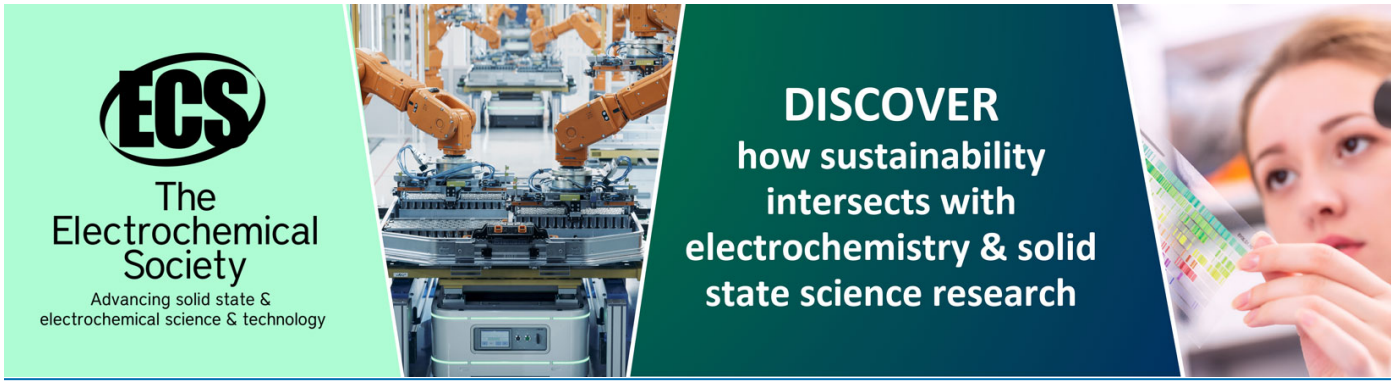
An Efficient Method for Automatic Generation of Labanotation Based on Bi-Directional LSTM



IOP Conf. Series: Journal of Physics: Conf. Series **1229** (2019) 012031 doi:10.1088/1742-6596/1229/1/012031

An Efficient Method for Automatic Generation of Labanotation Based on Bi-Directional LSTM

Xueyan Zhang [1,*], Zhenjiang Miao [1], Xiaonan Yang[2] and Qiang Zhang[1]

[1] Beijing Jiaotong University, School of Computer and Information Technology, Institute of Information Science, Haidian District, Beijing, China
[2] Center for Ethnic and Folk Literature and Art Derelopment.Ministry of Culture and Tourism.P.R.C, Beijing, China

[*]E-mail: 16120317@bjtu.edu.cn

Abstract. Labanotation uses a variety of graphic symbols to analyse and record human movements accurately and flexibly, which is an important means to protect traditional dance**.** In this paper, we introduce an efficient method for automatic generation of Labanotation from motion capture data by identifying human movements with bidirectional LSTM network (Bi-LSTM). Up to our knowledge, this is the first time that Bi-LSTM network has been introduced to the field of Labanotation generation. Compared with previous methods, Bi-LSTM used in our human movements recognition system learns context information for sequential data from not only the past but also the future directions. Combined with a newly designed discriminative skeleton-topologic feature, our approach has the ability to generate more accurate Labanotation than previous work. Experiment results on two public motion capture datasets show that our method outperforms state-of-the-art methods, demonstrating its effectiveness.

1. Introduction

Recently, the archiving and preserving of national assets such as folk dance and other performance art has become an important research topic [1]. Just like musicians record their compositions in the music score, choreographers can express their intensions in the dance notation. As a lively and logical system, Labanotation is one of the most widely used dance notation systems. It plays an important role in recording human movements, choreography, dance exercises and so on [2].

However, hand-writing Labanotation by observing human movements is a laborious and inefficient task. An effective solution is using computer technology to generate Labanotation automatically. In this paper, we propose a method based on Bi-LSTM network [3] for automatically generating Labanotation from motion capture data. The main components of our work are discriminative feature extraction and effective movement recognition. For the feature extraction, we use the skeleton-topologic feature [4]. For the recognition of human movements, we propose to apply the powerful recurrent neural network, which is expert in modelling context information for long periods and has obtained superior performance in speech processing and action recognition [5, 6]. Our method applies Bi-LSTM network to recognize the categories of lower limb movements, each of which corresponds to a specific type symbol in the Labanotation. Experiment results on two public motion capture datasets indicates that the proposed method achieves a much higher accuracy than previous work.

