Point to Segment Distance DTW for Online Handwriting Signals Matching

Elmokhtar Mohamed Moussa1,2, Thibault Lelore1 and Harold Mouchère2

1 MyScript SAS, Nantes, France
2 Nantes Université, Ecole Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

Keywords: Handwriting Matching, DTW Metric, Sampling Invariance.

Abstract: In this paper, we propose DTWseg, a modified DTW algorithm based on a point-to-segment distance instead of the euclidean point-to-point distance. Applying DTWseg to online handwriting matching proves to be advantageous compared to other algorithms as it is less sensitive to differences between signals sampling rates occurring due to acquisition frequencies or handwriting speed. It eliminates the need for a commonly practiced resampling that omits an important dynamic part of the ductus. Experiments on IRONOFF french words dataset and FLOWCHARTS dataset show DTWseg to be least impacted by sampling rate alterations. We also propose a new benchmark of state-of-the-art methods on offline handwriting to online conversion based on our new proposed metric.

1 INTRODUCTION

Pen and paper have been used for hundreds of years by humans to record their activities, ideas, creations... The digitization and processing of documents have changed their usage: historians can access ancient material easily, and companies created Electronic Document Management Systems to improve their process. These images of documents are called offline documents. In recent years, with the emergence of digital tablets and touch screens, new usages appear and new types of documents are created with handwritten digital content: online documents. Online documents and offline documents share the same downstream processing task: document classification, document segmentation, handwriting recognition, and writer identification, etc: but often with different approaches due to the different nature of their respective input source. On the one hand, offline content is stored as matrices of pixels and on the other hand, online documents are recorded as the pen trajectory on a surface tablet represented as a time series of x and y coordinates, in addition to other motion measures (pen pressure, velocity, etc.).

This work focuses on the comparison of handwritten samples of the online domain. It is useful for many applications including handwriting trajectory reconstruction from IMU-enhanced pen (Wehbi et al., 2022) where synchronization between the pen and recording tablet surface is challenging due to the difference in sampling rates and acquisition start and end. Other applications include signature verification (Sharma and Sundaram, 2018), template matching for content clustering, keyword (or shape) spotting (Szoke et al., 2005), and due to recent advances in research topics such as online handwriting synthesis (Graves, 2014), offline handwriting to online conversion (Kato and Yasuhara, 2000), the need for pertinent online handwriting quality evaluation metrics of the generated online handwriting is becoming more acute. Matching of two online signals is commonly done using linear interpolation alignment which deteriorates important temporal and spatial dynamics. DTW (Sakoe and Chiba, 1978) algorithm provide an elastic alignment capable of matching signals of different lengths. However, in our handwriting-specific case, it can portray a very negative matching for handwriting with similar directions and spatial arrangements as in Figure 1. We propose a new cost function based on the segment-to-point distance to compute DTW. It has the advantage to minimize the impact of the sampling rate. Our modified DTW is presented in Section 3. In Section 4, we experiment and compare our metric to classic DTW on different datasets and state-of-the-art offline handwriting to online conversion approaches.

850

Mohamed Moussa, E., Lelore, T. and Mouchère, H.
Point to Segment Distance DTW for Online Handwriting Signals Matching.
DOI: 10.5220/0011672600003411
In Proceedings of the 12th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2023), pages 850-855
ISBN: 978-989-758-626-2; ISSN: 2184-4313
Copyright © 2023 by SCITEPRESS – Science and Technology Publications, Lda. Under CC license (CC BY-NC-ND 4.0)
2 RELATED WORKS

DTW has been used in a wide range of document analysis applications. Here we focus on recent handwriting analysis subtopics: online handwriting synthesis and offline to online conversion. Other notable applications such as isolated character recognition (Bahlmann et al., 2002) can also be cited.

2.1 Online Handwriting Synthesis

This aims to generate the handwriting of a given text with neural networks (Graves, 2014). The generation is usually conditioned on a given writer and outputs convincing handwriting regarding the writing style, making online handwritten documents easily editable (Aksan et al., 2018). Matching the neural network predictions to ground truth online signals is essential to train such deep learning approaches. The MSE loss is ubiquitous in state-of-the-art methods, implying that the neural network has also to learn the exact correct sampling rate of the resampled ground truth signal, creating small misleading artifacts in many cases. Defining a loss function with more meaningful feedback on the general ductus rather than sampling rate artifact can help further improve the generative models. Following that line, (Ji and Chen, 2020) employed a CNN-LSTM discriminator combined with Graves Generator (Graves, 2014) in an adversarial training framework to obtain more realistic handwriting.

