Time Warping Under Dynamic Constraints With Application to Non-Stationary Action Alignment and Classification

Michalis Raptis Matteo Bustreo Stefano Soatto

:

Abstract

Action and event recognition from video require comparing temporal sequences of images, or of intermediate representations derived from them. Such a comparison should be insensitive to intrinsic temporal variations within the same class – for instance the speed of execution of a particular gesture – and at the same time retain the discriminative power that enable classifying different actions. In this paper, we propose a technique to compare temporal sequences that accounts for dynamic constraints implicit in the data generation process. Our technique is more flexible than those previously used for quasi-periodic actions such as walking gaits, but more discriminative than others based on dynamic time warping that do not satisfy dynamic constraints. We illustrate our approach on public datasets including stationary and non-stationary actions, using both motion-capture and image data. In all the experiments we have conducted, our approach outperforms competing ones. We highlight experiments where it exhibits limitations.

1. Introduction

Comparing time series is a problem of critical importance in the analysis of video for the detection and classification of actions or events of interest. These in turn are relevant to surveillance, environmental monitoring, and human-machine interfaces. In addition to the challenges of geometric and photometric variability common to other visual classification tasks, video analysis requires dealing with temporal variability, whereby the same event can occur at a variety of speeds, starting from a variety of initial instants and following a variety of velocity profiles. While geometric and photometric information present in *one image* is undoubtedly important (and indeed often sufficient) to recognize actions and events, the temporal evolution contains a significant amount of information, as illustrated eloquently by [9]. In this manuscript, therefore, *we concentrate on* the classification of *events that have distinct temporal signatures.* Comparison of time series is also key in a number of other disciplines, where a variety of tools have been developed from "dynamic time warping" in speech recognition [19] and temporal data mining [18] to Lyapunov exponents and non-linear embedding in chaotic physical and financial systems [10], to stochastic realization theory for control systems [6]. We argue, however, that *the analysis of motion imagery requires the development of dedicated tools,* because the models underlying other disciplines are either too restrictive or much too general. In fact, the assumption underlying most data-driven models in system identification is stationarity [14], which is obviously violated except for quasi-periodic gaits. On the opposite end, dynamic time warping (DTW) [18] reparametrizes the temporal axis in a way that is not compatible with physical constraints implicit in the data formation process. When we image the physical world, actions are performed by objects with masses and inertias, so their behavior can only generate velocity profiles that obey the resulting dynamic constraints.[1] Therefore, in this manuscript we introduce a *time warping model that accounts for dynamic constraints intrinsic in the hidden generative model* of an action or event. We call this **time warping under dynamic constraints** (TWDC).

1.1. Prior work

The comparison of time series takes a different form in different disciplines, and we refer the reader to textbooks cited above. Here we discuss the specific case of the analysis of video, where temporal variability can be addressed in the *representation*, by devising statistics of a video snippet that does not depend on its temporal evolution, or as part of the *matching process*, by defining

[1] Needless to say, even the whackiest actions we are likely to observe in applications to security, monitoring and surveillance are unlikely to approach the chaotic nature implicit in models for finance forecasting and astrophysics.

suitable distances or other discrepancy measures. Examples of the first approach include *averaging statistics*, where the video (or some pre-processing of it that reduces the effects of photometric variability, typically the extraction of silhouettes from background subtraction) is integrated against a kernel to arrive at a static feature [27]. These methods are specific to a particular image statistic (e.g. the silhouette) and do not generalize easily to other models (say affine moments). Another example is *instantaneous statistics*, where the value and derivative of a feature vector is computed at each instant and then quantized over time using a hidden Markov model [20]. These methods represent a coarse decimation of the original signal, so much of the information implicit in the dynamics is lost. Yet another example is to design *invariants* of the sequence. For instance, transfer functions are shift-invariant and are sufficient statistics for stationary processes [4, 2], but these methods do not generalize to non-stationary actions.