2. Related work

As a cross subject based on mathematics, mechanics and human anatomy, the study of automatic generation of Labanotation by computer technology is still in its infancy. Recently, some work of combining computer technology with Labanotation has been done.

First of all, some software assists people in drawing Labanotation is researched and developed, such as Laban Writer [7], Labanatory [8], Calanban [9] and LED&LINTEL [10]. Using these software to draw a Labanotation score, professionals only need to drag specific symbols to a specific area. Laban Writer [7] is currently the most widely used editor due to its practicality and convenience. However, it only supports installation and use on the Macintosh operating system.

Then, it is the research of automatic Labanotation generation from motion capture data. [11] conducts research in this area firstly. Their method based on spatial analysis, having the function of automatic generation for upper limb movements. However, the complex and important lower limb movements cannot be recognized by their methods. [12] proposes a computer-aided tool named GenLaban. It firstly performs the key-frames selection for human movements, then analyse human postures by the fixed rule. For the same movement, the amplitudes are various from different people. Therefore, this method of fixing parameters is not flexible enough.

Zhou [13, 14] proposes a key-fame matching method based on Dynamic Time Warping (DTW). Their main idea is comparing the sample with templates that pre-stored in the database and output the best matching category. Due to each person's height and weight are different, the same category of movements do not necessarily match the template. [15] presents a method based on Hidden Markov Models (HMMs), this approach train a Hidden Markov Model for each category of human movement. It achieves better results than other methods due to Hidden Markov Model choose the optimal decision of dynamic systems. However, as the categories of movements increase, it takes a long time to train Hidden Markov models.

3. Labanotation

Labanotation is a notation system used for analysing and preserving human movements, which is created by Rudolph Laban -- "the father of modern dance theory" in the early 20th century [2]. Nowadays, it has been not only used for the field of dance, but also in the mental medicine because of its versatility and no language barrier. A Labanotation score with 4 pages is shown in Figure 1. In general, Labanotation consists of two parts: vertical stuff and notation symbols.

Labanotation generally contains 9 or 11 vertical lines as its spectral structure. In practical applications, the number of columns depends on the actual situation. Each column represents a part of human body. Figure 2 shows the structure with 11 columns. For Labanotation symbols, 9 horizontal directions and 3 vertical levels make up 27 basic ones. Each symbol represents a kind of human motion. Figure 3 shows the basic symbols of Labanotation. More details on Labanotation please refer to [2].

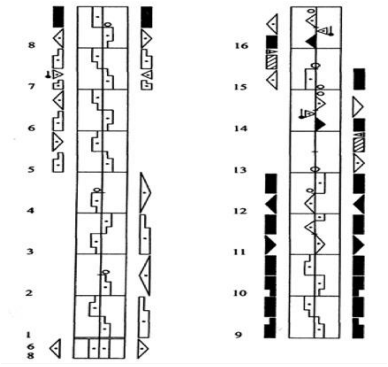


Figure 1. An example of Labanotation with 4 pages.

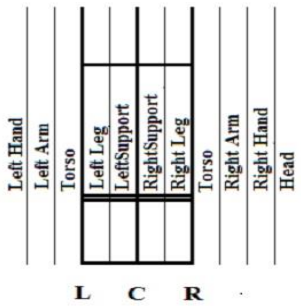


Figure 2. Structure of Labanotation.

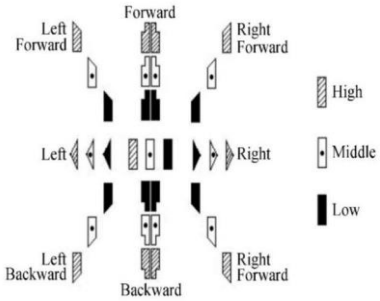


Figure 3. Basic symbols of Labanotation.

4. Proposed method

Each type of symbol in Labanotation corresponds to a category of human movements, so the core content of our method is to recognize human movements accurately.

LSTM and Bi-LSTM are briefly reviewed in the following. Then we describe the processing of motion data and the recognition of lower limb movement segments.

