Qualitative approximation to Dynamic Time Warping similarity between time series data

Blaž Strle, Martin Možina, Ivan Bratko
Faculty of Computer and Information Science
University of Ljubljana
Slovenia

Abstract

Dynamic time warping (DTW) is a method for calculating the similarity between two time series which can occur at different times or speeds. Although its effectiveness made it very popular in several disciplines, its time complexity of $O(N^2)$ makes it useful only for relatively short time series. In this paper, we propose a qualitative approximation Qualitative Dynamic Time Warping (QDTW) to DTW. QDTW reduces a time series length by transforming it to qualitative time series. DTW is later calculated between qualitative time series. As qualitative time series are normally much shorter than their corresponding numerical time series, time to compute their similarity is significantly reduced. Experimental results have shown improved running time of up to three orders of magnitude, while prediction accuracy only slightly decreased.

1. Introduction

Time series is a form of data that is present in virtually every scientific discipline and business application. It can be described as a sequence of observations, measured at successive times, spaced at (often uniform) time intervals. Dynamic Time Warping (DTW) (Sakoe and Chiba 1978) is a method for calculating the similarity between two time series which can occur at different times or speeds. Its ability to warp time axis and find optimal alignment between two time series has made it very popular. DTW has been used in several disciplines (Keogh and Pazzani 2001), such as: speech recognition, gesture recognition, data mining, robotics, manufacturing and medicine. In spite of its effectiveness, its time complexity of $O(N^2)$ makes it useful only for relatively short time series. This limitation can be overcome by reducing time series length. In qualitative modeling, numerical models can be seen as an abstraction of the real world and qualitative models are often viewed as a further abstraction of numerical models (Bratko 2000). In this abstraction, some quantitative information is abstracted away while keeping information that is relevant to the problem.

In this paper, we introduce a qualitative approximation Qualitative Dynamic Time Warping (QDTW) to DTW. QDTW reduces time series size by transforming it to a qualitative time series. As qualitative time series are usually much simpler and shorter than numerical time series, savings in running time are large.

The rest of this paper is structured as follows. Section 2 briefly reviews classic Dynamic Time Warping, including several techniques that make it more time efficient. In Section 3 we introduce and describe our modification to classic DTW. In Section 4, DTW and QDTW are experimentally evaluated on three domains and the results are discussed. Section 5 gives conclusions and future work.

2. Dynamic time warping

2.1 Dynamic Time Warping

In this section we briefly describe classic Dynamic Time Warping method. Dynamic Time Warping aligns two time series in the way some distance measure is minimized (usually Euclidean distance is used). Optimal alignment (minimum distance warp path) is obtained by allowing assignment of multiple successive values of one time series to a single value of the other time series and therefore DTW can also be calculated on time series of different lengths. Figure 1 shows examples of two time series and value alignment between them for Euclidean distance (left) and DTW similarity measure (right). Notice that the time series have similar shapes, but are not aligned in time. While Euclidean distance measure does not align time series, DTW does address the problem of time difference. By using DTW,

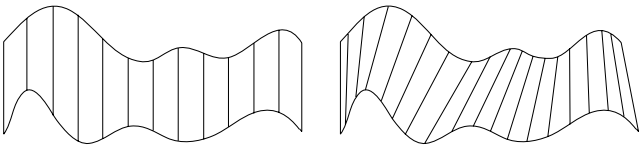

Figure 1: Example of two time series. Lines between time series show value alignment used by Euclidean distance (left) and Dynamic Time Warping similarity measure (right).

optimal alignment is found among several different warp paths. This can be easily represented if two time series $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_m)$, $a_i, b_j \in \mathbb{R}$ are arranged to form a n-by-m grid. Each grid point corresponds to an alignment between elements $a_i \in A$ and $b_j \in B$. A warp path $W = w_1, w_2, ..., w_k, ..., w_K$ is a sequence of grid points, where each w_k corresponds to a point $(i, j)_k$ - warp

path W maps elements of sequences A and B. A warp path is typically subject to several constraints:

- **Boundary conditions:** $w_1 = (1, 1)$ and $w_K = (n, m)$. This requires the warping path to start in first point of both sequences and end in last point of both sequences.
- **Continuity:** Let $w_k = (a, b)$ then $w_k - 1 = (a', b')$ where $a - a' \leq 1$ and $b - b' \leq 1$. This restricts the allowable steps in the warping path to adjacent cells.
- **Monotonicity:** Let $w_k = (a, b)$ then $w_k - 1 = (a', b')$ where $a - a' \geq 0$ and $b - b' \geq 0$. This forces the points in W to be monotonically spaced in time.

