JMLR: Workshop and Conference Proceedings 25:379–394, 2012 Asian Conference on Machine Learning

Improved sequence classification using adaptive segmental sequence alignment

Shahriar Shariat SSHARIAT@CS.RUTGERS.EDU
Vladimir Pavlovic VLADIMIR@CS.RUTGERS.EDU
CS Department, Rutgers University, Piscataway, NJ 08854

Editor: Steven C.H. Hoi and Wray Buntine

Abstract

Traditional pairwise sequence alignment is based on matching individual samples from two sequences, under time monotonicity constraints. However, in some instances matching two segments of points may be preferred and can result in increased noise robustness. This paper presents an approach to segmental sequence alignment based on adaptive pairwise segmentation. We introduce a distance metric between segments based on average pairwise distances, which addresses deficiencies of prior approaches. We then present a modified pair-HMM that incorporates the proposed distance metric and use it to devise an efficient algorithm to jointly segment and align the two sequences. Our results demonstrate that this new measure of sequence similarity can lead to improved classification performance, while being resilient to noise, on a variety of problems, from EEG to motion sequence classification.

1. Introduction

The task of asserting pairwise sequence similarities is central to many problems in machine learning. A family of alignment algorithms accomplishes this by measuring similarities between pairs of samples across two sequences and matching them under monotonicity (i.e., temporal ordering) constraints. Dynamic time warping (DTW) Berndt and Clifford (1994) is a common computational technique to tackle this problem. DTW and its variations have shown great results in many applications Ding et al. (2008).

DTW alignment algorithms are based on pairing of individual data points. That is, a sample at time t_i in sequence X is typically matched with only one other sample at time t_j in sequence Y. In many practical applications it may be more desirable (or robust) to establish pairing between groups of points: matching a temporal segment $X_k = [x_i, \ldots, x_{i+m}]$ to another segment of the contrasting sequence, $Y_l = [y_j, \ldots, y_{j+n}]$. For instance, one might be interested in not only calculating the distance but also retrieving locally similar segments of the contrasting sequences. In some instances considering groups of points instead of single samples and comparing their statistics is more robust to noise. Furthermore, in the case of non-causal time-series where local ordering of samples can change, such as in EEG recordings de Munck et al. (2007) or signals with general random time delays Blaum and Bruck (1994), one must employ a method that considers different permutations of the samples within a short period of time. In those cases point-to-point matching may yield suboptimal alignments.

In Shariat and Pavlovic (2011) the authors propose an approach, based on canonical correlation analysis (CCA), to handle the segmental alignment. They formulate an objective (IsoCCA) con-

strained properly to impose the time monotonicity. Although the results show strong resilience to noise, the objective does not provide a proper metric between the segments. This can cause the resulting segments to be unnecessarily short. Furthermore, the non-convexity of IsoCCA objective makes it increasingly sensitive to initial segmentation and model parameter choices. In another recent work, Ryoo (2011), the author proposes to find the best matching segments of the two sequences based on a probabilistic model. However, the algorithm does not handle gaps/insertions and, hence, does not consider a complete alignment model. Moreover, the author suggests empirically fixing all segment lengths, with the approach lacking clear means to handle data-driven segments. In practice, however, variable and data-adapted segments result in more robust alignments.

In this paper we propose a segmental alignment framework based on a probabilistic model and investigate its properties and robustness against noise in the context of sequence classification. The new contributions of this work are:

- We suggest a distance metric based on average pair-wise distances suitable for measuring similarity between two segments , aimed at segmental sequence alignment.
- Based on the proposed distance metric we develop our probabilistic alignment model by extending the pair-HMM formalism.

Through extensive experiments we show that the proposed method can lead to improved classification results on benchmark sequence classification tasks, classification of non-causal EEG signals, and recognition of activities from human motion data. This proposed approach is particularly resilient to the presence of noise where other similar approaches fail.

The paper is organized as follows: in Sec. 2 we discuss the metric property of IsoCCA and construct our segmental metric. In the following two sections (3, 4) the proposed model is discussed in detail and some of its properties are highlighted. In section 5 experimental results are presented. Section 6 concludes the paper with the discussion of our findings and some suggestions for future work.

2. Segment Matching Metric

In Shariat and Pavlovic (2011) the authors propose a segmental alignment method based on CCA, *i.e.* IsoCCA. Despite promising results, the proposed framework does not provide a proper metric between the segments. The reason for that lies in the fact that IsoCCA works by effectively finding the closet points of the convex hulls of the two segments of points. This results in a non-metric because the triangular inequality does not hold (cf. Fig. 1). Moreover in the case of overlapping convex hulls, their distance is zero even though the size of the common area can be very small resulting in unnecessarily small segments.

In some applications, as illustrated in Sec. 1, one is interested in matching unordered small segments of points. This naturally leads to matching two unordered sets of points where permutation is not a matter of concern. In addition to insensitivity to permutation, we seek to find a distance metric that suppresses the noise and is efficient to compute. Many distance metrics have been proposed to measure the distance between sets, c.f., Woznica et al. (2006). Often the proposed distances are based on non-linear functions (Hausdorff, for instance), which are computationally intensive. Moreover, Hausdorff-type distances can be highly insensitive to the content of the contrasting sets,

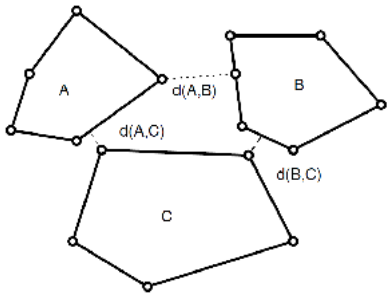


Figure 1: Taking closest distance of two convex hulls as a metric results in violating the triangular property

focusing instead on the boundary cases. Also their distance surface is rather flat giving very similar values for different input sets. Kernel distances Kondor (2003) are also not suitable when the set of points is small and therefore, in practice the estimated distribution is inaccurate. In the following we propose a distance based on average pair-wise distances.

