Revisiting Human Action Recognition: Personalization vs. Generalization

Andrea Zunino[1,2(✉)], Jacopo Cavazza[1,2], and Vittorio Murino[1,3]

[1] Pattern Analysis and Computer Vision (PAVIS), Istituto Italiano di Tecnologia, 16163 Genova, Italy
{andrea.zunino,jacopo.cavazza,vittorio.murino}@iit.it
[2] Electrical, Electronics and Telecommunication Engineering and Naval Architecture Department (DITEN), Università degli Studi di Genova, 16145 Genova, Italy
[3] Computer Science Department, Università di Verona, 37134 Verona, Italy

Abstract. By thoroughly revisiting the classic human action recognition paradigm, we analyzed different training/testing strategies, discovering that standard (cross-validating) testing strategies are not always the suitable validation procedures to assess an algorithm's performance. As a consequence, we design a novel action recognition architecture, applying a "personalized" strategy to learn how any subject performs any action. We discover that 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. Leveraging on that, we 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 score a favorable performance with respect to the state-of-the-art as to certify the soundness of our approach.

Keywords: Action recognition · Kinematic analysis · Generalization · Personalization strategy

1 Introduction

The video-based classification of human actions is a very complex task due to contextual clutter and noise, illumination variations, occlusions, and the implicit variability and complexity of actions. All these problems can be mitigated by the three-dimensional (3D) sensor technology, which allows to capture human motion at high spatial/temporal resolution (VICON), with good accuracy and low cost (Kinect). As a consequence, the development and improvement of computational approaches for 3D action recognition sharply rose in the recent year [12].

Within the context of 3D action recognition, this work undertakes a revisiting perspective, 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

S. Battiato et al. (Eds.): ICIAP 2017, Part I, LNCS 10484, pp. 469–480, 2017.
https://doi.org/10.1007/978-3-319-68560-1_42

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 [11], MSRC-Kinect12 [6] and HDM-05 [13]. 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 [15], handwriting [4], and gesture [10,19] recognition.

Among the few works which studied the variability within/across subjects, for instance, [1] did not register a strong impact of different subjects in daily activities classification, and [5] documented the stability of the performance on an *ad hoc* acquired dataset characterized by biometric homogeneity of the participants. Differently, in [16], 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 [17,20] for computer assisted rehabilitation tasks, as well as by [2] which performed a theoretical dissertation about within-subject and across-subjects noise using wearable motion sensors. Globally, [1,2,5,16,17,20] 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 robots and automatic systems.

These arguments are investigated through the following main contributions.

(*i*) We analyze the role of the individual subject in human action recognition. By considering MSR-Action3D [11], MSRC-Kinect12 [6] and HDM-05 [13] benchmark datasets, we propose a novel testing strategy, called ***Personalization***, where action classification is performed by considering the instances belonging to one specific subject at a time. We register a superior performance of ***Personalization*** while comparing it against ***One-Subject-Out***, which left out the data of one subject as the test set, and ***Cross-Validation***, where testing is performed on all subjects (which are also used for training).

(*ii*) In order to explain the latter performance and analyze the role of subjectiveness, we introduce a quantitative statistical analysis. This allows to evaluate the impact of retrieving in testing all the subjects used in the training phase, ultimately assessing the role played by either *inter-* or *intra-subject variability*.

(*iii*) Capitalizing on our improved understanding, we boost action recognition by learning the subject's identity. In particular, we propose a two-stage recognition pipeline (Fig. 1) where the preliminary estimation 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 methods [18].