From all possible warp paths DTW finds the optimal one:

$$DTW(A, B) = min_W[\sum_{k=1}^{K} d(w_k)]$$

Here $d(w_k)$ is the distance between elements of time series.

Algorithm The goal of DTW is to find minimal distance warp path between two time series. Dynamic programming can be used for this task. Instead of solving the entire problem all at once, solutions to sub problems (sub-series) are found and used to repeatedly find the solution to a slightly larger problem. Let $DTW(A, B)$ be the distance of the optimal warp path between time series $A = (a_1, a_2, ..., a_n)$ and $B = (b_1, b_2, ..., b_m)$ and let $D(i, j) = DTW(A', B')$ be the distance of the optimal warp path between the prefixes of the time series A and B:

$$D(0, 0) = 0$$

$$A' = (a_1, a_2, ..., a_i), B' = (b_1, b_2, ..., b_j)$$

$$0 \leq i \leq n, 0 \leq j \leq m$$

Then $DTW(A, B)$ can be calculated using the following recursive equations:

$$D(0, 0) = 0$$

$$D(i, j) = min(D(i - 1, j), D(i, j - 1),$$
$$D(i - 1, j - 1)) + d(a_i, b_j)$$

Here $d(a_i, b_j)$ is the distance between two values of the two time series (usually Euclidean distance is used).

The most common way of calculating $DTW(A, B)$ is to construct a $n*m$ cost matrix M, where each cell corresponds to the distance of the minimal distance warp path between the prefixes of the time series A and B (Figure 2):

$$M(i, j) = D(i, j)$$

$$1 \leq i \leq n$$

$$1 \leq j \leq m$$

We start by calculating all the fields with small indexes and then progressively continue to calculate fields with higher indexes:

```
for i = 1...n
  for j = 1...m
    M(i,j) = min(M(i-1,j), M(i,j-1), M(i,j)) + dst(a_i,b_j)
```

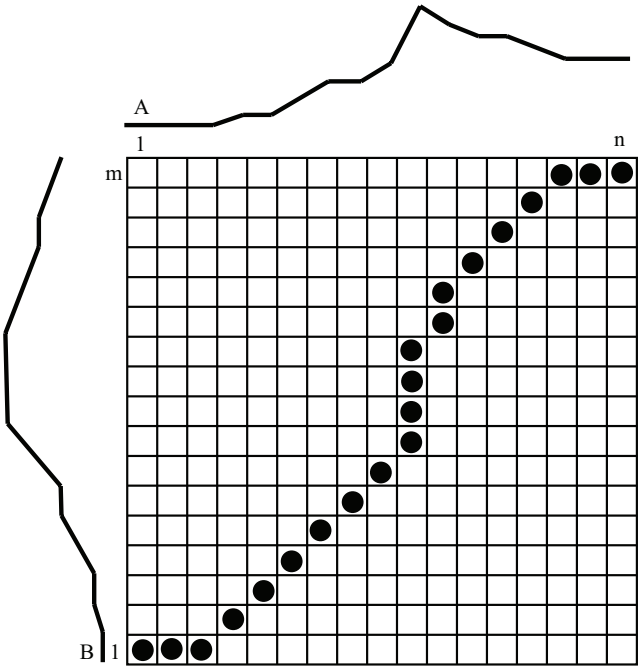

Figure 2: Minimal distance warp path between time series A and B.

The distance corresponding to the minimal distance warp path equals the value in the cell of a matrix M with the highest indexes $M(n, m)$. A minimal distance warp path can be obtained by following cells with the smallest values from $M(n, m)$ to $M(1, 1)$ (in Figure 2 the minimal distance warp path is marked with dots).