Formally, for $\mathcal{X}$ and $\mathcal{Y}$, two sets of points, we define

$$d(\mathcal{X},\mathcal{Y}) = \frac{1}{|\mathcal{X}||\mathcal{Y}|}\sum_{x_i\in\mathcal{X}}\sum_{y_j\in\mathcal{Y}}\|x_i - y_j\|_n \tag{1}$$

where $\|.\|_n$ is a convex norm between two points. It is trivial to show $d(\mathcal{X},\mathcal{Y}) \geq 0$ and $d(\mathcal{X},\mathcal{Y}) = d(\mathcal{Y},\mathcal{X})$. It is also straightforward to prove that (1) has the triangular property given the convexity of the norms. Equation (1) needs to be slightly modified to have definiteness property (i.e $d(x,y) = 0 \iff x = y$):

$$\mathcal{D}(\mathcal{X},\mathcal{Y}) = \frac{1}{|\mathcal{X}\cup\mathcal{Y}|}\left(\frac{1}{|\mathcal{X}|}\sum_{x_i\in\mathcal{X}}\sum_{y_i\in(\mathcal{Y}-\mathcal{X})}\|x_i - y_j\|_n + \frac{1}{|\mathcal{Y}|}\sum_{x_i\in(\mathcal{X}-\mathcal{Y})}\sum_{y_i\in\mathcal{Y}}\|x_i - y_j\|_n\right). \tag{2}$$

Equation (2) is symmetric, non-negative and definite due to empty sums in case of equality of $\mathcal{X}$ and $\mathcal{Y}$. To prove that (2) has triangular property, one can partition $(\mathcal{D}(\mathcal{X},\mathcal{Y}) + \mathcal{D}(\mathcal{Y},\mathcal{Z}) - \mathcal{D}(\mathcal{X},\mathcal{Z})) \geq 0$ into disjoint sets and observe that given triangular property of (1), the required inequality holds for 2. Note that in case of $\mathcal{X}\cap\mathcal{Y} = \emptyset$, (2) reduces to (1).We will show in the experimental results that even though the ordering of samples is not preserved within a short segment when modelled as a set, the proposed metric can be used for general purpose alignment. The metric also exhibits invariance to arbitrary temporal permutations. This can be beneficial for non-causal sequences that arise from random delays (e.g., EEG). However, it can also be desirable in video retrieval settings when, for instance, the direction of an activity is not a concern. In Sec. 4 we demonstrate why this metric is resilient to impulse noise when incorporated into an alignment algorithm and also how it can be computed efficiently.

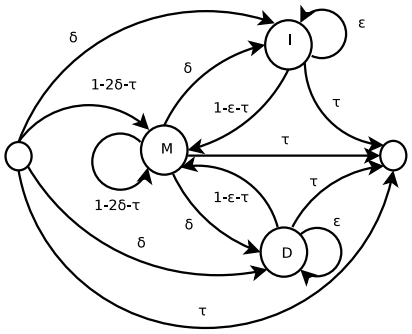


Figure 2: Segmental Pair-HMM diagram

3. Segmental Pair-HMM (SPHMM)

The Pair HMM, introduced by Durbin et al. (1997), can be seen as a probabilistic model defined on pairs of sequences (X, Y) that aims to describe their joint likelihood, $P(X, Y|alignment)$. As shown in Figure 2, PHMM has three states: M for match, I for insertion and D for deletion. Given two sequences of observations X and Y with n and m samples, the match state emits a pair of samples (x, y) $x \in X$, $y \in Y$. Insertion and deletion states emit $(x, -)$ and $(-, y)$ respectively where $-$ stands for a gap. This model implements an affine gap penalty which is more general than constant gap penalty usually used in DTW.

In the following we add the notion of segmentation to pair-HMM formalism. To define the segmentation structure consider a sequence X, an ordered finite set of samples. We define $\mathbf{s} : \mathbf{s}(X) = (s_1, s_2, \ldots s_{L_X})$ to be a mapping of the indexes of X to segment indexes. For example $\mathbf{s}(X)$ maps the sequence $(x_1 \ldots x_n)$ into segments $\big((x_1 \ldots x_{i_1}), (x_{i_1+1} \ldots x_{i_2}) \ldots (x_{i_{L_x-1}+1} \ldots x_{i_{L_x}})\big)$ where $i_{L_X} = n$. That is $s_1 : (1, 2, \ldots i_1) \rightarrow 1$, $s_{L_X} : (i_{L_X-1} + 1, i_{L_X-1} + 2, \ldots, i_{L_X}) \rightarrow L_X$ and in general $s_k : (i_{k-1} + 1, i_{k-1} + 2, \ldots, i_k) \rightarrow k$.Likewise, we define $\mathbf{s}(Y)$ for Y. From this point forward with $\mathbf{s}$ we represent the mapping of indexes of both sequences X and Y (i.e. $\mathbf{s} \equiv (\mathbf{s}(X), \mathbf{s}(Y))$). Therefore based on $\mathbf{s}$ we have

$$X = (\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_{L_X}) \tag{3}$$
$$Y = (\mathrm{Y}_1, \mathrm{Y}_2, \ldots, \mathrm{Y}_{L_Y}) \tag{4}$$

where $\mathrm{X}_k = \{x_i : i \in s_k^{-1}(X)\}$ where $s_k^{-1}(X)$ denotes the sample points in segment X_k and likewise for Y_k.

