Distance Measure for Querying Sequences of Temporal Intervals

Orestis Kostakis, Panagiotis Papapetrou, and Jaakko Hollmén
Aalto University School of Science, Department of Information and Computer Science
PO Box 15400, FI-00076 Aalto, Finland
Helsinki Institute for Information Technology (HIIT)
{orestis.kostakis,panagiotis.papapetrou,jaakko.hollmen}@aalto.fi

ABSTRACT

Time series representations are not always rich enough to describe the temporal activity, for instance, when the context and the relations of the observed elements are of interest. Sequences of temporal intervals use such intervals as primitives in their representation, and allow focusing on the temporal relations of these elements. This is a useful representation of data across many domains. Searching, indexing, and mining such sequences is essential for domain experts in order to discover useful information out of them. In this paper, we formulate the problem of comparing sequences of temporal intervals and propose a novel distance measure. We discuss the properties of the measure and study its robustness in the domain of sign language. Experiments on real data show that the measure is robust in terms of retrieval accuracy even for high levels of artificially introduced distortion.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, experimentation

Keywords

sequence, temporal intervals, distance measure, American Sign Language

1. INTRODUCTION

The improved ability to measure and store information and the interest to analyse temporal behaviour of systems and people are dominant drivers behind the growth of time series databases. The time series analysis [9] usually builds on the assumption that time series have a fixed sampling frequency, resulting in a equally spaced measurements in time. Here, we are interested in a special kind of temporal data, where the elements are temporal intervals. The main motivation of our work is sign language data, where temporal intervals are a natural way to represent its elements. Proficiency in interpreting and learning the gestures that form sign language must be acquired in order to obtain a deeper understanding of it.

Sequences of temporal intervals exist in many application domains, such as human motion databases, sign language data, human activity monitoring, and medical data. Their main characteristic is the fact that events are not necessarily instantaneous but they may have a time duration. Thus, there can be several temporal relations between events in these sequences. An example of such sequence is shown in Figure 1.

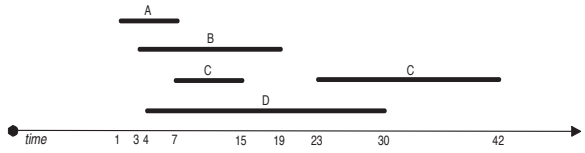


Figure 1: An example of a sequence of temporal intervals.

A review of temporal knowledge discovery paradigms and methods placing the temporal intervals into a proposed taxonomy knowledge discovery paradigms and methods [26].

Video technology can be used to transcribe a video frame sequence into a set of gestures. The transcription can be done manually by an expert or in an automated fashion by video analysis software — an interesting topic by itself. In this paper, however, we concentrate on the analysis of transcribed sequences. The individual gestures have a certain duration The are used in combination with others and carry the meaning of the words. If transcribed sequences are represented as sequential information, a faithful representation would take this duration information into account. Standard time series analysis, which would take these kind of relations into account explicitly, is not possible. This is the problem, which we try to alleviate by describing a suitable distance measure for collections of temporal intervals. Once a distance measure has been defined, a similarity can be computed between a transcribed (query) sequence and a database of transcribed sequences. For instance, given a query we may look up similar sequences, or train the user to make correct gestures in expressing a given word or sentence.

A very straightforward solution to comparing temporal interval sequences would be to map each one of them to

PETRA'11 May 25 - 27, 2011, Crete, Greece.

a sequence of instantaneous events by only considering the start and end points of each event interval. This would result in a simplification of the representation of sequences of interval-based events as they would be mapped to regular sequences of instantaneous events without the need for keeping track of the pair-wise event-interval relations. In addition, the size of the alphabet (i.e., set of all possible event labels) would double since each event interval label would be mapped to two instantaneous event labels corresponding to the start and end point of the event interval. Then, the solution to the problem would reduce to applying an existing distance/similarity measure for sequence matching, such as the edit distance [16]. The aforementioned solution may sound simple enough, though it is not correct, as it may lead to a large number of false positives. Consider the example shown in Figure 2. Obviously, the mapping for both sequences is the same: $\{A_{start},\ A_{start}, A_{end}, A_{end}\}$ and the edit distance would match them fully even though the relation between the two event intervals is different in the two sequences. Hence, we can deduce that in order to provide a robust similarity measure for such sequences, their representation should include additional information about the relations between the event intervals.

Figure 2: Two different sequences of temporal intervals where the mapping to a sequence of instantaneous events may produce the same representation for both.

The main focus of this paper is to study sequences of temporal intervals from the domain of *American Sign Language (ASL)*, since it is closely related to *assistive environments*. Our target is the development of a distance measure for comparing such sequences that could be further used for searching and indexing in large sequence databases, or for several data mining [10] tasks, such as summarisation, clustering, and classification. It should be noted, however, that our wider goal is not just to devise a distance measure but to define it in such a way that it provides a clear tie-in to significant patterns that may be found in each application domain (in our case we are looking for linguistically significant patterns). In other words, the measure should be able to discover new (and significant) patterns that domain experts would use in practice.