2.2 Improvements of Dynamic Time Warping

Although DTW's ability to find minimal distance warp path between time series makes it superior to simpler measures like Euclidean or Manhattan distance, its time complexity of $O(N^2)$ makes it useful only for relatively short time series. Many attempts to solve this issue have been proposed (Keogh and Pazzani 1999; Salvador and Chan 2007) which can be categorized as (Salvador and Chan 2007):

- constraints,
- data abstraction,

Constraints limit a minimum distance warp path search space by reducing allowed warp along time axis. Two most commonly used constraints are Sakoe-Chiba Band (Sakoe and Chiba 1978) and Itakura Parallelogram (Itakura 1975) which are shown in Figure 3.

Data abstraction speeds up the DTW algorithm by reducing the size of the input time series. Usually this technique speeds up DTW by a large constant factor for the price of a lower accuracy (Salvador and Chan 2007).

In this paper we are only interested in the data abstraction category. The data abstraction approach has already been used in (Keogh and Pazzani 1999) and (Salvador and Chan 2007). In (Salvador and Chan 2007), time series is reduced

several times and warp path found by DTW on lower resolution time series is used to calculate DTW on higher resolution time series.

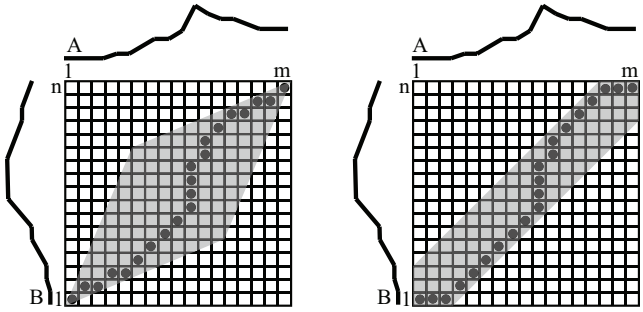

Figure 3: Itakura Parallelogram (left) and Sakoe-Chiba Band (right) constraints. Only shaded cells are used by DTW algorithm.

Data reduction is done by averaging adjacent pairs of points (data size is reduced by the factor of 2 every time resolution is decreased). In (Keogh and Pazzani 1999) a time series is approximated by a set of piecewise linear segments. The distance between segments is defined as the square of the distances of their means. Both of these approaches reduce time series size at the price of a lower accuracy. (Salvador and Chan 2007) compensate lower accuracy by calculating DTW several times on different resolution data, but data reduction part is still done at the price of information loss. Figure 4 shows a minimal distance warping path between sequences $(1, 2, 3, 4, 5)$ and $(5, 4, 3, 2, 1)$. Although they are very dissimilar, their mean values (shown as circles) are the same. This clearly shows drawbacks of data reduction by averaging, since the distance between these two segments would be 0.

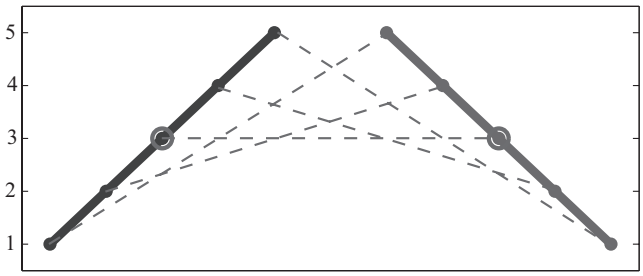

Figure 4: DTW between two time series. Circles represent mean time series value. Although the time series are not similar, their mean values are the same.

3. Qualitative Dynamic Time Warping (QDTW)

In our approach we would like to reduce time series size by removing information that is irrelevant for DTW. Our approach is based upon following theorem:

Theorem 1 *If two sequences A and B are qualitatively equal then*

$$DTW(A, B) \leq \varepsilon,$$

where

$$\varepsilon = \min(n * \text{maxdiff}(A)/2, m * \text{maxdiff}(B)/2).$$

Term $\text{maxdiff}(S)$ *is the maximal absolute difference between two adjacent elements in a time series S.*

We define two sequences to be qualitatively equal if both sequences are monotonic and their start and end values are equal. Figure 5 shows several examples of qualitatively equal sequences.

The theorem is based on the fact that in monotonic time series, the order in time (which a warp path has to respect) also corresponds to the order in the values. The theorem enables an approximation of $DTW(A, B)$ by qualitative DTW, described in the sequel. Suppose that time series A and B are samplings in time of two monotonic continuous functions of time. Then ε can be made arbitrarily small by increasing the density of sampling. Note that the sampling should be sufficiently dense w.r.t. the changes in the function value (not w.r.t. time). Consequently, if the "density approaches infinity" for any of the sequences A or B in Theorem 1, then $DTW(A, B)$ approaches 0.

