dataset, for example, using optimization methods like gradient descent or stochastic gradient descent. The training dataset is typically made up of pairs of input vectors (or scalars) and output vectors (or scalars), with the response key being usually denoised.

- Training set is a collection of data used in the learning process to train the parameters (i.e., weights) of a machine learning algorithm. A supervised learning algorithm for classification tasks examines the training dataset to assess, or learn, the best combinations of variables that will produce a successful predictive model. The aim of the training dataset is to create a trained (fitted) model that

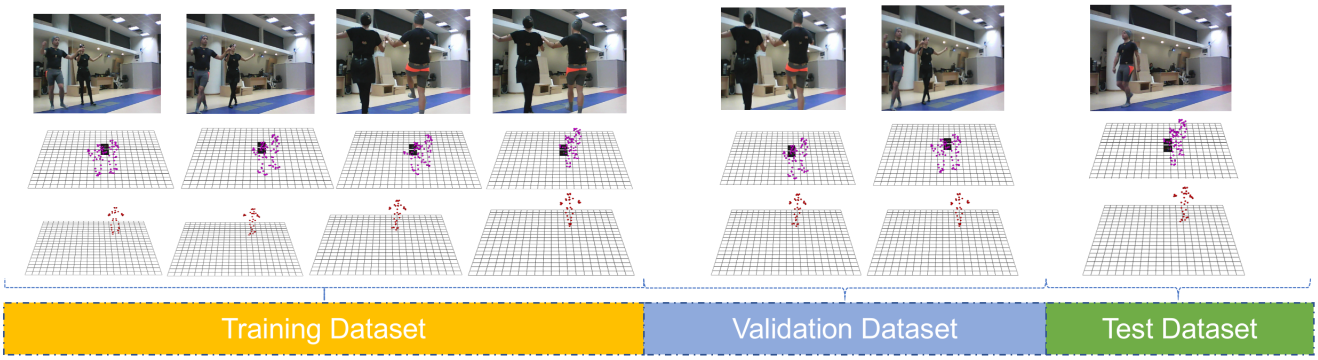


Figure 3.19 Training, Test and Validation datasets. Training dataset consists of sample data (RGB/point clouds) to fit the model. Validation dataset encompasses sample of data while defining models' hyper-parameters. Test dataset used to determine an unbiased evaluation of a final model fit on the training dataset.

accurate estimate unknown data. The fitted model's accuracy in classifying new data is estimated using new examples from the held-out datasets (validation and test datasets). The examples in the validation and test datasets should not be used to train the model to avoid issues like overfitting.

- Validation dataset is a collection of examples used to fine-tune a classifier's hyperparameters (or architecture). The number of hidden units in each layer is an example of a hyperparameter for neural networks. Validation set should have the same probability distribution as the training dataset, as should the testing set (as discussed above). When any classification parameter needs to be modified, a validation dataset, in addition to the training and test datasets, is needed to prevent overfitting. For example, if the most appropriate classifier for the problem is demanded, the training dataset is used to train the various candidate classifiers, the validation dataset is used to compare their outputs and choose which one to use, and the test dataset is used to obtain performance characteristics such as precision, recall and F1-score.

- Test set is a dataset that is unrelated to the training dataset but has the same probability distribution. If a model that fits the training dataset accurate also fits the test dataset, there has been limited overfitting. Overfitting is normally indicated by a better fit of the training dataset compared to the test dataset.

# Part II

# Content-based Sampling of Dance Sequences: Semantic Compression and Summarization

# Chapter 4

# Hierarchical Sparse Modelling Representation for Dance sequences Summarization

## 4.1 Introduction

In performing arts, such as choreography, dance and theatrical kinesiology, movements of human body signals and gestures are essential elements used to describe a storyline in an aesthetic and symbolic way. Although, we, as humans, can inherently perceive and decipher such human body signals in a natural way, this process is challenging for a computer system. One important aspect in the analysis of a performing dance is the automatic extraction of the choreographic patterns/elements since these elements provide an abstract and compact representation of the semantic information encoded in the overall dance storyline [120]. Such an abstract content representation is useful in many applications ranging from multimedia systems (e.g., indexing, browsing, content-based search and retrieval) [121] and education (e.g., teaching/learning of a dance choreography) [122], [141] to documentation and preservation of the ICH assets [123], [124].

Extraction of representative key frames, for an abstract description of a video sequence, is an important topic in multimedia research [52], [142], [143]. Actually, video summarization algorithms are content-based sampling procedures that reduce semantically unimportant or redundant content [144]. One of the first approaches towards video summarization is the extraction of scene (or shot) video segments within a video [145], [146]. In the following years, many other sophisticated algorithms have been proposed aiming at finding representative key frames to efficiently model the content of a video, usually through the application of clustering methods [52], [142], [147], [148], [149], [150] and [151]. These algorithms take visual data in the RGB or HSV color space and appropriately process them to extract feature-related transformations.

However, the recent advantages in software and especially hardware engineering have emerged several devices for capturing, storing and acquiring video content. The innovation of all these acquisition systems is that they capture, apart from the color, the depth information providing, therefore, new ways for modelling human body movements and gestures. Examples include Vicon [152], Kinect [40], PhaseSpace [153] and Xsens [154] architectures which have been used in many diverse application scenarios ranging from gaming, film, animation and the sports industry [155], [156]. Such devices detect and track in space and time a set of key points in order to form a three-dimensional (3D) representation of the human body motion. Exploiting the capabilities of the aforementioned devices, one could improve the performance and efficiency of video summarization, especially when it targets to the detection of choreographic patterns or the analysis of human motion trajectories.

### 4.1.1 Related Works

Works focusing on choreographic acquisition and modelling can be distinguished into those that deal with 3D digitization and capturing and those that mainly focus on the analysis and processing of dances.

Regarding 3D digitization, the work of [33] is considered as one of the first approaches in the field. In particular, this work introduces a 3D archive system for Japanese traditional performing arts. The graph-cuts algorithm is used to reconstruct the 3D model of the scene from multi-view videos. In the same context, the [34] digitizes Cypriot dances using the Phasespace Impulse X2 motion capture system. The architectures uses 8-cameras that are able to capture the 3D motion on modulated LEDs. In the same work, a video game is developed for making the teaching of Cypriot dances more attractive. In [38], the capturing architecture of the i-Treasure European Union funded project is described, mainly focusing on 3D digitization and analysis of rare European folkloric choreographies. A digitization framework suitable for tele-immersive applications of a dance is proposed in [35]. The purpose of this research is to design a creativity framework for dance choreography based on LMA (Laban Movement Analysis) [30]. Advanced motion captured architectures for digitizing folklore performing arts is presented in [36]. In this work, motion analysis algorithms are investigated with the main aim to transform the captured motion trajectories of the dancers into meaningful and semantically enriched LMA features.

Although 3D digitization technologies provide an efficient framework for documentation and preservation of the ICH artifacts of folklore dances, it has the limitation that the delivered 3D data are too large for processing, storing and archiving. For this reason, skeletonization is first performed, which is a process that emphasizes the geometrical and topological properties of the motion trajectories, extracting the medial axis. In this context, Kinect depth senors [37], Phasespace capturing [36] or Vicon [38] motion interface has been exploited.

Regarding choreographic analysis approaches, classification algorithms have been proposed on data expressing the human body movements. In this context, the work of [39] proposes a real-time classification system in detecting choreographed gesture classes. The input data have been acquired using the Kinect depth sensor [40], extracting a 3D wireframe skeleton of the dancers. Another dance classification approach is proposed in [41] using again data capturing from the Kinect sensor. In particular, the authors of [41] combine a Principal Component Analysis (PCA), acting as a feature selection process, with two classifiers; a Gaussian mixture and a hidden Markov model. A combination of principal component and Fisher's linear discriminant analysis, which is called fisherdance, is proposed in [42], for classifying Korean pop dances. The inputs are again from the Kinect sensor.

A dance recognition system is introduced in [43]. The platform compares an unknown move with a specified start and stop against known dance moves. The recognition method consists of a classification algorithm and a template matching using a database of model moves. Similarly, in the works of [44],

[37] a markerless tracking system, exploiting the principles of the Kinect sensor, is presented for motion trajectory interpretation and folklore dance pattern recognition.

Recently, video summarization algorithms have been proposed for choreographic motion trajectories [120]. These scheme exploits input data from a Vicon motion capturing interface and then applies a k-means classification algorithm to find out key frame representatives that abstractly model the choreography. In the broad research area of dance summarization, algorithms focusing on extracting key frames of human actions can be also considered. More specifically, the works of [45] and [46] introduce a classification framework for retrieving representative human actions, while the work of [47] proposes a hierarchical union of sub-spaces for human activity abstraction under a semi-supervised framework. In addition, the work of [48] proposes Histograms of Grassmannian Points for classifying multidimensional time-evolving data in dynamic scenes. A stylistic analysis of the variations of dance movements has been recently proposed in [49]. In addition, in the works of [50] and [157] emotional analysis and characterization of dance sequences are discussed.

### 4.1.2 Innovation and Originality

Video summarization algorithms are distinguished into two main categories. The first groups together video frames according to their similarity in feature space regardless of their temporal interrelations. Therefore, the extracted key representatives are estimated using only spatial properties of the content by globally processing a video sequence. Examples of such methods are the works of [52], [147], [151], [57], [158]. Instead, the second group of algorithms performs the key frame extraction process on the temporal fluctuations of the frame features focusing more on local, instead of global, properties of the visual content. An example of this category is the work of [159] that extracts the key frame representatives exploiting a curvature metric on the time trajectory of the features or the work of [160] that proposes spatial-temporal activity features or even the [47] that introduces a hierarchical sparse subspace clustering (HSSC) for human activity summarization. The last 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.

It is clear that the first group of algorithms is not suitable for a dance analysis since a choreography involves temporal variations and frame inter-relationships which are lost from a spatial-global processing. On the other hand, video synopsis focusing only on temporal feature fluctuations makes the derived summaries highly sensitive to noise and the micro-variations of dancer's steps. This leads to an over-representation modelling of the content, that is, to a large number of key frames. To overcome this problem, temporal-based summarization schemes use low-pass filters to smooth the feature trajectory and thus rejecting noisy key-frames [159]. However, the bandwidth of the low-pass filter significantly affects summarization performance and defining its proper value highly depends on the specific properties of the choreography, the tempo and the dancer's style.

For this reason, we introduced a spatio-temporal video summarization implemented under a hierarchical framework. More specifically, for a given dance video segment, initially global holistic descriptors are extracted to localize the key choreographic steps of the dance. Then, each segment is further decomposed into more detailed video sub-segments, refining the extracted initial (coarse) key representatives. In this way, we combine global features of the choreography with local-based descriptors that better capture the temporal attributes of a dance. This hierarchical video dance decomposition results in extracting a pyramid of key frames that provides a complete overview of a choreography, from a coarse to a fine description. Therefore, the proposed spatio-temporal hierarchical summarization scheme can be useful for various multimedia and computer graphics applications [28], such as fast browsing, storytelling, indexing and content-based retrieval.

Our analysis relies on 3D human skeleton points derived from the Vicon motion capturing interface. The advantage of directly handling 3D human skeleton points instead of raw depth data is that few data samples are involved in the processing of the dance sequences, making summarization much more efficient.

## 4.2 Choreographic Representation

First we extract kinematics attributes representing human body movements. This is performed by processing the 3D coordinates of the skeleton joints of the dancers. As we have described in Section 3 after the digitization process the RGB visual information has been transformed into discrete skeleton joints of $(x, y, z)$. These joints are processed using the methodology described in Section 3.6. In particular, for a joint the velocity, the acceleration are computed for a kinematics modelling of human body movements. More details about this process are described in Section 3.7. After this transformation, every skeleton joint is represented by the following attributes:

$$\boldsymbol{S}_k(t) = \begin{pmatrix} \vec{j}_k^L(t) \\ \vec{u}_k^L(t) \\ \vec{\gamma}_k^L(t) \end{pmatrix} = \begin{pmatrix} x_k^L(t) & y_k^L(t) & z_k^L(t) \\ u_k^x(t) & u_k^y(t) & u_k^z(t) \\ \gamma_k^x(t) & \gamma_k^y(t) & \gamma_k^z(t) \end{pmatrix} \tag{4.1}$$

where again variables $\gamma_k^x, \gamma_k^y, \gamma_k^z$ express the $x$, $y$, $z$ coordinates of the acceleration of the $k$-th joint. It is clear that $S_k(t)$ is a matrix of 3x3 elements.

In order to represent the kinematics of the whole dancer, we take the contribution of all the $M$ available joints. Therefore, a $3 \cdot M \times 3$ state matrix is constructed, expressing the kinematics of the dancer at time $t$.

$$\boldsymbol{S}(t) = \begin{pmatrix} \boldsymbol{S}_1(t) \\ \vdots \\ \boldsymbol{S}_M(t) \end{pmatrix} \tag{4.2}$$

## 4.3 Problem Formulation and Notation

The proposed hierarchical decomposition is graphically shown in Fig. 4.1. In this figure, we have illustrated the hierarchies of the first three layers. In a similar way, one can extend the decomposition to the next layers.

First, let us consider a choreographic video sequence consisting of $K$ image frames. Variable $t_0$ coincides with first choreographic frame, while variable $t_K$ with the last one. The choreographic video sequence is divided into sub-sequences (sub-segments) formed by the key frames. In this way, hierarchies

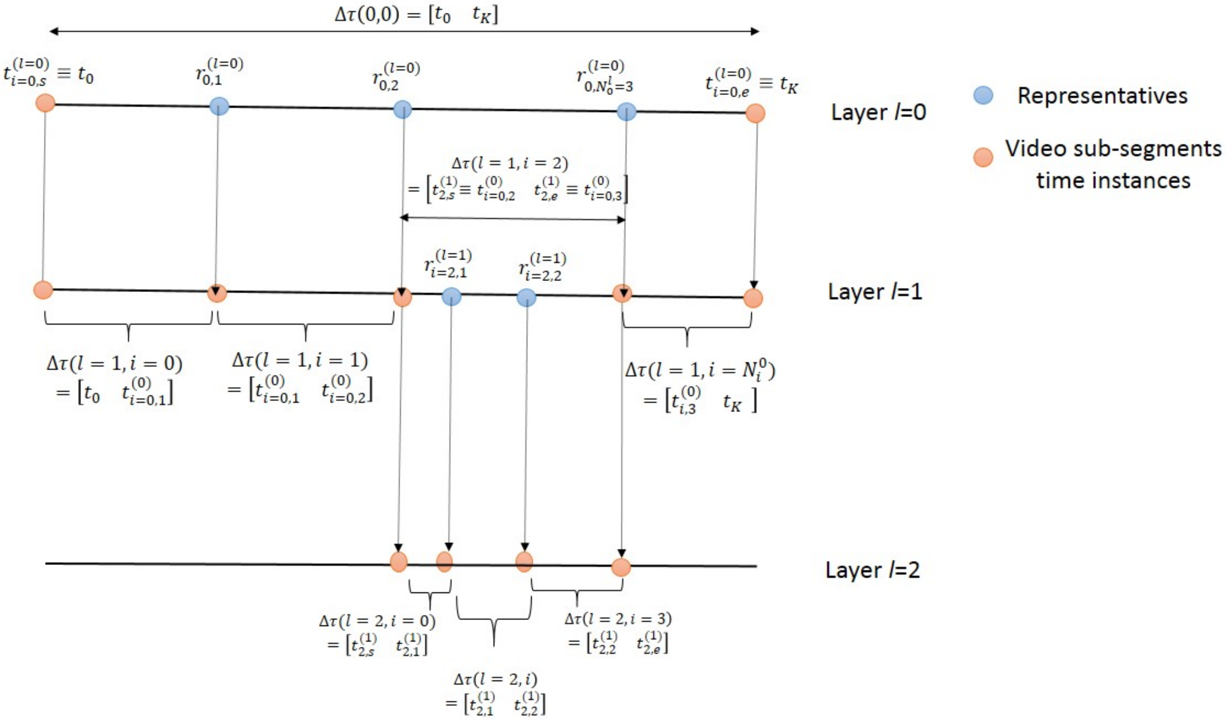

$\Delta\tau(0,0) = [t_0 \quad t_K]$
$t^{(l=0)}_{i=0,s} \equiv t_0$
$r^{(l=0)}_{0,1}$
$r^{(l=0)}_{0,2}$
$r^{(l=0)}_{0,N^1_0=3}$
$t^{(l=0)}_{i=0,e} \equiv t_K$
Layer l=0
Representatives
Video sub-segments time instances
$\Delta\tau(l=1, i=2)$
$= [t^{(1)}_{2,s} \equiv t^{(0)}_{i=0,2} \quad t^{(1)}_{2,e} \equiv t^{(0)}_{i=0,3}]$
$r^{(l=1)}_{i=2,1}$
$r^{(l=1)}_{i=2,2}$
Layer l=1
$\Delta\tau(l=1, i=0)$
$= [t_0 \quad t^{(0)}_{i=0,1}]$
$\Delta\tau(l=1, i=1)$
$= [t^{(0)}_{i=0,1} \quad t^{(0)}_{i=0,2}]$
$\Delta\tau(l=1, i=N^0_i)$
$= [t^{(0)}_{i,3} \quad t_K]$
Layer l=2
$\Delta\tau(l=2, i=0)$
$= [t^{(1)}_{2,s} \quad t^{(1)}_{2,1}]$
$\Delta\tau(l=2, i=3)$
$= [t^{(1)}_{2,2} \quad t^{(1)}_{2,e}]$
$\Delta\tau(l=2, i)$
$= [t^{(1)}_{2,1} \quad t^{(1)}_{2,2}]$

Figure 4.1 An example of the proposed hierarchical decomposition scheme.

of video sub-sequences are derived. Let us denote as $\Delta\tau(l,i)$ the $i$-th video sub-sequence at the $l$-th decomposition layer. Actually, each video sub-sequence $\Delta\tau(l,i)$ is a time interval defined as

$$\Delta\tau(l,i) = \left[ t^{(l)}_{i,s} \quad t^{(l)}_{i,e} \right] \tag{4.3}$$

where variable $t^{(l)}_{i,s}$ expresses the time instance of the first frame of the sub-sequence, while $t^{(l)}_{i,e}$ the time instance of the last frame. In case that $l = 0$ layer only one video segment exits and the $t^{(l=0)}_{i,s} \equiv t_0$ and $t^{(l=0)}_{i,e} \equiv t_K$. Therefore, in this case, we have that $\Delta\tau(0,0) = [t_0 \; t_K]$.

For each sub-sequence $\Delta\tau(l,i)$ a set of $N^{(l)}_i$ key representatives are extracted. We denote these representatives as $r^{(l)}_{i,j}, j = 1,2,\cdots,N^{(l)}_i$ where we recall that index $l$ refers to the $l$-th layer, $i$ to the $i$-th segment of this layer and $j$ to the $j$-th key representative within this segment. Therefore, the $r^{(l)}_{i,j}$ refers to the $j$-th representative of $\Delta\tau(l,i)$. Each representative $r^{(l)}_{i,j}$ is actually referring to a time instance $t^{(l)}_{i,j}$. Let us consider a set $\mathscr{T}^{(l)}_i$ that includes all the time instances $t^{(l)}_{i,j}$ of the representatives $r^{(l)}_{i,j}$. Let us also denote as $\tilde{\mathscr{T}}^{(l)}_i$ an augmented set that also includes the time instances of $t^{(l)}_{i,s}$ and $t^{(l)}_{i,e}$ that defines the sub-sequence $\Delta\tau(l,i)$. Therefore, we have that

$$\mathscr{T}^{(l)}_i = \left\{ \cdots t^{(l)}_{i,j} \cdots \right\} \; j = 1,2,\ldots,N^{(l)}_i \tag{4.4}$$

and

$$\tilde{\mathscr{T}}^{(l)}_i = \mathscr{T}^{(l)}_i \cup \{t^{(l)}_{i,s}, \, t^{(l)}_{i,e}\} \tag{4.5}$$

Using these time instances, the video sub-sequence $\Delta\tau(l,i)$ can be further decomposed into $N_i^{(l)}+1$ non-overlapping sub-segments, since the cardinality of the set $\|\tilde{\mathscr{T}}_i^{(l)}\| = N_i^{(l)}+2$ [see Eq. (4.5)]. Without loss of generality, we assume that the extracted representatives $r_{i,j}^{(l)}$ are sorted with respect to the time instances they refer to. That is, we have that $t_{i,j}^{(l)} < t_{i,j+1}^{(l)}, \forall j$. This way, an ascending order set is defined as

$$\begin{gathered}\mathscr{B}_i^{(l)} = \{t_{i,s}^{(l)} \le t_{i,1}^{(l)} \le \cdots \le t_{i,N_i^{(l)}}^{(l)} \le t_{i,e}^{(l)}\} = \\ = \{\cdots \le t_{i,j}^{(l)} \le \cdots\},\ j=0,1,2,\ldots,N_i^{(l)}+1\end{gathered} \tag{4.6}$$

### 4.3.1 Hierarchical Video Decomposition

Then, the $N_i^{(l)}+1$ video sub-segments are defined as follows:

$$\Delta\tau(l+1,m_i) = \left[t_{m_i}^{(l)} \quad t_{m_i+1}^{(l)},\right] \ m_i = 0,1,2,\ldots,N_i^{(l)} \tag{4.7}$$

In Eq.(4.7), notation $\Delta\tau(l+1,m_i)$ refers to the $m_i$-th video sub-segment of the $l+1$ layer that the $i$-th video segment at the $l$-th layer, i.e., the $\Delta\tau(l,i)$, is further decomposed into. Variables $t_{m_i}^{(l)} \equiv t_{i,m_i}^{(l)}$ and $t_{m_i+1}^{(l)} \equiv t_{i,m_i+1}^{(l)}$, with $t_{i,m_i}^{(l)}, t_{i,m_i+1}^{(l)} \in \mathscr{B}_i^{(l)}$.

Eq.(4.7) defines a set of $N_i^{(l)}+1$ video sub-spaces over which the interval $\Delta\tau(l,i)$ is further decomposed into. In the following, we denote as $\Delta\mathscr{D}(l,i)$ a set that contains all video sub-segments $\Delta\tau(l+1,m_i)$ that the segment $\Delta\tau(l,i)$ is decomposed into. Therefore, we have that

$$\begin{gathered}\Delta\mathscr{D}(l,i) = \mathit{Decomposed}\,\{\Delta\tau(l,i)\} \\ \text{or} \\ \Delta\mathscr{D}(l,i) = \{\cdots\Delta\tau(l+1,m_i)\cdots\},\ m_i = 0,1,2,\ldots,N_i^{(l)}\end{gathered} \tag{4.8}$$

It is clear that the decomposed video segments are mutually exclusive sets and their union equals the initial video segment $\Delta\tau(l,i)$.

$$\begin{gathered}\Delta\tau(l+1,m_i)\cap\Delta\tau(l+1,m_j) = \emptyset, \forall m_i \ne m_j \\ \text{and} \\ \bigcup_{i=0}^{N_i^{(l)}} \Delta\tau(l+1,m_i) = \Delta\tau(l,i)\end{gathered} \tag{4.9}$$

**Example:** Fig 4.1 presents an example of the proposed video decomposition framework. At the first layer three representatives are extracted to model the whole video sequence (marked in blue). Therefore, the initial video sequence is segmented into four further segments, 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. 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 $l=1$. Therefore, the video segment of $\Delta\tau(1,2)$ is further decomposed into three more sub-segments (see Fig. 4.1). This procedure is iteratively applied until the decomposition criterion identifies that no further decomposition is required.

### 4.3.2 Hierarchical Sparse Modelling Representation

In this section, we describe the algorithm for the estimation of the key representatives within a video segment $\Delta\tau(l,i)$. Our approach is based on a sparse representation modeling and it is based on the principles of the SMRS algorithm. The SMRS algorithm actually extracts a set of representative frames that can describe well the whole dance sequence. The approach tries to make the coefficient matrix as sparse as possible so as to achieve reconstruction of the whole dance sequence only from few data samples (key frames).

### 4.3.3 Sparse Modelling Representation Selection (SMRS)

In this section we describe how the representative frames $r_{i,j}^{(l)}$ are estimated using the SMRS algorithm. The basic idea behind this algorithm is that every image frame of the choreography can be expressed as a linear combination of one or more representative samples. In the following notation, we have removed the dependencies on variable $l$ since we refer to a given layer and a given time interval, that is, a video sub-sequence (sub-segment). Let us first assume that each image frame of the choreography, that is a posture of the dancer movement, is represented by a feature vector$\boldsymbol{f}(t)$, where $t = 1,2,\cdots K$ indicates the frame index of the choreography. In this paper, the feature vector$\boldsymbol{f}(t)$ expresses the position and kinematic properties as describe in Section 3.7 of every joint of the dancer. According to the statements of this Section, we have that

$$\boldsymbol{f}(t) = vec\{\boldsymbol{S}(t)\} \tag{4.10}$$

where the operator $vec(\cdot)$ transforms matrix $\boldsymbol{S}(t)$ into a vector by stacking up all the matrix rows. Recalling that $\boldsymbol{S}(t)$ is a $3\cdot M\times 3$ matrix (variable $M$ stands for the number of detected joints), and vector$\boldsymbol{f}(t)$ has size of $9\cdot M\times 1$.

The purpose of the key frame extraction algorithm is to estimate $N \ll K$ representatives that best reconstruct all the $K$ image frames of the choreography. For this reason, let us denote as $\boldsymbol{c}_i$, $i$=1,2,...,$K$ coefficient vectors that approximate the features of the $K$ image frames. Coefficients $\boldsymbol{c}_i$ are of $K\times 1$ size. The elements of the coefficient vectors $\boldsymbol{c}_i$ are estimated through the minimization of the following equation.

$$\sum_{i=1}^{K} \|\boldsymbol{f}(t_i) - \mathbf{F}\cdot\boldsymbol{c}_i\| = \|\mathbf{F}-\mathbf{F}\cdot\mathbf{C}\| \tag{4.11}$$

where matrix $\mathbf{F}$ contains all the feature of the $K$ choreographic frames, that is $\mathbf{F} = [\boldsymbol{f}(t_1)\cdots\boldsymbol{f}(t_K)]$. Additionally, matrix $\mathbf{C} = [\boldsymbol{c}_1\cdots\boldsymbol{c}_K]$ includes all the coefficient vectors $\boldsymbol{c}_i$ , $i = 1,2,\ldots, K$ in a matrix form. Matrix $\mathbf{F}$ is of size $9\cdot M\times 1$, while $\mathbf{C}$ of $K\times K$. Each row of matrix $\mathbf{C}$ expresses how the features of the representative corresponding to this row contribute to reconstruct the features of all the $K$-th frames as a linear combination. Usually, the Frobenius norm is considered to model $\|\cdot\|$. The Frobenius norm is defined as the square root of the sum of the absolute squares of the matrix elements.

Eq.(4.11) implies a linear relationship between each image frame and all the others. To choose $N \ll K$ representatives that best reconstruct the choreographic data, we enforce the following constraint for the matrix $\mathbf{C}$.

$$\|\mathbf{C}\|_{0,q} \leq N \tag{4.12}$$

where the norm $\|\cdot\|_{0,q}$ is defined as follows, $\|\mathbf{C}\|_{0,q} = \sum_{i=1}^{K} I(\|\boldsymbol{c}^i\| > 0)$, where $\boldsymbol{c}^i$ refers to the $i$-th row of matrix $\mathbf{C}$. The $I(\cdot)$ denotes an indicator function which returns 1 if the condition $\|\boldsymbol{c}^i\| > 0$ is true and otherwise it returns 0. In other words $\|\mathbf{C}\|_{0,q}$ counts the number of nonzero rows of $\boldsymbol{C}$. The indices of the nonzero rows of $\boldsymbol{C}$ corresponds to the $N$ out of $K$ representatives. The constraint of Eq.(4.12) implies that matrix $\boldsymbol{C}$ is block-sparse, having only $N$ out of $K$ rows nonzero.

To make the selection of representatives invariant with respect to a global translation of the data, we set an additional constraint regarding matrix $\boldsymbol{C}$.

$$\mathbf{1}^T \cdot \mathbf{C} = \mathbf{1}^T \tag{4.13}$$

Therefore, we have the following optimization problem,

$$\begin{gathered} \min_{\mathbf{C}} \|\mathbf{F} - \mathbf{F}\cdot\mathbf{C}\|_F \\ s.t. \\ \|\mathbf{C}\|_{0,q} \leq N \quad \text{and} \quad \mathbf{1}^T \cdot \mathbf{C} = \mathbf{1}^T \end{gathered} \tag{4.14}$$

Minimization of Eq.(4.14) is a NP-hard problem [57]. For this reason, one way to estimate matrix $\mathbf{C}$ is to relax the constraint of $\|\mathbf{C}\|_{0,q}$ to $\ell_1$-norm. Therefore, Eq.(4.14) is written as

$$\begin{gathered} \min_{\mathbf{C}} \|\mathbf{F} - \mathbf{F}\cdot\mathbf{C}\|_F \\ s.t. \\ \|\mathbf{C}\|_{1,q} \leq \tau \quad \text{and} \quad \mathbf{1}^T \cdot \mathbf{C} = \mathbf{1}^T \end{gathered} \tag{4.15}$$

where $\|\mathbf{C}\|_{1,q} \equiv \sum_{i-1}^{K} \|\mathbf{c}^i\|_q$. This means that the norm $\|\cdot\|_{1,q}$ expresses the sum of the $\ell_q$ norms of the rows of matrix $\mathbf{C}$, while scalar $\tau > 0$ is an appropriately chosen parameter. In this case, we have selected $\tau$ instead of $N$ since for the $N$-th optimal representatives the norm $\|\cdot\|_{1,q}$ is not necessarily bounded by $N$. in this paper, we have selected $q = 2$.

The optimization of Eq.(4.15) is performed using the Alternating Direction Method of Multipliers (ADMM) of [161]. Actually, this method comprises of iterative steps, taking into consideration the Lagrange multipliers of Eq.(4.15).

### 4.3.4 Key Frames Estimation

The main difficulty of applying the optimization strategy of the relaxed constraint problem of Eq.(4.15) instead of Eq.(4.14), is that the estimated matrix $\mathbf{C}$, as derived from the ADMM algorithm [161], has not only $N$ non-zeros rows out of the $K$ available. This is mainly due to the fact that the constraint $\|\mathbf{C}_{0,q}\| \leq N$ of Eq.(4.15) has been relaxed to $\|\mathbf{C}_{1,q}\| \leq \tau$, that is the $\ell_0/\ell_q$ norm has been substituted to $\ell_1/\ell_q$ norm. One way to estimate the most representative key frames is based on the values of the norm $\|\cdot\|_q$ for every row of $\mathbf{C}$. Particularly, the most important key frame is the one that has larger $\ell_q$-norm values of the respective row of the matrix $\mathbf{C}$ than the less important representatives. Therefore, the $N$ representatives are selected as the ones that satisfy the

$$\|\boldsymbol{c}^{i_1}\|_q \geq \|\boldsymbol{c}^{i_2}\|_q \geq \cdots \geq \|\boldsymbol{c}^{i_N}\|_q \tag{4.16}$$

where indices $i_1, i_2, \cdots, i_N$ corresponds to those rows of matrix $\mathbf{C}$ referring to the $N$ representatives.

### 4.3.5 Representative Error Modeling

In this section, we discuss the representativity capabilities of the extracted key-frames. As we have previously stated, each image frame $I(t)$ with a video segment $\Delta\tau(l,i)$ is represented by the feature vector $\boldsymbol{f}(t), \forall t \in \Delta\tau(l,i)$ [see Eq.(4.9)]. First, the minimum distance of any image frame within $\Delta\tau(l,i)$, that is $\forall I(t) : t \in \Delta\tau(l,i)$, with respect to the features of the representatives $r_{i,j}^{(l)} \in \mathscr{T}_i^{l}$ is computed. Then, the error modeling is defined as follows

$$\mathscr{E}_i^{(l)} = \min_{\forall t_{i,j}^{(l)} \in \mathscr{T}_i^{(l)}} \|\boldsymbol{f}(t) - \boldsymbol{f}(t_{i,j}^{(l)})\|_2, \; \forall t \in \Delta\tau(l,i) \tag{4.17}$$

where we recall that $\|\cdot\|_2$ is the $\ell_2$-norm and $t_{i,j}^{(l)}$ is $j$-th representative of the $i$-th segment at $l$-th layer. Based on Eq.(4.17), the average minimum representative error over all image frames within $\Delta\tau(l,i)$ is obtained as

$$\mathscr{E}_i^{(l)} = \sum_t \mathscr{E}_i^{(l)} / \|\Delta\tau(l,i)\|, \; \forall t \in \Delta\tau(l,i) \tag{4.18}$$

where the operator $\|\cdot\|$ denotes the cardinality of the video segment $\Delta\tau(l,i)$, that is the number of frames it has. Actually, Eq.(4.18) expresses a representation metric; the minimum possible error that the representatives can reconstruct the video segment of $\Delta\tau(l,i)$.

Algorithm 1 shows the main steps of the proposed spatio-temporal hierarchical summarization algorithm used to model the choreographic attributes of a dance.

## 4.4 Experimental Results

In this subsection, we define some objective metrics through which the evaluation of the proposed hierarchical algorithm was performed and compared against other techniques. We also present the way used to form the ground truth dataset.

### 4.4.1 Description of the adopted Dance sequences

As we have described in Section 3.5, under the framework of TERPSICHORE project [29], thirty (30) dance sequences was recorded using the Vicon motion capturing platform [152]. These dance sequences refer to five different choreographies (dances), each executed twice by three dancers (two male and one female). The recording process took place at the School of Physical Education and Sport Science of the Aristotle University of Thessaloniki. All sequences are Greek traditional folkloric dances, the selection of which was made by dance experts from the Aristotle University of Thessaloniki to fulfill (i) different types of complexities in the dance main patterns, (ii) linear and circular performances of the dance, (iii) different styles the choreography and (iv) different rhythmical tempos. The selection of different human sexes is due to the fact that the main steps of the dances are slightly different between men and women. For men, the dancing style is proud and imperious while for women modest and humble. All dancers are

```
**Initialization**
Set $l \rightarrow 0$; $i \rightarrow 0$;
Set $\Delta\tau(l,i) \rightarrow [t_0\, t_K]$;
Set Segment $\rightarrow \emptyset \bigcup \Delta\tau(l,i)$;
while $Segment \neq \emptyset$ do
    while $l \leq MaxLayers$ do
        $i \rightarrow 0$
        /* Take a video segment from Segment */
        $VG(l,i) \rightarrow Retrieveasegment(\text{Segment}, l, i)$;
        while $VG(l,i) \neq \emptyset$ do
            /* Apply the SMRS algorithm to extract key frames */
            $\mathscr{T}(l,i) \rightarrow SMRS(VG(l,i))$; -see Section 4.3.3
            $\tilde{\mathscr{T}}(l,i) \rightarrow Augment(\mathscr{T}(l,i), VG(l,i))$; -see Eq.(4.5)
            /* Order the extracted key representatives */
            $\mathscr{B}(l,i) \rightarrow Order(\tilde{\mathscr{T}}(l,i))$; -see Eq.(4.6)
            /* Create Sub-segments of VG */
            $\Delta\mathscr{D}(l+1) \rightarrow Decomposed(VG(l,i))$;
            /* Include $\Delta\mathscr{D}$ into Segment*/
            Segment $\rightarrow IncludeSet(\Delta\mathscr{D}(l+1), \text{Segment})$;
            /* Take the next segment at layer $l$ */
            $i \rightarrow i+1$;
            $VG(l,i) \rightarrow Retrieveasegment(\text{Segment}, l, i)$;
        end
        $l \rightarrow l+1$;
    end
end
```

**Algorithm 1:** The pseudocode of the proposed spatio-temporal hierarchical decomposition scheme

professional actors and each dance was executed twice per actor so as to record different paths of the same choreography.

The evaluation was also performed on three sequences recorded from Carnegie Mellon university and are available for free (at http://mocap.cs.cmu.edu/search.php?subjectnumber=%&motion=%). The sequences include 3D joint points of a human who performs (i) a long time theatrical kinesiology and (ii) two characteristic short-time dance pirouettes. The first dataset is adopted to evaluate the performance of our scheme to different types of movements as the theatrical ones. In this example, the actor is first walking across the scene. She/he then ascends a ladder, sits on the last ladder rung and performs a short-term acting on it by moving her/his head and standing up. She/he then descends the ladder and moving again on the scene. She/he then bends to go beneath an obstacle (e.g., a beam), after passing the obstacle she/he bends again. Then, she/he makes an about-face, moving below the obstacle and then walks on the scene again to ascend the ladder for a second time.

The other two datasets from Carnegie Mellow describes an actor performing dance pirouettes. The first is a dance choreography in which the actor performs a short-time dance including some pirouettes while the second includes at least three about-faces and rotations across the actor's axis.

### 4.4.2 Evaluation Metrics

The precision and recall metric is used to objectively assess the performance of all the evaluated algorithms. Precision measures the ratio of all relevant retrieved key frames over the total number of retrieved key frames by the use of an algorithm. Recall measures the ratio of all all relevant retrieved key frames over the total number of relevant frames in ground truth set. The main challenge in defining the precision and recall metrics in our cases is that key frames from a dance sequence should be ordered. This is due to the fact that the patterns composed of the main steps of a choreography should be specific for a given time internal depending on the music tempo and the type of the dance. In addition, the steps are ordered so that after the execution of a certain pattern (step) another specific step should be followed. For this reason, we define an ordered ground truth set, say $S^{gt} = \{\cdots \Delta\tau_i^{gt} \cdots\}$ that contains the $L$ ordered time internals within which the $i$-th representative frame should be in. For example, the set $S^{gt}$ contains time intervals with starts and ends points that model the main steps of a choreography such as the ones of Table 3.5. We assume that $N$ key frames have been retrieved by the application of a summarization algorithm, denoting as $r_j$. Each representative $r_j$ corresponds to a time instance $t_j$. We have removed variables $l$ and $i$ for simplicity since we are referring to a given video segment $\Delta\tau(l,i)$. It is clear that $N \neq L$ meaning that the number of extracted frames do not coincide with the number of the ground truth choreographic elements.

