A PAIR HIDDEN MARKOV SUPPORT VECTOR MACHINE FOR ALIGNMENT OF HUMAN ACTIONS

Names!

ABSTRACT

Alignment of human actions in videos is an important task for applications such as action comparison and classification. While well-established algorithms such as dynamic time warping are available for this task, they still heavily rely on basic linear cost models and heuristic parameter tuning. In this paper we propose a novel framework that combines the flexibility of the pair hidden Markov model (PHMM) with the effective parameter training of the structural support vector machine (SSVM). The framework extends the scoring function of SSVM to capture the similarity of two input sequences and introduces suitable feature and loss functions. The proposed approach is evaluated against state-of-the-art algorithms such as dynamic time warping (DTW) and canonical time warping (CTW) on pairs of human actions from the Weizmann and Olympic Sports datasets. The experimental results show that the proposed approach is capable of achieving an accuracy improvement of over 7 percentage points over the runner-up on both datasets.

***Index Terms*—** sequence alignment, pair HMM, dynamic time warping, structural SVM, loss functions.

1. INTRODUCTION

Sequence alignment is an important application in many research domains such as speech recognition, computer vision and bioinformatics. Its goal is to align the frames of two (or more) sequences of measurements so as to maximise their local similarity and facilitate the analysis of correspondences. Typical alignment methods employ a cost model to account for the "cost" of individual alignment operations and exploit dynamic programming algorithms to provide minimum-cost alignments. Dynamic time warping (DTW) is likely the most well-known alignment algorithm [1]. While it had been originally proposed for the alignment of time series, it has later found use in a number of other applications including gesture recognition [2], speech processing [3] and classification of genomic signals [4]. Over the years, many DTW extensions have been proposed and state-of-the-art performance is held by canonical time warping (CTW) [5] that leverages the use of canonical correlation analysis (CCA) [6] to find the most effective subspace for the alignment of the two sequences. However, DTW and CTW likewise are limited to linear cost models and do not provide explicit procedures for the training of the model's parameters. An improved model for the alignment of sequences is offered by the *pair hidden Markov model* (pair HMM, or PHMM for short) that is, at the same time, a variant of DTW and of a conventional HMM providing a full probabilistic treatment of the alignment problem [7]. PHMM is a generative sequential model that emits pairs of measurements and allows for model training under maximum likelihood or other estimation frameworks.

In this paper, we propose a novel alignment algorithm that combines the features of a PHMM with the effective parameter estimation of the *structural support vector machine* (SSVM) [8]. The new model - named pair hidden Markov support vector machine (PHMM-SSVM) offers several advantages over conventional alignment algorithms including: 1) the ability to learn the cost model from any sets of supervised or partially-supervised examples; 2) the use of a maximum-margin training objective that has a proven reputation for accurate prediction; (3) the flexibility of using any loss functions of choice during training; and (4) the possibility of using kernels to implement non-linear cost models. Our main contributions are a set of dedicated feature and loss functions that allow PHMM-SSVM to achieve remarkable alignment accuracy. The proposed model has been tested against DTW and CTW in a set of experiments on action alignment over pairwise versions of the Weizmann dataset [9] and the Olympic Sports dataset [10]. The experimental results show that the proposed approach is able to outperform existing models in terms of alignment accuracy.

2. PROBABILISTIC SEQUENCE ALIGNMENT AND MAXIMUM-MARGIN TRAINING

In the following, we describe the two main frameworks - the pair hidden Markov model (PHMM) and the structural support vector machine (SSVM) - that form the basis for the proposed integration.

2.1. Pair Hidden Markov Model

PHMM is a probabilistic model for pairwise sequence alignments. Given two sequences, $s = \{s_1, ..., s_i, ..., s_{L_s}\}$ and $t = \{t_1, ..., t_j, ..., t_{L_t}\}$, their alignment can be intuitively defined as a sequence of index pairs from the two sequences.

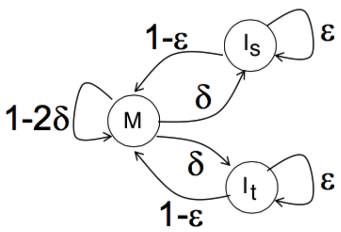

Fig. 1: PHMM state diagram.