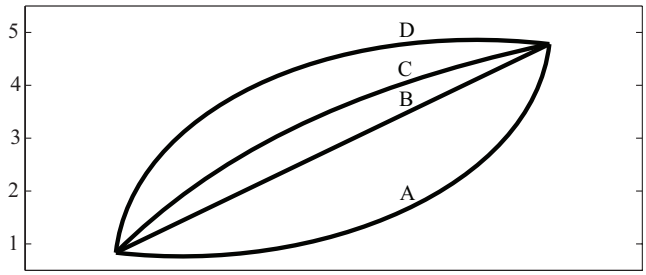

Figure 5: Four qualitatively equal sequences. DTW between any pair of them is 0.

QDTW transforms the original, numerical sequence to a qualitative sequence and then calculates DTW on the new sequence. Similar approach, where sequence is first transformed to a sequence of segments and their mean value is latter used to calculate DTW, was already proposed in (Keogh and Pazzani 1999). Main differences between approaches are in how segments are obtained and how this segments are latter used as input to the DTW. In our approach input sequences to the DTW consists of extreme points, that is the border points between the monotonic segments of the original curve (Figure 6). All monotonic segments are bound between two adjacent extreme points in the original sequence.

In our implementation, the program Qing (Žabkar et al. 2007) was used to extract the extreme points. Qing takes a sequence and a "persistence" parameter as input and returns a sequence of extreme points as output. Persistence parameter defines a minimal distance between extreme points (only extremes that differ more than persistence are returned).

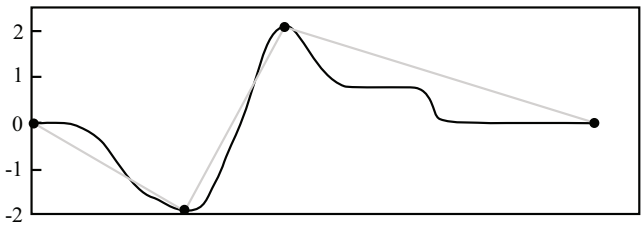

Figure 6: Example of a numerical sequence and its corresponding qualitative sequence where the black curve represents the original time series and the dots represent the extreme points - the border points between the three monotonic segments of the original curve: $(0, -2)$, $(-2, 2)$, $(2, 0)$. Sequence size is reduced from several points to only four points.

Consider two monotonic sequences $A = (a_1, a_2, ..., a_n)$, and $B = (b_1, b_2, ..., b_m)$. Then:

$$QDTW(A, B) = DTW((a_1, a_n), (b_1, b_m)),$$

where a_1, a_n, b_1, b_m are the extreme points. If $a_1 = b_1$ and $a_n = b_m$ then from the Theorem 1 following holds:

$$|QDTW(A, B) - DTW(A, B)| \leq \varepsilon.$$

When sequences are qualitatively equal, QDTW and DTW are almost equal (Theorem 1), otherwise problems can arise. There are two possible ways of violating the conditions for the applicability of Theorem 1:

- Extreme points do not coincide.
- Sequences are not monotonic.

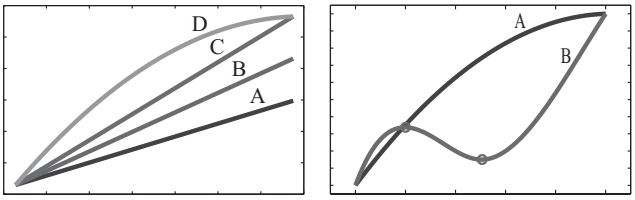

Figure 7: Possible violations of the conditions for the applicability of Theorem 1.

An example of monotonic sequences where the extreme points do not coincide is shown on the left side of Figure 7. It is obvious that DTW distance between base sequence A and any of the target sequences B, C, D is not necessarily the same as QDTW distance. More than in the actual values, we are interested in the distance order of target sequences B, C, D, when compared to base sequence A:

$$DTW(A, D) > DTW(A, C) > DTW(A, B),$$

$$QDTW(A, D) = QDTW(A, C) > QDTW(A, B).$$

When sequences with different extreme points (B, C) are compared to the base sequence (A), the order is preserved. In the case that target sequences have the same extreme points (C, D), QDTW cannot distinguish between them, when compared to the base sequence (A).