Then, we define an operator that returns for a given ground truth time interval, say $\Delta\tau_k^{gt}$ 0 if no retrieved key frame falls within this time interval and 1 if one or more key frames falls within this interval.

$$I_{rel}(\Delta\tau_k^{gt}) = \left\{ \begin{array}{ll} 0 & \text{if} \nexists\, t_j \in \Delta\tau_k^{gt} \;\; \forall j \\ & \\ 1 & \text{if} \exists\, t_j \in \Delta\tau_k^{gt} \;\; \forall j \end{array} \right\} \tag{4.19}$$

Eq. (4.19) means that if two or more key frames fall within the same ground truth time interval, only the first is counted as relevant whereas the rest ones are considered as irrelevant. This is due to the ordered nature of a dance sequence. In other words, many key frames depicting the same step of a choreography are erroneous since they do not contribute to the overall choreography. In case that some key frames do not correspond to any time interval, these are also ignore and are not counted as relevant. Thus, precision is defined as

$$\text{Pr} = \frac{\sum_{k=1}^{L} I_{rel}(\Delta\tau_k^{gt})}{N} \tag{4.20}$$

and recall as

$$\text{Re} = \frac{\sum_{k=1}^{L} I_{rel}(\Delta\tau_k^{gt})}{L} \tag{4.21}$$

where variable $N$ refers to the number of key frames and $L$ to the number of ground truth choreographic elements.

Ideally, Pr and Re should be 1 for an excellent retrieval. By combining both criteria, we can derived the F1-score as

$$\text{F1} = 2 * \frac{\text{Pr} * \text{Re}}{\text{Pr} + \text{Re}} \tag{4.22}$$

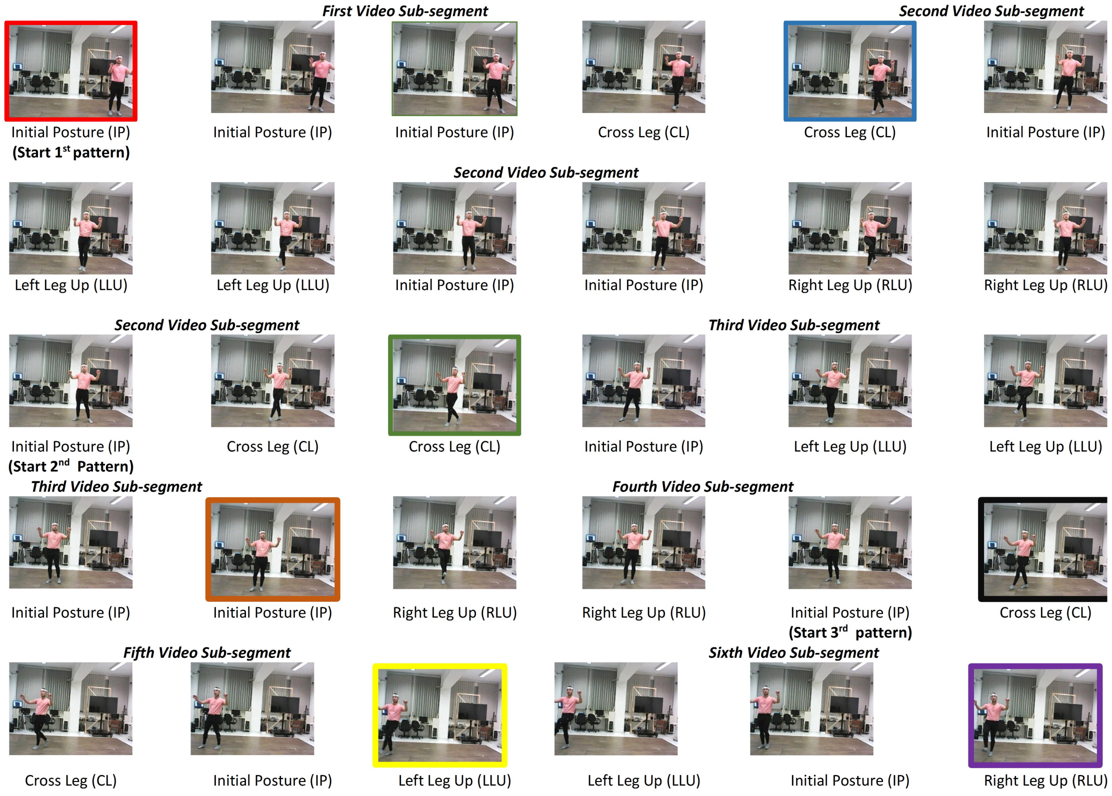


Figure 4.2 Thirty images of the video sequence of the dance Sirtos 3-Beat at constant time intervals. Images with a color frame correspond to the key frames extracted at the initial (coarse) layer using the proposed hierarchical summarization algorithm. We have used different colors for the image frames to better distinguish the video sub-segments on which decomposition takes place.

Values of F1-score close to one yields good retrieval performance, whereas low values indicate a performance which is not satisfactory.

### 4.4.3 Ground Truth Dataset

Regarding the Vicon dataset recorded at University of Thessaloniki, ground truth data was created by the respective dance experts. The ground truth includes the set of desired key frames as being specified by the experts and time instances of the choreography within which a key frame can be considered as relevant. The choreographic elements coincides with those of Table 3.5. Regarding the three examined datasets from Carnegie-Mellon University, we define some characteristic time intervals so as to describe the kinesiology and the dance patterns. Based on these ground truth time intervals, the objective evaluation of the proposed hierarchical scheme is carried out through measuring the precision, recall and F1-score. The same ground truth dataset was used for comparing the proposed algorithm with others proposed in the literature.

### 4.4.4 Experiments

Fig. 4.2 depicts thirty images of the dance video sequence of Sirtos 3-Beat from C-Dancer. Each image corresponds to a constant time interval. In this figure, we have also depicted the respective annotations

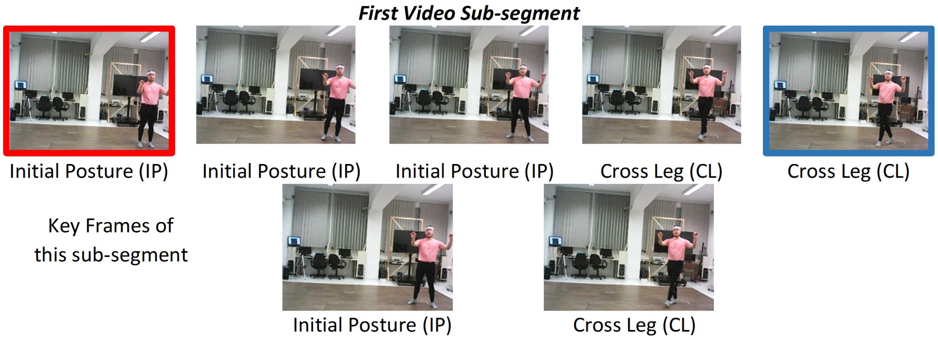


Figure 4.3 The key frames extracted for the first video sub-segment of the dance of Fig. 4.2. In this case, two key frames are extracted regarding the second layer of representation.

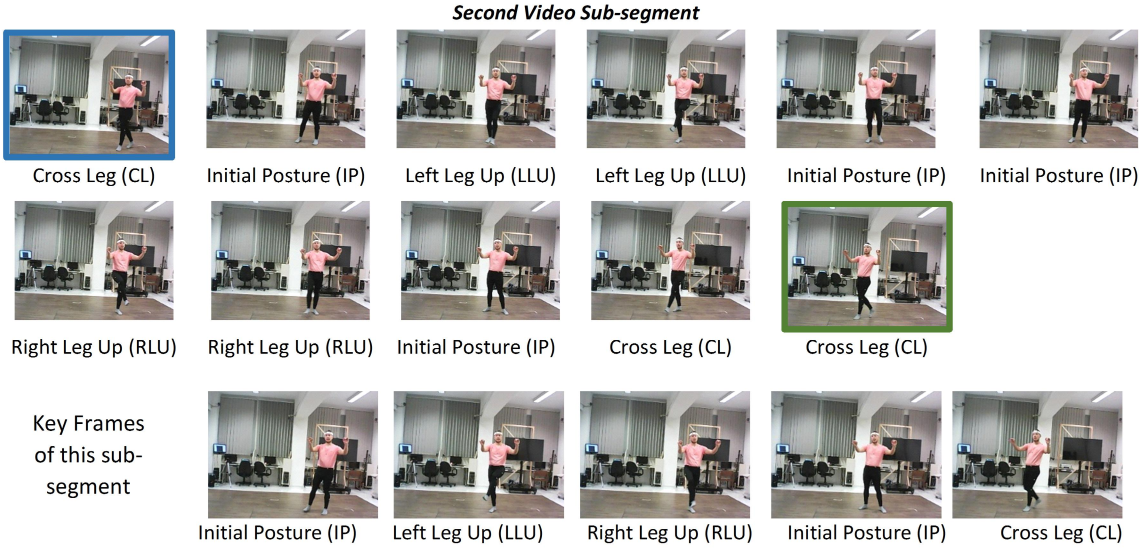


Figure 4.4 The key frames extracted for the second video sub-segment of the dance of Fig. 4.2. In this case, five key frames are extracted regarding the second layer of representation.

of the dance, followed the description of Table 3.5. As is observed, the depicted images refer to three repeated choreography patterns of Sirtos 3-beat dance. Each pattern is composed of six main steps namely: (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) (see Table 3.5). In this figure, we have also illustrated the extracted key frames at the first (initial-coarse) layer of the proposed hierarchical summarization algorithm. These key frames are shown with a color frame around each selected image. We have used different colors for the key frames to better distinguish the video sub-segments created on which decomposition takes place. In this particular example, six video sub-segments have been created. Figures 4.3 and 4.4 presents the representative key frames at the second layer of decomposition for the first two video sub-segments. As is observed, the second layer of representation better models the temporal choreography of the dance. Instead in the first decomposition layer some choreographic attributes are missing.

To better understand the way that our algorithm works regarding summarization of a dance, let us focus in the following on the specific choreographic patterns of the dance of Fig. 4.2 as being executed by Christos. These patterns can be represented in a symbolic form using the same notation

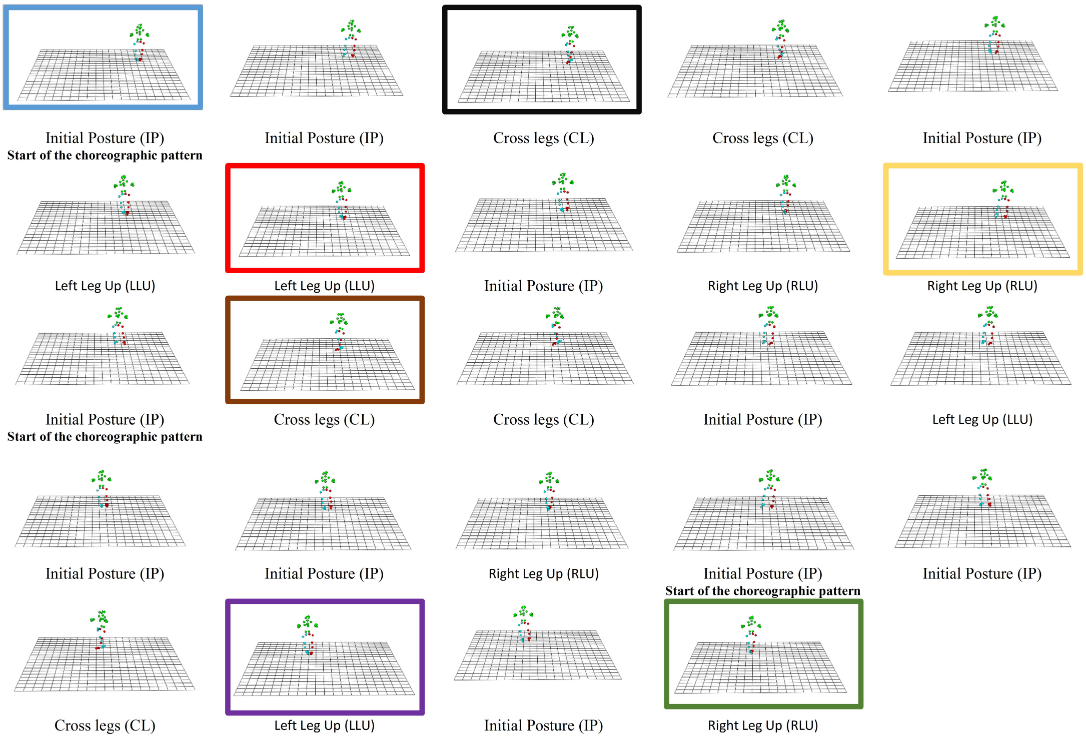


Figure 4.5 Skeleton data for Sirtos at 3-beat dance executed by dancer Ioanna. Images with a color frame correspond to the key frames extracted at the initial (coarse) layer using the proposed hierarchical algorithm. We have used different colors for the image frames to better distinguish the video sub-segments on which decomposition takes place.

as in Table 3.5: $\mathrm{IP}_1(1)$, CL(1), $\mathrm{IP}_2(1)$, LLU(1), $\mathrm{IP}_3(1)$, RLU(1) (first choreographic pattern); $\mathrm{IP}_1(2)$, CL(2), $\mathrm{IP}_2(2)$, LLU(2), $\mathrm{IP}_3(2)$, RLU(2) (second choreographic pattern); $\mathrm{IP}_1(3)$, CL(3), $\mathrm{IP}_2(3)$, LLU(3), $\mathrm{IP}_3(3)$, RLU(3) (third choreographic pattern). In this notation, numbering in parentheses refers to the index of the choreographic pattern (in our case 1-to-3, see also Fig. 4.2). Additionally, subscripts indicate the first, second and third IP position of the dancer within each choreographic pattern.

At the first layer, the extracted key frames model the following choreography: $\mathrm{IP}_1(1)$, CL(1), CL(2), $\mathrm{IP}_3(2)$, CL(3), LLU(3), and RLU(3). Therefore, some important choreographic elements are missing. On the contrary, at the second layer the extracted key frames model the following choreography: $\mathrm{IP}_1(1)$, CL(1), $\mathrm{IP}_2(1)$, LLU(1), RLU(1), (first choreographic pattern), $\mathrm{IP}_1(2)$, CL(2), $\mathrm{IP}_2(2)$, LLU(2), $\mathrm{IP}_3(2)$, RLU(2), RLU(2) (second choreographic pattern); $\mathrm{IP}_1(3)$, CL(3), $\mathrm{IP}_2(3)$, LLU(3), $\mathrm{IP}_3(3)$, $\mathrm{IP}_3(3)$, RLU(3) (third choreographic pattern). As is observed, almost all elements of the choreography have been identified (apart from the $\mathrm{IP}_3(1)$), increasing recall close to unity. However, few additional choreographic elements have been retrieved (e.g. twice the RLU(2) and RLU(3)), slightly decreasing the precision value. These statements are also verified in Table 4.1.

Fig. 4.5 presents the the choreographic patterns of the same dance (Sirtos at 3-beat) executed by a women dancer, Ioanna. In this particular example, instead of depicting the RGB images as we have done previously, we have illustrated the 3D point joints derived from the Vicon motion capturing system. This way, one can understand the geometry of the dance and how the extracted features are fluctuated in time and in 3D geometric space. In this figure, the annotation of the dance is also depicted followed

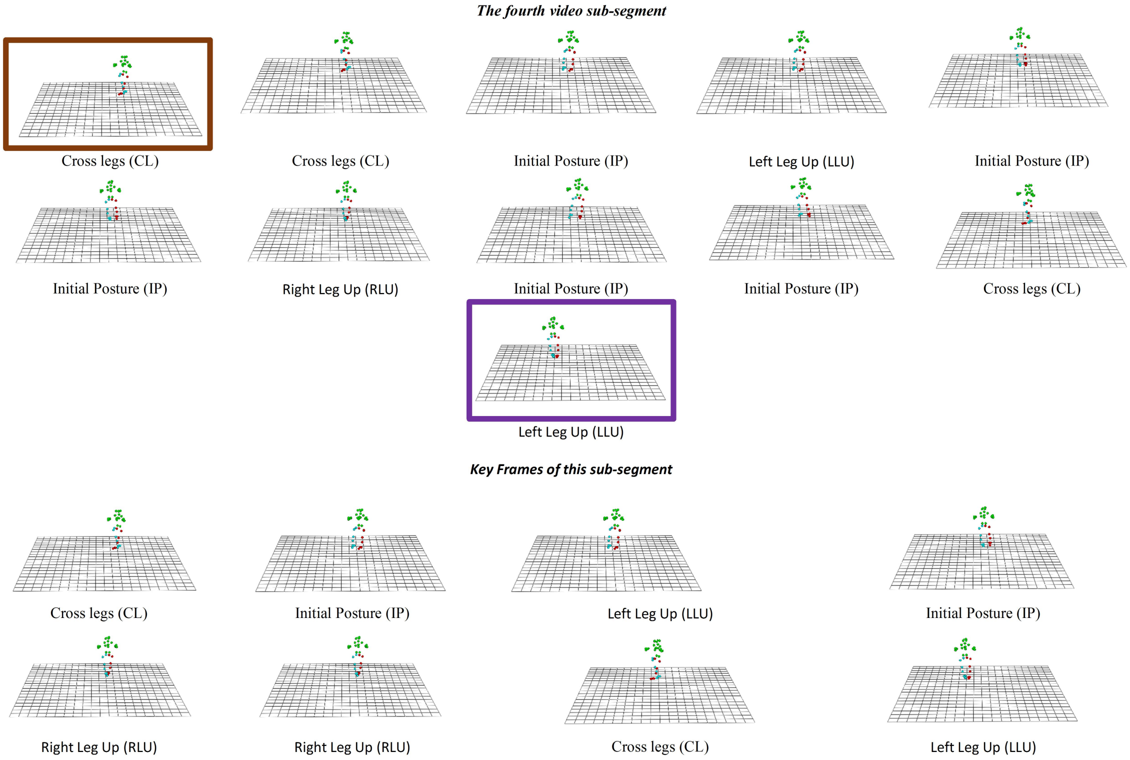


Figure 4.6 The key frames extracted for the fourth video sub-segment of the dance of Fig. 4.2. In this case, eight key frames are extracted regarding the second layer of representation.

the notation of Table 3.5. The extracted key frames at the first layer of hierarchy are marked with a color frame around them. As previously, we have assigned different colors to the key frames to better distinguish the video sub-segments over which the decomposition takes place. We can see that using only the key frames from the first layer of hierarchy, several choreographic patterns are missing. Instead, at the second decomposed level almost all dance elements are retrieved increasing a lot recall while at the same time keeping precision as high as possible. This is verified in Fig. 4.6 where the key frames of the fourth video sub-segment are depicted.

### 4.4.5 Comparisons with other methods

In Table 4.1, we present the precision, recall and F1-score values for all the recorded dances from the Vicon motion capturing system at Aristotle University of Thessaloniki under the activities of Terpsichore project [29]. The results have been obtained by averaging on the three dancers. In the same table, we also show comparisons against four other summarization methods proposed in the literature; the k-means clustering as proposed in [120], the conventional SMRS [57], a temporal-based video summarization scheme as presented in [159] and an hierarchical implementation of the k-means approach of [120].

We observe that the proposed hierarchical approach significantly increases the F1-score compared to all the other methods. This is due to the fact that recall values reach almost to unity while simultaneously keeping precision as high as possible. On the contrary, the approach in [120] seems to yield very high precision values but recall is significantly low meaning that many elements of the choreography cannot be retrieved. This is mainly due to the fact that temporal choreographic attributes are lost. We also

Table 4.1 Precision, recall values and F1 score for different summarization methods that use spatial, temporal or spatio-temporal attributes for the dance sequences recorded at University of Thessaloniki

| Type of Dance | Spatial-Driven Summarization | | Temporal-Drive Summarization | Spatio-Temporal Driven Summarization | |
|---|---|---|---|---|---|
| | k-means [120] | SMRS [57] | Temporal Variations of Feature Vector [159] | Hierarchical k-means | The proposed Method |
| | Average over three Dancers | | | | |
| Sirtos (3-Beat) | Pr=1.0<br>Re=0.22<br>F1=0.36 | Pr=0.88<br>Re=0.39<br>F1=0.54 | Pr=0.67<br>Re=0.44<br>F1=0.53 | Pr=0.36<br>Re=0.89<br>F1=0.52 | Pr=0.89<br>Re=0.94<br>F1=0.92 |
| Sirtos (5-Beat) | Pr=1.0<br>Re=0.21<br>F1=0.35 | Pr=0.63<br>Re=0.36<br>F1=0.45 | Pr=0.48<br>Re=0.71<br>F1=0.57 | Pr=0.30<br>Re=0.86<br>F1=0.44 | Pr=0.52<br>Re=0.93<br>F1=0.67 |
| Kalamatianos | Pr=0.63<br>Re=0.88<br>F1=0.73 | Pr=0.60<br>Re=0.88<br>F1=0.71 | Pr=0.63<br>Re=0.75<br>F1=0.68 | Pr=0.47<br>Re=1<br>F1=0.64 | Pr=0.65<br>Re=0.88<br>F1=0.75 |
| Trehatos | Pr=0.33<br>Re=0.17<br>F1=0.22 | Pr=0.67<br>Re=0.67<br>F1=0.67 | Pr=0.5<br>Re=0.28<br>F1=0.36 | Pr=0.30<br>Re=0.56<br>F1=0.39 | Pr=0.74<br>Re=0.89<br>F1=0.81 |
| Enteka | Pr=0.8<br>Re=0.16<br>F1=0.27 | Pr=0.89<br>Re=0.32<br>F1=0.47 | Pr=0.75<br>Re=0.36<br>F1=0.49 | Pr=0.80<br>Re=0.16<br>F1=0.27 | Pr=0.87<br>Re=0.84<br>F1=0.85 |

notice that for more complicated dances like Trehatos or Kalamatianos, the F1-score takes the lowest values compared to the simplest dances. This is quite justified since in these dances 3D geometry is very complicated in space and time, deteriorating summarization performance.

In Table 4.2, we illustrate the average error modelling of the thirty dance sequences captured at university of Thessaloniki, as obtained using the key frames of the first and of the second layer of hierarchy (see Section 4.3.5). This error is expressed as the norm of the features of the frames of the video sequence with the best assigned key frame. The results have been expressed in db while they have been averaged over all the three dancers. We notice that at the second layer of hierarchy a decrease in the error values is encountered indicating that the proposed hierarchical summarization better models the choreography. We also notice that for more complicated dances such as Trehatos and kalamatianos the error is higher due to the complexity of these dances.

In the following, we present the results for the Carnegie-Mellon university dataset. In particular, Fig. 4.7 illustrates the skeleton data of a theatrical kinesiology so as to provide an overview of this sequence. In this figure, we have depicted the results of the key frame extraction process at the first layer of hierarchy along with the respective annotation of the content. As is observed, the key frames of the first hierarchy provide a reliable representation of the choreography, though some elements are missing. These missing elements are detected on the second decomposition layer. Table 4.3 shows the average precision, recall and F1-score values for the three examined datasets of the Carnegie-Mellon university. In this table, we have also compared the results against the approach in [120], the conventional SMRS [57],

Table 4.2 Error modelling (expressed in db) for the thirty dance video sequences obtained using the key frames of the first and of the second layer of processing

| Type of Dance | One Level of Hierarchy(in db) | Two Levels of Hierarchy(in db) |
|---|---|---|
| Sirtos (3-Beat) | 28.72 | 25.17 |
| Sirtos (5-Beat) | 30.51 | 26.47 |
| kalamatianos | 31.58 | 27.45 |
| Trehatos | 33.42 | 30.81 |
| Enteka | 31.00 | 9.831 |

Table 4.3 Precision, recall values and F1 score for different summarization methods that use spatial, temporal or spatio-temporal attributes for the dance sequences of Carnegie Mellon University.

| Type of Dance | Spatial-Driven Summarization | | Time-Drive Summarization | Spatial-Time Driven Summarization | |
|---|---|---|---|---|---|
| | k-means [120] | SMRS [57] | Temporal Variations of Feature Vector [159] | Hierarchical k-means | The proposed Method |
| Theatrical Kinesiology | Pr=0.6<br>Re=0.22<br>F1=0.33 | Pr=1.0<br>Re=0.37<br>F1=0.54 | Pr=0.8<br>Re=0.51<br>F1=0.62 | Pr=0.73<br>Re=0.60<br>F1=0.66 | Pr=0.87<br>Re=0.94<br>F1=0.90 |
| Dance Pirouette 1 | Pr=0.78<br>Re=0.84<br>F1=0.81 | Pr=1.0<br>Re=0.84<br>F1=0.91 | Pr=0.75<br>Re=0.53<br>F1=0.62 | Pr=0.53<br>Re=0.53<br>F1=0.53 | Pr=0.83<br>Re=1<br>F1=0.90 |
| Dance Pirouette 2 | Pr=0.77<br>Re=0.77<br>F1=0.77 | Pr=1.0<br>Re=0.30<br>F1=0.47 | Pr=0.70<br>Re=0.61<br>F1=0.65 | Pr=0.23<br>Re=0.69<br>F1=0.34 | Pr=0.80<br>Re=0.69<br>F1=0.74 |

a temporal-based video summarization scheme as presented in [159] and an hierarchical implementation of the approach of the [120]. Again, the proposed hierarchical algorithm yields the highest F1-score meaning that it can effectively model the choreographic patterns of the datasets.

## 4.5 Discussion

Automatic extraction of a dance main choreographic patterns and steps are very important in performing arts since they can elicit the main structural components of a dance, identifying its style, assisting trainees towards a proper learning, and improving dance experts in their work for documenting the dance and relating it with the intangible components of the culture of a place. These choreographic patterns are in fact the semantics of a dance. Only the use of RGB color space for describing a dance choreography cannot properly represent the complicated 3D geometry of a dance. Inevitably, other capturing devices have been adopted such as Kinect or the Vicon motion system to capture the 3D points of the human joints forming human skeletons. In this paper, eight choreographies have been exploited. The first five refer to traditional Greek folkloric dances recorded in the premises of Aristotle University of Thessaloniki

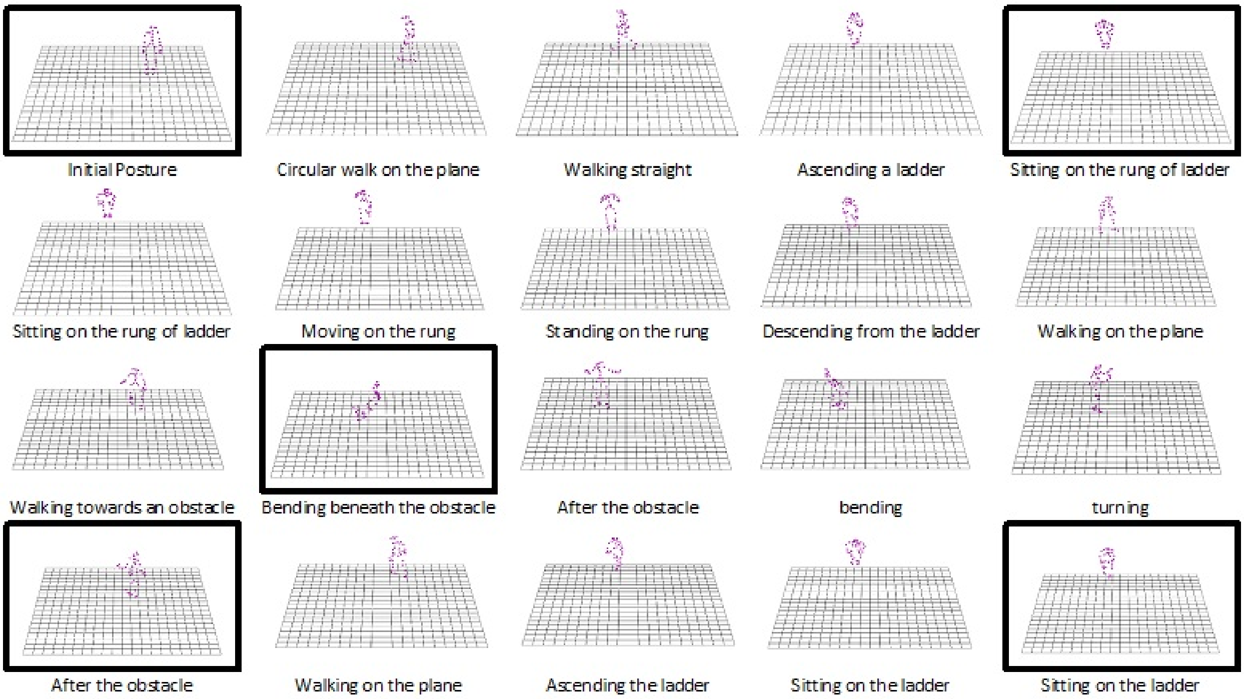


Figure 4.7 Skeleton data of a theatrical kinesiology of the Carnegie-Mellon University dataset, along with the estimated key frames at the first layer of hierarchy.

under the aegis of Terpsichore project. The other three are freely available 3D joints datasets from Carnegie-Mellon University referring to choro-theater performances.

Traditional video summarization algorithms cannot retrieve the choreographic patterns of a dance. This is due to the fact that these algorithms either spatially cluster together image content [120] or exploit the temporal fluctuation of the trajectory of the features [159]. Thus, the first category ignores the ordered sequence of the content while the second is trapped on micro-variations of the choreography. For this reason, in this paper a spatio-temporal decomposition of dance sequences is proposed based on a modification of the SMRS algorithm implemented under a hierarchical decomposition framework. The scheme initially extract global holistic descriptors that give a coarse representation of the choreography. Then, each video sub-segment is further decomposed to get a finer representation of the content. This hierarchical video dance decomposition results in pyramid of key frames that provide a complete overview of the choreography.

Experimental results and comparisons with other traditional summarization approaches indicate that our proposed hierarchical scheme reaches very high recall values close to one, meaning that almost all the main choreographic patterns are detected. Similarly, precision is kept as high as possible minimizing the number of noisy key frames being extracted. The results remain robust even for complicated dances and theatrical kinesiology performances.

# Chapter 5

# Physics-based key-frame selection for human motion summarization

## 5.1 Introduction

Due to the significance of Intangible Cultural Heritage (ICH) and its preservation, many international organizations (such as UNESCO) have focused on promoting research to encode, store, analyze and disseminate related content. Choreographic and kinesiology content holds a substantial position within ICH. The cultural significance of performing arts, and especially dance, along with the interdisciplinary interest in the study of ICH makes dance analysis a focal point of research. Several technological achievements, including pervasive video capturing devices and software, increased camera and display resolutions, cloud storage solutions, and motion capture technologies have generated advancements in capturing, documenting and storing ICH content with more efficiency and accuracy. In order to utilize the full potential of multimodal (text, image, video, 3D, MoCap) ICH data that are becoming increasingly available, the adaptation of the state-of-the-art technologies is needed to build new ones in the fields of the artificial intelligence (AI), computer vision, and image processing and connect the aforementioned scientific fields with the ICH. The adaptation of recent advancements for ICH content and specifically performing arts provides the opportunity for effective and efficient organization and management, fast indexing, retrieval and browsing but also automatic recognition and classification.

### 5.1.1 Related Work

The field of motion analysis has attracted the interest of several researchers. Part of the existing work focuses on the subfield of dance motion analysis, studying different facets of choreographic digitization and performing arts analysis. The US National Science Foundation has supported a program for developing a human/computer environment for tele-immersive dance [162], aiming at designing a creativity framework for choreography based on Laban Movement Analysis (LMA). In [33], a 3D archive system for Japanese dances is proposed based on multi-view videos and the graph-cut algorithm. In [163], motion analysis algorithms are investigated with a view to transforming captured motion trajectories of dancers into meaningful and semantically enriched LMA features. The skeletonization capability offered by motion capturing depth sensors (e.g. Microsoft Kinect and Asus Xtion), allows for methods that emphasize the geometrical and topological properties of motion trajectories. In [6] a dance motion analysis and synthesis framework is presented. In [68] a methodology for dance learning and evaluation using multi-sensors is proposed, where the robustness of skeletal tracking is improved through a fusion algorithm that splits the skeletal data into different body parts so as to allow view-invariant posture recognition. In [164], a technique for unexpected impacts into a motion capture driven animation system through the combination of a physical simulation is introduced, which responds to contact forces and a specialized routines determining the best plausible re-entry into motion library playback following the impact.

Regarding pose features, in [165], automated methods for extracting logically related motions from a data set are introduced converting them into an intuitively parameterized space of motions, whereas a pose distance metric on 3D motion data is proposed in [166]. In [167], a search algorithm for use with sampled motion data is proposed; additionally, a representation for motion data using a meaningful distance metric for poses is introduced. In [168], an approach to performance animation that employs video cameras and a small set of retro-reflective markers is introduced, in order to create a low-cost, easy-to-use system to learn pose distance metrics. In [169], efficient approaches for local and global motion matching, which are applicable even to very large databases, are presented. However, pose feature estimation, although closely related, cannot provide per se a summarization of choreographic sequences. In [8] an attempt for extracting representative frames from dance motion is made but the method does not take into consideration temporal interdependencies, whereas in [170] a comparison of classifiers for pose identification is performed.

In [171], automated methods for efficient indexing and content-based retrieval of motion capture data is presented. In [172], new methods for automatic classification and retrieval of motion capture data are presented facilitating the identification of logically related motions, whereas in [173], [174] methods for motion pattern classification based on hidden Markov models are proposed. In [175], a flexible, efficient method for searching arbitrarily complex motions in large motion databases is proposed. Again, the aforementioned approaches mainly focus on dynamic features and motion content-based retrieval/indexing, and therefore not on the extraction of key patterns of a choreographic sequence.

In this context, there are works in the literature exploiting motion features for content-based indexing and retrieval. In particular, in [176], a content-based 3D motion retrieval algorithm is proposed. In [177], an automatic segmentation of human motion data based on statistical properties of the motion is proposed as an efficient and robust alternative to hand segmentation. Additionally, in [178], an efficient method for fully automatic temporal segmentation of human motion sequences and similar time series is introduced, whereas in [179], a method based on dynamic time alignment of Gaussian mixture model clusters for matching actions in an unsupervised temporal segmentation is presented. In [180], a technique for multidimensional trajectory similarity estimation is proposed. Again, the main focus of the aforementioned works is on motion indexing, retrieval and segmentation, instead of dance summarization. The problem of learning motion primitives is addressed in [181] as a temporal clustering problem. In particular, an unsupervised hierarchical bottom-up framework called hierarchical aligned cluster analysis is presented. However, such an approach is mainly based on the statistical similarities of extracted features through the application of hierarchical clustering. Instead, in this paper, we consider kinematics-based variations of the choreographic trajectories, which take into account the physical attributes of the motion, e.g., by extracting key elements at key change points of joint velocity, acceleration, etc.

Recently, in the context of dance analysis a series of works have been presented regarding emotion and style modeling, as well as organization and annotation of motion data collections. In [51], an investigation through similarity between various emotional states with regard to the arousal and valence of the Russel's circumplex model is introduced, whereas a motion stylization technique for expressive mocap data, such as contemporary dances, is proposed in [182]. In terms of organization of large motion data collections, in [183], a system for automatically and efficiently annotating large unstructured collections of mocap data is proposed. In [184], a MotionExplorer was developed as an exploratory search system for large data collections of motion capture data. Finally, in [185], a scalable method for organizing the collection of motion capture data for overview and exploration is introduced.

### 5.1.2 Innovation and Originality

In this context, content summarization is an important and very useful application domain in the multimedia research community. As regards choreographic sequences, the automatic extraction of the choreographic elements is of significant interest, since such elements provide an abstract and compact representation of the semantic information encoded in the overall dance storyline.

In the literature, there is a large number of video summarization techniques focusing on extracting representative keyframes to summarize video segments depicting human activity, e.g. [186], [187], [188]. However, video synopsis focusing only on RGB information and temporal feature fluctuations makes the derived summaries highly sensitive to noise and to micro-variations of dancer steps, which makes them difficult to apply effectively to dance sequences. This can often lead to an over-representational modelling of the content, namely to a very large number of keyframes.

This chapter's contribution lies in the proposal of a framework for the selection of keyframes in 3D human motion data from real-world folklore choreographic sequences. Folklore dance summarization is a challenging problem mainly due to the application domain's particularities. This work is part of the European initiative "Terpsichore" [1], which aims to implement an innovative framework for the affordable digitization, modeling, analysis, archiving, e-preservation and presentation of ICH content related to European folk dances. The impact of an effective method for dance summarization is underscore when one conjunctly considers: the massive amounts of RGB-D and 3D skeleton data produced by video and motion capture devices; the huge number of different types of existing dances and variations thereof; the need for organizing, indexing, archiving, retrieving and analyzing dance-related cultural content in a tractable fashion and with lower computational and storage resource requirements, as well as the need for flexible and accessible tools for ICH dissemination and education.

In this chapter, we present two basic approaches which rest on different assumptions. The first approach is "time-independent" and simply extracts the key postures of which a given type of dance is composed, regardless of time and order of appearance in the dance. This clustering-based approach determines the salient "primitives" of a dance and helps reveal the basic characteristics of the nature and physiognomy of the dance. The second "time-involving" approach extracts representative summarized

[1] www.terpsichore-project.eu

sequences from long choreographic frame series through the calculation of the local extrema of kinematics-based feature trajectories. Such summaries can be used in more thorough analyses of dances in a variety of contexts (e.g. cultural, technical, academic, choreographic, spatial, commercial, educational). As a rough analogy to the film domain, the first approach would result in acquiring movie production stills, i.e. photographs that depict characteristic instances of the film, whereas the second approach would generate a short video with a plot summary, in other words a brief storytelling trailer. Both approaches are based on skeleton data and their dynamics and kinematics, but can be otherwise employed independently from each other according to the scope, goal and context at hand.

## 5.2 Physics-based Choreographic Movement Representation

The core of the acquisition framework designed for modeling the dancer motion trajectories in 3D space is Vicon, a motion capturing system used in several application domains, ranging from gaming, film production, clinical research and entertainment. In our implementation, ten Bonita B3 cameras are included, running the Nexus1.8.5.61009h software. The movement area is a 6.75 sq.m. More details about the Vicon system used is presented in Section 3.2.1.

