Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations

Mohamed E. Hussein[1], **Marwan Torki**[1], **Mohammad A. Gowayyed**[1], **Motaz El-Saban**[2]
[1]Department of Computer and Systems Engineering, Alexandria University, Alexandria, Egypt
{mehussein, mtorki, m.gowayyed}@alexu.edu.eg
[2]Microsoft Research Advanced Technology Lab Cairo, Cairo, Egypt
motazel@microsoft.com

Abstract

Human action recognition from videos is a challenging machine vision task with multiple important application domains, such as human-robot/machine interaction, interactive entertainment, multimedia information retrieval, and surveillance. In this paper, we present a novel approach to human action recognition from 3D skeleton sequences extracted from depth data. We use the covariance matrix for skeleton joint locations over time as a discriminative descriptor for a sequence. To encode the relationship between joint movement and time, we deploy multiple covariance matrices over sub-sequences in a hierarchical fashion. The descriptor has a fixed length that is independent from the length of the described sequence. Our experiments show that using the covariance descriptor with an off-the-shelf classification algorithm outperforms the state of the art in action recognition on multiple datasets, captured either via a Kinect-type sensor or a sophisticated motion capture system. We also include an evaluation on a novel large dataset using our own annotation.

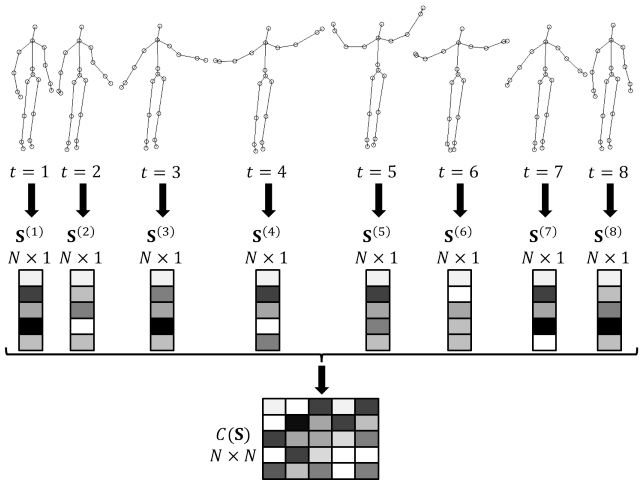

Figure 1: Construction of the covariance of 3D joints descriptor. A sequence of 3D joint locations of $T = 8$ frames is shown at the top for the "Start System" gesture from the MSRC-12 dataset. For the i^{th} frame, the vector of joint coordinates, $\mathbf{S}^{(i)}$ is formed. The sample covariance matrix is then computed from these vectors.

1 Introduction

Human action recognition is one of the many challenging problems targeted by machine vision researchers. It has many important applications in different domains. One of the most active such domains at the moment is interactive entertainment. A hive of activity around this domain was recently stimulated by the popularity of several gaming consoles with touch-less interfaces. For truly touch-less interface experience, a gaming console, such as Microsoft's XBox, deploys a low-cost depth sensor – the Kinect sensor. The depth data captured through the sensor can then be analyzed to estimate the player's body skeleton in real time [Shotton *et al.*, 2011a], which can further be analyzed to recognize his/her action or gesture. It was conjectured that using skeleton data alone for action recognition can perform better than using other low level image data [Yao *et al.*, 2011]. We already know that the approach works quite well in recognizing simple user gestures in gaming consoles. Nevertheless, the extent of success we can achieve with it and its utility in non-entertainment applications are not fully explored yet.

In this paper, we address the problem of representing a sequence of skeletal joint motions over time in a compact and efficient way that is highly discriminative for human action recognition. Particularly, we introduce a novel descriptor for human action recognition that is based on covariance matrices. As shown in Figure 1, the descriptor is constructed by computing the covariance matrix on the coordinates of body skeleton joints, sampled over time. To encode the temporal dependency of joint locations, we use multiple covariance matrices, each covering a sub-sequence of the input sequence, in a hierarchical fashion. We experimentally evaluated the descriptor on the task of human action recognition. We used multiple (recent and new) datasets of varying sizes and natures. In these experiments, classification using our descriptor either outperforms the state of the art or is the first to be reported. The benefit of the temporal hierarchy of descriptors becomes also evident from our experiments.

