Similarity-Based Processing of Motion Capture Data

Jan Sedmidubsky
Masaryk University
Brno, Czech Republic
xsedmid@fi.muni.cz

Pavel Zezula
Masaryk University
Brno, Czech Republic
zezula@fi.muni.cz

ABSTRACT

Motion capture technologies digitize human movements by tracking 3D positions of specific skeleton joints in time. Such spatio-temporal data have an enormous application potential in many fields, ranging from computer animation, through security and sports to medicine, but their computerized processing is a difficult problem. The recorded data can be imprecise, voluminous, and the same movement action can be performed by various subjects in a number of alternatives that can vary in speed, timing or a position in space. This requires employing completely different data-processing paradigms compared to the traditional domains such as attributes, text or images. The objective of this tutorial is to explain fundamental principles and technologies designed for similarity comparison, searching, subsequence matching, classification and action detection in the motion capture data. Specifically, we emphasize the importance of similarity needed to express the degree of accordance between pairs of motion sequences and also discuss the machine-learning approaches able to automatically acquire content-descriptive movement features. We explain how the concept of similarity together with the learned features can be employed for searching similar occurrences of interested actions within a long motion sequence. Assuming a user-provided categorization of example motions, we discuss techniques able to recognize types of specific movement actions and detect such kinds of actions within continuous motion sequences. Selected operations will be demonstrated by on-line web applications.

CCS CONCEPTS

• **Information systems** → *Similarity measures*; *Clustering and classification*; **Multimedia and multimodal retrieval**; • **Computing methodologies** → *Supervised learning by classification*;

KEYWORDS

motion capture data; similarity searching; subsequence matching; annotation; action detection; stream-based processing

ACM Reference Format:
Jan Sedmidubsky and Pavel Zezula. 2018. Similarity-Based Processing of Motion Capture Data. In *2018 ACM Multimedia Conference (MM '18), October 22–26, 2018, Seoul, Republic of Korea.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3240508.3241468

MM '18, October 22–26, 2018, Seoul, Republic of Korea

ACM ISBN 978-1-4503-5665-7/18/10.
https://doi.org/10.1145/3240508.3241468

1 INTRODUCTION

Motion capture data are multiple time series of 3D positions of human-skeleton joints recorded in a frame-by-frame manner. The interest in capturing these data is continuously growing and new application scenarios emerge in a variety of fields. For example, the data could be employed in military to virtually simulate a combat and conflict-resolving situations; in law-enforcement to identify suspicious subjects or events; in smart homes to detect anomalous behavior or body positions of elderly people; in sports to quantify the improvement of athlete's performance or to predict possible injuries; or in medicine to evaluate progress in rehabilitation or to discover movement disorders as indicators for choosing suitable treatments.

A great application potential together with a growing availability of capturing devices indicate a considerable increase of motion data volume in the near future. A capturing device – recording 3D positions of tens of skeleton joints simultaneously at the rate of 120 frames per second – can easily produce gigabytes of continuous data within a single day. Even though storing such quantity of data makes an issue, their intelligent management is a much more challenging problem. Content-based processing techniques, such as searching, organizing and analyzing, are crucial to fully exploit the data potential and make the expensively recorded data more accessible, valuable and reusable.

In this tutorial, we focus on search-based techniques that can efficiently localize relevant motions within a large data collection or relevant subsequences within a very long motion sequence. We also discuss classification techniques able to determine specific kind of movement actions with respect to a user-provided categorization of example motions. Such categorization can additionally be employed for semantic segmentation of long sequences, e.g., to detect user-specified motion events in real time or to provide the long-sequence annotation.

2 SIMILARITY CONCEPT

Intelligent processing of motion capture data requires to follow patterns used in real-life evolution and communication between species. There, recognition, learning and judgment presuppose an ability to categorize stimuli and classify situations by *similarity*, which is subjective and context-dependent. The common approach in processing complex, typically unstructured, digital data is to extract content-preserving structured *features* and use them for associative access. The most successful generic approach follows the *metric space* model [40], which has been already applied in numerous data processing domains [39]. However, contrary to more traditional data types such as text documents or images, the similarity in the motion-data domain has to additionally handle the dynamics of the time dimension [38].