The alternate approach consists of representing the video as just a collection of ordered frames (or some statistics of it), and then devise methods to compare video snaps that minimizes the effects of temporal variations. These include the Kullback-Leibler divergence between the sample video distributions, which are frought with computational difficulties, although recent advances make them more efficient [5], correlation kernels [25], and direct block-correlations [22, 27], also rather computationally intensive. These methods could be considered *discriminative* in the sense that they compare time series without regards to how they are generated, and the underlying model is implicit in the comparison algorithm. Dynamic time warping (DTW) [18] falls in this category, in that it mods out temporal variations as part of the matching process. It has been used successfully in other domains of vision research, from epipolar matching in calibrated stereo [17] to discrete-time action modeling [26], to handwriting [15] among others.

In between these two approaches there are *likelihood methods* that use one sequence to infer an underlying model, and then use this model to explain the data of the other sequence. In this approach, the more data are available (hence the better the estimate of the model) the worse the classification error is – an apparent paradox induced by the fact that the generalization model underlying this approach is trivial: Each realization models one sequence and noisy versions of it, without regard for the structure of the intrinsic variability that different realizations of the same process exhibit

More importantly, *none of the methods proposed so far compare temporal sequences in a way that explicitly takes into account the dynamics of the hidden process that generate the data*, though capturing those dynamics has been shown to lead to more robust classification results [16, 1]. Therefore, in this manuscript we propose a time-warping distance, following the lines of [18], that however takes into account dynamic constraints. Our work also relates to [23] who propose characterizing the space of activities as the quotient of a time series under time warpings, and from [11], who extend dynamic time warping to include temporal derivatives.

2. Formalization

The simplest instantiation of our problem can be formalized as searching for a distance $d(y_1, y_2)$ between two time series $y_i = \{y_i(t) \in \mathbb{R}^N\}_{t=1,\dots,T}$. For simplicity we will assume that the sequences have the same length, although all considerations extend to allow for different lengths. Among the simplest distances one could define is the $\mathbb{L}^2$ norm of the difference, $d_0(y_1, y_2) = \int_0^T \|y_1(t) - y_2(t)\|^2 dt$, which corresponds to a generative model where both sequences come from an (unknown) underlying process $\{h(t)\}$, corrupted by two different realizations of additive white zero-mean Gaussian "noise" (here the word noise lumps all unmodeled phenomena, not necessarily associated to sensor errors)

$$y_i(t) = h(t) + n_i(t) \quad i = 1, 2; \; t \in [0, \; T] \tag{1}$$

The $\mathbb{L}^2$ distance is then the (maximum-likelihood) solution for h that minimizes

$$d_0(y_1, y_2) = \min_h \phi_{data}(y_1, y_2|h) \doteq \sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt \tag{2}$$

subject to (1). Here h can be interpreted as the *average* of the two time series, and although in principle h lives in an infinite-dimensional space, no regularization is necessary at this stage, because the above has a trivial closed-form solution. However, later we will need to introduce regularizers, for instance of the form $\phi_{reg}(h) = \int_0^T \|\nabla h\| dt$. This admittedly unusual way of writing the $\mathbb{L}^2$ distance makes the extension to more general models simpler, as we discuss in the next subsections.

2.1. Dynamic time warping (review)

In this section we revisit dynamic time warping in a way that makes it amenable to the extentions we have discussed in the introduction. Consider an arbitrary

infinite-dimensional diffeomorphism x of the interval $[0, T]$, called a *time warping*, so that (1) becomes

$$y_i(t) = h(x_i(t)) + n_i(t) \quad i = 1, 2. \tag{3}$$

The data term of the cost functional we wish to optimize is still $\sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt$, but now subject to (3), so that minimization is with respect to the unknown functions x_1 and x_2 as well as h. Since the model is over-determined, we must therefore impose regularization [12] to compute the *time-warping distance*

$$d_1(y_1, y_2) = \min_{h \in \mathcal{H}, x_i \in \mathcal{U}} \phi_{data}(y_1, y_2 | h, x_1, x_2) + \phi_{reg}(h). \tag{4}$$

In order for $\tau \doteq x(t)$ to be a viable temporal index, x must satisfy a number of properties. The first is continuity (time, alas, does not jump); in fact, it is common to assume a certain degree of smoothness, and for the sake of simplicity we will assume that x_i is infinitely differentiable. The second is causality: The ordering of time instants has to be preserved by the time warping, which can be formalized by imposing that x_i be monotonic. Making the constraints more explicit, we can re-write the distance above as

$$\min_{h \in \mathcal{H}, x_i \in \mathcal{U}} \sum_{i=1}^{2} \int_0^T \|y_i(t) - h(x_i(t))\|^2 + \lambda \|\nabla h(t)\| dt \tag{5}$$

where λ is a tuning parameter that can be set equal to zero, for instance by choosing $h(t) = y_1(x_1^{-1}(t))$, and the assumptions on the warpings x_i are implicit in the definition of the set $\mathcal{U}$. This is an optimal control problem, that is solved globally using dynamic programming in a procedure called "dynamic time warping" (DTW).

