Adaptive Down-Sampling and Dimension Reduction in Time Elastic Kernel Machines for Efficient Recognition of Isolated Gestures

Pierre-François Marteau, Sylvie Gibet, Clément Reverdy

Adaptive Down-Sampling and Dimension Reduction in Elastic Kernel Machines for Efficient Recognition of Isolated Gestures

Pierre-Francois Marteau and Sylvie Gibet and Clément Reverdy

Abstract

In the scope of gestural action recognition, the size of the feature vector representing movements is in general quite large especially when full body movements are considered. Furthermore, this feature vector evolves during the movement performance so that a complete movement is fully represented by a matrix M of size DxT, whose element $M_{i,j}$ represents the value of feature i at timestamps j. Many studies have addressed dimensionality reduction considering only the size of the feature vector lying in $\mathbb{R}^D$ to reduce both the variability of gestural sequences expressed in the reduced space, and the computational complexity of their processing. In return, very few of these methods have explicitly addressed the dimensionality reduction along the time axis. Yet this is a major issue when considering the use of elastic distances which are characterized by a quadratic complexity along the time axis. We present in this paper an evaluation of straightforward approaches aiming at reducing the dimensionality of the matrix M for each movement, leading to consider both the dimensionality reduction of the feature vector as well as its reduction along the time axis. The dimensionality reduction of the feature vector is achieved by selecting remarkable joints in the skeleton performing the movement, basically the extremities of the articulatory chains composing the skeleton. The temporal dimensionality reduction is achieved using either a regular or adaptive down-sampling that seeks to minimize the reconstruction error of the movements. Elastic and Euclidean kernels are then compared through support vector machine learning. Two data sets

Pierre-Francois Marteau
IRISA (UMR 6074), Université de Bretagne Sud Campus de Tohannic, 56000 Vannes, France, e-mail: `firstname.nameATuniv-ubsDOTfr`

Sylvie Gibet
IRISA (UMR 6074), Université de Bretagne Sud Campus de Tohannic, 56000 Vannes, France, e-mail: `firstname.nameATuniv-ubsDOTfr`

Clément Reverdy at IRISA (UMR 6074), Université de Bretagne Sud Campus de Tohannic, 56000 Vannes, France, e-mail: `firstname.nameATuniv-ubsDOTfr`

that are widely referenced in the domain of human gesture recognition, and quite distinctive in terms of quality of motion capture, are used for the experimental assessment of the proposed approaches. On these data sets we experimentally show that it is feasible, and possibly desirable, to significantly reduce simultaneously the size of the feature vector and the number of skeleton frames to represent body movements while maintaining a very good recognition rate. The method proves to give satisfactory results at a level currently reached by state-of-the-art methods on these data sets. We experimentally show that the computational complexity reduction that is obtained makes this approach eligible for *real-time* applications.

1 Introduction

Gesture recognition is a challenging task in the computer vision community with numerous applications using motion data such as interactive entertainment, human-machine interaction, automotive, or digital home. Recently, there is an increasing availability of large and heterogeneous motion captured data characterized by a various range of qualities, depending on the type and quality of the motion sensors. We thus separate (i) databases of high resolution and quality built from expensive capturing devices and requiring a particular expertise, (ii) and low resolution and noisier databases produced with cheap sensors that do not require any specific expertise. Such databases open new challenges for assessing the robustness and generalization capabilities of gesture recognition algorithms on diversified motion data sets. Besides the quality of recognition, the complexity of the algorithms and their computational cost is indeed a major issue, especially in the context of real-time interaction.

We address in this paper the recognition of isolated gestures from motion captured data. As motion data are generally represented by high-dimensional time series, many approaches have been developed to reduce their dimension, so that the recognition process is more efficient while being still accurate. Among them, low-dimensional embeddings of motion data have been proposed that enable to characterize and parameterize action sequences. Some of them are based on statistical descriptors [Hussein et al., 2013], rely on relevant meaningful trajectories [Ofli et al., 2013], or characterize the style[Hussein et al., 2013]. In this paper we focus on dimensionality reduction along two complementary axes: the spatial axis representing the configuration of the skeleton at each frame, and the temporal axis representing the evolution over time of the skeletal joints trajectories. With such an approach, two main challenges are combined simultaneously:

- We use relevant trajectories (end-extremities) whose content may characterize complex actions;
- Considering that the temporal variability is of primary importance when recognizing skilled actions and expressive motion, we apply an adaptive temporal down-sampling to reduce the complexity of elastic matching.

These low-dimensional dual-based representations will be coupled with appropriate recognition algorithms that we expect to be more tractable and efficient.

Our recognition principle is based on a recent method that improves the performance of classical support vector machines when used with regularized elastic kernels dedicated to time series matching. Our objective is to show how the spatial and temporal dimensionality reductions, associated with such regularized elastic kernels significantly improve the efficiency, in terms of response time, of the algorithm while preserving the recognition rates.

The second section briefly presents the related works and state-of-the-art for isolated gesture recognition. In the third section we describe the nature of the motion data as well as the main pre-processing of the data. The fourth section gives the major keypoints of the method, positioning it in the context of multivariate sequential data classification. We present in the fifth part the evaluation of our algorithm carried out on two data sets with very distinct qualities and compare its performance with those obtained by some of the state-of-the-art methods. A final discussion is provided as well as some perspectives.

2 Related work

In human gesture recognition, there are two main challenges to be addressed: dimension reduction closely linked to feature descriptors, and recognition models which cover different aspects of dynamic modeling, statistical and machine learning approaches. In this section, we give a brief and non-exhaustive overview of the literature associated with each challenge.

Dimension reduction

The problem of dimensionality reduction (also called manifold learning) can be addressed with the objective to find a low-dimensional space that best represents the variance of the data without loosing too much information.Action descriptors have thus been defined for characterizing whole motion sequences, or punctual frames that need additional step of temporal modeling to achieve the recognition goal.

Numerous method are available to carry out such dimensionality reduction, the most popular being linear approaches such as Principal Component Analysis (PCA, [Jolliffe, 1986], [Masoud and Papanikolopoulos, 2003]), Linear Discriminant Analysis (LDA, [McLachlan, 2004]), or linear projections preserving locally neighborhoods (Locality Preserving Projection) [He and Niyogi, 2003]. Among nonlinear approaches, Locally Linear Embeddings (LLE, [Roweis and Saul, 2000]), Metric Multidimensional Scaling (MDS, [Kruskal and Wish, 1978]) and variants like Laplacian Eigenmap [Belkin and Niyogi, 2002], Isomap [Tenenbaum et al., 2000] have been implemented to embed postures in low dimensional spaces in which a more efficient time warp (DTW, see section 5) algorithm [?] can be used to classify movements. An extension of this method, called ST-Isomap, considers temporal relationships in local neighborhoods that can be propagated globally via a

shortest-path mechanism [Jenkins and MatariÄĞ, 2004]. Models based on Gaussian processes with latent variables are also largely used, for instance a hierarchical version has been recently exploited for gesture recognition [Han et al., 2010].

