Advancing Motion Words for Human Motion Classification

Author: **Mgr. David Procházka** Supervisor: **doc. RNDr. Vlastislav Dohnal, Ph.D.**

Faculty of Informatics, Masaryk University

Motivation

The growing volume of human motion data, captured by specialized sensors or extracted from ordinary video, requires efficient and effective representation. Its everyday applications range from **protecting cars in parking lots** by classifying human motion against a database of standard and suspicious behavior to **detecting falls in nursing homes** to trigger an immediate response. However, these high-dimensional data, typically represented by 3D skeleton sequences, incur non-trivial storage and processing costs.

The authors of [1] introduced a compact short-motion representation called **Motion Word** (MW). This one- or few-dimensional identifier is used to translate an existing skeleton sequence into a sequence of MWs. Using such a sequence in the action classification task **reduces the classification speed by two orders of magnitude** [1] and **improves the classification accuracy**.

Contributions

- We propose a **new type of MW** by partitioning the skeleton into five body parts.
- We define a concept of **joint relations** to target actions that involve an interaction of a set of joints.
- We replace DTW with edit distance as a distance function for comparing MW sequences. This change **improves classification accuracy by more than five percentage points**, **enables indexability of MW sequences** by numerous metric structures, and **applies to existing MWs**.
- We introduce a **post-processing filter** that targets misclassifications between **categories that differ in the number of action repetitions**.
- We define the concept of a **specialized classifier** that works on a subset of categories with stricter classification capabilities.
- We design a **Two-Stage Classification Framework** where a specialized classifier can refine a global classification decision, allowing **partial explainability** and **dynamic category addition and removal**.

Data and Evaluation Protocol

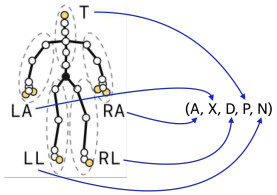
The evaluation is performed on **HDM05** [2], a dataset of 3D skeleton sequences captured at 120 frames per second, containing **2,345 actions** grouped into **130 categories**. The skeleton consists of 31 joints, tracked by a Vicon sensor. We normalize the actions, reduce their frame rate, and segment them. The motion words are then created over the **28,000 produced segments**.

We measure the effectiveness of MWs as the average accuracy of the **Weighted-Distance 4NN Classifier** [3] using a **leave-one-out** approach over the 2,345 action queries.

The framework is evaluated using **10-fold cross-validation** on a modification of HDM05, as suggested in [4]. The dataset is divided into ten folds, where nine are used for training and the rest for testing. The process is repeated once for each fold, and the overall classification accuracy is the average over all executions.

Composite Motion Word

The 3D skeleton sequence is **normalized** and **cut into fixed-size segments**, over which a **clustering** algorithm is invoked. Each body part is clustered independently. The cluster IDs are arranged in predefined positions in a 5-dimensional vector called a **Composite Motion Word** (CMW). Finally, two CMWs match if at least two body parts match.



Mapping of body parts to CMW. Adapted with modifications from [5].

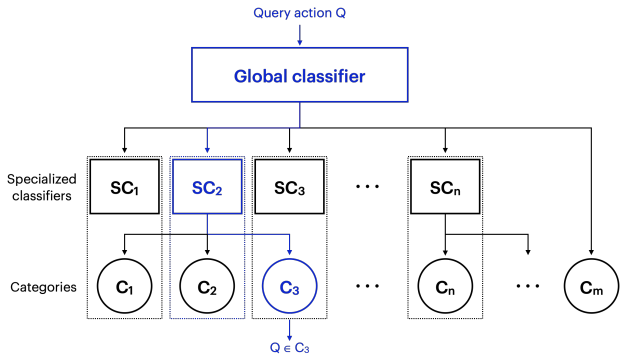
Classification Accuracy of Motion Words

We use the **edit distance** instead of the current DTW, effectively treating a MW sequence as a string. We introduce a **post-processing filter** for an environment where categories and their **actions differ in the number of repetitions**.

Motion Word	Accuracy
Segments as raw 3D skeleton data [1]	77.70%
Hard [1]	74.97%
Soft [1]	77.61%
Multi-Overlay [1]	80.30%
Composite + DTW	**75.57%**
Composite + Edit distance	**80.77%**
Composite + Edit distance + Filter	**81.75%**

Leaved-one-out using HDM05-130 [2].

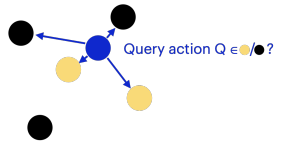
Two-Stage Classification Framework



The specialized classifier SC_2 **intercepts a misclassification** made by the global classifier and **refines the decision** by classifying the action into C_3.

Specialized Classification

Both global and specialized classifiers share the same **Weighted-Distance 4NN Classifier architecture** [3]. The specialized classifiers restrict MWs' matching by selecting **specific body part(s)** or focusing on the **neighborhood around the extremum of the action**. Through such restrictions, they **provide insight** into the classification decision.



A classifier issues a 4NN query to classify action Q, weights the distances to its neighbors [3], and selects the most likely category.

Classification Accuracy of the Framework

Method	Accuracy
Two-Stage Classification Framework	**90.02%**
Multi-Layer Perceptron [4]	95.59%
Hierarchical Bidirectional RNN [6]	96.92%

10-fold cross-validation using HDM05-65 [4].

Conclusion

We proposed a **Composite Motion Word** suited for situations where a user is interested in **motions specific to a subset of body parts**. **Edit distance** proved to be a **crucial replacement** for DTW, **increasing the accuracy** by over five percentage points and **offering a new perspective** on MW sequences. The accuracy of the **Two-Stage Classification Framework** is competitive with neural network approaches. Furthermore, the framework offers two key features: (i) **unsupervised core** – the specialized classifiers can be detached at will, and (ii) **classification insight** – realized through constrained matching.

References

[1] Jan Sedmidubsky et al. "Motion Words: A Text-Like Representation of 3D Skeleton Sequences". In: *ECIR 2020*. Lisbon, Portugal: Springer-Verlag, 2020, pp. 527–541.

[2] M. Müller et al. *Documentation Mocap Database HDM05*. Tech. rep. CG-2007-2. Universität Bonn, June 2007.

[3] Jan Sedmidubsky and Pavel Zezula. "Probabilistic Classification of Skeleton Sequences". In: *Database and Expert Systems Applications*. 2018, pp. 50–65.

[4] Kyunghyun Cho and Xi Chen. "Classifying and visualizing motion capture sequences using deep neural networks". In: *VISAPP*. Vol. 2. 2014, pp. 122–130.

[5] *Demonstration Application: Motion Data Processing*. 2013. URL: http://disa.fi.muni.cz/demo/motion-retrieval/ (visited on 02/09/2023).

[6] Yong Du, Wei Wang, and Liang Wang. "Hierarchical recurrent neural network for skeleton based action recognition". In: *CVPR*. 2015, pp. 1110–1118.