Given the segments defined by $\mathbf{s}$, the segmental alignment can be defined as the sequence of states (segment correspondences), q_t where $t = (t_x, t_y)$ is a tuple indexing the sequences' segment. For instance $t = (i, j)$ points to segment X_i of X and similarly X_j segment of Y. Therefore, an optimal alignment is $Q^* = q_{(1,1)} \cdots q_{(L_X, L_Y)}$, that maximizes the joint likelihood $P(X, Y|Q, \lambda, \mathbf{s})$ where λ are the model parameters. That is,

$$Q^* = \arg\max_Q P(X, Y|Q, \lambda, \mathbf{s}) \tag{5}$$

where

$$P(X,Y|Q,\lambda,\mathbf{s}) = \left(\prod_{t=(1,1)}^{p((L_X,L_Y))} \prod_{t' \in r(t)} b_{q_t}(\mathrm{X}_{t_1},\mathrm{Y}_{t_2}) a_{q_t q_{t'}} \right) b_{q_{(L_X,L_Y)}}(\mathrm{X}_{L_X},\mathrm{Y}_{L_Y}) \tag{6}$$

where $p(.)$ gives the preceding tuple of segment indexes while $r(.)$ is the successor operator. The prior on Q in (6), can be uniform or can encode traditional band-priors such as the Sakoe-Chiba band. Here $b_{q_t}(.)$ is the emission probability of emitting a pair of segments (X_{t_1}, Y_{t_2}) in state q_t and $a_{q_t q_{t'}}$ is the transition probability from state q_t to the state emitting the next pair. Basically, given segmentation $\mathbf{s}$, we simply treat the alignment problem as that of aligning sequences of segments. Then for instance the meaning of b_{q_t} changes from the likelihood of pairs of samples to the likelihood of pairs of segments. Equations (5-6) show that the optimal alignment is the Viterbi path for observing segmented and matched (X, Y) given the hidden Markov model in Fig. 2.

To define the probability $b_{q_t}(\mathrm{X}_{t_1},\mathrm{Y}_{t_2})$ we need to consider three cases, depending on the type of correspondence (M, I, or D). If the correspondence is of type M, then we can define

$$b_{q_t}(\mathrm{X}_{t_1},Y_{t_2}) \quad = exp(-\mathcal{D}(\mathrm{X}_{t_1},\mathrm{Y}_{t_2})) \cdot \Psi(|\mathrm{X}_{t_1}|,|\mathrm{Y}_{t_2}|) = \theta_{\mathrm{X}_{t_1},\mathrm{Y}_{t_2}}. \tag{7}$$

$\mathcal{D}(\mathrm{X}_{t_1},\mathrm{Y}_{t_2})$ is the distance between two segments, defined by (2). Ψ specifies the distribution of the corresponding segment lengths which can be learned from the data or assumed to have a certain distribution. If the correspondence is of type I we then define

$$b_{q_t}(\mathrm{X}_{t_1},-) = Pr(\mathrm{X}_{t_1},I) = exp(-\sigma_g|\mathrm{X}_{t_1}|) = \zeta_{\mathrm{X}_{t_1},-}, \tag{8}$$

In the case of D we can similarly define

$$b_{q_t}(-,\mathrm{Y}_{t_2}) = Pr(D,\mathrm{Y}_{t_2}) = exp(-\sigma_g|\mathrm{Y}_{t_2}|) = \zeta_{-,\mathrm{Y}_{t_2}} \tag{9}$$

where σ_g is scaling parameter. Given the observation likelihoods, it is possible to extend Durbin et al. (1997) non-segmental alignment Viterbi algorithm to the segmental model. To find the optimal segmentation $\mathbf{s}$ one can search over permissible segment lengths at each step of recursion in the Viterbi algorithm. This is equivalent to optimizing

$$Q^*, \mathbf{s}^* = \arg\max_{Q,\mathbf{s}} P(X,Y|Q,\lambda,\mathbf{s}), \tag{10}$$

which is our ultimate objective. To make the procedure computationally tractable one may impose a maximum constrain on the segment length.

3.1. Marginal matching likelihood

Let us define $\mathbf{S}$ to be the set of all possible segmentations of two sequences X and Y with m and n samples, respectively. Also assume that Π is the set of all segmental alignments between X and Y. Using the forward algorithm one can estimate the following

$$P(X,Y|\lambda) = \sum_{\mathbf{s}\in\mathbf{S}} \sum_{Q\in\Pi} P(X,Y|Q,\lambda,\mathbf{s}) P(\mathbf{s}) P(Q) \tag{11}$$

Computing the above is not tractable for every possible segmentation. We will assume $P(Q)$ and $P(\mathbf{s})$ to be uniform. Therefore, we approximate the joint probability of X and Y at each step

based on partial alignments and segmentations. The likelihood of matching two segments given all previous segmentations can be formulated as

$$P\left\{x_{1\ldots i}, y_{1\ldots j}|q_t = M, \lambda, \left(S^*(x_{1\ldots(i-k)}), S^*(y_{1\ldots(j-l)})\right)\right\} = \theta_{x_{(i-k)\ldots(i-1)}, y_{(j-l)\ldots(j-1)}} \max_{\mathbf{s}'\in\left(\mathbf{S}(x_{1\ldots(i-k)}), \mathbf{S}(y_{1\ldots(j-l)})\right)} \sum_{Q'\in\Pi_{(i-k),(j-l)}} P\left\{x_{1..(i-k)}, y_{1..(j-l)}|Q', \lambda, \mathbf{s}'\right\} \tag{12}$$

where

$$\left(S^*\left(x_{1\ldots i}\right), S^*\left(\left(y_{1\ldots j}\right)\right)\right) = \arg\max_{\mathbf{s}'\in(\mathbf{S}(x_{1\ldots i}), \mathbf{S}(y_{1\ldots j}))} \sum_{Q'\in\Pi_{i,j}} P(x_{1..i}, y_{1..j}|Q', \lambda, \mathbf{s}'). \tag{13}$$

In (12) and (13), x_i (y_i) denotes a sample in the sequence. Also, k and l are permissible segment lengths for X and Y. $\mathbf{S}(.)$ is the set of all segmentations while $S^*(.)$ denotes the approximated segmentation of the given input sequence. $\Pi_{i,j}$ is the set of all possible alignments of X and Y up to x_i and y_j. The first term on the right hand side of (12) is the likelihood of matching two segments the same as (7), while the second term finds the maximum marginalized likelihood over aligning partial sequences given all possible segmentations up to x_{i-k}, y_{j-l}. The same formulation can be defined for other states as well. The result of applying this recursive algorithm is the likelihood of all permissible segmentations for every pair of samples of contrasting sequences.

