Review of notations and techniques for digitising dance movement

Jordy Chetty
chtjor001@myuct.ac.za
University of Cape Town
Cape Town, Western Cape

ABSTRACT

In this literature we survey notations for formalising dance movements and dance digitisation systems. We also examine the effectiveness of such notations for conveying information and segments of the pipelines used in the digitisation process. We identify that dance notation schemes are complicated for dancers to record and reproduce choreography, and that existing systems are infeasible for consumer usage. To resolve this, we consider alternative notation schemes and procedures for the notation-to-animation and data-to-notation pipelines. The literature to follow is structured as follows. In section 1, we provide background theory on the topics in the paper. In section 2, we detail existing notation schemes and discuss their ability to accurate represent movements. In section 3, we examine methods to encode human movement with respect to dance. In section 4, we provide an overview of systems that have been created to generate symbolic representations of dance movements and to animate choreography using symbolic representations. Finally in section 5 we provide a conclusion to the paper and analyse the challenges and observations from the research in this paper.

CCS CONCEPTS

• **Computing methodologies** → **Shape inference**; **Animation**; *Shape representations*; Machine learning algorithms.

KEYWORDS

Artificial Intelligence, Computer Graphics, Animation, Formal Languages

1 INTRODUCTION

Traditionally learning to dance involves two parties, one that possesses competent knowledge about how movements are performed, other movements that can be performed in succession, and how they integrate within a rhythm count. Informally, this may be considered as the definition of the dance. The second party, the student, learns through observation or practice, or a combination thereof. Once adequate knowledge has been transferred to the student, dance movements can be easily identified, but conveying this information symbolically remains a challenging problem. The issue of encoding a dance style is challenged by variability in execution that can occur for a given motion, and the data that is required in order to accurately denote how to perform an action without losing information. The best efforts to achieve this store the position and orientation of vital limb information as time progresses along with a set of rules detailing where the movement can be performed. The set of data that is needed to convey this information is infeasible to write manually, and confusing for a learner to understand. As a result, no standardised notation exists to accurately encompass the full set of dance movements for all dance styles. Cillekens [3] states that this may possibly be attributed to the complexity of recording time-series three-dimensional data. Cumulative error due to the unreliability of human recollection over multiple generations may cause the contemporary structure of a dance style to vary from its original form. Entire movements may also be lost within a single generation. This is influential as dance styles have cultural significance. As movements are forgotten, heritage is lost. As a result, systems have been created to expedite the notation process and visualise notations through animations, but are greatly constrained making them impractical for general use.

In this paper we aim to understand the issues with current notation schemes and systems that transcribe dance into notation from data and create animations from notation. We examine the relationship between notations and systems that use these applications. We consider alternative approach to the data acquisition stage using Computer Vision techniques as an alternative to hardware-based motion capture to reduce the cost of using dance digitisation systems and to improve portability.

2 BACKGROUND

Segmentation of motion capture is the process of isolating a sequence of motion captured data, usually for the purpose of capturing a recognisable movement, such as a crossover in Salsa. In practice, motion capture systems record continuous motion and thereafter the data is segmented manually or via an automated process.

BioVision Hierarchy (BVH) is a format used to encode data recorded using motion capture systems. The header of the format describes skeletal information. Data in the file represents the recording and is specified in Euler angles.

Keypoints represent skeletal joints and are expressed in terms of coordinates. They can be used to drive the animation of models by using interpolation between two different keypoints. A set of two keypoints can represent a body limb implicitly, and a collection of limbs can represent a skeletal structure.

A time signature represents the meter of a piece of music. The meter is the 'recurring pattern of stresses and accents that provide the pulse or beat of music'. Two numbers are recorded at the beginning of a piece of music for this purpose, and represent the how many beats are in each measure. Salsa has a 4/4 measure, meaning that each measure must have beats totalling to four quarter notes.

3 FORMAL LANGUAGES IN DANCE

Substantial research has been done on symbolic representation of dance, although innovation in this area has proved difficult. As

mentioned in section 1, this arises from the challenges of detailing and conveying four-dimensional data [3]. [1] presents further issues with available formats, and stated that motion capture is currently the only truly accurate method to store human movement due to the inherent limitations of other representations. Other impediments involve synchronising movements with musical beats, and that many systems have also been designed to describe the motion of a single person but not multiple people which limits their usefulness to dance. Factoring these concerns into a representation only increase the complexity of it.

