Automatic transcription of Greek folk dance videos to labanotation based on autoencoders

Georgios Loupas[1], Theodora Pistola[1], Sotiris Diplaris[1], Christos Stentoumis[2], Konstantinos Ioannidis[1], Stefanos Vrochidis[1], and Ioannis Kompatsiaris[1]

[1] Information Technologies Institute - CERTH, Thessaloniki, Greece
{loupgeor, tpistola, diplaris, kioannid, stefanos, ikom}@iti.gr
[2] up2metric P.C., Athens, Greece
christos@up2metric.com

Abstract. We're creating an automatic system, based on autoencoders, to transcribe dance videos into labanotation [1], a movement notation system. Manual labanotation generation is a time-consuming process that requires specialized knowledge. Our system aims to save time and provide a valuable tool for choreographers, dancers, and anyone interested in documenting body movement. Our system analyzes RGB videos of dancers, isolates their movements, and generates labanotation as an image. The process involves extracting the 3D skeleton, segmenting movements, identifying them, and mapping them to Laban symbols. In our research, we focus on segmenting the movements of the dancer's lower body. We calculate the angles of the legs and use them as features to train an autoencoder. This approach, inspired by [2], has not been previously explored for human movement segmentation. Human movement segmentation remains a hard problem due to the temporal complexity among the high-dimensional motion features. Our work aims to automatically generate labanotation for Greek folk dances, contributing to the preservation and transmission of dance-related Intangible Cultural Heritage (ICH). The system is integrated into the CHROMATA online platform [3], which offers AI tools for analyzing, classifying, and annotating ICH content. This integration assists designers in creating immersive experiences based on ICH.

Keywords: Autoencoders · Labanotation · Folk Dances

1 Introduction

The segmentation of human motion is a complex problem. Segmenting and understanding human steps has numerous applications, such as choreography analysis and automatic notation in Labanotation. In this study, we tackle the problem of segmentation using an unsupervised approach. Unlike supervised methods, unsupervised methods do not require a large amount of annotated data. The lack of annotated data for training algorithms in generating Labanotation highlights the importance of approaching the problem in an unsupervised manner.

2 Related Work

Signal segmentation techniques find applications in various domains and problems involving the analysis of sequential data. Some of the areas where these techniques are used include speech processing, time series data, computer vision, bioinformatics, sensor data analysis, and medical signal processing.

2.1 Human Movement Segmentation

Motion segmentation has the ability to partition lengthy sequences of high-dimensional human motion data into fragments that are semantically independent in nature. This technique finds application in various domains, including gesture recognition, primitive modeling, action recognition, robotics, and human movement recognition [4]. Manual segmentation of motion capture data is a time-consuming and demanding task, as noted by previous research [5]. Traditional methods of human motion segmentation have predominantly concentrated on features associated with periodicity or angular velocity [6]. However, these features only capture elementary aspects of motion sequences and are limited in their ability to effectively segment complex motion patterns [7]. To address this limitation and extract higher-level feature information, researchers have proposed automated segmentation algorithms based on Kernel Principal Component Analysis (KPCA) and Probabilistic Principal Component Analysis (PPCA) [8] or Short-Term Principal Component Analysis (ST-PCA) [9] for analyzing high-dimensional time series data in motion capture. Machine learning techniques, including kernel k-means and spectral clustering, have also been applied to motion segmentation [10], [11]. An alternative approach introduced by Li et al. involves Temporal Subspace Clustering (TSC) [12]. Lin et al. have proposed an optimal control strategy based on reverse optimization criteria and residual estimation for determining segmentation points [13]. To improve segmentation effectiveness, a low-level time segmentation algorithm utilizing cosine distance has been developed [14]. Furthermore, researchers have explored Hierarchical Aligned Cluster Analysis (HACA) as a temporal clustering method for motion segmentation [15]. Lastly, rhythm-baseds segmentations has been proposed from a dance rhythm/beat perspective [16], [17].

2.2 Automatic Labanotation Generation

Regarding the automated extraction of Laban notation, some techniques focus on spatial analysis of movements [18], [19], [20],[17] for automated extraction of Laban notation. They involve motion segmentation to divide the motion into elemental movements and spatial analysis to map them to Laban symbols. However, these rule-based techniques have limitations in capturing the complexities of human body movements. In other methods, movements and their Laban symbols are recognized by comparing them to standard basic movements in a motion library, using Euclidean distance and Dynamic Time Warping (DTW) [21].

