Revisiting Human Action Recognition: Personalization vs. Generalization

Andrea Zunino*†, Jacopo Cavazza*† and Vittorio Murino*‡
firstname.lastname@iit.it
* Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, Via Morego 30, 16163, Genova, Italy
† Università degli Studi di Genova - Dipartimento di Ingegneria Navale, Elettrica, Elettronica e delle Telecomunicazioni, Via All’Opera Pia, 11A, 16145, Genova, Italy
‡ Università di Verona - Dipartimento di Informatica, Strada le Grazie 15, 37134, Verona, Italy

Abstract—By thoroughly revisiting the classic human action recognition paradigm, this paper aims at proposing a new approach for the design of effective action classification systems. Taking as testbed publicly available three-dimensional (MoCap) action/activity datasets, we analyzed and validated different training/testing strategies. In particular, considering that each human action in the datasets is performed several times by different subjects, we were able to precisely quantify the effect of inter- and intra-subject variability, so as to figure out the impact of several learning approaches in terms of classification performance. The net result is that standard testing strategies consisting in cross-validating the algorithm using typical splits of the data (holdout, k-fold, or one-subject-out) is always outperformed by a “personalization” strategy which learns how a subject is performing an action. In other words, it is advantageous to customize (i.e., personalize) the method to learn the actions carried out by each subject, rather than trying to generalize the actions executions across subjects. Consequently, we finally propose an action recognition framework consisting of a two-stage classification approach where, given a test action, the subject is first identified before the actual recognition of the action takes place. Despite the basic, off-the-shelf descriptors and standard classifiers adopted, we noted a relevant increase in performance with respect to standard state-of-the-art algorithms, so motivating the usage of personalized approaches for designing effective action recognition systems.

I. INTRODUCTION

Classification of actions or activities from videos or still images is actually a very complex task due to the high variability of the action/activity classes which can be performed differently by different people, and even differently by the same person at different times. Besides, other problems related to the context, clutter/noise, illumination variations, and occlusions make this task even more challenging. As a consequence, the two-dimensional (2D) and 2D+time actual informative structure of the image and video data is strongly affected by all the above issues, which complicate the design of reliable action recognition systems. For all the aforementioned reasons and because of the implying relevant practical applications, activity recognition is undoubtedly one of the most interesting and debated problems in computer vision and pattern recognition [1]. Fortunately, this scenario slightly improves owing to recently introduced three-dimensional (3D) sensor technology, which can nowadays be used to safely and accurately capture human motion at high-resolution in both spatial and temporal domains (e.g., VICON), with good accuracy, and at low cost (e.g., Kinect). Despite the richer information provided by 3D+time data, the task is only apparently easier as, although some of the problems linked to the appearance may result mitigated, the fine 3D information which can be recovered by such data may degrade the performance of the adopted recognition method.

All such aspects corroborate the continuous development and improvement of computational approaches for 3D action recognition, which can be categorized into three main classes, according to [2]. In the first one, the whole set of joints coordinates is modeled, for instance, by using first/second order finite difference schemes [3] or bag-of-word encodings [4]. Second, selection criteria of the most discriminant joint can be devised; for instance, [5] exploited mean and variance of joint angles and maximal angular velocity. Third, a temporal modeling of dynamics is performed through different techniques (e.g., autoregressive models [6] or non-parametric Bayesian inference [7]). Recently, [8] set the new state-of-the-art by considering a double layer of low- and high-level kernels to represent each action separately and to measure their mutual similarity, respectively.

In such context, this work undertakes a revisiting perspective of the action recognition paradigm, probing the principal
evaluation strategies applied in the literature on the most common, publicly available, benchmark datasets. Thus, we aim at providing a deep understanding about the challenges that have to be faced when devising classification protocols: such awareness leads us to introduce a new effective, yet simple, approach for action recognition. The experimental testbed we have chosen consists of 3 public datasets, namely MSR-Action3D [9], MSRC-Kinect12 [10] and HDM-05 [11]. Each has own peculiar traits, e.g., the amount and type of considered action classes or the number of skeletal joints. However, a common shared aspect is that a same action is performed by several subjects and a same subject actually performs each action more times. The variability of considered actions aim at reproducing real-world scenarios, while repeating actions and considering multiple actors allow to increase the learning methods in robustness and generalization, respectively. Usually, action recognition methods in the literature do not exploit the information associated to the subject identity, but they typically consider different splits of all action instances (e.g., k-fold cross-validation) in the training/testing phases. Nevertheless, such information is quite relevant, indeed discriminant, for the actual recognition of the actions since each human being shows peculiar features which are reflected in the way an action is performed. The former aspects have been rarely investigated and seldom quantified by previous recognition system to date and, to this end, we focus on two main aspects:

