Person Identification from Action Styles

Igor Kviatkovsky
Technion
Haifa, Israel
kviat@cs.technion.ac.il

Ilan Shimshoni
The University of Haifa
Haifa, Israel
ishimshoni@mis.haifa.ac.il

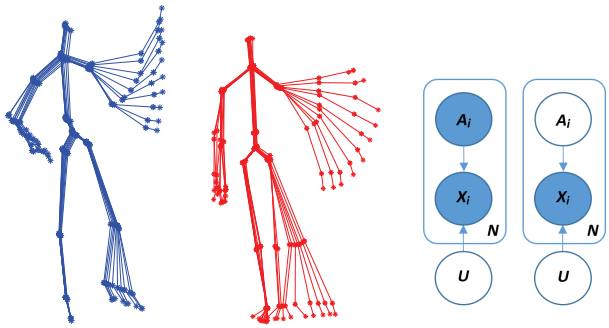
Ehud Rivlin
Technion
Haifa, Israel
ehudr@cs.technion.ac.il

Abstract

We consider a problem of identifying people based on their styles in performing actions from an arbitrary predefined set of action types. We present a generative model describing the action instance creation process and derive a probabilistic identity inference scheme, which implicitly includes action type inference as one of its components. Our experiments validate the power of the approach. We report high recognition rates on four publicly available action recognition datasets and one dataset for person authentication, on which we obtain state-of-the-art results. We make use of existing action representations and show that combining them with an action-specific Mahalanobis metric, learned from examples, improves the results.

1. Introduction

A typical question, action recognition is concerned with, is: "What is the performed action?". Another important question we consider in this work is: "Who is the person performing the action?". It was shown that gait motion patterns recorded using accurate motion capture (mocap) sensors, attached to skeletal joints, convey significant information which may be used to infer the subject's identity and demographic attributes [20, 21, 18]. In this work we address the problem of inferring persons' identities based on their style in performing arbitrary actions, captured using Kinect [17]. Unlike 3D skeletons captured using mocap, those recorded with Kinect are much noisier, and yet proved themselves useful for action recognition [24, 25, 29, 16]. However, a much higher precision is required for differentiating between instances of a particular action, performed by different users, than instances of different actions. Thus, it is uncertain that such noise is insignificant when actions of the same type, varying in style, are compared. Figs. 1(a) and 1(b) show two persons, a male and a female, performing a hand wave gesture. Looking at the trajectories of the limbs' joints, one can easily notice the stylistic differences between the performers, pointing on their gender as well as identity.



(a) (b) (c)

Figure 1. Two users – a male (a) and a female (b) – perform a "hand wav" gesture. (c) - Generative models describing the formation process of N action instances. X_i depends on the action label A_i and the user's identity label U. In the left diagram the labels A_i are known and the user's label U is unknown. In the right diagram the labels A_i and U are unknown.

1.1. Action Recognition

In the early works on action recognition articulated pose estimation was used to model the action, *e.g.*, [30]. However, due to the high complexity of the articulated body modeling, the attention has switched to appearance based approaches [2, 10, 3]. Thanks to the recent developments in low cost depth sensors and the accompanying skeleton tracking technologies [17], model based approaches regained their popularity [25, 29, 16, 24]. Yao *et al*. [31] show that pose-based action recognition typically outperforms appearance based methods even when pose estimates suffer from high levels of noise. Wang *et al*. [25] employ a multiscale Fourier temporal pyramid on pairwise relative joint positions gaining invariance to noise and temporal misalignment. Xia *et al*. [29] represent a pose as a histogram of normalized joint positions in 3D, and model an action as an evolution of poses using HMM. Ofli *et al*. [16] present a method for automatic selection of the most infor-

mative joints for a specific type of action. Hussein *et al*. [7] report impressive results using covariance descriptors extracted from sequences of absolute joint positions in multiple scales. In the recent work by Vemulapalli *et al*. [24] the authors propose a combination of Dynamic Time Warping (DTW) [23], a Fourier Temporal Pyramid (FTP) [25] and a linear SVM classifier to achieve state-of-the-art performance in action recognition from 3D skeletons.

1.2. Individual Recognition

Person identification based on motion is a widely studied topic in the computer vision community [26, 6, 15, 21]. Most of these efforts focus on analyzing motion of a particular type, *i.e*., locomotions, paying special attention to human gait due to its applicability in the surveillance applications domain. Early works on gait-based recognition used silhouette sequences extracted from video frames, *e.g*., [26]. The emerging depth sensors made other modalities, such as depth [6] and 3D skeletons [15], available for this task. Munsell *et al*. [15] use sequences of 3D skeletal representations from the Kinect to model two types of locomotions, *i.e*., walking and running. Given a test sequence, the authors use a two-step classification process in which they first classify the locomotion type and then apply a locomotion-specific identity classifier to identify the individual.

Inspired by the seminal work on biological motion perception by Johansson [8], many efforts were made to classify person identities and personal attributes from gait patterns, collected using accurate mocap sensors attached to major body joints [20, 21, 18]. Troje [20] uses gait patterns to classify the walker's gender. The gait patterns are modeled using the frequency and the amplitude of the trajectories' four major principal components. One of the conclusions of this work is that the walkers' gender is better identified using their gait dynamics than their skeletal structure. The follow-up work by Troje *et al*. [21] extends the approach to identify other personal attributes. Sigal *et al*. [18] automatically infer these attributes from videos of walking people using a 3D pose tracker.