4.1. Overview of LSTM and bidirectional LSTM

4.1.1 LSTM Unit. Compared with the basic feed forward neural networks, the biggest difference of the recurrent neural networks (RNNs) is the self-connected recurrent connections which are suitable for sequential dynamics.

Although in theory recurrent neural networks can capture long-term dependencies, they actually do not pass due to gradient vanishing or gradient explosion [16].

LSTM [17] is a kind of variant of RNN, which aims at solving the problem of gradient vanishing/explosion and learning the long-term context information. Through the design of several subtle gates, LSTM network controls the forgetting of information or passes it to the next step.

Each LSTM unit has a memory cell c^t, which is indexed by time t. Three sigmoidal gates: input gate i^t, forget gate f^t and output gate o^t control the reading or modification of c^t. Generally, the update formulas for an LSTM unit at time t are summarized as:

$$i^t = \sigma(W_{xi}x^t + W_{hi}h^{t-1} + W_{ci}c^{t-1} + b_i) \tag{1}$$

$$f^t = \sigma\left(W_{xf}x^t + W_{hf}h^{t-1} + W_{cf}c^{t-1} + b_f\right) \tag{2}$$

$$o^t = \sigma(W_{xo}x^t + W_{ho}h^{t-1} + W_{co}c^{t-1} + b_o) \tag{3}$$

$$c^t = f^t c^{t-1} + i^t \tan h(W_{xc}x^t + W_{hc}h^{t-1} + b_c) \tag{4}$$

$$h^t = o^t \tan h(c^t) \tag{5}$$

Where $\sigma(\cdot)$ represents the sigmoid function, x^t is the input vector and h^trepresents the output vector which contains all useful information at time t and before. All W matrices represent the weights of the connection between two units. b_i, b_f, b_o, b_c denotes the bias vectors.

4.1.2 Bidirectional LSTM

LSTM neural network only captures the previous information and has no knowledge of future, but for many sequence problems, understanding past and future information is very beneficial. Therefore, Bi-LSTM is a good solution and its effectiveness has also been proved [18].

Bi-LSTM performs information capturing on both forward and backward layers for each input. The two hidden layers learn the information for both directions and the final output is obtained by combining the results of the forward layer and the backward layer. Mathematical expressions as follows:

$$h_{f_t} = \sigma\left(W_{xh_f}x_t + W_{h_fh_f}h_{f_{t-1}} + b_{h_f}\right) \tag{6}$$

$$h_{b_t} = \sigma\left(W_{xh_b}x_t + W_{h_bh_b}h_{b_{t-1}} + b_{h_b}\right) \tag{7}$$

Where h_f and h_b represent the output of forward layer and the backward layer respectively. The final output in our work is the concatenation of these two parts.

4.2. Motion capture data processing and conversion

Recall that the file format for storing human motion data is BVH in our work, which defines a tree shaped human skeletal hierarchy to store motion data and has a total of 26 human joint nodes. It records the rotation data of each node relative to its parent node in the form of Euler angles, which is

the orientation data of each human joint during the motion process. For a more intuitive analysis, Euler angles are generally converted to 3D position data in Cartesian coordinate system.

Conversion rules are as follows: For any non-root node J and its precursor node P, assume that the initial offset of J is $(\mathrm{x}_0, \mathrm{y}_0, \mathrm{z}_0)$, the world coordinate position of node P is $(\mathrm{x}_p, \mathrm{y}_p, \mathrm{z}_p)$. Set node P to perform Euler angle rotation around the ZXY-axis. Human movement is accompanied by angular displacement. Set the angular displacement matrix R, then the new position of J relative P can calculated:

$$(\mathrm{x}_1, \mathrm{y}_1, \mathrm{z}_1) = R(\mathrm{x}_0, \mathrm{y}_0, \mathrm{z}_0) \tag{8}$$

Due to the tree shaped human skeletal hierarchy structure, the calculation of the current node position is actually a recursive process, it is the result of the interaction of all its precursor nodes and not just its immediate precursor node. Assume that the rotation matrix of all predecessor nodes of node J are $R_1, R_2, \cdots R_m$, then the position offset of node J should be:

$$(x_1, y_1, z_1) = \left(R_m \cdot \left(R_{m-1}\left(\cdots\left(R_1 \cdot (x_0, y_0, z_0)\right)\right)\right)\right) \tag{9}$$