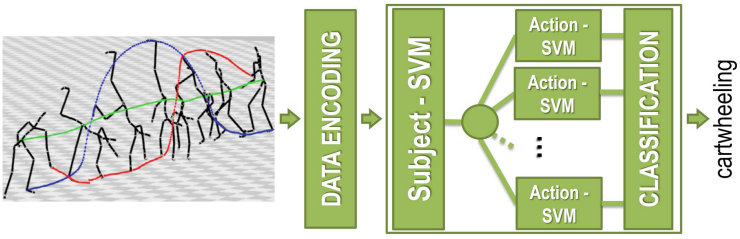

Fig. 1. As opposed to the generic recognition of an action performed by an unspecified human agent, we investigate a counterpart approach in which the action recognition accuracy is boosted by adopting a "personalization" 2-stage method, where the subject is first identified, followed by the actual classification of the action.

The rest of the paper is organized as follows. In Sect. 2, we present the considered datasets and the features adopted, and the evaluation strategies investigated

are reported in Sect. 3. Section 4 presents and widely discusses the experimental results, and we illustrate the aforementioned two-stage classification pipeline in Sect. 5. Finally, Sect. 6 draws the conclusions of this study.

2 Datasets and Feature Encoding

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

- **MSR-Action3D** [11] 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 2 or 3 times. In total, we used 544 sequences [8].
- **MSRC-Kinect12** [6] is a relatively large dataset of 3D skeleton data, recorded by means of a Kinect sensor. The dataset has 5881 sequences, containing 12 action classes performed by 30 different subjects. 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** [13], the number of skeleton joints is $J = 31$, each action is repeated 5 times on average by each of the 5 subjects involved during the acquisition through a VICON system. We followed the 14-classes experimental protocol of [8,18].

For all the aforementioned datasets, each trial can be formalized as a collection $\mathbf{S}$ of τ different acquisitions $\mathbf{p}(1), \ldots, \mathbf{p}(\tau)$. For any $t = 1, \ldots, \tau$, we denote with $\mathbf{p}(t)$ the column vector which stacks $\mathbf{p}_1(t), \ldots, \mathbf{p}_J(t) \in \mathbb{R}^3$, 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 [7,14]. In order to apply DTW, we evaluated the differences between any two joints collection $\mathbf{S} = [\mathbf{p}(1), \ldots, \mathbf{p}(\tau)]$ and $\mathbf{S}' = [\mathbf{p}'(1), \ldots, \mathbf{p}'(\tau')]$ through the following distance

$$d(\mathbf{p}(s), \mathbf{p}'(t)) = \frac{1}{J} \sum\nolimits_{j=1}^{J} \|\mathbf{p}_j(s) - \mathbf{p}'_j(t)\|, \tag{1}$$

where $\|\cdot\|$ is the Euclidean norm, $s = 1, \ldots, \tau$ and $t = 1, \ldots, \tau'$. The final similarity measure, provided by DTW to compare $\mathbf{S}$ and $\mathbf{S}'$, is $\delta(\mathbf{S}, \mathbf{S}')$ which is the minimum value of (1) computed over all the sequences of timestamps which optimally align $\mathbf{S}$ with $\mathbf{S}'$ (see [14] for more details).

Second, we also estimated the $n \times n$ covariance matrix

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

related to any trial $\mathbf{S}$, where $\overline{\mathbf{p}} = \frac{1}{\tau} \sum_{s=1}^{\tau} \mathbf{p}(s)$ averages all the τ coordinates and we denote $n = 3J$ for convenience. Since $\mathcal{C}$ is positive definite, we thus exploited the theory of the Riemannian manifold Sym_n^+ and projected (2) onto the tangent space to obtain $\widetilde{\mathcal{C}}$ [9]. Then, using the symmetry of $\widetilde{\mathcal{C}}$, we extracted its independent entries, yielding the following $n(n+1)/2$ vector

$$\mathrm{COV} = [\widetilde{\mathcal{C}}_{11}, \ldots, \widetilde{\mathcal{C}}_{1n}, \widetilde{\mathcal{C}}_{21}, \ldots, \widetilde{\mathcal{C}}_{2n}, \ldots, \widetilde{\mathcal{C}}_{nn}]. \tag{3}$$

Note that the usage of covariance is inspired by [18], 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 [8], 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[1] (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 [7] produced the training and testing Gram matrices given in input to the SVM.