In contrast to the surveillance applications' nature, where subjects are unaware of the underlying identification process, in the access-control authentication scenarios, subjects voluntarily provide the system with their identification samples. Gesture based biometrics is an evolving topic asking the question whether the human gesture may be used as an authentication modality. Lai *et al*. [9] demonstrate encouraging results using covariance descriptors extracted from silhouettes. In the follow-up work by Wu *et al*. [27] the authors demonstrate improved results by replacing the silhouettes with Kinect skeletons. In the recent work by Wu *et al*. [28] the benefits of using Kinect skeletons captured from multiple viewpoints are explored.

The "content-style" generative model, introduced by Tenenbaum and Freeman [19], represents an observation as a bilinear mixture of content and style. This approach was successfully applied in several domains including gait-based person recognition [22, 11]. The gait patterns were decomposed into "content", *i.e*., the periodic gait pattern, and "style", *i.e*., the personal stylistic variations, which was used to recognize the person. However, the primary focus of these works was to automatically synthesize novel (never-seen-before) graphic animations, rather than to identify the actor. Therefore, by decomposing action instances into content and style, they tried to simulate the actual action generation process. We are addressing a much simpler problem of discriminating between styles, where the discovery of underlying generative mechanism is not required. This leads to an improved performance comparing to the "content-style" separation based results.

1.3. Our Approach and Contributions

We present a general probabilistic framework, based on generative models, for user identity inference. We show that, in spite of the relatively high noise levels common to low cost pose estimation devices, the individual motion patterns collected from everyday actions as a single cue for person identification, do indeed have good discriminative properties. Moreover, we show that the identification accuracy is significantly improved when a combination of various action types is used. We evaluate our approach on four publicly available datasets for action recognition and one publicly available dataset for person authentication, on which we exhibit state-of-the-art performance. To the best of our knowledge, we are the first ones to report results on the problem of identity classification from general actions. Our results may serve as a baseline for further research in this direction.

Our approach differs from previous works in the following aspects:

1. Although reminiscent of the content-style separation paradigm [19], our generative model does not assume any specific dependence, *e.g*., bilinear, of the instance on content and style [22, 11]. Instead we use a non-parametric kernel density estimation to represent this dependence.

2. Unlike [21, 20, 18, 15] our method is not tailored towards any specific class of motion such as locomotion and therefore may be applied in a more general setting.

3. While [28] do consider identification from action rather than locomotion, their work is limited to identification in the "login" scenario, based on either user-defined action, which amounts to classical action recognition, or a single, shared-among-users, artificially designed "S" action. Besides showing that our

approach outperforms [28], we suggest identification from a wide variability of natural, everyday actions extending the applicability of the idea far beyond the "login" scenario. We also show that identification based on a variety of actions significantly improves the results.

The rest of the paper is organized as follows. Section 2 describes the general framework including the generative models and classifiers. In Section 3 action instance representations and distance metrics are discussed. In Section 4 the experimental results are presented. Section 5 presents a discussion and directions for future work.

2. Generative Model for User Identification

Under our model we restrict the allowed action types to a limited set of atomic actions $\mathcal{L}_a = \{\alpha_1, \ldots, \alpha_{N_a}\}$. Let $\mathcal{L}_u = \{u_1, \ldots, u_{N_u}\}$ denote the set of possible user class labels. We assume that the process of action instances creation is governed by one of the generative models in Fig. 1(c). $\{(X_i, A_i)\}_{i=1}^N$ denotes a set of N pairs of random variables, defined over $\mathbb{R}^d \times \mathcal{L}_a$, each associated with i's action instance representation and its label. Let U denote the random variable associated with the user identity class, defined over $\mathcal{L}_u$. Some scenarios assume that action labels are known, *i.e.*, the A_i variables are observed, while others do not. The diagrams in Fig. 1(c) correspond to these two scenarios.

Let us first assume that the action labels are given. Thus, $\{(\mathbf{x}_i, a_i)|\mathbf{x}_i \in \mathbb{R}^d, a_i \in \mathcal{L}_a, i = 2, \ldots, N\}$ denotes a set of N action instances, performed by a certain user $u \in \mathcal{L}_u$. According to the left diagram of Fig. 1(c) and [1]:

$$p(U|\{(\mathbf{x}_i, a_i)\}_{i=1}^N) \propto p(U) \prod_{i=1}^N p(\mathbf{x}_i|a_i, U), \qquad (1)$$

where $p(U)$ denotes a vector of prior probability values for all $u \in \mathcal{L}_u$ and $p(\mathbf{x}_i)$ is a shorthand notation for the probability $p(X_i = \mathbf{x}_i)$. The result $p(U|\{(\mathbf{x}_i, a_i)\}_{i=1}^N$ is a vector of posterior probability values.