More recent techniques employ Hidden Markov Models (HMMs) for labeling the lower extremities and Extra-Trees for the upper extremities [22]. Others utilize the dynamics of machine learning to train algorithms such as Neural Networks and Extreme Learning Machines (ELMs) [23], Recursive Neural Networks [24], [25], as well as seq2seq models with and without attention mechanisms [26], [27]. These approaches yield improved results compared to simple spatial analysis methods mentioned earlier but necessitate an adequate amount of training data.

3 Methodology

Our approach uses autoencoders to compare similarity in the latent space between successive frames, determining segmentation points in a 3D skeleton pose sequence. We extract motion information from knee-crotch angles, resulting in a time series of 2 variables.

The time series is divided into smaller chunks using a sliding window and fed into the autoencoder to obtain latent representations. This sliding window is continuously shifted with a constant stride until reaching the end of the signal.

Thus, essentially each time point is represented as a vector in the latent space, and due to the sliding window, temporal information is embedded. The autoencoder, in this context, acts as a feature extractor.

Given the latent representations, our algorithm then utilizes the concept that frames within a segment are expected to display a greater similarity compared to frames across different segments. The similarity between two feature vectors, $\mathbf{z_i}$ and $\mathbf{z_j}$, is is given by:

$$G(i,j) = \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\|\|\mathbf{z}_j\|} \tag{1}$$

So the self-similarity matrix is computed, and starting from a frame, the cosine similarity is taken into account with each subsequent frame until it reaches a minimum value. Since the similarity measure can be noisy and may exhibit several minimal points, an adaptive threshold is applied, so that a point is considered as a segment boundary only when it has a minimum and its value is lower than the threshold.

In our method, two autoencoder architectures were employed. A 1D convolutional autoencoder (CAE) was used as well as a full connected self-expressive autoencoder (SEA). The 1D CAE functions as a feature extractor that encapsulates the temporal information.

The SEA receives the representations generated by the 1D CAE as input and is trained to simultaneously minimize the mean squared error (MSE) and the self-expressive loss.

$$Loss = \|\mathbf{X} - \underline{\hat{\mathbf{X}}}\| + \|\mathbf{X} - \hat{\mathbf{X}}\| \tag{2}$$

where $\hat{\mathbf{X}}$ is the direct reconstruction and $\underline{\hat{\mathbf{X}}}$ is the self-expressed reconstruction. This objective aims to bring representations expressing similar segments of

the signal closer in the latent space while pushing apart dissimilar representations.

4 Experiments and Results

The data used in this study obtained from the Dance DB [1] and consists of 10 motion capture files of greek dances in BVH format of varying durations. From each BVH file, the three-dimensional coordinates of the joints were extracted. Each file was sampled to have a frame rate of 30 fps. To handle occasional sudden variations in joint coordinates, a median filter was used to eliminate outliers within the mostly noise-free data.

This study focuses on segmenting leg movements by extracting informative features such as knee and crotch angles, thus each leg sequence is represented as a time series with two variables. Separate experiments were conducted for the right and left leg due to different segmentation intervals. Due to the limited availability of annotated data and the difficulty of manual annotation, a rule-based algorithm was developed for motion segmentation, that detects motion changes using the minima of kinetic energy in the relevant limb. The rule-based algorithm, through observation, effectively identifies logical segments of movements by detecting changes that minimize the kinetic energy of the legs. While it serves as a suitable ground truth for evaluating accuracy, it may have limitations with complex and smooth movements.

Two experiments were conducted: one with noise-free signals and another with synthetic noise introduced in the skeletons. Segmentation results were obtained based on autoencoder representations as well as the characteristics of the original time series. The second experiment involved training autoencoders to denoise the signal and improve segmentation in the presence of synthetic noise. A comparison was made with the rule-based algorithm applied to the noisy signal, revealing its weaknesses in over-segmentation. The synthetic noise induced local minima in kinetic energy, potentially causing over-segmentation and exposing the weakness of the rule-based algorithm. However, the displacement information of each leg during movement was preserved, such as forward steps remaining predominantly forward with minimal alterations.

In the experiments, different window sizes, specifically 8, 16, and 32 frames, were considered, with a stride value set to one. The adaptive threshold for the i-th row is determined as the mean of the similarity values in the i-th row, specifically from the i-th column to the (i+96)-th column. This implies that, for a given frame, the adaptive threshold is computed as the average of the similarity values across subsequent frames up to a duration of approximately 3 seconds, or 96 frames, in the similarity matrix.