Other methods define discriminative features that best classify motion classes. This is the case in the work of [Yu and Aggarwal, 2009] that reduces the motion data to only five end-extremities of the skeleton (two feet, two hands and the head), thus giving some meaningful insight of the motion related to the action task. [Fothergill et al., 2012] and [Zhao et al., 2012] have in particular applied random forests to recognize actions, using a Kinect sensor, while [Ofli et al., 2013] recently proposed to automatically select the most informative skeletal joints to explain the current action. In the same line, [Hussein et al., 2013] use the covariance matrix for skeleton joint locations over time as a discriminative descriptor to characterize a movement sequence. Multiple covariance matrices are deployed over sub-sequences in a hierarchical fashion in order to encode the relationship between joint movement and time. In [Li et al., 2010] a simple bag of 3D points is used to represent and recognize gestural action. Similarly, in [Wang et al., 2012], *actionlets* are defined from Fourrier coefficients to characterize the most discriminative joints. Finally, it can be mentioned, among many existing applications that address the use of elastic distances into a recognition process, the recent work described in [Sempena et al., 2011], as well as the hardware acceleration proposed in [Hussain and Rashid, 2012]. However, to our knowledge, no work exploiting this type of distance has directly studied the question of data reduction along the time axis.

Gesture recognition Recognition methods essentially aim at modeling the dynamics of gestures. Some approaches, based on linear dynamic models [Veeraraghavan et al., 2004], have used autoregressive (AR) and autoregressive moving-average (ARMA) models to characterize the kinematics of movements, while other approaches, based on nonlinear dynamic models [Bissacco et al., 2007], have developed movement analysis and recognition scheme based on dynamical models controlled by Gaussian processes. [Mitra and Acharya, 2007] propose a synthesis of the major gesture recognition approaches relying on Hidden Markov Models (HMM). Histograms of oriented 4D normals have also been proposed in [Oreifej and Liu, 2013] for the recognition of gestural actions from sequences of depth images. [Wang et al., 2006] have exploited conditional random fields to model joint dependencies and thus increase the discrimination of HMM-like models. Recurrent neural network models have also been used [Martens and Sutskever, 2011]; among them, conditional restricted Boltzman's machines [Larochelle et al., 2012] have been studied recently in the context of motion captured data modeling and classification.

In this paper, we propose a new representation of human actions that results from a dual-based reduction method that occurs both spatially and temporally. We couple this representation to a SVM classification method associated with regularized elastic kernels.

3 Motion representation

We are working on isolated human motions acquired through motion captured databases. In recent years, there is an increasing availability of these databases, some of them being captured by high resolution devices (infrared marker-tracking system), such as those provided by CMU [cmu, 2003], HDM05 [Müller et al., 2007], and other ones captured by low-cost devices, such as MSRAction3D captured with the Microsoft Kinect system. With the first type of sensors, the acquisition process is expensive, as it necessitates for capturing skilled motion with a good accuracy several cameras with many markers located on an actor, and a post-processing pipeline which is costly in time and expertise. Instead, with the second type of sensors, the data acquisition is affordable and necessitates less time and expertise, but with a loss of accuracy which remains acceptable for some tasks.

After the recording, the captured data is filtered and reconstructed so that to eliminate most of the noise, data loss and labelling errors (markers inversion), and the output of this acquisition pipeline is generally a set of 3D-trajectories of the skeleton joints determined from the positions of markers. This kind of data is inherently noisy, mainly due to the quality of the sensors and the acquisition process, but also to approximations made during the reconstruction process. The modeling of the skeleton is indeed subject to some variation for different reasons: in particular the markers being positioned on cloths or on the skin of the actor's body, they can move during the capturing process, also the determination of the segment lengths, of the joints' centers and their associated rotation axis are not trivial and lead to modeling errors. To overcome these difficulties, the skeleton model is obtained through an optimization process such as the ones described in [de Aguiar et al., 2006], [O'Brien et al., 2000], or [Shotton et al., 2011]. The techniques based on a skeleton model hence convert 3D sensor data into Cartesian or angular coordinates that define the state of the joints over time with various accuracies.

Figure 1 presents two skeletons reconstructed from two very distinct capture systems. On the right, the skeleton is reconstructed from data acquired via the Microsoft Kinect, on the left from the Vicon-MX device used by the Max Planck Institute to produce the HDM05 datasets.

Thus, any skeletal-based model can be represented by a hierarchical tree-like structure composed of rigid segments (bones) connected by joints. A motion can be defined by a sequence of postures over time, formalized as a multivariate state vector describing a trajectory over time, i.e. a time series: $\{Y_t \in \mathbf{R}^k\}_1^T = [Y_1, ..., Y_T]$, where the k spatial dimension ($k = 3 \cdot N$, with N the number of joints) typically varies between 20 and 100 according to the capture devices and the considered task. As this state vector is obviously not composed of independent scalar dimensions, the spatio-temporal encoded redundancies open solutions for dimension reduction as well as noise reduction approaches. This is particularly relevant for motion recognition, as the objective is to aim at improving computation time and error rate.
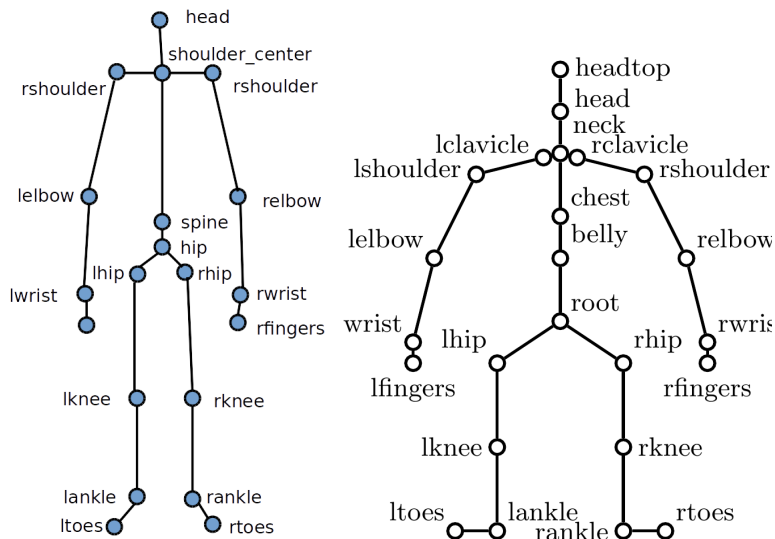
Fig. 1 Examples of skeletons reconstructed from motion data captured from the Kinect (left) and from the Vicon-MX device used by the Max Planck Institute (right).

4 Dimension reduction of motion capture data

Using elastic distances or kernels for recognition problems has proved to be very accurate and efficient. However, one of the main difficulty of such methods with time series data is to deal with their computational cost, in general quadratic with the length of the time series and linear with the *spatial* dimension characterized by the number of degrees-of-freedom. This high computational complexity is potentially limiting their use, especially when large amounts of data have to be processed, or when real-time constraint is required. We therefore expect that a dual dimensionality reduction, both on the time and spatial axis is particularly relevant, especially when associated to these techniques. Hence, for motion recognition using elastic distances, we propose to show that there exists a spatio-temporal redundancy in motion data that can be exploited.