The main contributions of this paper include:

- the formulation of the problem of comparing sequences of temporal intervals and the definition of a novel distance measure to solve this problem
- discussion of the properties of the proposed measure
- experimental evaluation of the distance measure and its robustness on real data from the field of American sign language

This paper is organized as follows: in Section 2, we provide the necessary background and definitions along with the problem formulation; in Section 3 we present the related work on sequences of temporal intervals, while in Section 4 we describe the proposed distance measure and its properties. Section 5 provides an extensive experimentation on real data. Finally, Section 6 summarizes and further discusses the findings of our paper.

2. BACKGROUND

The present work is based on the framework of Allen [4], where the temporal intervals are considered as primitive elements in a sequence, and where the relations between the temporal intervals are represented explicitly. In this Section, we formulate the problem, we present the seven relations among temporal intervals that we consider and, additionally, we describe the used notation.

2.1 Problem Formulation

Let $\mathcal{S}$=$\{S_1, \ldots, S_n\}$ be an ordered set of events occurring at time intervals, called *sequence of temporal intervals* or *e-sequence*. Given an alphabet of event labels σ, each $S_i = (E_i, t^i_{start}, t^i_{end})$ is called an *event interval*, where $E_i \in \sigma$ and t^i_{start},t^i_{end} denote the start and end time of E_i. The temporal order of the event intervals is ascending based on their start time; for intervals with the same start time the order is descending based on their end time. If ties still exist, alphabetical ordering is applied based on their labels. An instantaneous event E_i has $t^i_{start} = t^i_{end}$. An e-sequence of size k is a *k-e-sequence*. The e-sequence that is shown in Figure 1 corresponds to:

$$\mathcal{S} = \{(A, 1, 7), (B, 3, 19), (D, 4, 30), (C, 7, 15), (C, 23, 42)\}$$

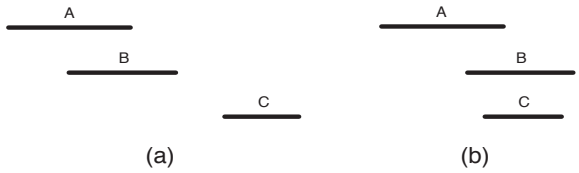


Figure 3: Two examples of arrangements of size 3 with alphabet $\sigma = \{A, B, C\}$.

In a database of e-sequences there may exist patterns of temporally related events; such patterns are called *arrangements*. An arrangement $\mathcal{A}$ of n events is defined as $\mathcal{A} = \{\mathcal{E}, \mathcal{R}\}$, where $\mathcal{E}$ is the set of event labels of all intervals that occur in $\mathcal{A}$, with $|\mathcal{E}| = n$, and $\mathcal{R} = \{R(E_1, E_2), R(E_1, E_3), \ldots, R(E_{n-1}, E_n)\}$ is the set of all interval relations $R(E_i, E_j)$ between each ordered pair (E_i, E_j), for $i = 1, \ldots, n$ and $j = i + 1, \ldots, n$, with $R(E_i, E_j) \in \mathcal{I}$. $\mathcal{I} = \{r_1, \ldots, r_{|\mathcal{I}|}\}$ denotes the set of all legal temporal relations that can exist between any pair of events and it will be defined in more detail in Section 2.2. The size of an arrangement $\mathcal{A} = \{\mathcal{E}, \mathcal{R}\}$ is equal to $|\mathcal{E}|$. Two examples of arrangements of size 3 are shown in Figure 3. For the first case (Figure 3(a)) $\mathcal{E} = \{A, B, C\}$ and $\mathcal{R} = \{overlap,\ follow,\ follow\}$ whereas for the second case (Figure 3(b)) $\mathcal{E} = \{A, B, C\}$ and $\mathcal{R} = \{overlap,\ overlap,\ contain\}$.

Based on the above formulation, we can now define problem of *comparing sequences of temporal intervals* as follows.

PROBLEM 2.1. *Given two e-sequences $\mathcal{S}$ and $\mathcal{T}$, define a distance measure D, such that $\forall \mathcal{S}, \mathcal{T}$:*

$$D(\mathcal{S},\mathcal{T}) \geq 0 \quad (1)$$
$$D(\mathcal{S},\mathcal{S}) = 0 \quad (2)$$
$$D(\mathcal{S},\mathcal{T}) = D(\mathcal{T},\mathcal{S}) \quad (3)$$

The degree to which the two e-sequences differ should be reflected in the value of $D(\mathcal{S},\mathcal{T})$ and should be in accordance with the knowledge obtained from domain experts.

2.2 Temporal Interval relations

Based on Allen's model for temporal intervals and their relations [4, 3, 24, 25] we consider the following relations for two interval events A and B. This is also presented in Figure 4.