Now let us assume that we are given N action instances $\{\mathbf{x}_i\}_{i=1}^N$ whose labels are unknown. The rightmost diagram in Fig. 1(c) describes the instance set creation process for this case. By marginalization over A_i we get:

$$p(U|\{\mathbf{x}_i\}_{i=1}^N) \propto p(U) \prod_{i=1}^N \sum_{a_i \in \mathcal{L}_a} \frac{p(\mathbf{x}_i|a_i, U)}{p(\mathbf{x}_i|a_i)} p(a_i|\mathbf{x}_i), \qquad (2)$$

where $p(a_i|\mathbf{x}_i)$ is obtained by applying any given action recognition algorithm on $\mathbf{x}_i$. Note that Eq. 1 is a special case of Eq. 2 assuming a perfect action recognition algorithm, assigning 1 to $p(a_i|\mathbf{x}_i)$ if and only if a_i is the $\mathbf{x}_i$'s true label, and 0 otherwise. In such case all the elements in the sum vanish except for one, while the denominator $p(\mathbf{x}_i|a_i)$ is constant with respect to U.

2.1. Classification

We propose to classify the user's identity using a MAP classifier corresponding to Eq. 2:

$$u^* = \underset{u \in \mathcal{L}_u}{\operatorname{argmax}}\, p(u) \prod_{i=1}^N \sum_{a_i \in \mathcal{L}_a} \frac{p(\mathbf{x}_i|a_i, u)}{p(\mathbf{x}_i|a_i)} p(a_i|\mathbf{x}_i). \qquad (3)$$

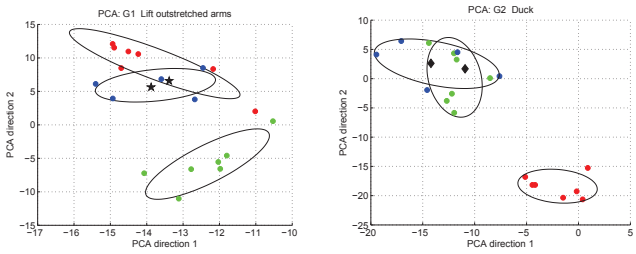
Assuming that we are given a set of labeled training samples, we use them to obtain a non-parametric estimate of the likelihood distribution $p(\mathbf{x}_1|a, u)$ for all pairs $(a, u) \in \mathcal{L}_a \times \mathcal{L}_u$. Let $\mathcal{D}_{a,u}$ denote the set of action instances of user u performing action a. Thus, applying a 1-nearest neighbor kernel density estimation (KDE), we obtain an estimator $\widehat{p}(\mathbf{x}|a, u)$ for $p(\mathbf{x}|a, u)$:

$$\widehat{p}(\mathbf{x}|a, u) \equiv \frac{1}{|\mathcal{D}_{a,u}|V}, \qquad (4)$$

where V is the volume of the D-dimensional[1] sphere of radius

$$r = \min_{\mathbf{x}' \in \mathcal{D}_{a,u} \setminus \{\mathbf{x}\}} d_a(\mathbf{x}, \mathbf{x}'),$$

centered at $\mathbf{x}$, and $d_a(\cdot, \cdot)$ measures the distance between action instances of class a. We show in Section 3.2, that the use of an action-specific distance measure allows us to take into account the instance variability of each particular action, when discriminating between users.



(a) (b)

Figure 2. Instance representations of the "Lift outstretched arms" (a) and "Duck" (b) actions from the MSRC-12 [5] dataset performed by three users. The representations are projected on the two major components of the PCA subspace.

2.2. Homogeneous and Heterogeneous Instance Sets

The MAP classifier given in Eq. 3 classifies the user based on a set of instances, $\{\mathbf{x}_i\}_{i=1}^N$, belonging to certain action types, regardless of whether the actual action labels

[1] D is the intrinsic data dimension, *i.e.*, the dimension of the actual manifold in which the data resides.

are given or not. We categorize the sets, of size larger than one, into two categories, namely the "homogeneous" and "heterogeneous" sets. A homogeneous set contains identically labeled action instances, *i.e.*, $\forall_{i,j}, a_i = a_j$, while all instances in a heterogeneous set are labeled differently, *i.e.*, $\forall_{i \neq j}, a_i \neq a_j$. Figs. 2(a) and 2(b) show the distributions of the "Lift outstretched arms" and the "Duck" action instances, from the MSRC-12 [5] dataset, respectively. These two action types utilize the torso and the legs differently, reflecting the various motion modalities of the human body. The training action instances for both action types (colored circles) come from three users colored with red, blue and green. Note that the "red" user performs the "Lift outstretched arms" action similarly to the "blue" one and differently from the "green" one, while the "green" user performs the "Duck" action similarly to the "blue" one and differently from the "red" one. Let us assume that the unknown user, at test time, is represented with a set of two action instances, marked with black "$\star$"s and "$\diamondsuit$"s. Using a homogeneous set of two stars for this task will result in high probability values for the "blue" and the "red" users. A similar ambiguous result – now for the "blue" and the "green" users – is expected if a homogeneous set of two diamonds is used. On the other hand, using a heterogeneous set of one star and one diamond will unambiguously identify the "blue" user. We remark that an alternative solution of increasing the homogeneous set size, N, is not feasible since N does not scale well with an increasing number of users and an increased area of the overlapping regions belonging to different users. We present extensive experimental validation of the above-mentioned claims in Section 4.3.