From this, the world coordinates of each human node can be obtained. Assume that the joint chain $J_0, J_1, \cdots J_r$, where J_r represents root node. The position offsets of these nodes relative to its immediate predecessor nodes are $O_0, O_1, \cdots O_{r-1}$ (root joint without precursor joint). Let P_{root} denote the root node's position, then the position of node J in the world coordinate system is:

$$P = P_{root} + O_{r-1} + \cdots + O_1 + O_0 \tag{10}$$

Through the above transformation, we can get each frame data of human body motion in world coordinates. However, the 3D spatial position data is just the position of each individual joint and lacks of relationship representation between the joints. To get a better representation of human movements, we carry out the feature extraction from the raw 3D world coordinates data. In this paper, we use the skeleton-topologic feature [4].

4.3. Recognition of Human Movements

The structure of Bi-LSTM network determines its powerful ability to capture contextual information and it can learn from both the past and future.

Our network structure mainly includes a Bi-LSTM layer and a fully connected layer. The skeleton-topologic feature [4] is used as input to the Bi-LSTM network. Each movement segment lasts 20 frames, each frame is represented by 4 vectors and each vector is characterized by 3D space coordinates. Hence, each movement segment is indicated by a 240 dimensional vector and the input size for the Bi-LSTM layer is $240 \cdot N$, where N represents the total number of samples in dataset. The layer of Bi-LSTM is mainly used to select information from the past and the future effectively, including the retention of useful information and the discarding of useless information. A dropout layer is added after the Bi-LSTM layer to prevent the over fitting. Then, the fully connected layer, using the softmax function as its activation function, is added to act as a classifier. In the training process of whole network, we set the batch-size to 32 and our goal is to minimize the value of the cross-entropy function. Moreover, we use the Adam [19] optimizer, which has faster convergence speed and better learning effect than other adaptive learning algorithms.

5. Experiment results

5.1. Datasets

In our work, we validate the efficiency and reliability of our method on two public datasets.

Dataset A: It is built by [15]. This dataset contains 16 categories of human lower limb movements and 6,400 motion capture data segments in total. It is noted that the 6,400 human movement segments

are uniformly distributed in each movement category. The duration of each movement segment is 20 frames, finally the entire data set has a total of 12,800 frames.

Dataset B: It is first introduced in our work [4]. This dataset contains 48 categories of human lower limb movements and 19,200 motion capture data segments in total. In order to be consistent with data set A, we also take 20 frames per movement segment and finally the data set has a total of 384,000 frames. Moreover, the human movement segments are also equally distributed in each category.

5.2. Evaluation

Each motion capture data segment in the dataset includes one single dance movement which corresponding to a specific Labanotation symbol. In the experiment, we randomly select half of each category sample in the data set as the training set and the rest as the test set.

Firstly, we compare our method with the key-fame matching method based on Dynamic Time Warping (DTW) proposed by [14] and the Hidden Markov Model (HMM) method proposed by [15] on dataset A. It should be pointed out that the features used in all the experiments are the same, which are the skeleton-based features [4].

Table 1 shows the performance comparison of our approach with the DTW-based key-fame matching method proposed by [14] and the HMM-based method proposed by [15]. It can be seen clearly that our method based on Bi-LSTM network achieves the best accuracy. On dataset A, our approach is about 4% better than the HMM-based method and 15% better than the key-fame matching method, which demonstrates the advantages of Bi-LSTM in processing time-series dynamic human motion data.

Table 1. Performance comparison of template method, HMM method and Bi-LSTM method

Accuracy (%) / Dataset / Methods	Dataset A (16 categories, 6400 segments)		Dataset B (48 categories, 19200 segments)	
	Left leg	Right leg	Left leg	Right leg
Template + DTW [14]	82.47	78.39	62.56	59.38
HMM [15]	91.52	92.26	90.31	89.25
Bi-LSTM (proposed)	**95.57**	**95. 68**	**95.91**	**96. 09**

To further verify the efficiency and robustness of the proposed approach, we carry out experiments on dataset B, which contains more categories of human movements and more samples compared with dataset A. It should be noted that the experiments performed on dataset A and dataset B are on the same Bi-LSTM network structure. It can be seen that even if the category and quantity of human movements are increased, our method still achieves high recognition accuracy, which indicates that our proposed method has strong generalization ability.

