# Beat Synchronous Dance Animation based on Visual Analysis of Human Motion and Audio Analysis of Music Tempo

Costas Panagiotakis[1], Andre Holzapfel[2], Damien Michel[3], and Antonis A. Argyros[4,3]

[1] Dept. of Commerce and Marketing, TEI of Crete, Greece, cpanag@csd.uoc.gr
[2] Universitat Pompeu Fabra, Barcelona, Spain, andre.holzapfel@upf.edu
[3] Institute of Computer Science, FORTH, Crete, Greece, michel@ics.forth.gr
[4] Computer Science Department, University of Crete, Greece, argyros@ics.forth.gr

**Abstract.** We present a framework that generates beat synchronous dance animation based on the analysis of both visual and audio data. First, the articulated motion of a dancer is captured based on markerless visual observations obtained by a multicamera system. We propose and employ a new method for the temporal segmentation of such motion data into the periods of dance. Next, we use a beat tracking algorithm to estimate the pulse related to the tempo of a piece of music. Given an input music that is of the same genre as the one corresponding to the visually observed dance, we automatically produce a beat synchronous dance animation of a virtual character. The proposed approach has been validated with extensive experiments performed on a data set containing a variety on traditional Greek/Cretan dances and the corresponding music.

## 1 Introduction

Synthesizing realistic human or animal motion is a very important research topic in computer animation with a high number of applications like virtual reality, computer games, movies and entertainment systems [1–3]. Motion synthesis algorithms usually should take into account several constraints in order to create realistic animations that are related with the virtual environment of animation. For example, to achieve realistic dance animation synthesis, the motion of the virtual character should synchronize with music. The rhythm of dance music can be considered to be based on several related periodicities. The periodicity which is the most convenient to cause a human to move his body is referred to as the beat. The problem of detecting the beat has recently attracted considerable research interest [1, 3, 4].

Usually, 3D motion dance data are available e.g. by capturing devices or by motion synthesis algorithms. So, a problem of a great interest is to synchronize them to a given dance music. Given proper and automatic synchronization, the realism of the resulting audiovisual experience is increased. Such a capability

is also expected to contribute to the development of important applications regarding the demonstration, study, teaching, spread and preservation of music and dances.

This paper addresses directly the aforementioned problem when the dance is periodic, meaning that it consists of repetitive motion patterns. The target case study considers traditional Cretan dances. Nevertheless, the adopted approach is applicable to a much broader class of dances. In order to solve the problem, we employ signal processing techniques for combined music and motion analysis, that is a vibrant and rapidly evolving field of research [5]. A growing trend in music and motion analysis is to tackle the problems globally and to exploit, whenever possible, the multimodal or multi-faceted aspects of music and motion.

A lot of research has been already devoted to the motion analysis of dance videos in order to estimate the rhythm of motion. In [6], a method of rhythmic information extraction from dance videos and music has been proposed. The rhythm of motion is estimated by the analysis of motion trajectories of points that are detected using an adaptation of the Shi-Tomasi (ST) corner detector [7]. Since the 2D visual information is not always sufficient to solve the problem of motion rhythm estimation with high accuracy, other methods [4] have been applied to 3D motion capture data. The method in [4] first detects rapid directional change of joints, estimating candidate beats and then transform this information to continuous motion signals using sequential cosine functions. Finally, the power spectrum density of signals is analyzed to estimate the dominant period.

In [8], two methods have been presented for segmenting periodic human motion capture data for mobile gait analysis. The first method is a model-based algorithm which operates directly on the joint angles detecting local minima and maxima. By considering pairs of successive minima and maxima, it is possible to identify distinct intervals in a periodic motion. However, this method suffers from some limitations, as it might still be possible to get conflicting segmentation sets between different joints. In such cases, deciding which joint produces more reliable results is a difficult task. The second method is a model-free, Latent Space algorithm, a dimensionality reduction method which first aggregates all the sensor data and transforms them into an 1D signal. Finally, the segmentation is given by the detection of local minima and maxima in the resulting 1D signal.