3.1 Space of Salsa Dance

In the case of Cuban Salsa, an attempt has been made by Renesse and Ecke [12] create a 'Space of Salsa Dance'. Objects within the space represent possible dance contexts. With a change to the original description, each object is defined by a quintuple as follows,

$$\langle P_{leader}, P_{follower}, A^{armpos}_{leader}, A^{armpos}_{follower}, C^{arm}_{arm} \rangle$$

Where P_{leader} and $P_{follower}$ describe the orientations of the leader and the follower by either 1 or 0. 1 is taken to mean that a person is facing the partner, and 0 to mean that they are facing away. A^{armpos}_{leader} and $A^{armpos}_{follower}$ denote the arm position of the leader and the follower. The superscript *armpos* represents the position of dancer's arm relative to their body and can take the value of either B (for 'back' or 'belly') or H (for 'head'). Finally C^{arm}_{arm} is used to denote the number of times that the hands of the dancers cross, from the leader's position. *arm* takes the value of L, for 'left hand'.R is similar. For multiple crosses, the values are duplicated. For example, C^{LL}_{RR} represents two crosses with the leaders left hand on top. If no crosses are made, 0 is written. It makes use of symmetries to expand the set of possible actions from the existing set of actions. These symmetries refer to the orientations of the leader and follower for a particular move. From this basic notation, variations of a dance move described by the language are added by introducing symmetry to the notation. Using this property and the model, the set of symbolic representations is extended. The fundamental purpose of the space is to show possible transitions between dance moves based on the salsa rule set. A point in the space is analogue to a dance move. Successive connections between points are transitions between the points.

3.2 The Salsa Method

This method was proposed by Boschetti and Lyons [10] as a way to learn Salsa turn patterns. The basic elements of the system are hand holds, directions, actions and positions, which they define as the 'Salsa Dictionary'. Hand holds are different configurations of the hand positions of the dancer. Directions denote the facing positions of the dancers relative to one another and the line of dance. Each dancer is represented by a horizontal or vertical arrow. Actions are dance moves performed by the dancer. The result of an action is called a position. Each element within the dictionary may have none or several variations which are represented using identifying symbols to provide further information about how to perform a move. Using these basic elements, they define a 'Salsa Language', the notation system for a choreography. A 5-by-3 matrix is used to record a single movement between a dancing pair. Each entry

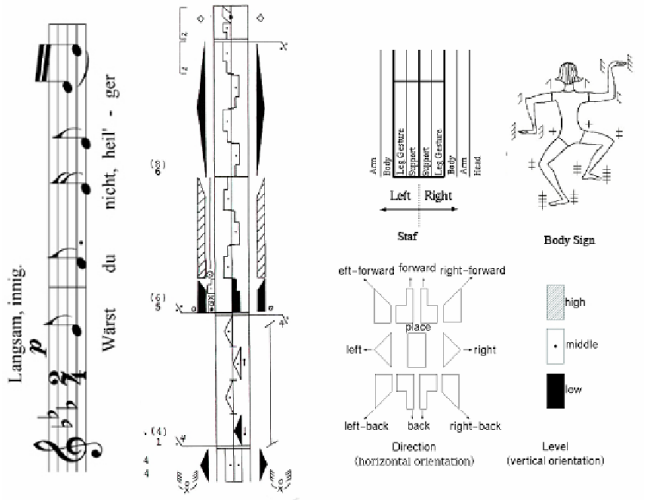

Figure 1: The Labanotation score (left) and basic Labanotation symbols (right). An XML representation of Labanotation, LabanXML, and its implementation on the notation editor LabanEditor2. [Nakamura, 2014]

provides information about the actions, movement and orientations of each dancer. A sense of timing of the actions of each person is given by the second and third column in the matrix which each represent a 4/4 measure of some arbitrary piece of music.

3.3 Labanotation

Labanotation a system to record, denote and depict human movement and motions. It consists of approximately 12000 distinct symbols [5]. It abstracts the body and the concepts of space, time and dynamics symbolically which can then be written in a score. The description of the four elements in a score provides a description of how a performance is conducted [6]. The notation is able to describe movement in arbitrary precision. Each movement is drawn using a vertical staff which is read left to right and top to bottom. The channels in the staff represent sides of the body and so any movement for a particular side is written in the associated channel. Gestures are recorded in all channels aside from the central one (called the support column), which is used to show which part of the body bears the weight of a movement as well as the transfer of weight from one movement to another. The staff is segmented horizontally to represent measures in a piece of music to encode rhythmical information. Direction symbols denote the direction of motion and are colour-coded to indicate vertical information. Variations to movements can be recorded using different accent symbols which represent different levels of emphasis. Additionally, the notation can be represented digitally using Labanotation Data (LND) format. LabanXML was also created to represent the notation in XML format.

