Down-Sampling coupled to Elastic Kernel Machines for Efficient Recognition of Isolated Gestures

Pierre-François Marteau, Sylvie Gibet, Clement Reverdy

Down-Sampling coupled to Elastic Kernel Machines for Efficient Recognition of Isolated Gestures

Pierre-Francois Marteau and Sylvie Gibet and Clément Reverdy
IRISA (UMR 6074), Université de Bretagne Sud
Campus de Tohannic, 56000 Vannes, France
Email: firstname.name AT univ-ubs DOT fr

***Abstract*—In the field of gestural action recognition, many studies have focused on dimensionality reduction along the spatial axis, to reduce both the variability of gestural sequences expressed in the reduced space, and the computational complexity of their processing. It is noticeable that very few of these methods have explicitly addressed the dimensionality reduction along the time axis. This is however a major issue with regard to the use of elastic distances characterized by a quadratic complexity. To partially fill this apparent gap, we present in this paper an approach based on temporal down-sampling associated to elastic kernel machine learning. We experimentally show, on two data sets that are widely referenced in the domain of human gesture recognition, and very different in terms of quality of motion capture, that it is possible to significantly reduce the number of skeleton frames while maintaining a 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. The computational complexity reduction makes this approach eligible for *real-time* applications.**

I. Introduction

During the past decade, gesture recognition has been a very active research field that has evolved in terms of improving motion capture technology and recognition methods, mostly based on machine learning techniques. Recently, the availability of low-cost consumer technology, often associated with game consoles, has helped to democratize the use of motion sensors, not only in the context of interactive video game, but also in various frameworks using gestural interaction. Hence high quality databases, built from expensive motion capture (*mocap*) devices requiring specific expertise, exist today alongside lower quality databases, i.e. containing noisier and less accurate data provided by new inexpensive sensors that require very little expertise. Therefore, heterogeneous databases of captured motion of various qualities are available to the scientific community, and comparing the robustness and generalization of recognition algorithms on these diverse motion databases is highly challenging. Beyond the quality of the recognition, the complexity of the algorithms and their response time is also a major issue, especially for real-time interaction.

In the context of the recognition of isolated gestures from motion captured data, we present in this paper a novel method that improves the performance of classical support vector machines when used with regularized elastic kernels. We also show how the temporal dimensionality reduction, associated with such elastic kernels significantly improves the efficiency of the algorithm and the recognition scores.

II. Motion captured data and sequence of skeletal poses

We focus on human motion data captured by various camera-based sensors (infrared marker-tracking system with high resolution, or webcam-style system). The data is uniformly preprocessed so that the captured data is finally reconstructed as a set of 3D-trajectories of the skeleton joints determined from the positions of markers captured on a real actor underlying more or less accurately the skeleton of the actor who produced the movement. The identification of the skeleton model from captured data is achieved through a mapping-optimization process such as the ones described in [1], [2], or [3]. The techniques based on a skeleton model hence convert 3D sensor data into Cartesian or angular coordinates that define with various accuracies the state of the joints over time.

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.

This sort of data is inherently noisy, mainly due to the data acquisition process (drifts, imprecision and shading, etc.) and sensor noise. Furthermore, due to the nature of the capture devices and in particular to the number of sensors that are operated, the nature of the reconstructed skeletons is also subject to some variation for two main reasons: in the one hand the morphology of the actor (segment lengths) is the main source of variability for a given capture device, and in the second hand, the number of joints (the number of degree of freedom) vary according to the capture device, leading to some additional noise due to the reconstruction of the skeletal data from sensor outputs.

Thus, any movement can be defined 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 new prospects for dimension reduction as well as noise reduction approaches, which is particularly relevant when considering motion recognition as one can targets gains both in terms of computation time and error rate.