This will allow us to validate the testing strategies using the same basic classification approach with two different descriptors.

3 Evaluation Strategies

We compare the following three testing modalities.

For testing, ***One-Subject-Out*** considers any action instance belonging to one subject separately, the system being training on the remaining ones. The final classification results average all the subject-out intermediate scores. This is in line with the protocols of [3,11,18].

In the ***Cross-Validation*** strategy, we performed a subject-balanced shuffling of data. Precisely, for each subject $\frac{2}{3}$ 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 the aforementioned partitions[2].

For the ***Personalization*** strategy, each model is trained on the action instances of a single subject at a time. To do this, we fix a subject and, for any action class, $\frac{2}{3}$ of samples are used in training, testing on the remaining $\frac{1}{3}$. Classification accuracies (in testing) are computed on each subject separately, finally fusing the single scores. As previously done, we average the classification results over 20 random splits of all the subject-specific instances.

[1] In all experiments, for the SVM cost parameter, we fixed $C = 10$.

[2] For the sake of clarity, please note that a test sample is never seen by the system in training.

4 Experimental Results and Discussion

In this Section, we compare *One-Subject-Out*, *Cross-Validation* and *Personalization*, using the descriptors of Sect. 2: the results related to DTW and COV are reported in Tables 1 and 2, respectively.

Table 1. DTW classification accuracies on the three MoCap datasets. Mean and standard deviation are reported in percentages for each testing strategies (best results are in bold).

Testing strategy	MSR-Action3D	MSRC-Kinect12	HDM-05
One-Subject-Out	28.42 ± 12.76	51.73 ± 17.58	92.39 ± 3.60
Cross-Validation	57.90 ± 3.07	66.93 ± 1.81	96.93 ± 1.72
Personalization	$\mathbf{81.75 \pm 2.71}$	$\mathbf{99.57 \pm 0.16}$	$\mathbf{97.59 \pm 0.85}$

Table 2. COV classification accuracies on the three MoCap datasets. Mean and standard deviation are reported in percentages for each testing strategies (best results are in bold).

Testing strategy	MSR-Action3D	MSRC-Kinect12	HDM-05
One-Subject-Out	70.49 ± 9.02	92.47 ± 6.01	87.78 ± 7.04
Cross-Validation	77.18 ± 3.59	98.57 ± 0.30	96.32 ± 1.97
Personalization	$\mathbf{92.46 \pm 1.09}$	$\mathbf{99.65 \pm 0.07}$	$\mathbf{99.02 \pm 0.98}$

In most case, the COV obtains higher performance with respect to DTW. 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*. Since common to both DTW and COV, such behavior is actually independent from the data representation.

It is worth noting that the *ranking in the accuracies obtained with the three different modalities is inversely depending on the number of the samples used in the training phase*.

Indeed, in both Tables 1 and 2, the lowest performance is always scored by *One-Subject-Out*, although such modality adopts the larger amount of training data if compared to either *Cross-Validation* or *Personalization*. The reason is that *One-Subject-Out* has to extrapolate more from the data, finding action-specific patterns which are also subject-invariant. Differently, the *Personalization* strategy is required to find action-specific patterns, totally neglecting intra-subject generalization. This helps explaining why *Personalization* obtains the best results for all datasets. Note that the latter fact occurs despite the *Personalization* strategy exploits the least number of samples within *One-Subject-Out* and *Cross-Validation*. In particular, by considering MSR-Action3D dataset (see Sect. 2), very few trials (and sometimes only one) are available per each action

class and subject. In spite of that, *Personalization* scores 92.46% and 81.75% with COV and DTW respectively, and outperforms all the other two strategies. Indeed, MSRC-Kinect12 and HDM-05 are almost saturated by *Personalization*: *e.g.*, 99.57 ± 0.16 of DTW and 99.02 ± 0.98 of COV respectively.