- **meet**(A,B) denotes the case where A meets B; B starts at the same time that A ends, $t^A_{end} = t^B_{start}$.
- **match**(A,B) denotes the case where A matches B; A starts and end at the same time as B, $t^A_{start} = t^B_{start}$ and $t^A_{end} = t^B_{end}$.
- **overlap**(A,B) denotes the case where A and B have overlapping parts; A starts before B and B starts before A has ended, $t^A_{end} > t^B_{start}$ and $t^A_{start} < t^B_{start}$.
- **contain**(A,B) denotes the case where A contains B; A starts before B starts and A ends after B ends, $t^A_{start} < t^B_{start}$ and $t^A_{end} > t^B_{end}$.
- **left contain**(A,B) denotes the case where A right-contains B; A and B start simultaneously and A ends after B, $t^A_{start} = t^B_{start}$ and $t^A_{end} > t^B_{end}$.
- **right contain**(A,B) denotes the case where A left-contains B; A starts before B starts but end simultaneously, $t^A_{start} < t^B_{start}$ and $t^A_{end} = t^B_{end}$.
- **follow**(A,B) denotes the case where A occurs before B; A ends before B begins, $t^A_{end} \leq t^B_{start}$.

We do not consider symmetric relations, i.e. B is contained by A, since these can be expressed by the above and thus, are redundant. Hence, based on the definitions in Section 2.1, $\mathcal{I} = \{meet, match, overlap, contain, left\ contain, right\ contain, follow\}$ and $|\mathcal{I}| = 7$.

Furthermore, as discussed in Papapetrou et. al [25] there may exist ambiguities between the aforementioned relations due to noise in the data; for simplicity, we do no consider this in our work.

3. RELATED WORK

Sequences of temporal intervals have been of great interest over the last decade. Existing studies, however, have focused merely on mining patterns and association rules, and not on comparing and querying such sequences.

Several approaches consider discovering frequent intervals in transactional databases [18, 28]. The intervals are, in many cases, not labelled and thus no relations between them are considered. Giannoti et. al [8] considers temporally annotated sequential patterns, where transitions from one event to another have a time duration. In addition, a graph-based approach [13] represents each temporal pattern by a graph; nonetheless, only two types of relations are considered (*follow* and *overlap*).

A generalized interval-based framework has been developed [15] with improved support counting techniques for mining interval-based episodes, without, however, considering any temporal relations or association rules between the events. Apriori-based techniques [11, 12, 20] for finding temporal patterns that and association rules on interval-based event sequences have been proposed, some [12] also applying interestingness measures to evaluate the significance of the findings. Another approach that considers sequences of interval-based events in a database is discussed in [14], however limited to certain forms of arrangements. Recent BFS-based and DFS-based approaches [29, 24, 23, 25] apply efficient pruning techniques thus reducing the inherent exponential complexity of the mining problem. In [30], a non-ambiguous event-interval representation is defined that considers the start and end points of each e-sequence and converts the interval-based representation to a sequential representation. Finally, there has been some recent work on mining semi-partial orders of time intervals [21].

Furthermore, several methods [6, 1] have been focusing on mining association rules from data that contains interval-based events. The support of the rules is measured only during these intervals. Moreover, in Ale et. al [2], the lifetime of an item is defined as the time between the first and the last occurrence and the temporal support is calculated with respect to this interval. In this way, the extracted rules are only active during a certain time, and outdated rules can be pruned by the user. Finally, Lu et. al [19] studies inter-transaction association rules by merging all itemsets within a sliding time window inside a transaction, whereas in [27] efficient techniques for mining spatio-temporal patterns are proposed.

Despite the active research in sequences of temporal intervals, the main focus of the existing literature is limited to mining patterns and association rules. To the best of our knowledge, there has been yet no robust distance or similarity measure for comparing such sequences.

4. DISTANCE MEASURE

In this section we define a distance measure for comparing two e-sequences. Focus is given on the relations among the event intervals, disregarding absolute time values. The steps involved in this computation include: (1) dropping the time stamps and representing each of the two e-sequences as an arrangement, (2) mapping each arrangement to a *relation matrix*. The two matrices are then compared to derive the distance score.

4.1 Relation Matrix

To define a relation matrix for an e-sequence $\mathcal{S}$ we should first map the e-sequence to an arrangement $\mathcal{A} = \{\mathcal{E}, \mathcal{R}\}$. The relation matrix $M_{\mathcal{A}}$ of $\mathcal{A}$ is a $|\mathcal{I}| \times |\sigma|^2$ integer-valued matrix that keeps track of the count of all $|\mathcal{I}|$ types of temporal relation pairs that may occur in the arrangement.

DEFINITION 4.1. *Given an arrangement $\mathcal{A}$, the corresponding relation matrix $M_{\mathcal{A}}$ is defined as follows:*

$$M_{\mathcal{A}}(i,j) = |R(E_k, E_l) = r_i|, \quad (4)$$

$\forall r_i \in \mathcal{I},\ j \in [1, |\sigma|^2],\ k \in [1, |\mathcal{E}|],\ l \in [k+1, \ldots, |\mathcal{E}|]$.

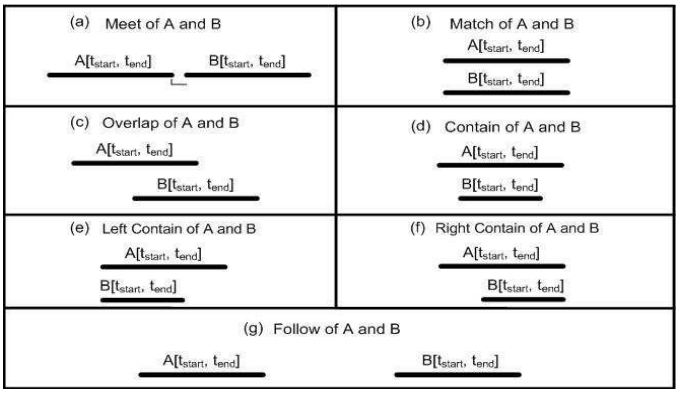


Figure 4: The seven temporal relations between two event intervals that are considered in this paper.