4. Discussion

In this section we discuss how the distance metric defined by (2) suppresses noise and affects the segment size as well as analyze the time complexity of the alignment algorithm.

4.1. Segment Size and Noise Suppression

Consider two discretely sampled continuous multivariate sequences, X and Y, that are to be sent through a noisy channel. In the source, both sequences are segmented and each segment is approximated by a line then the obtained lines are re-sampled and transmitted through the channel. To observe the mechanism of noise suppression based on the proposed distance we consider aligning of the two signals in the destination while an impulse is added to the one the sequences during transmission due to some interference. Formally, let X_k and Y_l be two of the line segments at the same time index where $x_{k,i} = \beta t_i + \xi$, $t_i = [1 \ldots .|\mathrm{X}_k|]$, $y_{l,j} = \beta t_j$, and $t_j = [1 \ldots |\mathrm{Y}_l|]$. Suppose an impulse corrupts X_k at $t = t_c$ ($1 \leq t_c \leq \min(|\mathrm{X}_k|, |\mathrm{Y}_l|)$) such that $x_{k,t_c} = y_{l,t_c} + \xi + \alpha$ (Fig.3) Assuming $\mathrm{X}_k \cap \mathrm{Y}_l = \emptyset$ (which is very probable given that the two sequences are continuous) the distance of two segments will be smaller than a point-to-point match only if the following inequality holds

$$\begin{aligned}\mathcal{D}(\mathrm{X}_k, \mathrm{Y}_l) &= \frac{1}{|\mathrm{X}_k||\mathrm{Y}_l|}\sum_{i=1}^{|\mathrm{X}_k|}\sum_{j=1}^{|\mathrm{Y}_l|}\|x_{k,i} - y_{l,j}\| \\ &\leq \frac{|\beta|}{|\mathrm{X}_k||\mathrm{Y}_l|}\left(\sum_{i=1}^{|\mathrm{X}_k|}\sum_{j=1}^{|\mathrm{Y}_l|}\|y_{l,i} - y_{l,j}\|\right) + \frac{(|\mathrm{X}_k| - 1)}{|\mathrm{X}_k|}|\xi| + \frac{1}{|\mathrm{X}_k|}|\xi + \alpha| < |\xi + \alpha|. \end{aligned} \tag{14}$$

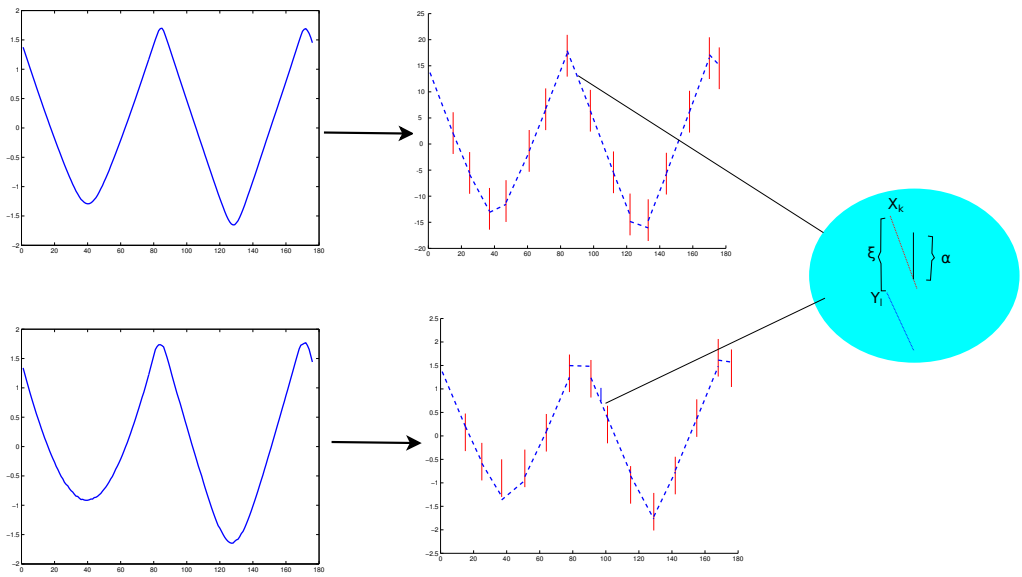


Figure 3: Piecewise linear approximation of a sequence based on fixed segments. The right plot shows the segments and the approximated lines (dashed lines). Two of the segments that are to be matched are magnified.