2.3. Framework Requirements

The MAP classifier introduced in Section 2.1 is general and assumes it is provided with a particular set of components. First, action instance representation, $\mathbf{x}$, should be selected, and action and user label sets, $\mathcal{L}_a$ and $\mathcal{L}_u$, defined. Second, a set of labeled training samples for each action-user pair has to be provided. Third, an action-specific distance measure $d_a(\cdot,\cdot)$ has to be defined. Note though that this is not a mandatory component and if missing, any generic distance measure, *e.g.*, L_2 norm, may be used at a price of lower accuracy. In addition, if action labels are unknown at test time, an action recognition algorithm giving $p(a|\mathbf{x})$ has to be provided. In Section 3 we discuss the choice of components used in our experiments.

3. Action Representation and Comparison

There exists a wide range of approaches for describing skeletal joint ensembles in 3D. Joint positions [7], normalized joint positions [29], joint angles [16], pairwise relative positions [25], geometric boolean features [14, 31] and points in Lie group [24] – all of these proved useful for action recognition and computer animation. In this work we adopt one of the simplest existing skeletal representations, namely the normalized joint positions (JP), due to its recent success in action recognition reported in [24]. The JP representation is constructed by taking the absolute joint positions in 3D and normalizing them with respect to the hip joint position. In the case of Kinect, the resulting representation is a (19×3)-dimensional vector.

3.1. Temporal Normalization

Rate variation is a well known problem in comparing action instances performed under different conditions (speed, style, etc.). Several methods were proposed to represent the action in a rate-invariant fashion. Veeraraghavan *et al*. [23] use a variant of Dynamic Time Warping (DTW) [13] to handle this issue. Wang *et al*. [25] build a Fourier Temporal Pyramid (FTP) and represent the action using the low frequency coefficients. In the recent work by Vemulapalli *et al*. [24], the authors combine the DTW and the FTP to obtain state-of-the-art performance in action recognition. While it is intuitive why temporal normalization helps action recognition it is not so with person identification. The difference in rate may be exactly what differentiates between users. In practice we saw that temporal normalization is indeed beneficial for user identification as well.

3.2. Action-Specific Metric Learning

Due to the highly constrained nature of the human skeleton structure and its dynamics [21, 30], most of the skeletal motion representations are redundant. Therefore, we reduce the instance dimensionality for each action class $a \in \mathcal{L}_a$ using Principal Component Analysis (PCA). The resulting representation is compact and less noisy, but considering our final goal we are interested in transformations improving the discriminative capabilities of the identity classifier. Therefore, we apply Linear Discriminant Analysis (LDA) on the PCA-transformed representations. The LDA maximizes the ratio of the between-class scatter to the within-class scatter so that the transformed instances of different identities fall far apart while those belonging to the same identity fall closer to each other. After applying LDA on samples shown at Fig. 2, the overlapping regions will be reduced while the clusters will be contracted as much as possible.

To conclude, let D_p and D_l denote the number of components selected in PCA and LDA, respectively. Given the PCA and the LDA transformations, $\mathbf{P}_a^{D_p \times D}$ and $\mathbf{L}_a^{D_l \times D_p}$ obtained from the D-dimensional instances of action $a \in \mathcal{L}_a$, we define $\mathbf{Z}_a^{D_l \times D} = \mathbf{L}_a \mathbf{P}_a$ as the composite PCA-LDA transformation for action class a.

Given the set of learned dimensionality reduction transformations, $\{\mathbf{Z}_a | a \in \mathcal{L}_a\}$, we define the distance between a pair of action instance representations $\mathbf{x}_i$ and $\mathbf{x}_j$ of class a,

as a squared Euclidean distance between their low dimensional representations. Thus,

$$d_a(\mathbf{x}_i, \mathbf{x}_j) = \|\mathbf{Z}_a\mathbf{x}_i - \mathbf{Z}_a\mathbf{x}_j\|_2^2 = (\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{M}_a(\mathbf{x}_i - \mathbf{x}_j) = \|\mathbf{x}_i - \mathbf{x}_j\|_{\mathbf{M}_a}, \quad (5)$$

where $\|\cdot\|_{\mathbf{M}_a}$ is a Mahalanobis distance and $\mathbf{M}_a = \mathbf{Z}_a^T\mathbf{Z}_a$. We use this action-specific distance measure for computing the likelihood estimator in Eq. 4. Note that the number of action instances in the training set for each action-user pair, $\mathcal{D}_{a,u}$, has to be larger than 3 in order for the PCA-LDA to work properly.

3.3. Action Recognition for User Identification

The MAP identity classifier in Eq. 3 depends on action recognition performance via the term $p(a|\mathbf{x})$. In our experiments we use one of the action classifiers proposed in [24], namely the linear SVM classifier trained in a one-vs-all fashion using normalized joint position (JP) representation (see [24] for details). We refer to this classifier as to JP-SVM. A given action instance, is classified as belonging to the class with the largest JP-SVM classification margin. We assign probability 1 to this class. An alternative approach of assigning each class a probability proportional to its margin, resulted in poor performance.

In our experiments we use the same datasets to train the user and the action classifiers. Thus, special care should be taken to ensures that, at test time, action classifiers are applied only on action instances from different users than those used during training. We solve this by using a leave-one-out approach to train the JP-SVM classifiers. That is, we train N_u action classifiers so that classifier number u is trained on action instances from all users except u. At test time, when computing the posterior $p(u|\{\mathbf{x}_i\}_{i=1}^N)$ we always use the action classifier which was not exposed to u's instances during training.