To effectively model the spatial and temporal evolutions of different motions, robust and sufficiently discriminative features need to be extracted [30]. To become invariant towards the subject's position, orientation and skeleton size, the input data are often normalized [22]. The normalized data are then processed to extract features on the level of frames or segments. The frame-based features [6] describe single-frame characteristics, for example, normalized distances of pairwise joints [35, 41], co-occurrence of joints [42] or other relational features [20]. The segment-based features describe a multiple-frame sequence by covariance matrices [33], fisher vectors [8], or learned representations extracted using convolutional neural networks [23], auto encoders [34] or support vector machines [11]. The learned representations [23, 29] generally achieve a higher descriptive power than hand-crafted features [20, 32].

The frame-based features are represented as a multi-dimensional time series whose length corresponds to the motion length. The time series of two different motions are compared by time-warping distance measures, such as the Dynamic Time Warping in [2]. On the other hand, the segment-based features have a fixed length and can be efficiently compared, for example, high-dimensional vectors by the Euclidean distance in [23] or bit strings by the Hamming distance in [34].

3 SEARCHING, CLASSIFICATION AND SEMANTIC SEGMENTATION

We consider that motion data appear in form of either a single *long* motion, or a collection of *short* motions. While the short motions represent semantically-indivisible actions (e.g., Rittberger jump taking 0.7 seconds), the long one relates to a more complex topic (e.g., figure-skating performance taking 3 minutes) and can contain many short actions. The long sequence can be processed either as a whole, or in the stream-based nature if the whole sequence is not known in advance or does not fit into main memory.

3.1 Similarity Searching

To search a collection of short motions, a k-nearest neighbor (kNN) query can be evaluated to obtain the k motions that are the most relevant to a user-provided query motion, based on similarity of their features. Since the collection can be large, multi-dimensional or metric-based index structures [40] can be employed to speed-up similarity search.

If the collection contains a long motion, subsequence search is applied to discover the long-motion parts that are similar to the query from both the content and length points of view. One way is to search for the long-motion frames whose features are similar to the features of selected query-motion frames. The retrieved sets of similar frames are then ranked in temporal order to identify query-relevant subsequences [26]. Since searching in frame-based features needn't be so effective, the segment-based features are extracted from overlapping [25] or disjoint [28] segments, which are detected in an unsupervised way [13, 37] from both the long motion and query. To identify query-relevant segments, sequential search can be used, such as the A-LTK method in [9] or string-matching-based algorithm in [5]. To improve scalability when searching in a large number of segments, an index structure is employed, for example, the trie-based structure in [12] or PPP-Codes index in [25].

3.2 Action Recognition

Action recognition, also referred to as action classification, is the problem of inferring the kind of movement action, based on a pre-classified collection of short motions. The class of a query motion can be recognized by a kNN classifier that searches the input collection to retrieve the k most query-relevant motions whose class labels are then ranked [24, 27]. Such kNN classifiers have been gradually effaced by the increasing success of neural networks that recognize the query class directly. Specifically, deep convolutional networks are trained by the 2D-motion-image features that are also classified by the network [14]. Most attempts suggest to employ the architecture of recurrent neural networks to better model the contextual dependency in the temporal domain [17]. This architecture can be enriched by the Long Short-Term Memory (LSTM) to better learn long-term temporal dependencies [15, 18, 30, 42]. To further handle the noise and occlusion of skeleton sequences, gating mechanisms are integrated to learn the reliability of the sequential data and accordingly adjust their effect on updating the long-term context information stored in LSTM cells [17]. To benefit from different architectures at the same time, the combination of convolutional and LSTM networks is proposed [21]. The recurrent networks are also enriched by attention-based mechanisms to additionally detect the most discriminative moments within an action [1, 18, 30].