Table 1: Transition probabilities table.

	M	S	T
M	$1-2\delta$	δ	δ
S	$1-\varepsilon$	ε	0
T	$1-\varepsilon$	0	ε

However, to simplify both notations and operations, the alignment is re-defined as a sequence of only three types of symbols: M ("match"), S ("insert a gap on sequence s") and T ("insert a gap on sequence t"). The symbols have the following meaning: assuming i and j to be the current indices over sequences s and t, respectively, 1) symbol M pairs frames s_i and t_j and then increments both indices; 2) symbol S pairs no frames and only increments index j; and, likewise, 3) symbol T pairs no frames and only increments index i. As a toy example, we show below a possible alignment for two short sequences from character set $\{A, B, C, D\}$:

$$\begin{array}{rcccccccc} s = & A & B & - & C & B & D & A & D \\ t = & A & B & D & C & B & - & A & D \\ y = & M & M & S & M & M & T & M & M \end{array} \quad (1)$$

In the above example, sequence y encodes the alignment, with the M symbols showing the matched frames (e.g, s_3 and t_4) and the S and T symbols accounting for the required gaps. The length of the alignment is bounded between $\max(L_s, L_t)$ and $L_s + L_t$.

In probability notation, a PHMM is a model for the joint probability, $p(s, t, y)$, of the two sequences and their alignment. Such a model can be used to infer an optimal alignment, $\bar{y}$, for the two sequences as $\bar{y} = \mathrm{argmax}_y\, p(s, t, y)$. Like for a conventional HMM, the joint probability of a PHMM factorises into a set of transition and emission probabilities. The transition probabilities are commonly defined as: (1) δ for transitions from M to either S or T; (2) ε for staying in S or T; (3) $1 - \varepsilon$ for transitions from either S or T to M. Note that the model bars direct transitions from S to T and the vice versa assuming that a pair of matched frames will always follow a run of gaps. Figure 1 shows the state diagram of the PHMM, while Table 1 shows the complete transition probabilities table.

To complete the model, we also need to define the emission probabilities. To this aim, we note the probability of emitting aligned pair (a, b) as $p_{a,b}$ and the probability of emitting measurement a against a gap as q_a. In the common case of numerical measurements, both p and q will be multi-variate likelihoods such as Gaussian distributions or mixture models.

Using a PHMM, the optimal alignment for a pair of sequences can be found via an equivalent Viterbi algorithm [11]. Its computational complexity, $O(\max(L_s, L_t))$, is only linear in the length of the sequences thus ensuring fast and efficient alignments. The main steps of the algorithm are given below, where the probability of reaching state $* = \{M, S, T\}$ at indices i and j over s and t is noted as $p^*(i, j)$.

Initialization: $p^M(1, 1) = p^S(1, 1) = p^T(1, 1) = p^*(0, j) = p^*(i, 0) = 1$.

Recurrence: $i = 1, ..., L_s, j = 1, ..., L_t$:

$$p^M(i,j) = p_{s_i,t_j} \max \left\{ \begin{array}{c} (1-2\delta)\; p^M(i-1, j-1) \\ (1-\varepsilon)\; p^S(i-1, j-1) \\ (1-\varepsilon)\; p^T(i-1, j-1) \end{array} \right. \quad (2)$$

$$p^S(i,j) = q_{s_i} \max \left\{ \begin{array}{l} \delta\; p^M(i-1, j) \\ \varepsilon\; p^S(i-1, j) \end{array} \right. \quad (3)$$

$$p^T(i,j) = q_{t_j} \max \left\{ \begin{array}{l} \delta\; p^M(i, j-1) \\ \varepsilon\; p^T(i, j-1) \end{array} \right. \quad (4)$$

Termination:

$$p(s, t, y) = \max(p^M(L_s, L_t), p^S(L_s, L_t), p^T(L_s, L_t)) \quad (5)$$

2.2. Structural SVM

Structural SVM is a powerful classifier that extends the notion of maximum-margin classification to the case of structured prediction. This case includes the classification of structures such as sequences and graphs and problems such as alignment and ranking. In the case of alignment, the problem is to learn a scoring function, $F(s, t, y)$, between input sequences s and t and output alignment y based on training samples of input-output pairs. The scoring function typically takes the form of a linear discriminant, $F(s, t, y) = w^\top \psi(s, t, y)$, that can be extended to non-linear mappings by the use of kernels. To learn an accurate alignment predictor, the training objective ensures that, for every training sample, the scoring function assigns its ground-truth alignment with a score higher than that of any other alignments by an appropriate margin.