In the following, let us denote as $\mathbf{J}_k^G = (x_i^G, y_i^G, z_i^G)$ the $k$-th joint out of the $N$=35 extracted by the Vicon architecture. Variables $x_i^G$,$y_i^G$ and $z_i^G$ indicate the coordinates of the respective joint with respect to the Vicon's reference coordination system. We assume that these joints have been obtained after density based filtering of the detected joints to remove possible noise from the acquisition process.

Our approach is defined as the study of the properties of the kinematics of the performers, consequently it is necessary to determine the basic modeling concepts and vectors to be used in the sequel. Considering the motion of each $\mathbf{J}_k^G$, $k$=1,2,...$N$ skeleton joint as the motion of a particle, and based on rigid body physically based modeling theory [189] we let function

$$\mathbf{s}_k(t) = \mathbf{J}_k^L(t) \tag{5.1}$$

$\mathbf{s}_k(t) = \mathbf{J}_k^L(t)$ denote the particle's location in local space at any time instance (frame). The function gives the velocity of the particle at time.

$$\mathbf{v}_k(t) = \mathbf{s}_k(t)' = \frac{d}{dt}\mathbf{s}_{k(t)} \tag{5.2}$$

Then,

$$\mathbf{a}_k(t) = \mathbf{v}_k(t)' = \frac{d}{dt}\mathbf{v}_k(t) = \frac{F_k(t)}{m_k} \tag{5.3}$$

provides the acceleration of the particle at time $t$, $F_k(t)$ is the is the force acting on the particle and is the particle's mass. For a system (dancer skeleton) with $N$ particles, we define the skeleton's location $\mathbf{S}(t)$, velocity $\mathbf{V}$(t) and acceleration $\mathbf{A}$(t) vectors respectively as:

$$\mathbf{S}(t) = \left[s_1(t) \cdots s_N(t)\right]^T \tag{5.4}$$

$$\mathbf{V}(t) = \frac{d}{dt}\mathbf{S}(t) = \left[v_1(t)\cdots v_N(t)\right]^T \tag{5.5}$$

$$\mathbf{A}(t) = \frac{d}{dt}\mathbf{V}(t) = \left[a_1(t)\cdots a_N(t)\right]^T \tag{5.6}$$

In the following we will use the above defined physically based vectors for the two types of summarization approaches.

## 5.3 Clustering-based time-independent representative selection

The main objective is to extract the most representative instances of the dance, its key postures, or, differently put, its basic primitives, regardless of their order in the sequence. We define our approach as an unsupervised clustering problem. Initially, we should mention that, depending on the type of dance to be summarized, it is possible that only a subset of the $N$ joints forming the location vector **S**(t) in Eq.(5.1) has to be considered as significant for the particular dance's moves. For example, in a choreography sequence where only the motion of the lower limbs is important (there are several such cases in Greek traditional folk dances, for instance), we will only take into consideration the respective joints in forming the feature location vector and consequently the velocity and acceleration vectors to follow. Let $q$=1,2,...,$M$ denote the joints which are significant for a particular dance ($M \leq N$). Then, the feature vector that represents the dancer motion at time t is given by:

$$\mathbf{f}(t) = \left[f_1(t)\cdots f_M(t)\right]^T = \left[s_1(t)\cdots s_M(t)\right]^T \tag{5.7}$$

Let us now consider a set $S$ containing similar dancing postures. Since a dance posture is represented by the feature vector $f(t_i)$, two dancing trajectories at two different captured time instances $\left[t_i,t_j\right]$, will belong to the same set $S$ only if:

$$\begin{aligned} [t_i,t_j] \in S \quad \text{if} \quad & D(f(t_i),f(t_j)) \leq D(f(t_i),f(t_j)) \\ & \forall[t_i,t_j] \in S \qquad \forall[t_i,t_j \notin S]. \end{aligned} \tag{5.8}$$

where $D(\cdot)$ is a distance metric. In our case the Euclidean distance is used. Eq. 5.8 states that two points $t_i$, $t_j$ on the dancer motion trajectory will belong to the same cluster only if the distance of the respective features vectors $f_i(t)$ and $f_j(t)$ is smaller than in the case of the two points belonging to a different cluster. Grouping the points of the dancing moment into $L$ different clusters, we identify the most salient "primitives", i.e. time instances $t_i$, $i = 1,2,...,L$ that best represent the choreography.

The aforementioned problem is actually an unsupervised clustering problem and can be approached through various clustering algorithms. Here, we have adopted the k-means++ algorithm [190], as it specifies a procedure to initialize the cluster centers before proceeding with the standard k-means optimization iterations, thus improving the clustering performance. Let us now consider as $f_c^i$ the centroid of the $i$-th formed cluster, that is:

$$f_c^i = \frac{\sum_{k \in S_i} f(t_k)}{\parallel S_i \parallel} \tag{5.9}$$

where the operator $\parallel \cdot \parallel$ denotes the cardinality of the generated set/cluster $S_i$ by the unsupervised clustering algorithm. Then, we select as the most representative instance $t_{si}$ among the samples of cluster the one that has a feature vector of minimum distance from the cluster centroid. Therefore, we have that:

$$t_{si} = \arg\min D(f(t_i), f_c^i), \forall t_i \in S_i \tag{5.10}$$

From the above, it becomes clear that we now have $L$ clusters of frames. For each of these groups, a representative frame (or time instance) is selected, say $t_{si}$. Thus, we have eventually extracted $L$ different frames (or time instances), i.e $\mathscr{C}$ = $[t_{s1}, t_{s2}, ..., t_{sL}]$, which represent suitably the dance sequence. In other words, $\mathscr{C}$ includes the basic primitives of the sequence, or, using the film analogy mentioned earlier, the characteristic “movie production stills” of the dance.

## 5.4 Summarization of motion capture sequences based on skeleton kinematics

We hereby focus on creating a summarized sequence of a dance, i.e. on briefly “telling its story”. Since a feature vector is assigned for each frame of a dance frame sequence, the vectors of all frames form a trajectory in a high dimensional feature space, which expresses their temporal variation. Thus, selecting the most representative frames within a sequence is equivalent to selecting appropriate curve points, able to represent the corresponding trajectory. Ideally, the selected curve points (summary) should provide sufficient information about the trajectory, so that it can be reproduced using some kind of interpolation. This can be achieved by extracting the time instances, i.e., the frame numbers, which reside in extreme locations of this trajectory. In our case, the magnitude of the second derivative of feature vectors for all frames within a sequence with respect to time is used as a curvature measure. The second derivative expresses the degree of acceleration or deceleration of an object that traces out the feature trajectory. Local maxima correspond to time instances of peak variation of the velocity, i.e., large acceleration or deceleration.

Eq.(5.5) and Eq.(5.6) define the velocity and acceleration vectors of the dancer skeleton. Since, however, time t is discrete (frame numbers), the first derivative (velocity) for joint q=1,2,...M is approximated as the difference of feature vectors between two successive frames:

$$\mathbf{v}_q(t) = \mathbf{f}_q^{'}(t) = \mathbf{f}_q(t+1) - \mathbf{f}_q(t) \tag{5.11}$$

Similarly, the 2nd derivative (acceleration) is approximated as:

$$\mathbf{a}_q(t) = \mathbf{v}_q^{'}(t) = \mathbf{v}_q(t+1) - \mathbf{v}_q(t) \tag{5.12}$$

Further, we define the measure to be investigated for local maxima as:

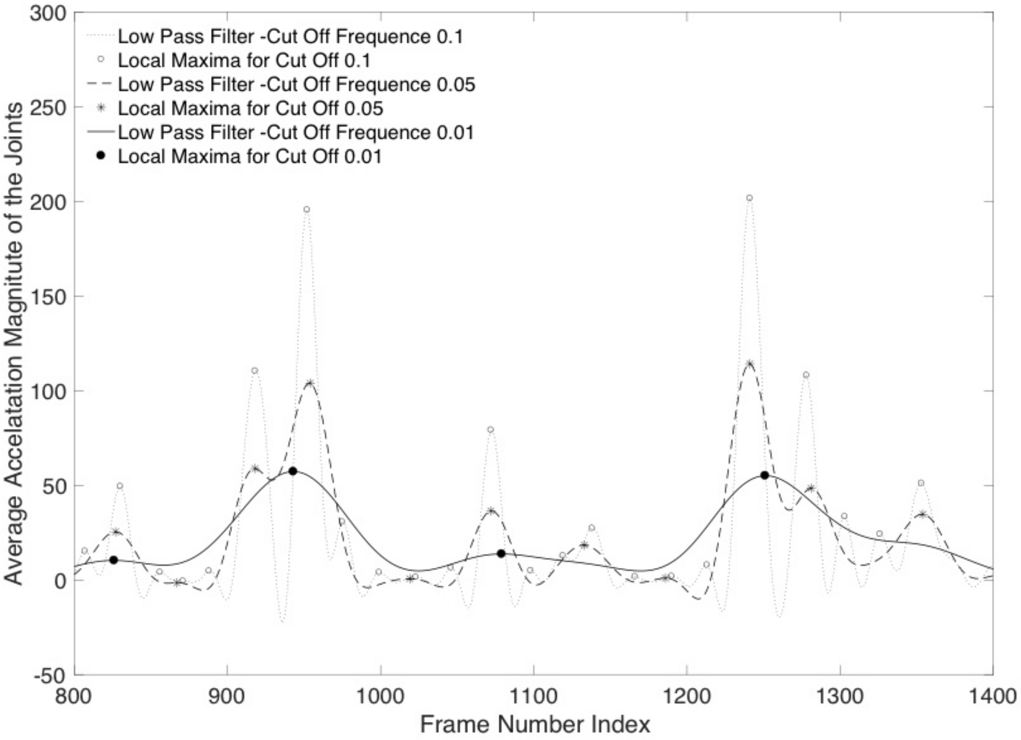


Figure 5.1 Average acceleration magnitude over time (frame number) for LPF cut-off frequency values fLPF=0.01, 0.05, and 0.1. As the cut-off frequency increases, the number of local maxima also increases, leading to longer, more informative summaries. .

$$\Gamma(t) = \sum_{q} | \mathbf{a}_q(t)| \tag{5.13}$$

The local maxima of $\Gamma(t)$ are considered as appropriate curve points, i.e. as the representative key frames that will constitute the summarized sequence of the dance. To obtain a smoother curve and remove noise, a LPF can be applied to $\Gamma(t)$.The appropriate value of the LPF's cut-off frequency varies according to the dynamics of each type of dance; e.g., a dance with abrupt and large variations in velocity of the body parts is bound to require a higher cut-off frequency, since these changes are represented in spectral content of higher frequencies in the frequency domain. Fig. 5.1 shows the diagram of the average acceleration measure $\Gamma(t)$. over time for different values of LPF cut-off frequency. As we can see, as the cut-off frequency increases, the number of local maxima (and therefore of salient frames selected) also increases, thus leading to longer, more inclusive summaries.

## 5.5 Performance Evaluation

### 5.5.1 Evaluation of the clustering-based time-independent keyframe selection-method

The acquisition of image and skeleton data is followed by a pre-processing step which removes noisy joint detections via a modified DBSCAN algorithm [191]. The filtered joints are transformed to a local coordinate system, and features are extracted representing the motion properties of the dance at each frame, i.e. time instance. In this particular choreography and without loss of generality, only the legs' joints are considered. This is due to the fact that in the examined Greek folk dances the arms tend to remain still with respect to the dancer's hip joint and are thus not an important attribute of this dance.

The optimal number of clusters for each instance of dance depends on the specific moves and postures and their variability and cannot be known with certainty in advance, although estimations can be made, especially by domain experts. Therefore, we have experimented with different numbers of clusters using the k-means++ method described in Section 4 and have evaluated the acquired results in terms of the clusters' consistency to deduce the optimal number of clusters in each case. For this, we have used two indices: Silhouette [192] and Davies-Bouldin index (DBI) [193]. The Silhouette index measures how similar an object (in our case, frame, i.e. posture) is to its own cluster (cohesion) compared to other clusters (separation). Here we calculate the average Silhouette value over all frames. The DBI is another measure of how well clustering has been done and is based on inherent features of the dataset. Good clustering solutions are denoted by high Silhouette values and low DBI values.

Fig. 5.2 shows the Silhouette and DBI values for Dancers F, C, and I for two different dances: Syrtos and Kalamatianos. Regarding Syrtos dance, we can see that both the Silhouette index and the Davies-Bouldin index indicate that the optimal number of clusters appears to be 4. This result is in agreement with the estimation of dance experts that the key postures that dominate the particular dance sequence are indeed four: (1) Left and right feet are opposite; (2) Left and right feet are crossed; (3) Left foot is raised; (4) Right foot is raised. Naturally, such results are dance-specific and heavily depend on the complexity and variance of each dance. As can be seen in the bottom of Fig. 5.2, the Silhouette and Davies-Bouldin indices for Kalamatianos dance indicate that a higher number of clusters, eight (8) in particular, is optimal in grouping the input data, which corresponds to a respective number of key postures in the dance.

### 5.5.2 Evaluation of the summarization of motion capture sequences method

For the evaluation of the kinematics-based summarization method, we have experimented with different dance sequences, also applying different cutoff frequencies to the employed LPFs. As Fig. 5.1 shows, the higher the cutoff frequency, the greater the number of key frames extracted and the longer the derived summary. It is important to note however, that the local maxima obtained for low cutoff frequencies, i.e. the most salient, important frames, also appear (in some cases with minimal deviations) as local maxima at higher cutoff frequencies. This allows deriving valid summaries of increasing lengths, by essentially following a hierarchical decomposition type of paradigm (see Fig. 7).

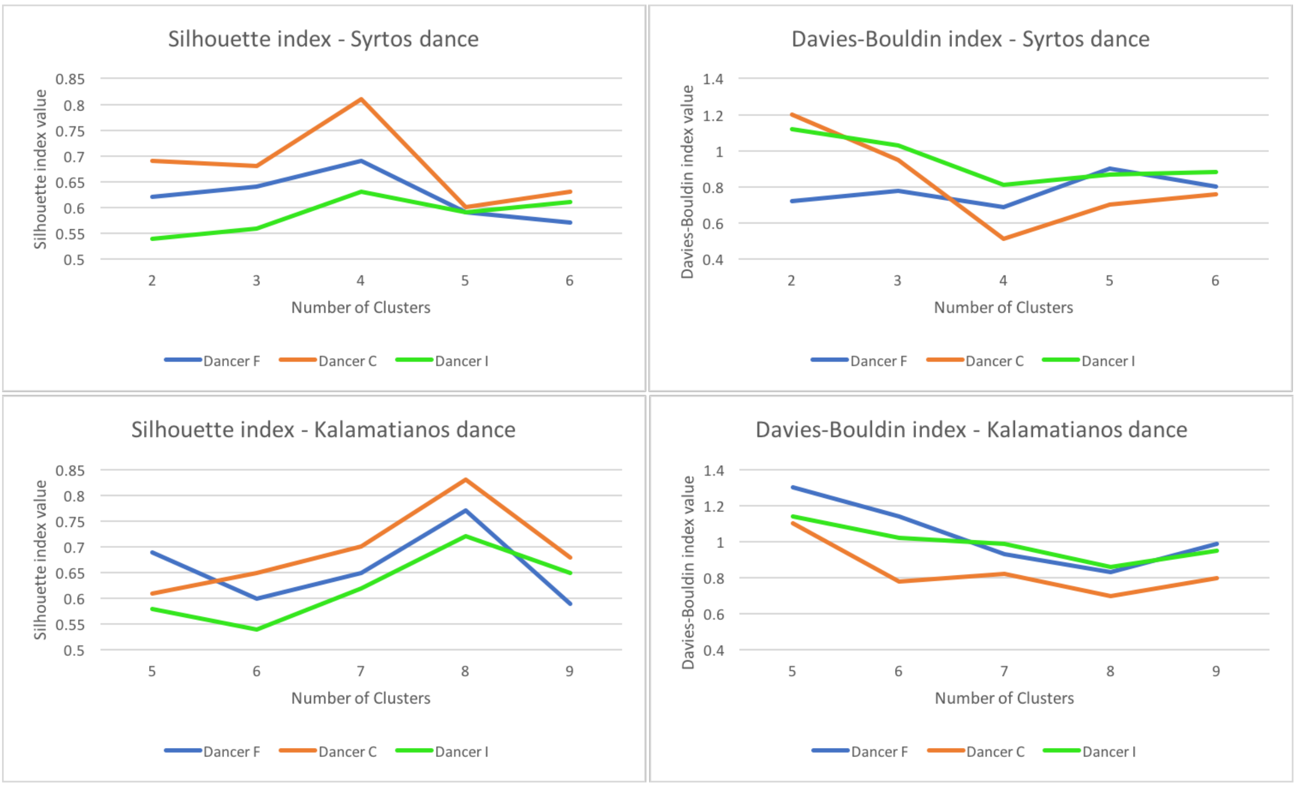


Figure 5.2 Evaluation of the clustering-based approach for different numbers of clusters based on the Silhouette and Davies-Bouldin indices. Top: Syrtos dance. Bottom: Kalamatianos dance.

Moreover, to objectively evaluate the summarization results, we use a ground truth dataset which consists of a set of characteristic key frames assessed by domain expert users. These target-key frames are compared against the ones derived by the proposed framework. Then, we calculate, as objective criteria, the precision and recall values. Since the estimated key frames differ in time w.r.t. the target ones, in this paper precision and recall are calculated as follows: Each estimated key frame by the summarization scheme is compared against the targets and the closest in time target is selected as the most suitable. Then, we evaluate the absolute error of the dance figure between the estimated key frame and the closest target frame for all joints. This indicates how close (in 3D space) the dance silhouette depicted in the estimated key frame is to the target frame. If this error is smaller than a threshold, the respective keyframe is considered as relevant. Otherwise it is considered as irrelevant. So precision is estimated as the ratio of relevant retrieved keyframes over the total number of retrievals, while recall over the total number of ground truth data. In our case, the threshold is adaptively estimated taking into account as reference point the mean square error (MSE) of all estimated key frames with respect to the target ones. More specifically, the threshold is set to be at 10% of the MSE values so that only truly relevant key frames in terms of the position of the dancer's silhouette are retained. Finally, F1-score is calculated as the harmonic mean of precision (Pr) and recall (Re), i.e. F1 = (2*Pr*Re)/(Pr+Re).

Table 5.1 presents the average precision and recall values for different frequency cut off values. We should state that higher frequency cut off values correspond to an extraction of a higher number of key frames. As is observed, recall slightly increases as the frequency cut off values increases, that is, for a higher number of estimated key frames since it is more probable for some of these key frames to be among the target ones. However, this increase is saturated, meaning that beyond a certain limit the

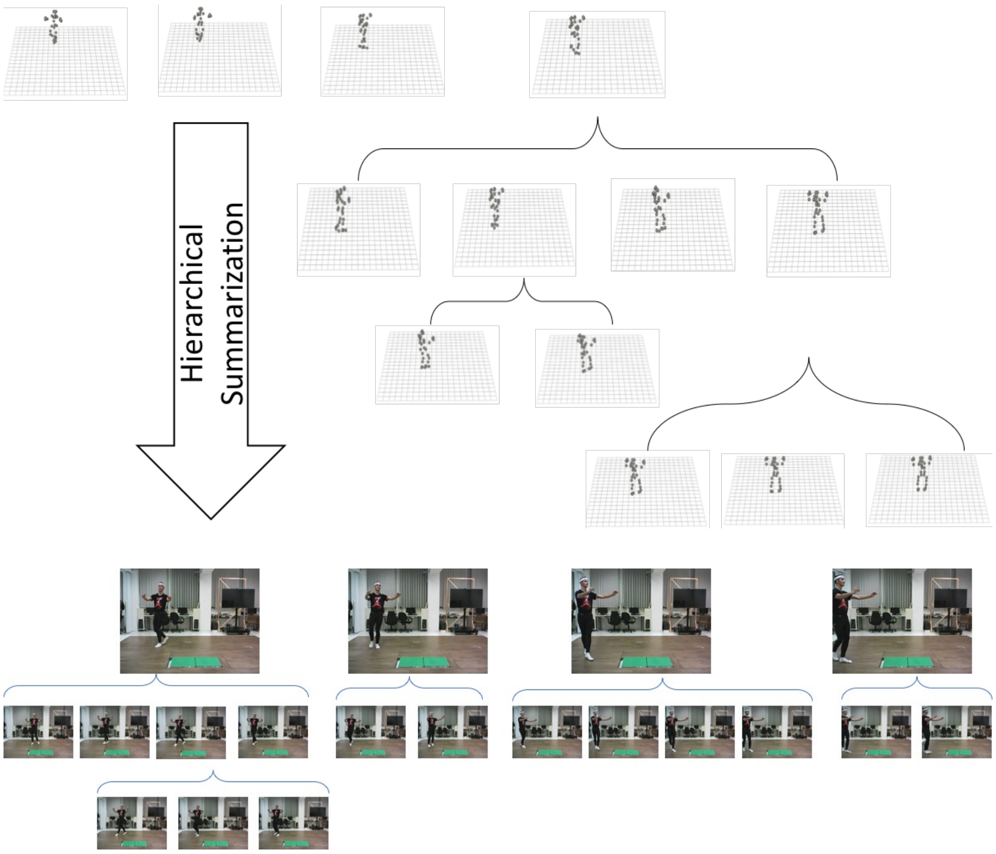


Figure 5.3 Example results of the kinematics-based summarization approach (top: 3D skeleton data – bottom: image data). Each line is part of a summary of different level of detail, acquired by applying different cutoff frequency values to the LPF, thus obtaining different numbers of local maxima. A frame can be hierarchically further decomposed to a larger number of frames, thus creating a longer, more informative summary.

improvement is actually negligible. On the other hand, precision increases up to a frequency cut off value of about 0.05 and then it decreases again since extraction of a higher number of key frames also increases the noise. In the sequel, we have compared the performance of the proposed kinematics-based method with two techniques in the literature: (i) a hierarchical k-means summarization technique [194], and (ii) SMRS technique [57]. Table 5.2 presents those results in comparison to the precision, recall and F1-score [195] of our proposed method (for the cutoff frequency yielding the best results) for three types of dances of the Terpsichore dataset: Syrtos, Kalamatianos and Trehatos, as well as for four sequences recorded from Carnegie Mellon University [2]. We observe that the proposed method yields a higher F1-score [195] in almost all cases, remaining satisfactorily robust even in the cases of complex choreographic sequences.

## 5.6 Discussion

Summarization of choreographic sequences is a little explored research topic, albeit a very challenging and potentially impactful one in the context of ICH intelligent e-preservation and promotion. Leveraging the capabilities of state-of-the-art motion capture technologies as well as physically based modeling

[2] publicly available at http://mocap.cs.cmu.edu

Table 5.1 Summarization performance metrics for different summarization methods (Terpsichore and CMU datasets). Precision (Pr) is estimated as the ratio of relevant retrieved keyframes over the total number of retrievals, while Recall (Re) over the total number of ground truth data. F1-score is the harmonic means of Precision and Recall.

| Sequences | Proposed Method | Hierarchical k-means [194] | SMRS [186] |
|---|---|---|---|
| Kalamatianos | Pr=0.63<br>Re=0.75<br>F1=0.68 | Pr=0.47<br>Re=0.92<br>F1=0.62 | Pr=0.59<br>Re=0.79<br>F1=0.67 |
| Syrtos | Pr=0.59<br>Re=0.65<br>F1=0.62 | Pr=0.36<br>Re=0.88<br>F1=0.52 | Pr=0.88<br>Re=0.39<br>F1=0.54 |
| Trehatos | Pr=0.69<br>Re=0.71<br>F1=0.70 | Pr=0.30<br>Re=0.56<br>F1=0.39 | Pr=0.67<br>Re=0.67<br>F1=0.67 |
| Dance Pirouette 1 (CMU) | Pr=0.75<br>Re= 0.53<br>F1=0.62 | Pr=0.51<br>Re=0.51<br>F1=0.51 | Pr=0.82<br>Re=0.54<br>F1=0.64 |
| Dance Pirouette 2 (CMU) | Pr=0.70<br>Re=0.61<br>F1=0.65 | Pr=0.23<br>Re=0.69<br>F1=0.34 | Pr=0.98<br>Re=0.32<br>F1=0.47 |
| Dance Pirouette 3 (CMU) | Pr=0.77<br>Re= 0.67<br>F1=0.72 | Pr=0.52<br>Re=0.47<br>F1=0.49 | Pr=0.89<br>Re=0.41<br>F1=0.56 |
| Dance Pirouette 4 (CMU) | Pr=0.71<br>Re= 0.63<br>F1=0.67 | Pr=0.59<br>Re=0.48<br>F1=0.53 | Pr=0.62<br>Re=0.58<br>F1=0.60 |

Table 5.2 Summarization performance metrics for different values of the LPF cutoff frequency.

| | Filter Cutoff Frequency | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Performance metrics | 0.01 | 0.02 | 0.03 | 0.04 | 0.05 | 0.06 | 0.07 | 0.08 |
| Precision | 28.1% | 45.3% | 58.9% | 64.4% | 58.8% | 47.1% | 33.2% | 22.3 % |
| Recall | 11.2% | 30.1% | 48.8% | 59% | 65.4% | 69.6% | 74.8% | 77.6% |
| F1-score | 16.01% | 36.2% | 53.4% | 61.6% | 61.9% | 56.2% | 46.0% | 34.6% |
| Compression Rate | 0.4% | 0.5% | 0.8% | 1.0% | 1.6% | 1.8% | 2.2% | 2.5% |
| MSE | 30.3db | 29.8db | 19.1db | 25.9db | 26.0db | 26.3db | 26.62db | 26.9db |

principles, we have designed, implemented and validated two approaches: a clustering-based method for the selection of the basic primitives of a choreography, and a kinematics-based approach that generates meaningful summaries at hierarchical levels of granularity. Future directions of this work include the extension of the kinematics model used to account for rotational motion, thus including angular velocity and acceleration, as well as exploring kinematics-based hierarchical sparse modelling approaches.

# Chapter 6

# Unsupervised 3D Motion Summarization Using Stacked Auto-Encoders

## 6.1 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 [196]. The existing video summarization techniques abstract the input data using three different approaches [197]. The first is the so-called *representative key-frame selection* that creates video summaries through a collection of representative key frames [198]. *The key subshot-oriented approach* selects the representative subshots of key-frames to form the video synopsis [199]. Finally, *the key object detection method* decomposes the whole video sequence into several single frames, each revealing representative objects in a given video sequence [200].

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 [7]. 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 [201],[202],[203].

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 [94] 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.

### 6.1.1 Related Work

The recent achievements of deep machine learning [94] 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 [204]. 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 SMRS algorithm [186] or its hierarchical implementation [7]. 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 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 [205]. Data compression can be achieved using other approaches, apart from SAE. The wavelet transform is one of these approaches [206]. 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 [207, 198]. 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 [7, 8, 202] 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

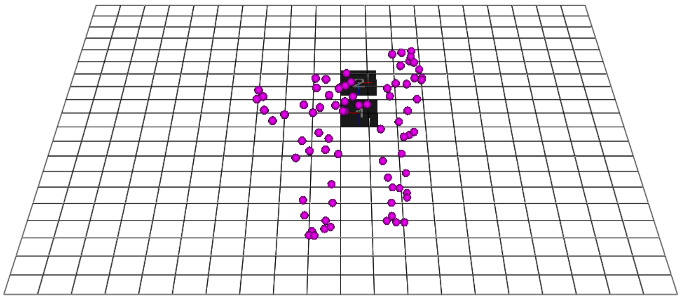


Figure 6.1 An example of geometric challenges due to the presence of multiple dancers.

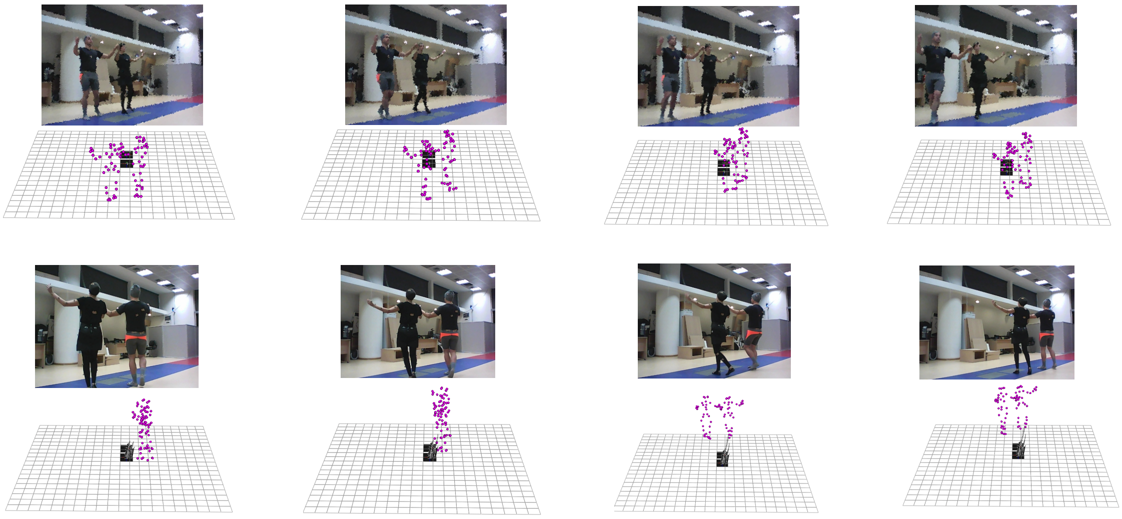


Figure 6.2 This visual sequences depict the motion capturing process. 3D skeletal data are obtained by the VICON motion capturing system (second and fourth row) and the respective RGB content (first and third row). This figure refers to Makedonikos dance sequence, executed by two dancers simultaneously.

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 6.1 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 6.2. 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.

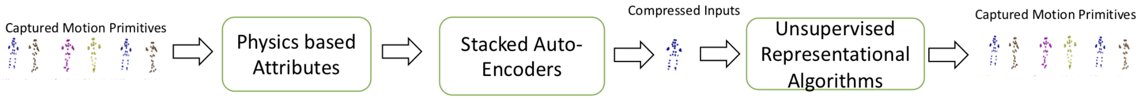

Captured Motion Primitives
Physics based Attributes
Stacked Auto-Encoders
Compressed Inputs
Unsupervised Representational Algorithms
Captured Motion Primitives

Figure 6.3 The proposed architecture for video dance summarization using stacked auto-encoders and representative algorithms.

### 6.1.2 Innovation and Originality

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.

This chapter 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.

## 6.2 The proposed summarization workflow

Figure 6.3 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 [202]. 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 [94, 205]. 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.

### 6.2.1 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 9.3). 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 6.2 show the humans' joints extracted both on RGB content (the first and the third row of Figure 6.2) and on a plane depicting the movement of the dancers in the space (second and fourth row of Figure 6.2). 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 6.2 where one can notice, by observing the frame content, that only one dancer appears to perform.

### 6.2.2 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 6.4 depicts the proposed SAE approach for input data dimensionality reduction. In the following Section 6.3, we analyze with more details the SAE structure adopted in this article.

### 6.2.3 Unsupervised Representational Sampling Algorithms

The last step of the proposed unsupervised video summarization algorithm employs traditional representational methods, such as the hierarchical SMRS [7], SMRS [186], K-OPTICS and Kennard Stone [208] for performing the final dance sequence summarization. K-OPTICS combines density-based and centroid based approaches [8, 118]. 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

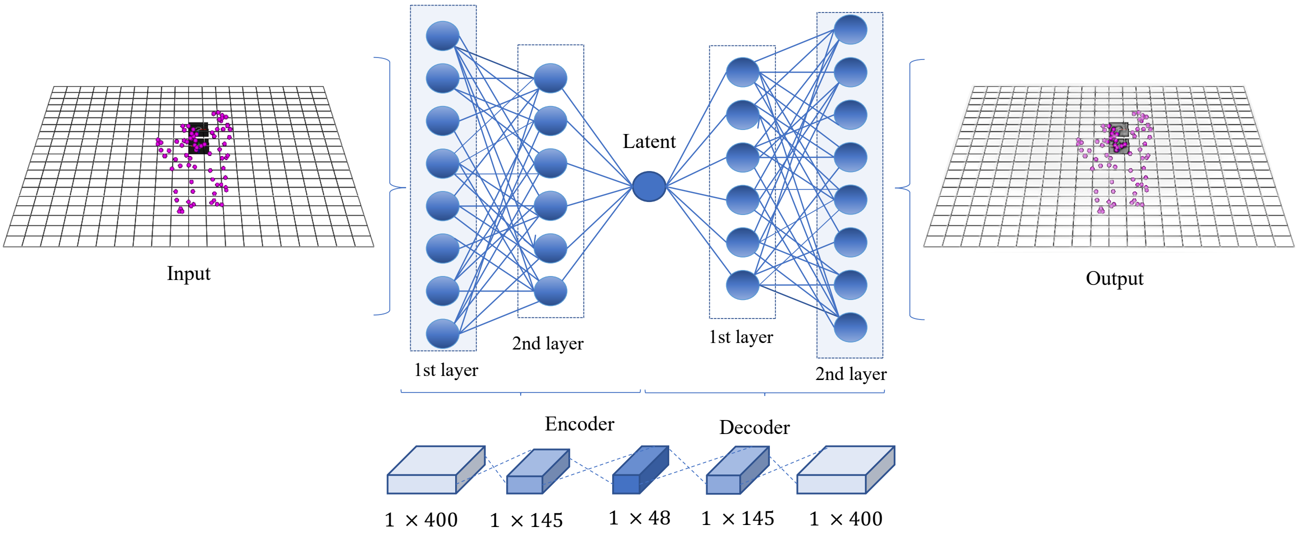


Figure 6.4 The structure of the proposed auto-encoder used for dimensionality reduction of the raw input signals.

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 [7], 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.

## 6.3 The Proposed SAE Scheme for Dance Sequence Summarization

The structure of the proposed SAE is depicted in Figure 6.4. 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 6.2.1). 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

$$\vec{h}_k^1 = f(W_1^T * \vec{x}_k + \vec{b}_1), \tag{6.1}$$

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.

$$min \sum_{i=1}^{Q} L(\vec{x}_k, \hat{\vec{x}}_k), \tag{6.2}$$

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.

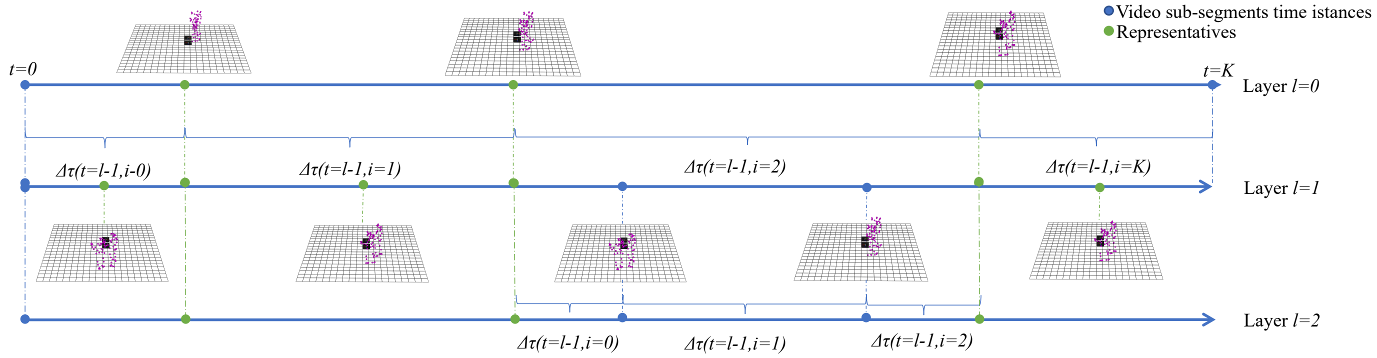


Figure 6.5 The architecture of the H-SMRS algorithm [7].

### 6.3.1 The Hierarchical-Sparse Modelling Representative Selection

A hierarchical implementation of the Sparse Modelling Representative Selection (SMRS) algorithm, say H-SMRS [7], 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 [186] 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) [161]. 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 [7] 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 6.5 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.

## 6.4 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 3.3.2). 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 6.2.1. 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 preprocess 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 (6.3)) 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 [208]. 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 summarization performs.

### 6.4.1 Evaluation Metrics

As we have stated above (see Sections 3.5, 3.3.2), 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,

$$\hat{l}(k) = \arg \min_{for\ all\ l \in G} |l - k| \ \forall \vec{r}_k \in R, \tag{6.3}$$

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,

$$\begin{aligned} \mu &= \frac{1}{K} \sum_{k=1}^{k=K} |\hat{l}(k) - k| \\ \mu_{max} &= \max_{\forall k \in R} |\hat{l}(k) - k|, \end{aligned} \tag{6.4}$$

where $\mu$ is the average time instance deviation among all $K$ extracted representatives and $\mu_{max}$ 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 $\vec{r}_k, k = 1, ..., K$.

$$\rho_i^{max} = \max_{\forall \vec{r}_k \in R} \rho(\vec{x}_i, \vec{r}_k) \ \forall \vec{x}_i, \tag{6.5}$$

where $\rho(\cdot)$ refers to the correlation coefficient of two vectors. The maximum the value $\rho$ 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.