3.3 Semantic Segmentation

Most recognition approaches classify only the short motions that correspond to a single action. Only few of them [3, 7, 33, 35, 36, 41] can detect and recognize actions within a long unsegmented motion. Such semantic segmentation is more difficult as the beginnings and endings of actions are unknown and have to be determined.

Similarly as in subsequence search, the long motion can be partitioned to extract segment-based features. The segment features are then used to search for the nearest matches within the features of the predefined class actions. If the computed similarity is high, the nearest-match class is considered as the segment label [7, 20]. The disadvantage is that many overlapping segments have to be processed and, when labeled, they do not have to straightforwardly mark the precise beginnings and endings of actions. Moreover, each segment has to be known before its processing begins, implying that labels are discovered with a slight delay.

To avoid such disadvantages, a per-class probability can be estimated for each frame by exploiting learned class representations. To enhance the quality, the contextual information of so-far scanned frames is continuously encoded, for example, in recurrent frame-based features [41], hidden states of auto encoders [4], deep beliefs [35] or LSTM-based neural networks [10, 31]. This enables detecting actions even before they finish, which makes the frame-based semantic segmentation suitable for early action detection [16, 19] or future action prediction [4, 10]. On the other hand, learning-based approaches require costly training and cannot dynamically react when the specification of target actions changes.

ACKNOWLEDGMENTS

This research was supported by ERDF "CyberSecurity, CyberCrime and Critical Information Infrastructures Center of Excellence" (No. CZ.02.1.01/0.0/0.0/16_019/0000822).

REFERENCES

[1] Fabien Baradel, Christian Wolf, and Julien Mille. 2017. Human Action Recognition: Pose-based Attention draws focus to Hands. In *ICCV Workshop on Hands in Action*.

[2] Mathieu Barnachon, Saïda Bouakaz, Boubakeur Boufama, and Erwan Guillou. 2014. Ongoing human action recognition with motion capture. *Pattern Recognition* 47, 1 (2014), 238–247.

[3] Said Yacine Boulahia, Eric Anquetil, Franck Multon, and Richard Kulpa. 2018. CuDi3D: Curvilinear displacement based approach for online 3D action detection. *Computer Vision and Image Understanding* (2018). https://doi.org/10.1016/j.cviu.2018.07.003

[4] Judith Butepage, Michael J. Black, Danica Kragic, and Hedvig Kjellstrom. 2017. Deep Representation Learning for Human Motion Prediction and Classification. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 6158–6166.

[5] Zhigang Deng, Qin Gu, and Qing Li. 2009. Perceptually Consistent Example-based Human Motion Retrieval. In *Symposium on Interactive 3D Graphics and Games (I3D '09)*. ACM, 191–198.

[6] Yong Du, Wei Wang, and Liang Wang. 2015. Hierarchical Recurrent Neural Network for Skeleton Based Action Recognition. In *Int. Conference on Computer Vision and Pattern Recognition (CVPR)*. 1110–1118.

[7] Petr Elias, Jan Sedmidubsky, and Pavel Zezula. 2017. A Real-Time Annotation of Motion Data Streams. In *19th International Symposium on Multimedia*. IEEE Computer Society, 154–161.

[8] Georgios Evangelidis, Gurkirt Singh, and Radu Horaud. 2014. Skeletal Quads: Human Action Recognition Using Joint Quadruples. In *22nd Int. Conference on Pattern Recognition (ICPR)*. 4513–4518.

[9] Y. Fang, K. Sugano, K. Oku, H. H. Huang, and K. Kawagoe. 2015. Searching human actions based on a multi-dimensional time series similarity calculation method. In *14th International Conference on Computer and Information Science (ICIS)*. 235–240. https://doi.org/10.1109/ICIS.2015.7166599

[10] Ashesh Jain, Amir Roshan Zamir, Silvio Savarese, and Ashutosh Saxena. 2016. Structural-RNN: Deep Learning on Spatio-Temporal Graphs. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), 5308–5317.

[11] Harshad Kadu and C.-C. Jay Kuo. 2014. Automatic Human Mocap Data Classification. *IEEE Transactions on Multimedia* 16, 8 (2014), 2191–2202.