- **Inter-subject variability**, which either refers to anthropometric differences of body parts or to incongruous personal styles in accomplishing the scheduled action. In practice, different subjects may perform the same (even very simple) action in different ways.
- **Intra-subject variability**, which represents the random nature of each single action class (e.g., throwing a ball), which can also be dictated by pathological conditions or environmental factors. In other words, this reflects the fact that a subject never performs an action in the same exact way.

Both aspects lead to the fact that a same action could not be performed exactly equal to itself, either it is executed by the same or different human beings. In this line, the additional information of subject identity has empirically demonstrated to be effective in customizing the classification on a specific user for speech [12], handwriting [13], and gesture [14] recognition.

Among the few works which studied the variability within/across subjects, for instance, [15] did not register a strong impact of different subjects in daily activities classification, and [16] documented the stability of the performance on an ad hoc acquired dataset characterized by biometric homogeneity of the participants. Differently, in [17], the performance of checking the correct execution of gymnastics sharply falls when the subject under testing is excluded from the training phase. A similar trend was registered by [18] and [19] for computer assisted rehabilitation tasks, as well as by [20] which performed a theoretical dissertation about within-subject and across-subjects noise using wearable motion sensors. Globally, [15], [16], [17], [18], [19], [20] did not mutually agree in their conclusions and, also, their investigation is actually limited by the use of private datasets explicitly designed for the considered application.

Despite some previous approaches grant in some way the importance of the knowledge of the human subject (especially for rehabilitation purposes, where the goal is directed to a specific subject), no study has been systematically reported to date on commonly used and publicly available datasets for general action/activity recognition. In other words, it is still an open problem to quantify how much those datasets are affected by inter- and intra-subject variability, and hence to figure out the impact of subjectiveness in action recognition to actually investigate the trade-off between personalization and generalization in the design of automatic systems.

These arguments are investigated in this paper through the following main contributions.

(i) Considering MSR-Action3D, MSRC-Kinect12 and HDM-05 benchmark datasets, we analyse different testing strategies, investigating the cases where 1) the data of one subject are left out for testing – **One-Subject-Out**, 2) the actions (properly split in separate sets) executed from all subjects are present in both training and in testing – **Cross-Validation**, and 3) the classification is performed over a simplified problem considering only the instances belonging to one specific subject at a time – **Personalization**. Such analysis is performed considering two different types of off-the-shelf encodings: covariance-based feature description and dynamic time warping, both used to estimate the mutual similarity of different action instances in the context of an SVM-based classification approach.

(ii) The role of subjectiveness is introduced and investigated, to estimate the balance between personalization and generalization (Figure 1). By means of a quantitative statistical analysis, we evaluate the effectiveness of retrieving in testing all the subjects used in the training phase by assessing the role played by either inter- or intra-subject variability.

(iii) Finally, we propose a two-stage recognition pipeline where the preliminary identification of the subject is followed by a subject-specific action classification. Overall, our new proposed pipeline shows a strong performance with respect to both Cross-Validation and One-Subject-Out strategies, also being superior to the state-of-the-art method [8]. This promising result may open a new paradigm for the development of action/activity recognition systems, also embracing the possibility to exploit these findings in the design of joint biometric, kinematic-driven authentication systems.

The rest of the paper is organized as follows. In Section II we present the considered datasets and the feature representations adopted, and the evaluation strategies investigated are reported in III. Section IV presents and widely discusses
the experimental results, and we illustrate the aforementioned two-stage classification pipeline in Section V. Finally, Section VI draws the conclusions of this study.

II. DATASETS & FEATURE ENCODING

Our investigation involves three publicly available MoCap datasets for activity recognition: MSR-Action3D, MSRC-Kinect12 and HDM-05. For all our experiments, we only used the 3D skeleton coordinates while the other data available (e.g., depth maps or RGB videos) were not taken into account. For the sake of clarity we briefly introduce each of them.