## 6.5 Choreographic Summarization Experiments

In this section, we present some results of different summarization algorithms on the above-mentioned dance sequences. In particular, Figure 6.6 demonstrates the results obtained on Syrtos (2 beat) dance sequence, consisting of more than 5000 frames, using as summarization algorithm the K-OPTICS. More

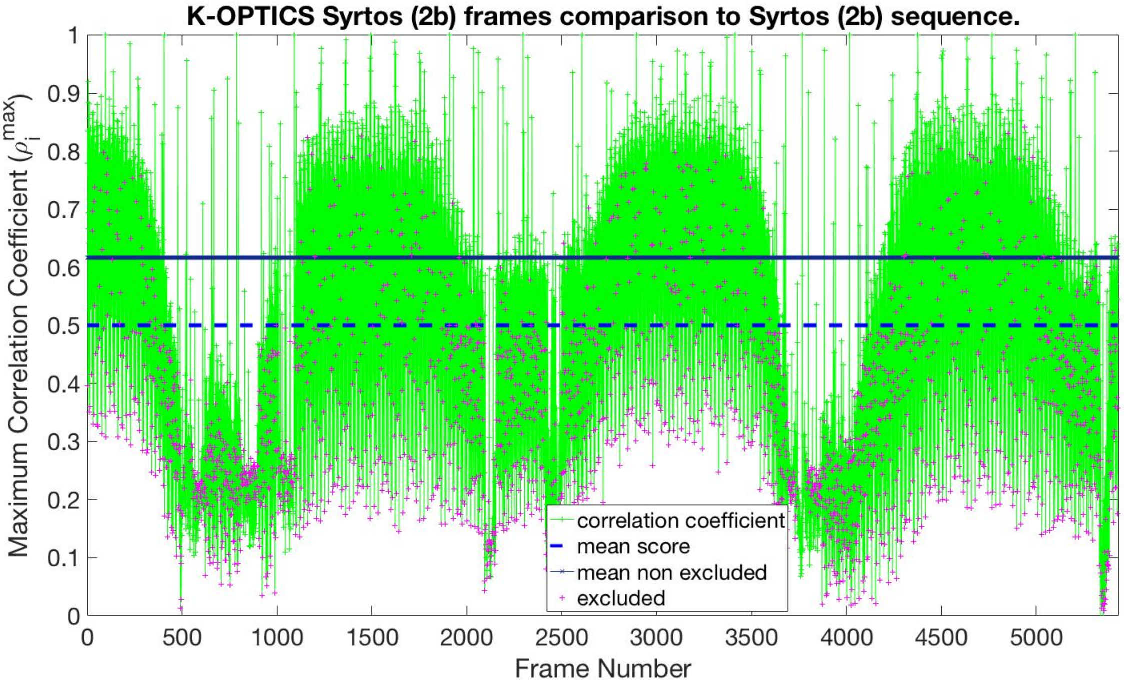


Figure 6.6 The maximum correlation scores $\rho_i^{max}$ for each frame of the original Syrtos at 2 beat dance sequence compared to the summarized one using K-OPTICS.

specifically, we extract 32 key-representatives using the K-OPTICS algorithm and then we calculate the maximum correlation score $\rho_i^{max}$ for each frame of Syrtos (2 beat) dance sequence against the 32 key frames extracted [see Equation (6.5)]. As shown in Figure 6.6, the average $\rho_i^{max}$ for all 5,000 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 $\rho_i^{max}$ 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 6.6 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 6.7 illustrates the summarization performance when the Kennard Stone sampling algorithm is applied over Syrtos (3 beat) dance sequence. Again, as in Figure 6.6, 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 6.1 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

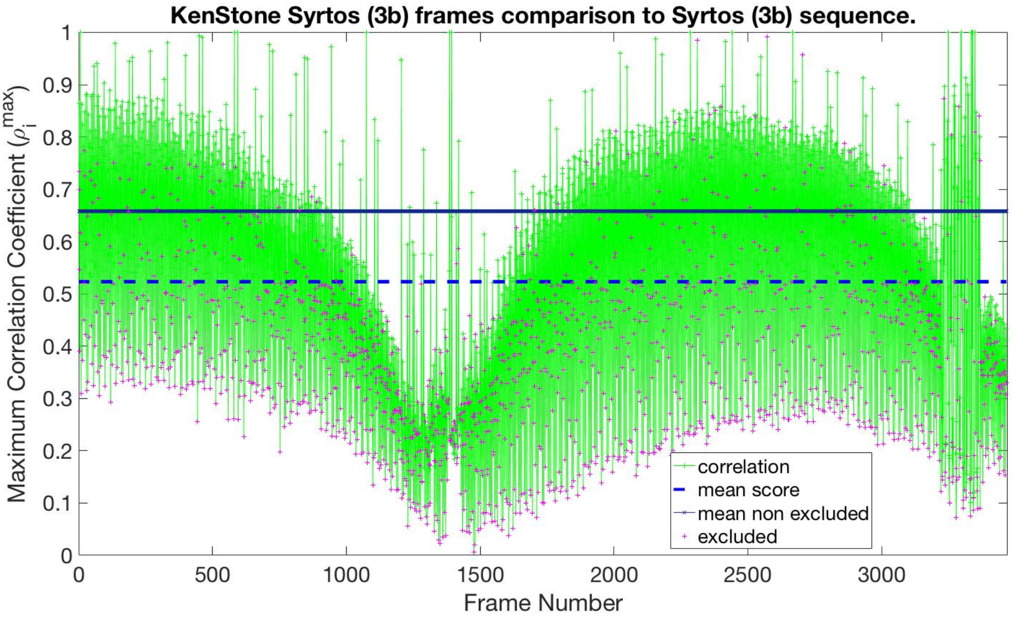


Figure 6.7 The maximum correlation scores $\rho_i^{max}$ for each frame of the original Syrtos at 3 beat dance sequence compared to the summarized one using Kennard Stone.

as representatives of the whole dance sequence variation. In this table, we have presented as bold the highest correlation values.

Figure 6.8 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. Since 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. 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 6.8, 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 mentioning 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 6.2 shows the average time deviation of key frames extracted by the four summarization algorithms and the ground truth data, that is, the value $\mu$, 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

Table 6.1 Maximum correlation coefficient scores ($\rho_i^{max}$) 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 |

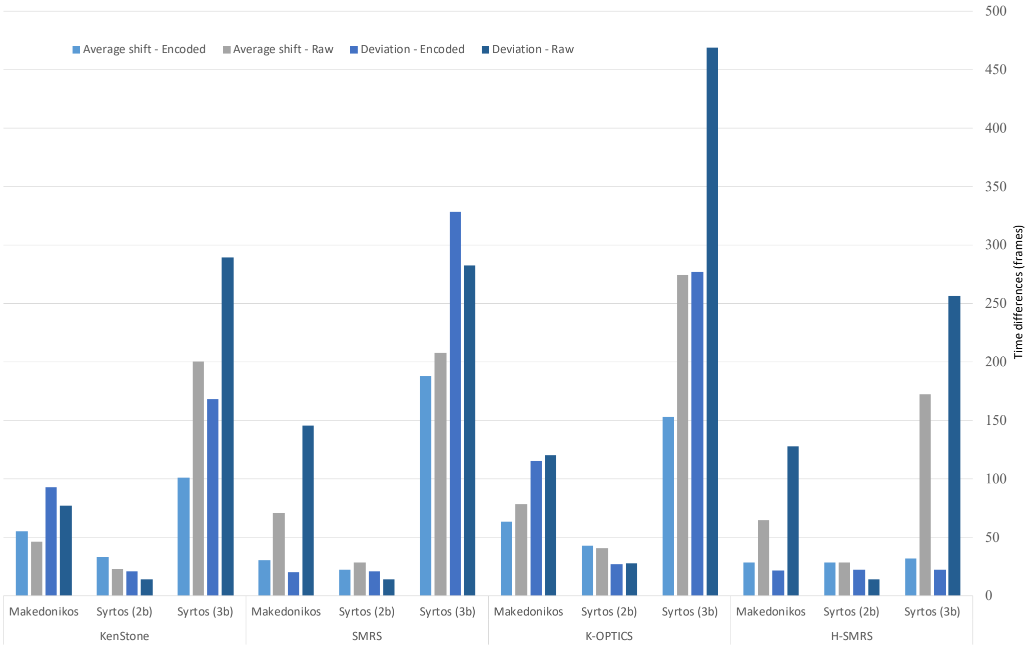


Figure 6.8 Data input type summarization impact when two dancers performed simultaneously for all the examined dance sequences.

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 6.2 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 | Aver. Shift With SAE | Average Shift without SAE | Standard Deviation with SAE | Standard Deviation without SAE |
|---|---|---|---|---|---|
| | Makedonikos | 0.46 | 0.38 | 0.78 | 0.64 |
| KenStone | Syrtos (2b) | 0.27 | 0.19 | 0.17 | 0.12 |
| | Syrtos (3b) | 0.84 | 1.67 | 1.4 | 2.41 |
| | Makedonikos | 0.25 | 0.59 | 0.17 | 1.21 |
| K- OPTICS | Syrtos (2b) | 0.19 | 0.23 | 0.17 | 0.12 |
| | Syrtos (3b) | 1.03 | 1.73 | 2.74 | 2.35 |
| | Makedonikos | 0.53 | 0.65 | 0.96 | 1 |
| SMRS | Syrtos (2b) | 0.34 | 0.36 | 0.23 | 0.23 |
| | Syrtos (3b) | 1.28 | 2.29 | 2.31 | 3.91 |
| | Makedonikos | 0.24 | 0.54 | 0.18 | 1.06 |
| H-SMRS | 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 6.3 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 6.3 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 6.9 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 6.10 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 6.10a) are close to the original skeletal data (see Figure 6.10b) 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.

Another important criterion is how results vary (fluctuate) from the average values, as depicted in Figure 6.8. This is also illustrated in Table 6.2 where the standard deviation of the average time shift is given. But in Table 6.4 we also present the minimum (best) and the maximum (worst) performance (that is, $\mu_{max}$ of Equation 6.4) for all the three dances. As we can see, $\mu_{max}$ 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.

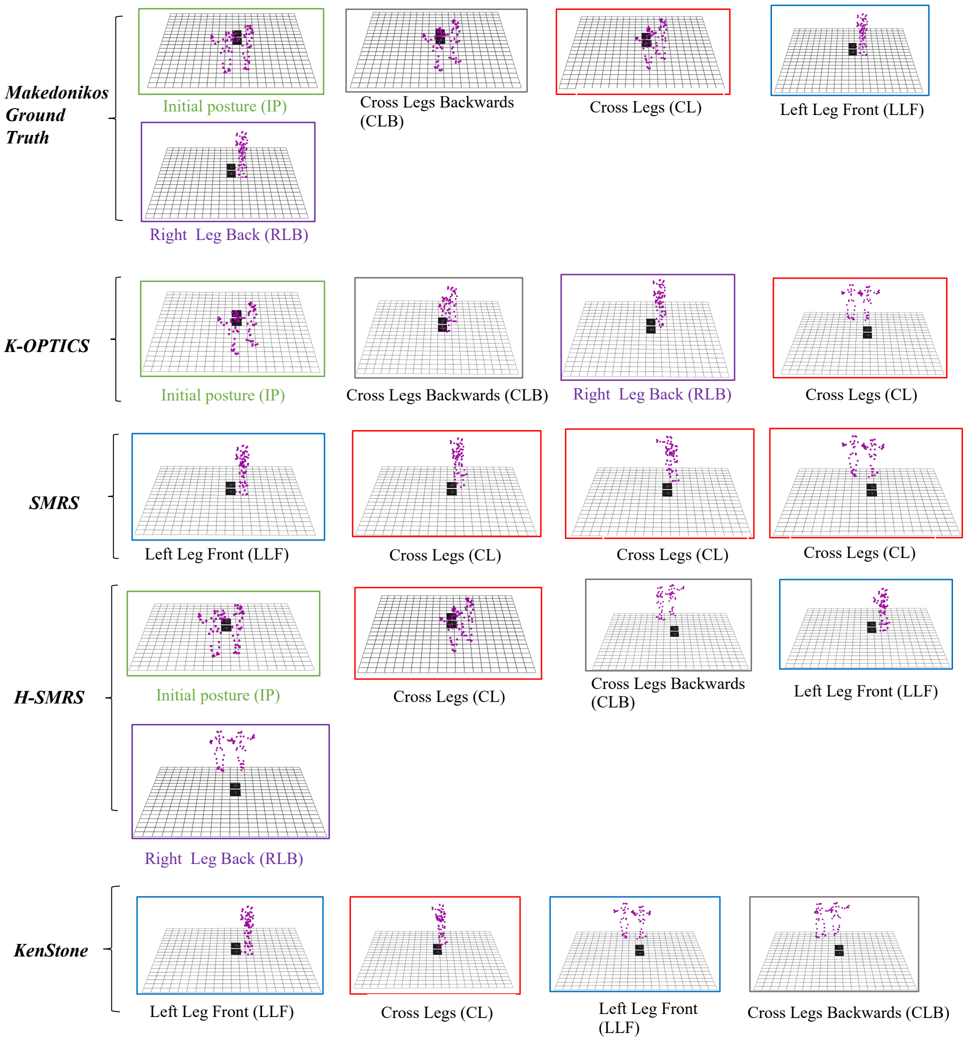


Figure 6.9 A visual representation of the key frames extracted by the four summarization algorithms than the ground truth ones in case of Makedonikos dance.

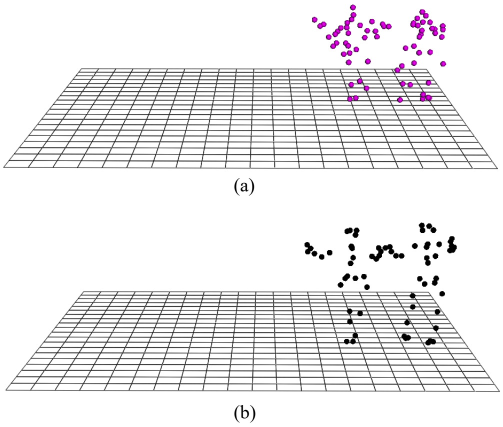


Figure 6.10 A representation of the decompressed data (**a**) relative to the original skeletal data (**b**), for the same time frame as for Syrtos (2b) dance sequence.

Table 6.4 The minimum (best) and maximum (worst) time deviation ($\mu_{max}$) 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 |

## 6.6 Discussion

In this chapter, we proposed 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 approach 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 approach 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.

# Part III

# Modelling and Analysis of Dance Sequences

# Chapter 7

# Choreographic Analysis using Dynamic Time Warping

## 7.1 Introduction

Intangible Cultural Heritage (ICH) is a prominent element of people's cultural identity as well as a significant aspect for growth and sustainability [38]. The expression of identity through Intangible Cultural Heritage takes many forms, among which folkloric dances hold a central position [209]. It is reasonable to consider that analyzing choreographic sequences is essentially a multidimensional modeling problem, given that both temporal and spatial factors should be taken into account. Research has been published in the literature pertaining to ICH preservation which focuses on the time element [210], [211], [212], [22]. Typical preservation acts include digitization, modelling and documentation.

Another important factor in preserving any type of performing arts, would be the development of an interactive framework that enhances the learning procedure of folklore dances. The recent advances in depth sensors, which have concluded to the development of low-cost 3D capturing systems, such as Microsoft Kinect [40] or Intel RealSense [213], permit easy capturing of human skeleton joints, in 3D space, which are then properly analyzed to extract dance kinematics [112]. The preservation of folk dances can be facilitated by modern Information and Communication Technologies by levarging recent developments in a variety of areas, such as storage, image and video processing, machine learning, cloud computing, crowdsourcing and automatic semantic annotation, to name a few [214].

Nevertheless, the digitization and the modelling of the information remains the most valuable task. Due to the tremendous growth of the motion capturing systems, depth cameras are a popular solution employed in many cases, because of their reliability, cost-effectiveness and usability and despite their limited range. Kinect is one of the most recognizable sensors in this category and in the choreography context can be used for recording sequences of points in 3D space for body joints at certain moments in time. Several recent research papers in the literature make use of such sensors for dance analysis, for example educational dance applications using sensors and gaming technologies [215], trajectory interpretation [216], advanced skeletal joints tracking [68], action or activity recognition [59, 56, 217–220, 94], key pose identification [221] and key pose analysis [1]. Apart from Kinect, another popular alternative motion capture system is VICON which is significantly more sophisticated and accurate [7], [112], [54].

### 7.1.1 Related Work

In [222], a comparison between abilities of the Kinect and VICON for gait analysis is introduced in the orthopaedic and neurologic field. In [223], the authors focus on the precision of the Kinect and the VICON motion capturing systems creating an application for rehabilitation treatments. In [224], the authors propose that the Kinect was able to accurately measure timing of clinically relevant movements in people with Parkinson disease. Contrary to the linear regression based approaches that have been carried out in the bio-medical field [223], [222], [224] regarding the similarities/dissimilarities and the precision of the adopted motion capturing system, in this work we follow a Dynamic Time Warping (DTW) approach in the kinesiology field. Moreover, the aforementioned approaches pertain to simple movement sequences i.e., knee flexion and extension, hip flexion and extension instead of our proposed choreographic dataset which includes more complex movements that combine several joints variations (see Table 3.5).

In [225], the authors introduce a motion classification framework using DTW. The aforementioned work utilizes DTW algorithm in order to classify motion sequences using the minimum set of bones (7 body joints). On contrary, our proposed framework uses 25 body joints analyzing the motion sequences using the DTW and Move-Split-Merge algorithms respectively. In [226], the authors propose an algorithm for 3D motion recognition which allows extensions of DTW with multiple sensors (view-point-weighted, fully weighted and motion-weighted) and can be employed in a variety of settings. DTW algorithm has also adopted in order to extract the kinessiology details from video sequences. In [227], the authors propose a video human motion recognition approach, which uses DTW to match motion projections in non-linear manifold space. In [228], the authors present a technique for motion pattern and action recognition, which employs DTW to match motion projections in Isomap non-linear manifold space.

Our proposed framework focuses on the similarity assessment of folkloric dances, using data from heterogeneous sources;i.e. data from high-cost devices like VICON and low-cost devices like Kinect II using predefined choreographic sequences. Research outcomes target on the underlying relationships among dances captured using the VICON and Kinect systems (see table 3.1).

VICON is a high-cost, motion capturing system, which exploits markers attached on dancers' joints to extract motion variations and the trajectory of a choreography. The VICON motion capturing system requires i) a properly equipped room of cameras and trackers, ii) experienced staff to manage the VICON devices, iii) a pre-capturing procedure, which is obligatory to calibrate the whole system. On the other hand, Kinect II is a low cost depth sensor, which requires no markers to extract the depth and humans' skeleton joints. This makes Kinect II applicable to non-professional users (everybody) from any environment (everywhere) and at any time. However, the captured trajectories are not as accurate as the ones extracted by the VICON system.

Consequently, the Kinect II device can be used as an in-home learning tool for most of dance choreographies by simple (non-experienced) users. This papers relates dance motion trajectories captured by the accurate VICON and the non accurate Kinect II system. A Dynamic Time Warping (DTW) methodology is adopted in order to find out similarities/dissimilarities between the two devices, considering as accurate reference dance motion trajectory the one derived from the VICON system. DTW algorithm can localize dance steps patterns which can not be accurately represented by the Kinect system and patterns that Kinect can be sufficiently described.

### 7.1.2 Contribution and Originality

The contribution of this work can be summarized in the following: Firstly, we present a comparative study on trajectory similarity estimation approaches, on data obtained by two types of sensors, using a complex dataset with challenging choreographic sequences, where joint movements are often varied and unstructured. Furthermore, the conducted experiments indicate that if significant levels of precision are ensured during initial data collection, design, development and fine-tuning of the system, then low-cost and widely popular motion capturing sensors, such as Kinect-II, suffice to provide a smooth and integrated experience on the user end, which would allow for relevant educational or entertainment applications to be adopted at scale.

## 7.2 The Proposed Methodology

In this work we investigate the possibility of utilizing skeleton data points as reference points, for the identification of dance choreographs. Data originates from professional motion capture equipment. These instances are used against corresponding skeletal data, recorded using low cost sensors.The proposed approach consists of the following steps: a)data capturing using high-end motion capture system, b) feature extraction, c) descriptive frames selection, for the database creation, d) data capturing using low-cost sensors, e) extraction of corresponding body joints and f) similarity assessment among the dance patterns between high-end and low-cost sensors.

The idea of spatial-temporal information management [229], [54] is applied, so that recorded dance sequences are summarized to a sequence of keyframes. This is achieved by employing an iterative clustering scheme, imposing time constraints. The proposed data managing scheme reduces the dance sequence to few keyframes, which are selected using density based clustering, in predefined time related subsets. It is important to note that noise or tempo variations do not affect the proposed approach. Given as set of keyframe sequences, for different dances, a comparison is performed among them. The sequences are signals containing information over dancer's joints' position and rotation. Signal similarity, employing the correlation measure is performed. Consequently, variations of the same dance should be easily identified, due to high similarity scores. Fig 8.1 depicts a block diagram of the proposed methodology.

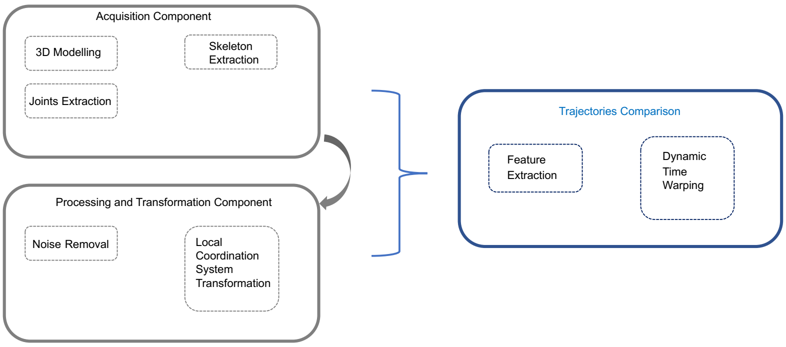


Figure 7.1 A block diagram of our proposed methodology.

## 7.3 Dynamic Time Warping for Dance Sequences Modelling

### 7.3.1 Dynamic Time Warping

Dynamic Time Warping [230] calculates an optimal match between two temporal sequences. DTW generated matching path is based on linear matching, but has specific conditions that need to be satisfied, in particular the conditions pertaining to continuity, boundary condition, and monotonicity. In the following a brief description on matching between curve points is provided. If $N_1$and $N_2$ are the number of points in two curves, then $i$-th point of curve 1 and the $j$-th point of curve 2 match if:

$$\frac{i-1}{N_1} * N_2 \leq j \leq \frac{i}{N_1} * N_2 \tag{7.1}$$

It should be mentioned that each point can match with maximum one point of the other curve. The boundary condition forces a match between the first points of the curve and a match between the last points of the curve. The continuity condition decides how much the matching can differ from the linear matching. The aforementioned condition is the heart of DTW. We formulate the aforementioned assumption as follows:

$$\frac{N_2}{N_1} * i - c * N_2 \leq j \leq \frac{N_2}{N_1} * i + c * N_2 \tag{7.2}$$

In the case that during the process of matching it is concluded that the $i$-th point of the first curve should match with the $j$-th point of the second curve, it is not possible: (i) that any point of the former with an index greater than $i$ matches with a point of the latter with an index smaller than $j$, and (ii) that any point of the former with an index smaller than $i$ matches with a point on the latter with index greater than $j$.

### 7.3.2 Kinect-II Evaluation using DTW

In our proposed methodology, we denote as reference sequences those are derived by the VICON motion capturing system. In addition, each choreographic sequence obtained by the low-cost sensor Kinect-II is contrasted to the VICON sequence. Our scope is to define the similarities/dissimilarities comparing the choreographic sequence for each dance using the DTW algorithm [230]. Furthermore, each choreographic sequence is depicted as a curve with different characteristics (e.g., duration, length). Our

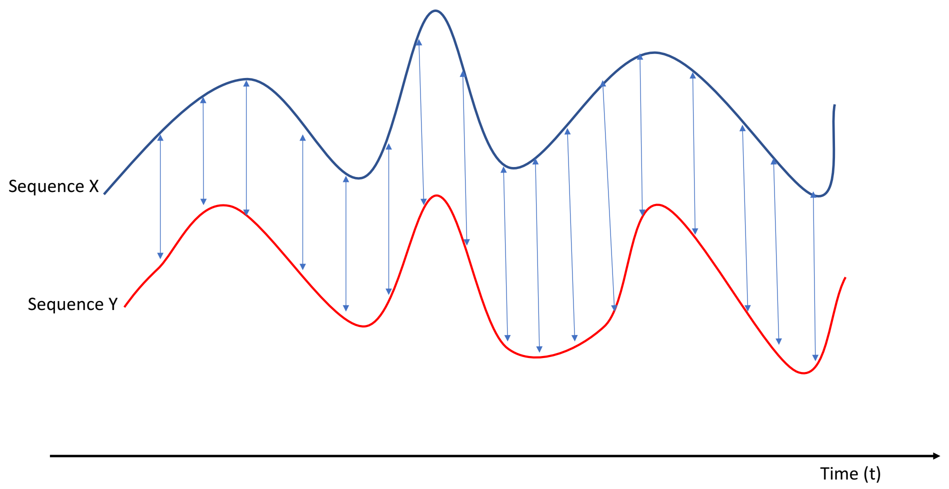


Figure 7.2 Time alignment of two choreographic sequences. Aligned points are depicted by the arrows.

proposed framework is to define the similarities/dissimilarities between the curves of the heterogeneous motion capturing systems. Every index of the choreographic sequence is matched with one (or more) indices of the other sequence for each dance. Fig. 7.2 depicts time alignment between two independent signals, in our framework the signals are obtained by the motion capturing systems. Let us denote, as $\vec{X}$ the sequences of the Kinect sensor and $\vec{Y}$ the sequences of the VICON accordingly. The $\vec{X}$ and $\vec{Y}$ enclosure the kinessiology features (body joints variations) for each dancer creating a motion database for the heterogeneous capturing system. In order to compare each feature, we define a local cost measure describing the similarity/dissimilarity of each feature. The cost matrix is defined as $P \in \mathbb{R}^{NxM}$ $P(n,m)$=$p(x_n,y_m)$. An $(N,M)$ dynamic warping path $p$= $(p_1,\cdots$ ,p$_s)$ determines an alignment between the $\vec{X}$ and $\vec{Y}$ vectors by assigning the element $x_{ns}$ of $\vec{X}$ to the element $y_{ms}$ of $\vec{Y}$. The vectors $\vec{X}$ and $\vec{Y}$ are denoted as follows:

$$\vec{X} = (x_1,\ldots,x_N) \tag{7.3}$$

$$\vec{Y} = (x_1,\ldots,x_M), \quad M \in \mathbb{N}. \tag{7.4}$$

In the following, we create a space defined by $F$. Then $x_n$, $y_m \in F$ for $n \in [1{:}N]$ and $m \in [1{:}M]$. In our framework, we define as $\vec{X}$ and $\vec{Y}$ the features which are obtained by the motion capturing system indicating every joint of the dancers body. Due to the heterogeneous motion capturing system, we should define the local coordination system. Fig. 7.3 depicts the transformation from the global coordination system to a local system for each motion capturing system, which is simultaneously a type of range fix that takes into consideration body parameters such as limb length. Inevitably, for the aforementioned constraints we denote as $\vec{C}_k^G = (x_k^G, y_k^G, z_k^G)$ the $k$-th joint out of the $M$=35 acquired by VICON system and $\vec{I}_l^G = (x_l^G, y_l^G, z_l^G)$ the $l$-th out of the $L$=25 obtained by the Kinect-II sensor respectively. Variables $x_i^G$, $y_i^G$ and $z_i^G$ indicate the coordinates of the respective $i$-th joint with regard to a reference point setting VICON architecture (in our case the center of the square surface). We have acquired the aforementioned

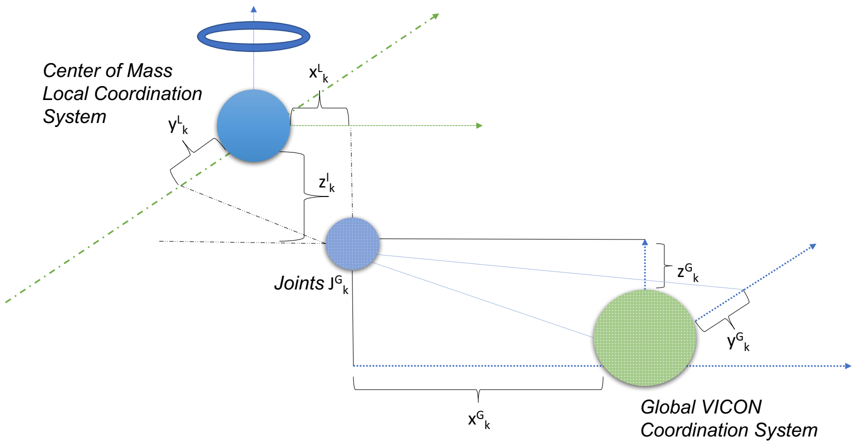


Figure 7.3 VICON global coordination system being transformed to a local coordination system. Its center is the center of mass of the dancer [8]. This allows for compensation of the dancer spatial positioning.

joints after applying a density-based filtering on the entirety of the detected joints so as to eliminate noise introduced during the acquisition procedure. The main difficulty in directly processing the extracted joints $\vec{C}_k^G$, $k$=1,2,...,$M$ is the coordinates system. Thus, we need to transform the $\vec{C}_k^G = (x_k^G, y_k^G, z_k^G)$ from the VICON coordinate system to a local coordinate system, the center of which is the center of mass of the dancer. We follow the same procedure for the Kinect-II architecture. This is obtained through the application of Eq. (7.5) on the joints coordinates $\vec{J}_k^G$,

$$\vec{C}_k^L = \vec{C}_k^G - \vec{C}_{cm} \tag{7.5}$$

$$\vec{I}_l^L = \vec{I}_l^G - \vec{I}_{cm} \tag{7.6}$$

where $\vec{H}_{cm}$ denotes the dancer's center of mass with regard to the coordination system expressed as:

$$\vec{C}_{cm} = \sum_{k=1}^{M} \frac{\vec{C}_k^L}{M} \tag{7.7}$$

$$\vec{I}_{cm} = \sum_{l=1}^{L} \frac{\vec{C}_l^L}{L} \tag{7.8}$$

and we recall that $M$, $L$ refers to the total number of joints extracted by the VICON and Kinect capturing system respectively.

Let us denote as cost matrix $p(\vec{X},\vec{Y})$=$p(\vec{C}_k^L,\vec{I}_l^L)$ the total cost of a warping path $p$ between $\vec{C}_k^L$ and $\vec{I}_l^L$.

$$\vec{p}(\vec{X},\vec{Y}) = \sum_{l=1}^{s} (p(x_n, y_m)) \tag{7.9}$$

The DTW distance between the $\vec{C}_k^L$ and $\vec{I}_l^L$ is defined ad follows:

$$D\vec{T}W(\vec{X},\vec{Y}) = \min p(\vec{C_k^L},\vec{I_l^L}) \tag{7.10}$$

### 7.3.3 Kinect-II Evaluation using Move-Split-Merge.

Motivated by the superiority of DTW for motion analysis shown in previous works e.g. against SVM [227], or approaches based on Locally Linear Embedding (LLE), Locality Preserving Projections (LPP) and LLP-HMM [228] we adopt DTW as our main reference algorithm. Moreover, we conduct further comparative experiments to also evaluate against a recent technique called Move-Split-Merge [231]. The Move-Split-Merge distance algorithm provides a means of measurement that resembles other distance-based approaches, where similarities/dissimilarities are computed by employing a series of operations for the transformation of a series "source" into a series "target". Move-Split-Merge algorithm utilizes as building blocks three fundamental operators. The Move operation is equivalent with a replacement operation, in which one value substitutes another. Split inserts an identical copy of a value immediately after its first instance, while Merge erases a value if it directly follows an identical value. Let us assume $X_i$=$(x_i,...,x_m)$ as a finite motion sequence of real numbers $x_i$. The move operation and the cost operation are defined as follows:

$$Move_{i,u}(X) = (x_1,...,x_{i-1},x_i+u,x_{i+1},...,x_m) \tag{7.11}$$

$$Cost(Move_{i,u}) = |u| \tag{7.12}$$

$$Split_i(X) = (x_1,...,x_{i-1},x_i,x_i,x_{i+1},...,x_m) \tag{7.13}$$

$$Cost(Split_i) = c \tag{7.14}$$

$$Merge_i(X) = (x_1,...,x_{i-1},x_{i+1},...,x_m) \tag{7.15}$$

$$Cost(Merge_i) = c \tag{7.16}$$

$$C(x_i,x_{i-1},y_j) = \begin{cases} c & \text{if} \quad x_{i-1} \leq x_i \leq y_j \quad or \quad x_{i-1} \geq x_i \geq y_j \\ c+min(|x_i-x_{i-1}|,|x_i-y_j|) & \text{otherwise} \end{cases} \tag{7.17}$$

## 7.4 Experimental Results

In our study, for capturing of the dancers' movement variations, we employ a multi-faceted motion capture system including one Kinect II depth sensor, the i-Treasures Game Design module (ITGD) module created in the context of i-Treasures project [38] and VICON motion capturing system. The

Table 7.1 The considered dances and their variations along with the length of each sequence for each of the three dancers. These dances were recorded using Kinect-II.

| Dance | Variation | Short Name | Duration | | |
|---|---|---|---|---|---|
| | | | Dancer 1 | Dancer 2 | Dancer 3 |
| Enteka | Straight | Syrt11Str8 | 749 | 807 | 858 |
| Kalamatianos | Circle | KalCirc | 655 | 593 | 561 |
| | Straight | KalStr8 | 304 | 378 | 455 |
| Makedonikos | Circle | MakCirc | 424 | 582 | 409 |
| | Straight | MakStr8 | 283 | 367 | 418 |
| Syrtos 2 beat | Circle | Syrt2Circ | 608 | 543 | 352 |
| | Straight | Syrt2Str8 | 623 | 639 | 334 |
| Syrtos 3 beat | Circle | Syrt3Circ | 608 | 964 | 947 |
| | Straight | Syrt3Str8 | 1366 | 678 | 511 |
| Trehatos | Circle | TrehCirc | 991 | 723 | 443 |
| | Straight | TrehStr8 | 315 | 295 | 355 |

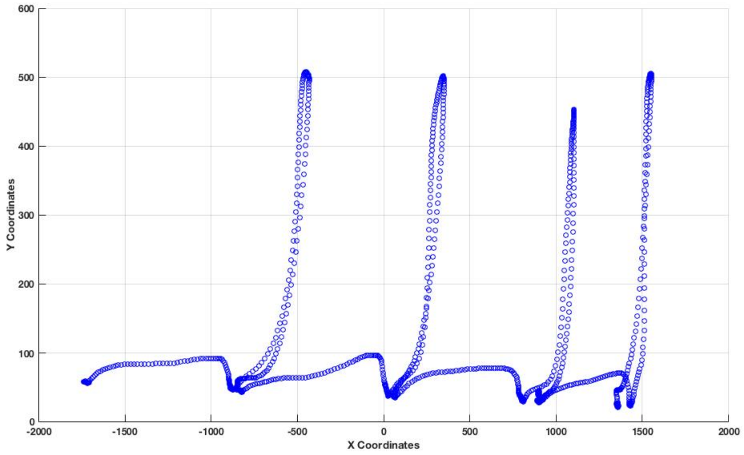


Figure 7.4 The coordinates of the trajectory of the left foot joint, which shows the rhythm of the dance performed by dancer 1.

ITGD module gives the possibility of recording and annotating mocap data acquired by a Kinect sensor. The employed algorithms were implemented in MATLAB. A variety of Greek folk dances with varying levels of complexity have been obtained. Three dancers (two men and one woman) each performed every dance twice: Once in a straight line and once in a semi-circular curved line. Fig. 7.6 and Fig. 7.7 depict the most representative postures of the Syrtos at 2 beats and Enteka dance respectively. Fig. 7.8 of Syrtos dance at 3 beats. Each choreographic posture indicates representative frames that summarizing the whole choreographic sequence providing the kinessiology patterns. Table 7.1 depicts the different duration of these dances across three different dancers.

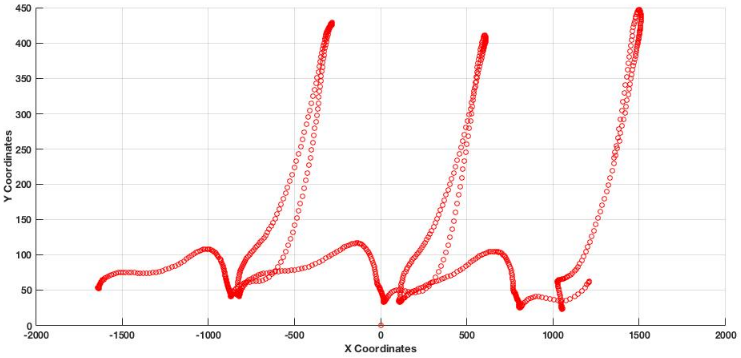


Figure 7.5 The coordinates of the trajectory of the left foot joint, which shows the rhythm of the dance performed by dancer 2.

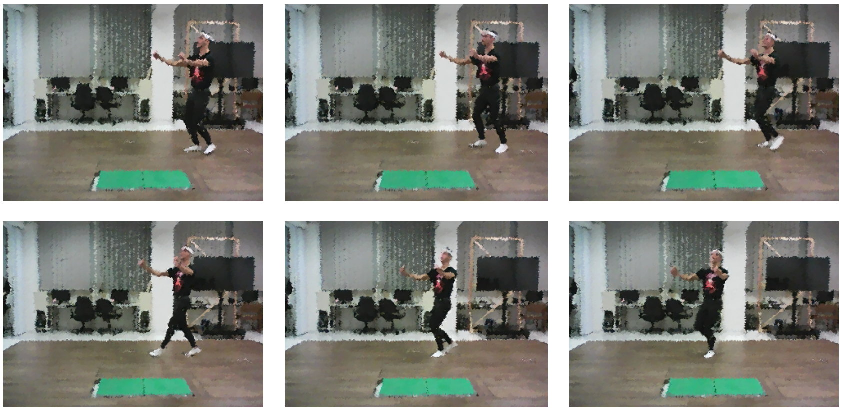


Figure 7.6 Illustration of Syrtos dance (2 beats, circular trajectory).

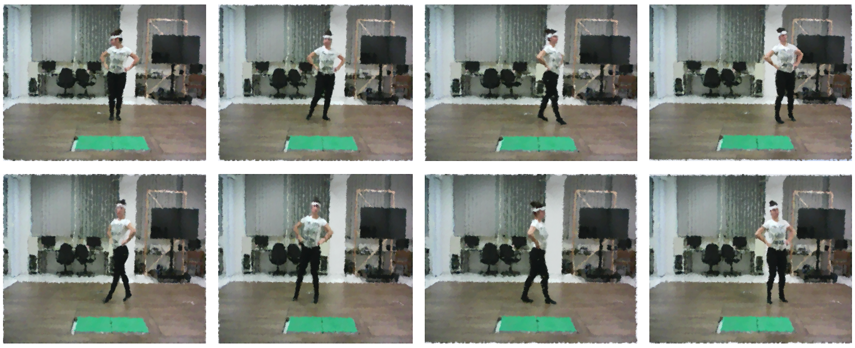


Figure 7.7 Illustration of Enteka dance performed by dancer 3.