The evaluation metrics used were Recall, Precision, F1-score, and R-value [28], with a tolerance of 16 frames. The R-value is a metric that takes into

[1] http://dancedb.cs.ucy.ac.cy, the Dance Motion Capture Database of the University of Cyprus.

account the over-segmentation that can lead to increased recall values, providing a more robust quantification of segmentation performance.

The 1D CAE consists of three 1d convolutional layers with max pooling and batch normalization and then a fully connected layer to produce the final encoding. The first layer has 32 filters, followed by a layer with 16 filters, and a final layer with 8 filters. Each layer uses a kernel size of 3 and a stride of 1.

Two layers were used as encoders of the SEA: the first layer has double the dimensions of the input, and the second layer has the same dimensions as the input representations. This configuration allows for a nonlinear transformation in the initial space of representations generated by the 1D CAE.

Training was performed using the Adam optimizer with a learning rate of 0.001. The models were trained until convergence, with early stopping applied after 20 epochs. A leave-one-out methodology was employed to train the autoencoders so we can evaluate the generalization capabilities of the learned representations. For our experiments, we used a NVIDIA GeForce RTX 3090 GPU.

The following results pertain to the average of the segmentation outcomes for both legs.

Method	**Window Size**	**Recall**	**Precision**	**F1**	**R-Value**
Initial features	-	**0.67**	0.62	**0.65**	0.69
CAE	8	0.63	0.65	0.64	0.69
SEA	8	0.64	**0.66**	**0.65**	**0.70**
CAE	16	0.59	0.63	0.61	0.67
SEA	16	0.60	0.63	0.61	0.67
CAE	32	0.52	**0.66**	0.58	0.64
SEA	32	0.55	**0.66**	0.60	0.66

Table 1: Performance of the time series representation and the autoencoder latent representations in noiseless conditions.

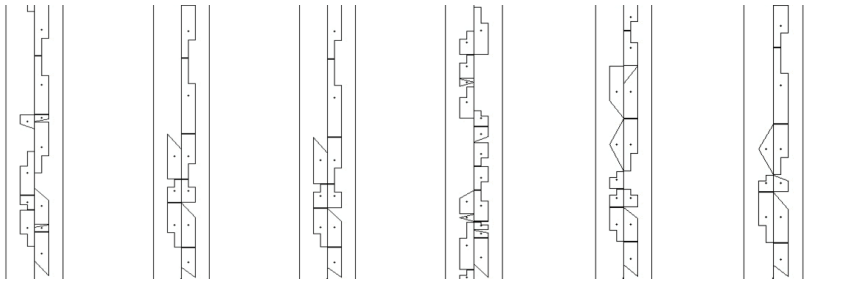
In noisless case, the algorithm applied to the initial time series has higher recall, while autoencoders achieve higher precision and similar F1 scores. R-value, indicating segmentation stability, is similar for both cases, suggesting comparable performance. Larger window sizes (e.g., 32) result in decreased performance, indicating difficulty in capturing fast-paced changes.

Method	**Window Size**	**Recall**	**Precision**	**F1**	**R-Value**
Rule-based	-	0.69	0.37	0.46	0.27
Initial features	-	0.57	0.61	0.59	0.65
CAE	8	**0.72**	0.62	**0.67**	0.69
SEA	8	0.67	**0.65**	0.66	**0.71**
CAE	16	0.66	0.61	0.63	0.67
SEA	16	0.69	0.62	0.65	0.68
CAE	32	0.57	0.61	0.59	0.66
SEA	32	0.62	0.62	0.62	0.68

Table 2: Performance of the rule-based algorithm, the time series representation and the autoencoder latent representations in noise conditions.

We observe that the rule-based algorithm performs poorly under noisy conditions, with over-segmentation and low R-value. Autoencoders achieve higher recall and outperform in all metrics. This demonstrates that the learned representations likely contain substantial information that makes them more resilient to noise. Increasing window size also leads to decreased performance.

A qualitative comparison of the generated Labanotation using the mentioned methods in relation to the ground truth is shown below. It can be observed that the networks manage to be more resilient to noise.



