International Journal of Semantic Computing
© World Scientific Publishing Company

Efficient Query Processing in 3D Motion Capture Gesture Databases via the Gesture Matching Distance

Christian Beecks[1], Marwan Hassani[1], Bela Brenger[2], Jennifer Hinnell[3], Daniel Schüller[2], Irene Mittelberg[2], Thomas Seidl[1]

[1]*Data Management and Exploration Group*
[1]*RWTH Aachen University, Germany*
[1]*{beecks,hassani,seidl}@cs.rwth-aachen.de*

[2]*Natural Media Lab*
[2]*RWTH Aachen University, Germany*
[2]*{brenger,schueller,mittelberg}@humtec.rwth-aachen.de*

[3]*Department of Linguistics*
[3]*University of Alberta, Canada*
[3]*hinnell@ualberta.ca*

Received (30 01 2016)
Revised (24 02 2016)

One of the most fundamental challenges when accessing gestural patterns in 3D motion capture databases is the definition of spatiotemporal similarity. While distance-based similarity models such as the *Gesture Matching Distance* on *gesture signatures* are able to leverage the spatial and temporal characteristics of gestural patterns, their applicability to large 3D motion capture databases is limited due to their high computational complexity. To this end, we present a lower bound approximation of the Gesture Matching Distance that can be utilized in an optimal multi-step query processing architecture in order to support efficient query processing. We investigate the performance in terms of accuracy and efficiency based on 3D motion capture databases and show that our approach is able to achieve an increase in efficiency of more than one order of magnitude with a negligible loss in accuracy. In addition, we discuss different applications in the digital humanities in order to highlight the significance of similarity search approaches in the research field of gestural pattern analysis.

Keywords: efficient query processing; spatiotemporal data; 3D motion capture data; gestural patterns; gesture signature; gesture matching distance; dynamic time warping.

1. Introduction

3D motion capture data is a specific type of multimedia data that is mainly used to record movements of humans, animals, or objects over time. This type of data has found widespread utilization in academia and industry, for instance, for entertaining purposes, medical applications, film-making, and video game development. One of the major advantages of 3D motion capture data is the capability of expressing spatiotemporal dynamics with the highest possible accuracy [1]. This property makes 3D motion capture data particularly useful for research into the domain of gestural pattern analysis.

A gestural pattern can be understood as a kinetic action involving hand, arm, and body configurations or movements over a certain period of time. A gestural pattern is represented either by extracting its characteristic features or utilizing the raw three-dimensional movement traces, the so-called trajectories. In order to maintain the high degree of exactness provided by utilizing 3D motion capture data, we represent gestural patterns by means of *gesture signatures* [1]. Gesture signatures are multidimensional trajectory representations which facilitate gestural pattern analysis with arbitrarily high exactness. Gesture signatures are able to adapt to the individual spatial and temporal properties of gestural patterns by allowing these patterns to differ in the number of included trajectories and their lengths as well as in the weighting scheme indicating the inherent relevance of the trajectories. In fact, gesture signatures provide an adaptable model-free approach which supports lazy query-dependent evaluation, i.e., no time-intensive training phase is needed prior to query processing.

In order to leverage the spatial and temporal characteristics of gestural patterns, we utilize the *Gesture Matching Distance* [1] for the similarity comparison of two gesture signatures. The Gesture Matching Distance is a distance-based similarity measure which quantifies the degree of dissimilarity between two differently structured gesture signatures by matching similar trajectories within the gesture signatures according to their spatial and temporal characteristics. To this end, the Gesture Matching Distance is parameterized with a distance measure between individual trajectories, such as the *Dynamic Time Warping* [2, 3], the *Levenshtein Distance* [4], the *Minimal Variance Matching* [5], the *Longest Common Subsequence* [6, 7], the *Edit Distance with Real Penalty* [8], the *Edit Distance on Real Sequences* [9], or the *Mutual Nearest Point Distance* [10].

Although the Gesture Matching Distance enables a user-customizable and adaptive similarity definition, it is accompanied by a high computation time complexity. The computation time complexity for a single distance computation between two gesture signatures is quadratic in the number of the underlying trajectories. Thus the applicability of this spatiotemporal similarity measure is limited to small-to-moderate 3D motion capture databases.

In this paper, we aim to counteract this efficiency issue and present a lower bound approximation of the Gesture Matching Distance [11] that can be utilized in an optimal multi-step query processing architecture [12]. Besides the theoretical investigation of this approximation, we benchmark the performance in terms of accuracy and efficiency and empirically show that the proposed lower bound approximation is able to achieve an increase in efficiency of more than one order of magnitude with a negligible loss in accuracy. In addition, we discuss different applications in digital humanities in order to highlight the significance of similarity search approaches in the research field of gestural pattern analysis.

The paper is structured as follows: Section 2 outlines related work with a particular focus on gestural pattern similarity by means of gesture signatures and the Gesture Matching Distance. In Section 3, we investigate the lower bound approximation of the Gesture Matching Distance. The optimal multi-step query processing algorithm is presented in Section 4. Experimental results are reported in Section 5, and a discussion of different applications in digital humanities with a particular focus on gesture research is included in Section 6. The conclusions are given in Section 7.

2. Related Work

2.1. *Gesture Signatures*

A *gesture signature* [1] is a lossless spatiotemporal representation of a gestural pattern which comprises different movement traces, the so-called *trajectories.* A trajectory can be thought of as a finite sequence of points in a multidimensional space. As we consider the three-dimensional Euclidean space $\mathbb{R}^3$, we define a trajectory $t \in \mathbb{T}$ as:

$$t : \{1, \ldots, n\} \to \mathbb{R}^3, \tag{1}$$

where $t(i) = (x_i, y_i, z_i) \in \mathbb{R}^3$ represents the coordinates of the movement trace at time $i \in [1, \ldots, n]$. The *trajectory space* $\mathbb{T} = \bigcup_{k \in \mathbb{N}} \{t | t : \{1, \ldots, k\} \to \mathbb{R}^3\}$ denotes the set of all finite trajectories.

Since a gestural pattern typically involves more than one trajectory within a certain period of time, we aggregate these trajectories by means of a *gesture signature* $S \in \mathbb{R}^{\mathbb{T}}$ which is defined as:

$$S : \mathbb{T} \to \mathbb{R}^{\geq 0} \text{ subject to } |\{t \in \mathbb{T} | S(t) \neq 0\}| < \infty. \tag{2}$$

A gesture signature is a function from the trajectory space $\mathbb{T}$ into the real numbers $\mathbb{R}$. It assigns each trajectory a non-negative weight indicating its relevance with respect to the corresponding gestural pattern. Possible weighting schemes include uniform weighting, motion distance weighting, and motion variance weighting [1]. The latter reflect the overall movement and vividness of a trajectory, respectively.

2.2. *Gesture Signature Distance Functions*

Gestural patterns typically maintain a high degree of idiosyncrasy meaning that the involved trajectories are almost unique. In order to quantify a similarity value between two differently structured gestural patterns, Beecks et al. [1] have investigated the idea of matching similar trajectories within the gestural patterns according to their spatial and temporal characteristics. To this end, the trajectories are compared by means of a trajectory distance function, such as the Dynamic Time Warping Distance, and the distances between matching trajectories are accumulated accordingly. Thus, given two gesture signatures $S_1, S_2 \in \mathbb{R}^{\mathbb{T}}$ and a trajectory distance function $\delta : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$, the *Gesture Matching Distance* between S_1 and S_2 is defined as:

$$\mathrm{GMD}_{\delta}(S_1, S_2) = \sum_{(t_1,t_2)\in \mathrm{m}^{\delta\text{-NN}}_{S_1\to S_2}} S_1(t_1)\cdot\delta(t_1,t_2) + \sum_{(t_2,t_1)\in \mathrm{m}^{\delta\text{-NN}}_{S_2\to S_1}} S_2(t_2)\cdot\delta(t_2,t_1), \tag{3}$$

where the *nearest-neighbor matching* $\mathrm{m}^{\delta\text{-NN}}_{S_1\to S_2} \subseteq \mathbb{T}\times\mathbb{T}$ between S_1 and S_2 is defined as $\mathrm{m}^{\delta\text{-NN}}_{S_1\to S_2} = \{(t_1,t_2)\in\mathbb{T}\times\mathbb{T} | S_1(t_1) > 0 \wedge S_2(t_2) > 0 \wedge t_2 = \mathrm{argmin}_{t\in\mathbb{T}}\delta(t_1,t)\}$.