Considering the quite rich literature on motion recognition, it appears that while some studies have shown success with dimensionality reduction on the spatial axis, very few have directly addressed a dimensionality reduction along the time axis , and much less work by combining the two approaches. [Keogh and Pazzani, 2000] has however proposed a temporal sub-sampling associated with dynamic time warping in the context of time series mining, followed later by [Marteau, 2009]. We address herein after these two lines of research.

4.1 Dimension reduction along the spatial axis

Dimension reduction (or manifold learning), is the process consisting of mapping high dimensional data to representations lying in a lower-dimensional manifold. This can be interpreted as mapping the data into a space characterized by a smaller dimension, from which most of the variability of the data can be reproduced.

We consider in this paper a more direct approach based on the knowledge of the mechanism underlying the production of motion data and the way human beings perceive and discriminate body movements. We make the assumption that the perception of human motion is better achieved in the so-called *task-space* represented by a selection of significant 3D joint trajectories. This hypothesis is supported by [Giese et al., 2008] who show that visual perception of body motion closely reflects physical similarities between joint trajectories. This is also consistent with the motor theory of motion perception presented in [Gibet et al., 2011]. Besides, we may reasonably accept that these joint trajectories embed sufficient discriminative information as inverse kinematics (widely used in computer animation and robotics) has shown to be very efficient and robust to reconstruct the whole skeleton movement from the knowledge of the end effector trajectories (hands, feet, head), possibly with the additional knowledge of mid-articulated joints trajectories (such as elbows and knees) and constraints. Hence a straightforward approach consists in constructing a motion descriptor that discards all joints information but the 3D positions of the mid and end effectors extremities. We thus select the 3D positions for the two wrists, the two ankles (the fingers and toes markers are less reliable in general), the two elbows, the two knees and the head. This leads to a time-varying descriptor lying in a 18D space, while a full body descriptor is embedded in a 60D, space for the Kinect sensor, significantly more for vicon settings in general.

4.2 Dimension reduction along the time axis

The straightforward approach we are developing to explicitly reduce dimensionality along the time axis consists in sub-sampling the movement data such that each motion trajectories takes the form of a fixed-size sequence of L skeletal postures. Then it becomes easy to perform a classification or recognition task by using elastic kernel machines on such fixed-size sequences. With such an approach, performance rates depend on the chosen degree of sub-sampling. This approach seems coarse since long sequences are characterized with the same number of skeletal poses than short sequences. For very short sequences, whose lengths are shorter than L, if any, we over-sample the sequence to meet the fixed-size requirement. But we consider this case as very marginal since we seek a sub-sampling rate much lower than the average sequence of movement length. In the following, we will experiment and compare uniform and adaptive down-sampling.

4.2.1 Uniform down-sampling

In order to explicitly reduce dimensionality along the time axis, our first straightforward approach here consists in down-sampling the motion data so that each motion trajectory takes the form of a fixed-size sequence of L skeletal postures, evenly distributed along the time axis. We refer to this approach as uniform down-sampling (UDS).

4.2.2 Adaptive down-sampling

The second approach is the so-called adaptive down-sampling (ADS) approach. Similarly to UDS, each motion trajectory takes the form of a fixed-size sequence of L skeletal postures, but these postures are not evenly distributed anymore along the time axis. They are selected such as to minimize a trajectory reconstruction criteria. Basically we follow the previous work by [Marteau and Gibet, 2006]. A data modeling approach is used to handle the adaptive sampling of the $\{Y_t\}$ multidimensional and discrete time series. More precisely, we are seeking an approximation Y_{θ^*} of Y such as:

$$\theta^* = \underset{\theta}{ArgMin}(E(\{Y_t\},\{Y_\theta,t\})) \tag{1}$$
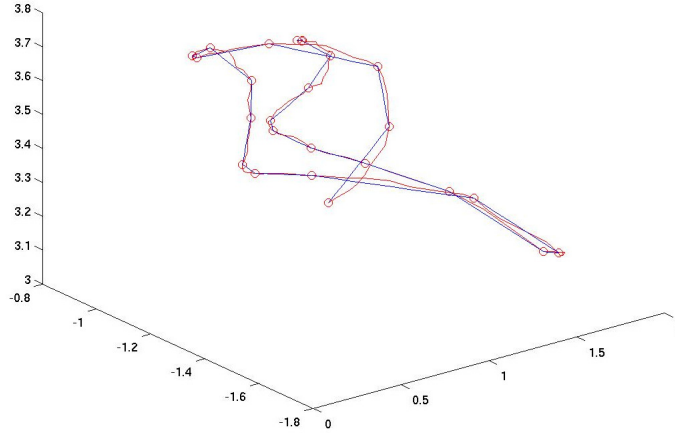
Fig. 2 Trajectory of the human wrist in the 3D Cartesian space adaptively dow-sampled with the localization of the 25 selected samples (red circles): motion capture data (blue) and reconstructed data by linear interpolation (red).

where E is the RMS error between Y and Y_θ selected among the set of piecewise linear approximations defined from Y. Since the optimal solution of the optimization problem defined by Eq. 1 is $O(L.n^2)$, where n is the length of the input time series,

we adopt a near to optimal solution as developed in [Marteau and Ménier, 2009] whose complexity is $O(n)$. As an example, in figure 2 the human wrist 3D trajectory is down-sampled using 25 samples positioned along the trajectory by minimizing the piecewise linear approximation.

5 Elastic kernels and their regularization

Dynamic Time Warping (DTW), [Velichko and Zagoruyko, 1970], [Sakoe and Chiba, 1971], by far the most used elastic measure, is defined as

$$d_{dtw}(X_p, Y_q) = d_E^2(x(p), y(q)) \tag{2}$$
$$+ \text{Min} \begin{cases} d_{dtw}(X_{p-1}, Y_q) & sup \\ d_{dtw}(X_{p-1}, Y_{q-1}) & sub \\ d_{dtw}(X_p, Y_{q-1}) & ins \end{cases}$$

where $d_E(x(p), y(q))$ is the Euclidean distance (possibly the square of the Euclidean distance) defined on $\mathbb{R}^k$ between the two postures in sequences X and Y taken at times p and q respectively.

When performed by a support vector machine (SVM) model, the optimization problem inherent to this type of learning algorithm is no longer quadratic. Moreover, the convergence towards the *optimorum* is no longer guaranteed, which, depending on the complexity of the task may be considered as detrimental.

Besides the fact that the DTW measure does not respect the triangle inequality, it is furthermore not possible to directly define a positive definite kernel from it. Hence, the optimization problem, inherent to the learning of a kernel machine, is no longer quadratic which could, at least on some tasks, be a source of limitation.