— The MSR-Action3D [9] dataset has 20 action classes of mostly sport-related actions (e.g., jogging or tennis-serve), performed by 10 subjects. J = 20 joints are extracted from the Kinect sensor data to model the human pose of the human agents. Each subject performs each action 16 times, on average. The available motion annotated gesture instances in total. Each subject accomplishes 3D skeleton data, recorded by means of a Kinect sensor.

 — MSRC-Kinect12 [10] is a relatively large dataset of 3D skeleton data, recorded by means of a Kinect sensor. The dataset has 594 sequences, containing 12 action classes performed by 30 different subjects, precisely there are 6244 annotated gesture instances in total. Each subject accomplishes each class of action 16 times, on average. The available motion files contain the trajectories estimated for J = 20 3D skeleton joints.

 — In HDM-05 [11], the number of skeleton joints is J = 31, and the dataset contains more than three hours of systematically recorded VICON MoCap data acquiring different types of gestures performed by 5 professional actors. Motion clips have been manually cut and annotated into roughly 100 different motion classes. In this work we have removed some severely corrupted samples (as in [21]) and, as in [8], we selected only the following 14 classes: clap above head, deposit floor, elbow to knee, grab high, hop both legs, jog, kick forward, lie down floor, rotate both arms backward, sit down chair, sneak, squat, stand up tie and throw basketball. In this setup each subject accomplishes each action 5 times, on average.

 For all the aforementioned datasets, each trial can be formalized as a collection S of J different acquisitions \( p(1), \ldots, p(\tau) \). For any \( t = 1, \ldots, \tau \), we denote with \( p(t) \) the column vector which stacks \( p_1(t), \ldots, p_J(t) \in \mathbb{R}^3 \), the three-dimensional \( x, y, z \) coordinates of the \( J \) skeletal joints. Using this notation, we now briefly introduce the two different representations for MoCap data.

 First, we investigated the usage of dynamic time warping (DTW), a classical tool to quantify the similarity across two different time series by means of alignment [22], [23]. In order to apply DTW, we evaluated the differences between any two joints collection \( S = [p(1), \ldots, p(\tau)] \) and \( S' = [p'(1), \ldots, p'('\tau')] \) through the following distance

\[
d(p(s), p'(t)) = \frac{1}{J} \sum_{j=1}^{J} ||p_j(s) - p'_j(t)||, \tag{1}
\]

where \( || \cdot || \) is the Euclidean norm, \( s = 1, \ldots, \tau \) and \( t = 1, \ldots, \tau' \). The final similarity measure, provided by DTW to compare \( S \) and \( S' \), is \( \delta(S, S') \) which is the minimum value of \( (1) \) computed over all the sequences of timestamps which optimally align \( S \) with \( S' \) (see [22] for more details).

 Second, we also estimated the \( n \times n \) covariance matrix

\[
C = \frac{1}{\tau-1} \sum_{t=1}^{\tau} (p(t) - \overline{p})(p(t) - \overline{p})^\top, \tag{2}
\]

related to any trial \( S \), where \( \overline{p} = \frac{1}{\tau} \sum_{s=1}^{\tau} p(s) \) averages all the \( \tau \) coordinates and we denote \( n = 3J \) for convenience. Since \( C \) is positive definite, we thus exploited the theory of the Riemannian manifold \( Sym^+_n \) and projected \( C \) onto the tangent space to obtain \( \tilde{C} \). Then, using the symmetry of \( \tilde{C} \), we extracted its independent entries, yielding the following \( n(n+1)/2 \) vector

\[
\text{COV} = [\tilde{C}_{11}, \ldots, \tilde{C}_{1n}, \tilde{C}_{21}, \ldots, \tilde{C}_{1n}, \ldots, \tilde{C}_{nn}] \tag{3}.
\]

Note that the usage of covariance is inspired by [8], which set the new state-of-the-art performance for action recognition from MoCap data. Also, our approach is similar to the case \( L = 1 \) in [21], where a \( L \)-layered temporal hierarchy of covariance descriptors is proposed, but differently from us, the projection stage onto the tangent space is not considered.