Beat detection in music aims at automatic estimation of the time instances where a human listener would tap his foot to the music. There have been several methods presented over the last decades. The approach of Klapuri [9] is widely considered as a state-of-the-art approach. While in the years after the publication of [9] quite innovative approaches were presented (*e.g.* [10]), no large improvement in general accuracy has been observed. As we aim at synchronising dance movement with audio signals in the context of traditional dances, we will apply the modification of the Klapuri method as proposed in [11]. This modification uses a signal representation derived from phase characteristics as input to a beat tracker similar to [9], which was shown to improve the alignment of the beat sequence to the audio signal in the case of traditional dances [11].

The beat detection in music has been used by several methods that aim to create new unseen dance animations that are synchronized with a given music [1,3]. In [1], a fast, greedy algorithm analyzes a library of stock motions and generates new sequences of movements that were not described in the library. The greedy algorithm with backtracking tries to find the best matching frame among the closest dance moves, take it as a greedy choice and repeats the same process. A second, genetic algorithm tries to optimize the dance sequence by taking a number of valid random dance figures as a population and applies the genetic operators of crossover and mutation to create new generations. In [3], the generation of dance performances is based on a given musical piece by matching the progressions of musical and motion patterns and by correlating musical and motion features. The proposed method uses similarity matrices for musical and motion sequences and matched the progressions of musical and motion contents by minimizing the difference between the two similarity matrices.

Most of the existing approaches that try to provide a temporal segmentation of human motion are heuristic and use simplifications or signal approximations without any global optimality criterion. Many approaches based on visual information use 2D tracking data and suffer from visual limitations like occlusions and noise. In addition, many approaches can be only applied to simple human motions (e.g. walking), where the period can be defined by the local minima and maxima of the signal. Moreover, certain methods synthesize new unseen animations that are synchronized with a given music.

On the contrary, in this paper, instead of creating new unseen animations (that usually requires a high number of 3D motion datasets), we solve the problem of synchronizing the 3D motion of a given dance with a given music. We have proposed an optimization approach that computes the optimal solution for the problem of temporal segmentation of human motion using 3D dance motion data. An advantage of the proposed method is that it can be applied to complex multidimensional signals such as those representing dance movements. We also apply an autocorrelation based criterion in order to segment the music signal into periods. This information is then used to produce beat synchronous dance animations. The experimental results show that the proposed method achieves very promising results. It should be also noted that the input to the algorithm is not marker-based motion capture data but rather data produced by a home-build markerless human articulation tracking data. In that sense, the proposed approach is also capable of tolerating noise in the representation of human motion.

The rest of the paper is organized as follows. Section 2 gives a brief overview of the proposed approach. Sections 3, 4 and 5 present the details of the three main building blocks of the proposed method, that is, temporal segmentation of periodic human motion, music beat detection algorithm and beat synchronous dance animation creation, respectively. The experimental results are given in Section 6. Finally, a summary of this work and the main directions of future work are provided in Section 7.

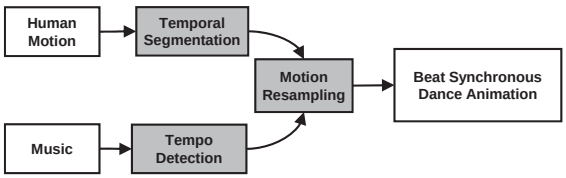



Fig. 1: Scheme of the proposed system architecture.