Regularized DTW: recent works [Cuturi et al., 2007], [Marteau and Gibet, 2014] allowed to propose new guidelines to regularize kernels constructed from elastic measures such as DTW. A simple instance of such regularized kernel, derived from [Marteau and Gibet, 2014] for time series of equal length, takes the following form, which relies on two recursive terms :

$$K_{rdtw}(X_p, Y_q) = K_{rdtw}^{xy}(X_p, Y_q) + K_{rdtw}^{xx}(X_p, Y_q)$$

$$K_{dtw}^{xy}(X_p, Y_q) = \frac{1}{3} e^{-\nu d_E^2(x(p), y(q))}$$
$$\sum \begin{cases} h(p-1, q) K_{rdtw}^{xy}(X_{p-1}, Y_q) \\ h(p-1, q-1) K_{rdtw}^{xy}(X_{p-1}, Y_{q-1}) \\ h(p, q-1) K_{rdtw}^{xy}(X_p, Y_{q-1}) \end{cases} \tag{3}$$

$$K_{rdtw}^{xx}(X_p, Y_q) = \frac{1}{3}$$
$$\sum \begin{cases} h(p-1, q) K_{rdtw}^{xx}(X_{p-1}, Y_q) e^{-\nu d_E^2(x(p), y(p))} \\ \Delta_{p,q} h(p, q) K_{rdtw}^{xx}(X_{p-1}, Y_{q-1}) e^{-\nu d_E^2(x(p), y(q))} \\ h(p, q-1) K_{rdtw}^{xx}(X_p, Y_{q-1}) e^{-\nu d_E^2(x(q), y(q))} \end{cases}$$

where $\Delta_{p,q}$ is the Kronecker's symbol, $\nu \in \mathbb{R}^+$ is a *stiffness* parameter which weights the local contributions, i.e. the distances between locally aligned positions, and $d_E(.,.)$ is a distance defined on $\mathbb{R}^k$.

The initialization is simply $K_{rdtw}^{xy}(X_0, Y_0) = K_{rdtw}^{xx}(X_0, Y_0) = 1$.

The main idea behind this line of regularization is to replace the operators min and max (which prevent the symmetrization of the kernel) by a summation operator ($\sum$). This leads to consider, not only the best possible alignment, but also all the best (or nearly the best) paths by summing up their overall cost. The parameter ν is used to control what we call nearly-the-best alignment, thus penalizing more or less alignments too far from the optimal ones. This parameter can be easily optimized through a cross-validation.

5.1 Normalization

As K_{rdtw} evaluates the sum on all possible alignment paths of the products of local alignment costs $e^{-d_E^2(x(p), y(p))/(2.\sigma^2)} \leq 1$, its values can be very small depending on the size of the time series and the selected value for σ. Hence, K_{DTW} values tend to 0 when σ tends towards 0, except when the two compared time series are identical (the corresponding Gram matrix suffers from a diagonal dominance problem). As proposed in [Marteau and Gibet, 2014], a manner to avoid numerical troubles consists in using the following *normalized* kernel:

$$\tilde{K}_{rdtw}(.,.) = exp\left(\alpha \frac{log(K_{rdtw}(.,.)) - log(min(K_{rdtw}))}{log(max(K_{rdtw})) - min(K_{rdtw})}\right)$$

where $max(K_{rdtw})$ and $min(K_{rdtw})$ respectively are the max and min values taken by the kernel on the learning data set and $\alpha > 0$ a positive constant ($\alpha = 1$ by default). If we forget the proportionality constant, this leads to take the kernel K_{rdtw} at a power $\tau = \alpha/(log(max(K_{rdtw})) - log(min(K_{rdtw})))$, which shows that the normalized ker-

nel $\tilde{K}_{rdtw} \propto K_{rdtw}^{\tau}$ is still positive definite ([Berg et al., 1984], Proposition 2.7).

We consider in this paper the non definite exponential kernel (Gaussian or Radial Basis Function (RBF) types) $K_{dtw} = e^{-d_{dtw}(.,.)/(2.\sigma^2)}$ constructed directly from the elastic measures d_{dtw}, the normalized regularized elastic kernel K_{rdtw}^{τ}, and the non-elastic kernel obtained from the Euclidean distance [1], i.e., $K_E(.,.) = e^{-d_E^2(.,.)/\sigma}$.

6 Experimentation

To estimate the robustness of the proposed approaches, we evaluate them on two motion capture databases of opposite quality, the first one, called *HDM05*, developed at the Max Planck Institute, the other one, called *MSR-Action3D*, at Microsoft research laboratories.

HDM05 data set [Müller et al., 2007] consists of data captured at 120Hz by a Vicon MX system composed of a set of reflective optical markers followed by six high-definition cameras and configured to record data at 120hz. The movement sequences are segmented and transformed into sequences of skeletal poses consisting of N = 31 joints, each associated to a 3D position (x, y, z). In practice the position of the root of the skeleton (located near its center of mass) and its orientation serving as referential coordinates, only the relative positions of the remaining 30 joints are used, which leads to represent each position by a vector $Y_T \in \mathbb{R}^k$, with $k = 90$. We consider two recognition/classification tasks: HDM05-1 and HDM05-2 that are respectively those proposed in [Ofli et al., 2012] (also exploited in the work of [Hussein et al., 2013]) and [Ofli et al., 2013]. For both tasks, three subjects are involved during learning and two separate subjects are involved during testing. For task HDM05-1, 11 gestural actions are processed: *{deposit floor, elbow to knee, grab high, hop both legs, jog, kick forward, lie down floor, rotate both arms backward, sneak, squat, and throw basketball}*. This constitutes 249 motion sequences. For task HDM05-2 , the subjects are the same, but five additional gestural actions are considered in addition to the previous 11: *{jump, jumping jacks, throw, sit down, and stand up}*. For this task, the data set includes 393 movement sequences in total. For both tests, the lengths of the gestural sequences are between 56 and 901 postures (corresponding to a movement duration between 0.5-7.5 sec.) .

MSR-Action3D data set: This database [Li et al., 2010] has recently been developed to provide a Kinect data *benchmark*. It consists of 3D depth image sequences (*depth map*) captured by the Microsoft Kinect sensor. It contains 20 typical interaction gestures with a game console that are labeled as follows *high arm wave, horizontal arm wave, hammer, hand catch, forward punch, high throw, draw x, draw*

---

[1] The Euclidean distance is usable only because a fixed number of skeletal positions is considered to characterize each movement, and this, irrespectively of their initial length

tick, draw circle, hand clap, two hand wave, side-boxing, bend, forward kick, side kick, jogging, tennis swing, tennis serve, golf swing, pickup & throw . Each action was carried out by 10 subjects facing the camera, 2 or 3 times. This data set includes 567 motion sequences whose lengths vary from 14 to 76 skeletal poses. The 3D images of size 640×480 were captured at a frequency of 15hz. From each 3D image a skeletal posture has been extracted with $N = 20$ joints, each one being characterized by three coordinates. As for the previous data set, we characterize postures relatively to the referential coordinates located at the root of the skeleton, which leads to represent each posture by a vector $Y_t \in \mathbb{R}^k$, with $k = 3 \times 19 = 57$. The task is to provide a cross-validation on the subjects, i.e. 5 subjects participating in learning and 5 subjects participating in testing, considering all possible configurations which represent 252 learning/testing pairs in total.