 For both representations, we used the support vector machine (SVM) for classification: when fed with COV, we normalized the data imposing zero mean and unit variance and we then used a linear kernel. Instead, the negative dynamic time warping kernel function [23] produced the training and testing Gram matrices given in input to the SVM.

III. EVALUATION STRATEGIES

As previously noticed, the common aspect across the considered datasets is that each subject performed the same action class several times (e.g., 2-3 times in the MSR-Action3D). Hence, we want to check if the additional knowledge of the identity of subject who is acting can boost the classification. Thus, we introduce the three different testing modalities adopted.

One-Subject-Out considers as testing data all the action instances belonging to one subject only, while the remaining trials are used in training. Consequently, the final classification results average all the subject-related intermediate scores. This is the more appropriate procedure in terms of generalization. For instance, it is fundamental for real-time 24h/7d video-surveillance application where the system should be able to recognize activities performed by never seen human agents. Moreover, One-Subject-Out is in line with the cross-subject test setting adopted in [8], [9].

In the Cross-Validation strategy we collect the data coming from any subject in a way that, for each of them, \( \frac{2}{3} \) of samples are used in training and the remaining \( \frac{1}{3} \) in testing. To guarantee robustness, the final classification results are averaged over 20 random choices for such a partition of the data. Such procedure can boost the action recognition accuracy since, during training, it exploits the information regarding

\(^2\)In all experiments, for the SVM cost parameter, we fixed \( C = 10 \).
all the subjects: in this way, the classifier can more easily
discriminates the action performed by a subject on which has
been already trained.

In the **Personalization** strategy, we suppose to have the same
number of classifiers as the number of the subjects present in
each dataset. Each classification model is specific of a given
subject and it is trained only on his/her activity instances: to
do this, once the classes with a number of trials less than 2 are
removed, we randomly choose $\frac{2}{3}$ of samples referring to every
action. Thus, we test on the remaining $\frac{1}{3}$. The final accuracy
score compares all the predicted and true labels, fusing the
results of all the specific classifiers. As previously done, we
average the classification results over 20 random splits of all
the subject-specific instances.

### IV. Experimental Results

In this Section, we compare **One-Subject-Out**, **Cross-
Validation** and **Personalization** by means of the two types of
features previously introduced in Section I. Table I and Table II
report the results related to DTW and COV, respectively.

In the majority of the comparisons, the COV descriptor obtains higher performance with respect to DTW. Nevertheless,
we can observe a common trend: the action classification
performance grows when switching from **One-Subject-Out** to **Cross-Validation**, reaching its peak with **Personalization**.
Moreover, such behavior is also independent from the features
adopted since showed by both DTW and COV. Additionally,
it is worth noting that the results reported in Tables I and II show that the accuracies obtained with the three different
modalities are inversely proportional to the number of the
samples used in the training phase. In all the cases, the lowest
performance is always scored adopting **One-Subject-Out**, although it has the largest number of training samples. This
result is expected, since **One-Subject-Out** has to extrapolate
more regular patterns from the data, namely finding subject-
invariant characteristics to recognize actions of unseen human
agents. In other words, in **One-Subject-Out** procedure, SVM
does not learn how a specific subject performs an action
but how a particular action is generally fulfilled. Conversely
the **Personalization** strategy obtains the best results for all
datasets. This procedure considers in the training step the
least number of samples, only regarding a specific subject.
In particular, if we focus on the acquisition details of MSR-
Action3D dataset (see Section I), very few trials per activity
performed by the same subject are available (sometimes only
a single action instance per subject is taken for training). Despite
this, **Personalization** scores 92.46% and 81.75% with COV
and DTW features respectively, and outperforms all the other
two strategies. Indeed, the other two datasets, MSRC-Kinect12
and HDM-05, are almost saturated using both descriptors (e.g.,
99.57 ± 0.16 of DTW on MSRC-Kinect12 and 99.02 ± 0.98
of COV on HDM-05).

**Cross-Validation** results deserve an own discussion since, in
all the comparisons, it always gives intermediate classification

---

A. Quantitative statistical analysis

To provide a better understanding about personalization vs.
generalization and inter- & intra-subject variability on the
considered datasets, we report some quantitative results
obtained by means of the following statistics.