Cross-Validation deserves an own discussion. Indeed, such strategy can be seen as a compromise between the two, since each subject is seen in both training and testing (as in *Personalization*) but is required to generalize across agents (as in *One-Subject-Out*). In terms of registered performance, *Cross-Validation* scores intermediately with respect to the other two strategies. Precisely, with respect to *One-Subject-Out*, *Cross-Validation* improves by margin: therefore, exploiting the same subject in both training and testing appears to be effective.

However, all *Cross-Validation* accuracies are always lower than the *Personalization* one, although the gap between them is sometimes very small (*e.g.*, *Cross-Validation* scores about 1% less with respect to *Personalization* on MSRC-Kinect12 dataset, see Table 2). Actually, this can be interpreted in the following manner: *adding many training samples belonging to different subjects does not always lead to an improvement, frequently confusing the (SVM) classifier.*

Evidently, the quality of the data is superior to quantity for the sake of performance. In the next Section, we will carry out a statistical analysis to characterize the concept of "quality" in terms of *inter-* and *intra-subject variability.*

4.1 Quantitative Statistical Analysis

Let us define the following statistics.

① $\mathbf{p_{subject}}$ For all testing action instances $\bar{\mathfrak{a}}$, which are correctly classified in *Cross-Validation*, consider the training action instance $\bar{\mathbf{a}}$ which is closest to $\bar{\mathfrak{a}}$. We call $\mathbf{p_{subject}}$ the (average) probability that both $\bar{\mathfrak{a}}$ and $\bar{\mathbf{a}}$ belongs to the same subject.

Clearly, $\mathbf{p_{subject}}$ measures how often a good prediction is obtained by exploiting the information exactly coming from the same subject. Hence, high/low $\mathbf{p_{subject}}$ values check if testing on the same subjects used for training gives a pros/cons for the classification, respectively.

② $\mathbf{p_{inter}}$ For each action class c, and for any instance $\mathfrak{a}_c$ of that class, consider the instance $\mathbf{a}_c$ (still belonging to the same class) which is closest to $\mathfrak{a}_c$ in the features space. While averaging on c, the frequency of that $\mathfrak{a}_c$ and $\mathbf{a}_c$ belonging to the *same subject* is denoted by $\mathbf{p_{inter}}$.

We can notice that $\mathbf{p_{inter}} \approx 0$ when *inter-subject variability* is negligible.

③ $\mathbf{p_{intra}}$ For any subject s and for any instance $\mathfrak{a}_s$, consider $\mathbf{a}_s$ which is the closest to $\mathfrak{a}_s$ within the ones in the dataset which belongs to the s-th subject. p_{intra} counts how frequently, $\mathfrak{a}_s$ and $\mathbf{a}_s$ belong to a *different action class.*

From the definition, if $\mathbf{p_{intra}} = 0$, all the trials of a given action and a given subject are almost identical and *intra-subject variability* is totally absent.

④ $\boldsymbol{\Delta}$ For each action class c, compute d_c as the maximal distance between two c-labelled elements in the dataset. Similarly, $d_{c,s}$ is the maximal distance of two c-labelled instances from the same subjet s. Define $\Delta_{c,s} = \frac{|d_{c,s} - d_c|}{d_c}$.

We have $0 \leq \Delta_{c,s} \leq 1$, where the extremal case $\Delta_{c,s} = 0$ correspond to a null *inter-subject variability*: since $d_{c,s} = d_c$, within the trials of class c, subjects are maximally shuffled (Fig. 2, left). Also, $\Delta_{c,s} = 1$ implies $d_{c,s} = 0$ which minimizes the *intra-subject variability* since all instances of class c from subject s collapse to a point (Fig. 2, right). We define $\boldsymbol{\Delta}$ as the average of all $\Delta_{c,s}$ over c and s. By construction, $\boldsymbol{\Delta}$ quantifies the relative importance between *inter-* and *intra-subject variability*, being the latter or the former preponderant on the other in case of low or high $\boldsymbol{\Delta}$ values, respectively.