6. Conclusion

In this paper, we propose a method based on bidirectional LSTM neural network for automatically generating Labanotation from motion capture data segments. As an advanced recurrent neural network, Bi-LSTM has the advantage of capturing long-term contextual information from both the past and the future, which is suitable for learning the sequential dynamic characteristics of human movements from motion capture data. Compared with the DTW-based key-fame matching method and HMM-based method, experiment results verify that our method have the superior performance in recognition accuracy and Labanotation generating. Therefore, the proposed method based on Bi-LSTM can generate Labanotation more accurately than previous method, which is more in line with practical application.

Acknowledgments

This work is supported by the NSFC 61672089,61273274,61572064, National Key Technology R&D Program of China 2012BAH01F03. We thank the support on data acquisition from Centre for Ethnic

and Folk Literature and Art Development of Ministry of Culture, Ministry of Culture and Tourism, P.R.C.

References

[1] Bie J H, Liang B E. A Literature Review on the Protection and Utilization of China's Intangible Cultural Heritage [J]. Tourism Forum, 2008, 17(7):603-609
[2] A. H. Guest. Labanotation: the system of analyzing and recording movement. Routledge, 2014
[3] Schuster M, Paliwal K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11):2673-2681
[4] X. Zhang, Z. Miao, and Q. Zhang. Automatic Generation of Labanotation Based On Extreme Learning Machine with Skeleton Topology Feature. In Signal Processing (ICSP), 2018 IEEE 14th International Conference on pages 510–515
[5] Donahue J, Hendricks L A, Guadarrama S, et al. Long-term recurrent convolutional networks for visual recognition and description[M]// AB initto calculation of the structures and properties of molecules /. Elsevier, 2017:85-91
[6] Du, Y.,Wang,W.,Wang, L.: Hierarchical recurrent neural network for skeleton based action recognition. In: Proc. IEEE Int' l Conf. Computer Vision and Pattern Recognition. (2015) p 1110–1118
[7] Laban Writer. http://dance.osu.edu/research/dnb/laban-writer
[8] Labanatory. http://www.labanatory.com
[9] Calaban. http://www.bham.ac.uk/calaban/contents.htm
[10] Edward F,HuntS, Politis G, et al. LED &LINTEL: A Windows Mini-Editor and Interpreter for Labanotation. http://donhe.topcities.com/pubs/led.heml
[11] Hachimura K, Nakamura M. Method of generating coded description of human body motion from motion-captured data[C]. //Robot and Human Interactive Communication, Proceedings. 10th IEEE International Workshop on. IEEE, 122 - 127
[12] W. Choensawat, M. Nakamura, and K. Hachimura. Genlaban: A tool for generating labanotation from motion capture data. Multimedia Tools and Applications, 74(23):10823–10846, 2015
[13] Z. Zhou. Research on automatic generation of labanotation based on dynamic programming. Master's thesis, Beijing Jiaotong University, 2017
[14] Z. Zhou, Z. Miao, and J.Wang. A system for automatic generation of labanotation from motion capture data. In Signal Processing (ICSP), 2016 IEEE 13th International Conference on pages 1031–1034
[15] M. Li, Z. Miao. Automatic Labanotation Generation from Motion-captured Data Based on Hidden Markov Models. The 4th Asia Conference on Pattern Recognition, 2017
[16] John F. Kolen, Stefan C. Kremer. Gradient Flow in Recurrent Nets: The Difficulty of Learning Long-term Dependencies [J]. 2001, 28(2):237-243
[17] Hochreiter S, Schmidhuber, Jürgen. Long Short-Term Memory [J]. Neural Computation, 1997, 9(8):1735-1780.
[18] Chris Dyer, Miguel Ballesteros,Wang Ling, Austin Matthews, and Noah A. Smith.2015. Transition-based dependency parsing with stack long short-term memory. In Proceedings of ACL-2015 (Volume 1: Long Papers), pages 334–343, Beijing, China, July
[19] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization [J]. Computer Science, 2014