3.4 Benesh Movement Notation

From [7], Benesh Movement Notation uses a five-line stave to represent different parts of the body. The notation is read from left to right, top to bottom. Limb position information is represented by

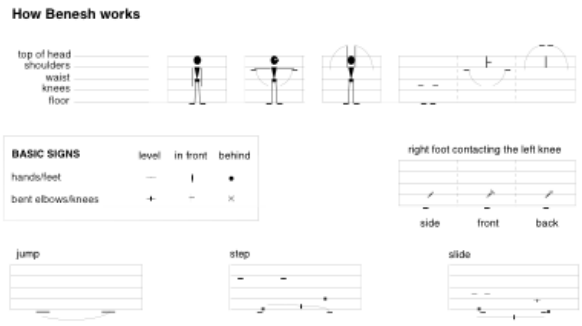

Figure 2: The basic elements of Benesh Movement Notation. [Benesh in Action, 2015]

abstract symbols and is written within the stave. Timing information is recorded above the stave and movement and orientation is recorded below the stave. The notation can also record information about the position of other dancers relative to the position of oneself below the stave. Vertical bars that divide segments of the stave show the progression of time. The similarity of the notation to musical notation allows it to be synchronized with any piece of music. In this case, the position of the bars in the score correspond to the position of the measures in the piece of music. Symbols can then be written within each measure to easily show which movements take place within a time-period.

3.5 Discussion

As with Labanotation and Benesh Movement Notation, both the dancer and the choreographer need to be competent with the syntax of the notations in order to reproduce and convey the ideas. This is inadequate as a choreographer may not be able to represent a choreography accurately, and similarly a dancer may not be able to reproduce the dance that was intended from the notation without a firm understanding of the notation. Due to the constraints of paper-based notation and human error, a high degree of information loss is inherent in the representations [17]. As a result, a considerable amount of ambiguity exists in the notation and thus some level of interpretation is required by the dancer. One can perhaps closely resemble a movement in some instances but there is no way to accurately determine whether some movement has been performed as intended exactly.

Under the 'Space of Salsa Dance', it is possible for multiple points to represent the same dance move. The notation itself does not convey information about how to perform a movement but rather serves to inform the user what can and cannot be done. Thus it assumes prior knowledge on Salsa in order to be used. Although accurate for the steps that it describes, the space is incomplete as it only describes moves for which the leader and follower arms are connected. Intricate dance movements are not possible in this scheme. This constrains the usefulness of the notation massively as the hands of the dancers in Salsa are disconnected numerous times throughout a choreography. Most harming is the fact that it doesn't provide information about the lower limbs [3]. On the other hand, the abstraction of movement allows a more compact representation of choreography in comparison to other existing notations. The way that the space is defined may allow other mathematical properties to be introduced which is a powerful feature, but currently this is not possible as the system does not obey fundamental mathematical properties to allow this.

The Salsa Method requires memorisation of symbols and their meanings along with visual figures that describe the symbols to be able to be used effectively. Information about a performance can be lost or prone to ambiguity as the same symbol is used to represent different actions in some instances. It is better suited towards advanced dancers and choreographers as a means to convey their ideas. Like the 'Space of Salsa Dance', it conveys little information about how a movement is performed. We also take note of the verbosity of the language in defining a piece of choreography. Using a 4/4 measure (as is most common in Salsa music) results in 21 measures for every minute of a song. This would mean that one would have to transcribe 11 different matrices to denote the choreography for each minute.

4 DANCE ANIMATION AND NOTATION SYSTEMS

Labanotation is most widely used in existing systems because of its popularity and ability to transcribe a wide range of movements. The bulk of research in dance synthesis and conversion applications have used this notation for symbolic representation. Systems in this domain can be broadly classified into three categories according to their core functionality; notation-transcribing systems, notation-to-animation systems, and data-to-notation systems. We concern ourselves with notation-to-animation systems and data-to-notation systems in this paper.