(a) GT (b) CAE (c) SAE (d) RB(N) (e) CAE(N) (f) SAE(N)

Fig. 1: Comparison of Labanotation Generation outputs, where GT refers to the Ground Truth, RB to the Rule-based algorith and (N) refers to noisy experiments

5 Conclusions and Future Work

Autoencoders seems to be effective for unsupervised motion segmentation, even in noisy conditions. The dynamics of the latent representations were observed, indicating potential for improvement with more complex architectures and additional cost functions. Proper motion segmentation is crucial for analyzing motion and transform it to labanotation where annotated data for training supervised techniques is not readily available. The use of autoencoders for denoising and segmenting motion could enhance also the results of rule-based algorithms in generating labanotation, yielding reliable outcomes regardless of the available data. In the future, we aim to explore more complex architectures. One idea is to employ transformer models, training them in a self-supervised manner using available non-labeled mocap files. Subsequently, fine-tuning could be performed on a small set of annotated data for labanotation generation, leading to improved performance.

Acknowledgments

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH-CREATE-INNOVATE (project code: T2EDK-01856).

References

1. Hutchinson, A., Guest, A. H., Hutchinson, W. A. (1977). Labanotation: or, kinetography Laban: the system of analyzing and recording movement (No. 27). Taylor Francis
2. Bhati, Saurabhchand Villalba, Jesús Żelasko, Piotr Dehak, Najim. (2020). Self-Expressing Autoencoders for Unsupervised Spoken Term Discovery. 4876-4880. 10.21437/Interspeech.2020-3000.
3. Pistola, Theodora Diplaris, Sotiris Stentoumis, Christos Stathopoulos, Evangelos Loupas, Georgios Mandilaras, Theodore Kalantzis, Grigoris Kalisperakis, Ilias Tellios, Anastasios Zavraka, Despoina Koulali, Panagiota Kriezi, Vera Vraka, Valia Venieri, Foteini Bacalis, Stratos Vrochidis, Stefanos Kompatsiaris, Ioannis. (2021). Creating immersive experiences based on intangible cultural heritage. 17-24. 10.1109/ICIR51845.2021.00012.
4. J. F. -S. Lin, M. Karg and D. Kulić, "Movement Primitive Segmentation for Human Motion Modeling: A Framework for Analysis," in IEEE Transactions on Human-Machine Systems, vol. 46, no. 3, pp. 325-339, June 2016, doi: 10.1109/THMS.2015.2493536.
5. G. Xia, H. Sun, L. Feng, G. Zhang and Y. Liu, "Human Motion Segmentation via Robust Kernel Sparse Subspace Clustering," in IEEE Transactions on Image Processing, vol. 27, no. 1, pp. 135-150, Jan. 2018, doi: 10.1109/TIP.2017.2738562.
6. Okada, N. Iwamoto, Naoya Fukusato, Tsukasa Morishima, Shigeo. (2015). Dance motion segmentation method based on choreographic primitives. GRAPP 2015 - 10th International Conference on Computer Graphics Theory and Applications; VISIGRAPP, Proceedings. 332-339. 10.5220/0005304303320339.
7. Y. Wang, X. Lin, L. Wu, W. Zhang, Q. Zhang and X. Huang, "Robust Subspace Clustering for Multi-View Data by Exploiting Correlation Consensus," in IEEE Transactions on Image Processing, vol. 24, no. 11, pp. 3939-3949, Nov. 2015, doi: 10.1109/TIP.2015.2457339.
8. Si-Xi Chen, Shu Chen, Jian-Wei Li, and Xin Chen. 2016. A hybrid P/KPCA-based approach for motion capture data automatic segmentation. J. Comp. Methods in Sci. and Eng. 16, 2 (2016), 197–206. https://doi.org/10.3233/JCM-160610
9. xu, Jianfeng Takagi, Koichi Yoneyama, Akio. (2010). Three-Dimensional Image Information Media: Beat Induction from Motion Capture Data Using Short-Term Principal Component Analysis. The Journal of The Institute of Image Information and Television Engineers. 64. 577-583. 10.3169/itej.64.577.
10. F. Zhou, F. De la Torre and J. K. Hodgins, "Aligned Cluster Analysis for temporal segmentation of human motion," 2008 8th IEEE International Conference on Automatic Face Gesture Recognition, Amsterdam, Netherlands, 2008, pp. 1-7, doi: 10.1109/AFGR.2008.4813468.
11. W. Wang, J. Shen, F. Porikli and R. Yang, "Semi-Supervised Video Object Segmentation with Super-Trajectories," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 41, no. 4, pp. 985-998, 1 April 2019, doi: 10.1109/TPAMI.2018.2819173.
12. S. Li, K. Li and Y. Fu, "Temporal Subspace Clustering for Human Motion Segmentation," 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 2015, pp. 4453-4461, doi: 10.1109/ICCV.2015.506.
13. J. F. -S. Lin, V. Bonnet, A. M. Panchea, N. Ramdani, G. Venture and D. Kulić, "Human motion segmentation using cost weights recovered from inverse optimal control," 2016 IEEE-RAS 16th International Conference on Humanoid

