SIGNAL DENOISING USING PRINCIPAL CURVES: APPLICATION TO TIME WARPING

Umut Ozertem, Deniz Erdogmus

Computer Science and Electrical Engineering Department
Oregon Health and Science University, Portland, OR 97239, USA
ozertemu@csee.ogi.edu, derdogmus@ieee.org

ABSTRACT

One of the most important problems with current time warping algorithms in the literature is sensitivity to noise. To improve the noise robustness of the current algorithms, we propose a denoising step based on the likelihood maximization of the pairwise signals. This approach is independent of the selection of the particular time warping algorithm, and can be coupled with any algorithm in the literature. Improvement in noise robustness not only brings increased robustness to current time warping applications, but also may trigger new application areas where the signals that need to be compared are buried in noise.

Index Terms— signal denoising, principal curves, time warping, nonlinear filtering

1. INTRODUCTION

Time warping finds use in many fields of time series analysis, mostly in biomedical and speech processing applications. A common problem in time series analysis is that although some signals show similar characteristics, their structure does not align in time axis. Dynamic time warping (DTW) is the first technique that aims to solve this problem [1]. DTW is very sensitive to noise, and presence of noise may lead to *singularities*.

Time warping literature is rich in terms of publications focusing on the noise robustness issue. Techniques include windowing of the time series signals to reduce the high frequency content, assuming that most of the high frequency content is noise [2, 3, 4]. This idea is based on constraining the space of allowable warpings. They are not guaranteed to converge, and they may prevent the optimal solution from being found.

It is probably safe to claim that almost all time warping algorithms are derivatives of original DTW algorithm. More recent approaches in the literature include derivative dynamic time warping (DDTW) [5], that uses the derivative of the signals rather than the original values, enhanced dynamic time warping (EDTW) [6], that brings a unifying view to DTW and hidden Markov models, and context dependent dynamic time warping (CDDTW) [7], that exploits application specific contextual characteristics of the signals to improve performance.

Principal curves are defined by Hastie [8, 9] as "*self-consistent finite length smooth curves passing through the middle of data*." The literature on principal curves has a variety of algorithm propositions, but there is not much work on the theoretical aspects. We recently proposed another definition for principal curves, which describes the principal curve in terms of the gradient and the Hessian of the data probability density [10], and we will use this definition throughout the paper.

We propose a principal curve based denoising scheme for the time warping algorithms. We rewrite the problem in a different feature space and propose using principal curve projections to implement a nonlinear nonparametric data driven denoising filter as a preprocessing step. The resulting preprocessing filter is nonparametric and employs kernel density estimation (KDE).

2. PRINCIPAL CURVES AND SIGNAL DENOISING

In this section, we will discuss how to utilize principal curves as a denoising filter. The feature in which the principal curve will be determined is described and principal curve projection methodology is derived. We define the principal curve as follows *"a point in the data feature space is on the principal curve if and only if the gradient is an eigenvector of the Hessian at this point and the remaining eigenvectors have negative eigenvalues"* [10]. The details of the principal curve definition and its properties are omitted due to restricted space. In summary, this definition generalizes the concept of local maximum to local ridge and the principal curve is defined as the local maximum likelihood ridge of the data pdf in the feature space.

In most template matching applications, the observed signals are generally compared with a noiseless template. Throughout the paper, we will consider the case of a noisy signal pair that represents a more realistic scenario; a noiseless template may not be available in all applications. For the applications that a noiseless template is available, one of the

This work is partially supported by NSF grants ECS-0524835, and ECS-0622239.

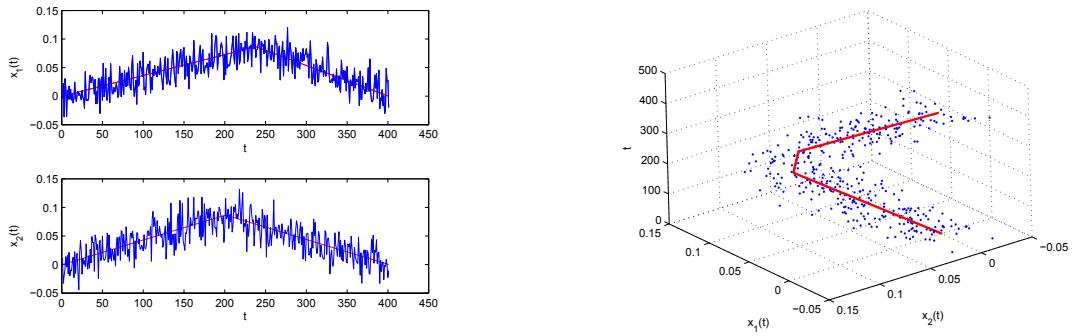

Fig. 1. Pair of noisy (SNR = 2dB) and noiseless warped signals in time-domain (left) and in feature-**r** space (right).