4.1 Notation-to-animation systems

LabanWriter is a notation-transcribing application which allows users to graphically create scores by way of a graphical interface. The files generated by this application were used by *LabanDancer* [15] to convert scores into graphical representations. The symbols in *LabanWriter* files were transformed and placed into a data structure which was then used as a reference to drive the animation of a three-dimensional model. The movement of the model was synthesised using IKAN, an inverse kinematics algorithm. The use of inverse kinematics as opposed to animation data presents possible inaccuracies in the performance of the animation as it is not data-driven. Furthermore, the limitations of *LabanWriter* directly affect the results of [15].

Laban Editor [8] has functionality for both creating scores and generating animations. Similarly to *LabanWriter*, scores can be written using an graphical interface. The application stores the movement data in LND format. Animation is performed by converting LND using motion conversion template files. These files generate the data required to drive the movement of a model. In contrast to [15], the system offers several conversion methods which enable the model to perform the animation differently by using different templates. [17] stated this as problematic as the quality of the animations relied on the quality of the template files available.

Both [15] and [8] only allow a single model to be animated at a time.

Life Forms, a proprietary software platform, addressed the issue of single character animation by allowing multiple figures to be animation using motion capture data. The system can import and export several popular motion capture file formats. It is packaged with additional motion captured data that can be used to extend animation sequences by using its blending functionality. Users can also edit models graphically, and create new poses from existing poses by editing the skeletal structures imported into the program. It can be integrated with popular applications such as *Maya*. It is more extensible and flexible than the applications previously described, but the support for Labanotation is limited.

4.2 Data-to-notation systems

[13] present a method to automate the conversion of motion captured data into Labanotation. It can only use BVH data as input. Their approach used three phases, motion data analysis to convert motion captured data into a set of three-dimensional coordinates, data analysis using spatial clustering and velocity thresholds to perform segmentation on the data, an auxiliary step to align the segments with the rhythmical information in Labanotation. The final stage converts the segments of motion into Labanotation scores in LND format. Spatial clustering from the BVH file was used to gain information about the center of gravity of the body. The velocity thresholds were used to identify the beginning and ending of movement sequences for the limbs of the body. It used the assumption that idle moments and shifts in motion represent the beginning and ending of distinct motion sequences respectively, which can be separated. The use of both together enabled the keypointing process to be automated, after which scores could be recovered. The problem with this method is that the precision of the conversion relies heavily on the accuracy of the data analysis phase, which in itself is reliant on the quality of the data fed into the system. Furthermore, the transformation of raw BVH in the first stage causes information loss due to the change in representation. Using the same data analysis method, the *GenLaban* system [2] addressed the concerns surrounding velocity threshold and spatial clustering by allowing the generated keypoints to be modified manually using a graphical interface to correct inaccuracies. This compensated for the tempo of different dance styles and the variation in performances of a choreography which may have been incorrectly keypointed initially, but did not address the fundamental issue at hand. Manual entries into the system still provide a means of erroneous keypointing. A quantisation phase was also added, comprising of motion direction analysis, bending analysis, weight support analysis and duration analysis. Each of the stages were used to improve the score generation and corresponded to different aspects of the score writing process which a transcriber would typically follow. Like *Life Forms*, the system can only use a subset of Labanotation to generate scores.

Recently stochastic and machine learning techniques have emerged as ways to refine the performance of such systems. [9] proposed a method to use Hidden Markov Models (HMM) for generating the Labanotation staff. Separate Markov Models were used for different movement categories. The results from their method yielded higher precision than previous methods for the same purpose. Additional HMM can added to the system to improve its accuracy, but this requires more training data and adds complexity to the system. In this regard, transcribing intricate movements becomes difficult. This was noticed by [17], who proposed the use of the Extreme Learning Machine (ELM) algorithm over HMM in the recognition process. This approach yielded both higher precision and lower computational times in comparison to [9]. Unlike [2, 13] which convert motion capture data into other representations, this method performs data analysis on the data directly. Taking in BVH data, the data analysis stage used the same method as [13] followed by feature extraction for use in the training of the model. This step addressed the information loss and noise in the data described previously by gathering skeleton-relevant data which was assumed to be more representative of movement and thus more accurate for the conversion process.

5 MOTION DATA GENERATION

The most crucial aspect of dance digitisation systems is the data acquisition stage. Hardware-based motion capture systems are most commonly used, but other systems are currently emerging that may be suitable alternatives to this while offering similar levels of precision.