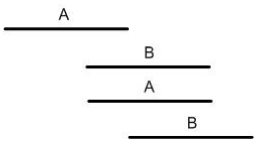
Rows of $M_{\mathcal{A}}$ correspond to relations among intervals (as defined in $\mathcal{I}$) and columns correspond to pairs of interval labels in σ. The value of each cell is the number of times a relation between the corresponding intervals occurs in $\mathcal{A}$. For example, $M_{\mathcal{A}}(1,1)$ denotes the number of times relation r_1 appears between (E_1, E_1) in $\mathcal{A}$.

For any arrangement $\mathcal{A}$ of size k, it holds that

$$\sum_{i=1}^{|\mathcal{I}|}\sum_{j=1}^{|\sigma|^2} M_{\mathcal{A}}(i,j) = \frac{k(k-1)}{2}. \quad (5)$$

To better illustrate this mapping, consider the example in Table 1 that demonstrates the relation matrix of the arrangement shown in Figure 5.

Figure 5: An arrangement with alphabet $\sigma = \{A, B\}$. Its corresponding relation matrix is shown in Table 1.



relation	{A,A}	{A,B}	{B,A}	{B,B}
meet	0	1	0	0
match	0	0	1	0
overlap	1	2	0	1
contain	0	0	0	0
left-contain	0	0	0	0
right-contain	0	0	0	0
follow	0	0	0	0

Table 1: The relation matrix $M_{\mathcal{A}}$ of arrangement $\mathcal{A}$ shown in Figure 5.

4.2 Arrangement Distance

Suppose we would like to compare e-sequences $\mathcal{S}$ and $\mathcal{T}$. We should first express these sequences with respect to their arrangement representation (say, $\mathcal{A}$ and $\mathcal{B}$, respectively), i.e., ignoring the event interval durations and considering only the temporal relations between the events. Then each arrangement will be mapped to its corresponding relation matrix representation; thus creating matrices $M_{\mathcal{A}}$ and $M_{\mathcal{B}}$. Now, the problem of comparing the original e-sequences is mapped to the problem of comparing their relation matrices.

Definition 4.2. *We define the following generalized arrangement distance function:*

$$\delta_p(\mathcal{A},\mathcal{B}) = \left(\sum_{i=1}^{|\mathcal{I}|}\sum_{j=1}^{|\sigma|^2} |M_{\mathcal{A}}(i,j) - M_{\mathcal{B}}(i,j)|^p\right)^{\frac{1}{p}}, \quad p \in \mathbb{N}^* \quad (6)$$

4.2.1 The Manhattan Distance

For $p = 1$, Equation 6 yields the entry-wise *Manhattan* distance between $M_{\mathcal{A}}$ and $M_{\mathcal{B}}$. Nonetheless, using this distance, it is possible that the comparison of two equal sized arrangements may yield a higher score than the comparison of one of these arrangements with a significantly smaller one since they are entirely different in size. For example, consider three arrangements $\mathcal{A}$, $\mathcal{B}$, $\mathcal{C}$ with 5, 10, and 10 event intervals, respectively. Effectively, this means that $\mathcal{A}$ contains 10 possible temporal relations, while $\mathcal{B}$ and $\mathcal{C}$ contain 45. Now suppose that $\mathcal{A}$ and $\mathcal{B}$ agree on 8 relations, whereas $\mathcal{B}$ and $\mathcal{C}$ agree on 25 relations. Then, $\delta_1(\mathcal{A},\mathcal{B}) = 2+37 = 39$, since they differ on the remaining 2 relations of $\mathcal{A}$ and the remaining 37 relations of $\mathcal{B}$. Similarly, $\delta_1(\mathcal{B},\mathcal{C}) = 20+20 = 40$. This suggests that $\mathcal{B}$ is more similar to $\mathcal{A}$ than $\mathcal{C}$.

Taking a closer look, however, one may easily notice that the total number of relations that may exist between $\mathcal{A}$ and $\mathcal{B}$ is 55, while 90 relations may exist between $\mathcal{B}$ and $\mathcal{C}$. This means that the former pair of arrangements agree on a smaller fraction of relations (i.e., $\frac{2\times 8}{55} \approx 0.28$) than the later pair (i.e., $\frac{2\times 25}{90} \approx 0.56$). This anomaly would propagate into further procedures, such as clustering, giving incorrect results.