The Gesture Matching Distance increases with decreasing similarity of the matching trajectories. The computation time complexity is quadratic in the number of trajectories, i.e. a single distance computation lies in $\mathcal{O}(|\{S_1(t) > 0\}_{t\in\mathbb{T}}|\cdot|\{S_2(t) > 0\}_{t\in\mathbb{T}}|\cdot\zeta)$ where ζ denotes the computation time complexity of the trajectory distance function δ.

In addition to the Gesture Matching Distance, other applicable signature distance functions [13, 14, 15, 16, 17] are the transformation-based *Earth Mover's Distance* [17], the correlation-based *Signature Quadratic Form Distance* [18], the matching-based *Hausdorff Distance* [19] and its variants [20, 21] as well as the *Signature Matching Distance* [22].

2.3. *Trajectory Distance Functions*

Fundamental to the question of how to model spatiotemporal similarity between gestural patterns comprising one or more trajectories, is the question of how to determine similarity between two individual trajectories. A common approach to comparing two trajectories is based on *Dynamic Time Warping* [2, 3]. The idea of this approach is to fit the trajectories to each other by aligning their coincident similar points and accumulating the corresponding point-wise distances. Given two trajectories $t_n : \{1,\ldots,n\} \to \mathbb{R}^3$ and $t_m : \{1,\ldots,m\} \to \mathbb{R}^3$ and a point-wise distance function $\delta : \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}$, the *Dynamic Time Warping Distance* between t_n and t_m is defined as:

$$\mathrm{DTW}_{\delta}(t_n,t_m) = \delta(t_n(n),t_m(m)) + \min\begin{cases}\mathrm{DTW}_{\delta}(t_{n-1},t_{m-1})\\ \mathrm{DTW}_{\delta}(t_n,t_{m-1})\\ \mathrm{DTW}_{\delta}(t_{n-1},t_m)\end{cases} \tag{4}$$

with

$$\mathrm{DTW}_{\delta}(t_0,t_0) = 0 \tag{5}$$

$$\mathrm{DTW}_{\delta}(t_i,t_0) = \infty \quad \forall 1 \leq i \leq n \tag{6}$$

$$\mathrm{DTW}_{\delta}(t_0,t_j) = \infty \quad \forall 1 \leq j \leq m. \tag{7}$$

The Dynamic Time Warping Distance is defined recursively by minimizing the distances δ between replicated points of the trajectories. In this way, the distance δ assesses the spatial proximity of two points while the Dynamic Time Warping Distance preserves their temporal order within the trajectories. By utilizing Dynamic

Programming, the computation time complexity of the Dynamic Time Warping Distance lies in $\mathcal{O}(n \cdot m)$.

In addition to Dynamic Time Warping described above, spatiotemporal similarity between trajectories can be assessed for instance by the *Levenshtein Distance* [4], the *Minimal Variance Matching* [5], the *Longest Common Subsequence* [6, 7], the *Edit Distance with Real Penalty* [8], the *Edit Distance on Real Sequences* [9], or the *Mutual Nearest Point Distance* [10].

2.4. *Other Approaches to Gestural Pattern Similarity*

Gestural patterns are mainly investigated in terms of gesture recognition, which aims at recognizing meaningful expressions of human motion including hand, arm, face, head, and body movements [23]. Many surveys [24, 25, 26, 27, 23, 28, 29, 30, 31] have been released in the past years, providing an extensive overview of the many facets of gesture recognition. Most approaches either rely on 2D video capture technology and, thus, computer vision techniques, cf. [32, 33], or on 3D motion capture technology, which provides higher accuracy and thus more potential for precise spatiotemporal similarity search. A recent survey of vision-based gesture recognition approaches can be found in [28]. Frequently encountered approaches for recognizing manual gestural patterns are based on *Hidden Markov Models* [34, 35, 36, 37] or more generally *Dynamic Bayesian Networks* [38]. More recent approaches are based for instance on *Feature Fusion* [39], on *Dynamic Time Warping* [40, 41], on *Longest Common Subsequences* [42], or on *Neural Networks* [43].

3. Lower Bound Approximation of the Gesture Matching Distance

In this section, we present the lower bound approximation of the Gesture Matching Distance. In order to derive this approximation, we will first investigate the theoretical properties of the underlying nearest-neighbor matching and then show how these findings lead to our proposal.

Suppose we are given two gesture signatures $S_1, S_2 \in \mathbb{R}^{\mathbb{T}}$ and a trajectory distance function $\delta : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$ that determines the dissimilarity between two individual trajectories. In general, a nearest-neighbor matching $\mathrm{m}^{\delta\text{-NN}}_{S_1 \to S_2} \subseteq \mathbb{T} \times \mathbb{T}$ assigns each trajectory $t \in \mathbb{T}$ from gesture signature S_1 to one or more trajectories $u, v, \ldots \in \mathbb{T}$ from gesture signature S_2. If the trajectories $u, v, \ldots$ are equally distant to trajectory t, i.e. if it holds that $\delta(t, u) = \delta(t, v) = \ldots$, trajectory t is matched to several nearest neighbors. In practice, however, the uniqueness of the distances between different trajectories most likely leads to exactly one nearest neighbor. If this is not the case, we assume that one of the nearest neighbors is selected non-deterministically. Based on this assumption, each nearest-neighbor matching $\mathrm{m}^{\delta\text{-NN}}_{S_1 \to S_2}$ between two non-empty gesture signatures S_1 and S_2 satisfies the following properties:

- *Left totality*:

$$\forall t \in \mathbb{T}, \exists u \in \mathbb{T} : S_1(t) > 0 \Rightarrow (t, u) \in \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}} \tag{8}$$

- *Right uniqueness*:

$$\forall t, u, v \in \mathbb{T} : (t, u) \in \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}} \wedge (t, v) \in \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}} \Rightarrow u = v \tag{9}$$

Intuitively, each trajectory $t \in \mathbb{T}$ that contributes to gesture signature S_1, i.e. which has a positive weight $S_1(t) > 0$, is matched to exactly one trajectory $u \in \mathbb{T}$ with $S_2(u) > 0$ from gesture signature S_2. These properties of a nearest-neighbor matching hold true irrespective of the underlying trajectory distance function δ. Thus, by replacing the trajectory distance function δ with another one, pairs of matching trajectories are subject to change, as shown in the following lemma.

Lemma 1. *Let $S_1, S_2 \in \mathbb{R}^{\mathbb{T}}$ be two gesture signatures and $\delta, \delta' : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$ be two trajectory distance functions. For the nearest-neighbor matchings $\mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}}$ and $\mathrm{m}_{S_1 \to S_2}^{\delta'\text{-NN}}$ between S_1 and S_2 it holds that:*

$$\forall t \in \mathbb{T}, \exists u, v \in \mathbb{T} : (t, u) \in \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}} \Leftrightarrow (t, v) \in \mathrm{m}_{S_1 \to S_2}^{\delta'\text{-NN}} \tag{10}$$

Proof. Let $S_1(t) \leq 0$. By definition of the nearest-neighbor matching it holds that $(t, u) \notin \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}}$ and that $(t, v) \notin \mathrm{m}_{S_1 \to S_2}^{\delta'\text{-NN}}$.

Let $S_1(t) > 0$. Suppose that $\exists u \in \mathbb{T}$ such that $(t, u) \in \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}}$. By definition of $\mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}}$ it then holds that $S_2(u) > 0$. Thus, $|\{t \in \mathbb{T} | S_2(t) \neq 0\}| > 0$. Consequently, by replacing δ with δ' there exists at least one trajectory $v \in \mathbb{T}$ with $S_2(v) > 0$ that minimizes $\delta'(t, v)$. Therefore, $(t, v) \in \mathrm{m}_{S_1 \to S_2}^{\delta'\text{-NN}}$. Suppose that $\forall u \in \mathbb{T}$ it holds that $(t, u) \notin \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}}$. Due to the fact that $S_1(t) > 0$ it follows that $\{t \in \mathbb{T} | S_2(t) \neq 0\} = \emptyset$. Consequently, by replacing δ with δ' if follows that $\forall v \in \mathbb{T}$ it holds that $(t, v) \notin \mathrm{m}_{S_1 \to S_2}^{\delta'\text{-NN}}$. This gives us the statement. □

Lemma 1 states that each trajectory t from gesture signature S_1 that matches a trajectory u from gesture signature S_2 according to a distance function δ also matches a trajectory v according to a distance function δ'. Due to the right uniqueness of the nearest-neighbor matching, we conclude that each pair of matching trajectories $(t, u) \in \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}}$ has exactly one counter pair $(t, v) \in \mathrm{m}_{S_1 \to S_2}^{\delta'\text{-NN}}$. This fact is summarized in the following corollary.