4. Experimental Results

We evaluate our approach on four well known datasets in the action recognition community and one public dataset for gesture-based user authentication [28].

In all our experiments we use the normalized joint positions (JP) representation described in Section 3. We investigate how the following temporal normalization approaches, applied to JP action representations, affect the classification performance:

1. **None.** No temporal normalization whatsoever.
2. **Dynamic Time Warping (DTW).** Normalize using a nominal curve [23, 24] obtained using DTW [13].
3. **Fourier Temporal Pyramid (FTP).** Represent the action with low frequency coefficients of a 3-level temporal pyramid [25].
4. **DTW+FTP.** First apply DTW and then FTP [24].

In all our experiments we use the DTW and FTP implementation provided by [24].

To evaluate the benefits of action-specific metric learning we use two metric types:

1. **L2.** No action-specific metric learning is performed, *i.e.*, $\mathbf{M}_a = \mathbf{I}, \forall a \in \mathcal{L}_a$.
2. **Mahalanobis** A Mahalanobis distance metric $\mathbf{M}_a$ is learned using the dimensionality reduction as described in Section 3.2. We set $D_p = 30$ and $D_l = 15$ in all our experiments except for the BodyLogin dataset, where we set $D_p = 40$ and $D_l = 39$.

We model the impact of action recognition performance on the identity classifier in Eq. 3 by considering two cases:

1. **Ground Truth (GT).** Assume that the action label is given by the dataset ground truth.
2. **Action Recognition (JP-SVM).** Assume that the action labels are unknown, but may be inferred using the action classifiers [24] discussed in Section 3.3. The actual average action recognition rate is reported in each experiment.

In all our experiments we report recognition results averaged over 10 random selections of action instances into training and test sets.

4.1. Action Recognition Datasets

Most of the existing human motion 3D skeletal datasets in the action recognition community contain the labels of the persons performing the action. This is done to prevent action recognition algorithms from overfitting towards actions of specific subjects during training. We use these labels for our task. Table 1 presents a summary of four action recognition datasets used in our experiments. For each dataset the number of action types, the number of users, the number of available action instances per user-action pair and the total number of available instances are specified.

MSRC-12 [5] is the largest dataset used in our experiments with 6243 action instance examples in total. We eliminated the user with label 18 from the dataset because no instances for the "Shoot" action were provided for him. Manual annotation of the first and the last frames of each action instance for this dataset is provided by [7].

The original MSR-Action3D [12] dataset contains recordings of 10 users performing 20 types of actions (1-3 recordings for each user-action pair). We chose a subset of 9 users (indexed 1-3,5-10), and 16 actions (see Table 1) so that each individual has at least two recordings for each action type. The rest of the datasets are used without modifications.

Dataset Name	Users Num.	Actions Num.	Inst. Per User-Act.	Inst. Tot.	Action Labels And Names
MSRC-12 [5]	29	12	8-10	6015	1.Lift outstretched arms; 2.Duck; 3.Push right; 4.Goggles; 5.Wind it up; 6.Shoot; 7.Bow; 8.Throw; 9.Had enough; 10.Change weapon; 11.Beat both; 12.Kick
UCFKinect [4]	16	16	5	1280	1.Balance; 2.Climb ladder; 3.Climb up; 4.Duck; 5.Hop; 6.Kick; 7.Leap; 8.Punch; 9.Run; 10.Step back; 11. Step front; 12. Step left; 13. Step right; 14. Twist left; 15. Twist right; 16. Vault
MSR-Action3D [12]	9	16	2-3	427	1.High arm wave; 2.Hor. arm wave; 3.Hammer; 4.Forward punch; 5.High throw; 6.Draw X; 7.Draw tick; 8.Draw circle; 9.Hand clap; 10.Two hand wave; 11.Side boxing; 12.Forward kick; 13.Jogging; 14.Tennis swing; 15.Tennis serve; 16.Golf swing
UTKinect [29]	9	10	2	180	1.Walk; 2.Sit down; 3.Stand up; 4.Pick up; 5.Carry; 6.Throw; 7.Push; 8.Pull; 9.Wave hands; 10.Clap hands

Table 1. Summary of action recognition datasets used in our experiments. In our results, we refer to actions using their labels specified in the rightmost column.

4.2. User Identification

We evaluate the performance of our approach for user identification on each one of the datasets in Table 1. In all our experiments we classify the user at test time using a single randomly selected action instance ($N = 1$). The number of training instances vary with dataset.

MSRC-12 dataset. In our first experiment we use the MSRC-12 dataset, to evaluate various combinations of temporal normalization. For each user-action pair we randomly select 4 action instances into the training set. Table 2 summarizes the true positive rate (TPR) for each one of the actions in the dataset, using all possible combinations of metric, action recognition and temporal normalization. Note that the average contribution of DTW to performance is higher than FTP, and the combination of both yields the best performance. Thus, in all our further experiments we use the DTW+FTP for temporal normalization. Note also that, as expected, the Mahalanobis distance metric outperforms the L2. We also see that the recognition accuracy degrades gracefully with the uncertainty introduced by action recognition. The actual JP-SVM's TPR in this experiment is 94% and is reported next to the algorithm name in the second column from left.