For this reason, we propose the following normalized version of δ_1:

Definition 4.3. *Given arrangements $\mathcal{A}$ and $\mathcal{B}$, the nor-*

malized manhattan distance is defined as follows:

$$\delta_{norm}(\mathcal{A},\mathcal{B}) = \frac{\sum_i \sum_j |M_{\mathcal{A}}(i,j) - M_{\mathcal{B}}(i,j)|}{\frac{|\mathcal{A}|(|\mathcal{A}|-1)}{2} + \frac{|\mathcal{B}|(|\mathcal{B}|-1)}{2}}$$
$$= 2 \times \frac{\sum_i \sum_j |M_{\mathcal{A}}(i,j) - M_{\mathcal{B}}(i,j)|}{|\mathcal{A}|(|\mathcal{A}|-1) + |\mathcal{B}|(|\mathcal{B}|-1)}$$
$$= \sum_{i=1}^{|\mathcal{I}|} \sum_{j=1}^{|\sigma|^2} \frac{|M_{\mathcal{A}}(i,j) - M_{\mathcal{B}}(i,j)|}{M_{\mathcal{A}}(i,j) + M_{\mathcal{B}}(i,j)} \quad (7)$$

For any pair of arrangements $\mathcal{A}$ and $\mathcal{B}$ the following three properties hold:

$$0 \leq \delta_{norm}(\mathcal{A},\mathcal{B}) \leq 1 \quad (8)$$
$$\delta_{norm}(\mathcal{A},\mathcal{A}) = 0 \quad (9)$$
$$\delta_{norm}(\mathcal{A},\mathcal{A}_{\emptyset}) = 1 \quad (10)$$

where $\mathcal{A}_{\emptyset}$ corresponds to the arrangement of the null e-sequence, i.e., a sequence without any intervals.

4.2.2 The Frobenius Norm

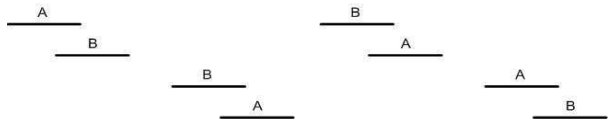
For $p = 2$, the distance expressed by Equation 6 is equal to the *Frobenius norm* of $M_{\mathcal{A}} - M_{\mathcal{B}}$:

$$\delta_2(\mathcal{A},\mathcal{B}) = \sqrt{\sum_{i=1}^{|\mathcal{I}|} \sum_{j=1}^{|\sigma|^2} |M_{\mathcal{A}}(i,j) - M_{\mathcal{B}}(i,j)|^2} \quad (11)$$

4.3 Properties

THEOREM 4.1. *δ_p is not metric.*

Proof: δ_p violates the identity of indiscernibles that should be satisfied by metric distance functions. Consider arrangements $\mathcal{A}$ and $\mathcal{B}$ shown in Figure 6. Clearly, $M_{\mathcal{A}}$ and $M_{\mathcal{B}}$ are identical even though they correspond to different arrangements.



(a) $\mathcal{A}$ (b) $\mathcal{B}$

Figure 6: Two arrangements for which δ_p violates the *identity of indiscernibles*.

It becomes apparent that neither the duration of temporal events nor the time separating them is taken into account. As a result, scaling the temporal values of a sequence does not affect the result when compared to another sequence.

In the case where basic arithmetic operations can be performed in constant time, i.e., addition, multiplication, and square root need $O(1)$ time units, the time needed to compare two arrangements $\mathcal{A}$ and $\mathcal{B}$ of sizes n and m respectively is $O(n^2 + m^2 + |\sigma|^2)$, assuming that both arrangements are defined over the same alphabet σ.

5. EXPERIMENTS

5.1 Dataset

We evaluated the robustness of the proposed distance measure on the American Sign Language database created by the National Center for Sign Language and Gesture Resources at Boston University[1]. The database contains a collection of 873 utterances, where each utterance associates a segment of video with a detailed transcription. Facial gestures play a crucial role in the grammar of ASL [5, 7, 17, 22], thus for our experiments we focused only on: specific types of gestures (e.g., raised eyebrows, head tilt forward), functional identification of clusters of these non-manual gestures that carry syntactic meaning (e.g., wh-question, negation), and part-of-speech identifications of manual signs (e.g., verb, wh-word), each one occurring over a time interval. A histogram of the e-sequence lengths in the ASL dataset is shown in Figure 7.

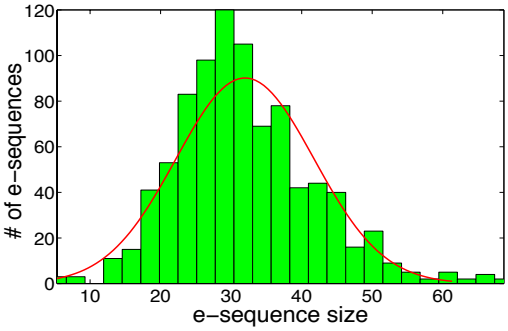


Figure 7: Histogram of the e-sequence lengths in the ASL dataset.

5.2 Setup

We tested the proposed measures using the ASL database. Queries were generated from the database sequences by adding noise to them. We generated 50 sets of 873 queries. For each set, a query was taken from a database sequence and then, each event interval in the query was shifted with probability p by a value bounded by a distortion factor of d. The chosen distortion factors were 10%, 20%, 30%, 40 and% 50% while probability levels varied from 0.1 to 1 with a step of 0.1. Put together, given a distortion level d as a percentage of the length of the whole query sequence, a random value under the uniform distribution is chosen in that integer interval. An interval has equal probability to be shifted back or forth in time. The durations of the intervals remained unaffected. Whether an interval would be affected by noise is determined independently by the probability parameter p.