Corollary 1. *Let $S_1, S_2 \in \mathbb{R}^{\mathbb{T}}$ be two gesture signatures and $\delta, \delta' : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$ be two trajectory distance functions. For each $(t, u) \in \mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}}$ there exists exactly one $(t, v) \in \mathrm{m}_{S_1 \to S_2}^{\delta'\text{-NN}}$.*

The corollary above implies that the cardinality of the nearest-neighbor matching between two gesture signatures S_1 and S_2 is fixed, i.e. it holds that $|\mathrm{m}_{S_1 \to S_2}^{\delta\text{-NN}}| = |\mathrm{m}_{S_1 \to S_2}^{\delta'\text{-NN}}|$ for any trajectory distance functions δ and δ'.

What remains to be shown is that the substitution of a trajectory distance function $\delta : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$ with a lower bound $\delta_{\text{LB}} : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$, which is a function that satisfies the following property for all trajectories $u, v \in \mathbb{T} : \delta_{\text{LB}}(u, v) \leq \delta(u, v)$, will lead to a lower bound approximation of the Gesture Matching Distance. To this end, we provide the following lemma.

Lemma 2. *Let $S_1, S_2 \in \mathbb{R}^{\mathbb{T}}$ be two gesture signatures and $\delta : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$ be a trajectory distance function with lower bound $\delta_{\text{LB}} : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$, i.e. it holds that $\forall u, v \in \mathbb{T} : \delta_{\text{LB}}(u, v) \leq \delta(u, v)$. The nearest-neighbor matching satisfies the following property:*

$$(t, u) \in \mathrm{m}^{\delta_{\text{LB}}\text{-NN}}_{S_1 \to S_2} \wedge (t, v) \in \mathrm{m}^{\delta\text{-NN}}_{S_1 \to S_2} \Rightarrow \delta_{\text{LB}}(t, u) \leq \delta(t, v) \tag{11}$$

Proof. Suppose it holds that $(t, u) \in \mathrm{m}^{\delta_{\text{LB}}\text{-NN}}_{S_1 \to S_2}$ and that $(t, v) \in \mathrm{m}^{\delta\text{-NN}}_{S_1 \to S_2}$. By definition of the nearest-neighbor matching it then holds that $u = \mathrm{argmin}_{t' \in \mathbb{T} \wedge S_2(t') > 0} \delta_{\text{LB}}(t, t')$ and that $v = \mathrm{argmin}_{t' \in \mathbb{T} \wedge S_2(t') > 0} \delta(t, t')$. Since it holds that $\min_{t' \in \mathbb{T} \wedge S_2(t') > 0} \delta_{\text{LB}}(t, t') \leq \min_{t' \in \mathbb{T} \wedge S_2(t') > 0} \delta(t, t')$ it follows that $\delta_{\text{LB}}(t, u) \leq \delta(t, v)$. □

Combining Corollary 1 and Lemma 2 finally leads to the proposal, as shown in the following theorem.

Theorem 1. (Lower Bound Approximation)
Let $S_1, S_2 \in \mathbb{R}^{\mathbb{T}}$ be two gesture signatures and $\delta : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$ be a trajectory distance function. For any lower bound $\delta_{\text{LB}} : \mathbb{T} \times \mathbb{T} \to \mathbb{R}^{\geq 0}$ of δ it holds that:

$$\mathrm{GMD}_{\delta_{\text{LB}}}(S_1, S_2) \leq \mathrm{GMD}_{\delta}(S_1, S_2) \tag{12}$$

Proof. The Gesture Matching Distance is defined as: $\mathrm{GMD}_{\delta}(S_1, S_2) = \sum_{(t_1, t_2) \in \mathrm{m}^{\delta\text{-NN}}_{S_1 \to S_2}} S_1(t_1) \cdot \delta(t_1, t_2) + \sum_{(t_2, t_1) \in \mathrm{m}^{\delta\text{-NN}}_{S_2 \to S_1}} S_2(t_2) \cdot \delta(t_2, t_1)$. By lower-bounding δ with δ_{LB} the number of summands stays the same since each $(t_1, t_2) \in \mathrm{m}^{\delta\text{-NN}}_{S_1 \to S_2}$ and $(t_2, t_1) \in \mathrm{m}^{\delta\text{-NN}}_{S_2 \to S_1}$ is replaced by exactly one $(t_1, t_2') \in \mathrm{m}^{\delta_{\text{LB}}\text{-NN}}_{S_1 \to S_2}$ and $(t_2, t_1') \in \mathrm{m}^{\delta_{\text{LB}}\text{-NN}}_{S_2 \to S_1}$, respectively (cf. Corollary 1). According to Lemma 2 it additionally holds that $\delta(t_1, t_2) \geq \delta_{\text{LB}}(t_1, t_2')$ and $\delta(t_2, t_1) \geq \delta_{\text{LB}}(t_2, t_1')$. We thus conclude that $\sum_{(t_1, t_2) \in \mathrm{m}^{\delta\text{-NN}}_{S_1 \to S_2}} S_1(t_1) \cdot \delta(t_1, t_2) \geq \sum_{(t_1, t_2') \in \mathrm{m}^{\delta_{\text{LB}}\text{-NN}}_{S_1 \to S_2}} S_1(t_1) \cdot \delta(t_1, t_2')$ and that $\sum_{(t_2, t_1) \in \mathrm{m}^{\delta\text{-NN}}_{S_2 \to S_1}} S_2(t_2) \cdot \delta(t_2, t_1) \geq \sum_{(t_2, t_1') \in \mathrm{m}^{\delta_{\text{LB}}\text{-NN}}_{S_2 \to S_1}} S_2(t_2) \cdot \delta(t_2, t_1')$, which gives us the statement. □

Theorem 1 shows that the lower bound approximation of the Gesture Matching Distance is attributed to the properties of its inherent trajectory distance function. How the resulting lower bound approximation is utilized in order to process queries in gestural pattern databases arising from 3D motion capture data efficiently is shown in the following section.

4. Efficient Query Processing with the Gesture Matching Distance

A fundamental approach underlying many query processing approaches is the *optimal multi-step algorithm* [12]. The idea of this algorithm consists in processing distance-based k-nearest-neighbor queries in multiple interleaved steps, where each step incrementally generates a candidate object with respect to a lower bound approximation which is subsequently refined by means of the exact distance function until the final results are obtained. The algorithm is optimal, i.e., the number of exact distance computations is minimized.

Algorithm 1 Optimal Multi-Step k-NN

```
1:  procedure $\mathrm{NN}_k(Q, \mathrm{GMD}_{\delta_{\mathrm{LB}}}, \mathrm{GMD}_{\delta}, \mathbb{D})$
2:      $\mathcal{R} \leftarrow \emptyset$
3:      $filterRanking \leftarrow ranking(Q, \mathrm{GMD}_{\delta_{\mathrm{LB}}}, \mathbb{D})$
4:      $S \leftarrow filterRanking.next()$
5:      while $\mathrm{GMD}_{\delta_{\mathrm{LB}}}(Q,S) \leq \max_{P \in \mathcal{R}} \mathrm{GMD}_{\delta}(Q,P)$ do
6:          if $|\mathcal{R}| < k$ then
7:              $\mathcal{R} \leftarrow \mathcal{R} \cup \{S\}$
8:          else if $\mathrm{GMD}_{\delta}(Q,S) \leq \max_{P \in \mathcal{R}} \mathrm{GMD}_{\delta}(Q,P)$ then
9:              $\mathcal{R} \leftarrow \mathcal{R} \cup \{S\}$
10:             $\mathcal{R} \leftarrow \mathcal{R} - \{\arg\max_{P \in \mathcal{R}} \mathrm{GMD}_{\delta}(Q,P)\}$
11:         $S \leftarrow filterRanking.next()$
12:     return $\mathcal{R}$
```

As shown in Algorithm 1, the first step consists in generating a ranking with respect to a query gesture signature $Q \in \mathbb{R}^{\mathbb{T}}$ by means of the lower bound approximation $\mathrm{GMD}_{\delta_{\mathrm{LB}}}$ (cf. line 3). Afterwards, this ranking is processed until $\mathrm{GMD}_{\delta_{\mathrm{LB}}}$ exceeds the exact distance of the k^{th}-nearest neighbor (cf. line 5), i.e. until it holds that $\mathrm{GMD}_{\delta_{\mathrm{LB}}}(Q,S) \not\leq \max_{P \in \mathcal{R}} \mathrm{GMD}_{\delta}(Q,P)$. The algorithm updates the result set $\mathcal{R}$ as long as gesture signatures $S \in \mathbb{D}$ with smaller distances $\mathrm{GMD}_{\delta}(Q,S) \leq \max_{P \in \mathcal{R}} \mathrm{GMD}_{\delta}(Q,P)$ have been found (cf. line 8).