[12] Mubbasir Kapadia, I-kao Chiang, Tiju Thomas, Norman I Badler, and Joseph T Kider Jr. 2013. Efficient Motion Retrieval in Large Motion Databases. In *ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games (I3D)*. ACM, New York, NY, USA, 19–28.

[13] Björn Krüger, Anna Vögele, Tobias Willig, Angela Yao, Reinhard Klein, and Andreas Weber. 2017. Efficient Unsupervised Temporal Segmentation of Motion Data. *IEEE Transactions on Multimedia* 19, 4 (April 2017), 797–812.

[14] Sohaib Laraba, Mohammed Brahimi, Joelle Tilmanne, and Thierry Dutoit. 2017. 3D skeleton-based action recognition by representing motion capture sequences as 2D-RGB images. *Computer Animation and Virtual Worlds* 28, 3-4 (2017).

[15] Chaolong Li, Zhen Cui, Wenming Zheng, Chunyan Xu, and Jian Yang. 2018. Spatio-Temporal Graph Convolution for Skeleton Based Action Recognition. In *32nd Conference on Artificial Intelligence (AAAI)*. AAAI Press.

[16] Sheng Li, Kang Li, and Yun Fu. 2018. Early Recognition of 3D Human Actions. *ACM Trans. Multimedia Comput. Commun. Appl.* 14, 1s, Article 20 (March 2018), 21 pages.

[17] Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. 2016. Spatio-Temporal LSTM with Trust Gates for 3D Human Action Recognition. In *European Conference on Computer Vision (ECCV)*. Springer International Publishing, Cham, 816–833.

[18] Jun Liu, Gang Wang, Ling-Yu Duan, Ping Hu, and Alex C. Kot. 2018. Skeleton Based Human Action Recognition with Global Context-Aware Attention LSTM Networks. *IEEE Transactions on Image Processing* 27, 4 (2018), 1586–1599.

[19] Shugao Ma, Leonid Sigal, and Stan Sclaroff. 2016. Learning Activity Progression in LSTMs for Activity Detection and Early Detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 1942–1950.

[20] Meinard Müller, Andreas Baak, and Hans-Peter Seidel. 2009. Efficient and Robust Annotation of Motion Capture Data. In *ACM SIGGRAPH/Eurographics Symposium on Computer Animation (SCA 2009)*. ACM Press, 17–26.

[21] Juan C. Nunez, Raul Cabido, Juan J. Pantrigo, Antonio S. Montemayor, and Jose F. Velez. 2018. Convolutional Neural Networks and Long Short-Term Memory for skeleton-based human activity and hand gesture recognition. *Pattern Recognition* 76 (2018), 80–94.

[22] Ronald Poppe, Sophie Van Der Zee, Dirk K. J. Heylen, and Paul J. Taylor. 2014. AMAB: Automated measurement and analysis of body motion. *Behavior Research Methods* 46, 3 (2014), 625–633.

[23] Jan Sedmidubsky, Petr Elias, and Pavel Zezula. 2017. Effective and Efficient Similarity Searching in Motion Capture Data. *Multimedia Tools and Applications* (2017), 1–22.

[24] Jan Sedmidubsky, Petr Elias, and Pavel Zezula. 2017. Enhancing Effectiveness of Descriptors for Searching and Recognition in Motion Capture Data. In *19th International Symposium on Multimedia*. IEEE Computer Society, 240–243.

[25] Jan Sedmidubsky, Petr Elias, and Pavel Zezula. 2018. Searching for variable-speed motions in long sequences of motion capture data. *Information Systems* (2018). https://doi.org/10.1016/j.is.2018.04.002

[26] Jan Sedmidubsky, Jakub Valcik, and Pavel Zezula. 2013. A Key-Pose Similarity Algorithm for Motion Data Retrieval. In *Advanced Concepts for Intelligent Vision Systems (ACIVS)*. Springer, 669–681.

[27] Jan Sedmidubsky and Pavel Zezula. 2018. Probabilistic Classification of Skeleton Sequences. In *29th International Conference on Database and Expert Systems Applications (DEXA)*. Springer, 1–15.