Ideally, we would like each noisy sequence (query) to be matched to the sequence from which it originated. Thus, we performed a test individually for all elements in our database for various probability and distortion levels. We performed a linear search in the database and computed all pair-wise distances between each query and each database sequence.

We compared the Normalized distance, the Manhattan distance, and the Frobenius distance in terms of:

[1]http://www.bu.edu/asllrp/cslgr/.

1. **nearest neighbour retrieval accuracy:** the fraction of noisy queries for which the originating sequence is retrieved.

2. **rank of nearest neighbour:** for each query, the number of database sequences with distance at least than or equal to that of the originating counterpart..

Our experiments were implemented in Java and were performed on a PC running Ubuntu Linux, equipped with Intel Core 2 Duo 2GHz and 4GB RAM. To distort and query each of the 873 sentences, for the 50 different pairs of noise and probability parameters, required approximately a total of 4 hours.

5.3 Results

As it is expected, when the distortion and probability parameter values are increased, the distance of the distorted queries to their originating sequences is increased, too. In Figures 8(a), 8(b), 8(c) we display how the average distance value fluctuates for each distance function. On the x-axis we show the level of distortion imposed to the queries and on the y-axis the average distance value of all queries to their originating counterpart. The error-bars above the curves display the maximum distance of a query to its original sequence whereas the error-bars below the curves correspond to the minimum value.

The results of our main experiment are presented in Figures 9(a), 9(b), 9(c). These figures display the percentage of the query searches that were able to return the originating sequences. Different curves represent different values for the probability parameter. The Normalized distance function turns out to be very robust by maintaining a success rate well over 99.5%. For the other two, the success rate dropped significantly with an increase of the parameter values.

Figure 10 displays a comparison of the success rates of the three functions for probability parameter values 0.6 and 1.0 and distortion parameter value 50%. The Normalized distance function has a clear advantage over the other two. Throughout our experiments, when using the normalized distance, only 2 or 3 out of 873 sentences were not matched correctly and this happened for the very high values of the distortion and probability parameters. One of the reasons that the Frobenius and Manhattan Distances yield poor results compared to the Normalized, is the Nearest-Neighbour anomaly we described in Section 4.2. While the Normalized function is a clear winner, among the other two the Frobenius distance performs better.

Another important issue is the rank of the nearest neighbour distribution which, as mentioned earlier, is the total number of samples in the database that have distance from the distorted query less than or equal to the distance of the query to the originating sequence; if a query is matched correctly, it has a rank of 1. The cumulative histograms of the ranks can be seen in Figure 11. While these do not provide any additional insight concerning the Normalized distance (additional to what is shown in Figure 10), they demonstrate an overall advantage of the Frobenius distance over the Manhattan.

6. SUMMARY AND CONCLUSIONS

Sequences of temporal intervals are the basis representation used in this work. This representation is used for time-related phenomena, where recorded events in time have a duration and can not be faithfully represented by an instantaneous event in time nor a plain time series representation.

The methodological contribution of our work is to propose a distance measure for sequences of temporal intervals, in order to quantify similarity between sequences. We propose transforming the sequence information to relational information between all pairs of intervals and counting how often certain pre-defined relations (such as overlapping, following etc.) hold for all the pairs. These counts are summarized in a relation matrix; a distance is calculated as a function of the relation matrices of the sequences. We use Manhattan distance, Frobenius distance, and a normalized distance to quantify the similarity between the matrices. Based on the experiments the normalized distance is a clear winner among the three, with the Frobenius distance being the second.

In the experiments, we empirically investigate the noise robustness of the proposed measure by artificially introducing distortion to the database queries and see how similar the original query and corresponding, distorted query are in the sense of the used distances. The applied contribution is to use the distance measure for signs from the American Sign Language represented as sequences of temporal intervals. The notion of similarity quantified by the distance measure opens up many possibilities, such as developing translation systems for the American Sign Language where reference sentences are queried from a large collection of sign language utterances or developing teaching aids for those learning the sign language.

Directions for future work include evaluating the general applicability of the proposed measures by testing their robustness on other application domains, such as network traffic, body sensor networks, epidemiological studies, etc. In addition, this novel distance measure enables widening the applicability of standard machine learning tasks, such as clustering or classification.

Acknowledgments

This work has been partially funded by the Centre of Excellence for Algorithmic Data Analysis Research (ALGODAN), at University of Helsinki and Aalto University School of Science in Finland.

7. REFERENCES

[1] T. Abraham and J. F. Roddick. Incremental meta-mining from large temporal data sets. In *ER '98: Proceedings of the Workshops on Data Warehousing and Data Mining*, pages 41–54, 1999.