It is important to note that **there is nothing "dynamic" about dynamic time warping,** other than its name. There is no requirement that the warping function x be subject to dynamic constraints, such as those arising from forces, inertia etc. However, some notion of dynamics can be coerced into the problem by characterizing the set $\mathcal{U}$ in terms of the solution of a differential equation. Following [18], as shown by [13], one can represent allowable $x \in \mathcal{U}$ in terms of a small, but otherwise unconstrained, scalar function u: $\mathcal{U} = \{x \in \mathcal{H}^2([0,\ T]) \ | \ddot{x} = u\dot{x};\ u \in \mathbb{L}^2([0, T])\}$ where $\mathcal{H}^2$ denotes a Sobolev space. If we define $\rho_i \doteq \dot{x}_i$ then $\dot{\rho} = u\rho$; we can then stack the two into $\xi \doteq [x,\ v]^T$, and $C = [1,\ 0]$, and write the data generation model as

$$\begin{cases} \dot{\xi}_i(t) = f(\xi_i(t)) + g(\xi_i(t)) u_i(t) \\ y_i(t) = h(C\xi_i(t)) + n_i(t) \end{cases} \tag{6}$$

as done by [13], where $u_i \in \mathbb{L}^2([0, T])$. Here f, g and C are given, and $h, x_i(0), u_i$ are nuisance parameters that are eliminated by minimization of the same old data term $\sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt$, now subject to (6), with the addition of a regularizer $\lambda \phi_{reg}(h)$ and an energy cost for u_i, for instance $\phi_{energy}(u_i) \doteq \int_0^T \|u_i\|^2 dt$. Writing explicitly all the terms, the problem of dynamic time warping can be written as

$$d_3(y_1, y_2) = \min_{h, u_i, x_i} \sum_{i=1}^{2} \int_0^T \|y_i(t) - h(C\xi_i(t))\| + \\ + \lambda \|\nabla h(t)\| + \mu \|u_i(t)\| dt \tag{7}$$

subject to $\dot{\xi}_i = f(\xi_i) + g(\xi_i) u_i$. Note, however, that this differential equation is only an expedient to (softly) enforce causality by imposing a small "time curvature" u_i.

In the next section we discuss how to enforce dynamic constraints in the comparison of two time series.

3. Time warping under dynamic constraints

Our strategy to enforce dynamic constraints in dynamic time warping is illustrated in Figure 1:

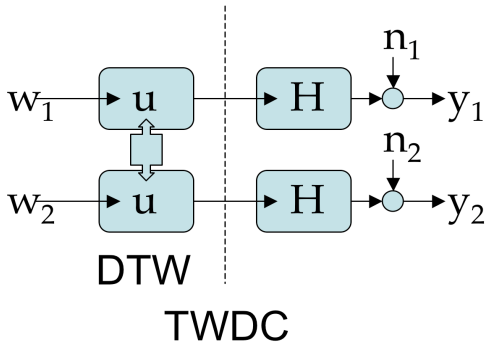

Figure 1. *Traditional dynamic time warping (DTW) assumes that the data come from a common function that is warped in different ways to yield different time series. In time warping under dynamic constraints (TWDC), the assumption is that the data are output of a dynamic model, whose inputs are warped versions of a common input function.*

Now, rather than the data being warped versions of some common function, as in (3), we will assume that *the data are outputs of dynamical models driven by inputs that are warped versions of some common function.* In other words, given two time series $y_i,\ i = 1, 2$, we will assume that there exist suitable matrices A, B, C,

state functions x_i of suitable dimensions, with their initial conditions, and a *common input* u such that the data are generated by the following model, *for some warping functions* $w_i \in \mathcal{U}$*:*

$$\begin{cases} \dot{x}_i(t) = Ax_i(t) + Bu(w_i(t)) \\ y_i(t) = Cx_i(t) + n_i(t). \end{cases} \tag{8}$$