We utilize the optimal multi-step algorithm as described above in order to efficiently query gesture signatures in 3D motion capture databases. To this end, we additionally subject the Dynamic Time Warping Distance to a bandwidth constraint, which limits the maximum permissible time difference between two aligned points of the trajectories, and lower-bound this variant, denoted as DTW_t, by LB_{Keogh} [44]. The advantage of this lower bound is its low computation time complexity, which is linear in the length of the trajectories. In fact, we approximate $\mathrm{GMD}_{\mathrm{DTW}}$ by $\mathrm{GMD}_{\mathrm{DTW}_t}$, which is then lower-bounded by means of $\mathrm{GMD}_{\mathrm{LB}_{Keogh}}$. The performance of this approach with respect to the qualities of accuracy and efficiency is empirically investigated in the following section.

5. Performance Analysis

In this section, we benchmark the accuracy and efficiency of the Gesture Matching Distance and its lower bound approximation by using the two following different spatiotemporal 3D motion capture databases.

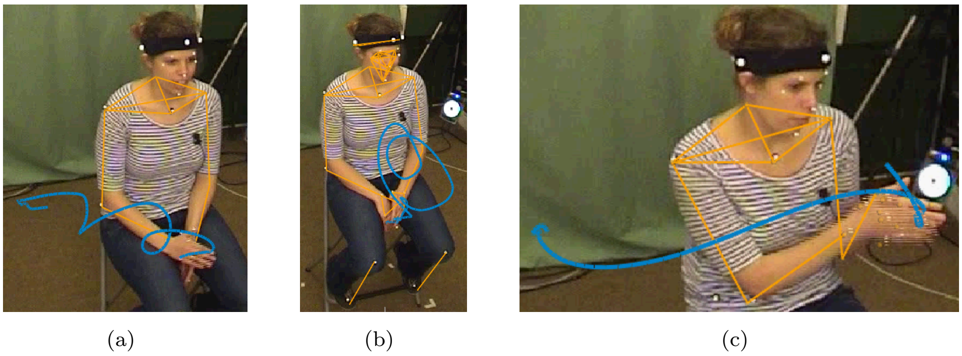

Fig. 1. Three example gestural patterns of different spatiotemporal movement types: (a) gesture of type spiral, (b) gesture of type circle, and (c) gesture of type straight. Blue trajectories indicate the main movements of the gestural patterns. Images are taken from [1].

The natural media corpus of 3D motion capture data [1] comprises three-dimensional motion capture data streams arising from eight participants during a guided conversation. The participants were equipped with a multitude of reflective markers which were attached to the body and in particular to the hands. The motion of the markers was tracked optically via cameras at a frequency of 100 Hz by making use of the Nexus Motion Capture Software from VICON. For evaluation purposes, we used the right wrist marker and two markers attached to the right thumb and right index finger each. The gestures arising within the conversation were classified by domain experts according to the following types of movement: *spiral*, *circle*, and *straight*. Example gestures of these movement types are sketched in Figure 1, which has been taken from [1]. A total of 20 gesture signatures containing five trajectories each was obtained from the 3D motion capture data streams. The trajectories of the gesture signatures have been normalized to the interval $[0, 1]^3 \subset \mathbb{R}^3$.

In addition to the 3D motion capture database described above, we utilized the *3D Iconic Gesture Dataset*[a] [45]. This dataset comprises 1,739 iconic gestures from 29 participants depicting entities, objects, and actions. Based on the provided 3D skeleton motion capture data, which was recorded via Microsoft Kinect, we randomly extracted up to 10,000 gesture signatures including between 2 and 10 trajectories in the three-dimensional Euclidean space $\mathbb{R}^3$ with a duration between 0.5 and 2.0 seconds. We additionally normalized the trajectories to the interval $[0, 1]^3 \subset \mathbb{R}^3$.

In the first series of experiments, we evaluated the accuracy of the proposed lower bound approximation of the Gesture Matching Distance in order to investigate the question of whether our proposal is able to find similar spatiotemporal patterns within 3D motion capture data streams accurately. To this end, we selected

[a] `http://projects.ict.usc.edu/3dig/`

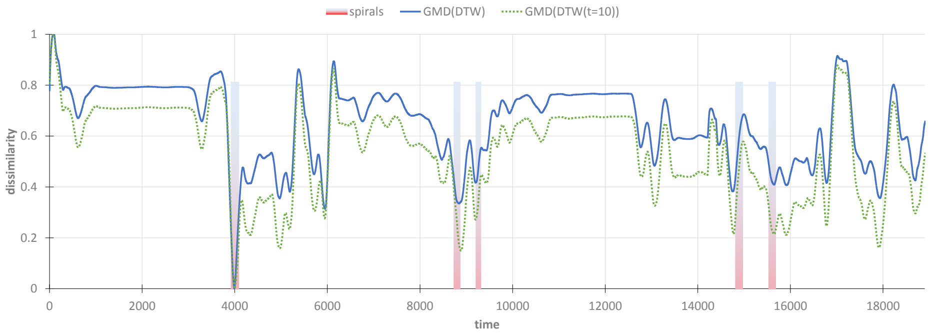

Fig. 2. Average dissimilarity values shown as a function of time with respect to gestural patterns of movement type *spiral*. Reddish time intervals depict gestural patterns included in the 3D motion capture data streams. The average dissimilarity of the exact Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}}$ is shown by the blue line, while the average dissimilarity of the approximate Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ is shown by the green dotted line.

different movement types and computed dissimilarity plots with respect to different gestural query patterns arising from the corresponding movement types in the natural media corpus. Based on the provided ground truth, we include one dissimilarity plot for the movement type spiral, which is shown in Figure 2, and two dissimilarity plots for the movement types straight and circle, which are shown in Figure 3 and Figure 4, respectively. The corresponding gestural patterns included in the 3D motion capture data streams are highlighted by means of reddish time intervals. The average dissimilarity values of the exact Gesture Matching Distance based on Dynamic Time Warping Distance $\mathrm{GMD}_{\mathrm{DTW}}$ are shown by blue lines, while the average dissimilarity values of the approximate Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$, where we fixed the maximum permissible time difference $t \in \mathbb{N}$ to a value of 10, are shown by green dotted lines.

As can be seen in the figures, the approximate Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ shows a behavior similar to the exact Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}}$. Both are able to respond to the corresponding queries with low dissimilarity values. In fact, the maximum absolute difference of dissimilarity values between $\mathrm{GMD}_{\mathrm{DTW}}$ and $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ is below a value of 0.242, while the average deviation is below a value of 0.11. We thus conclude that the approximate Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ is able to compete with the non-approximate Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}}$ in terms of accuracy.

In the second series of experiments, we evaluated the effect of the bandwidth constraint applied to the Dynamic Time Warping Distance, where we fixed the maximum permissible time difference $t \in \mathbb{N}$ to a value of 10. The precision in percentage of the approximate Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ with respect to the exact $\mathrm{GMD}_{\mathrm{DTW}}$ is summarized in Figure 5. The precision values are depicted as a function of the gesture signature length and the number of trajectories for different

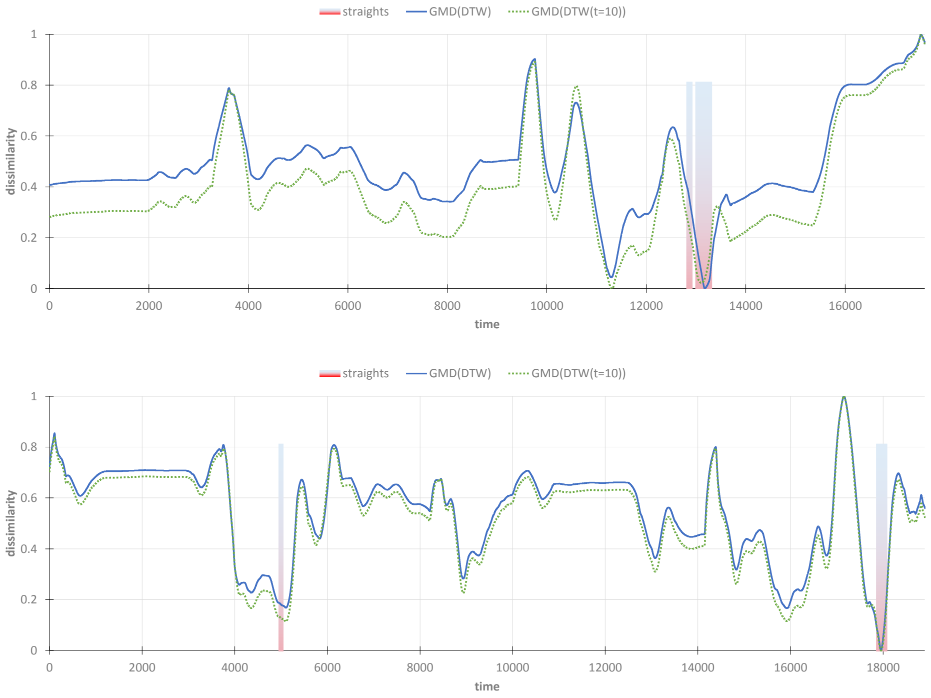

Fig. 3. Average dissimilarity values shown as a function of time with respect to gestural patterns of movement type *straight*. Reddish time intervals depict gestural patterns included in the 3D motion capture data streams. The average dissimilarity of the exact Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}}$ is shown by the blue line, while the average dissimilarity of the approximate Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ is shown by the green dotted line.