The paper is organized as follows: The remainder of this section summarizes the related work. In Section 2, we give background on the covariance descriptor. In Section 3, we explain the proposed Covariance of 3D Joints (Cov3DJ) descriptor, its different configurations, and efficient computations. Next, in Section 4, we present our experimental evaluation. Finally, we conclude the paper in Section 5.

1.1 Related Work

In human action recognition, there are three main challenges to be addressed: data capture, feature descriptors, and action modeling. In this section, we briefly summarize the literature associated with each challenge.

The first challenge is the availability and the quality of the captured data. Accurate skeleton data, captured using motion capture systems, such as the CMU MoCap database[1], and the HDM05 dataset [Müller *et al.*, 2007], are expensive to acquire. On the other hand, the Microsoft Kinect, and other low cost depth sensors, make the data acquisition affordable, with a loss of accuracy that is still acceptable for some applications. In addition to the depth maps produced by these sensors, the positions of skeletal joints can be estimated [Shotton *et al.*, 2011b]. Due to the low cost and widespread of such sensors, several skeletal datasets have been recently released [Li *et al.*, 2010; Fothergill *et al.*, 2012].

The second challenge in human action recognition is to find reliable and discriminative feature descriptions for action sequences. There are three common types of action descriptors: *whole sequence*, *individual frames* , and *interest points* descriptors. The latter two descriptors need additional steps of descriptor aggregation and temporal modeling in order to achieve the recognition goal.

An example of the methods that find a description of the whole sequence is the moments of Motion History Images [Bobick and Davis, 2001; Davis, 2001]. Examples of other methods that find a description for every image in a sequence, and defer the step of learning the dynamics, are the recent works of [Wang *et al.*, 2012b; Xia *et al.*, 2012]. In [Wang *et al.*, 2012b], a descriptor of relative positions between pairs of skeletal joints is constructed. The temporal modeling is done in the frequency domain via Fourier Temporal Pyramids. In [Xia *et al.*, 2012], a histogram of 3D joints descriptor in a frame is computed, a dictionary is built and the temporal modeling is done via HMM. Examples of methods that use interest point features is the spatio-temporal interest point features STIP [Laptev and Lindeberg, 2003]. However, local descriptors with depth data lack in its discrimination power due to the lack of texture in depth images. The work presented in [Gowayyed *et al.*, 2013], which uses histograms of displacement orientations, is the closest in spirit to the work presented in this paper.

The third challenge is modeling the dynamics of an action. Sequence analysis via generative models, such as HMMs [Xia *et al.*, 2012], or discriminative models, such as CRFs [Han *et al.*, 2010], are usually employed. In such methods, the joint positions or histograms of the joints positions are used as observations. Other recent approaches use recurrent neural networks [Martens and Sutskever, 2011], or Conditional Restricted Boltzman Machines [Mnih *et al.*, 2011]. Due to the large number of parameters to be estimated, these models need large amounts of data samples and training epochs to accurately estimate its model parameters.

2 The Covariance Descriptor

The covariance matrix for a set of N random variables is an $N \times N$ matrix whose elements are the covariance between every pair of variables. Let $\mathbf{X}$ be a vector of N random variables. The covariance matrix of the random vector $\mathbf{X}$ is defined as $\text{COV}(\mathbf{X}) = \text{E}[(\mathbf{X} - \text{E}(\mathbf{X}))(\mathbf{X} - \text{E}(\mathbf{X}))']$, where $\text{E}()$ is the expectation operator. The covariance matrix encodes information about the shape of the joint probability distribution of the set of random variables.

The covariance matrix was first introduced as a descriptor by Tuzel, *et al*. [2006]. In this work, the descriptor was used to describe a region of an image, where variables corresponded to different feature maps computed for the region, and samples of each variable corresponded to values of its feature map at different pixels in the region. The descriptor was applied successfully on object detection, texture classification, and pedestrian detection [Tuzel *et al.*, 2008].