Our goal is to find the distance between the time series by minimizing with respect to the nuisance parameters the usual data discrepancy $\sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt$ subject to (8), together with regularizing terms $\bar{\phi}_{reg}(u)$ and with $w_i \in \mathcal{U}$. Notice that this model is considerably different from one discussed in the previous section, as the state ξ earlier was used to model the temporal warping, whereas now it is used to model the data, and the warping occurs at the level of the input. It is also easy to see that the model (8), despite being linear in the state, includes (6) as a special case, because we can still model the warping functions w_i using the differential equation in (6). In order to write this *time warping under dynamic constraint* problem more explicitly, we will use the following notation:

$$y(t) = Ce^{At}x(0) + \int_0^T Ce^{A(t-\tau)}Bu(w(\tau))d\tau \doteq \\ \doteq L_0(x(0)) + L_t(u(w)) \tag{9}$$

in particular, notice that L_t is a convolution operator, $L_t(u) = F * u$ where F is the transfer function. We first address the problem where A, B, C (and therefore L_t) are given. For simplicity we will neglect the initial condition, although it is easy to take it into account if so desired. In this case, we define the distance between the two time series

$$d_4(y_1, y_2) = \min \sum_{i=1}^{2} \int_0^T \|y_i(t) - L_t(u_i(t))\| + \\ + \lambda \|u_i(t) - u_0(w_i(t))\| dt \tag{10}$$

subject to $u_0 \in \mathcal{H}$ and $w_i \in \mathcal{U}$. Note that we have introduced an auxiliary variable u_0, which implies a possible discrepancy between the actual input and the warped version of the common template. This problem can be solved in two steps: A deconvolution, where u_i are chosen to minimize the first term, and a standard dynamic time warping, where w_i and u_0 are chosen to minimize the second term. Naturally the two can be solved simultaneously.

3.1. Going blind

When the model parameters A, B, C are common to the two models, but otherwise unknown, minimization of the first term corresponds to blind system identification, which in general is ill-posed barring some assumption on the class of inputs u_i. These can be imposed in the form of generic regularizers, as common in the literature of blind deconvolution [7]. This is a general and broad problem, but beyond our scope here, so we will forgo it in favor of an approach where the input is treated as the output of an auxiliary dynamical model, also known as *exo-system* [8]. This combines standard DTW, where the monotonicity constraint is expressed in terms of a double integrator, with TWDC, where the actual stationary component of the temporal dynamics is estimated as part of the inference. The generic warping w, the output of the exo-system satisfies

$$\begin{cases} \dot{w}_i(t) = \rho_i(t), \quad i = 1, 2 \\ \dot{\rho}_i(t) = v_i(t)\rho_i(t) \end{cases} \tag{11}$$

and $w_i(0) = 0,\ w_i(T) = T$. This is a multiplicative double integrator; one could conceivably add layers of random walks, by representing v_i are Brownian motion. Combining this with the time-invariant component of the realization yields the generative model for the time series y_i:

$$\begin{cases} \dot{w}_i(t) = \rho_i(t), \quad i = 1, 2 \\ \dot{\rho}_i(t) = v_i(t)\rho_i(t) \\ \dot{x}_i(t) = Ax_i(t) + Bu(w_i(t)) \\ y_i(t) = Cx_i(t) + n_i(t). \end{cases} \tag{12}$$

Note that the actual input function u, as well as the model parameters A, B, C, are common to the two time series. A slightly relaxed model, following the previous subsection, consists of defining $u_i(t) \doteq u(w_i(t))$, and allowing some slack between the two; correspondingly, to compute the distance one would have to minimize the data term

$$\phi_{data}(y_1, y_2 | u, w_i, A, B, C) \doteq \sum_{i=1}^{2} \int_0^T \|n_i(t)\|^2 dt \tag{13}$$

subject to (12), in addition to the regularizers

$$\bar{\phi}_{reg}(v_i, u) = \sum_{i=1}^{2} \int_0^T \|v_i(t)\|^2 + \|\nabla u(t)\|^2 dt \tag{14}$$