The challenge with structural SVM is that the number of possible alignments for a given sequence pair is exponential in their length. This in turn leads to a highly-constrained learning objective that proves computationally infeasible even for relatively short sequences. However, Tsochantaridis *et al.* in [8] have shown that a very close approximation to the solution of SSVM can be obtained by using only a polynomial (i.e., easily feasible) number of constraints, and Joachims *et al.* in [12] have shown that this approach can also be used for the sequence alignment problem. Given a supervised training set of sequence pairs and alignments, $(s^i, t^i, y^i), i = 1 \ldots N$, the relaxed objective can be written as:

$$\begin{aligned} &\underset{w,\xi}{\text{argmin}} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi^i \quad s.t. \\ &w^\top\psi(s^i,t^i,y^i) - w^\top\psi(s^i,t^i,y) \geq \Delta(y^i,y) - \xi^i, \\ &\qquad i = 1\dots N, \ \ \forall y \in \mathcal{W} \end{aligned} \tag{6}$$

Like for a conventional SVM, objective (6) aims to strike a balance between the prediction error over the training set ($\sum_{i=1}^{N}\xi^i$) and a regulariser ($\|w\|^2$). The constraints ensure that the score assigned to the ground-truth alignment, y_i, is higher than that assigned to any other alignment y stored in a working set, $\mathcal{W}$, by margin $\Delta(y^i,y)$. At its turn, $\Delta(y^i,y)$ is a loss function that can be arbitrarily chosen to quantify the inaccuracy of incorrect alignments. The working set, $\mathcal{W}$, is populated using a constraint-violation approach that ensures that the solution for the relaxed objective (6) is "epsilon-close" to the solution of the complete objective. More details are provided in Section 3.3.

3. THE PROPOSED INTEGRATION: PHMM-SSVM

The integration of the PHMM in the SSVM framework (PHMM-SSWM hereafter) can be obtained by simply setting the PHMM's joint probability as:

$$p(s,t,y) \propto \exp(w^\top\psi(s,t,y)) \tag{7}$$

This restricts the emission probabilities to belong to the exponential family of distributions, which is however a very broad and encompassing family (Gaussian, Gamma, chi-squared etc). In addition, the assumption does not require the distribution to be normalised and therefore the w parameters can be chosen from a larger domain. Lastly, a non-linear parametrisation can be easily obtained by using kernels.

3.1. Parameter Vector and Feature Function

In the structural SVM framework, the score for a sample (s,t,y) is obtained from the product of a parameter vector, w, and a feature function, ψ, that provides a re-mapping of the given measurements and labels. As a common assumption, the score is assumed to be a decomposable function (a sum) over the individual labels of the assignment, $y_k, k = 1\dots|y|$. The parameter vector contains two sections: transition parameters w^{tr} and emission parameters w^{em}. The transition parameters are a (partial) 3×3 matrix indexed by labels y_{k-1} and y_k (transitions between symbols $S \rightarrow T$ and $T \rightarrow S$ are not allowed). As emission feature, we simply consider the absolute difference of measurements s_i and t_j in matching states; therefore, the emission parameters are a vector with the same dimensionality as the individual measurements, i.e. $w^{em}, s_i, t_j \in \mathbb{R}^D$. Logically, w^{em} should be assigned negative values during training so that more dissimilar measurements receive lower scores; however, we do not impose a negativity constraint on these parameters. With these assumptions, the score can be re-written as:

$$\begin{aligned} &w^\top\psi(s,t,y) = \sum_{k=1}^{|y|} w^{tr}_{y_{k-1},y_k} + w^{em\top}|s_i - t_j|\mathbf{I}[y_k = M] \\ &w^{tr}_{0,*} = 0; \quad \mathbf{I}[y_k = M] : i{+}{+}, j{+}{+}; \\ &\mathbf{I}[y_k = S] : j{+}{+}; \quad \mathbf{I}[y_k = T] : i{+}{+} \end{aligned} \tag{8}$$

where $\mathbf{I}$ is the indicator function and indices i and j, initially set to 1, are post-incremented according to the value of label y_k.