Recently, the same idea was generalized to video sequences by considering features of pixels in a volumetric spatio-temporal patch, and was applied to action recognition [Sanin *et al.*, 2013]. In this paper, we take a different approach inspired by the findings of [Yao *et al.*, 2011], in which pose data was found to outperform low-level appearance features in action recognition. Particularly, we use the pose data, represented by the body joint locations, sampled over time, as the variables on which the covariance matrix is computed.

3 The Covariance of 3D Joints (Cov3DJ) Descriptor

Suppose that the body is represented by K joints, and the action is performed over T frames. Let $x_i^{(t)}$, $y_i^{(t)}$, and $z_i^{(t)}$ be the x, y, and z coordinates of the i^{th} joint at frame t. Let $\mathbf{S}$ be the vector of all joint locations, that is $\mathbf{S} = [x_1, ..., x_K, y_1, ..., y_K, z_1, ..., z_K]'$, which has $N = 3K$ elements. Then, the covariance descriptor for the sequence is $\text{COV}(\mathbf{S})$. Typically, the probability distribution of $\mathbf{S}$ is not known and we use the sample covariance instead, which is given by the equation

$$C(\mathbf{S}) = \frac{1}{T-1} \sum_{t=1}^{T} (\mathbf{S} - \bar{\mathbf{S}})(\mathbf{S} - \bar{\mathbf{S}})' \ , \qquad (1)$$

where $\bar{\mathbf{S}}$ is the sample mean of $\mathbf{S}$, and the $'$ is the transpose operator.

The sample covariance matrix [2], $C(\mathbf{S})$, is a symmetric $N \times N$ matrix. For the descriptor, we only use its upper triangle. For example, for a skeleton with 20 joints, such as the one produced by the Kinect sensor (examples are in Figure 1), $N = 3 \times 20 = 60$. The upper triangle of the covariance matrix in this case is $N(N+1)/2 = 1830$, which is the length of the descriptor.

[1] http://mocap.cs.cmu.edu/

[2] Referred to just as 'covariance matrix' for the rest of the paper.

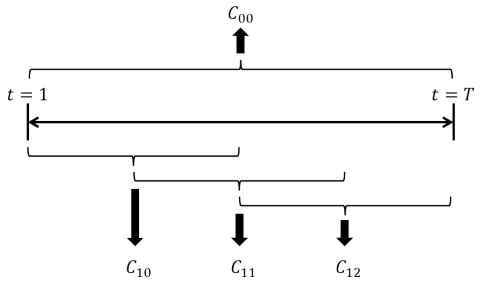

Figure 2: Temporal construction of the covariance descriptor. C_{li} is the i^{th} covariance matrix in the l^{th} level of the hierarchy. A covariance matrix at the l^{th} level covers $\frac{T}{2^l}$ frames of the sequence, where T is the length of the entire sequence.

3.1 Temporal Hierarchical Construction

The Cov3DJ descriptor captures the dependence of locations of different joints on one another during the performance of an action. However, it does not capture the order of motion in time. Therefore, if the frames of a given sequence are randomly shuffled, the covariance matrix will not change. This could be problematic, for example, when two activities are the reverse temporal order of one another, *e.g.* "push" and "pull".

To add the temporal information to Cov3DJ, we use a hierarchy of Cov3DJs, which is inspired by the idea of spatial pyramid matching [Lazebnik *et al.*, 2006] in 2D images. The hierarchical construction is shown in Figure 2. The top level Cov3DJ is computed over the entire video sequence. The lower levels are computed over smaller windows, overlapping or non-overlapping, of the entire sequence. Figure 2 shows only two levels in the hierarchy. Each covariance matrix is identified by two indices: the first is the hierarchy level index, and the second is the index within the level. The top level matrix covers the entire sequence and is denoted by C_{00}. A covariance matrix at level l is computed over $T/2^l$ frames of the sequence. The step from one window to the next is either the length of the window or half of it. If the step is half the window length, the windows overlap with one another. In Figure 2, covariance matrices in the second level overlap.

As we show in Section 4 adding more levels and allowing overlap enhances the ability of a classifier to distinguish among actions using the descriptor. However, the more layers we add and allowing overlap increases the length of the descriptor. For the descriptor configurations in Figure 2, a skeleton represented with 20 joints results in a descriptor of length $4 \times 1830 = 7320$.