## 2 Overview of the Proposed Approach

An overview of the proposed approach is illustrated in Fig. 1. The input to our method consists of (a) motion capture data, that is, the 3D position, orientation and articulation of the body parts of a human dancer while dancing a particular periodic dance and (b) the acoustic signal of a music that is compatible to that dance genre. The goal is to animate automatically a virtual/synthetic character who dances according to (a) but is also synchronized to the rhythm of (b). To achieve this result, the proposed approach employs three building blocks. The first one consists of an efficient algorithm for the temporal segmentation of the complex human motion capture data into the periods of dance. The second building block segments the acoustic signal into parts of duration equal to the music tempo period. Finally, a motion resampling algorithm is responsible for remapping each period of the motion capture data according to the estimated music tempo and for producing the beat synchronous dance animation. The following three sections present these building blocks in more detail.

## 3 Temporal Segmentation of Periodic Human Motion

The input to the temporal segmentation of human motion is the time series of the joint angles of an articulated human model. These joint angles are estimated by a recently proposed method [12] that relies on markerless, multicamera observations of a moving person. The employed method estimates accurately the parameters of an articulated human body model that has 11 joints and a total of 29 degrees of freedom.

Let $S \in \Re^{m,n}$ be the given multidimensional signal of captured human motion that contains the time series of the $m = 29$ degrees of freedom (i.e., joint angle) of the human motion. Let also $n$ be the number of temporal samples of each of these series. $S_i(j)$ denotes the $j$-sample of $i$-angle time series, $i \in \{1, ..., m\}$, $j \in \{1, ..., n\}$. Assuming that the human motion (dance in our case) is periodic, the goal of temporal segmentation is to segment $S$ into its periods. Let $T_p = \{t_0, t_1, t_2, ..., t_p\}$, $1 = t_0 < t_1 < t_2 < ... < t_p \leq n$ be a temporal segmentation of $S$ into $p$ segments. For each such segmentation, we define the following energy

function

$$E(T_p) = \sum_{i=1}^{m} \sum_{k=1}^{p-1} d(S_i(t_{k-1} : t_k - 1), S_i(t_k : t_{k+1} - 1)). \quad (1)$$

In Eq.(1), $d(.,.)$ denotes a function that computes the distance between the signal segments $S_i(t_{k-1} : t_k - 1)$ and $S_i(t_k : t_{k+1} - 1)$. In this work, the Pearson's distance [13] is used to implement $d(.,.)$. The Pearson's distance is defined based on Pearson's linear correlation coefficient $C(x, y)$ between the signals $x, y$, i.e.,

$$d(x, y) = 1 - C(x, y), \quad (2)$$

which is minimized when the signals' autocorrelation is maximized. Since the number of samples of each segment are not necessary equal, in order to estimate the autocorrelation between $S_i(t_{k-1} : t_k - 1)$ and $S_i(t_k : t_{k+1} - 1)$, $d(x, y)$ is estimated through a uniform resampling of the signal $S_i(t_k : t_{k+1} - 1)$, so that the resulting signal consists of $t_k - t_{k-1}$ samples.

According to our problem definition, it holds that the optimal temporal segmentation should minimize the energy $E(T_p)$ of Eq.(1). Thus, the temporal segmentation of the human motion amounts to estimating the segmentation $T_p$ with this property.

An extra complication arises by the fact that in the application domain we consider, the duration of each period slightly changes over the time. This means that for all $i \in \{1, ..., p\}$ there exists a small positive value $\alpha$ (e.g. $\alpha < 0.1$) such that

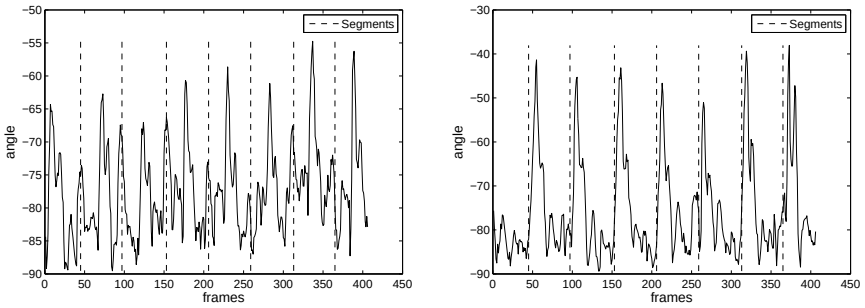
$$\frac{(i - \alpha) \cdot n}{p} \leq t_i \leq \frac{(i + \alpha) \cdot n}{p}. \quad (3)$$

The quantity $\frac{n}{p}$ corresponds to an upper bound estimation of the mean period of the motion signal. Then, the proposed method for temporal segmentation consists of two steps:

- Estimation of the number of periods $p$.
- Estimation of the optimal $T_p$ under the assumption that the signal consists of $p$ periods.