On the right hand side of Figure 7, a monotonic sequence is compared to a sequence that is not monotonic. If non monotonic part of sequence B (segment between two dots) is not detected (this can be due to high persistence parameter in the Qing algorithm), then both sequences have the same extreme points and $QDTW(A, B) = 0$, while $DTW(A, B) > 0$. On the other hand, if the decreasing part of sequence B is detected (small persistence), then sequence B is split into three segments by four extreme points. $QDTW(A, B)$ is calculated between the sequence of two extreme points from A and the sequence of four extreme points from B. As inner extreme points from B (bounding monotonically decreasing segment) have to map to extreme points from A, $QDTW(A, B)$ distance between A and B is quite large. With increasing number of short segments that map to one long segment, QDTW distance quickly increases. For now this represents the biggest problem of QDTW approach and should be solved in the future work.

Although, as we have shown, QDTW is not completely insensitive to information loss due to data reduction, we believe this will not significantly influence classification accuracy, and improved running time over DTW will more than compensate for slightly lower accuracy. The experimental evaluation that follows investigates this expectation.

4. Experimental evaluation

DTW is commonly used in time series classification domains. In these domains similarity or dissimilarity between time series determine whether time series belong to the same class or not. Therefore, similarity measure between time series is crucial part of the classification algorithm. Theorem 1 ensures that QDTW performs nearly the same as DTW if time series consist of qualitatively equal segments. This condition is rather strong. True applicability of QDTW can only be revealed with experimental evaluation on real world domains where conditions of Theorem 1 are not necessarily satisfied. With experimental evaluation, we would like to investigate how well QDTW performs in comparison to classic DTW in classification tasks. We are mostly interested in classification accuracy and execution time. The method was evaluated on three domains with different time series characteristics. Following data sets were used:

- **Australian Sign Language signs (High Quality) Data Set** (Kadous and Sammut 2002): The data set consists of the readings from 22 sensors that measure native signer hand position (11 sensors per hand) in time while signing one of 95 Auslan signs. For each Auslan sign 27 examples were recorded (total of 2565 examples). Due to DTWs high time complexity, only a subset of the original dataset was used. The subset consists of examples of the following ten signs: spend, lose, forget, innocent, Norway, happy, later, eat, cold, crazy.
- **Character Trajectories Data Set** (Asuncion and Newman 2007): The data set consists of 3-dimensional pen tip velocity trajectories which were recorded whilst writing individual characters. There are 20 different characters

in the data set. All of 2858 examples were captured by the same person using WACOM tablet. Due to the DTW time complexity only one seventh of the original examples were used (every seventh example from the original data set was included in the subset without changing the order of examples in the original dataset). All of the character labels (20) were included in the subset.

- **Character Recognition Data Set**: The data set consists of data from three sensors that measure the subject's hand acceleration while writing individual characters. There are 26 different characters in the data set. All of the 391 examples were obtained by the same person using tri-axis accelerometer.

4.1 Accuracy

In this section we are interested in how well QDTW performs in comparison to DTW and how different persistence settings effect classification accuracy. Classification was done using weighted k-nearest neighbor (k=3) algorithm using DTW or QDTW as similarity measure. The leave one out approach was used to estimate classification accuracy. QDTW method was evaluated using several relative persistence settings: 0.1, 0.2, 0.4 and 0.6. For each time series, persistence is obtained by multiplying relative persistence with the difference between time series maximum and minimum value.

As all the datasets consist of several variables (multivariate time series domains), any of these variables can be used for evaluation. Some of these variables are highly informative (similar examples belong to the same class while dissimilar examples belong to different classes) while others may not correlate with the class (random variables). On random variables, any similarity measure will behave similarly to a random similarity measure, so it makes sense to evaluate similarity measures only on highly informative variables. For this reason one variable, where DTW performs best, is used from each dataset to compare QDTW to DTW. These variables are: 'ryaw', 'y' and 'accY' from Australian Sign Language signs, Character Trajectories and Character Recognition datasets respectively. Classification accuracies using DTW and QDTW with different persistence settings are shown in Figure 8.