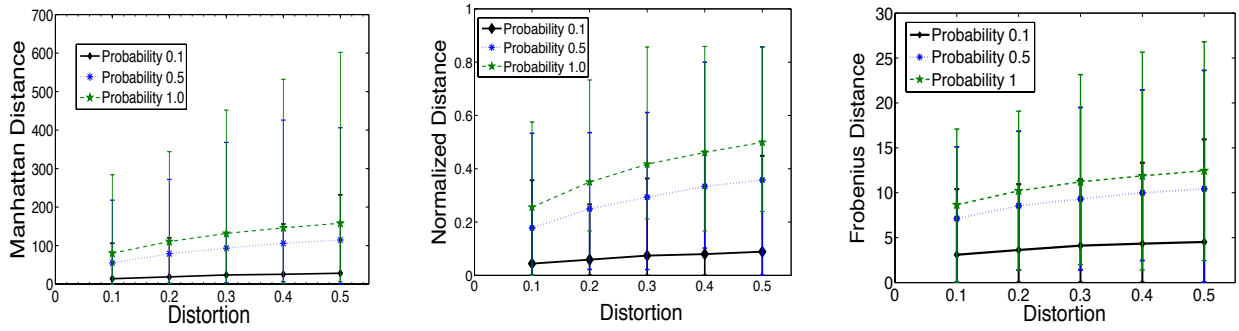
[2] J. M. Ale and G. H. Rossi. An approach to discovering temporal association rules. In *Proc. of the SAC*, pages 294–300, 2000.

[3] J. Allen and G. Ferguson. Actions and events in interval temporal logic. *Journal of Logic and Computation*, 1994.

[4] J. F. Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, November 1983.

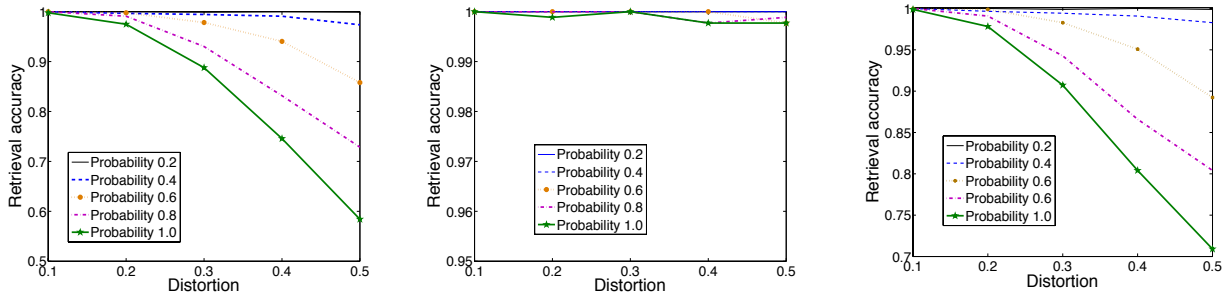
[5] C. Baker-Shenk. A micro-analysis of the nonmanual components of questions in American Sign Language. *Doctoral Dissertation*, 1983.

[6] X. Chen and I. Petrounias. Mining temporal features in association rules. In *Proc. of PKDD*, pages 295–300, London, UK, 1999. Springer-Verlag.



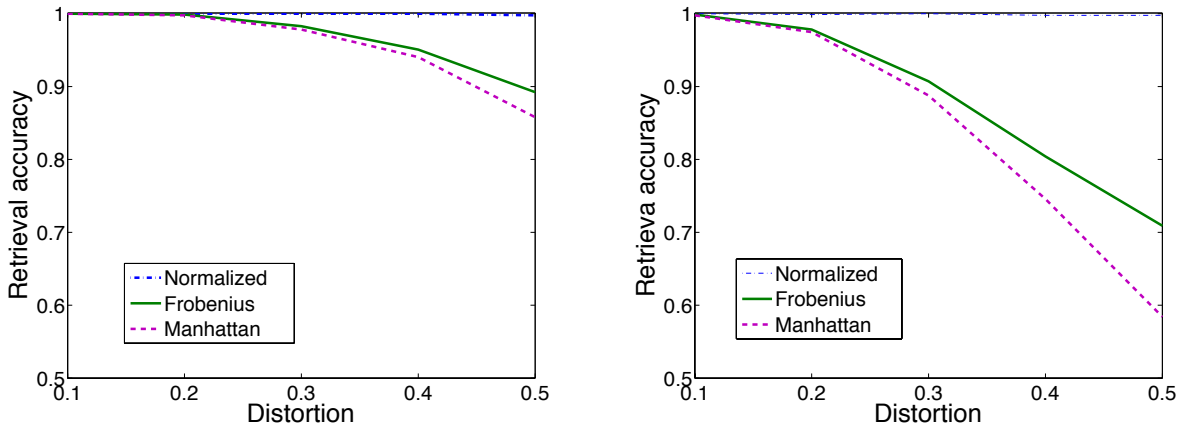
(a) Manhattan Distance (b) Normalized Distance (c) Frobenius Distance

Figure 8: Fluctuation of distance values for the distorted sequences. The curves correspond to the average distance of the noisy sequences to their originating counterparts. Error-bars show the minimum and maximum distance.



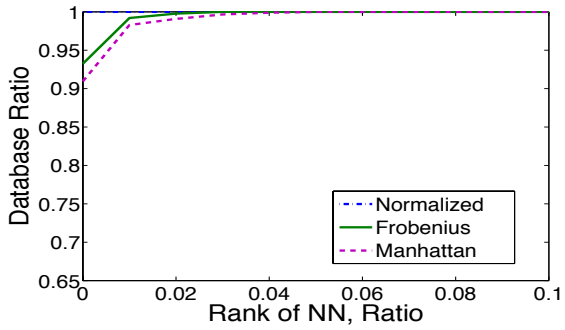
(a) Manhattan Distance (b) Normalized Distance (c) Frobenius Distance

Figure 9: Success ratio of matching the noisy sequences to their originating.