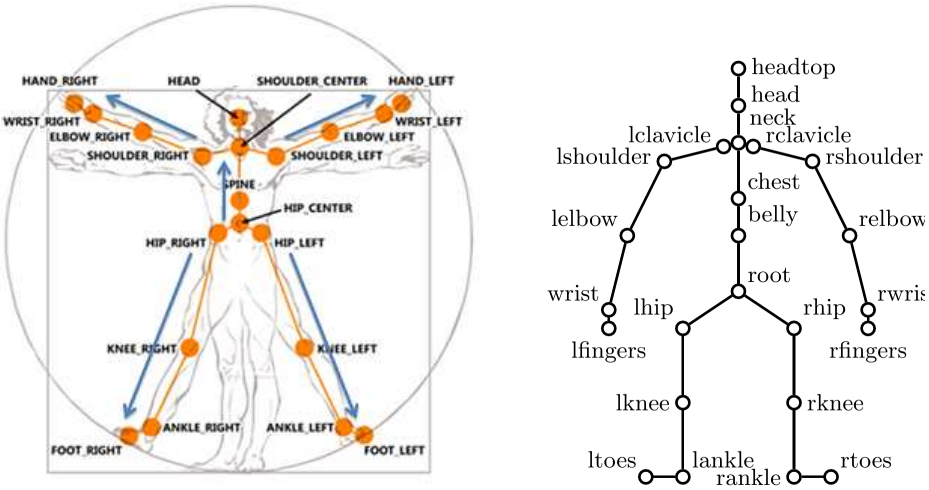

Fig. 1. Examples of skeletons reconstructed from motion data captured from the Kinect (left) and from the Vicon-MX device used by the Max Planck Institute (right).

III. Related works

Gesture analysis and recognition is very broad and recently very active area due to the democratization of low cost motion capture systems by camera. It covers aspects of signal and image processing, dynamic modeling, statistical and machine learning approaches. We hereinafter give a brief non-exhaustive overview of the main methods proposed for gesture analysis and recognition. These methods primarily focus on the extraction of features that significantly represent the kinematics and dynamics of the skeleton data characterized as a whole, or portions of it. Among them, some approaches, based on linear dynamic models [4], have exploited autoregressive (AR) models and autoregressive moving-average (ARMA) models to characterize the kinematics of movements, while other approaches, based on nonlinear dynamic models [5], have developed movement analysis and recognition scheme based on dynamical models controlled by Gaussian processes. [6] propose a synthesis of the major gesture recognition approaches relying on Hidden Markov Models (HMM). [7] 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 [8]; among them, conditional restricted Boltzman's machines [9] have been studied in the context of motion captured data modeling and classification.

Some methods address more specifically the problem of dimensionality reduction with an objective for reducing the variability while seeking an efficiency gain. In particular, the Principal Component Analysis (PCA) has been widely exploited for gesture analysis and recognition with the objective of reducing the dimensionality of motion data [10]. Other methods such as linear projections preserving locally neighborhoods (Locality Preserving Projection) [11], or their non-linear counter-parts such as ISOMAP [12], have been implemented to embed postures in low dimensional spaces in which a more efficient time warp (DTW, see section IV-B) algorithm, associated with the Hausdorff distance, can be used to classify movements. Other ad hoc methods for dimension reduction are also proving their efficiency: we can mention for example the recent work of [13] that proposes to only consider the trajectories of the 5 end-extremities of the skeleton (2 feet, 2 hands and the head). Models based on Gaussian processes with latent variables are also largely used, for instance a hierarchical version has been recently exploited for gesture recognition [14].

The identification of significant variables maximizing the discrimination of motion classes has also been widely explored. [15] and [16] have in particular applied random forests to recognize actions, using a Kinect sensor, while [17] recently proposed to automatically select the most informative skeletal joints to explain the current action. In the same line, [18] consider covariance matrices evaluated on some skeletal joints as discriminative descriptors to characterize a movement sequence. The use of sliding windows can be view as a dimension reduction along the time axis. In [19] a simple bag of 3D points is used to represent and recognize gestural action. Similarly, in [20], *actionlets* are defined from Fourrier's coefficients to characterize the most discriminative joints. Histograms of oriented 4D normals have been also proposed in [21] for the recognition of gestural actions and movements from sequences of depth images.

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 [22], as well as the hardware acceleration proposed in [23]. However, to our knowledge, no work exploiting this type of distance has directly studied the question of data reduction along the time axis.

IV. Downsampling movement sequences coupled to elastic kernel machines

When considering the use of elastic distances or kernels to benefit from their ability to deal with some form of temporal variability, we are rapidly confronted with their computational cost, in general quadratic with the length of the time series that are processed and linear with the *spatial* dimension (number of degrees-of-freedom). This high computational complexity is somehow limiting their use, especially when large amounts of

data has to be processed, or when so-called *real-time* constraint is required. It is therefore *a priori* particularly relevant to consider a dual dimensionality reduction, firstly on the time axis, and secondly on the spatial axis. Hence, in the context of *mocap* data processing, it seems useful to determine if a spatio-temporal redundancy of motion paths can be exploited, especially in the perspective of using elastic distances.