Fast Descriptor Construction

Creating multiple layers of the temporal hierarchy and allowing overlap dictates computing multiple covariance matrices for sub-sequences of the same sequence. Luckily, a dynamic programming approach can be deployed to make the the computation of every element of the matrix possible in constant time, after some pre-computations are performed. A similar idea was used in prior work with the names *integral images* for covariances on image patches [Tuzel *et al.*, 2008], and *integral videos* for covariances on spatio-temporal video patches [Sanin *et al.*, 2013]. The same concept can be applied in our case with the distinction that integrals are needed only on the time dimension, which we refer to as *integral signals*.

Following similar notation to [Tuzel *et al.*, 2008], we define the two integral signals $\mathbf{P}^{(t)}$ and $\mathbf{Q}^{(t)}$ as

$$\mathbf{P}^{(t)} = \sum_{i=1}^{t} \mathbf{S}^{(i)}, \;\; \mathbf{Q}^{(t)} = \sum_{i=1}^{t} \mathbf{S}^{(i)} \mathbf{S}^{(i)'} \;\; . \tag{2}$$

After some algebraic manipulation, we can reach the following formula for computing the covariance matrix of the range of frames from $t_1 + 1$ to t_2, inclusively.

$$C^{(t_1,t_2)}(\mathbf{S}) = \frac{1}{M-1}(\mathbf{Q}^{(t_1,t_2)} - \frac{1}{M}\mathbf{P}^{(t_1,t_2)}\mathbf{P}^{(t_1,t_2)'}) \;\; , \tag{3}$$

where $M = t_2 - t_1$, $\mathbf{Q}^{(t_1,t_2)} = \mathbf{Q}^{(t_2)} - \mathbf{Q}^{(t_1)}$, and $\mathbf{P}^{(t_1,t_2)} = \mathbf{P}^{(t_2)} - \mathbf{P}^{(t_1)}$. Details of the derivation are a straight forward simplification of the corresponding 2D version in [Tuzel *et al.*, 2008]. Having computed the signal integrals, $\mathbf{P}$ and $\mathbf{Q}$, we can compute the covariance matrix over any range of frames in time that is independent of the length of the range, using Equation 3.

It is worth noting here that integrating over one dimension only in integral signals, compared to integration over two and three dimensions in integral images and integral videos, respectively, is not just a simplification of mathematics and computational demands. It also leads to significantly less error accumulation on computing the integrals [Hussein *et al.*, 2008].

4 Experiments

We evaluated the discrimination power of our descriptor for action recognition. We performed this evaluation on three publicly available datasets. In one of them, we used our own annotation. Two of the datasets were acquired using a Kinect sensor, and one using a motion capture system. Details of the experiments are presented in the following sub-sections. In all experiments, we used a linear SVM classifier, using the LIBSVM software [Chang and Lin, 2011] with the descriptor. Before training or testing, descriptors are normalized to have unit L_2 norms. The covariance matrix is shift invariant by nature. To make it scale invariant, we normalized joint coordinates over the sequence to range from 0 to 1 in all dimensions before computing the descriptor.

4.1 MSR-Action3D Dataset

The MSR-Action3D dataset [Li *et al.*, 2010] has 20 action classes performed by 10 subjects. Each subject performed each action 2 or 3 times. There are 567 sequences in total, from which we used 544[3], each was recorded as a sequence of depth maps and a sequence of skeletal joint locations. Both types of sequences were acquired using a Kinect sensor. 20 joints were marked in the skeletal joint sequences as shown in Figure 3.

[3] [Li *et al.*, 2010] already excluded 10 sequences out of the 567 in their experiments. We excluded 13 more sequences that we found severely corrupted.