databases extracted from the 3D Iconic Gesture Dataset comprising 2k, 5k, and 10k gesture signatures. As can be seen in the figure, the precision values decrease with an increase in the length of the gesture signatures. At a gesture signature length of 0.5 seconds, the average precision stays at approximately 100%, which is reduced to approximately 93% when utilizing gesture signatures with a length of 2.0 seconds. An increase in the number of trajectories of the gesture signatures does not necessarily degenerate the performance of our approach. As observed empirically, gesture signatures comprising 6 trajectories always yield the highest precision values. This effect might be caused by the underlying movement traces of the corresponding trajectories. To sum up, the performance in terms of average precision of our proposal stays above 97%. Thus, the loss in accuracy is less than 3%, which is negligible in view of the increase in efficiency.

In the third series of experiments, we evaluated the query processing efficiency when utilizing the optimal multi-step algorithm as presented in Section 4 with the proposed lower bound approximation derived in Section 3. To this end, we investi-

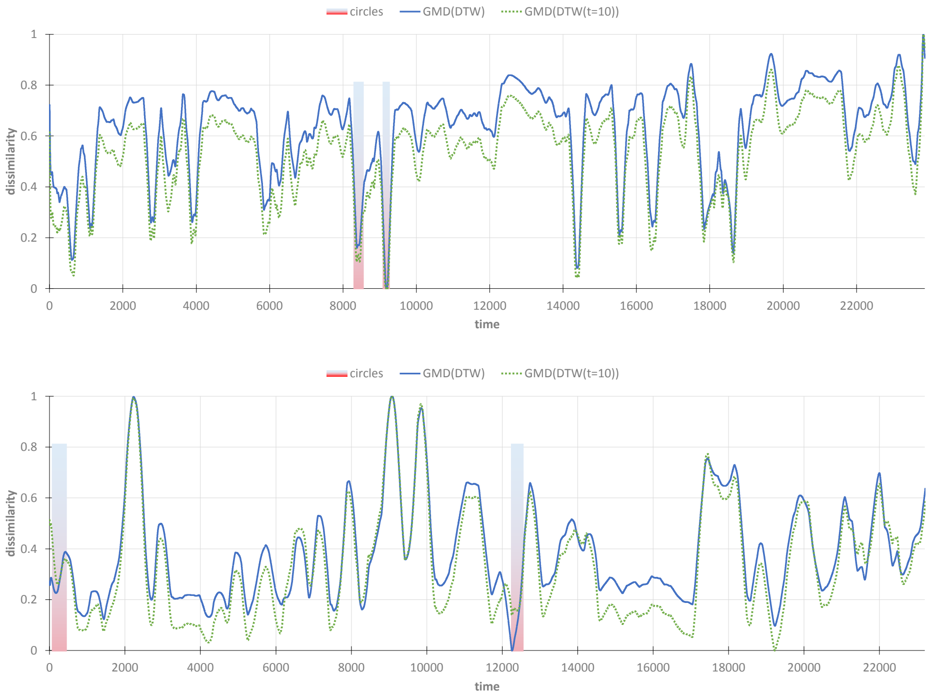

Fig. 4. Average dissimilarity values shown as a function of time with respect to gestural patterns of movement type *circle*. Reddish time intervals depict gestural patterns included in the 3D motion capture data streams. The average dissimilarity of the exact Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}}$ is shown by the blue line, while the average dissimilarity of the approximate Gesture Matching Distance $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ is shown by the green dotted line.

gated the average query response times needed for processing 100-nearest-neighbor queries in a database of 10k gesture signatures. As before, the length of the gesture signatures and the number of trajectories included in the gesture signatures are varied. The average query processing times in seconds are reported in Table 1. In general, the query response time increases by extending the length or the number of trajectories of the gesture signatures. As can be seen in the table, the sequential scan by means of the Gesture Matching Distance based on Dynamic Time Warping Distance $\mathrm{GMD}_{\mathrm{DTW}}$ shows the highest query response time. By utilizing gesture signatures comprising 10 trajectories with a length of 2 seconds, $\mathrm{GMD}_{\mathrm{DTW}}$ needs more than 176 seconds on average for query processing. This query processing time is reduced by more than one order of magnitude when processing the queries with the optimal multi-step algorithm based on the proposed lower bound approximation $\mathrm{GMD}_{\mathrm{LB}_{Keogh}}$. For the aforementioned parameters, our approach is able to complete query processing in 14 seconds on average. By increasing the size of the database to 100k gesture signatures comprising 10 trajectories with a length of 2 seconds each,

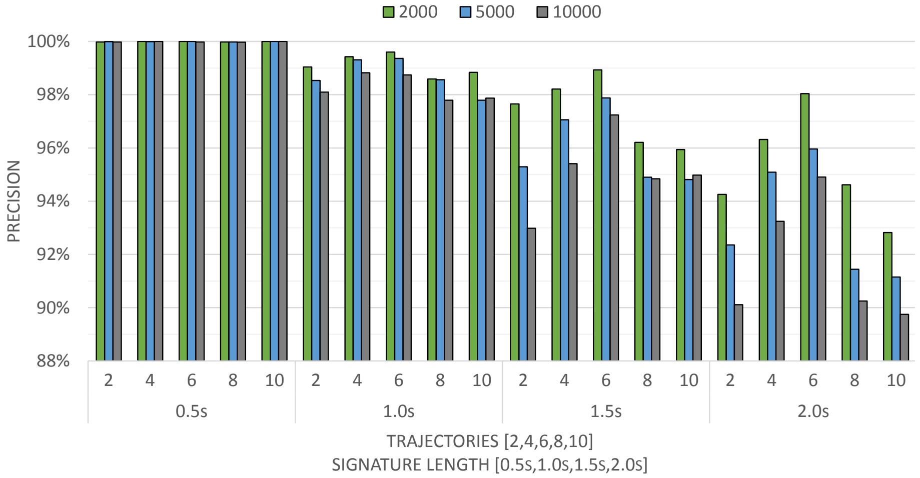

Fig. 5. Precision values in percentage of $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ in comparison to $\mathrm{GMD}_{\mathrm{DTW}}$ as function of gesture signature length and the number of trajectories. The database size is varied among 2k, 5k, and 10k gesture signatures.

the average query response times for the sequential scan with $\mathrm{GMD}_{\mathrm{DTW}}$ and that with $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ are approximately 1170 seconds and 445 seconds, respectively, whereas the optimal multi-step algorithm with the proposed lower bound approximation $\mathrm{GMD}_{\mathrm{LB}_{Keogh}}$ takes approximately 36 seconds. Thus, our approach is more than 12 times faster than the sequential scan with $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$ and more than 30 times faster than the sequential scan with $\mathrm{GMD}_{\mathrm{DTW}}$.

To conclude, the proposed lower bound approximation is able to achieve an increase in efficiency of more than one order of magnitude with a negligible loss in accuracy and thus enables efficient similarity search for gestural patterns in 3D motion capture databases. How the proposed approaches are utilized in the digital humanities, and in particular within the domain of gestural pattern analysis, is discussed by means of two research use cases in the following section.