A. Dimension reduction along the time axis

Considering the quite rich literature on gesture recognition, it is significant to note that while some studies have shown success with dimensionality reduction on the spatial axis, very few have directly addressed a reduction in dimensionality along the time axis *per se* to reduce the complexity of elastic matching. [24] explicitly mentioned a temporal sub-sampling associated with a dynamic time warping in the context of time series mining, followed later by [25]. In order to explicitly reduce dimensionality along the time axis, our straightforward approach here consists in sub-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. It becomes then easy to perform a classification or recognition task by using elastic kernel machines on such fixed-size sequences to assess performance rates depending on the degree of sub-sampling that is considered. Indeed, this approach is quite raw, as long sequences can be characterized with the same number of skeletal poses than short sequences. For very short sequences, whose length is shorter than L, if any, we over-sample the sequence in order to meet the fixed-size requirement. But we consider this case as very marginal here since we seek a sub-sampling rate much lower than the average length of the motion sequence.

B. Elastic kernels and their regularization

Dynamic Time Warping (DTW), [26], [27], by far the most used elastic measure, is defined as

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

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. Besides the fact that this measure does not respect the triangle inequality, it does not directly define a positive definite kernel. 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 [28], [29] allowed to propose new guidelines to regularize kernels constructed from elastic measures such as DTW. A simple instance of such regularized kernel, derived from [29] for time series of equal length, takes the following form, which relies on two recursive terms :

$$\begin{aligned} &\mathcal{K}_{rdtw}(X_p, Y_q) = K^{xy}_{rdtw}(X_p, Y_q) + K^{xx}_{rdtw}(X_p, Y_q) \\ &K^{xy}_{rdtw}(X_p, Y_q) = \tfrac{1}{3} e^{-\nu d_E^2(x(p), y(q))} \\ &\sum \begin{cases} h(p-1, q) K^{xy}_{rdtw}(X_{p-1}, Y_q) \\ h(p-1, q-1) K^{xy}_{rdtw}(X_{p-1}, Y_{q-1}) \\ h(p, q-1) K^{xy}_{rdtw}(X_p, Y_{q-1}) \end{cases} \\ &K^{xx}_{rdtw}(X_p, Y_q) = \tfrac{1}{3} \\ &\sum \begin{cases} h(p-1, q) K^{xx}_{rdtw}(X_{p-1}, Y_q) e^{-\nu d_E^2(x(p), y(p))} \\ \Delta_{p,q} h(p, q) K^{xx}_{rdtw}(X_{p-1}, Y_{q-1}) e^{-\nu d_E^2(x(p), y(q))} \\ h(p, q-1) K^{xx}_{rdtw}(X_p, Y_{q-1}) e^{-\nu d_E^2(x(q), y(q))} \end{cases} \end{aligned} \quad (2)$$

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 $K^{xy}_{rdtw}(X_0, Y_0) = K^{xx}_{rdtw}(X_0, Y_0) = 1$.

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.

Elastic kernels: we consider in this paper only the exponential kernel (Gaussian or RBF-type) constructed from the two previous elastic measures d_{dtw} and K_{rdtw}, and the non-elastic kernel obtained from the Euclidean distance [1], i.e. $K_{dtw}(.,.) = e^{-d_{dtw}(.,.)/\sigma}$, and $K_E(.,.) = e^{-d_E^2(.,.)/\sigma}$. For the regularized DTW kernel, a data dependent normalization heuristic is required and the final kernel takes the form:

$K_{rdtw}(.,.) = e^{\beta \mathcal{K}^{\alpha}_{rdtw}(.,.)/\sigma}$, with

- $\alpha = 1/log(max(\mathcal{K}_{rdtw}(.,.))/min(\mathcal{K}_{rdtw}(.,.)))$ and
- $\beta = exp(-\alpha \cdot log(min(\mathcal{K}_{rdtw}(.,.))))$,

where min and max are taken over all the training data pairs.

V. Experimentation

To estimate the robustness of the proposed approach, we evaluate it on two motion capture databases of opposite

[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

quality, the first one developed at the Max Planck Institute, the other at Microsoft research laboratories.

HDM05 data set [31] 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 [32] (also exploited in the work of [18]) and [17]. 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 [19] 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 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.