- Once fixed the **Cross-Validation** strategy, we aim at
checking which source of information is at disposal of SVM
when an exact prediction is performed. Hence, we compute the
probability $p_{\text{subject}}$ that, on average, any correctly classified
testing sample and the training sample closest to it both belong
to the same subject. Clearly, high/low $p_{\text{subject}}$ values check if
testing on the same subjects used for training gives a pros/cons
for the classification, respectively.

- For each action class, if considering one trial and its
nearest in the feature space, we compute the probability that,
on average, they belong to the same subject. We call it $p_{\text{inter}}$.
since absolutely quantifying inter-subject variability, whose impact is negligible if $p_{\text{inter}} \approx 0$.

- To do the same with the intra-subject variability, we compute $p_{\text{intra}}$ by checking whether, on average, for any fixed subject, every two samples, which are the closest ones in the feature space, actually refer to different classes of actions. From the definition, if $p_{\text{intra}} = 0$, all the trials of a given action and a given subject are almost identical and intra-subject variability is totally absent.

- Finally, we devise a comprehensive metric for inter and intra-subject variability. Considering each action class a separately, we find the two elements of maximal mutual distance, namely $d_a$. Also, we compute the maximal distance $d_{a,s}$ between all trials of action $a$ performed by subject $s$. Then, once set $\Delta_{a,s} = \frac{|d_{a,s} - d_a|}{d_a}$, this metric spans the extremal cases $\Delta_{a,s} = 0$ and $\Delta_{a,s} = 1$. If $\Delta_{a,s} = 0$, then $d_{a,s} = d_a$ and the inter-subject variability is minimized: instances belonging to the same action but different subjects are generally close to each other. Conversely, when $\Delta_{a,s} = 1$, it implies $d_{a,s} = 0$ which minimizes the intra-subject variability since all the trials of action $a$ and subject $s$ collapse into a single point. Clearly, it means that subject $s$ performs action $a$ in the same manner across all the different trials. We name $\Delta$ the final statistic, averaging $\Delta_{a,s}$ over $a$ and $s$. Globally, $\Delta$ measures which out of inter or intra-subject variability is overwhelming (see Figure 2 for a better understanding).

Referring to all the aforementioned statistics $p_{\text{subject}}, p_{\text{inter}}, p_{\text{intra}}$ and $\Delta$, a notion of “closeness” is involved. Clearly, it depends to the type of used feature: for COV, the distance is the Euclidean one, since induced by a linear kernel. Instead, for DTW, we use the dynamic time warping distance $\delta$, as introduced in Section [11] Table [III] reports the values of all the previous statistics for all the considered datasets. We only report the values related to COV since no remarkable differences are registered when moving to DTW[5]. Table [III] gives us useful insights about the classification results reported in Tables [II] and [II]. All the datasets show a similar behavior.

Actually, in all cases, $p_{\text{subject}}$ is extremely high (e.g., 0.89 for HDM-05) and it indicates that each subject performs any action according to his/her own peculiarities which are effective when discriminating on the same subject. However, adding the information referring to other subjects can be misleading and this is motivated by the fact that Personalization scores superior to Cross-Validation: e.g., using COV on HDM-05, 99.02% for the former and 96.32% for the latter (Table [II]).

Differently, the high values of $p_{\text{intra}}$ certify that, in all the datasets, the inter-subject variability is high and it frequently happens that different subjects have their own fashion to perform a same action, so classifying a given action of an unknown subject can therefore be difficult. As a consequence, this explains why, sometimes, One-Subject-Out gives very poor performance (e.g., Table [I], DTW on MSR-Action3D).

The scored valued for $p_{\text{intra}}$ attests that, in all the considered datasets, the intra-subject variability is totally not problematic and, almost surely, each subject repeats a single action coherently along his/her trials. Also, such trend is confirmed by the excellent performance scored by Personalization.

In conclusion, the registered values for $\Delta$ are, in all the cases, quite close to 1. It means that all the instances of a given subject performing a single action occupy a compact region in the feature space. This is an opposed situation to a sparser configuration where the subjects are extremely shuffled. In other words, for all the datasets, different subjects can be considered as different sub-problems: Personalization applies the divide et impera principle, separately solving each of them. Conversely, One-Subject-Out looks for patterns which have to be, at the same time, subject-invariant and action-specific. Despite the latter approach ensures much more generalization, it is also much more harder than the former: this is why Personalization outperforms One-Subject-Out.