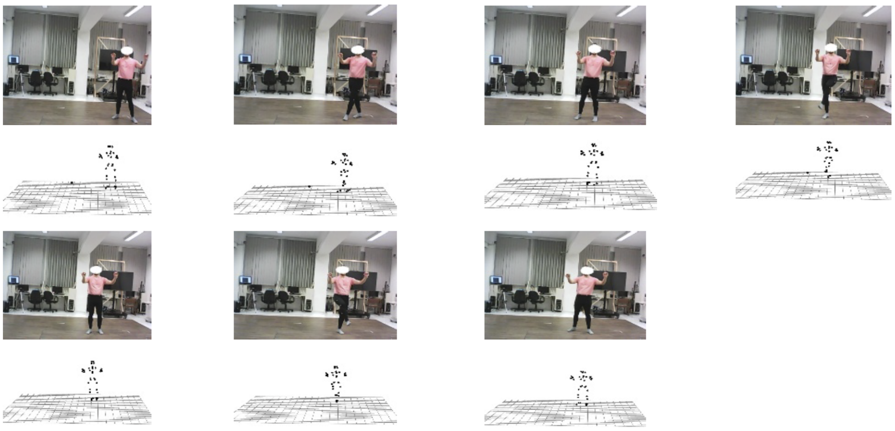


Figure 7.8 An instance that illustrates seven frames from the Syrtos at 3 beats dance.

### 7.4.1 Dataset Description

The dataset comprises six different folklore dances. For the Kinect capturing process, we use a single Kinect-II sensors placed in the front. Every dance is described by a set of consecutive image frames. Every frame $I_l$, $i = 1, \dots, l$ has a corresponding extensible mark-up language (XML) file with positions, rotations and confidence scores for 25 joints on the body (see Fig. 3.5) addition to timestamps. In Table 3.5, a brief description of the dances is provided [54]. After a series of processing steps, a skeleton from the VICON system is represented. In the discussed setting, ten Bonita B3 cameras were used. The capturing space was a square of 6.75 meters width, and the square's center constitutes the origin of the VICON coordinate system. We used a calibration wand with markers in order to optimize the calibration procedure. The dancers' movements were captured through the use of 35 markers at fixed positions on their bodies.

### 7.4.2 Similarity Analysis

Similarity analysis entails to a dance matching problem. Specifically, given a set of frames, from multiple body joints, captured using the Kinect, we try to identify the most closely related trajectories from the choreographic database. Assume that we have $n$ experienced dancers in the database. Then each time a new user performs a dance, the algorithm calculate the similarity scores among the newly recorded dance and the existing dances in the DB. Then, for each of the $n$ experienced dancers, we get the top 3 closest trajectories, given a distance metric. Thus, we have a total of $n$ times 3 dance suggestions. In this study we have 3 experienced dancers. Thus we had 9 dance suggestions every time. The similarity score (i.e. DTW or MSM) is then used to rank the results. Performance analysis focuses on how accurate the system is in matching correctly the recorded dance.

At first, we asked the dancer to execute a specific choreography. Since, VICON's frame rate is 4 times greater than Kinect, we have consider a sub-sample approach in a ratio 1 to 4; that way the frame rate matches the Kinect. Then, we exploit the similarity tests with existing entries in the database. Despite the variations in the trajectories, we expect that the movement itself will be similar among dancers. Thus, the similarity analysis has a solid base. Fig. 7.5 and 7.4 illustrate the left foot joint movement on the floor for two different dancers. As we observe, the choreographic pattern of each dance is extracted indicating not only the kinessiology variation of the dancers' joints but also the music tempo. The main patterns appears the same, despite the variations in descriptive characteristics (e.g. length and height).

Proposed approach's matching performance is displayed in fig. 7.9. Results illustrate the number of matches, for a specific recorded dance, to the existing dances in data base. There are three performance classes, denoted as Top3, Top6 and Top 9. Numbers 3, 6, and 9 indicate the number of the closest matched dances (from the database to the one currently performed). Recall that we have three professional dancers and each of them performed the same six dances. Thus, the highest possible score in category TopX is 3. Results indicate that the suggested methodology managed to match correctly at least once all the investigated dances, despite their complexity, as explained in [232].

Fig. 7.9 provides further insights to the similarity between the VICON and the Kinect-II sensors. The x axis depicts the name of each the dance (see Table 7.1) and the y axis the number of the matches according to the choreographic database. For example, Makedonikos in circular trajectory (MakCirc) Top9 score indicates that among the nine closest trajectory patterns, we have 3 matches with the Makedonikos dance captured by Kinect-II, one per dancer in the choreographic database. Consequently, Makedonikos dance captured by VICON system was matched to Makedonikos dance captured by Kinect-II; to an extent, most of the choreographies were successfully matched, by defining a score using the DTW or MSM algorithms, despite the differences in employed motion capture technologies.

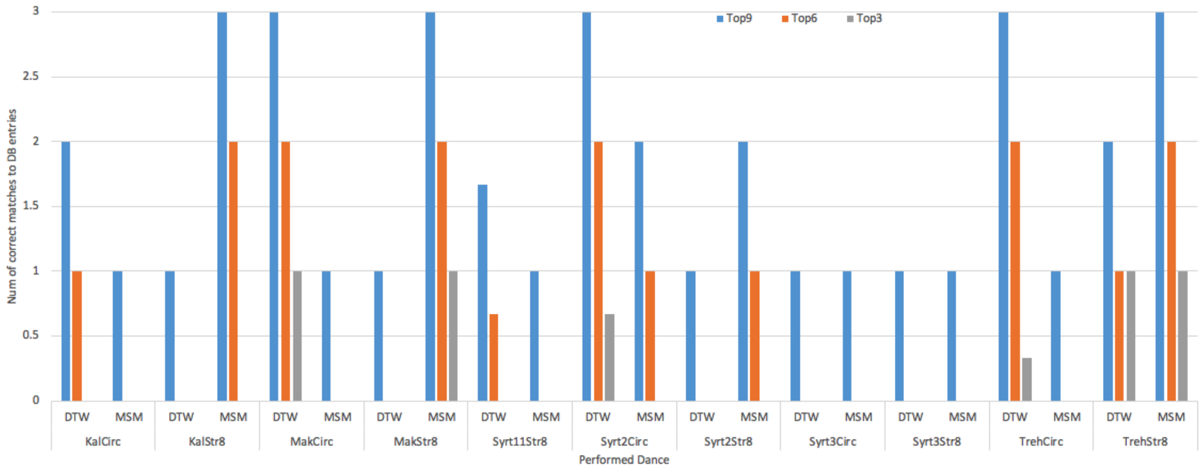


Figure 7.9 Performance illustration for the matching process.

## 7.5 Discussion

In this chapter, we explored the feasibility of pattern matching between heterogeneous motion capturing systems. The case study emphasizes on northern Greek folklore dances, which although complex and with several variations and particularities in their pattern, are characterized by elements of structure, contrary to chaotic versions of movement trajectories (e.g. [233]) in which similar explorations are far more difficult to perform. In this work, a two step process is adopted. The first step utilizes Kinect-II sensors, which provide dancer's skeleton feature values and a database is created. The second step involves the comparison of the trajectories in the database with a second database, created using VICON. The employed algorithms calculate similarity scores. According to these scores the algorithm provides a similar dance suggestion, for each of the dancers, in the choreographic database. The obtained results suggest that low-cost sensors such as Kinect-II can be utilized in the context of dance-related educational or entertainment applications, at least as part of the end-user side. Such a setup would however require the employment of a detailed and highly accurate dataset for training and development of the system, captured by a high precision system such as VICON. The conducted experiments indicate that if significant levels of precision are ensured during initial data collection, design, development and fine-tuning of the system, then low-cost and widely popular motion capturing sensors suffice to provide a smooth and integrated experience on the user end, which would allow for relevant educational or entertainment applications to be adopted at scale. Nevertheless, the proposed approach would not be appropriate for tasks that require great precision and accuracy in measurement of movement and positioning of individual joints, such as medical or rehabilitation applications.

# Chapter 8

# Bidirectional Long Short Term Memory for Dance Sequences Analysis

## 8.1 Introduction

One important element in preserving folklore performing arts is, apart from digitization, modelling and documentation, the development of *an interactive framework* that enhances the learning procedure of folklore dances. The recent advances in depth sensors which have concluded to the development of low-cost 3D capturing systems, such as Microsoft Kinect [40] or Intel RealSense [213], have permitted easy capturing of human skeleton joints in 3D space which are then properly analyzed to extract dance kinematics [112]. Using the aforementioned low-cost capturing interfaces, we can build *interactive serious-game platforms* to allow for the users to achieve a rich learning experience [234], [235]. ML algorithms are necessary elements in this direction since they offer the technological tools for evaluating and comparing users' movement with predefined choreographic structures (patterns). The purpose of an ML tool is to spatio-temporally analyze the captured 3D human joints (and the respective kinematics features of them) in order to identify the main choreographic patterns which are then compared against targeted dance motives. These ML tools can provide robust systems that can identify primitive choreographic postures and be coupled with serious games platforms as monitoring mechanisms that ensure the achievement of the serious games' learning goals.

Towards this context in this chapter, an educational game platform has been deployed where, in real-time constraints, a Kinect sensor recognizes the dance movements and correlates them to a Labanotation system. Labanotation is a framework that translates the spatial and temporal fluctuation of 3D human joints (i.e., 4D dimension, 3D geometry plus time) into predefined signs [30]. Although Laban interactive platforms have been studied in the literature, the main drawback of them is that they focused on a dance representation and visualization, failing in providing methods for evaluating a dance performance. Examples of such Laban-based tools include the LabanEditor [236] that gives the opportunity to non-experienced users to understand their movements or the [56] where an embodied learning interface is introduced interweaving Kinect sensing and Labanotation. However, the main limitations of such methods is that they do not incorporate machine learning tools in order to extract the main choreographic patterns useful for dance evaluation. In this paper, *a deep learning algorithm* has been adopted to evaluate the performance of a dance in an interactive Laban-based game platform so as to provide to end-users capabilities of assessing their dance steps (motives) against predefined structures (patterns).

### 8.1.1 Related Work

The use of computer technology for model and digitization of folklore performing arts has been recently studied in scientific literature. The works can be distinguished into the ones dealing with 3D digitization, choreographic analysis, and Labanotation.

Regarding 3D digitization of performing arts, one of the first approaches is presented in [33]. Specifically, this work introduces a 3D archive system for Japanese traditional dances. In [34], a digitization approach for Cypriot dances using the Phasespace Impulse X2 motion capture system is proposed. The architectures utilizes 8-cameras that are able to capture 3D motion on modulated LEDs. In [38], the capturing architecture schema of the i-Treasure European Union funded project is analyzed targeting on 3D digitization and analysis of rare European folkloric dances. The main limitation of the aforementioned approaches is that they require a marker capturing framework and the capturing process fails to include choreographic metadata. The first limitation is addressed in [40], where 3D wireframe skeleton structures are extracted based on a markerless interface, reducing, however, the overall digitization accuracy. The second limitation is addressed in [35] and [36] where the captured motion trajectories are transformed into meaningful and semantically enriched LMA features.

As far as choreographic analysis is concerned, classification algorithms have been proposed in order to analyze the captured digitized 3D data and then to identify the human body kinessiology entities. More specifically, the work of [41] combines Principal Component Analysis (PCA) and two classification schemes (specifically a Gaussian mixture and a hidden Markov model) for dance movement classification. Additionally, a combination of PCA and Fischer's linear discriminant analysis, for classifying Korean pop dances is introduced in [42]. In this context, style analysis algorithms have been proposed in [237], exploiting principles drawn from Labanotation. The method leverages knowledge from anatomy, kinesiology and psychology as that is incorporated in the Laban Movement Analysis. Finally, the works of [44], [37] introduce a markerless tracking system for motion trajectory identification and folklore dance pattern interpretation, while the [39] proposes a real-time classification system in detecting choreographed gesture classes.

Recently, summarization methods have been introduced for a more precise and representative choreographic analysis. These methods are capable of abstractly modelling a folklore dance and they are distinguished into two main categories. The first group spatially analyses motion captured features, while the second group relies on temporal fluctuations of the descriptors in order to extract the key choreographic postures. As far as the first group is concerned, the work of [8] introduces a key posture extraction framework exploiting spatial classification algorithms, such as the k-means. Instead, the works of [178] and [59] performs selection of the main dancer's postures using temporal segmentation algorithms. Particularly, the work of [178] relies on a neighborhood graph to partition a dance sequence into distinct activities and motion primitives according to self-similar structures, while the work of [59] detects variations of the kinematic-based motion characteristics. The main limitation of a spatially based summarization algorithm is that temporal inter-relationships of a dance are lost. On the contrary, temporally analysis algorithms are highly sensitive to noise and dancer's micro-movement variations. The aforementioned drawbacks are addressed in [7] where a spatio-temporally enriched summarization algorithm is considered. Spatio-temporal decomposition of a dance improves precision of extracting the main choreographic primitives since one the one hand spatial clustering identifies major choreographic postures, while one the other hand, temporal analysis identifies micro choreographic dancer's movements. Spatio-temporal hierarchical algorithms are also considered in [181].

Regarding Labanotation, several methods have been proposed in the literature for transforming the captured 3D motion into Laban scores [238], [236], [239], [56]. The work of [238] can be considered as one of the first approaches for automatic Labanotation. Improvements in terms of performance and accuracy have been considered in the works of [236], [239]. Recently, serious game platforms [240], [56] have been proposed for providing a friendly interface for educational purposes. These interactive platforms have two forms of operations; to make the user familiar to the Laban scores and to provide an educational framework of folklore dances.

### 8.1.2 Innovation and Originality

In this chapter we enhance the learning experience of folklore dances by introducing machine learning tools with the capability of providing a scalable quantifiable assessment of a choreography at different level of hierarchies; yielding a from coarse to fine evaluation. For this reason, initially the choreography is analyzed into representative 3D skeleton joints and then kinematics features are estimated to efficiently model these choreographic patterns. Then, pose identification and summarization methods are implemented with the main purpose of categorizing each dance sequence into choreographic primitives or extracting the main (key) choreographic pattern. Pose identification provides a detailed (fine) assessment of a dance, which, in the sequel, stimulates an assessment of a dance performance against ground truth data. On the other hand, summarization creates a coarse representation (and thus assessment) of the choreography.

Pose identification is implemented using a deep learning Long-Short Term Memory (LSTM) network with bi-directional functionalities. The objective is to use depth data to create a robust automatic posture identification system. To this end, only depth information was used, so as to ensure that the classification performance is only affected by the exprert dancers kinesiological capturing and not by miscellaneous information such as picture color and texture. Existing methods in modelling a choreography assume causal signal dependencies. A system is called causal when its outputs depend only on the past and the current input samples but not on future inputs. It is clear that a choreographic posture depends not only on the past and current dancers actions (steps) but also on future kinematic activities. For this reason, bi-directional forms of LSTM networks are adopted allowing both past (backwards) and future (forward) states to interact with the pose identification outputs.

As far as the choreographic summarization is concerned, the SMRS method is used, appropriately modified to support a hierarchical modelling of the dance sequence. in this way, the system is capable to assess the performance of a dancer at different levels of hierarchies. The proposed serious game platform supports Labanotation. This allows for dance professionals to qualitative evaluate a performance, to document the whole choreography and finally to recommend correction actions. It is clear that

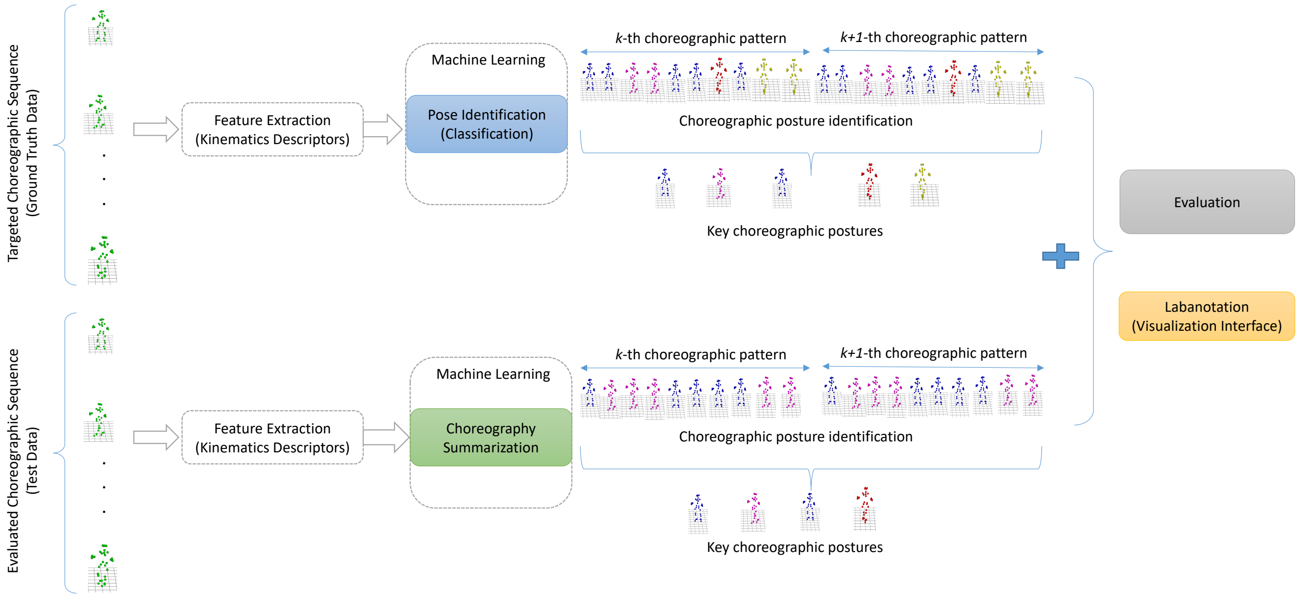

Targeted Choreographic Sequence (Ground Truth Data)
Evaluated Choreographic Sequence (Test Data)
Feature Extraction (Kinematics Descriptors)
Machine Learning
Pose Identification (Classification)
Choreography Summarization
k-th choreographic pattern
k+1-th choreographic pattern
Choreographic posture identification
Key choreographic postures
Evaluation
Labanotation (Visualization Interface)

Figure 8.1 The proposed system architecture for the interactive serious game platform incorporating machine learning for educational purposes.

these recommendations take into consideration the scalable quantitative metrics of the machine learning module. Finally, visualization tools are discussed with the capability of depicting dance performance in 3D skeleton joints and of encoding the dancer movements into Laban scores.

## 8.2 Educational System Architecture with ML cababilities

Fig.8.1 illustrates the main components of the proposed educational interactive serious game platform that incorporates machine learning techniques for choreographic postures identification and summarization. As is observed, the architecture is composed into the following subsystems.

1. ***Feature Extraction***: The purpose of this module is to encode the captured 3D skeleton points into kinematics descriptors for efficient representation of the choreography. In this approach the velocity and the acceleration of the dancers' skeleton joints are taken into account [112].

2. ***Machine Learning***: The use of this subsystem is to analyze the choreographic features, as derived from the feature extraction module, in order to provide a semantic encoding of the dance sequence. This module is discerned into two processes; *pose identification* and *choreographic summarization*. Pose identification incorporates deep learning classifiers [94], [219] and particularly bidirectional LSTM networks in order to classify each choreographic frame into distinct pose entities. The LSTM classifier feeds as inputs the kinematics features of 3D skeleton joints over a window of $p$ frames. On the other hand, the choreographic summarization module aims at processing the whole choreographic sequence and extract the key postures, that abstractly encode the dance. Dance summarization is performed using the SMRS on the kinematics features of the 3D skeleton joints. The aforementioned two ML components provide a scale-based representation of the choreography. In particular, pose identification provides a fine encoding of each choreographic frame instead summarization derives a rough representation of the dance sequence. This scale-based representation is

very important for educational purposes. Specifically, rough representation (through choreographic summarization) is actually an indicator of the overall users' performance to a given (ground truth) choreography. On the contrary, fine representation (through pose identification) actually provides a detailed evaluation, depicting additionally micro errors.

3. ***Evaluation***: The main purpose of this module is to incorporate objective metrics for comparing the test choreographic sequence against the ground truth one. Evaluation is performed over the classified postures (provided by pose identification), and the summarized choreographic entities (provided through choreographic summarization). Inevitably, ML techniques provide a high-level semantic representation of the choreographic sequence eliminating noisy effects in directly processing 3D skeleton data.

4. ***Labanotation (Visualization Interface)***: The final stage of the proposed interactive serious game platform is a Laban visualization engine with a main objective to transforming the detected choreographic entities into Laban scores [30]. Labanotation allows documentation and evaluation the of the whole procedure by the dance experts. Thus, it enables appropriate recommendation strategies.

## 8.3 Pose Identification

### 8.3.1 Feature extraction

Principles from the theory of rigid body dynamics [140] are exploited as far as the feature extraction process is concerned. In particular, the $xyz$ coordinates of each skeleton joint are transformed into the respective joint velocity and acceleration. More specifically, we have that $\vec{u}_k(t) = d\vec{s}_k^L(t)/dt$ as regards the velocity vector and $\vec{\gamma}_k(t) = d\vec{u}_k(t)/dt$ as regards the acceleration vector for the $k$-th joint. In this way, each choreographic frame $t$ is represented by $M$ feature vectors each of the form $\vec{f}_k(t) = [\vec{s}_k^L\ \vec{u}_k^L\ \vec{\gamma}_k^L]^T$ [112]. Therefore, the kinematics of the whole choreographic video sequence is given by a matrix $\vec{F}(t) = [\vec{f}_1^L(t) \cdots \vec{f}_M^L(t)]$.

### 8.3.2 Pose identification using Long-Short Term Memory (LSTM) Networks

Let us assume that we have $L$ available choreographic poses. Then, each frame $t$ is categorized to one, out of $L$ available, class according to the probabilities values $p_i(t)$, with $i = 1, \ldots, L$. Actually $p_i(t)$ expresses the probability that frame $t$ belongs to the $i$-th class. Particularly, we have that

$$\hat{c}(t) = \arg \max_{i \in 1, \ldots, L} p_i(t) \tag{8.1}$$

where $\hat{c}(t)$ expresses the class (i.e., the specific pose) that the $t$ frame belongs to. In the following, we denote as $\vec{p}(t) = [p_1(t) \ldots p_L(t)]^T$ a probability vector including all $p_i(t)$ at a image frame $t$.

Pose $\hat{c}(t)$ is a non-linear relationship of the 3D skeleton joints as the well as of the respective kinematic features $\vec{F}(t)$. However, for noise removal purposes and for the making pose identification process robust and stable with respect to time fluctuations, a non-linear moving average model is adopted. In statistics, a

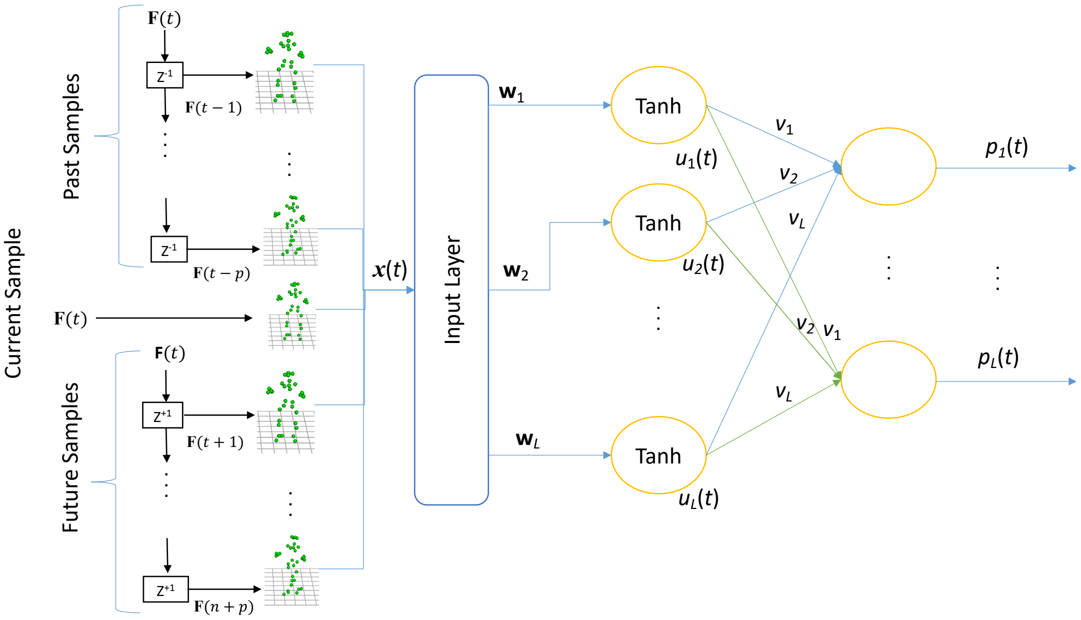


Figure 8.2 A feedforward neural network for modelling the unknown relationship of Eq. (8.2).

Moving Average (MA) filter predicts the value of a time series by taking into consideration responses over a time window $2 \cdot p + 1$, expressing the order of the model. It is clear that the choreographic posture at an image frame $t$ depends not only on the dancer movements at this and past frames but also on future samples. This means that the future dancer movements affects the current choreographic postures classification. Therefore, we have that

$$\vec{p}(t) = \vec{g}\Big(\vec{F}(t), \ldots, \vec{F}(t-p), \vec{F}(t+1), \cdots, \vec{F}(t+p)\Big) \tag{8.2}$$

The main difficulty in implementing Eq. (8.2) is that the non-linear vector valued function $\vec{g}(\dot{})$ is actually unknown. It has been proven, however, that a feedforward neural network with a Tapped Delay Line (TDL) input filter is able to approximate the model of Eq. (8.2) with any arbitrarily accuracy [241]. Fig. 8.2 presents the architecture of a feedforward neural netowrk for modelling the unknown relationship of Eq. (8.2). Mathematically speaking, the network models the probability vector $\vec{p}(t)$ as a relationship of $L$ hidden (latent) state units $u_i$.

$$\vec{p}(t) = \vec{u}^T(t) \cdot \vec{v}$$

$$\vec{u}(t) = \begin{bmatrix} u_1(t) \\ \vdots \\ u_L(t) \end{bmatrix} = \begin{bmatrix} tanh(\vec{w}_1^T \cdot \vec{x}(t)) \\ \vdots \\ tanh(\vec{w}_L^T \cdot \vec{x}(t)) \end{bmatrix} \tag{8.3}$$

In Eq. 8.3, vector $\vec{x}(t)$ denotes the input vector generated from the matrices $\vec{F}(t)$ after being vectorized. The outputs of the hidden neurons $u_i$ refers to a state, hidden vector regarding pose identification. Variables $\vec{w}_i$ are appropriate weight vectors derived from the training phase. These vectors regulates the importance of each input element $\vec{x}(t)$. Function $tanh(\cdot)$ denotes the hyperbolic tangent function. This means that each hidden state $u_i$ takes value between $[-1 \quad +1]$; values +1 indicates that the respective

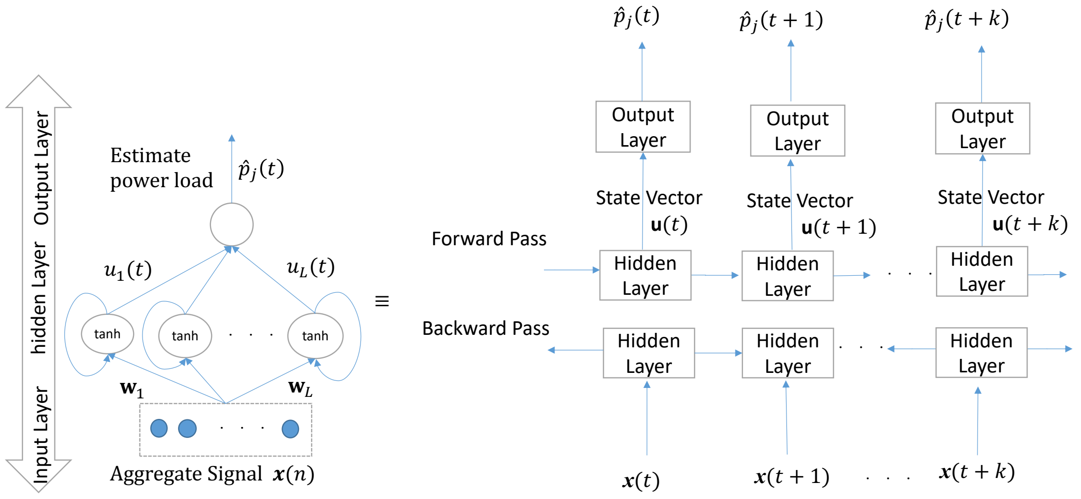


Figure 8.3 A bi-directional recurrent architecture unfolded through time to reveal the backward and forward passes.

hidden state contributes to the pose identification output, while values of -1 refers to no contribution. Finally, the identification outcome is a linear relationship of the hidden state vector $\vec{u}$ weighted by $\vec{v}$.

An alternative approach to model the non-linear relationship of Eq. (8.2) is to allow the hidden state variables to depend on either previous and future state values. In this way, we re-formulate the Eq. (8.3) in a way

$$u_i(t) = tanh(\vec{w}_i^T \cdot \vec{x}(t) + \vec{r}_{i,b}^T \cdot \vec{u}_i(t-1) + \vec{r}_{i,a}^T \cdot \vec{u}_i(t+1)) \tag{8.4}$$

where variables $\vec{r}_{i,b}$ refers to the weights regulate the backward pass of the network and $\vec{r}_{i,a}$ the forward pass.

Fig. 8.3 depicts the structure of a bi-directional recurrent neural network model [242]. In this figure, we have unfolded the network to reveal the backward and forward passes. It is clear that this type of network implements the relationship of Eq. (8.4).

### 8.3.3 Modelling the Long-range dependencies using LSTM architectures

The main limitation of the aforementioned modelling framework is that it fails to represent long-range dependencies. However, a choreography usually follows repeated patterns spanning over long-time periods. For this reason, bi-directional LSTM network is adopted for modelling the pose identification module. LSTMs are of similar structure to the bi-directional recurrent regression models but each node in the hidden layer is replaced by a memory cell, instead of a single neuron [243]. The structure of a single memory cell is depicted in Fig. 8.4 .

The memory cell contains the following different components (see Fig. 8.4): i) *the input node*, ii) *the input gate*, iii) *the forget gate* and v) *the output gate*. Each component applies a non-linear relationship on the inner product between the input vectors and respective weights (estimated through the training

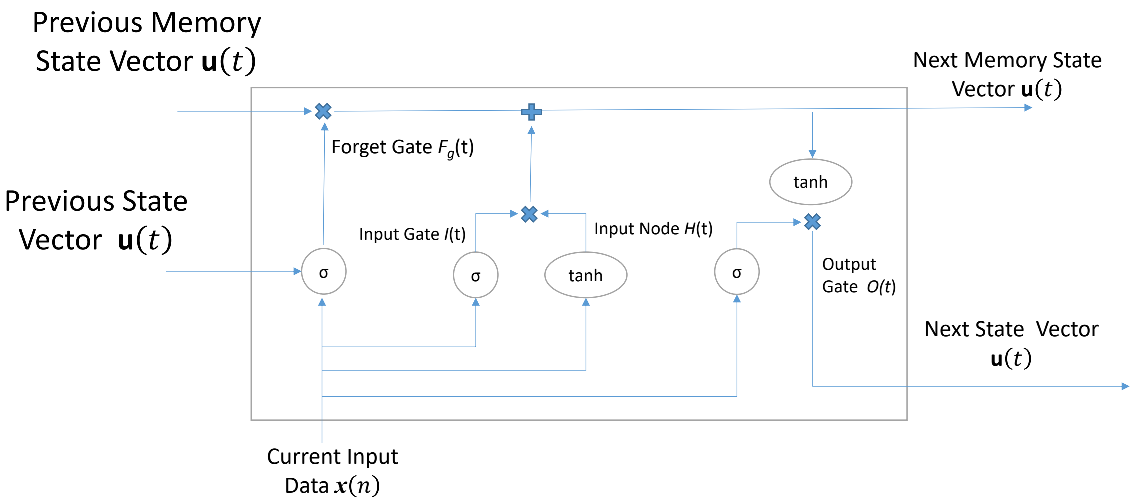

Previous Memory State Vector $\mathbf{u}(t)$
Next Memory State Vector $\mathbf{u}(t)$
Forget Gate $F_g$(t)
tanh
Previous State Vector $\mathbf{u}(t)$
Input Gate $I$(t)
Input Node $H$(t)
σ
σ
tanh
σ
Output Gate $O$(t)
Next State Vector $\mathbf{u}(t)$
Current Input Data $\boldsymbol{x}(n)$

Figure 8.4 The architecture of the memory cell for the Long-Short Term Memory (LSTM) network.

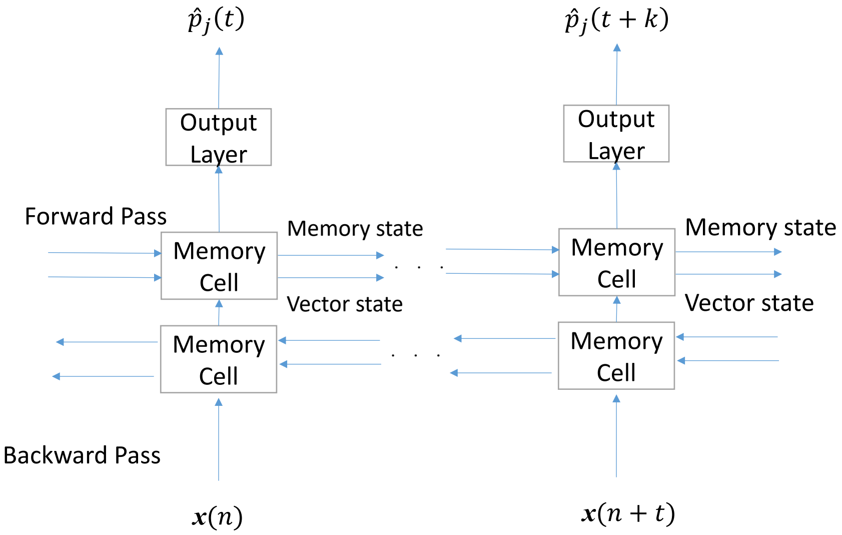


Figure 8.5 The architecture of Bi-directional LSTM used for pose identification.

process). Some of the components have the sigmoid function, expressed as $\sigma$ in Fig. 8.4 and some other the *tanh*.

*The forget gate*: The purpose of this component is to decide what information is throw out of the memory cell. The output ranges between 0 and 1, due to the sigmoid activation function. Values close to 0 means to dispose the incoming information while values close to 1 indicates that this information should be taken into consideration by the current memory cell.

*The input node*: The input node performs the same operation as a hidden neuron of a conventional recurrent regression model does. It appropriately combines (through a set of weights) the current input data and the previous vector states in order to decide whether the respective hidden state (latent variable) contributes or not (true or false) to the respective choreographic posture estimate.

*The input gate*: This gate regulates whether the respective hidden state is significant enough for the accurate estimation of current choreographic pose. It has the sigmoid function, meaning that its response range between 0 and 1. Values close to zero mean that this state is not significant at the respective time interval. The opposite happens for values close to one. This gates actually addresses problem related to the vanishing of the gradient slope of a tanh operator [243].

*The output gate*: This regulates whether the response of the current memory cell is "significant enough" to contribute to the next cell.

Fig. 8.5 illustrates the architecture of the proposed bi-directional LSTM for dance pose identification. The network includes backward and forward time instances to categorize the poses.

## 8.3.4 Extraction of dance sequences key-frames

The Sparse Modelling Representative Selection (SMRS) algorithm [57] extended in a way to support hierarchical implementation is adopted for choreographic summarization. The hierarchical implementation allows for a spatio-temporal extraction of key choreographic postures in contrast with the conventional SMRS algorithm where only spatial selection is considered. The spatial based modelling algorithms fail to index the temporal variations and the frame inter-relationships of the dance.

### 8.3.4.1 The Sparse Modelling Representative Selection

In this section, we briefly describe the SMRS algorithm used as a baseline for key choreographic postures selection. First, we vectorize the features $\vec{F}(t)$ by stacking up all rows. Therefore, we have that $\vec{d}(t) = vec(\vec{F}(t))$. Let us now denote as $\vec{D} = [\cdots \vec{d}(t) \cdots]$ a matrix that includes all vectorized features elements $\vec{d}(t)$ of the whole choreography (i.e., $\forall t$). The purpose of a summarization algorithm is to select a set of $N \ll Q$ representatives that best reconstruct the whole choreography (variable $Q$ refers to the total number of frames and $N$ to the extracted key postures). This is accomplished using the following equation.

$$\sum_{i=1}^{Q} \| \boldsymbol{d}(t_i) - \mathbf{D} \cdot_i \| = \|\mathbf{D} - \mathbf{D} \cdot \mathbf{C}\| \tag{8.5}$$

In Eq. (8.5), $\vec{c}_i$ is a coefficient vector regulates the similarity for every feature vector $\vec{d}(t)$ [57]. In order to estimate the best $N$ choreographic postures, we enforce the following constraint of the matrix $\vec{C}$, that is, $\|\vec{C}\|_0 \leq N$. Norm $\| \cdot \|_0$ counts the number of non-zero rows of matrix $\vec{C}$. Minimization of

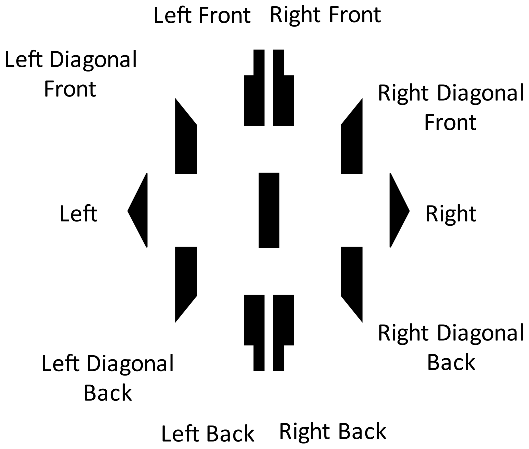


Figure 8.6 The eleven position of LABAN Motion Analysis used to represent a dancer's movement at the horizontal axis.