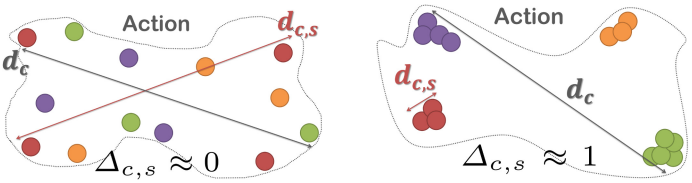

Fig. 2. In the feature space, we surround the region referring to a single action. Within, each point represents a trial and different colors relate to different subjects. Left: When $\Delta_{c,s} \approx 0$, *inter-subject variability* is minimized since, in general, trials from different subjects occupy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

In the definition of $\mathbf{p_{subject}}$, $\mathbf{p_{inter}}$, $\mathbf{p_{intra}}$ and $\boldsymbol{\Delta}$, a notion of "closeness" is involved. The latter depends on the exploited data representation. For COV, the distance is the Euclidean one, since induced by a linear kernel. Instead, for DTW, we use the dynamic time warping distance δ, as introduced in Sect. 2.

Discussion. Table 3 shows the values of our statistics in all the considered datasets. We only report the values related to COV since no remarkable differences are registered when moving to DTW[3].

Table 3. Quantitative evaluation of *inter* and *intra-subject variability*.

Dataset	$\mathbf{p_{subject}}$	$\mathbf{p_{inter}}$	$\mathbf{p_{intra}}$	$\boldsymbol{\Delta}$
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

In all cases, $\mathbf{p_{subject}}$ is extremely high (*e.g.*, 0.89 for HDM-05). Therefore, in *Cross-Validation* testing strategy, the performance is actually boosted by leveraging on how each subject perform a given action. Therefore, the scored $\mathbf{p_{subject}}$ values attest that the role of the subject is crucial in 3D action recognition.

[3] For instance, the value of p_{subject} for MSR-Actio3D is 0.77, for MSRC-Kinect12 is 0.97 and for HDM-05 is 0.85.

Inter-subject variability is a problem ($\mathbf{p_{inter}} > .85$). Thus, the same action is likely to be performed very differently by different subjects. This explains the difficulty of *One-Subject-Out* strategy.

On MSR-Action3D $\mathbf{p_{intra}}$ is low, being actually almost zero in the other cases. Especially in MSRC-Kinect12 and HDM-05, each subject identically repeats each action almost in the same way. As a consequence, *intra-subject variability* is not remarkably affecting the classification. Hence, even knowing one only action instance per subject can actually boost the recognition. This explains the favorable *Personalization* performance, despite the small data regime embraced.

Inter-subject variability is the actual burden to tackle, being totally overwhelming with respect to intra-subject one. The high values for $\boldsymbol{\Delta}$ (*e.g.*, 0.9 for MSRC-Kinect12) certify that the gap to fill across subject is actually remarkable, where the challenges related benchmark datasets analyzed can be intuitively imagined as in Fig. 2, right.

Globally, if we can automatically recognize the subject's identity of a training/testing instance, we can cast action recognition as an easier subproblem: we do not have to fill huge inter-subjects gaps, but just learning how to discriminate different actions of the same subjects (which are likely to be more separable). As we will prove in the next Section, such divide et impera strategy is very effective.

5 Divide et Impera. Two-Stage Recognition Pipeline

In comparison to *Cross-Validation* and *One-Subject-Out*, the *Personalization* strategy always achieves the best scores (Tables 1 and 2). As explained, this happens because *inter-subject variability* is highly problematic, being *intra-subject variability* small as in MSR-Action3D and eventually absent in the other cases. However, *Personalization* leverage on the unfavorable assumption: it requires the subject's identity to be known in order to classify the action.