UCFKinect dataset. For each user-action combination a random training set of size 4 is used. Table 3 summarizes the results for this dataset. Note that the Mahalanobis metric outperforms L2 for all action types, resulting in average TPR of 95%, regardless whether the action labels are given or not. Almost no accuracy degradation occurs as a result of applying action recognition, mainly due to the high JP-SVM's TPR of 98%.

MSRAction3D dataset. The number of available instances per each user-action pair limits the training set size to one, and so only the L2 norm is evaluated. Table 4 summarizes the results. Note that for some of the actions, *e.g.*, "High arm wave", a low TPR of 66% is obtained, while for others, *e.g.*, "Hand clap", a TPR as high as 100% is achieved. On average 90% TPR is achieved for the case of unknown action labels while the action recognition accuracy was 88%.

UTKinect dataset. Table 5 summarizes the results for this dataset. Again, a single action instance was used for training. Looking at the TPR values, note that this is a more challenging dataset – only 58% and 57% average TPR is obtained when action labels are known and unknown, respectively. Although the average results are not satisfactory, using heterogeneous sets of four randomly selected action types significantly improves the results. A TPR of 92% is obtained using a heterogeneous set of four topmost action types (highlighted in each row).

4.3. Homogeneous vs. Heterogeneous Sets

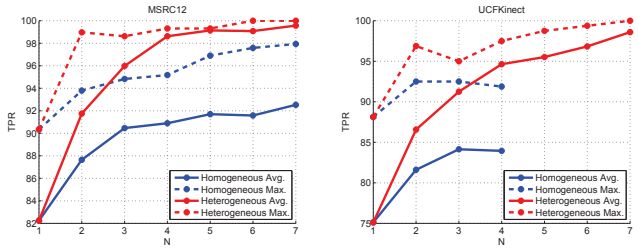
We now present experimental validation of the intuitive claims presented in Section 2.2. In the following experiments, a single action instance is randomly selected into the training set for each user-action pair. A set of $N \geq 1$ randomly selected action instances is used for classifying the user's identity at test time. In this experiment we assume that the instances' action labels are given. Figs. 3(a) and 3(b) show the average and the maximal TPR as a function of N using both homogeneous and heterogeneous sets, for the MSRC-12 and the UCFKinect datasets, respectively. For the homogeneous case, the "Avg." graph shows the TPR for randomly constructed homogeneous sets of size N, averaged across all action types. The upper bound "Max." graph shows the TPR for sets including instances from the topmost, in terms of TPR, action type, for each N. For the heterogeneous case, the "Avg." graph shows the average TPR for sets of size N, containing a random selection of action types. The "Max." graph shows the TPR for heterogeneous sets constructed of N topmost, in terms of TPR, action types. For both datasets, an average heterogeneous case outperforms the homogeneous one, and even outperforms the homogeneous' upper bound starting with $N = 3$ for MSRC-12, and $N = 4$ for UCFKinect. As expected, the

Metric	Act. Reco. (TPR%)	Temp. Model #	1	2	3	4	5	6	7	8	9	10	11	12	Avg.
L_2	**GT**	**None**	85	81	92	84	82	87	85	91	86	88	85	84	86
		FTP	87	88	96	90	91	92	88	98	89	97	92	91	92
		DTW	94	89	96	96	98	95	92	**99**	**98**	96	93	93	95
		DTW+FTP	**97**	**93**	**97**	**98**	**99**	**96**	**94**	99	97	**98**	**95**	**98**	**97**
	JP-SVM (94)	**None**	86	82	88	77	73	80	79	86	81	87	79	88	82
		FTP	85	85	96	87	83	86	87	**93**	89	93	83	89	88
		DTW	**94**	87	94	92	**92**	**91**	**92**	92	**98**	**96**	85	93	92
		DTW+FTP	92	**92**	**98**	**94**	89	91	90	93	96	96	**87**	**94**	**93**
Mahal.	**GT**	**None**	95	84	96	98	96	94	92	98	**98**	98	94	89	94
		FTP	97	94	**99**	**100**	95	**97**	94	98	98	**99**	96	94	97
		DTW	96	91	97	98	**99**	97	**96**	**99**	98	98	97	93	97
		DTW+FTP	**99**	**97**	99	99	98	97	94	99	98	99	**99**	**95**	**98**
	JP-SVM (94)	**None**	96	89	96	94	89	91	87	92	**97**	96	91	90	92
		FTP	94	90	97	94	**94**	**94**	92	93	97	**98**	**92**	93	94
		DTW	**97**	88	96	94	94	93	**95**	93	97	97	91	93	94
		DTW+FTP	97	**94**	**99**	**95**	92	93	92	**95**	97	98	91	**94**	**95**

Table 2. TPR in % for each one of the 12 actions in the MSRC12 [5] dataset. The "Avg." column is the average TPR across all 12 actions. The highlighted numbers indicate the maximal value in each cell.