A. Results and analysis

For the two considered tasks, we present the results obtained using a SVM classifier built from the LIBSVM library [33], the elastic kernels K_{dtw} and K_{rdtw}, and as a baseline the Euclidean distance kernel, K_E.

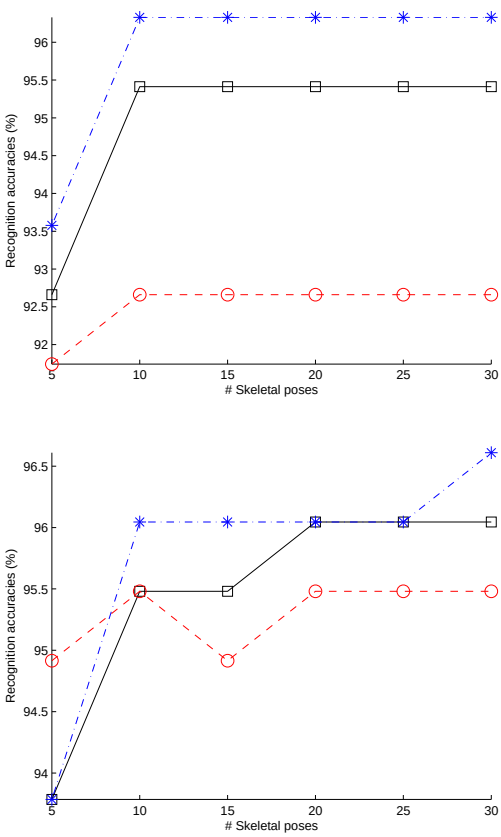

Fig. 2. Classification accuracies for HDM05-1 task (top), and HDM05-2 task (bottom), when the number of skeletal poses varies: K_E (red, circle, dash), K_{dtw} (black, square, plain), K_{rdtw} (blue, star, dotted).

Figure 2 presents the classification accuracies when the number of skeletal postures selected after downsampling varies between 5 and 30. Results for HDM05-1 task is presented in the top sub-figure, while results for HDM05-2 task is given in the bottom sub-figure. Figure 3 shows also the classification accuracies obtained on the MSRAction3D data set when the number of skeletal postures varies between 10 and 30: the accuracies obtained for a 10-fold cross validation on the training data is given in the top sub-figure. accuracies for the testing data is given in the bottom sub-figure.

On both figures, we observe that the sub-sampling does not catastrophically degrade the accuracies. High levels of down-sampling (e.g. 10 to 15 postures retained by movement, which represents an average compression ratio of 97 % for HDM05 and 70 % on MSRAction3D) lead to very satisfactory results (96-98 % for the two HDM05 tasks and 95 to 97 % for the MSRAction3D task on the learning data). The SVM classifier constructed on the basis of the regularized kernel K_{rdtw} produces the best recognition rate. We note that the MSRAction3D task is more difficult: 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 drop down to 82 % .

Table I gives for the MSRAction3D data set and for the

	K_E A	K_E T	K_{dtw} A	K_{dtw} T	K_{rdtw} A	K_{rdtw} 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

TABLE I. MEANS AND STANDARD DEVIATIONS OF CLASSIFICATION ACCURACIES ON THE MSRACTION3D DATA SET OBTAINED ACCORDING TO A CROSS-VALIDATION ON THE SUBJECTS (252 TESTS) A: ON THE TRAINING DATA, T: ON TEST DATA FOR A NUMBER OF SKELETAL POSTURES EQUAL TO 15.