Actually, in this Section we tackle this issue, obtaining an equivalently effective action recognition system, which is now able to operate in real-world conditions. The key is *learning the subject's identity.*

Inspired by our findings (Sect. 4.1), we posit that we can proficiently apply features designed for action representation in order to recognize the subject's identity. This originate a divide et impera paradigm where, first the subject's identity is recognized and then action recognition is performed using a subject-specific classifier, trained on the instance of a single subject only. Despite the reduced amount of data, the task should be easier to train due to the better separability of action classes when the subject's identity is fixed. Precisely, we propose the following two-stage pipeline (Fig. 1).

Stage 1. A unique SVM model (*subject-SVM*) recognizes subject's identity.
Stage 2. Within many subject-specific action classifiers (called *action-SVMs*), the final action recognition step is performed by the one corresponding to the subject identified in **Stage 1.**

For training *subject-SVM* and *action-SVMs*, we performed a $\frac{2}{3}/\frac{1}{3}$ random splitting for training and testing data related to any subject and any action. Obviously, for each of the *action-SVMs*, we used only the training and testing examples belonging to one subject at a time. During testing, the *subject-SVM* scores is used to select one of the *action-SVMs* (actually the one corresponding to the recognized subject): this is the model exploited for action classification.

To validate our proposed pipeline, both *subject-SVM* and *action-SVMs* are fed with COV features, more powerful than DTW. The results in Tables 4, 5 provide the mean and standard deviation of the accuracies scored in the two steps separately, over 20 different random partitions of the data.

Discussion. Since COV is designed for action recognition, it is suboptimal for subjects' identification. In fact, despite the classification performance we registered is still reliable (Table 4), when a subject is misclassified, the action classifier corresponding to another subject is used and performance can deteriorate.

Nevertheless, we only registered a 2% the drop with respect to *Personalization* strategy, which can be considered as our two-stage pipeline with perfect subject recognition in the first stage. Such performance is remarkable since, after all, *Personalization* requires the subjects' identity to be known, whereas we are effectively able to automatically learn it[4].

Although a comparison of our simple approach with more sophisticated approaches [3,8,18] is challenging, we score a favorable performance with respect to the state-of-the-art. Despite the simplicity of our pipeline, we only pay 6% on MSR-Action3D (96.9%, [18]). This is coherent with the fact that *intra-subject variability* is not totally absent in such a case ($\mathbf{p_{intra}} \approx 0.2$ in Table 3), therefore mining the underlying assumption of our approach. Differently, we are scoring

Table 4. Two-stage recognition pipeline - subject identification accuracies.

	MSR-Action3D	MSRC-Kinect12	HDM-05
subject-SVM	90.74 ± 2.41	85.18 ± 0.55	85.67 ± 3.18

Table 5. Two-stage recognition pipeline - action classification accuracies compared to SoA.

	MSR-Action3D	MSRC-Kinect12	HDM-05
action-SVMs	90.46 ± 1.17	$\mathbf{97.14 \pm 0.39}$	97.03 ± 1.36
SoA	**96.9** [18]	95.0 [3]	**98.1** [3]

[4] To have a better insight of the importance of the knowledge of the subject who is performing the action, we have conducted an experiment on MSRC-Kinect12 using COV features where we assume that the correct *action-SVM* is not available. Using the best *action-SVMs* belonging to all other subjects the performance drops from 97.14% to 80.68%.

almost on par with respect to [3] (98.1%) on HDM-05, also improving the state-of-the-art on MSRC-Kinect12 by about 2% (95.0%, [3]).

6 Conclusions

In this paper, we investigated the generalization capability of automatic activity recognition systems analyzing the proposed *Personalization* strategy in comparison with standard *Cross-Validation* and *One-Subject-Out* approaches. To this aim, we exploit classical representations (DTW and COV), with basic a classifier (linear SVM) on the MSR-Action3D, MSRC-Kinect12 and HDM-05 benchmark datasets.