Hence, we perform three classification tasks: HDM05-1, HDM05-2 and MSRAction3D, with or without a spatial dimensionality reduction while simultaneously considering a down-sampling on the time axis:

- The spatial dimensionality reduction is obtained by constructing a frame (skeletal pose) descriptor composed only with the end-effector trajectories in 3D (EED) comparatively to a full-body descriptor (FBD) that integrates all the joints trajectories that compose the skeleton. The FBD rests in a 90D space for HDM05 and in a 60D space for MSRAction3D. The EED rests in a 24D space (3D positions for 2 elbows, 2 hands, two knees and two feet) for the three tasks, leading to a data compression of 73% for HDM05 and 55% for MSRAction3D.

- The dimensionality reduction on the time axis is obtained through either a uniform down-sampling (UDS) or an adaptive down-sampling (ADS) based on a piecewise approximation of the FBD or EED trajectories. The number of skeletal poses varies from 5 to 30 for each trajectories, leading to an average data compression of 97% for HDM05 and 74% for MSRAction3D.

Fig. 3 Classification accuracies for HDM05-1 task as defined in [Ofli et al., 2012] uniform down-sampling, full body (top), adaptive down-sampling, full body (middle), adaptive down-sampling, end effector extremities (bottom), when the number of skeletal poses varies: K_E (red, circle, dash), K_{dtw} (black, square, plain), K_{rdtw}^{τ} (blue, star, dotted).
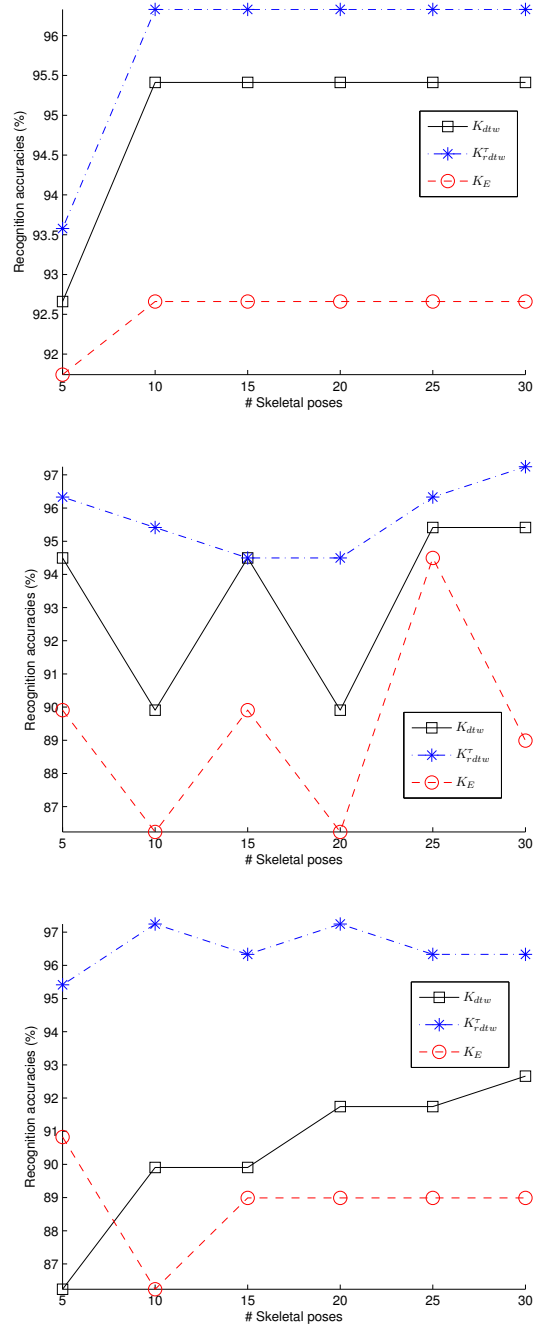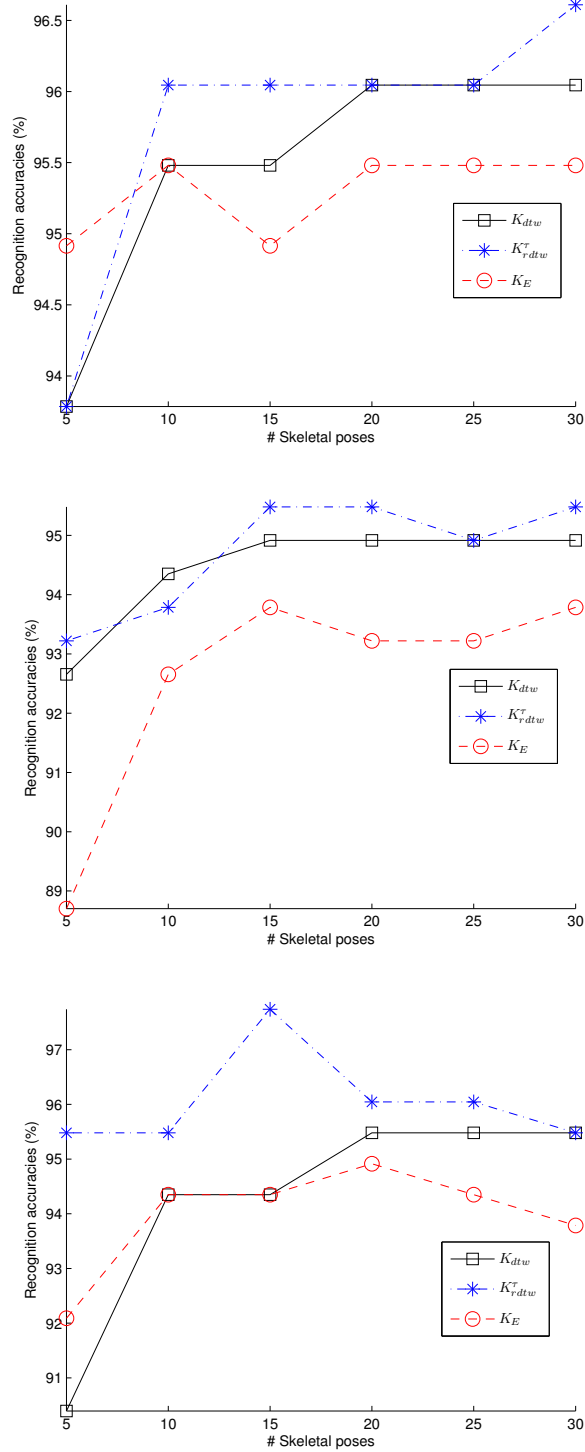
Fig. 4 Classification accuracies for HDM05-2 task as defined in [Ofli et al., 2013] uniform down-sampling, full body (top), adaptive down-sampling, full body (middle), adaptive down-sampling, end effector extremities (bottom), when the number of skeletal poses varies: K_E (red, circle, dash), K_{dtw} (black, square, plain), K_{rdtw}^{τ} (blue, star, dotted).