In comparison to DTW, QDTW (p=0.1) performed best on Australian Sign Language signs dataset where the difference between classification accuracies is only 0.01 (1.3%). QDTW performed worst on Character Recognition dataset where classification accuracy dropped by nearly 16% in comparison to DTW (from 0.88 for DTW to 0.74 for QDTW with persistence setting 0.1).

To evaluate how persistence affects classification accuracy, DTW and QDTW results for different relative persistence values (0.1, 0.2, 0.4 and 0.6) are ranked from best (1) to worst (5). For each dataset, average rank over all variables is calculated. Results are summarized in Table 1.

Table 1 confirms, as expected, that classification accuracy decreases with increasing relative persistence. The only domain where in some cases accuracy improved with increased relative persistence is Australian Sign Language domain.

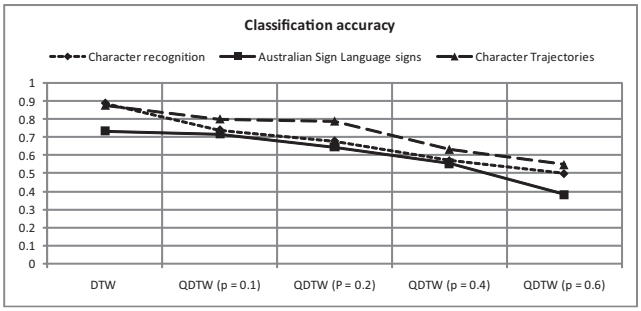

Figure 8: Classification accuracies for DTW and QDTW similarity measures where p denotes different relative persistence settings. Classification accuracies (shown from left to right) are for Australian Sign Language signs dataset: 0.73, 0.72, 0.64, 0.56, 0.38, for Character Trajectories dataset: 0.88, 0.80, 0.79, 0.63, 0.55 and for Character Recognition dataset: 0.88, 0.74, 0.68, 0.57, 0.50.

Table 1: Average rank for different relative persistence settings.

Method	Australian	Character Trajec.	Character Recog.
DTW	2.09	1	1
QDTW p=0.1	2.20	2	2
QDTW p=0.2	2.98	3	3
QDTW p=0.4	3.68	4.17	4
QDTW p=0.6	4.05	4.83	5

This can happen due to the presence of noise in some of its attributes, which can be removed only by more robust qualitative models. Overall, smaller relative persistence means larger classification accuracy in all evaluated datasets.

4.2 Efficiency

In this section we are interested in time efficiency of QDTW algorithm. Time efficiency is estimated with the number of distance calculations between two values of time series (size of the cost matrix M) which are needed for calculating DTW or QDTW similarity between two time series. Before calculating similarity, QDTW needs to transform time series to qualitative representation. As Qing is very efficient for qualitative modeling of time series, time to build qualitative models is insignificant in comparison to the time needed to calculate similarity and is thus omitted.

Time efficiency was estimated on all three datasets using variables 'ryaw', 'y' and 'accY' from Australian Sign Language signs, Character Trajectories and Character Recognition dataset respectively. For each dataset, similarity between all pairs of examples was calculated and average size of the cost matrix M ($M = m * n$, where m and n are time series lengths) is returned as a result. Figure 9 shows average size of the cost matrix M for calculating DTW and QDTW for all three domains.

From Figure 9, it is evident that QDTW was much faster than DTW on all three domains. Even for small persistence

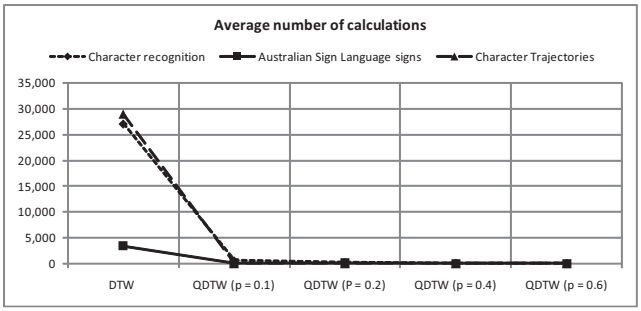

Figure 9: Average number of performed distance calculations between two values of time series when calculating similarity using DTW or QDTW where p is relative persistence setting. Average number of distance calculations (shown from left to right) are for Australian Sign Language signs dataset: 3382, 35, 25, 16, 9, for Character Trajectories dataset: 28893, 34, 30, 21, 13 and for Character Recognition dataset: 27058, 528, 182, 42, 19.

values, the savings in the number of distance calculations between two values of time series (size of the cost matrix M) are enormous (speed up by factor of nearly 100 on Australian Sign Language signs dataset , to nearly 850 on Character Trajectories dataset).