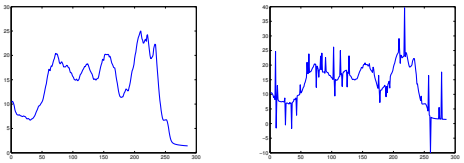
Note that since $\mathrm{X}_k \cap \mathrm{Y}_l = \emptyset$, the original distance described by (2) is reduced to (1). We used the convexity of the norm in the above. Therefore,

$$\sum_{i=1}^{|\mathrm{X}_k|}\sum_{j=1}^{|\mathrm{Y}_l|} \|y_{l,i} - y_{l,j}\| < \frac{|\mathrm{Y}_l|(|\mathrm{X}_k| - 1)\,[|\alpha + \xi| - |\xi|]}{|\beta|} \tag{15}$$

has to hold. One can observe that as long as $\alpha < -|\xi| - \xi$ or $\alpha > |\xi| - \xi$, by increasing $|\mathrm{X}_k|$ (or $|\mathrm{Y}_l|$) while the slope of the line (β) is kept constant, the left hand side of (15) grows quadratically while the right hand side grows linearly which leads to bounded segment length. Furthermore, if $|\beta| \to 0$ then as long as $|\mathrm{X}_k| > 1$, (15) is a tautology meaning that longer segment length is always favourable. Consequently, As β increases a point-to-point match becomes more likely. The result of such distance metric is that it flattens the signal around an impulse not only according to its neighbourhood but also to the contrasting sequence. This leads to a dynamic noise removal. Therefore, if the impulse is in fact a characteristic of the signal and not a noise, it will not be removed (similar to DTW) but in case of noisy impulse, it will be averaged and flattened.

4.2. Complexity

The time complexity of (10) is dependent both on the lengths of segments in each sequence and the length of the sequences themselves. Given that the number of states is fixed and small, one can prove that the time complexity of the Viterbi (or forward) algorithm is $O(l_1 l_2 mn)$ where l_1 and l_2 are the maximum segment lengths for each sequence and n and m are the lengths of the sequences. To compute the distance between two segments, one can employ the summed area table technique



(*a*) Original (*b*) $\omega = 100\%$

Figure 4: Sample of a sequence from UCR dataset (Coffee) with different levels of noise.

Crow (1984) to improve the performance. That is, the pairwise distances of all pairs of samples are pre-calculated and the summed area table is constructed. Then within the matching procedure only a few additions are required to compute the distance. Usually, l_1 and l_2 are not too long relative to the sequence lengths. Thus, the overall time complexity is typically a constant factor away from that of the regular DTW.

5. Experimental Results

In this section we show the utility of SPHMM through extensive experiments. We first examine our proposed approach the benchmark data. We use the first dataset (data1) from the UC Riverside "time-series classification page" (http://www.cs.ucr.edu/~eamonn/time_series_data/). To show that our method is able to deal with non-causal time-series we also apply it to a publicly available EEG data set. Finally, we show that SPHMM can improve the performance of activity recognition classification on a subset of CMU MoCap data.

The parameters of SPHMM are determined empirically, as described below. Euclidean distance is used as the measure of distance between two samples. We observed that L_1 norm can slightly, but not significantly, improve the results in case of excessive noise but we do not include those results. Throughout this section l_1 and l_2 denote the maximum allowed lengths of the segments. We have also assumed the scaling parameter of gap operations (equations (8) and (9)) to be $\sigma_g = 1$. In all experiments the classifier is the 1-Nearest Neighbour (1-NN), with the similarity measure defined by one of the three methods (two traditional, DTW or PHMM, and SPHMM) .

5.1. Benchmark Data

In order to compare our proposed approach to DTW and demonstrate the applicability of our method to general sequences, we tested SPHMM on the first subset of time-series from the UC Riverside time-series repository that contains 20 datasets. To be able to test the noise resilience of SPHMM, we have added impulse noise to all sequences. Additive noise process is Gaussian $N(0, \omega\sigma_i)$ where σ_i is the standard deviation of feature i and ω is the power degree of the noise. We have added the noise to time points chosen uniformly at random such that the noise does not cover more than 20% of the sequence duration (Fig. 4).

We conducted the experiment on original data and one noisy version of data with $\omega = 100\%$. For every sequence, we have sampled from the noise process three times independently and added the resulting samples to the corresponding time-series. Different similarity measures are then applied to each noisy version of the data and the resultant recognition accuracies are averaged over different noisy versions of each dataset and reported. The results are shown in Tab.1, We compared

Table 1: UCR time-series classification accuracy.

	Original			$\omega = 100\%$		
	DTW	PHMM	SPHMM	DTW	PHMM	SPHMM
Lighting7	72.60	75.34	**79.45**	43.51	53.97	**73.74**
OSULeaf	58.26	65.70	**66.12**	47.11	55.15	**68.18**
OliveOil	**86.67**	**86.67**	**86.67**	28.89	28.89	**32.22**
SwedishLeaf	79.68	80.64	**85.28**	27.84	46.43	**57.81**
Trace	**100.00**	**100.00**	**100.00**	73.67	75.83	**88.67**
Two Patterns	**100.00**	**100.00**	**100.00**	88.22	89.86	**99.96**
fish	83.43	**86.86**	**86.86**	26.28	57.62	**68.19**
synthetic control	**99.33**	96.67	97.33	92.78	92.89	**93.22**
wafer	98.00	99.76	**99.79**	84.01	89.60	**99.39**
yoga	83.80	**84.00**	**84.00**	65.10	67.97	**78.20**
50words	75.16	80.00	**80.44**	57.21	74.12	**77.87**
Adiac	60.36	**60.87**	**60.87**	7.08	12.79	**37.02**
Beef	50.00	**53.33**	**53.33**	40.00	50.00	**53.33**
CBF	99.44	**99.89**	**99.89**	74.37	85.93	**98.00**
Coffee	**82.14**	78.57	**82.14**	57.14	63.22	**76.78**
ECG200	80.00	**91.00**	**91.00**	77.00	81.00	**85.00**
FaceAll	**79.94**	77.51	79.41	67.89	69.05	**77.20**
FaceFour	82.95	89.77	**90.91**	52.65	69.78	**89.77**
Gun Point	90.67	**98.00**	**98.00**	69.33	73.78	**83.55**
Lighting2	88.52	**90.16**	**90.16**	61.97	76.89	**86.89**
Average	82.55	84.74	85.58	57.1	65.74	76.25

Table 2: Statistics of match operations for UCR time-series. μ indicates the mean, σ stands for standard deviation of segments' length and $\max$ shows the largest extracted segment.