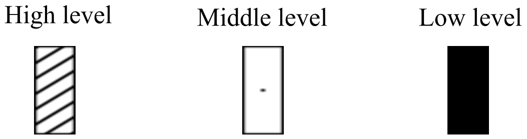


Figure 8.7 The three main categories used by the Labanotation interface to model a dancer's movement in vertical axis.

Eq. (8.5) subject to the constraint of $||\vec{C}||_0 \leq N$ is a NP-hard problem [57]. For this reason, we relax the hard constraint into an $\ell_1$-norm, that is $||\vec{C}||_1 \leq \tau$. In this case, we select $\tau$ instead of $N$, since $\ell_1$-norm is not necessarily bounded by $N$. The Alternating Direction Method of Multipliers (ADMM) of [161] is adopted for solving Eq. (8.5) subject to constraint $||\vec{C}||_1 \leq \tau$.

#### 8.3.4.2 Hierarchical sparse modelling

A hierarchical implementation of the SMRS algorithm is adopted for key choreographic postures extraction. In particular, first the SMRS algorithm is applied on the whole choreographic data. In this way, a set of key representative frames is extracted, expressing specific time instances of the choreography. Then, the detected time intervals are further decomposed to create hierarchies of key postures. This is accomplished by applying the SMRS algorithm on the created sub-time intervals (expressed by the key postures of the previous layer of processing). This results in a spatio-temporal summarization scheme. More specifically, the first layers of key postures show a coarse (not accurate) representation of the choreography. Instead, the last layers provide a more detailed (fine) representation [7].

## 8.4 The Labanotation Interface

### 8.4.1 A dancer's movement representation

The Labanotation visualizes the kinessiology variations of a dance. It is like a music score that describes a song. Laban motion analysis uses pre-determined symbols that model the motion attributes of a dancer. In

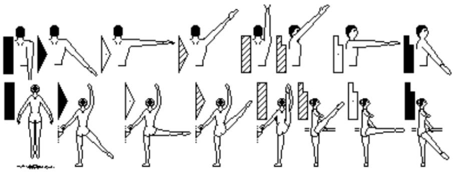


Figure 8.8 The adopted Laban codes combining both the horizontal and vertical axis (3D space). In particular, this example combines the signs of both Figs. 8.6 and 8.7 respectively [9].

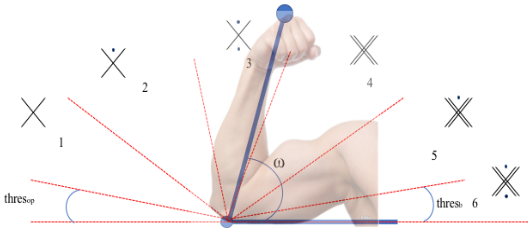


Figure 8.9 The six different signs adopted for modelling the bending angles of a dancer's joint.

the proposed Labanotation interface, basic Laban symbols have been created including points of direction as well as the respective bending (systole) degrees. The symbols of Laban have been created using the software platform of Inkscape. We have created eleven different positions as far as the horizontal axis is concerned and regarding a specific human joint out of $M$ available. These positions are depicted in Fig. 8.6. We denote these symbols as follows: the Left (L), Left Diagonal Front (LDF), Left Diagonal Back (LDB) and the Left Front/Back (LF/LB) positions. In the same manner, we have the Right Front (RF), the Right Diagonal Front (RDF), the Right (R), the Right Diagonal Back (RDB) and the Right Back (RB) codes. Apart from the horizontal representation, a dancer's movement is modelled as far as the vertical axis is concerned (see Fig. 8.7). We have used three main categories for such vertical modelling. Fig. 8.8 illustrates the Laban symbols in 3D space, by combining both the horizontal and vertical axis.

### 8.4.2 A dancer's bending (systole) angles representation

Apart from the representation of a dancer's movements, we need to model the systole angle. The adopted encoding framework is depicted in Fig. 8.9. In particular, the whole bendable territory, is divided into equal-sized spaces corresponding to the bending degrees. The first symbols defines zero degree of bending while the last one 180 degrees. Fig. 8.9 clarifies the bending degrees symbols according to the humans' joint systole.

## 8.5 The Evaluation Interface

As far as the assessment of the choreography is concerned, initially a professional dancer is recorded by the interactive serious game platform. The extracted 3D skeleton joints as well as the respective features are fed into the pose identification deep learning module. This way, each image frame is categorized into one into of the $L$ available postures, $c_p(t)$. In the next step, a non-professional user is recorded. Again 3D skeleton joints are extracted along with the respective features. This information is then fed into the same pose identification architecture for assessing the choreography of the non-experienced user. In particular, let us denote as $c_u(t)$ the estimated postures of the user sequence $t$ (test sequence) for every frame $t$.

In this chapter, a scalable evaluation framework for choreography assessment is adopted. In particular, the output of the summarization module provides an overall (coarse) evaluation of the dancer's performance. On the other hand, the pose identification module is responsible for a more detailed dance assessment. Regarding, the fine assessment process each frame of the test dance sequence is compared against the professional dance sequence. Therefore, an detailed performance error $E_d$ is defined as follows:

$$E_d = \|c_p(t) - c_u(t)\|_2 \tag{8.6}$$

In this same context, we define the coarse performance error from the summarized choreographic elements. More specifically, the test sequence is fed to the summarization module. This module is responsible for extracting a set of key postures, $c_u(t_i^l)$. In this notation, variable $t_i$ refers to the time instance of the $i$-th key postures at the $l$ hierarchy. It should be mentioned that the higher a hierarchy is the more choreographic postures are extracted and therefore a more detailed assessment is considered. In this case, the coarse performance error is defined as:

$$E_c^l = \|c_p(t_i^l) - c_u(t_i^l)\|_2 \tag{8.7}$$

where $t_i^l$ refers to time instances where the extracted key postures are identified at layer $l$. We recall that as we increase the level of hierarchy $l$ a more detailed choreography assessment is built (i.e., for large $l$ the error $E_c^l$ is close to the $E_d$). Therefore, we result to a scalable assessment framework.

## 8.6 Experiments

In this section, we present the evaluation test-bed used for assessing the proposed interactive serious game platform for dance learning.

### 8.6.1 Dataset description

We use data sets of TERPSICHORE project [29]. The data sets contain recordings from Greek traditional folklore dances, performed by professionals. Five different folklore choreographies have been recorded each is performed by three experts (two male and one female). We chose male and female expert-dancers since for those particular dances, the choreographic performance between men and women is 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. In all recording a Kinect-II sensor has been exploited for creating 3D skeleton joints. In Section 3.5, we presented a detailed description of the recorded dances.

### 8.6.2 Performance Evaluation Metrics

To objectively evaluate dance performance, we adopt four different metrics such as the *Accuracy*, *Precision*, *Recall* and the combination of Precision and Recall criteria as a single metric, the *F1-Score*. Precision measures the ratio of all relevant retrieved key frames over the total number of retrieved key frames by the use of an algorithm. Recall measures the ratio of all all relevant retrieved key frames over the total number of relevant frames in ground truth set. The main challenge in defining the precision and recall metrics in our cases is that key frames from a dance sequence should be ordered. This is due to the fact that the patterns composed of the main steps of a choreography should be specific for a given time internal depending on the music tempo and the type of the dance.

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive + False Positive}} \tag{8.8}$$

and recall as

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive + False Negative}} \tag{8.9}$$

where variable $N$ refers to the number of key frames and $L$ to the number of ground truth choreographic elements. Ideally, Pr and Re should be 1 for an excellent retrieval. By combining both criteria, we can derived the F1-score as

$$\text{F1} = 2 * \frac{\text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8.10}$$

Similarly, Accuracy is defined as the ratio

$$\text{Accuracy} = \frac{\text{TP}_{+}\text{TN}}{\text{TP + FP + TN + FN}} \tag{8.11}$$

### 8.6.3 Training/Testing Evaluation Methodology

In this section, we present the methodology adopted in this paper to construct the training and the testing datasets through which the classifiers of section are assessed. The initial dataset consists of 8149 frames of expert dancers performing the choreographies described in Table 1. This includes all the joints captured by the Kinect sensor. The Spine Base joint is used to transform of all other body joints from a global system to a local one, beginning in the Spine Based joint. Thus the position of the joints is not sensitive to the dancer's position in respect with the Kinect sensor. To simulate non-expert performances, the dataset was augmented by adding noise both to the axial and lateral measurements. In particular, the existing data set was used as base for the creation of an additonal observations by adding random noise. The upper bound of the noisy data was set as 10 and 20% of the original values respectively. Thus we create a new syntehtic dataset of both expert and non-expert performances consisting of 24447 frames of Kinect data. This dataset is broken down in training and test datasets following the 80-20 rule. 20% of each "noisy" capturing is used for testing while the remaining dataset is used for training. 10% of the training dataset is used for cross validation purposes. The initial groundtruth dataset includes performance by both men and women professional dancers. This is due to the fact that the kinesiological capturing of choreography

Table 8.1 Performance evaluation of the proposed LSTM model for pose identification compared with other learning methods.In this table, we have provided the effect of memory, that is number of dance frames feeding the network.

| Algorithms | Method | Metrics | | | |
|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F1-score |
| SVM | No Memory | 41,34% | 31,04% | 36,33% | 33,48% |
| | 5 Frames | 50,19% | 39,77% | 46,98% | 43,08% |
| | 10 Frames | 69,15% | 59,03% | 62,71% | 60,82% |
| kNN | No Memory | 21,46% | 13,26% | 16,23% | 14,60% |
| | 5 Frames | 31,96% | 24,94% | 32,78% | 28,33% |
| | 10 Frames | 34,57% | 26,63% | 34,12% | 29,92% |
| Neural 1 | No Memory | 57,14% | 46,32% | 51,96% | 48,98% |
| | 5 Frames | 55,66% | 45,44% | 58,14% | 51,01% |
| | 10 Frames | 59,68% | 49,08% | 64,67% | 55,81% |
| Neural 2 | No Memory | 60.56% | 49,83% | 59,79% | 54,36% |
| | 5 Frames | 59,09% | 48,53% | 62,26% | 54,55% |
| | 10 Frames | 62,36% | 51,47% | 63,92% | 57,03% |
| CNN | No Memory | 57,27% | 49,43% | 60,88% | 54,56% |
| | 5 Frames | 69,99% | 65,15% | 65,05% | 65,10% |
| | 10 Frames | 74,47% | 72,65% | 65,96% | 69,14% |
| LSTM | No Memory | 59,35% | 48,62% | 66,45% | 53,46% |
| | 5 Frames | 76,20% | 66,98% | 69,74% | 71,30% |
| | 10 Frames | 81,06% | 74,22% | 71,20% | 77,49% |

slightly changes (position of hands, horizontal movement in comparison to the Kinect sensor) based on the gender of the performer.

Regarding the performance of the proposed interactive serious game platform in assessing the choreography of a non-expert user. For this reason, initially a annotated choreography is loaded by the system. This choreography has been performed by a professional dancer. Moreover, the main choreographic elements are available such as the ones depicted in Table 3.5. We assume that each choreography is associated with a given tempo to avoid synchronization issues. In the following, evaluation test-bed is performed either for pose identification or for choreographic summarization, providing, therefore, a scalable assessment framework.

### 8.6.4 Pose identification

First, a Long Short Term Memory Network (LSTM) is trained to learn a specific choreographic pattern. In particular, the input of the network is the kinematics features extracted from the 3D-skeleton joints of Kinect-II. Table 8.1 depicts the performance of the proposed LSTM model for estimating the main choreographic primitives of a dance. In this table, we have used as evaluation metrics the Precision, Recall and F1-score used in information retrieval [244]. A comparison with different shallow learning paradigms is also presented. Particularly, we have compared the proposed bi-directional LSTM with a) a Support Vector Machine (SVM), b) k-Nearest Neighbor (kNN), c) two feedforward neural networks (of

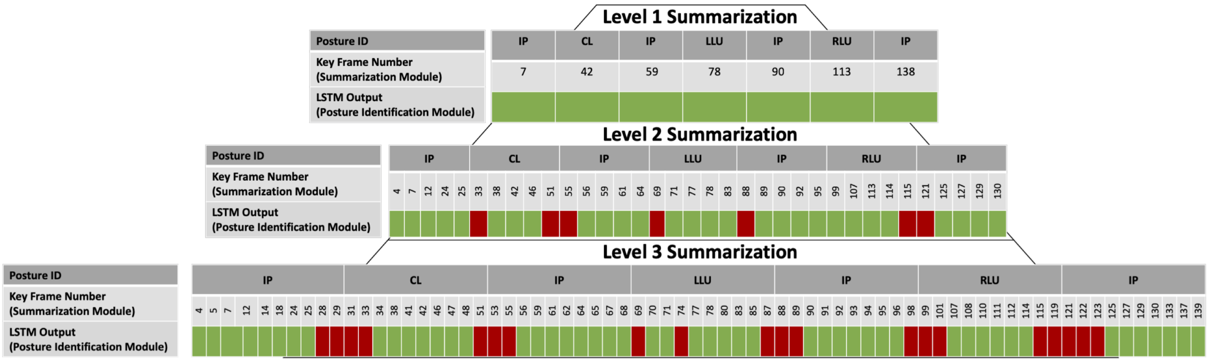
Table 8.2 The effect of the proposed pose identification for performing a detailed assessment score on a choreography against a target.

| Algorithms | Method | Metrics | | | |
|---|---|---|---|---|---|
| | | Accuracy | Precision | Recall | F1-score |
| SVM | No Noise | 69,15% | 59,03% | 62,71% | 60,82% |
| | 10% Noise | 55,18% | 65,52% | 38,85% | 48,78% |
| | 20% Noise | 25,87% | 25,84% | 18,65% | 21,67% |
| kNN | No Noise | 34,57% | 26,63% | 34,12% | 29,92% |
| | 10% Noise | 23,33% | 24,86% | 19,52% | 21,87% |
| | 20% Noise | 8,01% | 7,80% | 6,23% | 6,93% |
| Neural 1 | No Noise | 59,68% | 49,08% | 64,67% | 55,81% |
| | 10% Noise | 52,21% | 58,40% | 45,24% | 50,99% |
| | 20% Noise | 32,53% | 34,87% | 26,23% | 29,95% |
| Neural 2 | No Noise | 62,36% | 51,47% | 63,92% | 57,03% |
| | 10% Noise | 57,03% | 65,18% | 46,72% | 54,43% |
| | 20% Noise | 40,64% | 44,53% | 32,70% | 37,72% |
| CNN | No Noise | 74,47% | 72,65% | 65,96% | 69,14% |
| | 10% Noise | 60,34% | 54,34% | 61,51% | 57,70% |
| | 20% Noise | 54,28% | 46,50% | 59,60% | 52,24% |
| LSTM | No Noise | 81,06% | 74,22% | 71,20% | 72,69% |
| | 10% Noise | 75,30% | 66,53% | 54,31% | 59,81% |
| | 20% Noise | 56,15% | 65,61% | 42,37% | 51,49% |

two configurations- one with a single hidden layer of 10 neurons/layer and one with two hidden layers, each with 10 neurons/layer) and d) on Convolutional Neural Network (CNN) (with 1 convolutional and one fully connected layer).

As is observed, the proposed bi-directional LSTM model for pose identification outperforms the compared shallow learning paradigms. In this table, we have also presented the effect of memory in terms of the number of dance frames that feed the learning model, i.e., the effect of the tapped delay line filter, on classification performance. As is observed, the tapped delay line filter increases pose identification performance. It should be mentioned that for the evaluation test-bed framework, we have used professional dancers. The model has been trained using data from a given dance realization and it is tested using data from other realizations of the same dancer. It is worth noting that the performance for the medium memory window (5 frames) slightly decreases for the case of the Feedforward Neural Networks 1 and 2. These two networks have a rather simple structure therefore this slight degradation in performance could be put down to their tendency to overfit and their limited ability to generalize as opposed to the more complex architectures (including deep ones) examined.

Then, a non-expert user performs the given choreographic sequence. Again, the kinematics features, extracted from the 3D skeleton joints, and their descriptors are the inputs to the LSTM model. The purpose of the network is to categorize the choreography of a simple user into choreographic basic units defined by the experts, i.e., the annotated choreography. In order to set-up this experiment, a noise is added to the 3D skeleton joints of the ground truth data. Specifically, the additive noise models the errors



Level 1 Summarization

| Posture ID | IP | CL | IP | LLU | IP | RLU | IP |
|---|---|---|---|---|---|---|---|
| Key Frame Number (Summarization Module) | 7 | 42 | 59 | 78 | 90 | 113 | 138 |
| LSTM Output (Posture Identification Module) | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |

Level 2 Summarization

| Posture ID | IP | | | | | CL | | | | | IP | | | | | LLU | | | | | IP | | | | | RLU | | | | | IP | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Key Frame Number (Summarization Module) | 4 | 7 | 12 | 24 | 25 | 33 | 38 | 42 | 46 | 51 | 55 | 56 | 59 | 61 | 64 | 69 | 71 | 77 | 78 | 83 | 88 | 89 | 90 | 92 | 95 | 99 | 107 | 113 | 114 | 115 | 121 | 125 | 127 | 129 | 130 |
| LSTM Output (Posture Identification Module) | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 |

Level 3 Summarization

| Posture ID | IP | | | | | | | | | | CL | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Key Frame Number (Summarization Module) | 4 | 5 | 7 | 12 | 14 | 18 | 24 | 25 | 28 | 29 | 31 | 33 | 34 | 38 | 41 | 42 | 46 | 47 | 48 | 51 |
| LSTM Output (Posture Identification Module) | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 |

| Posture ID | IP | | | | | | | | | | LLU | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Key Frame Number (Summarization Module) | 53 | 55 | 56 | 59 | 61 | 62 | 64 | 65 | 67 | 68 | 69 | 70 | 71 | 74 | 77 | 78 | 80 | 83 | 85 | 87 |
| LSTM Output (Posture Identification Module) | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 |

| Posture ID | IP | | | | | | | | | | RLU | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Key Frame Number (Summarization Module) | 88 | 89 | 90 | 91 | 92 | 93 | 94 | 95 | 96 | 98 | 99 | 101 | 107 | 108 | 110 | 111 | 112 | 114 | 115 | 119 |
| LSTM Output (Posture Identification Module) | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟥 | 🟥 |

| Posture ID | IP | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Key Frame Number (Summarization Module) | 121 | 122 | 123 | 125 | 127 | 129 | 130 | 133 | 137 | 139 |
| LSTM Output (Posture Identification Module) | 🟥 | 🟥 | 🟥 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 | 🟩 |

Figure 8.10 A schematic representation of the hierarchical summarization approach adopted. As the level of summarization increases the number of extracted key frames that provides an abstract representation of the choreography also increases. Thus, a more detailed representation of the dance is derived.

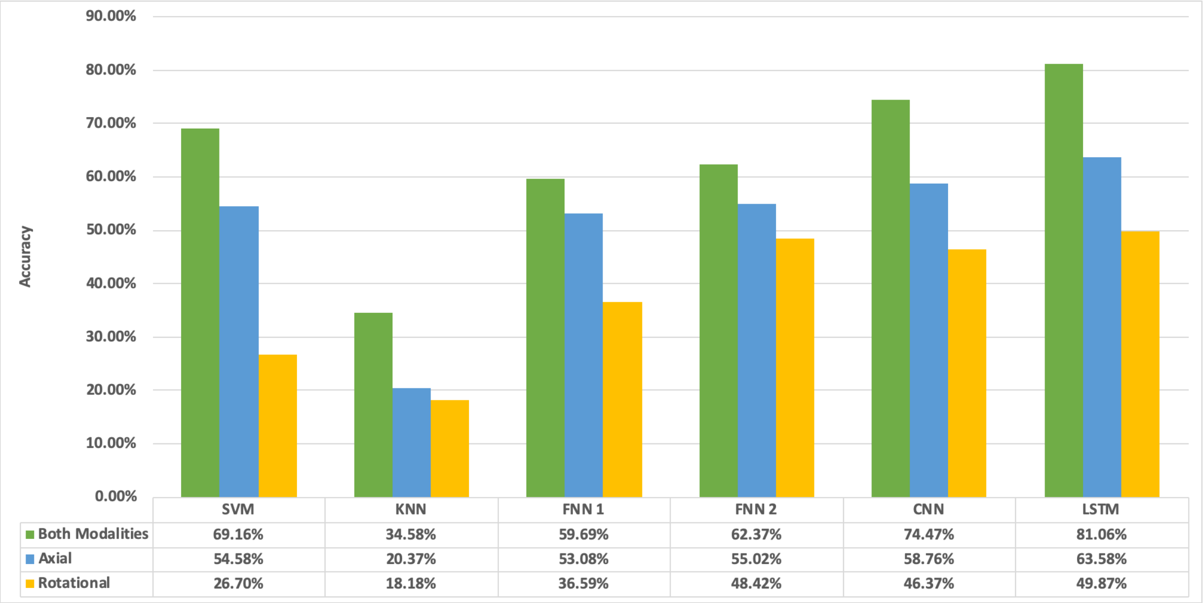
done by the non-expert user as far as the given choreography is concerned. It is clear that the more noise we add the worse a choreography is. Table 8.2 presents the performance of the pose identification module versus different levels of noise. As is observed, the more the noise we add, we worse the classification accuracy is. However, the performance is better for the LSTM model in the sense that it retains a robust behavior against the noise added. Therefore, it is better for assessing a choreography of non-expert user against a target choreography of an expert dancer. The bi-directional LSTM model retains a more robust behavior against noise. Therefore, it better models the mistakes happened in a choreography by non expert dancers. We also observe better generalization performance over noise compared to other shallow learning networks. This is an important aspect in evaluating a choreography mainly due to the fact that the shallow learning modules fails to provide a proportional evaluation score of the choreographic errors.

### 8.6.5 Comparative results against different feature selections

Three different sets of features were used to evaluate the performance of the proposed classifier. Firstly, the classifier was tested using only the axial data of the depth sensor. Secondly, only the rotation data were used, and finally all the features were used. Figure 8.11 depicts the results of such feature selection. It is clear that selecting both axial and rotation data from the depth sensor is better than using only one of the depth parameters.

### 8.6.6 Choreographic Main Primitives Estimation

The aforementioned assessment provides a detailed evaluation of the choreography of a non-expert user compared to a ground truth dataset, providing an evaluation score per dance frame. However, usually, a generic assessment is necessary for learning purposes. This way, the system provides an abstract recommendation of the performance of a sequence. In order to do this generic evaluation, we exploit results of the choreographic summarization module. This module is able to extract a set of key (main) choreographic patterns (motives) that best model the whole choreography. The assessment in this case is performed as follows. First, a non-expert user performs the dance given a choreography. In our evaluation test-bed, the expert results are corrupted by noise added on the extracted 3D skeleton joints. Then, the



| | SVM | KNN | FNN 1 | FNN 2 | CNN | LSTM |
|---|---|---|---|---|---|---|
| 🟩 Both Modalities | 69.16% | 34.58% | 59.69% | 62.37% | 74.47% | 81.06% |
| 🟦 Axial | 54.58% | 20.37% | 53.08% | 55.02% | 58.76% | 63.58% |
| 🟨 Rotational | 26.70% | 18.18% | 36.59% | 48.42% | 46.37% | 49.87% |

Figure 8.11 Accuracy of different learning models versus the different feature selection methodologies.

algorithm of Section 8.3.4.2 is applied to detect a set of key frames organized in an hierarchical manner. These frames are compared against the ones provided by the CNN pose identification module and the error of Eq.(8.7) is estimated. Fig. 8.10 shows a schematic representation of the proposed hierarchical dance summarization approach. At level 1, only the most salient dance poses are selected as key frames to abstractly model the whole choreography. At higher summarization levels, more key dance poses are selected to refine the choreographic representation through key frame extraction. Table 8.3 depicts the error achieved, as far as dance assessment is concerned, for different summarization levels. Initially, only seven frames have been extracted to represent the choreography. The assessment error over all these seven frames is 100% meaning that the non-professional user has carried out correctly the main steps of the choreography. At higher levels (level 2, and 3), more key frames are detected, (35 and 70 respectively). In this case, we observe that the performance error is reduced meaning that the non-expert user makes some mistakes in executing the choreography. Therefore, the summarization module provides to the interactive serious game platform a coarse to fine assessment useful for learning purposes. This fine to coarse assessment framework is depicted in Fig. 8.10. Particularly, in this figure, we depict the extracted key frames along with the pose identification performance. Green cells indicate that the detected pose is in accordance to the ground truth choreography (no error is accomplished). Instead, red cells indicates the time instances where mistakes in the choreography are encountered. Using these red cells, the performance error is computed.

## 8.7 Visualization interface

The visualization system is responsible for creating a LABAN modelling of a choreography. This is important for delivering the performance of the simple user to dance experts for further recommendation, suggestions and corrections. Fig. 8.12 shows the main menu developed for this particular educational

Table 8.3 Caption Performance error with respect to the number of hierarchical levels used for the summarization.

| Level of Detail | Number of Frames | Correctly Classified | Performance (%) |
|---|---|---|---|
| Level 1 | 7 | 7 | 100% |
| Level 2 | 35 | 28 | 80% |
| Level 3 | 70 | 50 | 71,43% |

game. The visualization system depicts the skeleton and the joints in the 3D space as well as a vector showing the direction that the user is looking at. The kinetic coordinate system is used in the three-dimensional Kinect system and the resulting calculated directional vector is placed in the center of the chest. The axes of the diagram are calculated in meters. This visualization enables 3D viewing, so that the user can view and comprehend the recorded posture and the form that is visualized into the symbolic representation of Labanotation.

Fig. 8.13(a) presents the interface regarding the RGB captured image and the respective skeleton as obtained by the Kinect depth sensor. In this particular scenario, the user raises up his arms. Another example is depicted in Fig. 8.13(b). In this figure, the user makes another movement, raising his right leg. It is clear that the Kinect sensor captures the respective movement and encodes it into skeleton data.

The visualization interface provides the capability of viewing the human movement at different views. This is depicted in Fig. 8.15, where the recorded skeleton is shown in front, side, and top view. In this visualization, each joint is depicted in a different shade, as defined by the Labanotation. A snapshot of the proposed interactive serious game platform is presented in Fig. 8.14. In this figure, we illustrate the captured 3D skeleton joints as well as the Laban symbols as presented in Section 8.4.

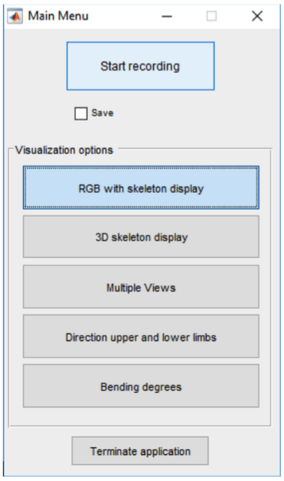
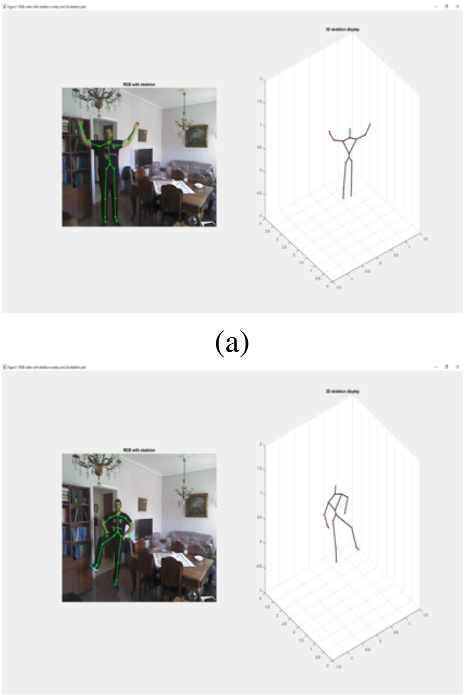


Figure 8.12 The main menu of the proposed embodied educational serious game incorporating Kinect-II sensor and Laban movement analysis for dance training [10].



(a)

(b)

Figure 8.13 Two snapshots of the visualization interface depicting human skeleton overlaid on RGB data for different human postures [10].

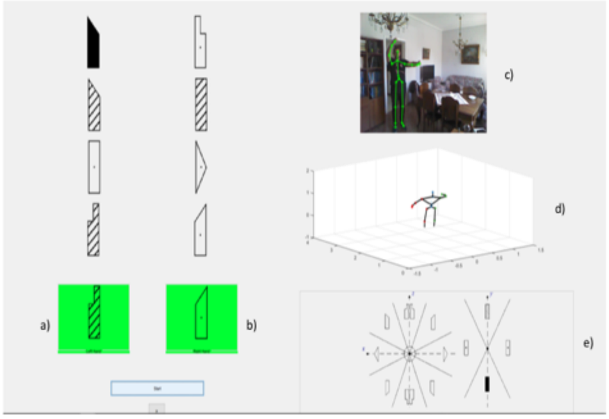


Figure 8.14 A snapshot of the interactive serious game interface [10].

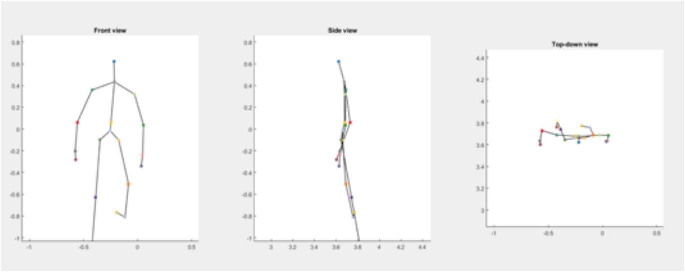


Figure 8.15 The multiple view interface of the proposed serious game platform, allowing users to depict the 3D captures human skeletons [10].

## 8.8 Discussion

In this chapter we have incorporated Artificial Intelligence (AI) in an interactive serious game platform for learning choreographies. Specifically, our proposed framework enhances cognitive and kinesthetic functions through Computer-supported collaborative learning. Additionally, the proposed framework is useful for simple users who want to enter the kinesiology and choreographic segment. AI is capable of providing an assessment interface of a performing choreography by being trained in ground truth data. Two AI modules have been implemented; *pose identification* and *summarization.* Pose identification exploits a bi-directional LSTM model with input data 3D kinematics descriptors (e.g., velocity and acceleration) of a dancer's skeleton joints in order to estimate the main choreographic primitives of a dance. On the other hand, the summarization module is based on the Sparse Modelling Representative Selection (SMRS) algorithm implemented under a hierarchical framework. The summarization interface provides a coarse to fine assessment which is important for dance learning. This way, we can assess a choreography at different representational levels. The coarse levels show the main steps of a dance while the detailed levels provide an assessment per dance frame. Experimental results indicate that the proposed LSTM model for pose identification is robust against other shallow learning techniques. In addition, a visualization interface that supports Laban movements analysis is also adopted to enhance the interactive communication between the game-like platform and, experienced or not, users.

# Chapter 9

# Adaptable deep non-linear Autoregressive Moving Average Filters (ARMA) filter for choreographic modelling

## 9.1 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 [245]. 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 [7]. 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 [246]. 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 [247], [248]. Therefore, CNN networks have been used for human action recognition [249], [250].

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, therefore, to address the dynamic characteristics of a dance.

### 9.1.1 Related Work

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 [86], pose identification [87], action prediction [88], emotion recognition [89] and background subtraction [90]. In [91], an unsupervised approach is proposed for modelling human activities, while in [7], summarization of folklore dances have been introduced using an hierarchical SMRS algorithm. In this context, the work of [92] has introduced an action recognition framework exploiting dense trajectories. Finally, in [93] 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 [94]. In [95], a CNN neural network model have been introduced for human activity analysis, while the work of [96] uses RGB-D and skeleton data for activity analysis. In [97], the authors introduce a two-stream convolutional neural network structure for action recognition in videos. In this context, the work of [98] introduces a three-stream CNN for action recognition modelling, while the work of [99] proposes CNNs structures on depth maps and postures for human action recognition. Finally, Makantasis el al. [100] introduces a behavioural understanding approach for industrial environments, while in [101], 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 [102] exploiting concepts of visual attention algorithms. In this context, the work of [103] introduces a background modeling algorithm using CNN structures. Similarly, in [104], the authors introduce methods of Mixture of Gaussians to face background dynamics. In [105], 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 [106] 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.

### 9.1.2 Innovation and Originality

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 [241].

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).

## 9.2 An Auto Regressive Moving Average-Enriched CNN for Choreography Modeling

Fig. 9.1 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

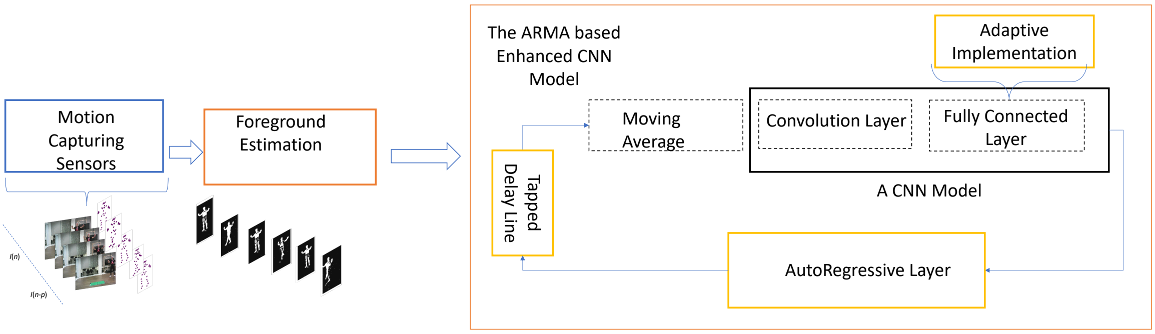


Figure 9.1 The overall proposed architecture adopted in this paper for choreographic modeling.

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.

## 9.2.1 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

$$\begin{aligned} y(n) = g(x(n), x(n-1), \cdots, x(n-q), \\ y(n-1), \cdots, y(n-p)) + e(n) \end{aligned} \tag{9.1}$$

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 (9.1) with any degree of accuracy [241].

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 [247]. 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 [251].

However, conventional CNN structures have not designed to approximate a non-linear ARMA filter as the one of Eq. (9.1). For this reason, in this paper, we extent the conventional CNN models to have ARMA characteristics.

### 9.2.2 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 TDL layer to the network. This is illustrated in Fig. 9.2. 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 (9.1). 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 $x(n-j),\ j=0,\cdots,q-1$.

### 9.2.3 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. 9.2, where the classification output feeds back to the input layer. Actually, the AR behavior implements the second part of (9.1), that is the non linear function of $y(n)$ is related with its own previous values $y(n-1), \cdots, y(n-p)$.

### 9.2.4 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.9.2 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

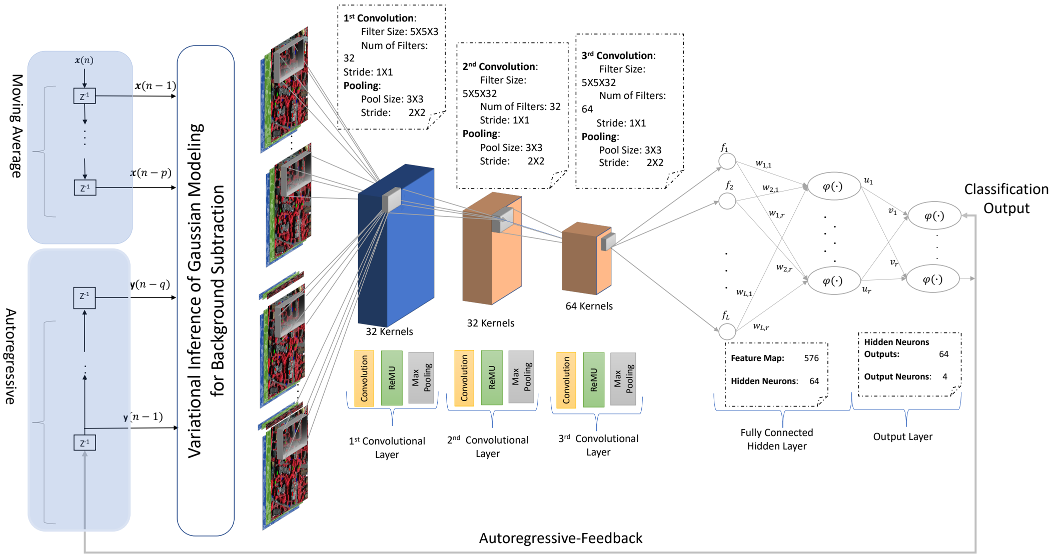

Moving Average
x(n)
Z-1
x(n − 1)
Z-1
x(n − p)
Autoregressive
Z-1
y(n − q)
Z-1
y(n − 1)
Variational Inference of Gaussian Modeling for Background Subtraction
1st Convolution: Filter Size: 5X5X3 Num of Filters: 32 Stride: 1X1 Pooling: Pool Size: 3X3 Stride: 2X2
2nd Convolution: Filter Size: 5X5X32 Num of Filters: 32 Stride: 1X1 Pooling: Pool Size: 3X3 Stride: 2X2
3rd Convolution: Filter Size: 5X5X32 Num of Filters: 64 Stride: 1X1 Pooling: Pool Size: 3X3 Stride: 2X2
32 Kernels
32 Kernels
64 Kernels
Convolution
ReMU
Max Pooling
1st Convolutional Layer
2nd Convolutional Layer
3rd Convolutional Layer
$f_1$
$f_2$
$f_L$
$w_{1,1}$
$w_{2,1}$
$w_{1,r}$
$w_{2,r}$
$w_{L,1}$
$w_{L,r}$
$\varphi(\cdot)$
$u_1$
$u_r$
$v_1$
$v_r$
Classification Output
Feature Map: 576
Hidden Neurons: 64
Hidden Neurons Outputs: 64
Output Neurons: 4
Fully Connected Hidden Layer
Output Layer
Autoregressive-Feedback

Figure 9.2 The architecture of the proposed ARMA-CNN used for choreographic modeling in this paper

ARMA-CNN model. Fig. 9.2 presents the proposed ARMA-CNN architecture adopted for choreographic modeling.