3.2. Loss Functions

The common way to measure the inaccuracy of a predicted alignment is by use of a Hamming distance between the ground-truth alignment, y, and the prediction, $\bar{y}$. This function is often referred to as Q-loss function in the alignment literature and noted as $\Delta_Q(y,\bar{y})$ [13]. The Q-loss is decomposable over the individual operations in the alignments as $\Delta_Q(y,\bar{y}) = 1 - \sum_{k=1}^{|y|}\delta(y_k,\bar{y}_k)$. At its turn, $\delta(y_k,\bar{y}_k)$ returns $1/N$ (N: number of frame matches in the ground truth) when a ground-truth match is correctly predicted and 0 otherwise. In practice, we compute the loss by explicitly unfolding all the frame indices over sequences s and t in both the ground truth and the predicted alignment.

Another useful loss function is the Q_4-loss: this is a more lenient loss function that counts a match as correct even if the indices of the matching frames in the prediction are shifted by ± 2 compared to those in the ground truth. In the experiments, we report results in terms of both Q-loss and Q_4-loss. In addition, during the annotation of the training set we annotate the ground-truth alignment only for some "key" frames that we can match with high confidence (e.g., apex phases of movements). The loss is measured only against such key-frame matches.

3.3. Most-Violated Constraints

In learning a structural SVM model for alignments, one should ensure that the ground-truth alignments receive scores higher than all other, possible alignments. However, the number of possible alignments is exponential in the length of the input sequences, leading to an unmanageable number of constraints even for sequences of relatively short length. The solution proposed by [8, 12] is a relaxed problem (6) that only considers the sub-set of the "most-violated constraints", i.e. the constraints that set the value of penalty ξ^i for each sample, $i = 1\dots N$. The solution is proven to be ϵ-close to that of the fully-constrained problem, where ϵ is a small constant (set to 0.01 in our experiments) that can be made arbitrary smaller at the cost of only a polynomial increase in the number of iterations of the solver. The most-violated constraint for each sample is identified from a re-writing of the constraints:

$$\begin{aligned}
&w^\top \psi(s^i, t^i, y^i) - w^\top \psi(s^i, t^i, y) \geq \Delta(y^i, y) - \xi^i \ \ \forall y \\
&\rightarrow \xi^i \geq -w^\top \psi(s^i, t^i, y^i) + w^\top \psi(s^i, t^i, y) + \Delta(y^i, y) \ \ \forall y \\
&\rightarrow y^{*i} = \underset{y}{\operatorname{argmax}} (w^\top \psi(s^i, t^i, y) + \Delta(y^i, y))
\end{aligned} \tag{9}$$

As (9) shows, the alignment y^{*i} setting the value of penalty ξ^i can be found by a modified version of the inference, known as "loss-augmented" inference since it adds up the loss function to the score. Given that the loss function is decomposable frame-by-frame, the efficient Viterbi algorithm can also be used for this maximisation. Algorithm 1 shows the main steps of the learning procedure.

Algorithm 1: Learning algorithm: main steps.

```
Input  : Measurement sequences $s^i, t^i$ and
         ground-truth alignment $y^i$, $i = 1 \dots N$;
         parameter $\epsilon$
$\mathcal{W} = \emptyset, \ w = 0, \ \xi = 0$
repeat
    foreach $i = 1 \dots N$ do
        $y^{*i} \leftarrow \operatorname{argmax}_y (w^\top \psi(s, t, y) + \Delta(y^i, y))$;
        if $\xi_i = [w^\top (\psi(s^i, t^i, y^{*i}) - \psi(s^i, t^i, y^i)) +$
        $\Delta(y^i, y^{*i})] > \xi_i^{prev} + \epsilon$ then
            $\mathcal{W} \leftarrow \mathcal{W} \cup y^{*i}$;
        end
    end
    $(w, \xi) = \operatorname{argmin}_{w,\xi} \frac{1}{2} \|w\|^2 + C \sum_{i=1}^N \xi^i \ s.t. \ \mathcal{W}$;
until $\xi$ unchanged;
Output: Model $w$
```