V. TWO-STAGE RECOGNITION PIPELINE

In all the experiments related to Personalization, we assume that the subject who is performing the given action is known a priori, so that we can easily choose the correct classification model trained over the specific human agent. In real-world applications, Personalization can be replaced with a two-stage process where 1) a unique SVM model (subject-SVM) classifies the identity of the subject and 2) the final stage of action recognition is performed by means of a set of subject-specific SVM classifiers (action-SVMs). Both subject-SVM and action-SVMs are trained on the same random partition of the data (collecting $2/3$ of the trial from any subject and action). Actually, for subject-SVM, no preliminary subject-dependent splitting is required, while, the trials belonging to a single subject at a time are used in the case of action-SVMs. In the testing phase, the subject-SVM predicts the agent label that indexes which action-SVMs classifier has to be used for the final action or activity recognition: when a subject is

![Fig. 2. In the feature space, we surround the region referring to a single action. Within, each point represents a trial and different colors relate to different subjects. Left: When $\Delta_{a,s} \approx 0$, inter-subject variability is minimized since, in general, trials from different subjects occupy nearby positions. Right: The case $\Delta_{a,s} \approx 1$ minimizes the intra-subject variability because all the instances of the same subject are compactly clustered.](Image 43x726 to 156x786)

### Table III

| Dataset         | $p_{\text{subject}}$ | $p_{\text{inter}}$ | $p_{\text{intra}}$ | $\Delta$  |
|-----------------|-----------------------|--------------------|---------------------|----------|
| MSR-Action3D    | 0.78                  | 0.86               | 0.19                | 0.71     |
| MSRC-Kinect12   | 0.97                  | 0.97               | 0.01                | 0.90     |
| HDM-05          | 0.89                  | 0.95               | 0.01                | 0.74     |

4For instance, the value of $p_{\text{subject}}$ for MSR-Actio3D is 0.77, for MSRC-Kinect12 is 0.97 and for HDM-05 is 0.85.
misclassified, the wrong model is used and, obviously, the action classification may be wrong.

To validate our proposed pipeline, both subject-SVM and action-SVMs are fed with COV features, more powerful than DTW. The results in Table IV provide the mean and standard deviation of the accuracies scored in the two steps separately, over 20 different random partitions of the data. Since COV were originally suited for action recognition, such descriptors are not optimal for human identification. In fact, in all the datasets, the average accuracy of subject-SVM (second row in Table IV) does not exceed 91% and, clearly, more suitable biometric descriptors could improve it. Finally the action-SVM results (in bold in Table IV) deviate from the classification scores of Cross-Validation and Personalization (Table III). e.g. in MSRC-Kinect12, the action recognition score of Personalization and the two-stage pipeline is 99.65 ± 0.07 and 97.14 ± 0.39, respectively. As a final remark, it is notable that, despite the off-the-shelf feature employed, the mean scored performance in action classification of the two-stage pipeline is superior to the state-of-the-art method [8] on MSRC-Kinect12 (92.3%) and HDM05 (96.8%).

VI. CONCLUSIONS

In this paper, we investigated the generalization capability of automatic action and activity recognition systems focusing on Personalization in a comparison with standard Cross-Validation and One-Subject-Out strategies. To this aim, we exploited DTW and COV on MSR-Action3D, MSRC-Kinect12 and HDM-05 benchmark MoCap datasets.

From the experiments, One-Subject-Out resulted the more challenging strategy, although being able to ensure a major generalizable. Differently, despite Cross-Validation was actually boosted from the usage of the same subject in both training and testing, the additional information relative to the other subjects could mislead. Finally, the Personalization strategy, gave the highest performance, despite the lowest number of instances used in training.

In addition, we also provided several quantitative statistics to measure inter and intra-class variability on the considered datasets: as a result, the latter is almost marginal, while the former is the actual burden that has to be tackled when devising new techniques.

Finally, we proposed a two-step classification pipeline by first identifying the subject and second using subject-specific classifiers for the actual action recognition. This promising result may pave the way of a new paradigm for the design of action/activity recognition systems, also embracing the possibility to exploit these findings for custom human-specific action recognition systems.