which yields a combined optimization problem

$$d_5(y_1, y_2) = \min_{u \in \mathbb{L}^2, A, B, C} \sum_{i=1}^{2} \int_0^T (\|y_i(t) - Cx_i(t)\|^2 + \\ + \|v_i(t)\|^2 + \|\nabla u(t)\|^2) dt \tag{15}$$

subject to (12). This distance can be either computed in a globally optimal fashion on a discretized time domain using dynamic programming, or more simply we can run a gradient descent algorithm based on the first-order optimality conditions. In the next section we report experiments on real and simulated data that illustrate the power and limitations of the approach proposed.

4. Experiments

We performed several experiments on controlled synthetic datasets (not reported for reasons of space) and with publicly available sets of real data, both from motion capture and from image sequences, for both stationary and non-stationary actions. Here we report a set of representative results that illustrate the characteristics of our approach as it compares with DTW and other published results on such public datasets; some of the experiments reported also highlight the limitation of our approach.

4.1. Stationary sequences

In order to set a baseline and compare our approach against existing ones, we have first used the popular CMU MoCap Dataset for the case of quasi-periodic sequences. The data is provided as a set of joint angle trajectories on a skeletal model of the human body, obtained by a motion capture system. It contains instances of 23 individuals walking and running. We restrict our observation period to just one walking cycle and we perform a variance normalization [21]. Of all the joint angle trajectories,for simplicity, we have selected a subset of 6, corresponding to lower-body joints. Despite these decimations, the correct classification rate using a simple nearest-neighbor classifier based on TWDC was 100%. The resulting confusion matrix (pairwise distances between each data pair, organized into a matrix, with dark intensity indicating low distance) is shown in Fig. 2. The sequences have been organized so that walking sequences occupy the upper quadrant, whereas running sequences are in the lower one.

In order to make a baseline comparison with DTW, we generate a score by looking at the first k nearest neighbors, and summing the number of classification errors based on the k-th neighbor, instead of the nearest, with k going from 2 to 10. DTW achieves a cumulative score of 25, whereas with TWDC it was 13. Although it should be obvious *a-priori* that our approach should improve on DTW, because it includes it as a subset, this simple experiment suffices to validate this hypothesis.

In the next experiment we used a more challenging dataset, provided by UCLA [2], which includes

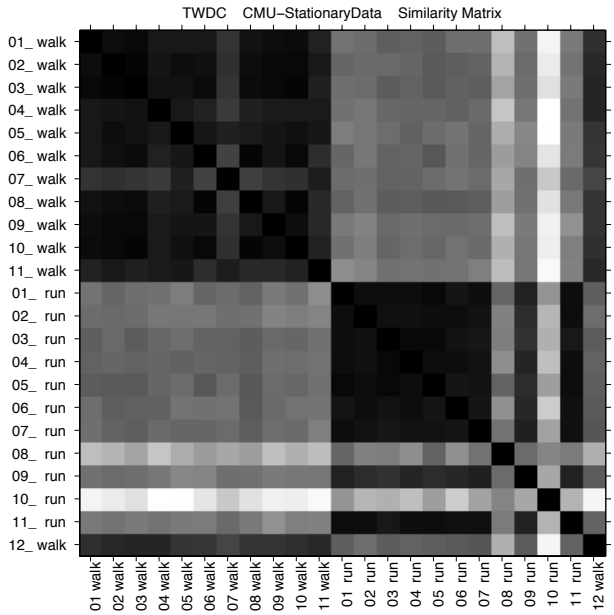

Figure 2. *Confusion matrix for walking/running classification in the CMU MoCap dataset. The distance between 12 walking and 11 running sequences is visualized as an intensity value, with dark being small and light being large. Classification based on first nearest neighbor yields 100% correct. To compare with DTW, we sum the classification errors obtained by using the k-th neighbor, instead of the nearest neighbor, with $k = 2, \ldots, 10$. DTW achieves a score of 25, whereas TWDC performs better with a score of 13.*

sequences of limping that are more subtle and hence harder to discriminte from walking. Our approach outperforms both DTW as well as the results reported by [2] in most cases. The confusion matrix is shown in Fig. 3 and the following table shows correct classification performance for the three actions available: (walking, running, limping).