Fig. 5 Classification accuracies for the MSRAction3D data set, uniform down-sampling, full body (top), adaptive down-sampling, full body (middle), adaptive down-sampling, end effector extremities (bottom), when the number of skeletal poses varies: K_E (red, circle, dash), K_{dtw} (black, 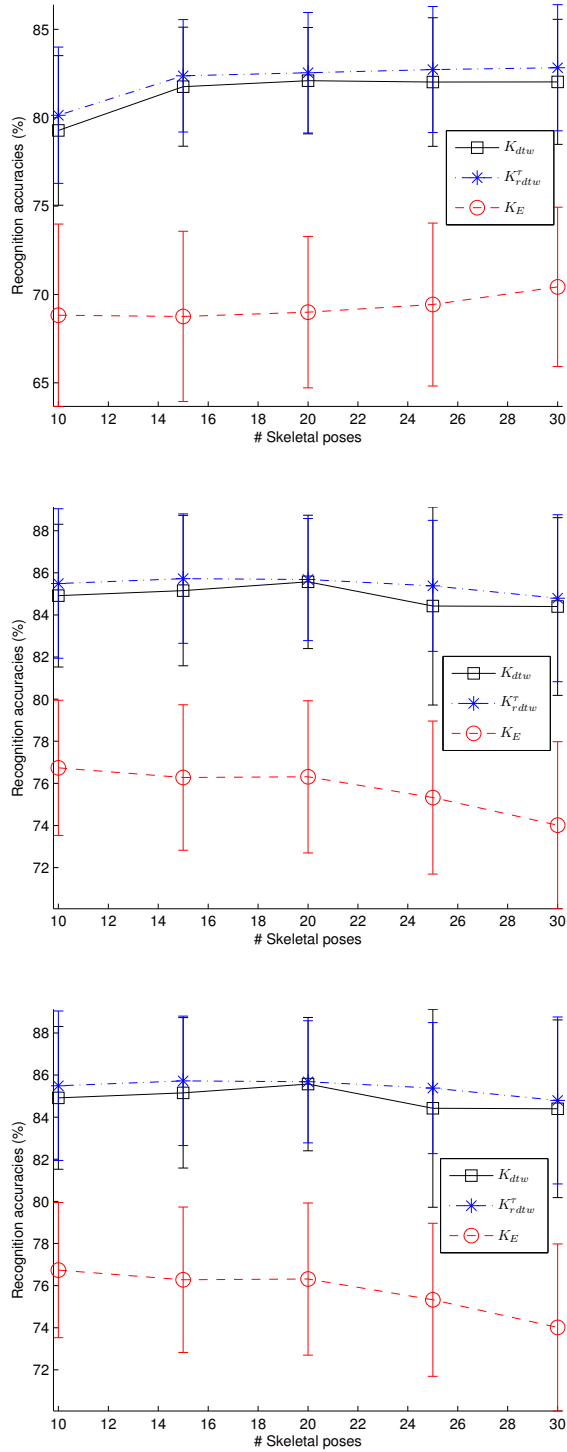square, plain), K_{rdtw}^{τ} (blue, star, dotted). Additionally, the cross validation on subjects (252 tests) allows to show the variance of the results.

6.1 Results and analysis

For the three considered tasks, we present the results obtained using a SVM classifier built from the LIBSVM library [Chang and Lin, 2011], the elastic non definite kernel $K_{dtw} = e^{-d_{dtw}(.,.)/(2.\sigma^2)}$, the elastic definite kernel K_{rdtw}^τ, and as a baseline, the Euclidean RBF kernel, $K_E = e^{-d_E^2(.,.)/(2.\sigma^2)}$.

Figures 3, 4 and 5 present the classification accuracies for respectively the HDM05-1, HDM05-2 and MSRAction3D tasks for the test data when the number of skeletal postures selected after down-sampling varies between 5 and 30. For these three figures, the top sub-figure presents classification accuracies when the FBD (Full Body) descriptor associated to a uniform down sampling (UDS) are used, the middle sub-figure classification accuracies when the FBD (Full Body) descriptor associated to an adaptive down sampling (ADS) are used, and the bottom sub-figure gives classification accuracies when the EED (End Extremities) descriptor associated to an adaptive down sampling (ADS) is used.

On all three figures, we observe that the down-sampling does not degrade the classification accuracies when the number of poses is over 10, and for some tasks it may even significantly improves the accuracy. This is likely due to the noise filtering effect of the down-sampling. High levels of down-sampling (e.g. when 10 to 15 postures are retained to describe movements, which represents an average compression ratio of 97 % for HDM05 and 70 % on MSRAction3D) lead to very satisfactory results (90 to 98 % for the two HDM05 tasks: Figures 3 and 4, and 82 to 86 % for the MSRAction3D task: 5). Best results are obtained for a number of skeletal poses between 15 and 20, when the EED descriptor is used in conjunction with an adaptive down sampling. The SVM classifier constructed on the basis of the regularized kernel K_{rdtw}^τ produces the best recognition rates (>= 96 % for the two HDM05 tasks). We note that the MSRAction3D task is much more difficult since it consists in a cross validation based on the performing subjects. Much lower performance are obtained for the SVM built on the basis of the Euclidean distance; in addition, if very good classification rate (96 %) is obtained on the training data, due to the noisy nature of Kinect data and the inter subject variability, the recognition rate on the test data falls down from 82 to 86 % .

Table 1 gives for the MSRAction3D data set and for the SVM based on K_E, K_{dtw} and K_{rdtw}^τ kernels, means and standard deviations, obtained on the training data (L) and testing data (T), of recognition rates (classification accuracies) when performing the cross-validation over the 10 subjects (5 train - 5 test splits leading to 252 tests), for the full body descriptor (FBD) and the end extremities descriptor (EED) associated either to a uniform down-sampling (UDS) or an adaptive Down Sampling (ADS). For this test, movements are represented as sequences of 15 skeletal postures. The drop of accuracies between learning and testing phases is due, on this dataset, to the large inter subjects variability of movement performances. Nevertheless, our experiment shows that the best average classification accuracies (obtained in general with minimal variance) are obtained for the most compact movement rep-

FBD-UDS	K_E (L)	K_E (T)	K_{dtw} (L)	K_{dtw} (T)	K_{rdtw}^{τ} (L)	K_{rdtw}^{τ} (T)
Mean	87,71	69,73	96,04	81,41	96,65	82,50
Stand. dev.	2,34	5,73	1,36	5,04	1,13	3,22
FBD-ADS	K_E (L)	K_E (T)	K_{dtw} (L)	K_{dtw} (T)	K_{rdtw}^{τ} (L)	K_{rdtw}^{τ} (T)
Mean	85,06	76,48	92,13	84,72	91,89	85,09
Stand. dev.	3,01	3,18	2,74	3,31	2,66	3,58
EED-ADS	K_E (L)	K_E (T)	K_{dtw} (L)	K_{dtw} (T)	K_{rdtw}^{τ} (L)	K_{rdtw}^{τ} (T)
Mean	92.46	76.27	97.14	85.16	**97.19**	85.72
Stand. dev.	3.45	0.96	3.57	5.04	0.93	3.07

Table 1 Means and standard deviations of classification accuracies on the MSRAction3D data set obtained according to a cross-validation on the subjects (5-5 splits, 252 tests) (L): on the training data, (T): on test data for a number of skeletal postures equal to 15.

resentation, i.e. when the EED descriptor is used associated to an adaptive down-sampling. This is true both for the training and testing datasets.