Besides comparison of QDTW to DTW, we are also interested in how different persistence settings effect time efficiency. Figure 10 shows average number of performed distance calculations between two values of time series for different relative persistence values. It can be seen from Figure

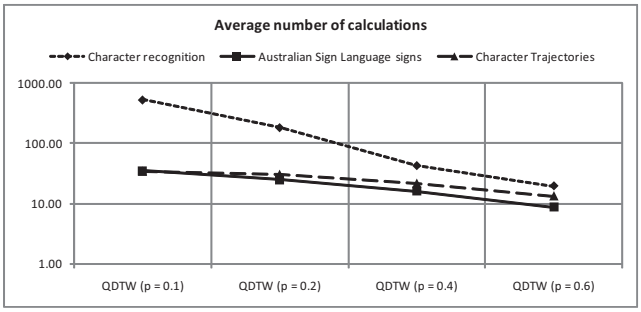

Figure 10: Average number of performed distance calculations between two values of time series (shown on logarithmic scale) when calculating QDTW similarity with different relative persistence settings (p).

10 that the average number of performed distance calculations between two values of time series is decreasing with higher persistence values. The results also show that similar persistence values on different domains do not necessarily mean similar savings in time. This means QDTWs performance is not only persistence dependent but also domain dependent.

5. Conclusions and future work

In this paper, we have stated a new theorem (Theorem 1), which explains when time series data can be reduced without loss of information relevant to DTW. Shortcomings of data reduction by averaging have been explained and new algorithm QDTW (Qualitative Dynamic Time Warping) have been introduced. QDTW is a modification of DTW algorithm, which is based on Theorem 1. It transforms time series data into qualitative series and thus significantly reduces data size. Experimental results have shown up to 1000 times speed-up with respect to the DTW algorithm. These significant improvements in efficiency are often obtained at acceptable loss in classification accuracy. QDTW major drawbacks are its inability to guarantee bounds on deviations from the optimal warp path solution, and its domain dependent efficiency. In future work, we will try to improve QDTW accuracy by reducing errors due to violations of the conditions for the applicability of Theorem 1. Special attention will be devoted to problems which arise due to non-monotonicity of segments, which is sometimes discovered by QING, while sometimes it is not. In these cases, we are comparing sequences with large number of short segments and sequences with small number of long segments, which usually results in a poor estimation of distance given by QDTW.

Acknowledgments

This work was partly funded by the X-Media project (www.x-media-project.org) sponsored by the European Commission as part of the Information Society Technologies (IST) programme under EC grant number IST-FP6-026978, and Slovene Agency for Research and Development (ARRS).

References

Asuncion, A., and Newman, D. 2007. UCI machine learning repository.

Bratko, I. 2000. *Prolog Programming for Artificial Intelligence*. Addison Wesley.

Itakura, F. 1975. Minimum prediction residual principle applied to speech recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 23(1):67–72.

Kadous, M. W., and Sammut, S. C. 2002. Temporal classification: Extending the classification paradigm to multivariate time series. Technical report, University of New South Wales, School of Computer Science and Engineering.

Keogh, E. J., and Pazzani, M. J. 1999. Scaling up dynamic time warping to massive dataset. In *PKDD '99: Proceedings of the Third European Conference on Principles of Data Mining and Knowledge Discovery*, 1–11. London, UK: Springer-Verlag.

Keogh, E. J., and Pazzani, M. J. 2001. Derivative dynamic time warping. In *In First SIAM International Conference on Data Mining (SDM2001*.

Sakoe, H., and Chiba, S. 1978. Dynamic programming algorithm optimization for spoken word recognition. *Acoustics, Speech and Signal Processing, IEEE Transactions on* 26(1):43–49.

Salvador, S., and Chan, P. 2007. Toward accurate dynamic time warping in linear time and space. *Intell. Data Anal.* 11(5):561–580.

Žabkar, J.; Jerše, G.; Mramor, N.; and Bratko, I. 2007. Induction of qualitative models using discrete morse theory. In *Proceedings of the 21st Workshop on Qualitative Reasoning*.