The number of periods $p$ can be automatically computed by getting the global maximum of the amplitude signal of Fast Fourier Transform of a given motion signal. $p$ can be also estimated by the music beat detection (see Section 4) under the assumption that the source music corresponding to the given dance motion signal is available. Alternatively, $p$ can be given by the minimization of mean of $E(\hat{T}_s)$ over the periods, where $\hat{T}_s$ denotes the uniform time segmentation into $s$ periods. This means that the duration of each period is equal to $\frac{n}{s}$. We restrict the search space to periods between $0.25s$ and $1.25s$, which relates to the range of possible tempo in music. In notation,

$$p = argmin_{s \geq 2} \frac{E(\hat{T}_s)}{s}. \quad (4)$$




(a) (b)

Fig. 2: The proposed temporal segmentation of the left and right knee angles.

Having estimated the number of periods $p$ (see Eq. (4)), the goal of the proposed method is to find the segmentation $T_p$ that minimizes $E(T_p)$ under the constraint of Eq.(3). Let $D(u, v)$, $v > u$ be a metric that measures how periodic is the signal segment that corresponds to the time interval $[u, v]$. $D(u, v)$ takes its minimum value of zero if the segment from $u$ to $v$ is periodic. According to the definition of $E(T_p)$ (see Eq.(1)), $D(u, v)$ is given by

$$D(u,v) = \min_{-\alpha \leq \psi \leq \alpha} \sum_{i=1}^{m} d(S_i(u : v-1), S_i(v : w)), \tag{5}$$

where $w = v - 1 + \lceil (1 + \psi) \cdot (v - u) \rceil$. In order to reduce the computation cost (see also the constraint of Eq. (3)) $D(u, v)$ is computed according to Eq. 5 only when $\frac{(1-2\alpha)\cdot n}{p} \leq v - u \leq \frac{(1+2\alpha)\cdot n}{p}$. Otherwise, $D(u, v)$ is set to $\infty$.

Then, we construct a graph $G$ as follows. A node $u \in \{1, ..., n\}$ of $G$ corresponds to the time instance $u$. The weight of edge $u \sim v$ is given by $D(u, v)$. Finally, we add the virtual node $n + 1$ that is connected with the last time instances $\{\lfloor n - \alpha \cdot \frac{n}{p} \rfloor, ..., n\}$ with an almost zero weight, since we don't know the end of the signal periods. Then, the global minimum of $E(T_p)$ under the constraint of Eq.(3) is given by the sum of weights of the shortest path between the nodes 1 and $n + 1$. The nodes (time instances) of the shortest path correspond to the optimal solution for the problem of signal segmentation. Using the Dijkstra's algorithm [14], the time complexity for the shortest path computation is $O(log(N) \cdot E)$, where $N$ and $E$ are the number of nodes and edges of $G$, respectively. This cost can be reduced to $O(Nlog(N))$, since $E = O(N)$ due to the constraint of Eq.(3).

Figure 2 illustrates the proposed temporal segmentation using as input the angles of the left and right knees of a dancer dancing the traditional Cretan dance "Siganos" (see Section 5). As it can be observed, although the signals are quite complex and exhibit some small differences in period synchronization, the proposed method successfully segments both of them.