Metric	Act. Reco. (TPR%)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Avg.
L_2	**GT**	93	94	95	90	94	88	96	75	96	89	91	96	98	83	93	97	92
	JP-SVM (98)	93	94	95	90	94	86	96	72	96	88	91	96	98	81	93	97	91
Mahal.	**GT**	95	99	99	92	97	97	95	94	99	94	94	96	98	91	98	98	96
	JP-SVM (98)	95	99	99	92	97	94	95	91	99	93	94	96	98	87	96	98	95

Table 3. TPR in % for each one of the 16 action in the UCFKinect [4] dataset. The "Avg." column is the average TPR across all 16 actions.



(a) (b)

Figure 3. TPR obtained using homogenous vs. heterogeneous labeled action instance sets for the MSRC-12(a) and the UCFKinect(b) datasets, as a function of set size N (see text for details). Due to limited number of action instances for each action-user combination in the UCFKinect dataset the homogeneous graphs terminate at $N = 4$.

selection of the best performing action types for the heterogeneous set results in the best performance.

4.4. User Authentication

The BodyLogin [28] dataset contains 3D skeletal sequences of 40 users performing two types of gestures. The sequences were simultaneously captured from different view points (center, right, left), using three Kinect devices. Unlike the previously discussed datasets, created with the action recognition problem in mind, this one was especially created for gesture-based user authentication. In [28] the authors discuss two user identification scenarios. In the first one, users are identified based on the way they execute one common-to-all "S" gesture. In the second one, users are identified based on their personal (user-defined) gestures. The latter is considered more of an action-recognition-based approach and is less challenging for authentication than the former, where a single action is shared among all users. To simulate variations in gestures performed by the same user, different interfering scenarios ("degradations") were used when recording the data. For example, a "User Memory" degradation contains gestures recorded one week apart by the same user, simulating variations caused by user's memory. A "Personal Effect" degradation contains gestures performed by the same user, while wearing heavier clothes, such as a raincoat or carrying a backpack. This degradation simulates variations caused by changes in user's clothing. Refer to [28] for details regarding all possible degradation.

We compare our results in user identification, based on the "S" gesture, to those reported in [28]. Our approach uses the DTW+FTP temporal normalization and the Mahalanobis distance, learned using PCA-LDA with the following parameters: $D_p = 40$ and $D_l = 39$. We follow the

Act. Reco. (TPR%)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Avg.
GT	86	94	98	97	100	100	97	93	100	99	90	92	91	90	82	88	93
JP-SVM (88)	66	94	93	90	90	100	97	88	100	99	84	86	91	90	81	88	90

Table 4. TPR in % for each one of the 16 selected actions from the MSRAction3D [12] dataset. The "Avg." column is the average TPR across all 16 actions.

Act. Reco. (TPR%)	1	2	3	4	5	6	7	8	9	10	Avg.	Rand. 4	Best 4
GT	53	**67**	47	35	46	51	62	**77**	**80**	**63**	58	87	92
JP-SVM (96)	57	**67**	47	35	46	40	62	**77**	**80**	**63**	57	83	92

Table 5. TPR in % for each one of the 10 action in the UTKinect [29] dataset. The highlighted numbers indicate four topmost TPR values in each row. The "Avg." column is the average TPR across all 10 actions. The "Rand. 4" and "Best 4" columns are the average TPRs when heterogeneous sets of four randomly selected action types and four topmost action types are used, respectively.

leave-one-out evaluation protocol used in the experiments on the closed-set identification problem, where a nearest neighbor with the Cov3DJ [7] descriptor was used.

Table 6 summarizes the results. Interestingly, the most common, central, viewpoint results in the worst average TPR. We get a TPR of 100% using either left or right viewpoint when no degradations are applied. The TPR slightly drops when a "Personal Effect" degradation takes place and reaches about 90% when the effect of "User Memory" degradation is considered. Our method outperforms Cov3DJ in each of the camera viewpoints and acquisition scenarios. Moreover, the rightmost column reports the best results achieved in [28] by fusing the information from all three Kinect views. All, except one, of our single-view results outperform the multiple-view results of [28].

5. Discussion and Conclusions

In this work we introduced a general framework for person identification based only on motion patterns, generated while performing arbitrary actions from a predefined set. The accuracy of our approach on four publicly available action recognition datasets and one person authentication dataset is high enough to be practical, and is improved even more when different action types are used for classification.

This direction opens space for future research. Even though the classification problem we considered in this works is person identification, the approach is general and may be used for other problems, *e.g.*, gender classification, medical state estimation and even action recognition. Developing new action representations and distance metrics, tailored specifically for the task of user identification, is another potential line of research.

Camera	Left		Right		Center		Multi. View
Method	**Cov3DJ [28]**	**Our**	**Cov3DJ [28]**	**Our**	**Cov3DJ [28]**	**Our**	**Best [28]**
Train/Test							
No degradation/No degradation	98.5	**100.0**	97.0	**100.0**	96.5	**100.0**	98.0
No degradation/Personal Effects	93.5	**96.5**	94.0	**99.0**	88.9	**94.0**	94.0
No degradation/User Memory	80.5	**90.5**	80.0	**90.5**	82.5	**84.5**	85.5
No degradation/Reproducibility	71.5	**88.5**	74.0	**91.0**	77.5	**83.0**	79.5
No degradation/All of the above	86.0	**92.8**	86.2	**93.7**	86.4	**86.7**	89.1
Everything/Everything	98.7	**99.9**	98.9	**99.9**	98.7	**99.8**	99.5
Average	88.1	**94.7**	88.4	**95.7**	88.4	**91.3**	90.9

Table 6. TPR in % of our method compared to that of [28] for the "S" action under various degradation. The highlighted numbers indicate the best methods for each combination of degradation and viewpoint. Our approach outperforms Cov3DJ in all categories. (Note that we converted the results, originally reported in terms of Correct Classification Error (CCE), to TPR).