5.1 Motion capture

Both industrial-grade and consumer hardware systems are available to capture movement. These systems can either be marker-based or markerless. Marker-based systems can be separated into active and passive systems. Active systems use colour-changing sensors whereas passive systems do not change colour. These types of systems offer greater precision in capturing but require technical specialists, large dedicated environments and expensive equipment and set up costs. As a result, they are most commonly used within industry where precision is required for realism, such as in film and gaming industries. In comparison, general consumer systems such as the line of Kinect sensors from Microsoft use markerless technology which use software processes for keypoint detection. They have few environment constraints, are inexpensive in comparison to marker-based devices and use software-calibration techniques as opposed to on-body sensors. The inherent limitation of markerless systems is lower accuracy because the data gathering process uses some degree of estimation to locate the positions of limbs and joints. The types of systems that use these devices are able to accept some loss in precision.

5.2 Pose estimation

A fundamental problem in Computer Vision is pose estimation, which relates to finding the transformation of a two-dimensional object. Initial research focused on extracting two-dimensional poses from image data. Nowadays the primary focus is geared towards recovering three-dimensional poses from two-dimensional input. From [18], the new challenges that are faced as a result of this are accounting for variations in appearance of figures in the input data, the angle and viewpoint of the camera which took the image or video, and data quality degradation caused by external entities

Figure 3: Results using the LiveCap system. LiveCap: Real-time Human Performance Capture from Monocular Video. [Habermann et al, 2019]

and self-occlusions of the camera. Sensitivity to occlusions is particularly of concern to our research as they are common in dance. Different approaches have been put forward to address the effects of the above while simultaneously improving the accuracy of existing techniques. The significance of this is that pose estimation techniques may be able to be used rather than motion capture hardware systems which would make the data-capturing process portable and remove the need for dedicated hardware systems. We will only consider state-of-the-art that use monocular (single-source) imagery as it is most relevant to our research.

[14] presents a method to estimate three-dimensional poses using a self-supervised learning approach. The learning task comprised two separate tasks of 2D-to-3D pose estimation and 3D-to-2D pose projection. The projection stage was used to improve the accuracy of the estimation by maintaining consistency between the 3D estimations generated by the method and the original 2D data. The method was tested using the Human3.6M dataset against other pose estimation methods across three different scenarios for testing and performed the best in two of the three scenarios. The method significantly outperformed other methods in scenario two, where it was applied under intentional information constraints. *BodyNet* [11] proposed using a neural-network. Four separately trained networks were used in combination to estimate pose from an image source. Most recently, *MonoPerfCap* [16] and *LiveCap* [4] both allow performance to be captured in real-time. The data generated from both systems is able to be rendered from any viewpoint, as an image or video. They are very similar to one another. During testing, [4] achieved greater pose estimation accuracy over [16], but performed worse for surface reconstruction on the same dataset.

6 CONCLUSIONS

Several notations exist for notating dance movements but Labanotation has become most widely accepted, and is thus used in most dance digitisation systems. Other notation schemes are either more complex, or not descriptive enough to be able to record dance movement sufficiently. The use of Labanotation in these types of systems is debatable as they only implement a subset of it due to its inherent complexity. Fundamentally the expressiveness of existing systems is dependent on the notation that is used to represent movement. On the other hand, the expressiveness of notation schemes increases as the complexity of the notation. This makes designing effective systems difficult. Furthermore, the usage of systems is constrained by overhead costs, and also limit the environments in which dance movement can be recorded. Presenting a system to balance the complexity and expressiveness at a low cost is a problem that has still not been solved. Recent breakthroughs in Computer Vision have resulted in systems that are able to provide three-dimensional keypointing and animation from two-dimensional imagery. This may be a possible alternative to hardware-based systems but has not been explored as yet which may solve some of the problems presented.

REFERENCES

[1] Tom Calvert. 2016. Approaches to the representation of human movement: notation, animation and motion capture. In *Dance Notations and Robot Motion*. Springer, 49–68.

[2] Worawat Choensawat, Minako Nakamura, and Kozaburo Hachimura. 2015. GenLaban: A tool for generating Labanotation from motion capture data. *Multimedia Tools and Applications* 74, 23 (01 Dec 2015), 10823–10846. https://doi.org/10.1007/s11042-014-2209-6

[3] William Cillekens. 2010. *Dance animation synthesis through the use of footstep driven notation*. Master's thesis. Utrecht University, Utrecht, The Netherlands.