2.2 Offline Conversion to Online Handwriting

Given a static offline handwritten document, the goal here is to recover the temporal information of the pen trajectory and thus an online document. This could be used mainly for two reasons: use the existing online tool as a recognizer; or allow the user to edit his content as a vectorized image (instead of a flat image). The retrieved trajectory should be as faithful as possible to the writer’s offline document. Such systems (Chan, 2020) are often evaluated using word error rate when intended for offline recognition. WER presents a useful insight into the semantic coherence of online reconstruction. However it presents a major drawback, depending on the complexity of the recognition systems, the correct word can be recognized even if online reconstruction is unfaithful e.g. slanting and rotation don’t usually affect recognizers. Root Mean Square Error (RMSE) and Dynamic Time Warping algorithm (DTW) (Sakoe and Chiba, 1978) are commonly proposed as evaluation metrics.
(Hassaïne et al., 2013; Phan et al., 2015; Dinh et al., 2016; Archibald et al., 2021; Mohamed Moussa et al., 2021; Diaz et al., 2022) of the reconstructed online signal w.r.t. to the ground-truth online signal. RMSE (cf. equation 1) is a one-to-one mapping that measures the distance between two temporal signals \( x_i \) and \( \hat{x}_j \) of lengths \( N \) and \( M \) respectively. If \( N = M \) and the signals are well aligned (same frequency and in phase) RMSE is a straightforward measure of the distance between them. Nevertheless, in many cases where the signals are not perfectly aligned (stretched or compressed at different time windows, out of phase etc.) the pairing becomes far less obvious.

\[
\text{RMSE} = \sqrt{\frac{1}{N} \sum_{j=1}^{N} (x_j - \hat{x}_j)^2} \quad (1)
\]

3 PROPOSED METRIC

We first present the classical DTW and then our proposed DTW-seg.

3.1 DTW

DTW algorithm computes the optimal alignment between two signals of different lengths. It allows for elastic one-to-many matching. A cumulative cost \( C_{N \times M} \) matrix is constructed using a L2 distance function \( f(x_i, \hat{x}_j) = \| x_i - \hat{x}_j \|^2 \). DTW algorithm finds a warping path \( w = \{w_p = (i, j) \in [N \times M]\}_{p=1}^P \) minimizing equation 2 and satisfying the following constraints:

- boundaries: \( w_1 = (1, 1) \wedge w_P = (N, M) \)
- monotonicity: let \( w_p = (i, j) \) and \( w_{p+1} = (i', j') \) then \( i \leq i' \wedge j \leq j' \)
- continuity: \( i' \leq i+1 \wedge j' \leq j+1 \)

This optimization is solved by recursively updating the cumulative cost matrix with equation 3.

\[
\begin{align*}
\text{DTW}(x, \hat{x}) &= \min_{w \in W} \left\{ \sum_{p=1}^{P} f(w_p) \right\} \\
&= f(w_p) = f(x_i, \hat{x}_j) \\
C_{ij} &= f(x_i, \hat{x}_j) + \min \left\{ C_{i,j-1}, C_{i-1,j-1}, C_{i-1,j} \right\} \\
\end{align*}
\quad (2)
\]

DTW distance is defined as the cumulative cost of \( w \) equal to \( C_{NM} \) normalized by the warping path length. Figure 3 shows an example of DTW algorithm output. In this work, we focus on defining a sampling rate invariant cost function \( f \).

3.2 DTW-seg

Let \( x_i \) a point and \([\hat{x}_j, \hat{x}_{j+1}]\) a segment between two consecutive elements, we define a point to segment cost function, as illustrated in Figure 3, by:

\[
\begin{align*}
\bar{a} &= \hat{x}_j \hat{x}_{j+1}, \quad \bar{b} = \hat{x}_j x_i, \quad \bar{c} = \hat{x}_{j+1} x_i \\
g(x_i, [\hat{x}_j, \hat{x}_{j+1}]) &= \begin{cases} 
\bar{a} \cdot \bar{b} < 0, & f(x_i, \hat{x}_j) \\
\bar{a} \cdot \bar{c} < 0, & f(x_i, \hat{x}_{j+1}) \quad (4) \\
\text{else,} & \|\text{proj}_\bar{b}\bar{c}\|^2
\end{cases}
\end{align*}
\]