noise terms can be dropped. The observed signals are

$$\begin{array}{rcll} x_1(t) & = & s(f(t)) + n_1(t), & t \in \{t_1, \ldots, t_N\} \\ x_2(t) & = & s(t) + n_2(t), & t \in \{t_1, \ldots, t_N\} \end{array} \quad (1)$$

where $f(t)$ is the sought time warping function and $n_1(t)$ and $n_2(t)$ are additive noise. For simplicity and without loss of generality, we assume that the discrete-time samples of the signals $x_1(t)$ and $x_2(t)$ have the same length and are sampled at the same rate - a resampling procedure using well-established methods can precede the proposed technique if this assumption is invalid. We build the feature vector of the data as

$$\mathbf{r}_i = \begin{bmatrix} x_1(t_i) \\ x_2(t_i) \\ t_i \end{bmatrix}, \quad t \in \{t_1, \ldots, t_N\} \quad (2)$$

Figure 1 shows the structure of **r** for some realizations of noisy signal pairs $x_1(t)$, and $x_2(t)$. Figure 1a shows the time series signals along with their noiseless counterparts $s(t)$, and $s(f(t))$. For illustrative purposes, here we used a simple piecewise linear signal. The structure of the noiseless signal pairs (red) is also shown in Figure 1b. Note that, for unimodal additive noise signals $n_1(t)$, and $n_2(t)$, the data structure in **r** space clearly shows the pairwise signal characteristics with a perturbation around a predominant shape. This observation becomes clearer considering the structure of the noiseless signals in the **r** domain (red). We propose to use the principal curve projections of the data samples to approximate the noiseless signal characteristics.

To find the principal curve of the data in **r**, one can directly use our earlier proposition [10]. But here we have an easier problem, where only the samples of the principal curve at time indices $t_1 < \ldots < t_N$ are sufficient. Constraining the time axis, namely the third dimension, one can write a subspace likelihood maximization algorithm to find the principal curve. At the peak of the pdf on any constrained space $t = t_0$, the gradient is parallel with one of the eigenvectors of the Hessian; hence, the point is on the principal curve. We use KDE to estimate the density, which is

$$p(\mathbf{r}) = N^{-1} \sum_{i=1}^{N} K_\Sigma(\mathbf{r} - \mathbf{r}_i) \quad (3)$$

where $K_\Sigma(\cdot)$ is typically a Gaussian kernel function with covariance Σ. While variable-full-covariance KDE would prove more outlier robust and accurate, the computational complexity trade-off needs to be considered. For the rest of the derivations in the paper, we assume that fixed-circular (isotropic) Gaussian kernels are used. in many cases Taking the derivative of (3) with respect to **r**, and equating it to zero, one obtains

$$\mathbf{r} \sum_{i=1}^{N} G_{\sigma^2}(\mathbf{r} - \mathbf{r}_i) - \sum_{i=1}^{N} \mathbf{r}_i \, G_{\sigma^2}(\mathbf{r} - \mathbf{r}_i) = \mathbf{0} \quad (4)$$

Solving for **r** yields the well-known mean shift update, which is a fixed-point, EM-type likelihood maximization rule: [12]

$$\mathbf{r} \leftarrow \frac{\sum_{i=1}^{N} \mathbf{r}_i \, G_{\sigma^2}(\mathbf{r} - \mathbf{r}_i)}{\sum_{i=1}^{N} G_{\sigma^2}(\mathbf{r} - \mathbf{r}_i)} \quad (5)$$

To constrain the iterations to the specific time-instant of interest, the update should be projected back onto the $t = t_0$ plane.

$$\mathbf{r} \leftarrow \mathbf{A} \frac{\sum_{i=1}^{N} \mathbf{r}_i \, G_{\sigma^2}(\mathbf{r} - \mathbf{r}_i)}{\sum_{i=1}^{N} G_{\sigma^2}(\mathbf{r} - \mathbf{r}_i)} + \mathbf{b} \quad (6)$$

The projection matrix **A**, and translation vector **b** are

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ t_0 \end{bmatrix} \quad (7)$$