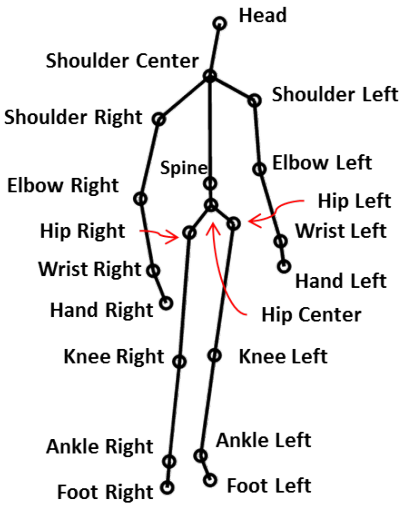

Figure 3: Skeleton joint locations and names as captured by the Kinect sensor.

Method	Acc. (%)
Rec. Neural Net. [Martens and Sutskever, 2011]	42.50
Hidden Markov Model [Xia *et al.*, 2012]	78.97
Action Graph [Li *et al.*, 2010]	74.70
Random Occupancy Patterns [Wang *et al.*, 2012a]	86.50
Actionlets Ensemble [Wang *et al.*, 2012b]	88.20
Proposed Cov3DJ	**90.53**

Table 1: Comparative results on the MSR-Action3D dataset.

We used the typical experimental setup on this dataset [Li *et al.*, 2010], which divides the action classes into three action sets, each containing 8 action classes, with some overlap between action sets. Classifiers are trained to distinguish among actions in the same set only. The reported accuracy is the average over the three sets.

Several studies have already been conducted on the MSR-Action3D dataset. Table 1 shows the classification rate of our approach compared to the state-of-the-art methods[4]. Our results in this table correspond to using three levels of the descriptor while allowing overlap in the second and third levels. Our approach achieves 90.53% classification rate, exceeding the second best approach by more than 2%. It is worth noting that we only rely on joint locations in our approach, while other algorithms, such as [Li *et al.*, 2010; Wang *et al.*, 2012a], use the depth maps. Moreover, our descriptor construction and classification algorithm are considerably simpler than the Ensemble of Actionlets used in [Wang *et al.*, 2012b]; and, our encoding of temporal information is also considerably simpler than HMMs, used in [Xia *et al.*, 2012]. Therefore, the effectiveness of our approach is fostered by its simplicity compared to other state of the art methods, which shows its practical advantage.

Next, we use the same dataset to evaluate the effect of changing the parameters of descriptor construction. The results in Table 2 show the classification accuracy for different levels in the temporal hierarchy while enabling or disabling overlap. In general, we can deduce that adding more levels enhances the descriptor's discrimination power, and hence, the classification accuracy. The overlap also enhances the classification accuracy. Another observation is that even with one level, Cov3DJ outperforms all algorithms in Table 1, except for the Actionlets Ensemble [Wang *et al.*, 2012b]. With only two levels and overlap, Cov3DJ outperforms all other methods in the table.

[4]The entry for RNN in Table 1 was copied from [Wang *et al.*, 2012b].

	$L=1$	$L=2$	$L=3$	$L=2$, OL	$L=3$, OL
AS1	88.04	86.96	86.96	88.04	88.04
AS2	78.57	81.25	84.82	83.93	89.29
AS3	95.24	94.29	93.33	94.29	94.29
Mean	87.28	87.50	88.37	88.75	**90.53**

Table 2: Results on MSR-Action3D using different levels in the temporal hierarchy. Adding levels and allowing overlap (marked with OL) enhances classification accuracy in general. Best results for $L=3$ with overlapping.

Metaphoric Gestures	No. of Insts.	Iconic Gestures	No. of Insts.
Start system	508	Duck	500
Push right	522	Goggles	508
Wind it up	649	Shoot	511
Bow	507	Throw	515
Had enough	508	Change weapon	498
Beat both	516	Kick	502

Table 3: Gesture classes in the MSRC-12 dataset and the number of annotated instances from each class.

4.2 MSRC-12 Kinect Gesture Dataset

To test the proposed approach when a large number of training instances is available, we experimented on the MSRC-12 dataset [Fothergill *et al.*, 2012]. MSRC-12 is a relatively large dataset for action/gesture recognition from 3D skeleton data, recorded using a Kinect sensor. The dataset has 594 sequences, containing the performances of 12 gestures by 30 subjects. There are $6,244$ annotated gesture instances in total. Table 3 shows the 12 gesture classes in the dataset and the number of annotated instances from each. The gesture classes are divided into two groups: metaphoric gestures, and iconic gestures.