## 4 Music Beat Detection

As our main focus lies upon the synchronization of movement data to a music signal of a traditional dance, we apply a beat tracking algorithm that was tailored towards specific properties of the music signals at hand [15]. The central aspect of the method presented in [11] is the usage of group delay defined as the derivative of the phase spectrum over frequency. The method takes advantage of the fact that the average group delay, also referred to as phase slope function, can provide insight about the position of impulses that are caused by note onsets. The phase slope function is used to obtain a signal representation, which emphasizes time instances of instrument note onsets and can be used to track the beat in a music signal. The parameters of the phase slope computation are the same as those presented in [11]. Onset candidates are determined separately in four frequency bands, which results in four band-wise onset signals $\mathbf{y}_c[n], c = 1...4$.

For the estimation of beat times from the band-wise onset signals, an algorithm based on the method proposed by Klapuri et al. [9] has been used. The algorithm first determines a tempo trajectory for a piece of music, and then aligns a sequence of impulses having a period related to that tempo to the music signal. The tempo trajectory is obtained by computing a weighted sum of $\mathbf{y}_c[n]$

$$\mathbf{y}[n] = \sum_{c=1}^{4}(6-c)\mathbf{y}_c[n] \tag{6}$$

and then, by weighting $\mathbf{y}[n]$ with the spectral flux at each sample $n$:

$$\mathbf{y}_{flux}[n] = \mathbf{y}[n]\sum_{\omega} HWR(|X(\omega,n)| - |X(\omega,(n-1))|). \tag{7}$$

In Eq.(7), HWR denotes a half wave rectification and $X(\omega, n)$ denotes the (short time) Fourier transform of the signal as used in the group delay computation. In order to obtain a set of tempo periods, the sample autocorrelation of $\mathbf{y}_{flux}[n]$ is computed in rectangular windows of $t_{win} = 8s$ length with a hop size of $1s$. The obtained sequence of autocorrelation vectors describes the development of rhythmic content over the duration of a piece. In the following, the tempo periods $\beta$ have been estimated using a *Hidden Markov Model*(HMM) as described in [9]. This results in a sequence of beat period estimations $\beta[k]$, with $k = 1...N$, with $N$ being the number of autocorrelation vectors for a piece.

In order to align a beat pulse with the signal, we compute the likelihood of an alignment phase $\Phi[k]$ in analysis frame $k$

$$P(\hat{\mathbf{r}}_{\tilde{\mathbf{y}}\mathbf{y}}|\Phi[k] = l) = \sum_{c=1}^{4}(6-c)\sum_{n=0}^{8f_o}\tilde{\mathbf{y}}_k[n+l]\mathbf{y}_c[(k-4)f_o + n] \tag{8}$$

where $\tilde{\mathbf{y}}_k$ is a reference pulse train of $t_{win}f_o + 1$ samples length, having an impulse at the middle position and a period equal to $\beta[k]$. Thus, just like in the estimation of the beat period, an eight second length window has been used. The weighted sum of the band wise correlations as computed in (8) is then used in an HMM framework as suggested in [9].

| $MSig_1$ | $MSig_2$ | $MAl_1$ | $MSyr_1$ | $MSyr_2$ |
|---|---|---|---|---|
| 794 - 827 | 682 - 710 | 570 - 571 | 388 - 450 | 388 - 400 |

Table 1: The number of frames for the beat synchronous dance animations.

## 5 Beat Synchronous Dance Animation

Having segmented the given human motion and music into periods, the next task is to create a beat synchronous dance animation. To achieve this, we resample the motion signal so that it becomes equal to the target music tempo. More specifically, let $T_p = \{t_0, t_1, t_2, ..., t_p\}$ and $T'_p = \{t'_0, t'_1, t'_2, ..., t'_p\}$ be the temporal segmentations of the human motion and target music signals, respectively. In order to get beat synchronous dance animation, the $i$-segment $[t_i, t_{i+1}]$ of the motion signal should be resampled with $r_i = \frac{t'_{i+1} - t'_i}{t_{i+1} - t_i}$ oversampling rate. If $r_i = 1$, then there are no changes in the resulting motion signal. In order to avoid rapid changes on sampling rate on the borders of segments, a continuous oversampling rate $r(t)$ at time $t \in [t_i, t_{i+1}]$ of the motion signal can be used, which is defined as

$$r(t) = \frac{\sum_{k=-1}^{1} w(t, i+k) \cdot r_{i+k}}{\sum_{k=-1}^{1} w(t, i+k)}, \tag{9}$$