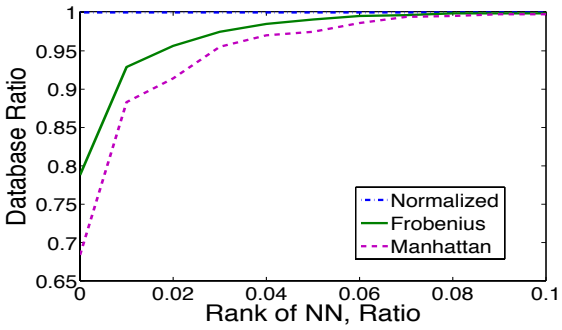


(a) Probability 0.6 (b) Probability 1.0

Figure 10: Comparison of the distance functions with respect to the success ratio of matching the noisy sequences to their originating.



(a) Probability 0.6, distortion 50%



(b) Probability 1.0, distortion 50%

Figure 11: Comparison of the cumulative histograms for the rank of nearest neighbour for each distance measure. Note that ranks are denoted as a ratio of the database size.

[7] G. R. Coulter. American Sign Language typology. *Doctoral Dissertation*, 1979.

[8] F. Giannotti, M. Nanni, and D. Pedreschi. Efficient mining of temporally annotated sequences. In *SDM*, 2006.

[9] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, 1994.

[10] D. J. Hand, H. Mannila, and P. Smyth. *Principles of Data Mining*. MIT Press, 2001.

[11] F. Höppner. Discovery of temporal patterns - learning rules about the qualitative behaviour of time series. In *Proc. of PKDD*, pages 192–203, 2001.

[12] F. Höppner and F. Klawonn. Finding informative rules in interval sequences. In *Advances in Intelligent Data Analysis, Proc. of the 4th International Symposium*, pages 123–132, 2001.

[13] S.-Y. Hwang, C.-P. Wei, and W.-S. Yang. Discovery of temporal patterns from process instances. *Computers in Industry*, 53(3):345–364, 2004.

[14] P. Kam and A. W. Fu. Discovering temporal patterns for interval-based events. In *DaWaK*, pages 317–326, 2000.

[15] S. Laxman, P. Sastry, and K. Unnikrishnan. Discovering frequent generalized episodes when events persist for different durations. *IEEE Transactions on Knowledge and Data Engineering*, 19(9):1188–1201, 2007.

[16] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics*, 10(8):707–710, 1966.

[17] S. K. Liddell. American Sign Language syntax. *The Hague: Mouton*, 1980.

[18] J.-L. Lin. Mining maximal frequent intervals. In *Proc. of SAC*, pages 624–629, 2003.

[19] H. Lu, J. Han, and L. Feng. Stock movement prediction and n-dimensional inter-transaction association rules. In *Proc. of the ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pages 12:1–12:7, 1998.

[20] C. Mooney and J. F. Roddick. Mining relationships between interacting episodes. In *Proc. of SDM*, 2004.

[21] F. Mörchen and D. Fradkin. Robust mining of time intervals with semi-interval partial order patterns. In *SDM*, pages 315–326, 2010.

[22] C. Neidle, J. Kegl, D. MacLaughlin, B. Dawn, and R. G. Lee. The syntax of American Sign Language: Functional categories and hierarchical structure. 2000.

[23] P. Papapetrou, G. Benson, and G. Kollios. Discovering frequent poly-regions in dna sequences. In *Proc. of the IEEE ICDM Workshop on Data Mining in Bioinformatics*, 2006.

[24] P. Papapetrou, G. Kollios, S. Sclaroff, and D. Gunopulos. Discovering frequent arrangements of temporal intervals. In *Proc. of IEEE ICDM*, pages 354–361, 2005.

[25] P. Papapetrou, G. Kollios, S. Sclaroff, and D. Gunopulos. Mining frequent arrangements of temporal intervals. *Knowledge and Information Systems (KAIS)*, pages 133–171, 2009.

[26] J. F. Roddick and M. Spiliopoulou. A survey of temporal knowledge discovery paradigms and methods. *IEEE Transactions of Knowledge and Data Engineering*, 14(4):750–767, July/August 2002.

[27] I. Tsoukatos and D. Gunopulos. Efficient mining of spatiotemporal patterns. In *Proc. of the SSTD*, pages 425–442, 2001.

[28] R. Villafane, K. A. Hua, D. Tran, and B. Maulik. Knowledge discovery from series of interval events. *Intelligent Information Systems*, 15(1):71–89, 2000.

[29] E. Winarko and J. F. Roddick. Armada - an algorithm for discovering richer relative temporal association rules from interval-based data. *Data Knowl. Eng.*, 63(1):76–90, 2007.

[30] S.-Y. Wu and Y.-L. Chen. Mining nonambiguous temporal patterns for interval-based events. *IEEE Transactions on Knowledge and Data Engineering*, 19(6):742–758, 2007.