	Original			$\omega = 100\%$		
	μ	σ	max	μ	σ	max
Lighting7	1.02	0.21	5	1.09	0.42	5
OSULeaf	1.00	0.09	5	1.13	0.56	5
OliveOil	1.00	0.00	1	1.00	0.01	2
SwedishLeaf	1.01	0.15	5	1.05	0.27	5
Trace	1.00	0.06	4	1.05	0.27	5
Two Patterns	1.05	0.29	5	1.10	0.39	5
fish	1.00	0.00	1	1.03	0.18	5
synthetic control	1.07	0.31	5	1.12	0.38	5
wafer	1.01	0.11	5	1.05	0.28	5
yoga	1.00	0.02	5	1.03	0.18	5
50words	1.00	0.07	5	1.09	0.43	5
Adiac	1.00	0.01	3	1.03	0.21	5
Beef	1.00	0.00	1	1.01	0.23	5
CBF	1.03	0.21	5	1.13	0.44	5
Coffee	1.60	1.54	5	2.68	1.16	5
ECG200	1.02	0.18	5	1.06	0.30	5
FaceAll	1.04	0.26	5	1.09	0.36	5
FaceFour	1.06	0.38	5	1.10	0.44	5
Gun Point	1.00	0.01	2	1.05	0.26	5
Lighting2	1.02	0.19	5	1.09	0.43	5

the proposed approach to DTW and pair-HMM (where no segmentation is applied) with the warping band. The warping band for all methods is set to $\rho = 15\%$ of the length of the sequence. The parameters for SPHMM and PHMM are set to $\delta = 0.4$ or $\delta = 0.1, \epsilon = .1, \tau = 0.01$ and $l_1 = l_2 = 5$. For OliveOil, SwedishLeaf, Synthetic control, Beef, ECG200, FaceAll, FaceFour, Adiac and Wafer, $\delta = 0.1$ resulted in better performance while $\delta = 0.4$ showed a better result on the remaining datasets in the training phase. The parameters are not changed for noisy data experiments.

One can see in Tab.1 that PHMM is superior to DTW in 17 cases and SPHMM is superior or on par with PHMM in all cases and superior to DTW in 18 cases in the original, noise-free setting. However, as soon as the noise is introduced, SPHMM shows significantly better performance compared to both DTW and PHMM even though PHMM still outperforms DTW.

Table 2 shows some statistics of alignments for the original and noisy sequences resulted from the experiment. It shows the average (μ), standard deviation (σ) and the maximum ($\max$) length of segments for match operations for all alignments. It could be beneficial to increase the maximum segment length as in many instances segments of length 5 have been extracted. However, the results still show significant improvement while the computational time is tractable with a maximum segment length as small as 5. We do not show the statistics for gap operations but generally their mean segment length reduces as the noise level elevates since they are replaced by match operations. One can observe that the average length of match usually increases as the noise level elevates. Also note that the average segment length is close to 1, i.e. the traditional PHMM sample-to-sample matching. This is not unexpected as the chosen data does not result from the random delay processes.

Table 3: Recognition rates for EEG dataset. The first row shows the maximum segment length. For each maximum segment length the mean accuracy and standard deviation over different folds are reported.

	$l = 1$		$l = 5$		$l = 10$		$l = 20$	
	Acc	St.dev	Acc	St.dev	Acc	St.dev	Acc	St.dev
SPHMM	74.7	2.61	75.1	2.97	78.47	2.35	82.64	1.35
DTW	74.4	1.78	N/A					
CTW	75.52	1.01	N/A					

Table 4: Accuracy results for different fixed segmentations

$l_{fixed} = 5$		$l_{fixed} = 10$		$l_{fixed} = 20$		$l_{fixed} = 30$	
Acc	St.dev	Acc	St.dev	Acc	St.dev	Acc	St.dev
70.62	2.14	72.79	2.16	73.89	2.62	72.64	2.17

On the other hand, and due to noise (inherent or artificial), it is advantageous to have intermittently extended segments, as evident from the second column of Tab. 2.

It is interesting to note that SPHMM's recognition rates are better than the best reported recognition rates for DTW in UC Riverside "time-series classification page" in 17 cases. In those experiments DTW is finely tuned with additional optimal band selection.

5.2. EEG Signal Classification

We next applied our adaptive segmental alignment model to EEG signals to show its effectiveness in case of non-causal and noisy time-series. We used the P300 dataset described in Hoffmann et al. (2005). Each subject is exposed to 6 different images, one of which is the target image. Dataset consists of 9 subjects. Four session are held for each subject. In each session six runs are conducted such that the set of all 6 images is shown at least 20 times to each subject where one of the images is the target in each run. We chose subject 1 and target 2 for our experiment. In each fold of cross-validation we keep one session as training and the remaining three are used as the test set such that every session is used as training once. 1-NN is used as the classifier. We applied the default pre-processing on the data except that we increased the sub-sampling rate to 128 from 32 to acquire longer signals (129 samples). As recommended, we only kept 8 channels. We have compared SPHMM against DTW and CTW Zhou and de la Torre (2009). The spatial embedding included by CTW is a reasonable choice for aligning EEG signals. We have applied SPHMM with different maximum lengths to demonstrate that the longer segments and permutation invariance of the distance metric can result in improved recognition rates.

The results are shown in Tab. 3. As expected the accuracy does not show significant improvement over DTW for maximum segment lengths of 5. However, for longer segments SPHMM becomes significantly more accurate. Optimal performance of DTW was achieved without a warping band. To assess the effects of adaptive segmentation and alignment we also tested against sequences pre-segmented into fixed length segments. The results are shown in Tab. 4. Adaptive segmentation remains advantageous especially for longer segment lengths.