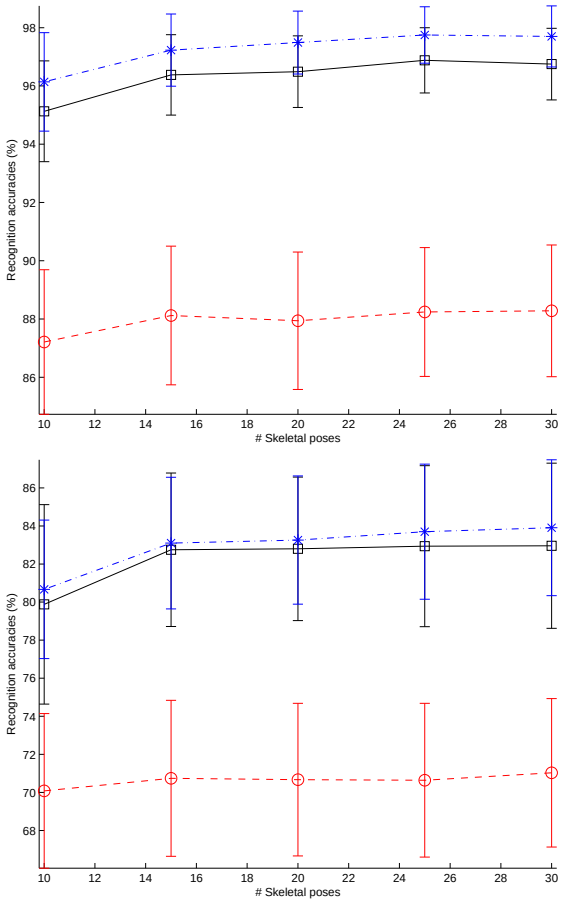

Fig. 3. Classification accuracies for the MSRAction3D data set, on the learning data (top) and testing data (bottom) when the number of skeletal poses varies: K_E (red, circle, dash), K_{dtw} (black, square, plain), K_{rdtw} (blue, star, dotted).

SVM based on K_E, K_{dtw} and K_{rdtw} kernels, means and standard deviations, obtained on the training data (A) and testing data (T), of recognition rates (classification accuracies) when performing the cross-validation over the 10 subjects (252 configurations). For this test, movements are represented as sequences of 15 skeletal postures. The drop of accuracies between Learning and testing is due, on this dataset, to the large inter subjects variability of movement realizations.

For comparison, table II gives results obtained by different methods of the state-of-the-art and compare them with the performance of our SVM constructed from the Regularized DTW associated with a down-sampling of 15 postures. To that end, we have reimplemented the Cov3DJ approach [18] 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 [32]	84.40
Cov3DJ, L = 3 [18]	95.41
$SVMK_{rdtw}$, 15 poses	**96.33**
HDM05-2	Accuracy (%)
SMIJ [17], 1-NN	91.53
SMIJ [17], SVM	89.27
$SVMK_{rdtw}$, 15 poses	**96.05**
MSR-Action3D	Accuracy (%)
Cov3DJ, L=3 [18]	72.33 ± 3.69 [2]
HON4D, [21],	82.15 ± 4.18
$SVMK_{rdtw}$ 15 poses,	**83.10 $\pm$ 3.46**

TABLE II. COMPARATIVE STUDY.

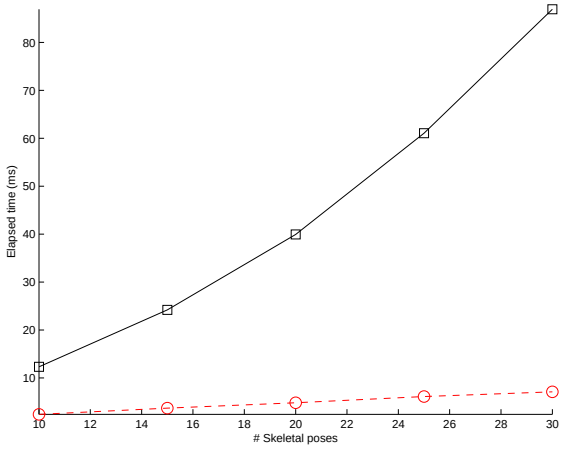

Fig. 4. Elapsed time as a function of the number of skeletal poses (10 to 30 poses): i) Euclidean Kernel, Red/round/dotted line, ii) Elastic kernel (RDTW), Black/square/plain line.

Finally, in Figure 4, 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 is less than 25 milliseconds when 15 poses are considered, which effectively meets easily *real-time* requirements.

VI. CONCLUSION AND PERSPECTIVES

In the context of isolated gesture recognition, where few studies explicitly consider dimension reduction along the time axis, we have presented a simple approach based on sub-sampling motion sequences coupled to the exploitation of elastic kernel machines. 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. The temporal redundancy is therefore high and apparently not critical for the discrimination of the selected

[2] according to our own implementation of Cov3DJ

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 performance gain (compared to kernel based on the Euclidean distance) which is 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. This study opens perspectives to the use of more sophisticated elastic kernels [25] associated to adaptive sampling techniques [34] [35] capable of extracting the most significant and discriminant skeletal poses in movement sequences.