[28] Jan Sedmidubsky, Pavel Zezula, and Jan Svec. 2017. Fast Subsequence Matching in Motion Capture Data. In *21st European Conference on Advances in Databases and Information Systems (ADBIS)*. Springer, 1–14.

[29] Roshan Singh, Jagwinder Kaur Dhillon, Alok Kumar Singh Kushwaha, and Rajeev Srivastava. 2018. Depth based enlarged temporal dimension of 3D deep convolutional network for activity recognition. *Multimedia Tools and Applications* (2018). https://doi.org/10.1007/s11042-018-6425-3

[30] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2016. An End-to-End Spatio-Temporal Attention Model for Human Action Recognition from Skeleton Data. *CoRR* abs/1611.06067 (2016). http://arxiv.org/abs/1611.06067

[31] Sijie Song, Cuiling Lan, Junliang Xing, Wenjun Zeng, and Jiaying Liu. 2018. Spatio-Temporal Attention-Based LSTM Networks for 3D Action Recognition and Detection. *IEEE Transactions on Image Processing* 27, 7 (July 2018), 3459–3471.

[32] Bin Sun, Dehui Kong, Shaofan Wang, Lichun Wang, Yuping Wang, and Baocai Yin. 2018. Effective human action recognition using global and local offsets of skeleton joints. *Multimedia Tools and Applications* (2018). https://doi.org/10.1007/s11042-018-6370-1

[33] Chang Tang, Wanqing Li, Pichao Wang, and Lizhe Wang. 2018. Online human action recognition based on incremental learning of weighted covariance descriptors. *Information Sciences* 467 (2018), 219–237. https://doi.org/10.1016/j.ins.2018.08.003

[34] Yingying Wang and Michael Neff. 2015. Deep signatures for indexing and retrieval in large motion databases. In *8th ACM SIGGRAPH Conference on Motion in Games*. ACM, 37–45.

[35] D. Wu and L. Shao. 2014. Leveraging Hierarchical Parametric Networks for Skeletal Joints Based Action Segmentation and Recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 724–731.

[36] Yan Xu, Zhengyang Shen, Xin Zhang, Yifan Gao, Shujian Deng, Yipei Wang, Yubo Fan, and EricI-Chao Chang. 2017. Learning multi-level features for sensor-based human action recognition. *Pervasive and Mobile Computing* 40 (2017), 324–338.

[37] Xiaomin Yu, Weibin Liu, and Weiwei Xing. 2017. Behavioral segmentation for human motion capture data based on graph cut method. *Journal of Visual Languages & Computing* 43 (2017), 50–59.

[38] Pavel Zezula. 2015. Similarity Searching for the Big Data. *Mob. Netw. Appl.* 20, 4 (2015), 487–496. https://doi.org/10.1007/s11036-014-0547-2

[39] Pavel Zezula. 2016. Similarity Searching for Database Applications. In *Advances in Databases and Information Systems (ADBIS)*. Springer International Publishing, Cham, 3–10.

[40] Pavel Zezula, Giuseppe Amato, Vlastislav Dohnal, and Michal Batko. 2006. *Similarity Search: The Metric Space Approach*. Advances in Database Systems, Vol. 32. Springer-Verlag. 220 pages.

[41] Xin Zhao, Xue Li, Chaoyi Pang, Quan Z. Sheng, Sen Wang, and Mao Ye. 2014. Structured Streaming Skeleton – A New Feature for Online Human Gesture Recognition. *ACM Trans. Multimedia Comput. Commun. Appl.* 11, 1s (2014), 22:1–22:18.

[42] Wentao Zhu, Cuiling Lan, Junliang Xing, Wenjun Zeng, Yanghao Li, Li Shen, and Xiaohui Xie. 2016. Co-occurrence Feature Learning for Skeleton Based Action Recognition Using Regularized Deep LSTM Networks. In *30th AAAI Conference on Artificial Intelligence (AAAI '16)*. AAAI Press, 3697–3703.