For comparison, table 2 gives results obtained by different methods of the state-of-the-art and compare them with the performance of an SVM that exploits the regularized DTW kernel (K_{rdtw}^{τ}) associated to the end extremity descriptor (EED) and an adaptive down-sampling (ADS) of 15 skeletal poses. To that end, we have reimplemented the Cov3DJ approach [Hussein et al., 2013] to get, for the MSRAction3D data set, the average result given by a 5-5 cross-validation on the subjects (252 tests). This comparative analysis shows that the SVM constructed from regularized DTW kernel provides results slightly above the current state-of-the-art for the considered data sets and tasks.

HDM05-1	Accuracy (%)
SMIJ [Ofli et al., 2012]	84.40
Cov3DJ, L = 3 [Hussein et al., 2013]	95.41
$SVM - K_{rdtw}^{\tau}$, EED, ADS with 15 poses	**96.33**
HDM05-2	**Accuracy (%)**
SMIJ [Ofli et al., 2013], 1-NN	91.53
SMIJ [Ofli et al., 2013], SVM	89.27
$SVMK_{rdtw}^{\tau}$, EED, ADS with 15 poses	**97.74**
MSR-Action3D	**Accuracy (%)**
Cov3DJ, L=3 [Hussein et al., 2013]	72.33 ± 3.69 [2]
HON4D, [Oreifej and Liu, 2013],	82.15 ± 4.18
$SVMK_{rdtw}^{\tau}$ 15 poses,	$\mathbf{85.72 \pm 3.07}$

Table 2 Comparative study on the MSRAction3D dataset, according to a cross-validation on the subjects (5-5 splits, 252 tests).

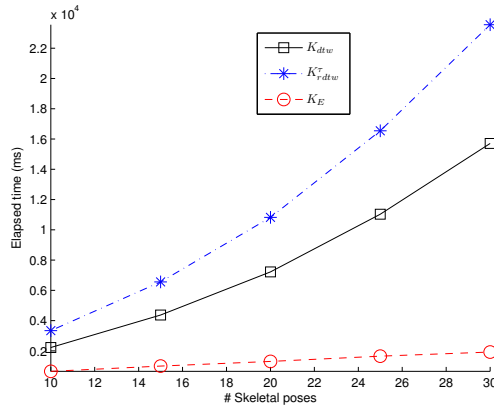[2] according to our own implementation of Cov3DJ

Fig. 6 Elapsed time as a function of the number of skeletal poses (10 to 30 poses): i) RBF Euclidean Kernel, Red/round/dashed line, ii) RBF DTW kernel, Black/square/plain line, iii) normalized regularized DTW kernel(K_{rdtw}^{τ}), blue/star/dotted line.

Finally, in Figure 6, we give the average CPU elapsed time for the processing of a single gestural MSRAction3D action when varying the number of retained skeletal poses. The test has been performed on an Intel Core i7-4800MQ CPU, 2.70GHz. Although the computational cost for the elastic kernel is quadratic, the latency for the classification of a single gestural action using a SVM-K_{rdtw}^{τ} is less than 25 milliseconds when 15 poses are considered, which effectively meets easily *real-time* requirements.

7 Conclusion and perspectives

In the context of isolated action recognition, where few studies explicitly consider dimension reduction along both the spatial and time axes simultaneously, we have presented a recognition model based on the dimensionality reduction of the skeletal pose descriptor and the down-sampling of motion sequences coupled to elastic kernel machines. Two ways of down-sampling have been considered: a uniform down-sampling that evenly selects samples along the time axis and an adaptive down-sampling based on a piecewise linear approximation model. The dimensionality reduction of the skeletal pose descriptor is straightforwardly obtained by considering only end effector trajectories, which is consistent with some sensorimotor perceptual evidence about the way human beings perceive and interpret motion. On the data sets and tasks that we have addressed, we have shown that, even when quite important down-sampling is considered, the recognition accuracy only slightly degrades. In any case, best accuracies are obtained when an adaptive down-sampling is used on the end effector 3D trajectories. The temporal redundancy is therefore

high and apparently not critical for the discrimination of the selected movements and tasks. In return, the down-sampling benefits in terms of computational complexity is quadratic with the reduction of the number of skeletal postures kept along the time axis.

Furthermore, the elasticity of the kernel provides a significant performance gain (comparatively to kernel based on the Euclidean distance) which is very important when the data are characterized by high variability. Our results show that a SVM based on a regularized DTW kernel is very competitive comparatively to the state-of-the-art methods applied on the two tested data sets, even when the dimension reduction on the time axis is important. The down-sampling and dimensionality reduction of the descriptor ensures that this approach meets the real-time constraint of gesture recognition.

This study opens perspectives to the use of elastic kernels constructed from more sophisticated time elastic distances [Marteau, 2009] that cope explicitly with time stamped data, associated to adaptive sampling techniques such the one developed in this paper or more sophisticated techniques capable of extracting the most significant and discriminant skeletal poses in movement sequences, based on semantic segmentation. We also aim at testing these powerful tools to more complex tasks, where skilled gestures are studied, or/and expressive variations are considered.

References

[cmu, 2003] (2003). Motion capture database, carnegie mellon university. //http://mocap.cs.cmu.edu/.

[Belkin and Niyogi, 2002] Belkin, M. and Niyogi, P. (2002). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396.

[Berg et al., 1984] Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*, volume 100 of *Graduate Texts in Mathematics*. Springer-Verlag, New York.

[Bissacco et al., 2007] Bissacco, A., Chiuso, A., and Soatto, S. (2007). Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(11):1958–1972.

[Chang and Lin, 2011] Chang, C.-C. and Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27.

[Cuturi et al., 2007] Cuturi, M., Vert, J.-P., Birkenes, O., and Matsui, T. (2007). A Kernel for Time Series Based on Global Alignments. In *Proceedings of ICASSP'07*, pages II–413 – II–416, Honolulu, HI. IEEE.

[de Aguiar et al., 2006] de Aguiar, E., Theobalt, C., and Seidel, H.-P. (2006). Automatic learning of articulated skeletons from 3d marker trajectories. In et al., B., editor, *ISVC*, volume 4291 of *Lecture Notes in Computer Science*, pages 485–494. Springer.

[Fothergill et al., 2012] Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1737–1746, New York, NY, USA. ACM.

[Gibet et al., 2011] Gibet, S., Marteau, P.-F., and Duarte, K. (2011). Toward a motor theory of sign language perception. In Efthimiou, E., Kouroupetroglou, G., and Fotinea, S.-E., editors, *Gesture Workshop*, volume 7206 of *Lecture Notes in Computer Science*, pages 161–172. Springer.

[Giese et al., 2008] Giese, M. A., I.M., T., and S., E. (2008). Metrics of the perception of body movement. *Journal of Vision*, 8(9):1–18. reviewed.

[Han et al., 2010] Han, L., Wu, X., Liang, W., Hou, G., and Jia, Y. (2010). Discriminative human action recognition in the learned hierarchical manifold space. *Image Vision Comput.*, 28(5):836–849.