6. Applications in Digital Humanities: Two Gesture Research Use Cases

Recent studies in linguistics and cognitive science have shown that the workings of the human mind are intricately bound to the workings of the human body [46, 47, 48, 49]. Concomitantly, research has shown that co-speech behavior is highly conventionalized and intricately tied to structures in the speech stream [50, 51, 52, 53, 54, 55]. Together, these conventionalized multimodal constructions [56, 57, 58] of speech and body movement convey the semantics and pragmatics of the message. Moving away from a long-time focus on text or speech in isolation, linguists – especially cognitive linguists – have increasingly targeted full-bodied,

Table 1. Average query response time in seconds for processing 100-nearest-neighbor queries in a database of 10k gesture signatures.

signature length	trajectories	opt. multi-step $\mathrm{GMD}_{\mathrm{LB}_{Keogh}}$	seq. scan $\mathrm{GMD}_{\mathrm{DTW}_{t=10}}$	seq. scan $\mathrm{GMD}_{\mathrm{DTW}}$
2.0s	10	14.0	67.5	176.4
	8	8.4	42.4	110.7
	6	3.6	24.1	64.1
	4	1.6	12.1	33.0
	2	0.5	3.1	8.3
1.5s	10	8.3	44.3	97.0
	8	5.0	29.3	62.3
	6	2.2	16.2	34.2
	4	1.3	9.6	20.3
	2	0.3	2.3	5.0
1.0s	10	4.6	21.4	34.0
	8	5.0	22.7	36.0
	6	1.3	10.0	15.8
	4	0.8	5.9	9.2
	2	0.2	1.5	2.4
0.5s	10	2.8	14.8	16.1
	8	1.6	7.9	8.2
	6	0.8	5.4	6.0
	4	0.3	1.6	1.8
	2	0.1	0.6	0.7

multimodal interaction as their object of study. However, due to the challenges inherent in studying multimodal data, which requires recording and tediously annotating full conversations, gesture studies to date have largely been based on case studies of one or two individuals in conversation. With recent advances in digital data, such as the availability of a few large video-based language corpora that provide audio-video streams with time-aligned closed-captioning text[b], linguists now have hundreds of spontaneous conversations available to them for study. However, the annotation of body movement remains complex and highly reliant on qualitative measures based on an annotator's visual examination of a video played at reduced speed.

3D motion capture, however, provides a radically different lens through which to both capture and examine multimodal data. Here we describe two research programs

[b] For example, UCLA's Little Red Hen lab is an international research team using a proprietary online corpus of more than 200,000 hours of broadcast television: `https://sites.google.com/site/distributedlittleredhen/home`

that use 3D motion capture data to investigate certain structures in embodied representations (i.e. gesture, head movement, and other modalities) and how these are co-articulated with structures in the speech stream. The aim of each study is the understanding of the interaction between conceptual structure, linguistic structure – i.e. the speech, and embodied/physical structure.

6.1. *Aspectual contours: Matching verb types with gestural movement types*

In the study presented in [59, 60], 3D motion capture data was used to investigate the gestural profiles corresponding to linguistic utterances conveying grammatical phenomenon known as *aspect* – or the linguistic phenomenon that captures how speakers modulate the "temporal contour" of an event [61]. *Contour* implies a shape in space, making aspect a natural grammatical category through which to explore the trajectories of co-speech gestures. Aspect encodes the ways in which an event can be construed dynamically by performing additional computations without losing the character of the original event [62]. For example, inherent in the meaning of verbs in English such as *sneeze* is the interpretation of the event as a bounded, punctual, single episode. However, aspect is dynamic and can be altered in interaction with grammatical elements that have aspectual force. Using a normally bounded verb such as *sneeze* in a progressive construction (*-ing*) renders the event unbounded and yields an iterative interpretation, as in *He is sneezing* or *He keeps sneezing* [63, 64].

In a study of the co-speech gestures associated with aspect-marking auxiliary verbs in English, [53] examined constructions in English headed by the auxiliary co-verbs *continue*, *keep*, *start*, *stop* and *quit* (e.g. keep sneezing, stop talking, quit smoking, etc.). The goal of the study was to determine if gestures correlated with these auxiliary constructions were conventionalized across speakers, and if so, what the conventionalized features of the gestures are for each construction. Results showed a statistically significant correlation between both the timing and the form of the gesture and the aspect marked in the auxiliary verb in the speech stream. The gesture profile of the auxiliary *keep*, for example, was characterized by longer onset times (i.e., a greater latency between onset of gesture and onset of the auxiliary verb in the speech) and repeated gesture strokes, many of which were cyclic or spiral in trajectory. This study and others [65, 60] have noted that prototypical movement profiles are readily recognizable in co-speech gesture given certain linguistic cues.

In a follow-up study [59, 60], 3D motion capture data was used to explore the forms that emerged as aspect-marking profiles in [53]. The motion capture data increased the sophistication of the analysis by allowing us to investigate the degree of similarity between the gesture profiles corresponding to spoken utterances, as well as providing more nuanced visualizations of the movement traces and temporal dynamics of these gestures. Eight native speakers of North American English were recorded in the Natural Media Lab of RWTH Aachen University in casual conversation with a confederate, who remained the same for all participants. Participants recounted and interpreted a short movie and conversed about topics such as habits and hobbies, resulting in approximately six hours of recorded interaction. To analyze the data, we identified those discourse sequences in which the trajectory, direction, and form of the gesture trace (circle, spiral, arc, etc.) reflected one of the conventionalized, aspectually-charged forms established in the previous research [53].

The computational analyses of gesture similarity by means of a distance-based similarity model [1] enables us to recognize in a quantitative manner (rather than relying on visual assessment) which trajectory type a gesture has. This proves most useful in differentiating forms, for example, a spiral and circle, which differ only in displacement in space for the former, or lack thereof for the latter. Such a distinction is difficult to make unequivocally using manual annotation which relies on a researcher's observation of a video (possibly involving poor camera angles of the gesture) played at reduced speed.

6.2. *Establishing multimodal clusters in dialogic travel-planning data*

Brenger et al. [66] investigated multimodal constructions that may be observed when interlocutors utilize their gesture spaces for spatial-geographical orientation during collaborative travel planning (e.g. planning an Interrail trip through Europe). The basis for the study was the Multimodal Speech & Kinetic Action Corpus (MuSKA), compiled in the Natural Media Lab of RWTH Aachen University [67, 68]. As in the previous use case, several data streams were recorded and aligned in the Natural Media Lab (audio, video and 3D motion capture data), though in this study the recorded conversations were informal dialogues between friends. In speech, indicating potential travel destinations and routes typically involves the use of highly context-dependent indexical expressions such as certain functional *closed-class items* [63] or *shifters* [69]. Examples include prepositions, pronouns, demonstratives, and connectors. The assumption underlying this study was that, in spoken German discourse, the use of place names and indexical expressions – such as prepositions (e.g., *nach* (to), *von* (from), *bei* (at)) and locative or directional adverbials (e.g., *da* (there), *hier* (here), *rber* (over)) – would correlate with distinct kinds of gestures, namely *locating* and *routing* gestures.

More specifically, the study's target structures were prepositional phrases such as constructions combining prepositions and adverbials (e.g., PREP + ADV such as, *nach hier*, *nach da* (to here/there)) or prepositions and nouns (e.g., PREP + N such as *von Norden* (from the north), *nach Paris* (to Paris)). We also included adverbial phrases comprising locative and directional adverbial such as $\text{ADV}_{\text{locative}}$ + $\text{ADV}_{\text{directional}}$ (e.g., *da rber* (over there), *hier hin* (to here)).

The "travel planning"-sub-corpus contains 60 minutes of annotated discourse data, with speech transcripts coded for shifters and the adverbial and prepositional phrases in which they occur. The video data were coded for gestural shifts exhibiting locating or routing functions. In three dialogues (42 minutes in total), 300 gesture-accompanied occurrences of locative prepositions and adverbials were identified (130 place names; 170 combinations of prepositions with either locative or directional adverbials. PREP + $\text{ADV}_{\text{locative}}$ or $\text{ADV}_{\text{directional}}$). Regarding spatial orientation and gestural charting, we observed two main strategies: a) indicating places (cities, countries) through *locating gestures*; and b) tracing trajectories through *routing gestures*. We hypothesized that whereas prepositional phrases entailing place names or locative adverbs correlate with indexical locating gestures, deictic adverbial phrases may co-occur with both locating gestures and routing gestures containing specific directional movement information that is not necessarily specified in the concurrent speech [70, 71, 72, 52]. In addition to analyzing gestural patterns and multimodal clusters with the help of the computational methods presented in this paper, we are currently working on appropriate ways to visualize the data in the form of heat maps.