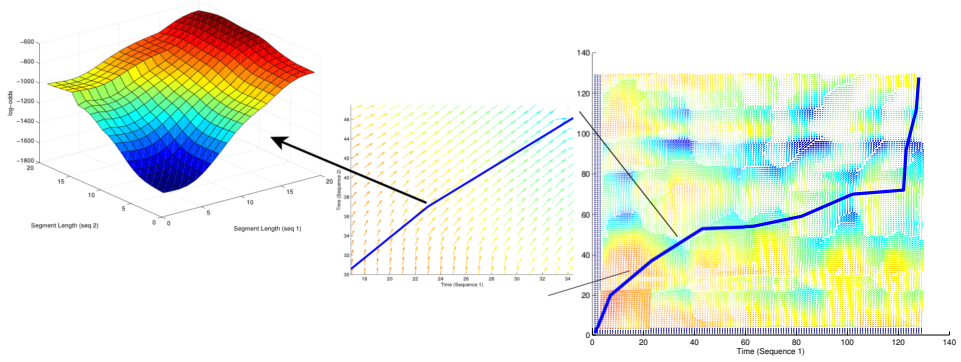


Figure 5: Segment length distribution for all positions for two EEG sequences. A certain portion of the graph is magnified. Smaller graph shows the likelihood of all possible segmentations for a single position (24,38) in the alignment matrix

Segment Length Distribution: Based on (12) we estimated the likelihood of all possible segmentations in aligning two EEG sequences for a maximum segment length of 30 and visualized it in Fig 5. The right-most graph depicts a vector field where each vector points to the most likely segment length (result of 12) at the corresponding position in the warping matrix and darker color indicates higher likelihood. The optimal alignment path is shown in the graph. A small portion of the graph is magnified in the middle graph, and then with the left-most graph depicting an example of the likelihood of all possible segmentations for a single position (24,38) selected by the alignment algorithm as a match operation. The chosen segment length at that position is 16 and 20 which has the highest likelihood and is the same segmentation selected by the alignment algorithm. This indicates the approximated forward algorithm can potentially be used to learn an improved local segmentation model.

Figure 6 shows the histogram of selected segment lengths for all pairs of sequences by aligning all recordings of two full sessions for target 2. The maximum segment length is set to length of the sequence to observe which segment lengths are selected without being limited to an upper bound. Since likely segments were mostly below the length of 20 we only show that potion of the histogram. Segment length of 1 and 1 is the most likely segment length. If this was not the case it would be very unlikely that DTW could result in any successful alignment.

5.3. Motion Capture Data

To contrast our approach with IsoCCA we tested SPHMM on MoCap sequences in the same setting. We used the same selection of sequences as Shariat and Pavlovic (2011). Namely, 62 sequences containing more than 40000 frames of 8 different actions from CMU MoCap dataset (http://mocap.cs.cmu.edu): walking, runing, boxing, jumping, marching, dancing, sitting down and shaking hands. Each class contains 7, 10, 8, 6, 10, 10, 7 and 4 sequences, respectively. Classes were selected with actions performed by different subjects. The dimensionality of data is reduced from 62 to 10 using PCA while keeping 99.8% of the energy. We compared SPHMM to IsoCCA, DTW and CTW Zhou and de la Torre (2009). 1-NN is used as the classifier to find the closest

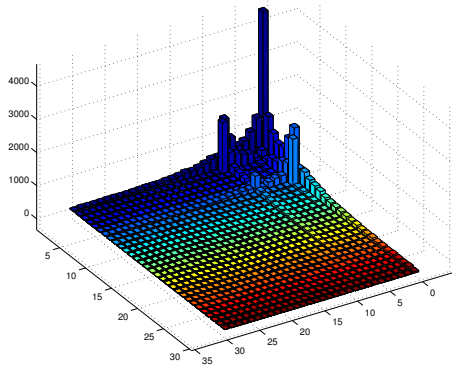


Figure 6: The distribution of segment lengths selected by alignment algorithm for all pairwise matches with maximum length of 129. Note that no segments of length over 30 were ever chosen.

Table 5: Accuracy of fixed segmentation

	$l_{fixed} = 5$	$l_{fixed} = 10$	$l_{fixed} = 20$	$l_{fixed} = 50$
Accuracy	80.65	74.19	77.42	69.35

sequence to any given query in a leave-one-out setting. Parameters for SPHMM are empirically set to $\delta = 0.001, \epsilon = .1, \tau = 0.01$ and $l_1 = l_2 = 10$. In DTW Sakoe-Chiba constraint with $\rho = 13\%$ is imposed to improve its performance in classification. For higher levels of noise we have permitted more gap operations for DTW by increasing warping window to $\rho = 18\%$. CTW is applied on the original 62 dimensional data set as it showed a better performance on it. As mentioned in Shariat and Pavlovic (2011), CTW is unable to achieve better results than DTW. The recognition accuracies are shown in Tab. 6.

Our method shows significantly higher performance compared to the other methods. The segmental approach was able to recognize proper segments of sequences and match them to their corresponding segments on the contrasting sequence. As an example, in Fig. 7, we have shown a portion of the alignment of two boxing sequences. Segments are separated by red lines and matched segments are indicated by arrows. Segments with no arrow pointing to them are either deleted or inserted based on the sequence one may take as reference. One can observe that similar actions are distinguished and matched. This can be explained by the fact that if the two partitions are similar and do not change drastically, the segment length tends to be longer (ref. Sec. 4). Another interesting observation is that the direction of action is ignored. Last match depicted in the figure, shows the correspondence of two punching actions one in forward and the other one in backward direction. In an action recognition task one is typically interested in retrieving actions regardless of their direction. However, the change of direction can sometimes introduce practical difficulties.