Each sequence in the dataset is a recording of one subject performing one gesture for several times in a row. The ground truth annotation for each sequence marks the *action point* of the gesture, which is defined in [Nowozin and Shotton, 2012] as "a single time instance at which the presence of the action is clear and that can be uniquely determined for all instances of the action". For a real-time application, such as a game, this is the point at which a recognition module is required to detect the presence of the gesture.

The ground truth annotation of the dataset is designed for experimenting on the task of action detection, in which it is required to locate instances of different actions in a given

video sequence. We wanted to benefit from the large volume of the dataset without moving away from the task of action recognition. Therefore, we needed to know the start and end of each gesture instance, not just the action point.

To perform our action recognition experiments, we manually annotated the sequences of the dataset to mark the onset and offset of each gesture instance.[5] To make this task easy, we made use of the action point annotation. We developed a simple tool to facilitate locating the boundaries of each action instance starting the search always from the marked action point.

The lengths of the gesture instances – *i.e*. the number of frames between the onset and offset of the gesture – resulting from our annotation range from 13 frames to 492 frames. The lower end of this range correspond to legitimate instances of the gesture "wind it up", which is sometimes performed by the subject multiple times in a row, and the ground truth marks each as a separate instance. The higher end of the range, however, typically correspond to an odd performance of the gesture, *e.g*. dancing and moving back and forth while performing the gesture with unnecessary repetitions, or an extra slow performance of the gesture. Such odd long instances constitute a very small fraction of the dataset. Only 40 instances in the entire dataset are longer than 200 frames. We included all instances in our experiments. The median length of an instance is 80 frames. If we consider the ending of the gesture instance to be the action point, instead of the offset point, the median length becomes 40 frames and the maximum becomes 440 frames. Given the wide range of gesture lengths, choosing a fixed length sequence ending at an action point, *e.g*. 35 frames as in [Fothergill *et al.*, 2012], to represent an action is not quite convincing. While 35 frames is shorter than more than half of the instances, it may include more than two consecutive instances of the short gestures, such as "wind it up".

In the following subsections, we first present the experimental results using our own annotation, which are the first such results on this dataset. Then, we compare to a recently published experiment on the same dataset in an action recognition setting [Ellis *et al.*, 2013], in which gesture boundaries where considered to be the mid-points between consecutive action points. In all these experiments, we vary the number of levels in the temporal hierarchy, between 1 and 2, and enable or disable overlap in the latter. Given the large volume of the dataset, we could not experiment on three hierarchical levels of the descriptor due to the limitations of our system's configurations. We completely disregard the type of instructions given to each subject [Fothergill *et al.*, 2012] in all our experiments.

Leave-One-Out Experiments

In this experiment, we used all action instances from 29 subjects for training and the action instances of the remaining subject for testing. We performed the experiment 30 times, excluding one subject in each run. The benefit of such setup is two fold: First it allows for testing the inter-subject generalization of the approach while using as much data as possible for training. Second, it allows for detecting problematic subjects and analyzing the sources of some of the classification errors.

[5] This annotation can be downloaded from http://www.eng.alexu.edu.eg/ mehussein/msrc12_annot4rec/.

	$L = 1$	$L = 2$	$L = 2$, OL
Leave One Out	92.7	93.6	93.6
50% Subject Split	90.3	91.2	91.7
$1/3$ Training	97.7	97.8	97.9
$2/3$ Training	98.6	98.7	98.7
[Ellis *et al.*, 2013]'s	89.6	90.9	91.2

Table 4: Classification accuracy results for experiments on the MSRC-12 dataset with different experimental setups and different descriptor configurations. The numbers shown are percentages. For explanation of each experimental setup, refer to subsections of Section 4.2.

The results of the experiment are shown in the first row of Table 4. The values in the table are the average classification rate over the 30 runs. This average value ranges from 92.7% to 93.6%, slightly increasing with increasing the descriptor's length (by adding an extra level to the hierarchy and allowing overlap). The high classification rate verifies the inter-subject discrimination power of the descriptor.

