Computers & Graphics xxx (2018) xxx–xxx

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Q1 Section on Computer Graphics for Serious Games and Virtual Environments

Spatio-temporal summarization of dance choreographies

Q2 Ioannis Rallis[a], Nikolaos Doulamis[a,1,*], Anastasios Doulamis[a], Athanasios Voulodimos[b], Vassilios Vescoukis[a]

[a] National Technical University of Athens, Athens, 9 Heroon Polytechniou Street, 15773 Zografou, Greece
[b] University of West Attica, Athens, Agiou Spyridonos Street, 12243 Egaleo, Greece

ARTICLE INFO

Article history:
Received 5 January 2018
Revised 11 March 2018
Accepted 9 April 2018
Available online xxx

Keywords:
Dance summarization
Spatio-temporal clustering
Hierarchical decomposition
Skeleton 3D joints
SMRS

ABSTRACT

An important issue in performing dance analysis is the automatic extraction of its choreographic patterns, since these elements provide an abstract representation of the semantics of the dance and encode 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 often repeated in time. Additionally, 3D geometry is lost using color information. For this reason, in this paper we propose a new dance summarization scheme of 3D motion captured data (in the form of skeleton joints coordinates) recorded using the Vicon motion capture system. The proposed key frame extraction method implements a hierarchical scheme that exploits spatio-temporal variations of dance features. Initially, global holistic descriptors are extracted to localize the key choreographic steps of a dance (coarse representation). Then, each segment is further decomposed into finer sub-segments to improve dance representativity (fine representation). The 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 on thirty folkloric dance sequences recorded at the Aristotle University of Thessaloniki under the framework of Terpsichore project representing five different choreographies and on publicly available datasets from Carnegie–Mellon University, which depict performances on theatrical kinesiology. Comparisons with other traditional video summarization methods indicate a clear superiority of the proposed hierarchical spatio-temporal decomposition scheme.

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 [1]. Such an abstract content representation is useful in many applications ranging from multimedia systems (e.g., indexing, browsing, content-based search and retrieval) [2] and education (e.g., teaching/learning of a dance choreography) [3,4] to documentation and preservation of the Intangible Cultural Heritage (ICH) assets, [5].

Extraction of representative key frames for an abstract description of a video sequence, is an important topic in multimedia research [6,7]. Actually, video summarization algorithms are content-based sampling procedures that reduce semantically unimportant or redundant content. One of the first approaches towards video summarization is the extraction of scene (or shot) video segments within a video [8]. 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 [9–11] and [12]. These algorithms take visual data in the RGB or HSV color space and appropriately process them to extract feature-related transformations.

However, recent advances in software and especially hardware engineering have led to the emergence of several new devices for capturing, storing and acquiring video content. The innovation of these acquisition systems is that they capture, apart from color, the

* Corresponding author.
E-mail addresses: irallis@central.ntua.gr (I. Rallis), ndoulam@cs.ntua.gr (N. Doulamis), adoulam@cs.ntua.gr (A. Doulamis), thanosv@mail.ntua.gr (A. Voulodimos), v.vescoukis@cs.ntua.gr (V. Vescoukis).
[1] Editor-in-Chief, Computers and Graphics Journal.

https://doi.org/10.1016/j.cag.2018.04.003
0097-8493/

depth information providing, therefore, new ways for modelling human body movements and gestures. Examples include Vicon, Kinect and PhaseSpace systems which have been used in many diverse application scenarios including gaming, film, animation and the sports industry. Such devices detect and track in space and time a set of key points in order to form a three-dimensional (3D) representation of human body motion. Exploiting the capabilities of the aforementioned devices, one could improve the performance and efficiency of video summarization, especially when it targets the detection of choreographic patterns or the analysis of human motion trajectories.

Video summarization algorithms are distinguished into two main categories. The first one 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 Doulamis et al. [12–15]. 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 [16], which extracts key frame representatives utilizing a curvature metric on the time trajectory of the features. Other examples include the work of Laganire et al. [17] that proposes spatial-temporal activity features, and [18], which 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 to micro-variations of dancer steps. This leads to an over-representation modelling of the content, i.e., 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 reject noisy key-frames [16]. However, the bandwidth of the low-pass filter significantly affects summarization performance and the definition of its proper value highly depends on the specific properties of the choreography, the tempo and the dancer's style.

In this paper, we introduce a spatio-temporal summarization algorithm that considers 3D motion captured data, instead of *RGB* information, represented by 3D joints that model human skeleton. In particular, in our 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 proposed spatio-temporal approach is implemented under a hierarchical framework. More specifically, for a given dance segment, initially global holistic descriptors are extracted to localize the key choreographic steps of the dance, derived from the 3D human joints. Then, each segment is further decomposed into more detailed sub-segments, refining the extracted initial (coarse) key representatives. In this way, we combine global with local modelling that better capture the temporal attributes of a dance. This hierarchical 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, such as fast browsing, storytelling, indexing and content-based retrieval.

1.1. Previous 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 Hisatomi et al. [19] 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 [20] digitizes Cypriot dances using the Phasespace Impulse X2 motion capture system. This architecture uses 8-cameras that are able to capture 3D motion on modulated LEDs. In the same work, a video game is developed for making the teaching of Cypriot dances more attractive. In [21], 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 dance is proposed in [22]. The purpose of this research is to design a creativity framework for dance choreography based on LMA (Laban Movement Analysis) [23]. Advanced motion captured architectures for digitizing folklore performing arts is presented in [24]. In this work, motion analysis algorithms are investigated with the main aim of transforming 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 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 [25], Phasespace capturing [24] or Vicon [21] motion interface can be exploited.

Regarding choreographic analysis approaches, classification algorithms have been proposed on data expressing human body movements. In this context, the work of Raptis et al. [26] proposes a real-time classification system in recognizing choreographed gesture classes. The input data have been acquired using the Kinect depth sensor [27], extracting a 3D wireframe skeleton of dancers. Another dance classification approach is proposed in [28] using again data captured from the Kinect sensor. In particular, the authors of [28] combine 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 Fishers linear discriminant analysis, which is called fisherdance, is proposed in [29], for classifying Korean pop dances, using Kinect sensor as the input data source.

A dance recognition system is introduced in [30]. 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 [25,31] 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 [1]. This scheme exploits input data from a Vicon motion capturing system and then applies a k-means clustering 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 work of Wu et al. [18] proposes a hierarchical union of

Please cite this article as: I. Rallis et al., Spatio-temporal summarization of dance choreographies, Computers & Graphics (2018), https://doi.org/10.1016/j.cag.2018.04.003