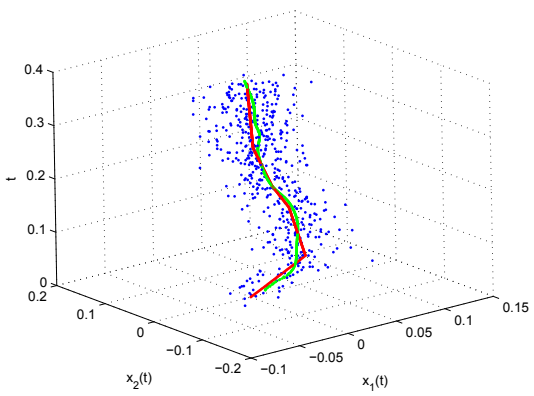

Fig. 2. The noisy (blue) and noiseless (red) versions of two realizations of $r_{synthetic}(t)$ in the $\mathbf{r}$ domain, along with the principal curve of the data (green).

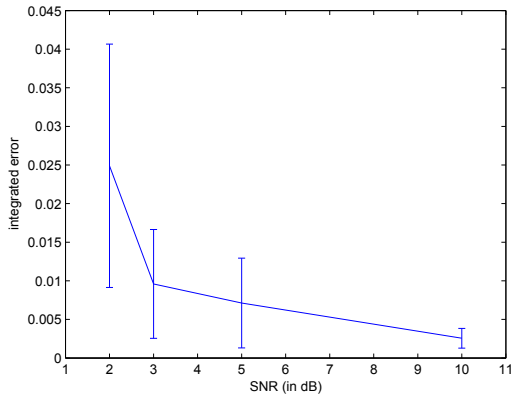

Fig. 3. Mean and standard deviation of the approximation error between estimated and true warping functions over 100 random realizations of the data.

where t_0 is the time index of the initial sample $\mathbf{r}_i$. Iterating (6) until convergence (achieved when the gradient of (3) becomes an eigenvalue of its Hessian at the current point, a condition checked by monitoring the angle between the gradient and its multiplication with the Hessian), one can project any sample $\mathbf{r}_i$ onto principal curve, obtaining $\widetilde{\mathbf{r}}_i$.

$$\widetilde{\mathbf{r}}_i \approx \begin{bmatrix} s(f_{12}(t_i)) \\ s(t_i) \\ t_i \end{bmatrix}, \quad t \in \{t_1, \ldots, t_N\} \tag{8}$$

Principal curve iterations given in (6) has a complexity of $O(N)$ per sample per iteration. This complexity could be reduced by truncating the Gaussian kernels in the KDE, thus iterating each trajectory based on the nearby data neighbors.

An important point here is the selection of the bandwidth of the Gaussian kernel. The selection of kernel function is a well studied topic, and literature on kernel density estimation and kernel machines is rich in techniques that extend from local neigborhood distances based heuristic approaches to maximum likelihood based principled methods [11]. Here we will use a leave-one-out cross validation maximum likelihood approach to select the kernel bandwidth. The specific selections for the experiments will be given in the experimental results section.

After principal curve projections, any time warping algorithm can be employed to $\widetilde{\mathbf{r}}$. Since our aim is to introduce the principal curve denoising concept, rather than optimizing implementation details for a particular application, for simplicity, here we use the original DTW algorithm for the following demonstrations [1].

3. EXPERIMENTAL RESULTS

To be able to control the amount of noise, and provide results at different noise levels, we prefer to present results on synthetic signals. Consider the following piecewise linear signal

$$r_{synthetic}(t) = \begin{cases} \frac{t}{t_1} & : 0 \leq t \leq t_1 \\ \frac{1-t}{1-t_1} & : t_1 \leq t \leq 1 \end{cases} \tag{9}$$

where t_1 is uniformly distributed between 0.1 and 0.9. We generate realizations of this random signal and add Gaussian noise of different powers to obtain 10dB, 5dB, 3dB, and 2dB SNR-levels. For a 2dB signal, in Figure 2 we present the structure of the feature space along with the noiseless signal (red) and the approximation provided by principal curve (green). Obviously, as the noise level decreases, the accuracy of the principal curve approximation gets better. In Figure 3, we present the results of 100 Monte Carlo simulations for signals of different noise levels. We evaluate the integrated error between the noiseless signal structure in $\mathbf{r}$ domain, and the principal curve. Mean and ± 2 standard deviances of the error is given for 10dB, 5dB, 3dB, and 2dB.