Inspection of the individual errors in each of the 30 runs revealed that the most problematic gesture instances belonged to subject number 2. By inspecting the gesture classes with high error rate for this subject, we found that in most cases, the subject performed the gesture with unrelated movements, for example, dancing or walking while the gesture should be performed only by hands.

50% **Subject Split**

In the next experiment, we test the sensitivity of the classifier to reducing the number of training samples. We trained 20 different classifiers, each on a random selection of half the persons for training and the other half for testing. The average correct classification rate, as shown in the second row of Table 4, ranges from 90.3% to 91.7%. Despite using only half the instances for training compared to around 97% in the leave-one-out experiment, the reduction in the classification accuracy is less than 3%

From the experiments above, it is clear that the discrimination power of our descriptor is larger on the MSRC-12 dataset. This is despite the larger number of gestures and despite the larger differences among subjects' performances of the same gesture. This can be attributed to the larger number of instances available to train the classifiers.

$1:2$ **Instance Split**

In the final experiment, we test how much the classifier's performance can be enhanced if samples from all subjects are used in training and testing. The instances of each gesture class are randomly split between training and testing. The splits are done by random sampling without replacement. In other words, no instance can be shared between the training and testing sets. Two different split ratios are used: either $1/3$ of the instances are used for training and the rest for testing, or $2/3$ are used for training and the rest for testing. 20 different random splits from each ratio are generated in each ratio.

The results for this experiment are shown in the third and forth rows of Table 4. When one third of the data is used for training, the accuracy is around 98%. When two thirds are used for training, the accuracy goes up to around 99%.

From this experiment, we can see that a significant portion of the error we saw in the previous experiments were due to inter-person variations in performing the gesture. This could be due to giving different types of instructions to different users upon collecting the dataset [Fothergill *et al.*, 2012].

Using Mid-Action Points as Gesture Boundaries

In this section, we compare to the results of [Ellis *et al.*, 2013], in which a 4-fold cross validation experiment was conducted on the MSRC-12 dataset. Following their same setup, the midpoint between two consecutive action points were used to divide a video into gesture instances, while using the first and last frames of a video as the boundaries for the first and last gesture instances. The results of this experiment are shown in the last row of Table 4. The classification rate reported in [Ellis *et al.*, 2013] is 88.7%, which is slightly inferior to our basic configuration. Our best configuration achieves 91.2% accuracy in this experiment.

4.3 HDM05-MoCap Dataset

Similar to [Ofli *et al.*, 2012], we experimented with our approach on a Motion Capture dataset, namely the HDM05 database [Müller *et al.*, 2007]. There are three main differences between this dataset and the preceding two datasets: First, it is captured using motion-capture sensors, which leads to much less noise than in the data acquired by a Kinect sensor. Second, the number of joints recorded is 31 instead of 20. This leads to a longer descriptor since the size of the covariance matrix in this case is 93×93. Third, the frame rate is much higher, 120 fps instead of 15 or 30 fps as in the preceding two datasets.

We used the same setup in [Ofli *et al.*, 2012] with the same 11 actions performed by 5 subjects. We had 249 sequences in total. We used 3 subjects (140 action instances) for training, and 2 subjects (109 action instances) for testing. The set of actions used in this experiment is: *deposit floor, elbow to knee, grab high, hop both legs, jog, kick forward, lie down floor, rotate both arms backward, sneak,squat, and throw basketball*

The results in Table 5 show that the most basic configuration of our descriptor outperforms the best configuration of the SMIJ approach [Ofli *et al.*, 2012]. The results also show that the more levels we add to the temporal hierarchy the better classification accuracy we can achieve.

We can observe how the classification accuracy with the HDM05 dataset is significantly better than the classification accuracy with the MSR-Action3D dataset (Section 4.1) although the numbers of training samples used in both datasets are comparable. This can be attributed to the much lower level of noise in HDM05's data, and the extra information available with the higher frame rate and larger number of joints.

Method	Accuracy(%)
SMIJ [Ofli *et al.*, 2012]	84.40
Cov3DJ $L = 1$	92.66
Cov3DJ $L = 2$	93.57
Cov3DJ $L = 3$	**95.41**

Table 5: Classification accuracy on the HDM05 dataset with various configurations of Cov3DJ compared to the baseline method.