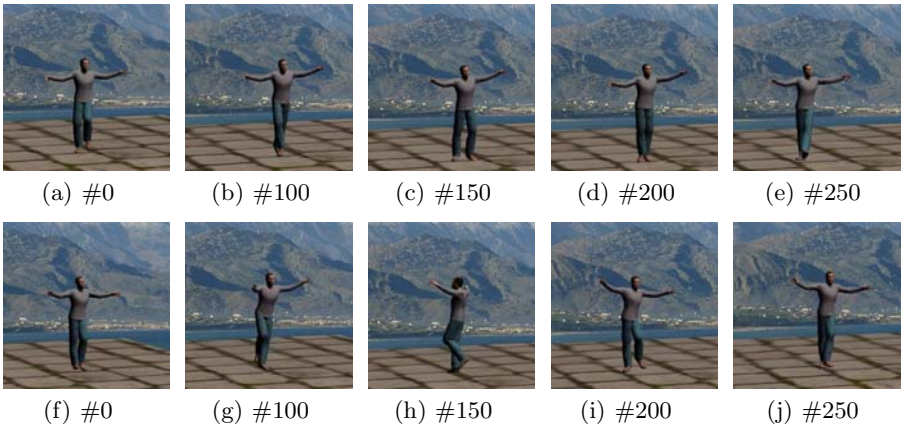
where $\delta = \frac{t_p - t_0}{p}$ is equal with the mean period of motion signal. In Eq.(9), $w(t,k) = exp(\frac{-2 \cdot (t - (t_{i+k} + t_{i+k+1})/2)^2}{\delta^2})$. In addition, the application of $r(t)$ keeps the animation beat synchronous.

## 6 Experimental Results

The proposed method has been evaluated on a data set consisting of several traditional Greek/Cretan dances. A professional dancer has performed a variety of such dances. His performance was recorded by a fully calibrated and synchronized multicamera system operating at $25Hz$. Three types of dances were investigated, namely "Siganos", "Maleviziotis" and "Syrtos" which are danced in 6, 8 and 12 steps, respectively. Like most of traditional Cretan dances, a theme can be repeated with a practically infinite number of variations. The articulated motion of the dancer was tracked with a recently proposed method [12], which estimates the 3D position, orientation and full articulation of the human body based on the markerless observations provided by the camera system.

We employed motion recordings of Siganos ($MSig_1, MSig_2$), Maleviziotis ($MMal_1$) and Syrtos ($MSyr_1, MSyr_2$). The number of temporal samples of $MSig_1$, $MSig_2$, $MMal_1$, $MSyr_1, MSyr_2$ is 750, 638, 658, 445 and 417, respectively.

Each of these recordings has been synchronized using two audio recordings of different tempos. We have used two audio recordings for each dance type: Siganos ($ASig_1, ASig_2$), Syrtos ($ASyr_1, ASyr_2$) and Maleviziotis ($AMal_1, AMal_2$), respectively. Table 1 presents the number of frames of the beat synchronous dance



(a) #0 (b) #100 (c) #150 (d) #200 (e) #250

(f) #0 (g) #100 (h) #150 (i) #200 (j) #250

Fig. 3: ((**a**) - (**e**)) Frames of synchronized dance animation for $MSig_2$ under $ASyr_1$. ((**f**) - (**j**)) Frames of synchronized dance animation for $MSyr_2$ under $ASyr_2$.

animations using the two corresponding audio recordings. So, the number of frames of $MSig_1$ with $ASig_1$ and $ASig_2$ is 794 and 827, respectively. Figures 3(a)-3(e) and 3(f)-3(j) show sample frames where the motion of $MSig_2$ was aligned to the music of $ASig_1$ and the motion of $MSyr_2$ was aligned to the music of $ASyr_2$, respectively. A more complete set of video results containing ten beat synchronous dance animation videos can be downloaded at `http://alturl.com/64ihx`. As it can be verified, the proposed framework provides beat synchronous dance animations of good quality, in all tested cases.