[4] Marc Habermann, Weipeng Xu, Michael Zollhöfer, Gerard Pons-Moll, and Christian Theobalt. 2019. LiveCap: Real-Time Human Performance Capture From Monocular Video. *ACM Trans. Graph.* 38, 2, Article 14 (March 2019), 17 pages. https://doi.org/10.1145/3311970

[5] Don Herbison-Evans. 1988. Dance, Video, Notation and Computers. *Leonardo* 21, 1 (1988), 45–50. http://www.jstor.org/stable/1578415

[6] Ann Hutchinson and Ann Hutchinson Guest. 1970. *Labanotation: Or, Kinetography Laban: the System of Analyzing and Recording Movement*. Number 27. Taylor & Francis.

[7] Benesh International. 2001. The Benesh Movement Notation score. Retrieved April 29, 2019 from https://www.royalacademyofdance.org/documents/benesh-docs/BMNscore.pdf

[8] K. Kojima, K. Hachimura, and M. Nakamura. 2002. LabanEditor: Graphical editor for dance notation. In *Proceedings. 11th IEEE International Workshop on Robot and Human Interactive Communication*. 59–64. https://doi.org/10.1109/ROMAN.2002.1045598

[9] M. Li, Z. Miao, and C. Ma. 2017. Automatic Labanotation Generation from Motion-Captured Data Based on Hidden Markov Models. In *2017 4th IAPR Asian Conference on Pattern Recognition (ACPR)*. 793–798. https://doi.org/10.1109/ACPR.2017.55

[10] SalsaIsGood. 2001. A Dictionary for Salsa and Mambo moves. Retrieved April 24, 2019 from http://salsaisgood.com/dictionary/main_Salsa_Method.htm

[11] Gul Varol, Duygu Ceylan, Bryan Russell, Jimei Yang, Ersin Yumer, Ivan Laptev, and Cordelia Schmid. 2018. BodyNet: Volumetric Inference of 3D Human Body Shapes. In *The European Conference on Computer Vision (ECCV)*.

[12] Christine von Renesse and Volker Ecke. 2011. Mathematics and Salsa dancing. *Journal of Mathematics and the Arts* 5, 1 (2011), 17–28. https://doi.org/10.1080/17513472.2010.491781 arXiv:https://doi.org/10.1080/17513472.2010.491781

[13] J. Wang and Z. Miao. 2018. A method of automatically generating Labanotation from human motion capture data. In *2018 24th International Conference on Pattern Recognition (ICPR)*. 854–859. https://doi.org/10.1109/ICPR.2018.8545306

[14] Keze Wang, Liang Lin, Chenhan Jiang, Chen Qian, and Pengxu Wei. 2019. 3D Human Pose Machines with Self-supervised Learning. *CoRR* abs/1901.03798 (2019). arXiv:1901.03798 http://arxiv.org/abs/1901.03798

[15] Lars Wilke, Tom Calvert, Rhonda Ryman, and Ilene Fox. [n. d.]. From dance notation to human animation: The LabanDancer project. *Computer Animation and Virtual Worlds* 16, 3âĂŘ4 ([n. d.]), 201–211. https://doi.org/10.1002/cav.90 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cav.90

[16] Weipeng Xu, Avishek Chatterjee, Michael Zollhöfer, Helge Rhodin, Dushyant Mehta, Hans-Peter Seidel, and Christian Theobalt. 2018. MonoPerfCap: Human Performance Capture From Monocular Video. *ACM Trans. Graph.* 37, 2, Article 27 (May 2018), 15 pages. https://doi.org/10.1145/3181973

[17] Xueyan Zhang, Zhenjiang Miao, and Qiang Zhang. 2018. Automatic Generation of Labanotation Based On Extreme Learning Machine with Skeleton Topology Feature. In *2018 14th IEEE International Conference on Signal Processing (ICSP)*. IEEE, 510–515.

[18] Xiaowei Zhou, Menglong Zhu, Spyridon Leonardos, Konstantinos G. Derpanis, and Kostas Daniilidis. 2015. Sparseness Meets Deepness: 3D Human Pose Estimation from Monocular Video. *CoRR* abs/1511.09439 (2015). arXiv:1511.09439 http://arxiv.org/abs/1511.09439