Comparison of gait classification performances in k-nearest neighbor matching

Model Used	k = 3	k = 5	k = 7
DTW	(63.6%, 63.6%, 0)	(63.6%, 63.6%, 0)	(63.6%, 63.6%, 0)
[2]	(86.0%, 98.7%, 15.0%)	(88.6%, 98.7%, 15.0%)	(**93.9%**, 98.7%, 17.5%)
TWDC	(90.9%, **100%**, 33.3%)	(90.9%, **100%**, 33.3%)	(90.9%, **100%**, **44.4%**)

The table shows the percent correct classification for three actions (walking, running, limping), for k-nearest neighbor classification with $k = 3, 5, 7$ using DTW, TWDC and the results reported by [2]. Our approach outperforms both approaches on average, and in every category and nearest neighbor count, except for walking with $k = 7$ where [2] performs better. Note, however, that such an approach relies on the stationarity assumption, and therefore fails to perform on the scenarios considered in the next subsection.

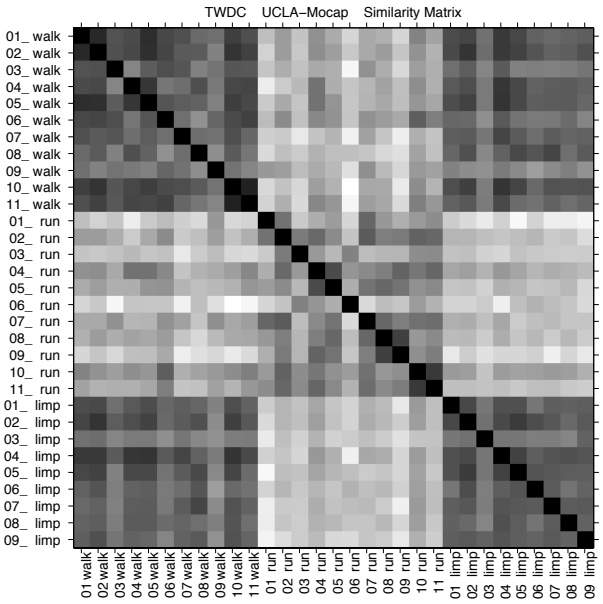

Figure 3. *Confusion matrix for the UCLA dataset [2]. There are 31 sequences of (walking, running, limping). Classification performance using k nearest neighbor is reported in the previous table.*

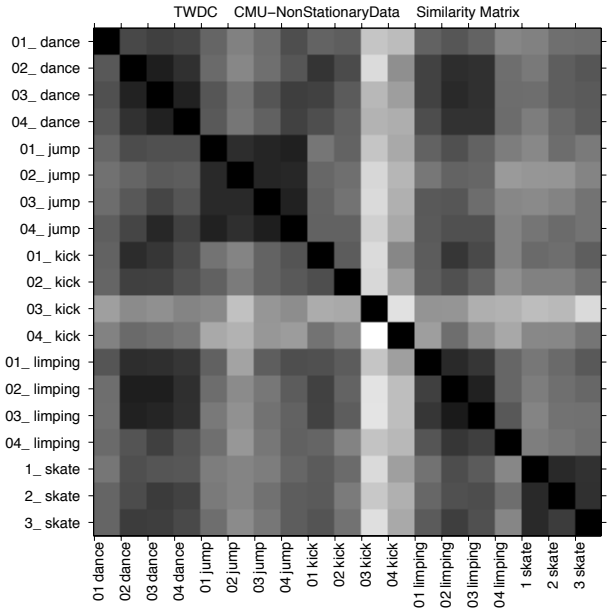

Figure 4. *Confusion matrix for 19 sequences and 5 different activities (dancing, jumping, kicking, limping, skating) from the CMU MoCap dataset. Numerical classification scores are reported in the previous table.*

4.2. Non-stationary sequences

In this section we put to the test the functioning of our approach on sequences of non-stationary actions such as dancing, jumping, kicking, limping and skating, also taken from the CMU MoCap dataset. Here, algorithms that rely on the assumption of stationarity cannot be employed; because we are not aware of other published results on this section of the MoCap dataset, we simply report our results, summarized in the confusion matrix in Fig. 4 as well as the following table.