## 7 Conclusions

In this work, we proposed a framework that generates beat synchronous dance animation combining complex human motion capture data with an audio signal of a target music. The proposed approach has been successfully tested on variety of dances containing cyclic activities such as traditional Greek/Cretan dances. The proposed method yields the optimal solution for the problem of temporal segmentation into periods of multidimensional signal applied on complex human motion data such as dance movements. Regarding future work, we plan to apply the temporal signal segmentation to other types of periodic signals (e.g. ECG or geophysical signals) in order to segment them into periods. Moreover, the proposed method can be used as a last part of a motion synthesis scheme. Thus, by providing audio input only, the envisioned system will be able to provide new unseen motions of beat synchronous realistic animations of a virtual dancer.

## Acknowledgments

This research has been co-financed by the European Union (European Social Fund - ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALIS-UOA- ERASITECHNIS.

## References

1. Alankus, G., Bayazit, A.A., Bayazit, O.B.: Automated motion synthesis for dancing characters. Computer Animation and Virtual Worlds **16** (2005) 259–271
2. Panagiotakis, C., Tziritas, G.: Snake terrestrial locomotion synthesis in 3d virtual environments. The Visual Computer **22** (2006) 562–576
3. Kim, J.W., Fouad, H., Sibert, J.L., Hahn, J.K.: Perceptually motivated automatic dance motion generation for music. Computer Animation and Virtual Worlds **20** (2009) 375–384
4. Kim, T., Park, S., Shin, S.: Rhythmic-motion synthesis based on motion-beat analysis. In: ACM Transactions on Graphics (TOG). Volume 22. (2003) 392–401
5. Essid, S., Richard, G.: Fusion of multimodal information in music content analysis. In: Multimodal Music Processing. (2012)
6. Chu, W., Tsai, S.: Rhythm of motion extraction and rhythm-based cross-media alignment for dance videos. IEEE Transactions on Multimedia (2012) 1–1
7. Shi, J., Tomasi, C.: Good features to track. In: CVPR. (1994) 593–600
8. Valtazanos, A., Arvind, D., Ramamoorthy, S.: Comparative study of segmentation of periodic motion data for mobile gait analysis. In: Wireless Health 2010, ACM (2010) 145–154
9. Klapuri, A.P., Eronen, A.J., Astola, J.T.: Analysis of the meter of acoustic musical signals. IEEE Trans. on Audio, Speech, and Language Processing **14** (2006) 342–355
10. Böck, S., Schedl, M.: Enhanced Beat Tracking with Context-Aware Neural Networks. In: Proceedings of the 14th International Conference on Digital Audio Effects (DAFx-11), Paris, France (2011)
11. Holzapfel, A., Stylianou, Y.: Beat tracking using group delay based onset detection. In: Proc. of ISMIR - International Conference on Music Information Retrieval. (2008) 653–658
12. Michel, D., Panagiotakis, C., Argyros, A.A.: Tracking human body articulations with multiple rgbd sensors. Technical report, FORTH-ICS (2013)
13. Fulekar, M.: Bioinformatics: applications in life and environmental sciences. Springer Verlag (2009)
14. Dijkstra, E.: A note on two problems in connexion with graphs. Numerische mathematik **1** (1959) 269–271
15. Holzapfel, A.: Similarity methods for computational ethnomusicology. PhD thesis, University of Crete, Greece (2010)