4. EXPERIMENTS

The following experiments evaluate the proposed PHMM-SSVM model in the temporal alignment of action videos against DTW [1] and a state-of-the-art algorithm, CTW [5]. In the first experiment, we compare the performance in aligning the "jump" action from different subjects of the Weizmann dataset [9]. In the second experiment, we compare the "clean-and-jerk" action performed by 11 subjects from the challenging Olympic Sports dataset [10]. For the SSVM training, we have set parameters C to 10 and ϵ to 0.01, with no noticeable sensitivity. Results are reported in terms of both Q-loss and Q_4-loss (see section 3.2).

4.1. Weizmann Dataset

The Weizmann dataset contains ten actions performed by nine actors. While it has been long saturated in terms of action recognition accuracy, according to [5] it is still probing for testing alignment accuracy. In this experiment, we follow [5] and first subtract the background from the videos and then process the resulting frames by the Euclidean distance transform [14], preserving 99% of the energy by retaining the top 416 principal components. As test data, we have formed 13 video pairs from action "jump" selecting different subject pairs and annotating their key-frames manually. We have then randomly picked 6 as training set and the others as test set, yet ensuring that the subjects pairs in the test set did not appear in the training.

	1 - Q-loss ($\%$)	1 - Q_4-loss ($\%$)
PHMM-SSVM	68.6%	98.0%
DTW	41.2%	72.6%
CTW	60.8%	96.1%

Table 2: Alignment accuracy for action "jump" in the Weizmann dataset.

Table 2 reports the alignment accuracy on the test set as the one-complement of the Q and Q_4 losses. The table clearly shows that PHMM-SSVM achieves higher accuracy in terms of Q-loss than both DTW (27.4 percentage points) and CTW (7.8 percentage points). Since the Q_4-loss is more lenient, its accuracy is generally higher for all algorithms: however, the proposed PHMM-SSVM still achieves the highest accuracy and the ranking is unvaried.

4.2. Olympic Sports Dataset

The Olympic Sports dataset is a more challenging dataset of real sport videos from YouTube. In this dataset, we chose action "clean-and-jerk" (a specialty of weightlifting) since the manual alignment of its key-frames is relatively certain. We created 55 pairs of ground-truth alignments and split them into 27 pairs for training and 28 for test. As measurements, we computed dense feature descriptors for each frame of the video sequences usig Laptev's STIP extractor [15]. We then computed a bag-of-words for each frame with $1,000$ bins using the VLFeat library [16]. Note that the specific choice of features is not the focus of this paper.

Table 3 reports the alignment accuracy on the test set. Again, the table clearly shows that PHMM-SSVM achieves higher accuracy in terms of Q-loss than both DTW (13.1 percentage points) and CTW (7.6 percentage points). While the Q_4-loss tends to reduce these differences, PHMM-SSVM still achieves the highest accuracy in terms of Q_4-loss and the ranking is again unvaried. Figure 2 shows an example of the ground-truth and predicted alignments: a) the top two rows show six manually-matched key-frames from the two sequences. The frames from the first sequence are used as "template" and those from the second represent the ground-truth alignment; b) the third row shows the alignment pre-