To evaluate the effects of the denoising on the final results, for a particular realization of $r_{synthetic}(t)$, we compare the correct time warping function with time warping functions of the noisy signals and the principal curve denoising results. Figure 4a shows the results regarding to noisy signals - again for SNR levels of 10dB, 5dB, 3dB, and 2dB - along with the time warping function of the noiseless signal. Figure 4b shows the same for the same signals employing the denoising step. Clearly, using the principal curve projections instead of the data samples provides an improvement in noise robustness.

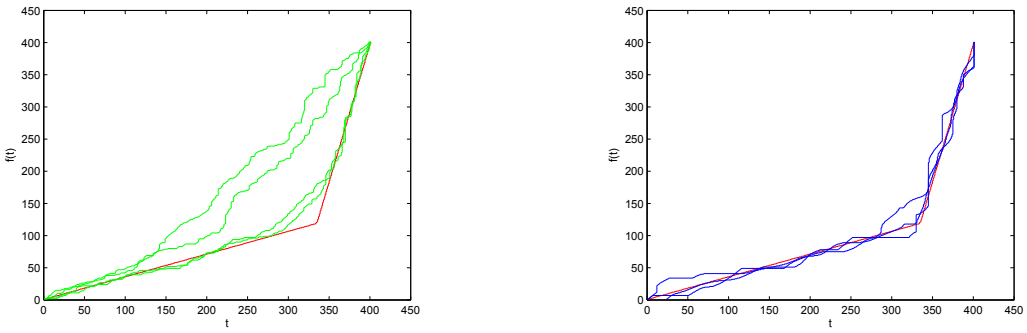

Fig. 4. The comparison of time warping function estimates for noisy data(left) and the principal curve denoised data (right) overlaid on the true piecewise linear time warping function.

4. DISCUSSION

One of the most important problems with time warping algorithms in the literature is sensitivity to noise. We propose a nonlinear signal denoising technique, which is purely nonparametric. The denoising principle is independent of the input features or the time warping algorithm used.

The aim of the paper is not to improve the results of any specific state-of-the-art time warping application, but to propose a domain independent nonparametric preprocessing tool to improve the noise robustness. Therefore, here we simply test our system with synthetic test signals in which we can control the noise power. For simplicity, we demonstrated time warping results using the signals themselves as the input features and the original DTW algorithm to find the time warping function; however spectograms or any other feature can also be utilized; one can couple any time warping algorithm with the proposed technique.

5. REFERENCES

[1] H. Sakoe and S. Chiba, "A Dynamic Programming Approach to Continuous Speech Recognition," in *Proceedings of International Congress on Acoustics*, Budapest, Hungary, 1971.

[2] L. Rabnier, A. Rosenberg, and S. Levinson, "Consideration in Dynamic TIme Warping Algorithms for Discrete Word Recognition," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. ASSP-26, pp. 575-582, 1978.

[3] D. Berndt, and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," *Workshop on Knowledge Discovery in Databases*, 1994.

[4] C. Myers, L. Rabnier, and A. Rosenberg, "Performance Tradeoffs in Dynamic Time Warping Algorithms Isolated Word Recognition," *IEEE Transactions on Acoustics, Speech and Signal Proccessing*, vol. ASSP-28, pp. 623-635, 1980.

[5] E. J. Keogh, and M. J. Pazzani, "Derivative Dynamic Time Warping," *in First SIAM International Conference on Data Mining*, 2001.

[6] E. Yaniv, and D. Burshtein, "An Enhanced Dynamic Time Warping Model for Iproved Estimation of DTW Parameters," *IEEE Transactions on Speech and Audio Processing*, vol 11., no. 3, 2003.

[7] A. Pikrakis, S. Theodoridis, and D. Kamarotos, "Recognition of Isolated Musical Patterns Using Context Dependent Dynamic Time Warping," *IEEE Transactions on Speech and Audio Processing*, vol 11, no.3, 2003.

[8] T. Hastie, "Principal Curves and Surfaces," Ph.D. Thesis, Stanford Univ., 1984.

[9] T. Hastie, W. Stuetzle, "Principal Curves," *Jour. Am. Statistical Assoc.*, vol. 84, pp. 502-516, 1989.

[10] D. Erdogmus, U. Ozertem, "Self-Consistent Locally Defined Principal Surfaces," *International Conference on Acoustics, Speech and Signal Processing*, vol. 2, pp. II549-II552, 2007.

[11] R. O. Duda , P. E. Hart , and D. G. Stork, "Pattern Classification (2nd Edition)", Wiley, 2000.

[12] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790-799, 1995.