### 9.2.5 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

$$w_a = w_b + dw \tag{9.2}$$

In Eq.(9.2) $dw$ refers to a small perturbation of the network weights. Eq. (9.2) 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. 9.3. A frequency domain approach is adopted for estimating the main choreographic pattern as in [252]. Let us denote that using the method of [252], the main choreographic pattern have been estimated as

$$\gamma = \{c_1(n_s), \cdots, c_L(n_e)\} \tag{9.3}$$

In Eq. 9.3 $c_i(t)$ 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.

$$y_{w_a}(n) \approx c_i(n)\ \forall c_i(n) \in \gamma \tag{9.4}$$

In Eq. (9.4), $y_{w_a}(n)$ denotes the response of the network at the time instance $n$ of using the new adapted weights $w_a$. Eq. (9.4) means that the network response should respect the main choreographic pattern sequence.

Using the assumption of Eq. (9.2), 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

$$e_i(n) = A_i \cdot dw \tag{9.5}$$

In Eq. (9.5) 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,

$$e_i(n) = y_{w_a}(n) - y_{w_b}(n) \tag{9.6}$$

Solving Eq. (9.5) 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. (9.4)).

### 9.2.6 The Optimization Procedure

The main problem of solving Eq.(9.5) 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.9.2). Therefore, there is no a unique solution of solving Eq. (9.5).

To address this limitation, an additional constraint is introduced in this paper. Particularly, we select among all possible solutions that satisfy Eq. (9.5), the one that yields a minimum modification of the small perturbations $dw$. This means that we have the following constraint optimisation framework

$$\begin{gathered} min \; \|dw\| \\ \text{subject to} \\ c_i(n+1) = A_i \cdot dw \end{gathered} \tag{9.7}$$

Solving Eq. (9.7), 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 (9.5). Instead, the previous network knowledge should be modified as discusses in [241].

### 9.2.7 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 [253]. 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 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:

$$P(X_t) = \sum_{i=1}^{K} \omega_{i,t} * \eta(X_t, \mu_{i,t}, \Sigma_{i,t}) \tag{9.8}$$

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. (9.8), we have denoted as scalar for simplicity purposes (More information can be found at [253]). In addition, in Eq. (9.8), $X_t$ 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. (9.8) are probability distributions instead of scalar. Therefore, better function approximations are achieved, improving background/foreground separation performance as it is discussed in [253].

## 9.3 Experimental Evaluation

### 9.3.1 Description of the dataset used

For evaluating and comparing the proposed algorithm against state-of-the-art methods folklore video sequences are used as presented in Table 3.5. 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 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 3.5 describes 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. 9.3 visually depicts the main choreographic primitives of two dance

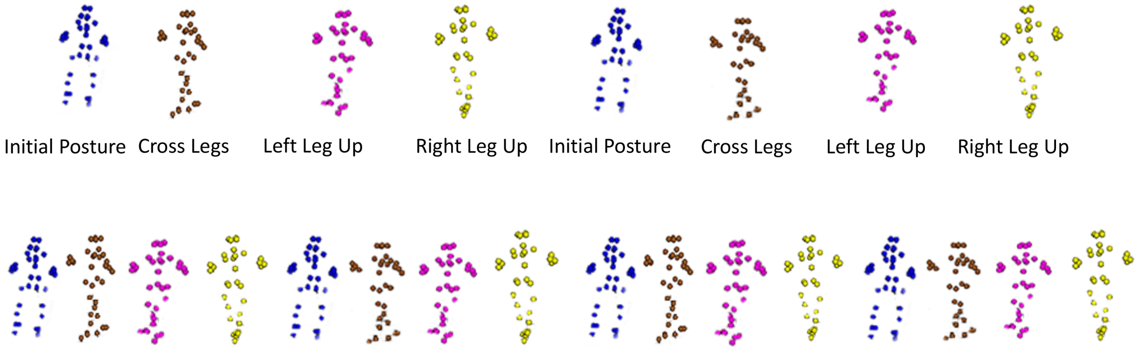


Figure 9.3 Choreographic primitives of two dance sequences.

sequences. As is observed, the choreographic primitives same similarities with each other, imposing difficulties in the recognition process.

### 9.3.2 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. 9.4 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 learning network structures in terms of choreographic modeling. The effect of background modeling and therefore foreground separation is depicted in Table 9.1. 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. 9.4 the results are obtained using the background separation algorithm.

The effect of the background modeling and therefore, the foreground estimation is depicted in Fig.9.5. Background removal is very important for choreographic modeling, since irrelevant to the choreography content is discarded. Fig. 9.6 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).

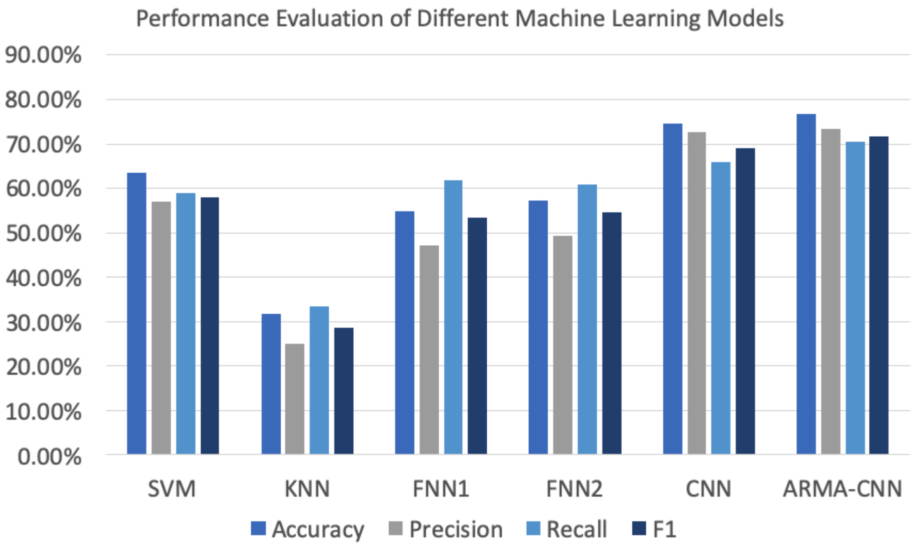


Figure 9.4 Performance Evaluation of different machine learning network set-ups for choreographic primitive classification

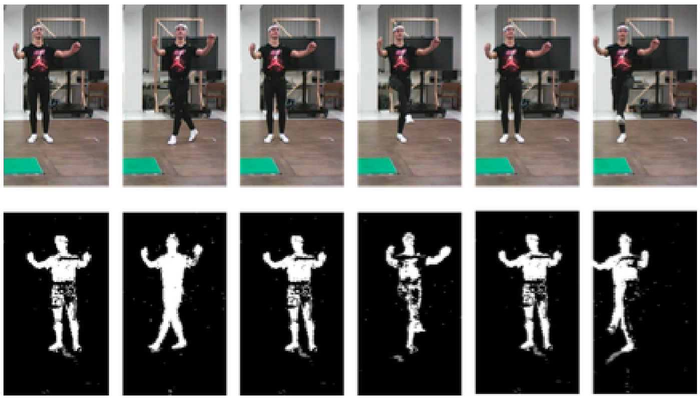
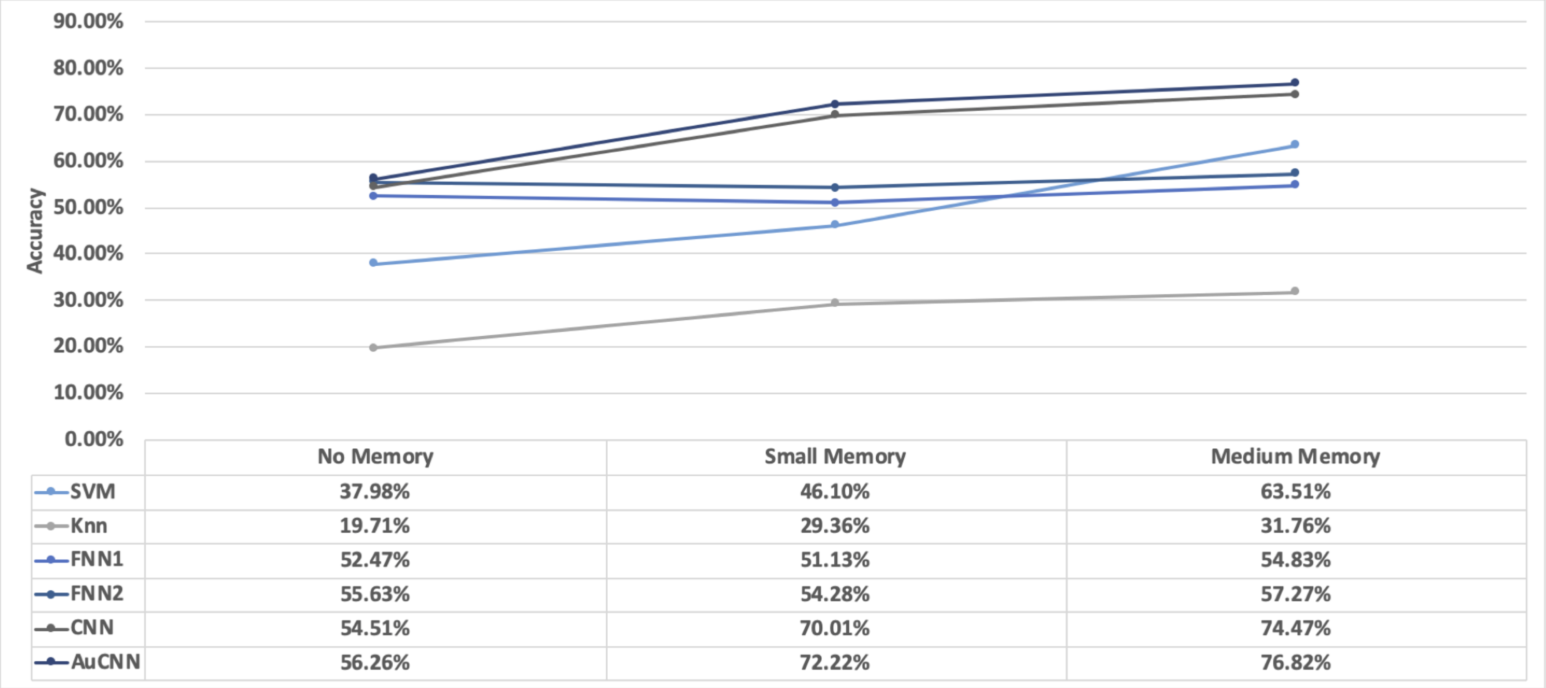


Figure 9.5 Simulation results regarding background/foreground estimation.

Table 9.1 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

| 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% |



| | No Memory | Small Memory | Medium Memory |
|---|---|---|---|
| SVM | 37.98% | 46.10% | 63.51% |
| Knn | 19.71% | 29.36% | 31.76% |
| FNN1 | 52.47% | 51.13% | 54.83% |
| FNN2 | 55.63% | 54.28% | 57.27% |
| CNN | 54.51% | 70.01% | 74.47% |
| AuCNN | 56.26% | 72.22% | 76.82% |

Figure 9.6 The effect of memory, that is the length of the tapped delay filter, on the choreographic modeling performance

## 9.4 Discussion

This chapter 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.

# Part IV

# Conclusions and Future Work

# Chapter 10

# Conclusions of the Thesis

## 10.1 Summary

This thesis was concerned with the development of *(a) new methods for improving the extraction of choreographic primitives taking into account time series analysis, (b) identification algorithms for extraction representative postures from choreographic sequences and (c) semantic representation and notation techniques.*

- Part I presented the theoretical background regarding ICH and the principles with respect to the mathematical modelling of folklore choreographic sequences. Moreover, in Chapters 1, 2, 3 the recent trends on choreographic representation in terms of machine learning, video summarization, pose identification and dance annotation are described. Additionally, this part presents the adopted sensors network (Vicon/Kinect motion capturing systems), the technical specifications, the kinessiological modelling and the annotation of the Greek folklore dances. It is important to mention, that our approach encompasses more than thirty folkloric dance sequences recorded at the Aristotle University of Thessaloniki and at the School of Physical Education and Sport Science of the University of Thessaly in Trikala under the TERPSICHORE project (see Chapter 3.4). These choreographic datasets encompass more than 83662 RGB images and point clouds records compatible with various databases.

- Part II presents the adopted techniques for content-based sampling of the selected folklore choreographic sequences. This part is oriented on the semantic compression and the video summarization taking into consideration the complexity of the spatio-temporal sequences. In particular, Chapter 4 exploited a hierarchical scheme that implements spatio-temporal variations of the dance features. Firstly, global holistic descriptors are defined to localize the key choreographic steps of a dance (a coarse representation). Secondly, each segment is further decomposed into finer sub-segments to improve dance representativity (fine representation). Chapter 5 describes an abstract representation of the semantic details of choreographic sequences taking into consideration a key-frame selection algorithm. Chapter 6 compares the summarization performances taking into account four sampling algorithms all implemented under a SAE scheme's projected data. Specifically, a SAE framework followed by a hierarchical SMRS algorithm implemented to summarize choreographic sequences.

- Part III (Chapters 7, 8, 9) focused on modelling and analysis of folklore choreographic sequences. Chapter 7 explored the feasibility of pattern matching between heterogeneous motion capturing systems. In this chapter, a trajectory interpretation in folklore sequences is described. The conducted experiments indicate that if significant levels of precision are ensured during initial data collection, design, development and fine-tuning of the system, then low-cost and widely popular motion capturing sensors suffice to provide a smooth and integrated experience on the user end, which would allow for relevant educational or entertainment applications to be adopted at scale. Nevertheless, the proposed approach would not be appropriate for tasks that require great precision and accuracy in measurement of movement and positioning of individual joints. Chapter 8 focuses on the enhancement of the learning experience of folklore dances by introducing machine learning tools with the capability of providing a scalable quantifiable assessment of a choreography at different level of hierarchies; yielding a from coarse to fine evaluation. For this reason, initially the choreography is analyzed into representative 3D skeleton joints and then kinematics features are estimated to efficiently model these choreographic patterns. Then, pose identification and summarization methods are implemented with the main purpose of categorizing each dance sequence into choreographic primitives or extracting the main (key) choreographic pattern. Pose identification provides a detailed (fine) assessment of a dance, which, in the sequel, stimulates an assessment of a dance performance against ground truth data. On the other hand, summarization creates a coarse representation (and thus assessment) of the choreography. Chapter 9 describes an adaptable autogressive and moving average layer (R-ARMA) into a conventional CNN filter to model the dynamic behavior of a choreography. In addition, to face the choreography dynamics, we introduced 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 indicated the efficiency of the proposed model against conventional deep machine learning filters.

### 10.1.1 Innovation and Originality

The work presented in the previous Chapters was achieved with the ultimate goal of highlighting the emblematic role of ICH, the effective use of emerging machine learning techniques and the implementation of image processing algorithms. The main contributions of this thesis are listed below.

1. **Development of two folklore choreographic datasets** (see Sections 3.3.1, 3.3.2, 3.5). Our approach included thirty folkloric dance sequences recorded at the Aristotle University of Thessaloniki under the framework of TERPSICHORE project representing five different choreographies.

2. **A method that matches trajectories' patterns, existing in a choreographic database, to new ones originating from different sensor types such as VICON and Kinect II**. The main objective of this approach is to evaluate the performance between heterogeneous motion capturing systems (see Section 7).

3. **A key frame extraction framework that implements a hierarchical scheme exploiting spatio-temporal variations of the dance features is introduced**. In Section 4 we introduced a spatio-

temporal video summarization implemented under a hierarchical framework. This hierarchical video dance decomposition results in extracting a pyramid of key frames that provides a complete overview of a choreography, from a coarse to a fine description. The advantage of directly processing 3D human skeleton points instead of raw depth data is that few data samples are involved in the processing of the dance sequences, making summarization much more efficient.

4. **A machine learning method exploiting deep learning paradigms is proposed (see Section 8)**. This proposed framework proposes a choreographic summarization architecture based on SMRS in order to abstractly represent the performing choreography through a set of key choreographic primitives. We have modified the SMRS algorithm in a way to extract hierarchies of key representatives. Choreographic summarization provides an efficient tool for a coarse quantitative evaluation of a dance. Moreover, hierarchical representation scheme allows for a scalable assessment of a choreography. The serious game platform supports advanced visualization toolkits using Labanotation in order to deliver the performing sequence in a formal documentation.

5. **Development of a method to address dynamic limitations of choreogpahic sequences (see Section)**. We introduced 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.

6. **Development of a deep stacked auto-encoder (SAE) scheme followed by a H-SMRS algorithm proposed to summarize dance video sequences** (see Section 9). 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. 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 representation 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.

### 10.1.2 Future Prospects

Although, during this dissertation, many approaches carried out in the areas of digitization, recording and modeling of ICH. During the completion period of this dissertation, more scientific approaches, new algorithms and topics of interest for further research have emerged.

***Adaptation of Generative Adversarial Networks to create choreographic sequences***

Since the introduction of deep learning, researchers have proposed content generation systems using deep learning and proved that they are competent to generate convincing motion content and kinesiological output, including music, rhythm and choreographic patterns [254]. These deep learning-based algorithms imitate and reproduce patterns with same statistics with the training set [255]. The framework of Generative Adversarial Networks (GAN) can generate choreographic patterns that imitate choreographic sequences but do not belong to motion training dataset [53], [256], [257], [258] [259]. Further research on this will facilitate the generation of choreographic sequences of less noisy point clouds/RGB images.

***Emotional style dancers representation and modelling***

The dancers' expressions, the emotions and the style of the performers are crucial to categorize the choreographic patterns. The research of human behaviour under different emotional states is important to define different personalities, moods or emotion variations [260].

***Implementation of U-Net on choreographic data for classification purposes.***

There's a huge assent that deep networks requires many thousand annotated training samples. Particularly, within the substance of ICH, folklore choreographic sequences are usually not annotated. The finding that pre-training a network on a rich source set can offer better performance once fine-tuned on a usually much smaller target set, has been instrumental to numerous applications [261], [262]. Nowadays, very little is known about its usefulness in 3D point cloud understanding. I observe this as an opportunity considering the effort required for annotating data in 3D. At this direction, i aim at facilitating research on 3D representation learning. Different from previous works focusing on high-level scene understanding tasks.

# Bibliography

[1] Eftychios Protopapadakis, Athina Grammatikopoulou, Anastasios Doulamis, and Nikos Grammalidis. Folk dance pattern recognition over depth images acquired via kinect sensor. *3D ARCH-3D Virtual Reconstruction and Visualization of Complex Architectures*, 2017.

[2] Omid Alemi and Philippe Pasquier. Machine learning for data-driven movement generation: a review of the state of the art. *arXiv preprint arXiv:1903.08356*, 2019.

[3] Ioannis Rallis, Eftychios Protopapadakis, Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Georgios Bardis. Choreographic pattern analysis from heterogeneous motion capture systems using dynamic time warping. *Technologies*, 7(3):56, 2019.

[4] Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: Archive of motion capture as surface shapes. In *International Conference on Computer Vision*, pages 5442–5451, October 2019.

[5] Eftychios Protopapadakis, Ioannis Rallis, Anastasios Doulamis, Nikolaos Doulamis, and Athanasios Voulodimos. Unsupervised 3d motion summarization using stacked auto-encoders. *Applied Sciences*, 10(22):8226, 2020.

[6] Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, KangKang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. Emotion Control of Unstructured Dance Movements. In *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, SCA '17, pages 9:1–9:10, New York, NY, USA, 2017. ACM. event-place: Los Angeles, California.

[7] Ioannis Rallis, Nikolaos Doulamis, Anastasios Doulamis, Athanasios Voulodimos, and Vassilios Vescoukis. Spatio-temporal summarization of dance choreographies. *Computers & Graphics*, 73:88–101, 2018.

[8] Ioannis Rallis, Ioannis Georgoulas, Nikolaos Doulamis, Athanasios Voulodimos, and Panagiotis Terzopoulos. Extraction of key postures from 3d human motion data for choreography summarization. In *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 94–101. IEEE, 2017.

[9] Christian Griesbeck. Introduction to labanotation. *http://www. rz. uni-frankfurt. de/griesbec/LABANE. HTML*, 1996.

[10] Ioannis Rallis, Nikolaos Bakalos, Nikolaos Doulamis, Anastasios Doulamis, and Athanasios Voulodimos. Bidirectional long short-term memory networks and sparse hierarchical modeling for scalable educational learning of dance choreographies. *The Visual Computer*, pages 1–16, 2019.

[11] Henrietta Marrie. The unesco convention for the safeguarding of the intangible cultural heritage and the protection and maintenance of the intangible cultural heritage of indigenous peoples. *" Intangible heritage"*, pages p–169, 2009.

[12] Anastasios Doulamis, Marinos Ioannides, Nikolaos Doulamis, Andreas Hadjiprocopis, Dieter Fritsch, Olivier Balet, Martine Julien, Eftychios Protopapadakis, Kostas Makantasis, Guenther Weinlinger, Paul S. Johnsons, Michael Klein, Dieter Fellner, Andre Stork, and Pedro Santos. 4D

reconstruction of the past. In *First International Conference on Remote Sensing and Geoinformation of the Environment (RSCy2013)*, volume 8795, page 87950J. International Society for Optics and Photonics, August 2013.

[13] Andreas Aristidou, Panayiotis Charalambous, and Yiorgos Chrysanthou. Emotion analysis and classification: Understanding the performers' emotions using the lma entities. 34(6):262–276, 2015.

[14] Ioannis Rallis, Athanasios Voulodimos, Nikolaos Bakalos, Eftychios Protopapadakis, Nikolaos Doulamis, and Anastasios Doulamis. Machine learning for intangible cultural heritage: A review. *Visual Computing for Cultural Heritage*, page 104, 2020.

[15] Sophia Labadi. *UNESCO, cultural heritage, and outstanding universal value: Value-based analyses of the World Heritage and Intangible Cultural Heritage Conventions*. Rowman & Littlefield, 2013.

[16] M Chapuis. Preserving our heritage, improving our environment, vol. i, 20 years of eu research into cultural heritage, eur 22050 en, 2009. *M. Chapuis, A. Lydon and A. Brandt-Grau, Preserving our heritage, improving our environment*, 2.

[17] Jean-Claude Piris. *The Lisbon Treaty: a legal and political analysis*. Cambridge University Press, 2010.

[18] Aleksandra Marinković, Aleksandra Mirić, and Filip Mirić. European policy on digitisation of cultural heritage from 2005. onwards. *Facta Universitatis, Series: Law and Politics*, pages 97–103, 2016.

[19] Comité des Sages. For a europe of civic and social rights. *Brussels: European Commission*, 1996.

[20] Despoina Karavia and Andreas Georgopoulos. Placing intangible cultural heritage. In *2013 Digital Heritage International Congress (DigitalHeritage)*, volume 1, pages 675–678. IEEE, 2013.

[21] Anastasios Doulamis, Marinos Ioannides, Nikolaos Doulamis, Andreas Hadjiprocopis, Dieter Fritsch, Olivier Balet, Martine Julien, Eftychios Protopapadakis, Kostas Makantasis, Guenther Weinlinger, et al. 4d reconstruction of the past. In *First international conference on remote sensing and geoinformation of the environment (RSCy2013)*, volume 8795, page 87950J. International Society for Optics and Photonics, 2013.

[22] Athanasios Voulodimos, Nikolaos Doulamis, Dieter Fritsch, Konstantinos Makantasis, Anastasios Doulamis, and Michael Klein. Four-dimensional reconstruction of cultural heritage sites based on photogrammetry and clustering. *Journal of Electronic Imaging*, 26(1):011013, 2016.

[23] Anthony Shay and Barbara Sellers-Young. *The Oxford Handbook of Dance and Ethnicity*. Oxford University Press, April 2016. Google-Books-ID: cPnmDAAAQBAJ.

[24] Nikolaos Doulamis, Anastasios Doulamis, Charalabos Ioannidis, Michael Klein, and Marinos Ioannides. Modelling of Static and Moving Objects: Digitizing Tangible and Intangible Cultural Heritage. In Marinos Ioannides, Nadia Magnenat-Thalmann, and George Papagiannakis, editors, *Mixed Reality and Gamification for Cultural Heritage*, pages 567–589. Springer International Publishing, Cham, 2017.

[25] Michalis Raptis, Darko Kirovski, and Hugues Hoppe. Real-time Classification of Dance Gestures from Skeleton Animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '11, pages 147–156, New York, NY, USA, 2011. ACM. event-place: Vancouver, British Columbia, Canada.

[26] Kosmas Dimitropoulos, Sotiris Manitsaris, Filareti Tsalakanidou, Spiros Nikolopoulos, Bruce Denby, Samer Al Kork, Lise Crevier-Buchman, Claire Pillot-Loiseau, Martine Adda-Decker, Stéphane Dupont, Joëlle Tilmanne, Michela Ott, Marilena Alivizatou, Erdal Yilmaz, Leontios Hadjileontiadis, Vasileios Charisis, Olivier Deroo, Athanasios Manitsaris, Ioannis Kompatsiaris, and Grammalidis Nikos. Capturing the intangible: An introduction to the i-treasures project. 01 2014.

[27] Eftychios Protopapadakis, Athanasios Voulodimos, Anastasios Doulamis, Stephanos Camarinopoulos, Nikolaos Doulamis, and Georgios Miaoulis. Dance Pose Identification from Motion Capture Data: A Comparison of Classifiers. *Technologies*, 6(1):31, March 2018.

[28] Fotis Liarokapis. An augmented reality interface for visualizing and interacting with virtual content. *Virtual Reality*, 11(1):23–43, 2007.

[29] Terpsichore Project. http://terpsichore-project.eu/, 2019.

[30] Janis Pforsich. *Handbook for Laban Movement Analysis*. 1977.

[31] K. Kojima, K. Hachimura, and M. Nakamura. LabanEditor: Graphical editor for dance notation. In *11th IEEE International Workshop on Robot and Human Interactive Communication Proceedings*, pages 59–64, September 2002.

[32] Ann Hutchinson, William Ambrose Hutchinson, and Ann Hutchinson Guest. *Labanotation: Or, Kinetography Laban : the System of Analyzing and Recording Movement*. Taylor & Francis, 1970. Google-Books-ID: Tq1YRDuJnvYC.

[33] Kensuke Hisatomi, Miwa Katayama, Kimihiro Tomiyama, and Yuichi Iwadate. 3d archive system for traditional performing arts. *International journal of computer vision*, 94(1):78–88, 2011.

[34] Efstathios Stavrakis, Andreas Aristidou, Maria Savva, Stephania Loizidou Himona, and Yiorgos Chrysanthou. Digitization of cypriot folk dances. pages 404–413, 2012.

[35] R.M. Sheppard, M. Kamali, R. Rivas, M. Tamai, Z. Yang, W. Wu, and K. Nahrstedt. Advancing interactive collaborative mediums through tele-immersive dance (ted): A symbiotic creativity and design environment for art and computer science. In *Proc.of the 2008 ACM International Conference on Multimedia, with co-located Symposium and Workshops*, pages 579–588, 2008.

[36] Andreas Aristidou, Efstathios Stavrakis, Panayiotis Charalambous, Yiorgos Chrysanthou, and Stephania Loizidou Himona. Folk dance evaluation using laban movement analysis. *Journal on Computing and Cultural Heritage (JOCCH)*, 8(4):1–19, 2015.

[37] Eftychios Protopapadakis, Athina Grammatikopoulou, Anastasios Doulamis, and Nikolaos Grammalidis. Folk dance pattern recognition over depth images acquired via kinect sensor. In *Proc. of the International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences - ISPRS Archives*, volume 42, pages 587–593, 2017.

[38] Kosmas Dimitropoulos, Sotiris Manitsaris, Filareti Tsalakanidou, Spiros Nikolopoulos, Bruce Denby, Samer Al Kork, Lise Crevier-Buchman, Claire Pillot-Loiseau, Martine Adda-Decker, Stéphane Dupont, Joëlle Tilmanne, Michela Ott, Marilena Alivizatou, Erdal Yilmaz, Leontios Hadjileontiadis, Vasileios Charisis, Olivier Deroo, Athanasios Manitsaris, Ioannis Kompatsiaris, and Grammalidis Nikos. Capturing the intangible: An introduction to the i-treasures project. 01 2014.

[39] Michalis Raptis, Darko Kirovski, and Hugues Hoppe. Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 147–156, 2011.

[40] Zhengyou Zhang. Microsoft kinect sensor and its effect. *IEEE multimedia*, 19(2):4–10, 2012.

[41] Aymeric Masurelle, Slim Essid, and Gaël Richard. Multimodal classification of dance movements using body joint trajectories and step sounds. In *2013 14th international workshop on image analysis for multimedia interactive services (WIAMIS)*, pages 1–4. IEEE, 2013.

[42] Dohyung Kim, Dong-Hyeon Kim, and Keun-Chang Kwak. Classification of k-pop dance movements based on skeleton information obtained by a kinect sensor. *Sensors*, 17(6):1261, 2017.

[43] Travis T Simpson, Susan L Wiesner, and Bradford C Bennett. Dance recognition system using lower body movement. *Journal of applied biomechanics*, 30(1):147–153, 2014.

[44] Apostolos Laggis, Nikolaos Doulamis, Eftychios Protopapadakis, and Andreas Georgopoulos. A low-cost markerless tracking system for trajectory interpretation. *The International Archives of Photogrammetry, Remote Sensing and Spatial Information Sciences*, 42:413, 2017.

[45] Ioannis Mademlis, Anastasios Tefas, and Ioannis Pitas. A salient dictionary learning framework for activity video summarization via key-frame extraction. *Information Sciences*, 432:319–331, 2018.

[46] Ioannis Mademlis, Anastasios Tefas, Nikos Nikolaidis, and Ioannis Pitas. Summarization of human activity videos via low-rank approximation. *Signal Processing (ICASSP 2017)*, 5:9.

[47] T. Wu, P. Gurram, R.M. Rao, and W.U. Bajwa. Hierarchical union-of-subspaces model for human activity summarization. In *Proc. of the IEEE International Conference on Computer Vision*, pages 1053–1061, 2016.

[48] Kosmas Dimitropoulos, Panagiotis Barmpoutis, Alexandros Kitsikidis, and Nikos Grammalidis. Classification of multidimensional time-evolving data using histograms of grassmannian points. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(4):892–905, 2016.

[49] Andreas Aristidou, Efstathios Stavrakis, Margarita Papaefthimiou, George Papagiannakis, and Yiorgos Chrysanthou. Style-based motion analysis for dance composition. *Visual Computer*, pages 1–13, 2017.

[50] Andreas Aristidou, KangKang Yin, Qiong Zeng, Daniel Cohen-Or, Baoquan Chen, Efstathios Stavrakis, and Yiorgos Chrysanthou. Emotion control of unstructured dance movements. In *Proc. of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation*, 2017.

[51] Andreas Aristidou, Panayiotis Charalambous, and Yiorgos Chrysanthou. Emotion analysis and classification: Understanding the performers' emotions using the lma entities. In *Computer Graphics Forum*, volume 34, pages 262–276. Wiley Online Library, 2015.

[52] Anastasios D Doulamis, Nikolaos D Doulamis, and Stefanos D Kollias. A fuzzy video content representation for video summarization and content-based retrieval. *Signal Processing*, 80(6):1049–1067, 2000.

[53] Caroline Chan, Shiry Ginosar, Tinghui Zhou, and Alexei A. Efros. Everybody Dance Now. *arXiv:1808.07371 [cs]*, August 2018. arXiv: 1808.07371.

[54] Ioannis Rallis, Nikolaos Doulamis, Athanasios Voulodimos, and Anastasios Doulamis. Hierarchical sparse modeling for representative selection in choreographic time series. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1023–1027. IEEE, 2018.

[55] Björn Krüger, Anna Vögele, Tobias Willig, Angela Yao, Reinhard Klein, and Andreas Weber. Efficient unsupervised temporal segmentation of motion data. *IEEE Transactions on Multimedia*, 19(4):797–812, 2017.

[56] Ioannis Rallis, Apostolos Langis, Ioannis Georgoulas, Athanasios Voulodimos, Nikolaos Doulamis, and Anastasios Doulamis. An embodied learning game using kinect and labanotation for analysis and visualization of dance kinesiology. In *2018 10th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 1–8. IEEE Computer Society, 2018.

[57] Ehsan Elhamifar, Guillermo Sapiro, and René Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1607, 2012.

[58] Kanav Kahol, Priyamvada Tripathi, and Sethuraman Panchanathan. Automated gesture segmentation from dance sequences. In *Sixth IEEE International Conference on Automatic Face and Gesture Recognition, 2004. Proceedings.*, pages 883–888. IEEE, 2004.

[59] Athanasios Voulodimos, Nikolas Doulamis, Anastasios Doulamis, and Ioannis Rallis. Kinematics-based extraction of salient 3d human motion data for summarization of choreographic sequences. In *2018 24th international conference on pattern recognition (ICPR)*, pages 3013–3018. IEEE, 2018.

[60] Athanasios Voulodimos, Ioannis Rallis, and Nikolaos Doulamis. Physics-based keyframe selection for human motion summarization. *Multimedia Tools and Applications*, December 2018.

[61] Björn Krüger, Jochen Tautges, Andreas Weber, and Arno Zinke. Fast Local and Global Similarity Searches in Large Motion Capture Databases. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, SCA '10, pages 1–10, Goslar Germany, Germany, 2010. Eurographics Association. event-place: Madrid, Spain.

[62] Jehee Lee, Jinxiang Chai, Paul S. A. Reitsma, Jessica K. Hodgins, and Nancy S. Pollard. Interactive Control of Avatars Animated with Human Motion Data. In *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '02, pages 491–500, New York, NY, USA, 2002. ACM. event-place: San Antonio, Texas.

[63] Jinxiang Chai and Jessica K. Hodgins. Performance Animation from Low-dimensional Control Signals. In *ACM SIGGRAPH 2005 Papers*, SIGGRAPH '05, pages 686–696, New York, NY, USA, 2005. ACM. event-place: Los Angeles, California.

[64] Takaaki Shiratori, Atsushi Nakazawa, and Katsushi Ikeuchi. Dancing-to-Music Character Animation. *Computer Graphics Forum*, 25(3):449–458, 2006.

[65] Alexandros Kitsikidis, Kosmas Dimitropoulos, Deniz Uğurca, Can Bayçay, Erdal Yilmaz, Filareti Tsalakanidou, Stella Douka, and Nikos Grammalidis. A Game-like Application for Dance Learning Using a Natural Human Computer Interface. In Margherita Antona and Constantine Stephanidis, editors, *Universal Access in Human-Computer Interaction. Access to Learning, Health and Well-Being*, Lecture Notes in Computer Science, pages 472–482. Springer International Publishing, 2015.

[66] Sam Ferguson, Emery Schubert, and Catherine J. Stevens. Dynamic Dance Warping: Using Dynamic Time Warping to Compare Dance Movement Performed Under Different Conditions. In *Proceedings of the 2014 International Workshop on Movement and Computing*, MOCO '14, pages 94:94–94:99, New York, NY, USA, 2014. ACM. event-place: Paris, France.

[67] Nikolaos Bakalos, Eftychios Protopapadakis, Anastasios Doulamis, and Nikolaos Doulamis. Dance Posture/Steps Classification Using 3D Joints from the Kinect Sensors. In *2018 IEEE 16th Intl Conf on Dependable, Autonomic and Secure Computing, 16th Intl Conf on Pervasive Intelligence and Computing, 4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech)*, pages 868–873, August 2018.

[68] Alexandros Kitsikidis, Kosmas Dimitropoulos, Stella Douka, and Nikos Grammalidis. Dance analysis using multiple kinect sensors. In *Computer Vision Theory and Applications (VISAPP), 2014 International Conference on*, volume 2, pages 789–795. IEEE, 2014.

[69] Dushyant Mehta, Helge Rhodin, Dan Casas, Pascal Fua, Oleksandr Sotnychenko, Weipeng Xu, and Christian Theobalt. Monocular 3d human pose estimation in the wild using improved cnn supervision. In *2017 International Conference on 3D Vision (3DV)*, pages 506–516, 2017.

[70] Ioannis Kavouras, Eftychios Protopapadakis, Anastasios Doulamis, and Nikolaos Doulamis. Skeleton Extraction of Dance Sequences from 3D Points Using Convolutional Neural Networks Based on a New Developed C3D Visualization Interface. In Michael E. Auer and Thrasyvoulos Tsiatsos, editors, *The Challenges of the Digital Transformation in Education*, Advances in Intelligent Systems and Computing, pages 267–279. Springer International Publishing, 2019.

[71] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *arXiv:1611.08050 [cs]*, November 2016. arXiv: 1611.08050.

[72] Luka Crnkovic-Friis and Louise Crnkovic-Friis. Generative Choreography using Deep Learning. *arXiv:1605.06921 [cs]*, May 2016. arXiv: 1605.06921.

[73] Bernhard Kohn, Aneta Nowakowska, and Ahmed Nabil Belbachir. Real-time body motion analysis for dance pattern recognition. In *2012 IEEE computer society conference on computer vision and pattern recognition workshops*, pages 48–53. IEEE, 2012.

[74] Taoran Tang, Jia Jia, and Hanyang Mao. Dance with Melody: An LSTM-autoencoder Approach to Music-oriented Dance Synthesis. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pages 1598–1606, New York, NY, USA, 2018. ACM. event-place: Seoul, Republic of Korea.

[75] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional GANs. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8798–8807, Salt Lake City, UT, USA, June 2018. IEEE.

[76] Aymeric Masurelle, Slim Essid, and Gaël Richard. Multimodal classification of dance movements using body joint trajectories and step sounds. In *2013 14th international workshop on image analysis for multimedia interactive services (WIAMIS)*, pages 1–4. IEEE, 2013.

[77] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5967–5976, Honolulu, HI, July 2017. IEEE.

[78] Panagiotis Barmpoutis, Tania Stathaki, and Stephanos Camarinopoulos. Skeleton-based human action recognition through third-order tensor representation and spatio-temporal analysis. *Inventions*, 4(1), 2019.