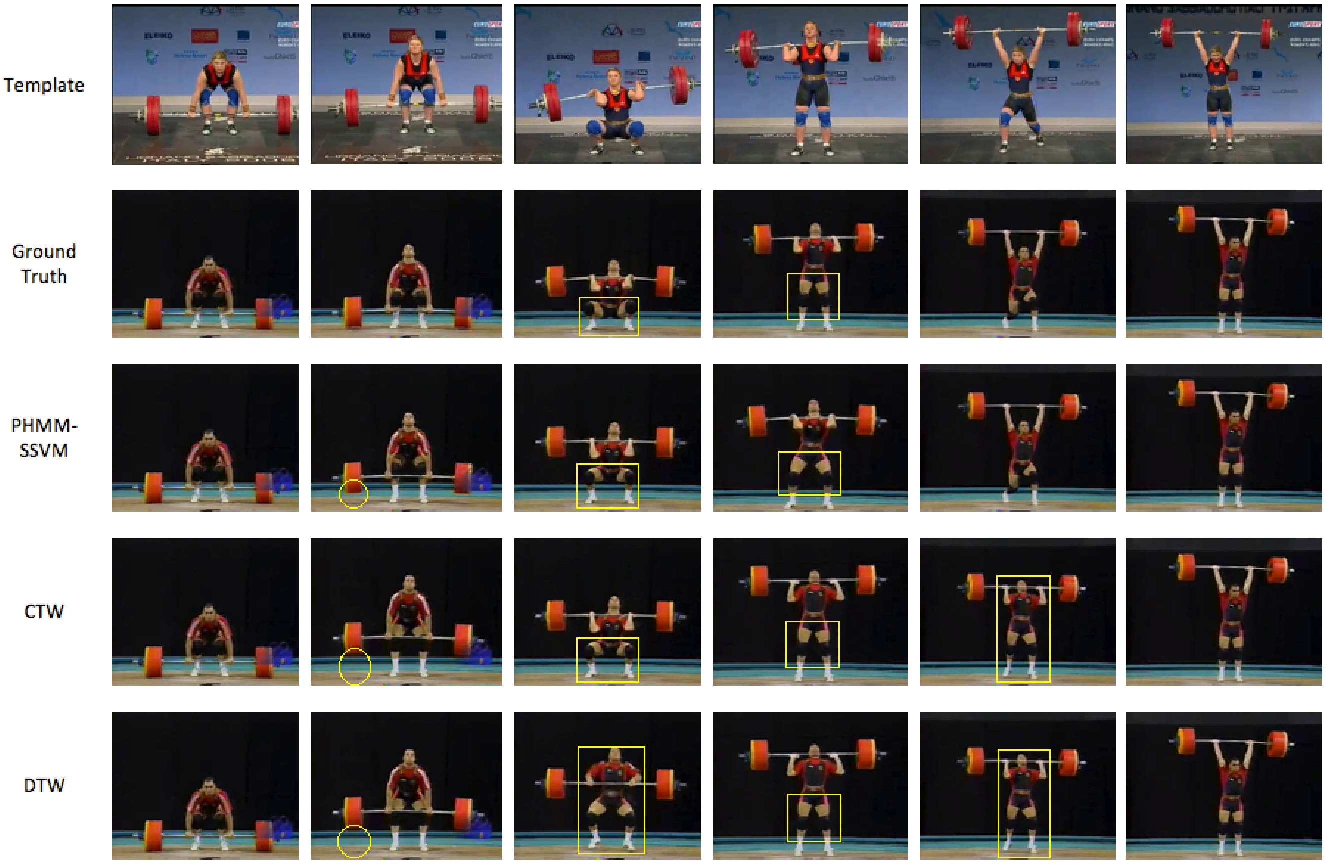

Fig. 2: Example of ground-truth and predicted alignments from the Olympic Sports dataset: top two rows): six matched key-frames from two clean-and-jerk sequences. The first row is used as template and the frames of the second row show the ground-truth alignment; third row): alignment predicted by the proposed PHMM-SSVM; bottom two rows): alignments predicted by CTW and DTW. The superimposed ellipses and rectangles visually highlight the alignment errors.

	1 - Q-loss (%)	1 - Q_4-loss (%)
PHMM-SSVM	50.0%	74.2%
DTW	36.9%	69.6%
CTW	42.4%	70.5%

Table 3: Alignment accuracy for action "clean-and-jerk" in the Olympic Sports dataset.

dicted by the proposed PHMM-SSVM; and c) the bottom two rows show the alignments predicted by CTW and DTW, respectively. The superimposed ellipses and rectangles visually highlight the alignment errors. This figure shows that the alignment predicted by PHMM-SSVM only mildly differs from the ground truth and is more accurate than those returned by DTW and CTW.