[He and Niyogi, 2003] He, X. and Niyogi, P. (2003). Locality preserving projections. In *In Advances in Neural Information Processing Systems 16*. MIT Press.

[Hussain and Rashid, 2012] Hussain, S. and Rashid, A. (2012). User independent hand gesture recognition by accelerated dtw. In *Int. Conf. on Informatics, Electronics Vision (ICIEV)*, pages 1033–1037.

[Hussein et al., 2013] Hussein, M. E., Torki, M., Gowayyed, M. A., and El-Saban, M. (2013). Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*.

[Jenkins and MatariÄǦ, 2004] Jenkins, O. C. and MatariÄǦ, M. J. (2004). A spatio-temporal extension to isomap nonlinear dimension reduction. In *In The International Conference on Machine Learning (ICML 2004*, pages 441–448.

[Jolliffe, 1986] Jolliffe, I. (1986). *Principal component analysis*. Springer series in statistics. Springer-Verlang.

[Keogh and Pazzani, 2000] Keogh, E. J. and Pazzani, M. J. (2000). Scaling up dynamic time warping for datamining applications. In *Proc. of the Sixth ACM SIGKDD*, KDD '00, pages 285–289, New York, NY, USA.

[Kruskal and Wish, 1978] Kruskal, J. and Wish, M. (1978). *Multidimensional Scaling*. Sage Publications.

[Larochelle et al., 2012] Larochelle, H., Mandel, M., Pascanu, R., and Bengio, Y. (2012). Learning algorithms for the classification restricted boltzmann machine. *J. of Machine Learning Research*, 13:643–669.

[Li et al., 2010] Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3d points. In Press, I. C., editor, *Proc. IEEE Int'l Workshop on CVPR for Hum. Comm. Behav. Analysis*, pages 9–14.

[Marteau, 2009] Marteau, P. F. (2009). Time warp edit distance with stiffness adjustment for time series matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(2):306–318.

[Marteau and Gibet, 2005] Marteau, P. F. and Gibet, S. (2005). Adaptive sampling of motion trajectories for discrete task-based analysis and synthesis of gesture. In *LNAI Proc. of Int. Gesture Workshop*, pages 224–235. Springer.

[Marteau and Gibet, 2006] Marteau, P.-F. and Gibet, S. (2006). Adaptive sampling of motion trajectories for discrete task-based analysis and synthesis of gesture. In Gibet, S., Courty, N., and Kamp, J.-F., editors, *Gesture in Human-Computer Interaction and Simulation*, volume 3881 of *Lecture Notes in Computer Science*, pages 224–235. Springer Berlin Heidelberg.

[Marteau and Gibet, 2014] Marteau, P.-F. and Gibet, S. (2014). On Recursive Edit Distance Kernels with Application to Time Series Classification. *IEEE Trans. on Neural Networks and Learning Systems*, pages 1–14.

[Marteau and Ménier, 2009] Marteau, P.-F. and Ménier, G. (2009). Speeding up simplification of polygonal curves using nested approximations. *Pattern Anal. Appl.*, 12(4):367–375.

[Martens and Sutskever, 2011] Martens, J. and Sutskever, I. (2011). Learning recurrent neural networks with hessian-free optimization. In *ICML*, pages 1033–1040.

[Masoud and Papanikolopoulos, 2003] Masoud, O. and Papanikolopoulos, N. (2003). A method for human action recognition. *Image Vision Comput.*, 21(8):729–743.

[McLachlan, 2004] McLachlan, G. (2004). *Discriminant Analysis and Statistical Pattern Recognition*. Wiley Series in Probability and Statistics. Wiley.

[Mitra and Acharya, 2007] Mitra, S. and Acharya, T. (2007). Gesture recognition: A survey. *Trans. Sys. Man Cyber Part C*, 37(3):311–324.

[Müller et al., 2007] Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., and Weber, A. (2007). Documentation mocap database hdm05. Technical Report CG-2007-2, Universität Bonn.

[O'Brien et al., 2000] O'Brien, J. F., Bodenheimer, R. E., Brostow, G. J., and Hodgins, J. K. (2000). Automatic joint parameter estimation from magnetic motion capture data. In *Proceedings of Graphics Interface 2000*, pages 53–60.

[Ofli et al., 2012] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2012). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. In *CVPR Workshops*, pages 8–13. IEEE.

[Ofli et al., 2013] Ofli, F., Chaudhry, R., Kurillo, G., Vidal, R., and Bajcsy, R. (2013). Sequence of the most informative joints (smij): A new representation for human skeletal action recognition. *Journal of Visual Communication and Image Representation*, 0(0):1–20.

[Oreifej and Liu, 2013] Oreifej, O. and Liu, Z. (2013). Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. *2013 IEEE CVPR*, pages 716–723.

[Roweis and Saul, 2000] Roweis, S. T. and Saul, L. K. (2000). Nonlinear dimensionality reduction by locally linear embedding. *SCIENCE*, 290:2323–2326.

[Sakoe and Chiba, 1971] Sakoe, H. and Chiba, S. (1971). A dynamic programming approach to continuous speech recognition. In *Proceedings of the 7th International Congress of Acoustic*, pages 65–68.

[Sempena et al., 2011] Sempena, S., Maulidevi, N., and Aryan, P. (2011). Human action recognition using dynamic time warping. In *Int. Conf. on Electrical Engineering and Informatics (ICEEI)*, pages 1–5.

[Shotton et al., 2011] Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., Kipman, A., and Blake, A. (2011). Real-time human pose recognition in parts from single depth images. In *Conf. on Computer Vision and Pattern Recognition*, CVPR '11, pages 1297–1304. IEEE.

[Tenenbaum et al., 2000] Tenenbaum, J. B., de Silva, V., and Langford, J. C. (2000). A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319.

[Veeraraghavan et al., 2004] Veeraraghavan, A., Chowdhury, A. K. R., and Chellappa, R. (2004). Role of shape and kinematics in human movement analysis. In *CVPR (1)*, pages 730–737.

[Velichko and Zagoruyko, 1970] Velichko, V. M. and Zagoruyko, N. G. (1970). Automatic recognition of 200 words. *International Journal of Man-Machine Studies*, 2:223–234.

[Wang et al., 2012] Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *IEEE int. conf. CVPR*, pages 1290–1297.

[Wang et al., 2006] Wang, S. B., Quattoni, A., Morency, L., Demirdjian, D., and Darrell, T. (2006). Hidden conditional random fields for gesture recognition. In *IEEE int. conf. CVPR*, volume 2, pages 1521–1527.

[Yu and Aggarwal, 2009] Yu, E. and Aggarwal, J. (2009). Human action recognition with extremities as semantic posture representation. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, 0:1–8.

[Zhao et al., 2012] Zhao, X., Song, Z., Guo, J., Zhao, Y., and Zheng, F. (2012). Real-time hand gesture detection and recognition by random forest. In Zhao, M. and Sha, J., editors, *Communications and Information Processing*, volume 289, pages 747–755. Springer Berlin Heidelberg.