REFERENCES

[1] E. de Aguiar, C. Theobalt, and H.-P. Seidel, "Automatic learning of articulated skeletons from 3d marker trajectories." in *ISVC*, ser. Lecture Notes in Computer Science, B. et al., Ed., vol. 4291. Springer, 2006, pp. 485–494.

[2] J. F. O'Brien, R. E. Bodenheimer, G. J. Brostow, and J. K. Hodgins, "Automatic joint parameter estimation from magnetic motion capture data," in *Proceedings of Graphics Interface 2000*, May 2000, pp. 53–60.

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

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

[5] A. Bissacco, A. Chiuso, and S. Soatto, "Classification and recognition of dynamical models: The role of phase, independent components, kernels and optimal transport," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 11, pp. 1958–1972, 2007.

[6] S. Mitra and T. Acharya, "Gesture recognition: A survey," *Trans. Sys. Man Cyber Part C*, vol. 37, no. 3, pp. 311–324, May 2007.

[7] S. B. Wang, A. Quattoni, L. Morency, D. Demirdjian, and T. Darrell, "Hidden conditional random fields for gesture recognition," in *IEEE int. conf. CVPR*, vol. 2, 2006, pp. 1521–1527.

[8] J. Martens and I. Sutskever, "Learning recurrent neural networks with hessian-free optimization," in *ICML*, 2011, pp. 1033–1040.

[9] H. Larochelle, M. Mandel, R. Pascanu, and Y. Bengio, "Learning algorithms for the classification restricted boltzmann machine," *J. of Machine Learning Research*, vol. 13, pp. 643–669, Mar. 2012.

[10] O. Masoud and N. Papanikolopoulos, "A method for human action recognition," *Image Vision Comput.*, vol. 21, no. 8, pp. 729–743, 2003.

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

[12] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, p. 2319, 2000.

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

[14] L. Han, X. Wu, W. Liang, G. Hou, and Y. Jia, "Discriminative human action recognition in the learned hierarchical manifold space," *Image Vision Comput.*, vol. 28, no. 5, pp. 836–849, May 2010.

[15] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin, "Instructing people for training gestural interactive systems," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ser. CHI '12. New York, NY, USA: ACM, 2012, pp. 1737–1746.

[16] X. Zhao, Z. Song, J. Guo, Y. Zhao, and F. Zheng, "Real-time hand gesture detection and recognition by random forest," in *Communications and Information Processing*, M. Zhao and J. Sha, Eds. Springer Berlin Heidelberg, 2012, vol. 289, pp. 747–755.

[17] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy, "Sequence of the most informative joints (smij): A new representation for human skeletal action recognition," *Journal of Visual Communication and Image Representation*, vol. 0, no. 0, pp. 1–20, 2013.

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

[19] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *Proc. IEEE Int'l Workshop on CVPR for Hum. Comm. Behav. Analysis*, I. C. Press, Ed., 2010, pp. 9–14.

[20] J. Wang, Z. Liu, Y. Wu, and J. Yuan, "Mining actionlet ensemble for action recognition with depth cameras," in *IEEE int. conf. CVPR*, 2012, pp. 1290–1297.

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

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

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

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

[25] P. F. Marteau, "Time warp edit distance with stiffness adjustment for time series matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 306–318, 2009.

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

[27] H. Sakoe and S. Chiba, "A dynamic programming approach to continuous speech recognition," in *Proceedings of the 7th International Congress of Acoustic*, 1971, pp. 65–68.

[28] M. Cuturi, J.-P. Vert, O. Birkenes, and T. Matsui, "A Kernel for Time Series Based on Global Alignments," in *Proceedings of ICASSP'07*. Honolulu, HI: IEEE, April 2007, pp. II–413 – II–416.

[29] P.-F. Marteau and S. Gibet, "On constructing positive elastic kernels with application to time series classification," IEEE Transactions on Neural Networks and Learning Systems, June 2014

[30] D. Haussler, "Convolution kernels on discrete structures," University of California, Santa Cruz, Tech. Rep., 1999, technical Report.

[31] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Krüger, and A. Weber, "Documentation mocap database hdm05," Universität Bonn, Tech. Rep. CG-2007-2, June 2007.

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

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

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

[35] P.-F. Marteau and G. Ménier, "Speeding up simplification of polygonal curves using nested approximations." *Pattern Anal. Appl.*, vol. 12, no. 4, pp. 367–375, 2009.