Replacing \( f \) by \( g \) as a cost function in equations 2 and 3 we define \( DTW_{seg} \) which minimizes the alignment cost changes w.r.t. variation in the sampling rate. The special case that needs to be mentioned, as illustrated by figure 4, is that of a point distance to a segment between a stroke end and the next stroke start. In this case, the segment is considered invalid and thus omitted.
4 EXPERIMENTS

To demonstrate the sensitivity to resampling of RMSE, DTW and our DTW$_{seg}$ metric, we experiment with different spatial resampling strategies:

- Equidistant linear resampling with distance $d$;
- Simple moving average (SMA) with previous 2 points:

$$x'_i = \frac{x_i + x_{i-1}}{2}$$

We used the validation set of IRONOFF (Viard-Gaudin et al., 1999) containing 19,888 words and the FLOWCHARTS (Awal et al., 2011) validation set with 172 flowcharts. Table 1 shows that DTW$_{seg}$ is relatively small after oversampling or moving average transformations compared with classic DTW. The aforementioned transformations degrade the spatial information of the signals the least compared to subsampling yet the reported DTW and RMSE are high. DTW is observed to be the highest when subsampling IRONOFF with $d = 10$ in comparison, DTW$_{seg}$ is one and a half folds smaller. For FLOWCHARTS when subsampling with $d = 15$, DTW$_{seg}$ is 3 folds smaller than DTW.

In addition, we use our metric to benchmark state-of-the-art offline handwriting to online conversion approaches, namely, (Chan, 2020), (Diaz et al., 2022) and (Archibald et al., 2021). Using their public official implementations. We also include an internal rule-based method based on a smoothness criterion. Table 2 shows the results of their evaluations on the validation set of IRONOFF dataset. Synthetic offline images, with a stroke width randomly chosen between one and three pixels, are rendered from the ground truth online. We observe that the three metrics rank the different approaches in the same manner. All of the previously mentioned approaches predict oversampled online signals therefore they have a bigger DTW alignment cost compared to DTW$_{seg}$. In fact, (Chan, 2020) approach is ranked second, closely trailing behind our Internal approach. It is to be noted that (Archibald et al., 2021) is based on a data-driven CNN-LSTM trained only on English IAM (Martí and Bunke, 2002) dataset. A finetuning on the training set of IRONOFF could have helped the network to adapt to unseen french words, yielding better results. No meta-parameters tuning for (Diaz et al., 2022) ap-
proach was performed for a fairer comparison. Figure 5 illustrates inference results for different SoTA approaches. Figure 5a shows that (Archibald et al., 2021) prediction is overall the best in this particular instance. In fact, figures 5c and 5b tend to oversimplify small loops, the latter is also missing a portion of the last small ending. Since all of the mentioned methods infer oversampled signals, $DTW_{seg}$ is shown to be the metric that best evaluates the inherent signal directions and spatial arrangement with minimal regard to sampling frequencies.

5 DISCUSSION

This work focuses on improving the matching of similar online handwriting signals with different sampling frequencies. This variability occurs when recording simultaneously on multiple devices or due to the natural variance in human writing velocity. Another challenging extension, which is out of the scope of this pa-

per, is the invariance to stroke direction inversion (e.g. crossing a t with a left-to-right or right-to-left stroke) and stroke permutation (e.g. letter x in Figure 2). In fact, DTW’s strict continuity constraints make it such that those small handwriting preferences are assigned a very important alignment cost which can hinder the performance of downstream tasks. (Archibald et al., 2021) employs a DTW loss function that finds the permutation of consecutive pairs of strokes and stroke direction that minimizes the alignment cost. This approach does not deal with longer-range permutations such as crossing or dotting. (Li et al., 2013) proposed a more complete multi-stroke DTW based on the A* star algorithm to overcome the combinatorial explosion of alignment hypothesis. However, it is still difficult to upscale to the word level and beyond.
6 CONCLUSIONS

In this paper, we presented DTWseg, a modified DTW algorithm based on a segment-to-point cost function dedicated to online handwriting matching. We showed that classical matching approaches such as RMSE and DTW distance overstate the sampling rate’s importance. DTWseg, on the other hand, matches more closely signals differing in sampling rates. We also benchmark SoTA for offline to online conversion with DTWseg. In future work, we will study the definition of a loss function (Cuturi and Blondel, 2018) based on DTWseg to train a neural network for the offline to online conversion online task. We hypothesize that DTWseg provides more meaningful information as its gradient pushes the network’s predictions to be closer to the signal as a whole rather than a single point from the signal.

REFERENCES