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. 3

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie. Behavior recognition via sparse spatio-temporal features. In *PETS*, pages 65–72, 2005. 1

[3] A. Efros, A. Berg, G. Mori, and J. Malik. Recognizing action at a distance. In *ICCV*, pages 726–733, 2003. 1

[4] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola, Jr., and R. Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *IJCV*, 101(3):420–436, 2013. 7

[5] S. Fothergill, H. Mentis, P. Kohli, and S. Nowozin. Instructing people for training gestural interactive systems. In *SIGCHI*, pages 1737–1746, 2012. 3, 4, 5, 7

[6] M. Hofmann, S. Bachmann, and G. Rigoll. 2.5d gait biometrics using the depth gradient histogram energy image. In *IEEE Fifth International Conference on Biometrics: Theory, Applications and Systems (BTAS 2012)*, pages 23–26, 2012. 2

[7] 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*, pages 2466–2472, 2013. 2, 4, 5, 8

[8] G. Johansson. Visual perception of biological motion and a model for its analysis. *Perception Psychophysics*, 14(2):201–211, 1973. 2

[9] K. Lai, J. Konrad, and P. Ishwar. Towards gesture-based user authentication. In *IEEE Conf. Advanced Video and Signal-Based Surveillance (AVSS)*, pages 282–287, 2012. 2

[10] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003. 1

[11] C. Lee and A. Elgammal. Gait style and gait content: Bilinear models for gait recognition using gait re-sampling. In *International Workshop on Automatic Face and Gesture Recognition (AFGR)*, pages 147–152, 2004. 2

[12] W. Li, Z. Zhang, and Z. Liu. Action recognition based on a bag of 3D points. In *CVPRW*, 2010. 5, 8

[13] M. Müller and T. Röder. Motion templates for automatic classification and retrieval of motion capture data. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pages 137–146, 2006. 4, 5

[14] M. Müller, T. Röder, and M. Clausen. Efficient content-based retrieval of motion capture data. *TOG*, 24:677–685, 2005. 4

[15] B. C. Munsell, A. Temlyakov, C. Qu, and S. Wang. Person identification using full-body motion and anthropometric biometrics from kinect videos. In *ECCV Workshops*, pages 91–100, 2012. 2

[16] 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 *CVPRW*, pages 8–13, 2012. 1, 4

[17] 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 *CVPR*, pages 1297–1304, 2011. 1

[18] L. Sigal, D. Fleet, N. Troje, and M. Livne. Human attributes from 3D pose tracking. In *ECCV*, pages III: 243–257, 2010. 1, 2

[19] J. B. Tenenbaum and W. T. Freeman. Separating style and content with bilinear models. *Neural Comput.*, 12(6):1247–1283, 2000. 2

[20] N. Troje. Decomposing biological motion: a framework for analysis and synthesis of human gait patterns. *Journal of Vision*, 5:371–387, 2002. 1, 2

[21] N. Troje, C. Westhoff, and M. Lavrov. Person identification from biological motion: effects of structural and kinematic cues. *Perception Psychophysics*, 64(4):667–675, 2005. 1, 2, 4

[22] M. Vasilescu. Human motion signatures: analysis, synthesis, recognition. In *Proc. Int. Conf. on Pattern Recognition (ICPR)*, pages 456–460, 2002. 2

[23] A. Veeraraghavan, A. Srivastava, A. Roy Chowdhury, and R. Chellappa. Rate-invariant recognition of humans and their activities. *IEEE Transactions on Image Processing*, 18(6):1326–1339, 2009. 2, 4, 5

[24] R. Vemulapalli, F. Arrate, and R. Chellappa. Human action recognition by representing 3D skeletons as points in a lie group. In *CVPR*, 2014. 1, 2, 4, 5

[25] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, pages 1290–1297, 2012. 1, 2, 4, 5

[26] L. Wang, T. Tan, H. Ning, and W. Hu. Silhouette analysis-based gait recognition for human identification. *PAMI*, 25(12):1505–1518, 2003. 2

[27] J. Wu, J. Konrad, and P. Ishwar. Dynamic time warping for gesture-based user identification and authentication with kinect. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 2371–2375, 2013. 2

[28] J. Wu, J. Konrad, and P. Ishwar. The value of multiple viewpoints in gesture-based user authentication. In *CVPRW*, pages 90–97, 2014. 2, 3, 5, 7, 8, 9

[29] L. Xia, C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3D joints. In *CVPRW*, pages 20–27, 2012. 1, 4, 8

[30] Y. Yacoob and M. Black. Parameterized modeling and recognition of activities. *CVIU*, 73:232–247, 1999. 1, 4

[31] A. Yao, J. Gall, G. Fanelli, and L. Van Gool. Does human action recognition benefit from pose estimation? In *BMVC*, 2011. 1, 4