5. CONCLUSION

In this paper, we have presented a novel approach for sequence alignment and showed its effectiveness in aligning human actions in videos. The proposed method - named pair hidden Markov support vector machine (PHMM-SSVM) - integrates the probabilistic formulation of the pair HMM with the effective parameter training of structural SVM. The proposed integration includes dedicated feature and loss functions suitable to achieve accurate alignments. Experimental results over two probing video datasets show that PHMM-SSVM achieves higher accuracy than both a standard dynamic programming solution (DTW) and a state-of-the-art algorithm (CTW), with improvements over the runner-up of more than 7 percentage points. In addition, while in this work we have used the proposed model for aligning actions in videos, there are no standing limitations to its general use in any other domain. In the near future, we plan to extend our approach to semi-supervised settings with limited ground-truth annotation.

6. REFERENCES

[1] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice-Hall Signal Processing Series, Englewood Cliffs, New Jersey, 1993.

[2] D. M. Gavrila and L. S. Davis, "Towards 3-d model-based tracking and recognition of human movement: a multi-view approach," in *In International Workshop on Automatic Face- and Gesture-Recognition. IEEE Computer Society*, 1995, pp. 272–277.

[3] C. Myers, L. Rabiner, and A. Rosenberg, "Performance tradeoffs in dynamic time warping algorithms for isolated word recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 28, no. 6, pp. 623–635, 1980.

[4] Helena Skutkova, Martin Vtek, Petr Babula, Ren Kizek, and Ivo Provaznik, "Classification of genomic signals using dynamic time warping.," *BMC Bioinformatics*, vol. 14, no. S-10, pp. S1, 2013.

[5] Feng Zhou and Fernando De la Torre, "Canonical time warping for alignment of human behavior," in *Advances in Neural Information Processing Systems Conference (NIPS)*, December 2009.

[6] T. W. Anderson, Ed., *An Introduction to Multivariate Statistical Analysis*, Wiley, 1984.

[7] Richard Durbin, Sean Eddy, Anders Krogh, and Graeme Mitchison, "Biological sequence analysis: probabilistic models of proteins and nucleic acids," 1998.

[8] I. Tsochantaridis, T. Joachims, T. Hofmann, and Y. Altun, "Large margin methods for structured and interdependent output variables," *JMLR*, vol. 6, pp. 1453–1484, 2005.

[9] Lena Gorelick, Moshe Blank, Eli Shechtman, Michal Irani, and Ronen Basri, "Actions as space-time shapes," *Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 12, pp. 2247–2253, December 2007.

[10] Juan Carlos Niebles, Chih wei Chen, and Li Fei-fei, "Modeling temporal structure of decomposable motion segments for activity classification," in *in Proc. 11th European Conf. Comput. Vision*, 2010, pp. 392–405.

[11] Matthew S. Ryan and Graham R. Nudd, "The viterbi algorithm," Tech. Rep., Coventry, UK, UK, 1993.

[12] Thorsten Joachims, Tamara Galor, and Ron Elber, "Learning to align sequences: A maximum-margin approach," in *New Algorithms for Macromolecular Simulation, B. Leimkuhler, LNCS Vol. 49, Springer*, 2005.

[13] Chun-Nam John Yu, Thorsten Joachims, Ron Elber, and Jaroslaw Pillardy, "Support vector training of protein alignment models.," *Journal of Computational Biology*, vol. 15, no. 7, pp. 867–880, 2008.

[14] Calvin R. Maurer, Rensheng Qi, Vijay Raghavan, and Senior Member, "A linear time algorithm for computing exact euclidean distance transforms of binary images in arbitrary dimensions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 265–270, 2003.

[15] Ivan Laptev and Tony Lindeberg, "Space-time interest points," in *IN ICCV*, 2003, pp. 432–439.

[16] Andrea Vedaldi and Brian Fulkerson, "Vlfeat: An open and portable library of computer vision algorithms," in *Proceedings of the 18th ACM International Conference on Multimedia*, New York, NY, USA, 2010, MM '10, pp. 1469–1472, ACM.