Results obtained in the classification of non stationary signals of the MoCap Dataset in k-nearest neighbor matching (k=3)

Dance	Jump	Kick	Limp	Skate
100%	100%	50%	25%	100%

The table shows that some actions are rather simple to classify. Others, however, are more subtle, for instance limping and kicking. The latter in particular is quite short, so partial matching with other actions (such as dance) reveals considerable similarity that only photometric context (e.g. the presence of a ball) can disambiguate. Furthermore, as we discuss in the next section, there are range (scale) transformation that we do not model explicitly – for we decide to concentrate on time domain transformations – that play a confounding role in classification.

In the next experiment we tested our algorithm on the dataset presented by [3], that consists in binary sequences of images obtained from background subtraction from a stationary camera pointed in front of a scene where subjects were performing a series of non-stationary actions. Direct comparison with the metric used by [3] is not possible, since they employ a representation that compounds temporal information via averaging, rather than by warping. Therefore, classification results are affected by the representation as well as by the metric, and there is no way to disentangle the two. Nevertheless, we can compare the overall classification results, summarized in the confusion matrix in Fig. 5. For simplicity we have used only 3 coarse features, corresponding to the height and to the width of upper and lower part of a bounding box of the silhouette . Despite this brutal simplification, we achieve classification rates comparable with [3] summarized in the following table.

Results obtained in the classification of non stationary signals of the Weizmann Database in k-nearest neighbor matching (k=3)

Run	66.7%	**Jump**	55.6%	**Wave1**	77.8%
Walk	88.9%	**pJump**	33.4%	**Wave2**	100%
Side	88.9%	**Jack**	100%	**Bend**	100%

Additionally, our results are obtained without any particular attention to spatial modeling to spatial model-

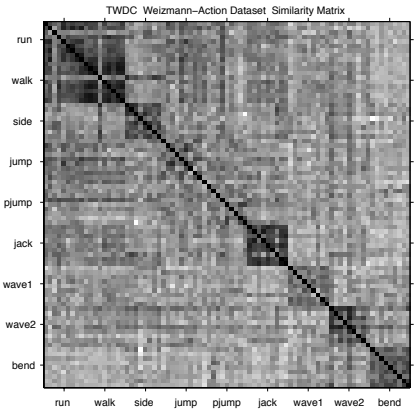

Figure 5. *Confusion matrix for 81 sequences and 9 different activities (run, walk, side, jump, pjump, jack, wave1, wave2, bend) from the Weizmann dataset. Numerical classification scores are reported in the previous table.*

ing and normalization, just by considering deformations of the temporal axis. Clearly, there is a lot to be gained from the use of more sophisticated representations such as those used by [3, 24], but this simple experiment suffices to validate the flexibility and power of our approach on image-based data sets.

5. Discussion

We have introduced "time warping under dynamic constraints" that is an optimization scheme to find the time domain deformation of a time series that best fits another time series while respecting their intrinsic dynamics. This is achieved by modeling each time series by a different realization of a dynamical model driven by time-warped versions of the same unknown input. We have illustrated the relationship to standard "dynamic time warping" and shown empirically that classification performance is improved when dynamic constraints are taken into account.

On publicly available datasets, our approach mostly performed as well as expected. On some tasks, however, performance was unexpectedly low. This was due in part to sampling issues, as some of the non-stationary actions were available over a long sequence, whereas others were available in short snippets. Another shortcoming of our approach, just because we have decided to focus on time domain deformations, is the fact that we do not explicitly model range (amplitude) transformations. It is obvious that these should also be taken into account, and there are many ways to do so, depending on the domain and on the data representation selected. For instance, for image sequences one can pre-process them to normalize for contrast scalings, and again one could do so as part of the representation or as part of the matching process. This of course is not an issue on binary images since the range is normalized, but it is an issue on geometric conversions of the silhouette where there are, for instance, scale or affine variations. Again, one could employ affine-invariant representations (e.g. Fourier-Mellin moments) or optimize with respect to the best matching affine transformation during matching (or average with respect to a given procrustean distribution).