Average match segment length for MoCap was 3.70 with standard deviation of 4.05 showing that many (relatively) long segments are selected. Again to assert the efficacy of adaptive segment length determination we compared our main results against fixed segmentation (Tab. 5). The results are significantly inferior to adaptive SPHMM. Based on table 5 we assume that adaptive segmentation with maximum segment length of 20 may result in an even a better performance.

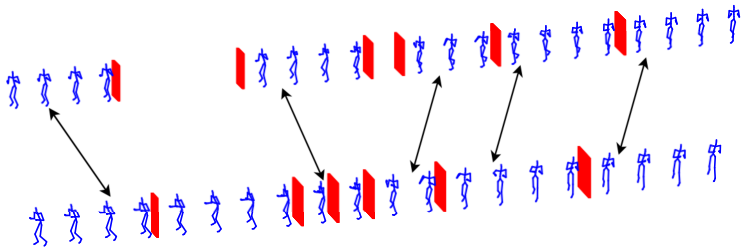


Figure 7: A portion of alignment of two boxing sequences. Segments are separated by red lines. The matched segments are indicated by arrows. Those segments with no arrow pointing to them are either deleted or inserted.

Table 6: Accuracy of SPHMM versus other methods

	SPHMM	IsoCCA	DTW	CTW
Accuracy	90.32	87.10	82.26	50.64

To assess the noise resilience of the SPHMM compared to other methods we added impulse noise in the same way described in section 5.1 except that the spread of the noise is restricted to 5% of the sequence. The noise is added only to the query sequences and the experiment setting is as above. To investigate whether a noise removal pre-processing can improve the performance of DTW beyond SPHMM, we apply a median filter on the data and show its performance with DTW-NR along with the accuracies of DTW, IsoCCA and SPHMM in Fig. 8. The noise level in Fig. 8 starts form $\omega = .2$ to make the noise removal performed on the query for DTW more meaningful. Obviously, noise removal on clean data will result in loss of information and leads to degraded performance for DTW. One can observe the stability of the classification accuracy of SPHMM in presence of different levels of noise. The noise removal can elevate the performance of DTW at high noise levels but it reduces the accuracy in lower levels of noise.

6. Conclusion

In this paper we presented a probabilistic model for segmental sequences alignment. We showed that a modified pair-HMM, in conjunction with a proper segment metric, can lead to effective joint segmentation and segmental alignment. Our experimental results showed high accuracy particularly when confronted with high levels of noise where DTW does not perform well even after noise removal pre-processing. Additionally, the invariance to local permutation has enabled our algorithm to perform well on non-causal signals. Compared to IsoCCA, our metric-based approach displays improved classification performance while having a reasonable computational complexity in practice.

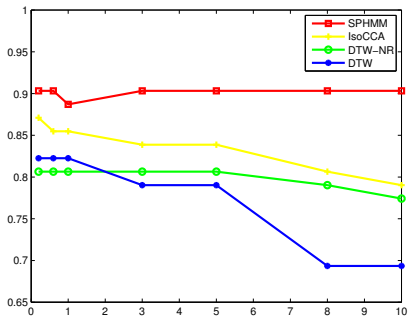


Figure 8: Comparing recognition accuracy of SPHMM versus other methods in presence of noise. Horizontal axes shows the level of noise.

References

Donald J. Berndt and James Clifford. Using dynamic time warping to find patterns in time series. In *KDD Workshop*, pages 359–370, 1994.

M. Blaum and J. Bruck. Coding for delay-insensitive communication with partial synchronization. *Information Theory, IEEE Transactions on*, 40(3):941 –945, may 1994. ISSN 0018-9448.

Franklin C. Crow. Summed area tables for texture mapping. *SIGGRAPH '84: Proceedings of the 11th annual conference on Computer graphics and interactive techniques*, pages 207–211, 1984.

J.C. de Munck, S.I. Gonçalves, L. Huijboom, J.P.A. Kuijer, P.J.W. Pouwels, R.M. Heethaar, and F.H. Lopes da Silva. The hemodynamic response of the alpha rhythm: An eeg/fmri study. *NeuroImage*, 35(3):1142 – 1151, 2007. ISSN 1053-8119.

H. Ding, G. Trajcevski, P. Scheuermann, X. Wang, and E. Keogh. Querying and mining of time series data: experimental comparison of representations and distance measures. *Proceedings of the VLDB Endowment*, 1(2):1542–1552, 2008.

Richard Durbin, Sean Eddy, Anders Krogh, and G. Mitchison. *Biologica Sequence Analysis. Probabilistic model of proteins and nuclear acids*. Cambridge University press, 1997.

U. Hoffmann, G. Garcia, J. Vesin, K. Diserens, and T. Ebrahimi. A Boosting Approach to P300 Detection with Application to Brain-Computer Interfaces. In *Proceedings of the IEEE EMBS Conference on Neural Engineering*, ISCAS. SPIE, 2005.

Risi Kondor. A kernel between sets of vectors. *MACHINE LEARNING-INTERNATIONAL*, 2003.

M. S. Ryoo. Human activity prediction: Early recognition of ongoing activities from streaming videos. *IEEE Conference on Compupter vision*, 2011.

Shahriar Shariat and Vladimir Pavlovic. Isotonic CCA for Sequence Alignment and Activity Recognition. *IEEE Conference on Compupter vision*, 2011.

A. Woznica, A. Kalousis, and M. Hilario. Distances and (indefinite) kernels for sets of objects. In *Data Mining, 2006. ICDM '06. Sixth International Conference on*, pages 1151 –1156, dec. 2006.

F. Zhou and F. de la Torre. Canonical time warping for alignment of human behavior. *Advances in Neural Information Processing Systems (NIPS)*, pages 1–9, 2009.