From the experiments, *One-Subject-Out* resulted the more challenging strategy, although being able to ensure a better generalization. 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. 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, by using subject-specific classifiers for action recognition. This paradigm can be applied to general surveillance tasks, by monitoring the activities of unknown subjects by means of the model corresponding to the most similar training subject. Additionally, this opens to the design of custom human-robotic systems and novel authentication procedures.

References

1. Bao, L., Intille, S.S.: Activity recognition from user-annotated acceleration data. In: Ferscha, A., Mattern, F. (eds.) Pervasive 2004. LNCS, vol. 3001, pp. 1–17. Springer, Heidelberg (2004). doi:10.1007/978-3-540-24646-6_1
2. Barshan, B., Yurtman, A.: Investigating inter-subject and inter-activity variations in activity recognition using wearable motion sensors. Comput. J. **59**, 1345–1362 (2015)
3. Cavazza, J., Zunino, A., San Biagio, M., Murino, V.: Kernelized covariance for action recognition. In: ICPR (2016)
4. Connell, S., Jain, A.: Writer adaptation for online handwriting recognition. PAMI **24**(3), 329–346 (2002)
5. Dalton, A., ÓLaighin, G.: Comparing supervised learning techniques on the task of physical activity recognition. IEEE J. Biomed. Health Inf. **17**(1), 46–52 (2013)
6. Fothergill, S., Mentis, H.M., Kohli, P., Nowozin, S.: Instructing people for training gestural interactive systems. In: ACM Conference on Computer-Human Interaction (2012)

7. Gudmundsson, S., Runarsson, T.P., Sigurdsson, S.: Support vector machines and dynamic time warping for time series. In: IJCNN (2008)
8. Hussein, M., Torki, M., Gowayyed, M., El-Saban, M.: Human action recognition using a temporal hierarchy of COV descriptors on 3d joint locations. In: IJCAI (2013)
9. Jayasumana, S., Hartley, R., Salzmann, M., Li, H., Harandi, M.: Kernel methods on the riemannian manifold of symmetric positive definite matrices. In: CVPR (2013)
10. Joshi, A., Ghosh, S., Betke, M., Pfister, H.: Hierarchical bayesian neural networks for personalized classification. In: NIPS workshop (2016)
11. Li, W., Zhang, Z., Liu, Z.: Action recognition based on a bag of 3d points. In: CVPR workshop (2010)
12. Lo Presti, L., La Cascia, M.: 3D skeleton-based human action classification: a survey. Pattern Recogn. **53**, 130–147 (2016)
13. Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., Weber, A.: Documentation mocap database HDM-05. Technical report CG-07-2, June 2007
14. Albrecht, T.: Dynamic time warping. In: Information Retrieval for Music and Motion, pp. 69–84. Springer (2009)
15. Shinoda, K., Lee, C.H.: A structural bayes approach to speaker adaptation. IEEE Trans. Speech Audio Process. **9**, 276–287 (2001)
16. Taylor, P.E., Almeida, G.J.M., Kanade, T., Hodgins, J.K.: Classifying human motion quality for knee osteoarthritis using accelerometers. In: EMBC (2010)
17. Tormene, P., Giorgino, T., Quaglini, S., Steanelli, M.: Matching incomplete time series with dynamic time warping: an algorithm and an application to post-stroke rehabilitation. Artif. Intell. Med. **45**(1), 11–34 (2009)
18. Wang, L., Zhang, J., Zhou, L., Tang, C., Li, W.: Beyond covariance: feature representation with nonlinear kernel matrices. In: ICCV (2015)
19. Yao, A., Van Gool, L., Kohli, P.: Gesture recognition portfolios for personalization. In: CVPR (2014)
20. Yurtman, A., Barshan, B.: Automated evaluation of physical therapy exercises using multi-template dynamic time warping on wearable sensor signals. Comput. Methods Programs Biomed. **117**(2), 189–207 (2014)