Robots (Humanoids), Cancun, Mexico, 2016, pp. 1107-1113, doi: 10.1109/HUMANOIDS.2016.7803409.
14. Y. Yang, J. Chen, Y. Zhan, X. Wang, J. Wang and Z. Liu, "Low Level Segmentation of Motion Capture Data Based on Cosine Distance," 2015 3rd International Conference on Computer, Information and Application, Yeosu, Korea (South), 2015, pp. 26-28, doi: 10.1109/CIA.2015.14.
15. F. Zhou, F. De la Torre and J. K. Hodgins, "Hierarchical Aligned Cluster Analysis for Temporal Clustering of Human Motion," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 3, pp. 582-596, March 2013, doi: 10.1109/TPAMI.2012.137.
16. W. -T. Chu and S. -Y. Tsai, "Rhythm of Motion Extraction and Rhythm-Based Cross-Media Alignment for Dance Videos," in IEEE Transactions on Multimedia, vol. 14, no. 1, pp. 129-141, Feb. 2012, doi: 10.1109/TMM.2011.2172401.
17. C. Cui, J. Li, D. Du, H. Wang, P. Tu and T. Cao, "The Method of Dance Movement Segmentation and Labanotation Generation Based on Rhythm," in IEEE Access, vol. 9, pp. 31213-31224, 2021, doi: 10.1109/ACCESS.2021.3060103.
18. Guo, H., Miao, Z., Zhu, F., Zhang, G., Li, S. (2014, November). Automatic labanotation generation based on human motion capture data. In Chinese Conference on Pattern Recognition (pp. 426-435). Springer, Berlin, Heidelberg.
19. Wang, J., Miao, Z., Guo, H., Zhou, Z., Wu, H. (2017). Using automatic generation of Labanotation to protect folk dance. Journal of Electronic Imaging, 26(1), 01102
20. Ikeuchi, K., Ma, Z., Yan, Z., Kudoh, S., Nakamura, M. (2018). Describing upper-body motions based on labanotation for learning-from-observation robots. International Journal of Computer Vision, 126(12), 1415-1429.
21. Zhou, Z., Miao, Z., Wang, J. (2016, November). A system for automatic generation of labanotation from motion capture data. In 2016 IEEE 13th International Conference on Signal Processing (ICSP) (pp. 1031-1034). IEEE.
22. Li, M., Miao, Z., Ma, C. (2019). Dance movement learning for labanotation generation based on motion-captured data. IEEE Access, 7, 161561-161572.
23. Zhang, X., Miao, Z., Zhang, Q. (2018, August). Automatic generation of Labanotation based on extreme learning machine with skeleton topology feature. In 2018 14th IEEE International Conference on Signal Processing (ICSP) (pp. 510-515). IEEE.
24. Zhang, X., Miao, Z., Yang, X., Zhang, Q. (2019, May). An efficient method for automatic generation of labanotation based on bi-directional lstm. In Journal of Physics: Conference Series (Vol. 1229, No. 1, p. 012031). IOP Publishing.
25. Hao, S., Miao, Z., Wang, J., Xu, W., Zhang, Q. (2019, September). Labanotation generation based on bidirectional gated recurrent units with joint and line features. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 4265-4269). IEEE.
26. Li, M., Miao, Z., Ma, C. (2020, May). Sequence-to-sequence labanotation generation based on motion capture data. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4517-4521). IEEE.
27. Li, M., Miao, Z., Xu, W. (2021). A CRNN-based attention-seq2seq model with fusion feature for automatic Labanotation generation. Neurocomputing, 454, 430-440.
28. Räsänen, Okko Laine, Unto Altosaar, Toomas. (2009). An improved speech segmentation quality measure: The R-value. Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH. 1851-1854. 10.21437/Interspeech.2009-538.