[79] Andreas Aristidou and Yiorgos Chrysanthou. Motion Indexing of Different Emotional States Using LMA Components. In *SIGGRAPH Asia 2013 Technical Briefs*, SA '13, pages 21:1–21:4, New York, NY, USA, 2013. ACM. event-place: Hong Kong, Hong Kong.

[80] Anastasios Ballas, Tossaporn Santad, Kingkarn Sookhanaphibarn, and Worawat Choensawat. Game-based system for learning labanotation using microsoft kinect. In *2017 IEEE 6th global conference on consumer electronics (GCCE)*, pages 1–3. IEEE, 2017.

[81] Kozaburo Hachimura, Katsumi Takashina, and Mitsu Yoshimura. Analysis and evaluation of dancing movement based on lma. In *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, pages 294–299. IEEE, 2005.

[82] Durell Bouchard and Norman Badler. Semantic Segmentation of Motion Capture Using Laban Movement Analysis. In Catherine Pelachaud, Jean-Claude Martin, Elisabeth André, Gérard Chollet, Kostas Karpouzis, and Danielle Pelé, editors, *Intelligent Virtual Agents*, Lecture Notes in Computer Science, pages 37–44. Springer Berlin Heidelberg, 2007.

[83] Haris Zacharatos, Christos Gatzoulis, Yiorgos Chrysanthou, and Andreas Aristidou. Emotion Recognition for Exergames Using Laban Movement Analysis. In *Proceedings of Motion on Games*, MIG '13, pages 39:61–39:66, New York, NY, USA, 2013. ACM. event-place: Dublin 2, Ireland.

[84] Feng Zhou, Fernando De la Torre, and Jessica K Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):582–596, 2012.

[85] Andreas Aristidou, Efstathios Stavrakis, Margarita Papaefthimiou, George Papagiannakis, and Yiorgos Chrysanthou. Style-based motion analysis for dance composition. *The visual computer*, 34(12):1725–1737, 2018.

[86] Jezekiel Ben-Arie, Zhiqian Wang, Purvin Pandit, and Shyamsundar Rajaram. Human activity recognition using multidimensional indexing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(8):1091–1104, 2002.

[87] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid. P-cnn: Pose-based cnn features for action recognition. In *Proceedings of the IEEE international conference on computer vision*, pages 3218–3226, 2015.

[88] Simon Hadfield and Richard Bowden. Hollywood 3d: Recognizing actions in 3d natural scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3398–3405, 2013.

[89] Yin Fan, Xiangju Lu, Dian Li, and Yuanliu Liu. Video-based emotion recognition using cnn-rnn and c3d hybrid networks. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, pages 445–450. ACM, 2016.

[90] Massimo Piccardi. Background subtraction techniques: a review. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 4, pages 3099–3104. IEEE, 2004.

[91] Timo Milbich, Miguel Bautista, Ekaterina Sutter, and Bjorn Ommer. Unsupervised video understanding by reconciliation of posture similarities. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4394–4404, 2017.

[92] Heng Wang, Alexander Kläser, Cordelia Schmid, and Liu Cheng-Lin. Action recognition by dense trajectories. 2011.

[93] Maheshkumar H Kolekar and Deba Prasad Dash. Hidden markov model based human activity recognition using shape and optical flow based features. In *2016 IEEE Region 10 Conference (TENCON)*, pages 393–397. IEEE, 2016.

[94] Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018, 2018.

[95] Ming Zeng, Le T Nguyen, Bo Yu, Ole J Mengshoel, Jiang Zhu, Pang Wu, and Joy Zhang. Convolutional neural networks for human activity recognition using mobile sensors. In *6th International Conference on Mobile Computing, Applications and Services*, pages 197–205. IEEE, 2014.

[96] Pushpajit Khaire, Praveen Kumar, and Javed Imran. Combining cnn streams of rgb-d and skeletal data for human activity recognition. *Pattern Recognition Letters*, 115:107–116, 2018.

[97] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*, pages 568–576, 2014.

[98] Liangliang Wang, Lianzheng Ge, Ruifeng Li, and Yajun Fang. Three-stream cnns for action recognition. *Pattern Recognition Letters*, 92:33–40, 2017.

[99] Aouaidjia Kamel, Bin Sheng, Po Yang, Ping Li, Ruimin Shen, and David Dagan Feng. 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, 2019.

[100] Konstantinos Makantasis, Anastasios Doulamis, Nikolaos Doulamis, and Konstantinos Psychas. Deep learning based human behavior recognition in industrial workflows. In *2016 IEEE ICIP*, pages 1609–1613. IEEE, 2016.

[101] Chuang Gan, Naiyan Wang, Yi Yang, Dit-Yan Yeung, and Alex G Hauptmann. Devnet: A deep event network for multimedia event detection and evidence recounting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2568–2577, 2015.

[102] Konstantinos Makantasis, Anastasios Doulamis, and Nikolaos Doulamis. Vision-based maritime surveillance system using fused visual attention maps and online adaptable tracker. In *2013 14th international workshop on image analysis for multimedia interactive services (WIAMIS)*, pages 1–4. IEEE, 2013.

[103] Mohammadreza Babaee, Duc Tung Dinh, and Gerhard Rigoll. A deep convolutional neural network for video sequence background subtraction. *Pattern Recognition*, 76:635–649, 2018.

[104] Sriram Varadarajan, Paul Miller, and Huiyu Zhou. Region-based mixture of gaussians modelling for foreground detection in dynamic scenes. *Pattern Recognition*, 48(11):3488–3503, 2015.

[105] Filippo Maria Bianchi, Daniele Grattarola, Cesare Alippi, and Lorenzo Livi. Graph neural networks with convolutional arma filters. *arXiv preprint arXiv:1901.01343*, 2019.

[106] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, pages 3844–3852, 2016.

[107] Antoine Liutkus, Angélique Dremeau, Dimitrios Alexiadis, Slim Essid, and Petros Daras. Analysis of Dance Movements Using Gaussian Processes: Extended Abstract. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, pages 1375–1376, New York, NY, USA, 2012. ACM. event-place: Nara, Japan.

[108] Alexandros Kitsikidis, Kosmas Dimitropoulos, Stella Douka, and Nikos Grammalidis. Dance analysis using multiple kinect sensors. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 789–795. IEEE, 2014.

[109] D. Kim, M. Jang, Y. Yoon, and J. Kim. Classification of Dance Motions with Depth Cameras Using Subsequence Dynamic Time Warping. In *2015 8th International Conference on Signal Processing, Image Processing and Pattern Recognition (SIP)*, pages 5–8, November 2015.

[110] Ioannis Rallis, Ioannis Georgoulas, Nikolaos Doulamis, Athanasios Voulodimos, and Panagiotis Terzopoulos. Extraction of key postures from 3d human motion data for choreography summarization. In *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 94–101. IEEE, 2017.

[111] Ioannis Rallis, Nikolaos Doulamis, Anastasios Doulamis, Athanasios Voulodimos, and Vassilios Vescoukis. Spatio-temporal summarization of dance choreographies. *Computers & Graphics*, 73:88–101, 2018.

[112] Athanasios Voulodimos, Ioannis Rallis, and Nikolaos Doulamis. Physics-based keyframe selection for human motion summarization. *Multimedia Tools and Applications*, pages 1–17, 2018.

[113] S. Dewan, S. Agarwal, and N. Singh. Spatio-Temporal Laban Features for Dance Style Recognition. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2911–2916, August 2018.

[114] Jiaman Li, Yihang Yin, Hang Chu, Yi Zhou, Tingwu Wang, Sanja Fidler, and Hao Li. Learning to generate diverse dance motions with transformer. *CoRR*, abs/2008.08171, 2020.

[115] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Learn to dance with AIST++: music conditioned 3d dance generation. *CoRR*, abs/2101.08779, 2021.

[116] Andreas Aristidou, Panayiotis Charalambous, and Yiorgos Chrysanthou. Emotion Analysis and Classification: Understanding the Performers' Emotions Using the LMA Entities. *Comput. Graph. Forum*, 34(6):262–276, September 2015.

[117] Alexandros Kitsikidis, Kosmas Dimitropoulos, Stella Douka, and Nikos Grammalidis. Dance analysis using multiple kinect sensors. In *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, volume 2, pages 789–795. IEEE, 2014.

[118] Eftychios Protopapadakis, Athanasios Voulodimos, Anastasios Doulamis, Stephanos Camarinopoulos, Nikolaos Doulamis, and Georgios Miaoulis. Dance pose identification from motion capture data: a comparison of classifiers. *Technologies*, 6(1):31, 2018.

[119] Anastasios D Doulamis, Athanasios Voulodimos, Nikolaos D Doulamis, Sofia Soile, and Anastasios Lampropoulos. Transforming intangible folkloric performing arts into tangible choreographic digital objects: The terpsichore approach. In *VISIGRAPP (5: VISAPP)*, pages 451–460, 2017.

[120] Ioannis Rallis, Ioannis Georgoulas, Nikolaos Doulamis, Athanasios Voulodimos, and Panagiotis Terzopoulos. Extraction of key postures from 3d human motion data for choreography summarization. In *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 94–101. IEEE, 2017.

[121] Damien Tardieu, Xavier Siebert, Stéphane Dupont, Barbara Mazzarino, and B Blumenthal. An interactive installation for browsing a dance video database. In *2010 IEEE International Conference on Multimedia and Expo*, pages 1624–1628. IEEE, 2010.

[122] Jacky CP Chan, Howard Leung, Jeff KT Tang, and Taku Komura. A virtual reality dance training system using motion capture technology. *IEEE transactions on learning technologies*, 4(2):187–195, 2010.

[123] Matija Marolt, Janez Franc Vratanar, and Gregor Strle. Ethnomuse: Archiving folk music and dance culture. In *IEEE EUROCON 2009*, pages 322–326. IEEE, 2009.

[124] Eike Falk Anderson, Leigh McLoughlin, Fotis Liarokapis, Christopher Peters, Panagiotis Petridis, and Sara De Freitas. Developing serious games for cultural heritage: a state-of-the-art review. *Virtual reality*, 14(4):255–275, 2010.

[125] Patsorn Sangkloy Bhavishya Mittal Sean Dai James Hays Daniel Castro, Steven Hickson and Irfan Essa. Let's dance: Learning from online dance videos. In *eprint arXiv:2139179*, 2018.

[126] ACCAD. Open motion data project. https://accad.osu.edu/research/motion-lab/system-data, 2021. Online;.

[127] UCL. Affectme: Affective multimodal engagement. http://web4.cs.ucl.ac.uk/uclic/people/n.berthouze/AffectME.html, 2021. Online;.

[128] Carnegie Mellon. Cmu graphics lab motion capture database. http://mocap.cs.cmu.edu/, 2021. Online;.

[129] Cologne MoCap. Mocap database of th köln - university of applied science. https://mocap.web.th-koeln.de/, 2021. Online;.

[130] Meinard Müller, Tido Röder, Michael Clausen, Bernhard Eberhardt, Björn Krüger, and Andreas Weber. Documentation mocap database hdm05. 2007.

[131] Leonid Sigal, Alexandru O Balan, and Michael J Black. Humaneva: Synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. *International journal of computer vision*, 87(1-2):4, 2010.

[132] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359, 2008.

[133] NUS. Nus capture database. http://mocap.cs.sfu.ca/nusmocap.html, 2021. Online;.

[134] SIG. Center for computer graphics. https://fling.seas.upenn.edu/~mocap/cgi-bin/i.php, 2021. Online;.

[135] Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. Ntu rgb+d: A large scale dataset for 3d human activity analysis. In *IEEE Conference on Computer Vision and Pattern Recognition*, June 2016.

[136] Jun Liu, Amir Shahroudy, Mauricio Perez, Gang Wang, Ling-Yu Duan, and Alex C. Kot. Ntu rgb+d 120: A large-scale benchmark for 3d human activity understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

[137] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. The kit whole-body human motion database. In *International Conference on Advanced Robotics (ICAR)*, pages 329–336, 2015.

[138] Christian Mandery, Ömer Terlemez, Martin Do, Nikolaus Vahrenkamp, and Tamim Asfour. Unifying representations and large-scale whole-body motion databases for studying human motion. *IEEE Transactions on Robotics*, 32(4):796–809, 2016.

[139] Arseny A Sokolov, Peter Zeidman, Michael Erb, Frank E Pollick, Andreas J Fallgatter, Philippe Ryvlin, Karl J Friston, and Marina A Pavlova. Brain circuits signaling the absence of emotion in body language. *Proceedings of the National Academy of Sciences*, 117(34):20868–20873, 2020.

[140] David Baraff. Rigid body simulation. In *Proc. of the SIGGRAPH Course Notes*, volume 19, pages 1–68, 1992.

[141] Sara De Freitas, Genaro Rebolledo-Mendez, Fotis Liarokapis, George Magoulas, and Alexandra Poulovassilis. Learning as immersive experiences: Using the four-dimensional framework for designing and evaluating immersive learning experiences in a virtual world. *British Journal of Educational Technology*, 41(1):69–85, 2010.

[142] Costas Panagiotakis, Anastasios Doulamis, and Georgios Tziritas. Equivalent key frames selection based on iso-content principles. *IEEE Transactions on circuits and systems for video technology*, 19(3):447–451, 2009.

[143] Y.S. Avrithis, A.D. Doulamis, N.D. Doulamis, and S.D. Kollias. Stochastic framework for optimal key frame extraction from mpeg video databases. *Computer Vision and Image Understanding*, 75:3–24, 1999.

[144] Nikolaos Doulamis and Anastasios Doulamis. Non-sequential multiscale content-based video decomposition. *Signal processing*, 85(2):325–356, 2005.

[145] Boon-Lock Yeo and Bede Liu. Rapid scene analysis on compressed video. *IEEE Transactions on circuits and systems for video technology*, 5(6):533–544, 1995.

[146] Farshid Arman, Remi Depommier, Arding Hsu, and M-Y Chiu. Content-based browsing of video sequences. In *Proceedings of the second ACM international conference on Multimedia*, pages 97–103, 1994.

[147] Chong-Wah Ngo, Yu-Fei Ma, and Hong-Jiang Zhang. Video summarization and scene detection by graph modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(2):296–305, 2005.

[148] Arthur G Money and Harry Agius. Video summarisation: A conceptual framework and survey of the state of the art. *Journal of visual communication and image representation*, 19(2):121–143, 2008.

[149] Zheng Lu and Kristen Grauman. Story-driven summarization for egocentric video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[150] Yanwei Fu, Yanwen Guo, Yanshu Zhu, Feng Liu, Chuanming Song, and Zhi-Hua Zhou. Multi-view video summarization. *IEEE Transactions on Multimedia*, 12(7):717–729, 2010.

[151] Nikolaos Doulamis, Anastasios Doulamis, Yannis Avrithis, Klimis Ntalianis, and Stefanos Kollias. Efficient summarization of stereoscopic video sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 10:501–517, 2000.

[152] Markus Windolf, Nils Götzen, and Michael Morlock. Systematic accuracy and precision analysis of video motion capturing systems—exemplified on the vicon-460 system. *Journal of Biomechanics*, 41(12):2776–2780, 2008.

[153] Jeremy Davis, James Doebbler, John Junkins, Matthew Vavrinax, and John Vian. Characterizing and calibrating the novel phasespace camera system. In *Proc. AIAA Guidance, Navigation, and Control Conference*, 2011.

[154] OptiTrak. Optitrack motion capture system. http://optitrak.com, 2017. Online;.

[155] Y. Kim, S. Baek, and B.-C. Bae. Motion capture of the human body using multiple depth sensors. *ETRI Journal*, 39:181–190, 2017.

[156] Alexandra Pfister, Alexandre M. West, Shaw Bronner, and Jack Adam Noah. Comparative abilities of microsoft kinect and vicon 3d motion capture for gait analysis. *Journal of Medical Engineering & Technology*, 38(5):274–280, 2014.

[157] Andreas Aristidou, Panayiotis Charalambous, and Yiorgos Chrysanthou. Emotion analysis and classification: Understanding the performers' emotions using the lma entities. 34(6):262–276, 2015.

[158] Fadi Dornaika and Ihab Kamal Aldine. Decremental sparse modeling representative selection for prototype selection. *Pattern Recognition*, 48(11):3714–3727, 2015.

[159] Nikolaos Doulamis and Anastasios Doulamis. Non-sequential multiscale content-based video decomposition. *Signal processing*, 85(2):325–356, 2005.

[160] Robert Laganière, Raphael Bacco, Arnaud Hocevar, Patrick Lambert, Grégory Païs, and Bogdan E Ionescu. Video summarization from spatio-temporal features. In *Proceedings of the 2nd ACM TRECVid Video Summarization Workshop*, pages 144–148, 2008.

[161] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn*, 3:1–122, 2011.

[162] Renata M. Sheppard, Mahsa Kamali, Raoul Rivas, Morihiko Tamai, Zhenyu Yang, Wanmin Wu, and Klara Nahrstedt. *Advancing interactive collaborative mediums through tele-immersive dance (TED): A symbiotic creativity and design environment for art and computer science*, pages 579–588. 12 2008.

[163] Andreas Aristidou, Efstathios Stavrakis, Panayiotis Charalambous, Yiorgos Chrysanthou, and Stephania Loizidou Himona. Folk dance evaluation using laban movement analysis. *Journal on Computing and Cultural Heritage*, 8:20:1–20:19, 08 2015.

[164] Victor Brian Zordan, Anna Majkowska, Bill Chiu, and Matthew Fast. Dynamic response for motion capture animation. *ACM Trans. Graph.*, 24(3):697–701, July 2005.

[165] Lucas Kovar and Michael Gleicher. Automated extraction and parameterization of motions in large data sets. In *ACM Transactions on Graphics (ToG)*, volume 23, pages 559–568. ACM, 2004.

[166] Cheng Chen, Yueting Zhuang, Feiping Nie, Yi Yang, Fei Wu, and Jun Xiao. Learning a 3d human pose distance metric from geometric pose descriptor. *IEEE Transactions on Visualization and Computer Graphics*, 17(11):1676–1689, 2011.

[167] Kevin Forbes and Eugene Fiume. An efficient search algorithm for motion data using weighted pca. In *Proceedings of the 2005 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 67–76. ACM, 2005.

[168] Jinxiang Chai and Jessica K Hodgins. Performance animation from low-dimensional control signals. In *ACM Transactions on Graphics (ToG)*, volume 24, pages 686–696. ACM, 2005.

[169] Björn Krüger, Jochen Tautges, Andreas Weber, and Arno Zinke. Fast local and global similarity searches in large motion capture databases. In *Proceedings of the 2010 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 1–10. Eurographics Association, 2010.

[170] Eftychios Protopapadakis, Athanasios Voulodimos, Anastasios Doulamis, Stephanos Camarinopoulos, Nikolaos Doulamis, and Georgios Miaoulis. Dance pose identification from motion capture data: A comparison of classifiers. *Technologies*, 6(1), 2018.

[171] Meinard Müller, Tido Röder, and Michael Clausen. Efficient content-based retrieval of motion capture data. In *ACM Transactions on Graphics (ToG)*, volume 24, pages 677–685. ACM, 2005.

[172] Meinard Müller and Tido Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 137–146. Eurographics Association, 2006.

[173] Athanasios S. Voulodimos, Dimitrios I. Kosmopoulos, Nikolaos D. Doulamis, and Theodora A. Varvarigou. A top-down event-driven approach for concurrent activity recognition. *Multimedia Tools and Applications*, 69(2):293–311, Mar 2014.

[174] Nikolaos D. Doulamis, Athanasios S. Voulodimos, Dimitrios I. Kosmopoulos, and Theodora A. Varvarigou. Enhanced human behavior recognition using hmm and evaluative rectification. In *Proceedings of the First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams*, ARTEMIS '10, pages 39–44, New York, NY, USA, 2010. ACM.

[175] Mubbasir Kapadia, I-kao Chiang, Tiju Thomas, Norman I Badler, Joseph T Kider Jr, et al. Efficient motion retrieval in large motion databases. In *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, pages 19–28. ACM, 2013.

[176] Feng Liu, Yueting Zhuang, Fei Wu, and Yunhe Pan. 3d motion retrieval with motion index tree. *Computer Vision and Image Understanding*, 92(2-3):265–284, 2003.

[177] Jernej Barbič, Alla Safonova, Jia-Yu Pan, Christos Faloutsos, Jessica K Hodgins, and Nancy S Pollard. Segmenting motion capture data into distinct behaviors. In *Proceedings of Graphics Interface 2004*, pages 185–194. Canadian Human-Computer Communications Society, 2004.

[178] Anna Vögele, Björn Krüger, and Reinhard Klein. Efficient unsupervised temporal segmentation of human motion. In *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 167–176. Eurographics Association, 2014.

[179] Matthew Field, David Stirling, Zengxi Pan, Montserrat Ros, and Fazel Naghdy. Recognizing human motions through mixture modeling of inertial data. *Pattern Recognition*, 48(8):2394–2406, 2015.

[180] Eftychios Protopapadakis, Athanasios Voulodimos, and Nikolaos Doulamis. Multidimensional trajectory similarity estimation via spatial-temporal keyframe selection and signal correlation analysis. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, PETRA '18, pages 91–97, New York, NY, USA, 2018. ACM.

[181] Feng Zhou, Fernando De la Torre, and Jessica K Hodgins. Hierarchical aligned cluster analysis for temporal clustering of human motion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):582–596, 2012.

[182] Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, KangKang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. Emotion control of unstructured dance movements. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 1–10, 2017.

[183] Meinard Müller, Andreas Baak, and Hans-Peter Seidel. Efficient and robust annotation of motion capture data. In *Proceedings of the 2009 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 17–26. ACM, 2009.

[184] Jürgen Bernard, Nils Wilhelm, Björn Krüger, Thorsten May, Tobias Schreck, and Jörn Kohlhammer. Motionexplorer: Exploratory search in human motion capture data based on hierarchical aggregation. *IEEE transactions on visualization and computer graphics*, 19(12):2257–2266, 2013.

[185] Songle Chen, Zhengxing Sun, and Yan Zhang. Scalable organization of collections of motion capture data via quantitative and qualitative analysis. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 411–418. ACM, 2015.

[186] Ehsan Elhamifar, Guillermo Sapiro, and René Vidal. See all by looking at a few: Sparse modeling for finding representative objects. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1600–1607, 2012.

[187] Robert Laganière, Raphael Bacco, Arnaud Hocevar, Patrick Lambert, Grégory Païs, and Bogdan E. Ionescu. Video summarization from spatio-temporal features. In *Proceedings of the 2Nd ACM TRECVid Video Summarization Workshop*, TVS '08, pages 144–148, New York, NY, USA, 2008. ACM.

[188] Tong Wu, Prudhvi Gurram, Raghuveer M Rao, and Waheed U Bajwa. Hierarchical union-of-subspaces model for human activity summarization. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[189] David Baraff. Physically based modeling: Principles and practice implicit methods for differential equations. In *SIGGRAPH*, volume 97, pages E1–E4, 1997.

[190] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[191] Konstantinos Makantasis, Anastasios Doulamis, Nikolaos Doulamis, and Marinos Ioannides. In the wild image retrieval and clustering for 3d cultural heritage landmarks reconstruction. *Multimedia Tools and Applications*, 75(7):3593–3629, 2016.

[192] Peter J Rousseeuw. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[193] David L Davies and Donald W Bouldin. A cluster separation measure. *IEEE transactions on pattern analysis and machine intelligence*, (2):224–227, 1979.

[194] Kohei Arai and Ali Ridho Barakbah. Hierarchical k-means: an algorithm for centroids initialization for k-means. *Reports of the Faculty of Science and Engineering*, 36(1):25–31, 2007.

[195] Yiming Yang and Xin Liu. A re-examination of text categorization methods. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 42–49. ACM, 1999.

[196] Nikolaos D Doulamis, Anastasios D Doulamis, Y Avrithis, and Stefanos D Kollias. A stochastic framework for optimal key frame extraction from mpeg video databases. pages 141–146, 1999.

[197] Yujia Zhang, Xiaodan Liang, Dingwen Zhang, Min Tan, and Eric P Xing. Unsupervised object-level video summarization with online motion auto-encoder. *Pattern Recognition Letters*, 130:376–385, 2020.

[198] Anastasios D Doulamis, Nikolaos D Doulamis, and Stefanos D Kollias. On-line retrainable neural networks: improving the performance of neural networks in image analysis problems. *IEEE Transactions on Neural Networks*, 11(1):137–155, 2000.

[199] Anastasios D Doulamis and Nikolaos D Doulamis. Optimal content-based video decomposition for interactive video navigation. *IEEE Transactions on Circuits and Systems for Video Technology*, 14(6):757–775, 2004.

[200] Nuno Vasconcelos and Andrew Lippman. A spatiotemporal motion model for video summarization. In *Proceedings. 1998 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No. 98CB36231)*, pages 361–366. IEEE, 1998.

[201] Andreas Aristidou, Daniel Cohen-Or, Jessica K Hodgins, Yiorgos Chrysanthou, and Ariel Shamir. Deep motifs and motion signatures. *ACM Transactions on Graphics (TOG)*, 37(6):1–13, 2018.

[202] Athanasios Voulodimos, Ioannis Rallis, and Nikolaos Doulamis. Physics-based keyframe selection for human motion summarization. *Multimedia Tools and Applications*, 79(5):3243–3259, 2020.

[203] Andreas Aristidou and Joan Lasenby. Inverse kinematics: a review of existing techniques and introduction of a new fast iterative solver. 2009.

[204] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[205] Eftychios Protopapadakis, Athanasios Voulodimos, Anastasios Doulamis, Nikolaos Doulamis, Dimitrios Dres, and Matthaios Bimpas. Stacked autoencoders for outlier detection in over-the-horizon radar signals. *Computational intelligence and neuroscience*, 2017, 2017.

[206] Stéphane Mallat. *A wavelet tour of signal processing*. Elsevier, 1999.

[207] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice-Hall, Inc., 2007.

[208] Agus Saptoro, Moses O Tadé, and Hari Vuthaluru. A modified kennard-stone algorithm for optimal division of data for developing artificial neural network models. *Chemical Product and Process Modeling*, 7(1), 2012.

[209] Anthony Shay and Barbara Sellers-Young. Dance and ethnicity. In *The Oxford Handbook of Dance and Ethnicity*. 2016.

[210] Marinos Ioannides, A Hadjiprocopi, Nikolaos Doulamis, Anastasios Doulamis, Eft Protopapadakis, Kostas Makantasis, Pedro Santos, Dieter Fellner, Andre Stork, Olivier Balet, et al. Online 4d reconstruction using multi-images available under open access. *ISPRS Photogr. Rem. Sens. Spat. Inf. Sc.*, 2:169–174, 2013.

[211] Georgia Kyriakaki, Anastasios Doulamis, Nikolaos Doulamis, Marinos Ioannides, Konstantinos Makantasis, Eftichios Protopapadakis, Andreas Hadjiprocopis, Konrad Wenzel, Dieter Fritsch, Michael Klein, et al. 4d reconstruction of tangible cultural heritage objects from web-retrieved images. *International Journal of Heritage in the Digital Era*, 3(2):431–451, 2014.

[212] Anastasios D Doulamis, Nikolaos D Doulamis, Konstantinos Makantasis, and Michael Klein. A 4d virtual/augmented reality viewer exploiting unstructured web-based image data. In *VISAPP (2)*, pages 631–639, 2015.

[213] Leonid Keselman, John Iselin Woodfill, Anders Grunnet-Jepsen, and Achintya Bhowmik. Intel realsense stereoscopic depth cameras. *arXiv preprint arXiv:1705.05548*, 2017.

[214] Nikolaos Doulamis, Anastasios Doulamis, Charalabos Ioannidis, Michael Klein, and Marinos Ioannides. Modelling of static and moving objects: digitizing tangible and intangible cultural heritage. In *Mixed Reality and Gamification for Cultural Heritage*, pages 567–589. Springer, 2017.

[215] Alexandros Kitsikidis, Kosmas Dimitropoulos, Erdal Yilmaz, Stella Douka, and Nikos Grammalidis. Multi-sensor technology and fuzzy logic for dancer's motion analysis and performance evaluation within a 3d virtual environment. In *International Conference on Universal Access in Human-Computer Interaction*, pages 379–390. Springer, 2014.

[216] Apostolos Laggis, Nikolaos Doulamis, Eftychios Protopapadakis, and Andreas Georgopoulos. A low-cost markerless tracking system for trajectory interpretation. In *ISPRS International Workshop of 3D Virtual Reconstruction and Visualization of Complex Arhitectures*, 2017.

[217] Athanasios Voulodimos, Dimitrios Kosmopoulos, Galina Veres, Helmut Grabner, Luc Van Gool, and Theodora Varvarigou. Online classification of visual tasks for industrial workflow monitoring. *Neural Networks*, 24(8):852 – 860, 2011.

[218] Dimitrios I. Kosmopoulos, Athanasios S. Voulodimos, and Theodora A. Varvarigou. Robust human behavior modeling from multiple cameras. In *2010 20th International Conference on Pattern Recognition*, pages 3575–3578, 2010.

[219] Nikolaos Doulamis and Athanasios Voulodimos. Fast-mdl: Fast adaptive supervised training of multi-layered deep learning models for consistent object tracking and classification. In *2016 IEEE International Conference on Imaging Systems and Techniques (IST)*, pages 318–323, Oct 2016.

[220] Nikolaos D. Doulamis, Athanasios S. Voulodimos, Dimitrios I. Kosmopoulos, and Theodora A. Varvarigou. Enhanced human behavior recognition using hmm and evaluative rectification. In *Proceedings of the First ACM International Workshop on Analysis and Retrieval of Tracked Events and Motion in Imagery Streams*, ARTEMIS '10, pages 39–44, New York, NY, USA, 2010. ACM.

[221] Ioannis Rallis, Ioannis Georgoulas, Nikolaos Doulamis, Athanasios Voulodimos, and Panagiotis Terzopoulos. Extraction of key postures from 3d human motion data for choreography summarization. In *2017 9th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 94–101. IEEE, 2017.

[222] Alexandra Pfister, Alexandre M West, Shaw Bronner, and Jack Adam Noah. Comparative abilities of microsoft kinect and vicon 3d motion capture for gait analysis. *Journal of medical engineering & technology*, 38(5):274–280, 2014.

[223] Adso Fern'ndez-Baena, Antonio Susín, and Xavier Lligadas. Biomechanical validation of upper-body and lower-body joint movements of kinect motion capture data for rehabilitation treatments. In *2012 fourth international conference on intelligent networking and collaborative systems*, pages 656–661. IEEE, 2012.

[224] Brook Galna, Gillian Barry, Dan Jackson, Dadirayi Mhiripiri, Patrick Olivier, and Lynn Rochester. Accuracy of the microsoft kinect sensor for measuring movement in people with parkinson's disease. *Gait & posture*, 39(4):1062–1068, 2014.

[225] K. Adistambha, C. H. Ritz, and I. S. Burnett. Motion classification using dynamic time warping. In *2008 IEEE 10th Workshop on Multimedia Signal Processing*, pages 622–627, Oct 2008.

[226] Hyo-Rim Choi and TaeYong Kim. Combined dynamic time warping with multiple sensors for 3d gesture recognition. *Sensors*, 17(8):1893, 2017.

[227] Nazlı Ikizler and Pınar Duygulu. Human action recognition using distribution of oriented rectangular patches. In *Workshop on Human Motion*, pages 271–284. Springer, 2007.

[228] Jaron Blackburn and Eraldo Ribeiro. Human motion recognition using isomap and dynamic time warping. In *Workshop on Human Motion*, pages 285–298. Springer, 2007.

[229] Anastasios Doulamis, Nikolaos Doulamis, Charalabos Ioannidis, Christina Chrysouli, Nikos Grammalidis, Kosmas Dimitropoulos, Chryssy Potsiou, Elisavet Konstantina Stathopoulou, and Marinos Ioannides. 5d modelling: An efficient approach for creating spatiotemporal predictive 3d maps of large-scale cultural resources. *ISPRS Annals of Photogrammetry, Remote Sensing & Spatial Information Sciences*, 2015.

[230] Donald J Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD workshop*, volume 10, pages 359–370. Seattle, WA, 1994.

[231] A. Stefan, V. Athitsos, and G. Das. The move-split-merge metric for time series. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1425–1438, June 2013.

[232] Eftychios Protopapadakis, Athanasios Voulodimos, and Nikolaos Doulamis. Multidimensional trajectory similarity estimation via spatial-temporal keyframe selection and signal correlation analysis. In *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, pages 91–97. ACM, 2018.

[233] Luigi Fortuna, Mattia Frasca, and Cristoforo Camerano. Strange attractors, kinematic trajectories and synchronization. *International Journal of Bifurcation and Chaos*, 18(12):3703–3718, 2008.

[234] Hui-mei Justina Hsu. The potential of kinect in education. *International Journal of Information and Education Technology*, 1(5):365, 2011.

[235] Alexandros Kitsikidis, Kosmas Dimitropoulos, Deniz Uğurca, Can Bayçay, Erdal Yilmaz, Filareti Tsalakanidou, Stella Douka, and Nikos Grammalidis. A game-like application for dance learning using a natural human computer interface. In *International Conference on Universal Access in Human-Computer Interaction*, pages 472–482. Springer, 2015.

[236] Kazuya Kojima, Kozaburo Hachimura, and Minako Nakamura. Labaneditor: Graphical editor for dance notation. In *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*, pages 59–64. IEEE, 2002.

[237] Andreas Aristidou, Efstathios Stavrakis, Margarita Papaefthimiou, George Papagiannakis, and Yiorgos Chrysanthou. Style-based motion analysis for dance composition. *The visual computer*, 34(12):1725–1737, 2018.

[238] Kozaburo Hachimura and Minako Nakamura. Method of generating coded description of human body motion from motion-captured data. In *Robot and Human Interactive Communication, 2001. Proceedings. 10th IEEE International Workshop on*, pages 122–127. IEEE, 2001.

[239] Jiaji Wang, Zhenjiang Miao, Hao Guo, Ziming Zhou, and Hao Wu. Using automatic generation of labanotation to protect folk dance. *Journal of Electronic Imaging*, 26(1):011028, 2017.

[240] Anastasios Ballas, Tossaporn Santad, Kingkarn Sookhanaphibarn, and Worawat Choensawat. Game-based system for learning labanotation using microsoft kinect. In *Consumer Electronics (GCCE), 2017 IEEE 6th Global Conference on*, pages 1–3. IEEE, 2017.

[241] Anastasios D Doulamis, Nikolaos D Doulamis, and Stefanos D Kollias. 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, 2003.

[242] Mike Schuster and Kuldip K Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

[243] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[244] Gobinda G Chowdhury. *Introduction to modern information retrieval.* Facet publishing, 2010.

[245] Richard Kurin. Safeguarding intangible cultural heritage in the 2003 unesco convention: a critical appraisal. *Museum international*, 56(1-2):66–77, 2004.

[246] Ioannis Rallis, Nikolaos Bakalos, Nikolaos Doulamis, Athanasios Voulodimos, Anastasios Doulamis, and Eftychios Protopapadakis. Learning choreographic primitives through a bayesian optimized bi-directional lstm model. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1940–1944. IEEE, 2019.

[247] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[248] Konstantinos Makantasis, Anastasios Doulamis, Nikolaos Doulamis, and Antonis Nikitakis. Tensor-based classifiers for hyperspectral data analysis. *arXiv preprint arXiv:1709.08164*, 2017.

[249] Gül Varol, Ivan Laptev, and Cordelia Schmid. Long-term temporal convolutions for action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6):1510–1517, 2018.

[250] Aouaidjia Kamel, Bin Sheng, Po Yang, Ping Li, Ruimin Shen, and David Dagan Feng. 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, 2019.

[251] Nikolaos Bakalos, Ioannis Rallis, Nikolaos Doulamis, Anastasios Doulamis, Eftychios Protopapadakis, and Athanasios Voulodimos. Choreographic pose identification using convolutional neural networks. In *2019 11th International Conference on Virtual Worlds and Games for Serious Applications (VS-Games)*, pages 1–7. IEEE, 2019.

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

[253] Konstantinos Makantasis, Antonios Nikitakis, Anastasios D Doulamis, Nikolaos D Doulamis, and Ioannis Papaefstathiou. 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, 2017.

[254] Yuzhu Dong, Andreas Aristidou, Ariel Shamir, Moshe Mahler, and Eakta Jain. Adult2child: Motion style transfer using cyclegans. In *Motion, Interaction and Games*, pages 1–11. 2020.

[255] Maria Kaselimi, Nikolaos Doulamis, Athanasios Voulodimos, Anastasios Doulamis, and Eftychios Protopapadakis. Energan++: A generative adversarial gated recurrent network for robust energy disaggregation. *IEEE Open Journal of Signal Processing*, 2:1–16, 2020.

[256] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara Berg. Dance dance generation: Motion transfer for internet videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019.

[257] Maria Kaselimi, Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. A robust to noise adversarial recurrent model for non-intrusive load monitoring. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3335–3339. IEEE, 2021.

[258] Maria Kaselimi, Athanasios Voulodimos, Eftychios Protopapadakis, Nikolaos Doulamis, and Anastasios Doulamis. Energan: A generative adversarial network for energy disaggregation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1578–1582. IEEE, 2020.

[259] Maria Kaselimi, Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. A generative adversarial gated recurrent network for power disaggregation & consumption awareness.

[260] Andreas Aristidou, Qiong Zeng, Efstathios Stavrakis, KangKang Yin, Daniel Cohen-Or, Yiorgos Chrysanthou, and Baoquan Chen. Emotion control of unstructured dance movements. In *Proceedings of the ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 1–10, 2017.

[261] Athanasios Voulodimos, Eftychios Protopapadakis, Iason Katsamenis, Anastasios Doulamis, and Nikolaos Doulamis. A few-shot u-net deep learning model for covid-19 infected area segmentation in ct images. *Sensors*, 21(6):2215, 2021.

[262] Iason Katsamenis, Eftychios Protopapadakis, Athanasios Voulodimos, Anastasios Doulamis, and Nikolaos Doulamis. Transfer learning for covid-19 pneumonia detection and classification in chest x-ray images. *medRxiv*, 2020.