REFERENCES

[1] R. Poppe, “A survey on vision-based human action recognition,” Image Vision Comput., vol. 28, no. 6, pp. 976–990, Jun. 2010.
[2] L. Lo Presti and M. La Cascia, “3D skeleton-based human action classification: A survey,” Pattern Recognition, vol. 53, pp. 130 – 147, 2016.
[3] M. Sandri, M. Leordeanu, and C. Sminchisescu, “The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection,” in ICCV, 2013.
[4] L. Xia, C.-C. Chen, and J. Aggarwal, “View invariant human action recognition using histograms of 3d joints,” in CVPR workshop, 2012.
[5] F. Olth, R. Chaudhry, G. Kurillo, R. Vidalt, and R. Bajcsy, “Sequence of the most informative joints (SMI): A new representation for human skeletal action recognition,” in CVPR workshop, 2013.
[6] R. Slama, H. Wannous, M. Daoudi, and A. Srivastava, “Accurate 3d action recognition using training on the grassmann manifold,” Pattern Recognition, vol. 48, no. 2, pp. 556–567, 2015.
[7] N. Raman and S. J. Maybank, “Action classification using a discriminative multilevel hdp-hmm,” Neurocomputing, vol. 154, pp. 149–161, 2015.
[8] L. Wang, J. Zhang, L. Zhou, C. Tang, and W. Li, “Beyond covariance: Feature representation with nonlinear kernel matrices,” in ICCV, 2015.
[9] W. Li, Z. Zhang, and Z. Liu, “Action recognition based on a bag of 3d points,” in CVPR workshop, 2010.
[10] S. Fotherhill, H. M. Mentis, P. Kohli, and S. Nowozin, “Instructing people for training gestural interactive systems,” in ACM Conference on Computer-Human Interaction, 2012.
[11] M. Müller, T. Röder, M. Clausen, B. Eberhardt, B. Kräger, and A. Weber, “Documentation mocap database HDM-05,” Universität Bonn, Tech. Rep. CG-2007-2, June 2007.
[12] K. Sanfins and C.-H. Lee, “A structural bayes approach to speaker adaptation,” IEEE Transactions on Speech and Audio Processing, vol. 9, pp. 276–287, 2001.
[13] S. Connell and A. Jain, PAMI, vol. 24, no. 3, pp. 329–346, 2002.
[14] A. Yao, L. Van Gool, and P. Kohli, “Gesture recognition portfolios for personalization,” in CVPR, 2014.
[15] L. Bao and S. S. Intille, “Activity recognition from user-annotated acceleration data,” in Pervasive Computing, 2004.
[16] A. Dalton and G. Laighin, “Comparing supervised learning techniques on the task of physical activity recognition,” IEEE Journal of Biomedical and Health Informatics, vol. 17, no. 1, pp. 46–52, 2013.
[17] P. E. Taylor, G. J. M. Almeida, T. Kanade, and J. K. Hodgins, “Classifying human motion quality for knee osteoarthritis using accelerometers,” in EMBC, 2010.
[18] P. Tormene, T. Giorgino, S. Quaglini, and M. Steanelli, “Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation,” Artificial Intelligence in Medicine, vol. 45, no. 1, pp. 11 – 34, 2009.
[19] A. Yurtman and B. Barshan, “Automated evaluation of physical therapy exercises using multi-template dynamic time warping on wearable sensor signals,” Computer Methods and Programs in Biomedicine, vol. 117, no. 2, pp. 189 – 207, 2014.
[20] B. Barshan and A. Yurtman, “Investigating inter-subject and inter-activity variations in activity recognition using wearable motion sensors,” The Computer Journal, 2015.
[21] M. Hussein, M. Torki, M. Gowayed, and M. El-Saban, “Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations,” JICAI, 2013.
[22] M. Müller, “Dynamic time warping,” in Information Retrieval for Music and Motion. Springer, 2007.
[23] S. Gandumussud, T. P. Runarsson, and S. Sigurdsson, “Support Vector Machines and Dynamic Time Warping for Time Series,” in IJCNN, 2008.
[24] S. Jayasumana, R. Harlley, M. Salzmann, H. Li, and M. Harandi, “Kernel Methods on the Riemannian Manifold of Symmetric Positive Definite Matrices,” in CVPR, 2013.