6.3. *Insights and Future Directions*

In both case studies outlined here – and indeed throughout gesture studies, whether working with motion capture data or video data – the methodology continues to require manually searching of corpus data and annotated ELAN files for linguistic phrases and then comparing the corresponding gestures to each other in terms of their spatiotemporal similarity. Thus, in the travel-planning study, the main effort lay in manually identifying spatiotemporal aspects and properties of corresponding gestures that allowed them to be regarded as routing or placing gestures – a very time consuming venture. However, the strength of the approach presented in this paper is in its goal of integrating the various audio and video data streams and annotated transcripts into a motion-capture driven multimodal database that can be efficiently searched with the types of query processing algorithms presented here. This would enable the semi-automatic search for gestures' spatial and temporal characteristics with their co-occurring linguistic structures. The computational identification of inter-gestural similarity would dramatically speed up the search process and thus enable future gesture researchers to explore larger corpora than is currently possible with manual searching. With regards to computational gesture signatures, these investigations could additionally be extended to the identification of any gestural pattern – locating and routing gestures, aspect-marking forms, and others, rather than relying on the linguistic target phrases that currently drive the searches.

Part of the value of interdisciplinary approaches to complex, dynamic multimodal data such as the collaborations presented here lies in the reciprocity of the collaborations: not only are speech and gesture data in multiple streams a welcome, and increasingly necessary, challenge for computer scientists, the computational approach is also increasingly crucial for linguists and gesture researchers. For instance, one prerequisite to identifying relatively stable patterns of correlated linguistic and gestural structures is that the multimodal cluster is used with a relatively high frequency. The computational methods applied to multimodal speech and gesture data suggested here would thus also enable linguists and gesture researchers to contribute to the advancement of the still young area of multimodal cluster analysis and thus to predict communicative behavior in certain utterance contexts. The inclusion of an aligned similarity search for syntactic structures and phrases, coupled with the presented similarity search for kinetic movement patterns, could take this promising venture one step further when it comes to identifying time-elastic multimodal clusters in larger multimodal corpora.

7. Conclusions

In this paper, we have addressed the issue of efficiently accessing gestural patterns in 3D motion capture data based on spatiotemporal similarity. To this end, we modeled gestural patterns by means of gesture signatures and investigated a lower bound approximation of the Gesture Matching Distance. Our approach is able to achieve an increase in efficiency of more than one order of magnitude with a negligible loss in accuracy. We thus claim that the proposed distance-based approach to gestural pattern analysis enables the semi-automatic investigation of large heterogeneous motion capture data archives.

References

[1] C. Beecks, M. Hassani, J. Hinnell, D. Schüller, B. Brenger, I. Mittelberg, and T. Seidl, "Spatiotemporal similarity search in 3d motion capture gesture streams," in *Proceedings of the 14th International Symposium on Spatial and Temporal Databases (SSTD 2015)*, 2015.

[2] J. Blackburn and E. Ribeiro, "Human motion recognition using isomap and dynamic time warping," in *Human Motion–Understanding, Modeling, Capture and Animation.* Springer, 2007, pp. 285–298.

[3] J. Yang, Y. Li, and K. Wang, "A new descriptor for 3d trajectory recognition via modified cdtw," in *Automation and Logistics (ICAL), 2010 IEEE International Conference on.* IEEE, 2010, pp. 37–42.

[4] M. Hahn, L. Krüger, and C. Wöhler, "3d action recognition and long-term prediction of human motion," in *Computer Vision Systems.* Springer, 2008, pp. 23–32.

[5] L. J. Latecki, V. Megalooikonomou, Q. Wang, R. Lakaemper, C. A. Ratanamahatana, and E. Keogh, "Elastic partial matching of time series," in *Knowledge Discovery in Databases: PKDD 2005.* Springer, 2005, pp. 577–584.

[6] M. Vlachos, M. Hadjieleftheriou, D. Gunopulos, and E. Keogh, "Indexing multidimensional time-series with support for multiple distance measures," in *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining.* ACM, 2003, pp. 216–225.

[7] M. Vlachos, G. Kollios, and D. Gunopulos, "Elastic translation invariant matching of trajectories," *Machine Learning*, vol. 58, no. 2-3, pp. 301–334, 2005.

[8] L. Chen and R. Ng, "On the marriage of lp-norms and edit distance," in *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30.* VLDB Endowment, 2004, pp. 792–803.

[9] L. Chen, M. T. Özsu, and V. Oria, "Robust and fast similarity search for moving object trajectories," in *Proceedings of the 2005 ACM SIGMOD international conference on Management of data*. ACM, 2005, pp. 491–502.

[10] S. Fang and H. Chan, "Human identification by quantifying similarity and dissimilarity in electrocardiogram phase space," *Pattern Recognition*, vol. 42, no. 9, pp. 1824–1831, 2009.

[11] C. Beecks, M. Hassani, F. Obeloer, and T. Seidl, "Efficient distance-based gestural pattern mining in spatiotemporal 3d motion capture databases," in *Proceedings of the 15th International Conference on Data Mining Workshops*, 2015.

[12] T. Seidl and H.-P. Kriegel, "Optimal multi-step k-nearest neighbor search," in *Proceedings of the ACM SIGMOD International Conference on Management of Data*, 1998, pp. 154–165.

[13] C. Beecks, "Distance-based similarity models for content-based multimedia retrieval," Ph.D. dissertation, RWTH Aachen University, 2013. [Online]. Available: http://darwin.bth.rwth-aachen.de/opus3/volltexte/2013/4807/

[14] C. Beecks, S. Kirchhoff, and T. Seidl, "On stability of signature-based similarity measures for content-based image retrieval," *Multimedia Tools and Applications*, vol. 71, no. 1, pp. 349–362, 2014.

[15] C. Beecks and T. Seidl, "On stability of adaptive similarity measures for content-based image retrieval," in *Proceedings of the International Conference on Multimedia Modeling*, 2012, pp. 346–357.

[16] C. Beecks, M. S. Uysal, and T. Seidl, "A comparative study of similarity measures for content-based multimedia retrieval," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2010, pp. 1552–1557.

[17] Y. Rubner, C. Tomasi, and L. J. Guibas, "The earth mover's distance as a metric for image retrieval," *International Journal of Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[18] C. Beecks, M. S. Uysal, and T. Seidl, "Signature quadratic form distance," in *Proceedings of the ACM International Conference on Image and Video Retrieval*, 2010, pp. 438–445.

[19] F. Hausdorff, *Grundzüge der Mengenlehre.* Von Veit, 1914.

[20] D. P. Huttenlocher, G. A. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.

[21] B. G. Park, K. M. Lee, and S. U. Lee, "Color-based image retrieval using perceptually modified hausdorff distance," *EURASIP Journal of Image and Video Processing*, vol. 2008, pp. 4:1–4:10, 2008.

[22] C. Beecks, S. Kirchhoff, and T. Seidl, "Signature matching distance for content-based image retrieval," in *Proceedings of the ACM International Conference on Multimedia Retrieval*, 2013, pp. 41–48.

[23] S. Mitra and T. Acharya, "Gesture recognition: A survey," *Trans. Sys. Man Cyber Part C*, vol. 37, no. 3, pp. 311–324, May 2007. [Online]. Available: http://dx.doi.org/10.1109/TSMCC.2007.893280

[24] N. A. Ibraheem and R. Z. Khan, "Article: Survey on various gesture recognition technologies and techniques," *International Journal of Computer Applications*, vol. 50, no. 7, pp. 38–44, 2012.

[25] R. Z. Khan and N. A. Ibraheem, "Survey on gesture recognition for hand image postures." 2012, pp. 110–121.

[26] J. LaViola, "A survey of hand posture and gesture recognition techniques and technology," *Brown University, Providence, RI*, 1999.

[27] J. Liu and M. Kavakli, “A survey of speech-hand gesture recognition for the development of multimodal interfaces in computer games,” in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2010, pp. 1564–1569. [Online]. Available: http://dx.doi.org/10.1109/ICME.2010.5583252

[28] S. S. Rautaray and A. Agrawal, “Vision based hand gesture recognition for human computer interaction: a survey,” *Artificial Intelligence Review*, vol. 43, no. 1, pp. 1–54, 2015.

[29] S. Ruffieux, D. Lalanne, E. Mugellini, and O. A. Khaled, “A survey of datasets for human gesture recognition,” in *Human-Computer Interaction. Advanced Interaction Modalities and Techniques - 16th International Conference*, ser. Lecture Notes in Computer Science, vol. 8511. Springer, 2014, pp. 337–348.

[30] R. Watson, “A survey of gesture recognition techniques,” Trinity College Dublin, Department of Computer Science, Tech. Rep., 1993.

[31] Y. Wu and T. S. Huang, “Vision-based gesture recognition: A review,” in *Gesture-based communication in human-computer interaction.* Springer, 1999, pp. 103–115.

[32] T. B. Moeslund and E. Granum, “A survey of computer vision-based human motion capture,” *Computer vision and image understanding*, vol. 81, no. 3, pp. 231–268, 2001.