Our contribution is obviously only a piece of the puzzle of building an effective, robust and reliable machine to classify actions and events from video, but we feel that our handling of time domain transformations is well suited for this task as it represents a sound tradeoff between the simplicity of the model and its flexibility: It is not as simple as simple linear stationary models, but it is not as general as fully non-linear models of chaotic dynamics that are employed in other disciplines such as finance or astrophysics.

References

[1] J. Alon, S. Sclaro, G. Kollios, and V. Pavlovic. Discovering clusters in motion time-series data, 2003. 2

[2] A. Bissacco. Modeling and learning contact dynamics in human motion. June 2005. 2, 5, 6

[3] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri. Actions as Space-Time Shapes. *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, 2, 2005. 6, 7

[4] M. Brand. Subspace mappings for image sequences. In *Proc. Workshop Statistical Methods in Video Processing*, June 2002. 2

[5] A. B. Chan and N. Vasconcelos. Probabilistic kernels for the classification of auto-regressive visual processes. 2005. 2

[6] K. De Coch and B. De Moor. Subspace angles and distances between arma models. *Proc. of the Intl. Symp. of Math. Theory of Networks and Systems*, 2000. 1

[7] B. Giannakis and J. Mendel. Identification of nonminimum phase systems using higher order statistics. *IEEE Trans. on Acoustic, Speech and Signal Processing*, 37(3):360–377, 1989. 4

[8] A. Isidori. *Nonlinear Control Systems*. Springer Verlag, 1989. 4

[9] G. Johansson. Visual perception of biological motion and a model for its analysis. 14:201–211, 1973. 1

[10] H. Kantz and T. Schreiber. *Nonlinear Time Series Analysis*. Cambridge University Press, 2004. 1

[11] E. J. Keogh and M. J. Pazzani. Dynamic time warping with higher order features. In *Proceedings of the 2001 SIAM Intl. Conf. on Data Mining*, 2001. 2

[12] A. Kirsch. An introduction to the mathematical theory of inverse problems. *Springer-Verlag, New York*, 1996. 3

[13] C. F. Martin, S. Sun, and M. Egerstedt. Optimal control, statistics and path planning. 1999. 3

[14] R. Martin. A metric for arma processes. *IEEE Trans. on Signal Processing*, 48(4):1164–1170, 2000. 1

[15] M. E. Munich and P. Perona. Conitnuous dynamic time warping for translation-invariant curve alignment with applications to signature verification. In *ICCV*, 1999. 2

[16] V. Pavlovic and J. Rehg. Impact of dynamic model learning on classification of human motion. pages 788–795. 2

[17] M. Pollefeys. 3D model from images. *ECCV tutorial lecture notes, Dublin, Ireland, 2000*, 2000. 2

[18] J. O. Ramsey and B. W. Silverman. *Functional Data Analysis*. Springer Verlag, 2005. 1, 2, 3

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

[20] Yang Song, Xiaolin Feng, and Pietro Perona. Towards detection of human motion. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2000)*, pages 810–817, 2000. 2

[21] R. Tanawongsuwan and A. Bobick. Gait recognition from time-normalized joint-angle trajectories in the walking plane. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pages 726–731, 2001. 5

[22] Yaron Ukrainitz and Michal Irani. Aligning sequences and actions by maximizing space-time correlations. In *ECCV (3)*, pages 538–550, 2006. 2

[23] A. Veeraraghavan, R. Chellappa, and A. K. Roy-Chowdhury. The function space of an activity. 2006. 2

[24] Ashok Veeraraghavan, Amit K. Roy-Chowdhury, and Rama Chellappa. Matching shape sequences in video with applications in human movement analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(12):1896–1909, 2005. 7

[25] S.V.N. Vishwanathan, R. Vidal, and A. J. Smola. Binet-cauchy kernels on dynamical systems and its application to the analysis of dynamic scenes. *International Journal of Computer Vision*, 2005. 2

[26] A. D. Wilson and A. F. Bobick. Parametric hidden markov models for gesture recognition. In *IEEE Trans. on Pattern Analysis and Machine Intelligence*, volume 21(9), pages 884–900, Sept. 1999. 2

[27] Lihi Zelnik-Manor and Michal Irani. Statistical analysis of dynamic actions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(9):1530–1535, 2006. 2