[33] T. B. Moeslund, A. Hilton, and V. Krüger, “A survey of advances in vision-based human motion capture and analysis,” *Computer vision and image understanding*, vol. 104, no. 2, pp. 90–126, 2006.

[34] C. Keskin, A. Erkan, and L. Akarun, “Real time hand tracking and 3d gesture recognition for interactive interfaces using hmm,” *ICANN/ICONIPP*, vol. 2003, pp. 26–29, 2003.

[35] Y. Nam and K. Wohn, “Recognition of hand gestures with 3d, nonlinear arm movement,” *Pattern recognition letters*, vol. 18, no. 1, pp. 105–113, 1997.

[36] A. Psarrou, S. Gong, and M. Walter, “Recognition of human gestures and behaviour based on motion trajectories,” *Image and Vision Computing*, vol. 20, no. 5, pp. 349–358, 2002.

[37] H.-I. Suk, B.-K. Sin, and S.-W. Lee, “Hand gesture recognition based on dynamic bayesian network framework,” *Pattern Recognition*, vol. 43, no. 9, pp. 3059–3072, 2010.

[38] ——, “Recognizing hand gestures using dynamic bayesian network,” in *Automatic Face & Gesture Recognition, 2008. FG’08. 8th IEEE International Conference on.* IEEE, 2008, pp. 1–6.

[39] J. Cheng, C. Xie, W. Bian, and D. Tao, “Feature fusion for 3d hand gesture recognition by learning a shared hidden space,” *Pattern Recognition Letters*, vol. 33, no. 4, pp. 476–484, 2012.

[40] T. Arici, S. Celebi, A. S. Aydin, and T. T. Temiz, “Robust gesture recognition using feature pre-processing and weighted dynamic time warping,” *Multimedia Tools Appl.*, vol. 72, no. 3, pp. 3045–3062, 2014. [Online]. Available: http://dx.doi.org/10.1007/s11042-013-1591-9

[41] S. Bodiroža, G. Doisy, and V. V. Hafner, “Position-invariant, real-time gesture recognition based on dynamic time warping,” in *Proceedings of the 8th ACM/IEEE international conference on Human-robot interaction.* IEEE Press, 2013, pp. 87–88.

[42] H. Stern, M. Shmueli, and S. Berman, “Most discriminating segment–longest common subsequence (mdslcs) algorithm for dynamic hand gesture classification,” *Pattern Recognition Letters*, vol. 34, no. 15, pp. 1980–1989, 2013.

[43] H. Hasan and S. Abdul-Kareem, “Static hand gesture recognition using neural networks,” *Artificial Intelligence Review*, vol. 41, no. 2, pp. 147–181, 2014.

[44] E. J. Keogh, "Exact indexing of dynamic time warping," in *Proceedings of 28th International Conference on Very Large Data Bases*, 2002, pp. 406–417.

[45] A. Sadeghipour, L.-P. Morency, and S. Kopp, "Gesture-based object recognition using histograms of guiding strokes," in *Proceedings of the British Machine Vision Conference*, 2012.

[46] B. Bergen and K. Wheeler, "Grammatical aspect and mental simulation," *Brain and Language*, vol. 112, no. 3, pp. 150–158, 2010.

[47] R. W. Gibbs Jr, *Embodiment and cognitive science.* Cambridge University Press, 2005.

[48] C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and J. Bressem, Eds., *Body – Language – Communication: An international handbook on multimodality in human interaction: Vol. 2*, ser. Handbücher zur Sprach- und Kommunikationswissenschaft. Berlin and Boston: De Gruyter Mouton, 2014, vol. 38.2.

[49] C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and S. Teßendorf, *Body - Language - Communication: An International Handbook on Multimodality in Human Interaction. (Handbooks of Linguistics and Communication Science 38).* Berlin/ Boston: De Gruyter Mouton, 2013.

[50] J. Bressem, "A linguistic perspective on the notation of form features in gestures," in *Body – Language – Communication: Vol. 1*, ser. Handbücher zur Sprach- und Kommunikationswissenschaft, C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and S. Teßendorf, Eds. Berlin and Boston: De Gruyter Mouton, 2013, vol. 38.1, pp. 1079–1098.

[51] C. Debras, "L?expression multimodale du positionnment interactionnel (multimodal stance-taking)." Ph.D. dissertation, 2013.

[52] E. Fricke, *Origo, Geste und Raum.* Mouton de Gruyter, 2007.

[53] J. Hinnell, "Multimodal aspectual constructions in north american english: A corpus analysis of aspect in co-speech gesture using little red hen," in *International Society of Gesture Studies (ISGS)*, 2014.

[54] I. Mittelberg, "The exbodied mind. cognitive-semiotic principles as motivating forces in gesture," in *Body – Language – Communication: Vol. 1*, ser. Handbücher zur Sprach- und Kommunikationswissenschaft, C. Müller, A. Cienki, E. Fricke, S. H. Ladewig, D. McNeill, and S. Teßendorf, Eds. De Gruyter Mouton, 2013, vol. 38.1, pp. 750–779.

[55] S. Schoonjans, "Modalpartikeln als multimodale konstruktionen. eine korpusbasierte kookkurrenzanalyse von modalpartikeln und gestik im deutschen," Ph.D. dissertation, 2014.

[56] A. E. Goldberg, *Constructions at work: The nature of generalization in language.* Oxford and New York: Oxford University Press, 2006. [Online]. Available: http://search.ebscohost.com/login.aspx?direct=true\&scope=site\&db=nlebk\&db=nlabk\&AN=215621

[57] F. Steen and M. Turner, *Multimodal construction grammar. Language and the Creative Mind.* CSLI, 2013, pp. 255–274.

[58] E. Zima, "Gibt es multimodale Konstruktionen? Eine Studie zu [V (motion) in circles] und [all the way from X PREP Y]," 2014.

[59] J. Hinnell, C. Beecks, M. Hassani, T. Seidl, and I. Mittelberg, "Multimodal auxiliary constructions in english: A quantitative image-schema analysis of aspectual contours in gesture," in *12th Conference on Conceptual Structure, Discourse and Language (CSDL)*, 2014.

[60] I. Mittelberg, J. Hinnell, C. Beecks, M. Hassani, and T. Seidl, "Emerging grammar in gesture: A motion-capture data analysis of image-schematic aspectual contours

in north american english speaker-gesturers.” in *International Cognitive Linguistics Conference (ICLC)*, 2015.

[61] B. Comrie, *Aspect: An introduction to the study of verbal aspect and related problems.* Cambridge university press, 1976, vol. 2.

[62] W. Frawley, *Linguistic Semantics.* Lawrence Erlbaum Associates, 1992. [Online]. Available: https://books.google.de/books?id=uyavMKhIfV8C

[63] L. Talmy, *Towards a Cognitive Semantics.* MIT Press, 2000.

[64] B. Heine and T. Kuteva, *World Lexicon of Grammaticalization.* Cambridge University Press, 2002. [Online]. Available: https://books.google.de/books?id=Ua3vSiz0gaEC

[65] A. Cienki, *Image schemas and mimetic schemas in cognitive linguistics and gesture studies*, ser. Benjamins Current Topics. John Benjamins Publishing Company, 2015. [Online]. Available: https://books.google.de/books?id=wxCqCgAAQBAJ

[66] B. Brenger, D. Schüller, M. Priesters, and I. Mittelberg, “3d heat maps of multimodal travel planning: Correlating prepositional and adverbial phrases with locating and routing gestures,” Accepted abstract for International Society for Gesture Studies (ISGS) Conference, 2016.

[67] B. Brenger, “Head gestures in dialogue - identification and computational analysis of motion-capture data profiles of speakers’ and listeners’ communicative action.” 2015.

[68] B. Brenger and I. Mittelberg, “Shakes, nods and tilts. motion-capture data profiles of speakers? and listeners? head gestures,” in *Proceedings of the 3rd Gesture and Speech in Interaction (GESPIN) Conference*, 2015.

[69] R. Jakobson, “Shifters, verbal categories and the russian verb,” in *Word and Language*, ser. Selected Writings. De Gruyter, 1971, vol. II. [Online]. Available: https://books.google.de/books?id=ASkcAAAAIAAJ

[70] H. H. Clark, “Pointing and placing,” in *Pointing. Where language, culture, and cognition meet*, S. Kita, Ed. Lawrence Erlbaum Assoc., 2003, pp. 243–268.

[71] K. Cooperrider and R. Núñez, “Across time, across the body: Transversal temporal gestures,” *Gesture*, vol. 9, no. 2, pp. 181–206, 2009.

[72] K. R. Coventry, T. Tenbrink, and J. E. Bateman, *Spatial language and dialogue.* Oxford University Press, 2009, vol. 3.