5 Conclusion and Future Directions

We introduced a novel descriptor for action sequences consisting of 3D skeletal joint movements. The descriptor, named Cov3DJ, is based on the covariance matrix of 3D joint locations over the entire sequence. Cov3DJ has a fixed length, independent from the sequence's length. Temporal information can be effectively encoded in the descriptor by incorporating multiple Cov3DJs for sub-windows of the sequence, that are possibly overlapping, in a temporal hierarchy. Cov3DJ is also efficient to compute over multiple overlapping windows using integral signals.

We evaluated the discrimination power of the descriptor on the task of human action recognition from skeleton data. Despite the simplicity of the descriptor, training an off-the-shelf linear SVM classifier on it outperforms the state of the art methods in multiple dataset. We achieved a classification rate of 90.5% on the MSR-Action3D dataset, and 95.4% on HDM05 MoCap dataset. In addition, we experimented on the newly introduced MSRC-12 dataset, with our own annotation, achieving up to 93.6% cross-subject classification rate.

While the descriptor is both scale and translation invariant, it is not rotation or reflection invariant. We believe that rotation invariance can easily be achieved by transforming joint locations to the subject's local coordinate frame, defined based on the current pose, instead of using the camera frame. Alternatively, rotation invariant features, such as angles or velocities, can be used instead of joint locations. Reflection invariance, on the other hand, is more challenging to achieve. However, in some applications, actions performed by the left and right parts of the body are distinct, in which case, reflection invariance is not desired. Finally, a more flexible temporal sub-division may help achieve better performance.

Acknowledgment

This research project was sponsored by a grant from the Microsoft Research Advanced Technology Lab Cairo, and was hosted by VT-MENA Research Center in Alexandria University.

References

[Bobick and Davis, 2001] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 23(3):257 –267, mar 2001.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines.

ACM Transactions on Intelligent Systems and Technology, 2:27:1–27:27, 2011.

[Davis, 2001] J.W. Davis. Hierarchical motion history images for recognizing human motion. In *Detection and Recognition of Events in Video, 2001. Proceedings. IEEE Workshop on*, pages 39 –46, 2001.

[Ellis *et al.*, 2013] Chris Ellis, Syed Zain Masood, Marshall F. Tappen, Joseph J. Laviola, Jr., and Rahul Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *Int. J. Comput. Vision*, 101(3):420–436, February 2013.

[Fothergill *et al.*, 2012] Simon Fothergill, Helena Mentis, Pushmeet Kohli, and Sebastian Nowozin. Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746, 2012.

[Gowayyed *et al.*, 2013] Mohammad A. Gowayyed, Marwan Torki, Mohamed E. Hussein, and Motaz El-Saban. Histogram of oriented displacements (hod): Describing 3d trajectories of human joints for action recognition. In *International Joint Conference on Artificial Intelligence*. AAAI Press, 2013.

[Han *et al.*, 2010] Lei Han, Xinxiao Wu, Wei Liang, Guangming Hou, and Yunde Jia. Discriminative human action recognition in the learned hierarchical manifold space. *Image and Vision Computing*, 28(5):836 – 849, 2010.

[Hussein *et al.*, 2008] M. Hussein, F. Porikli, and L. Davis. Kernel integral images: A framework for fast non-uniform filtering. In *IEEE Conference on Computer Vision and Pattern Recognition, 2008.*, june 2008.

[Laptev and Lindeberg, 2003] I. Laptev and T. Lindeberg. Space-time interest points. In *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, pages 432 –439 vol.1, 2003.

[Lazebnik *et al.*, 2006] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 2*, pages 2169–2178, 2006.

[Li *et al.*, 2010] Wanqing Li, Zhengyou Zhang, and Zicheng Liu. Action recognition based on a bag of 3d points. In *IEEE International Workshop on CVPR for Human Communicative Behavior Analysis*, 2010.

[Martens and Sutskever, 2011] J. Martens and I. Sutskever. Learning recurrent neural networks with hessian-free optimization. In *Proc. 28th Int. Conf. on Machine Learning*, 2011.

[Mnih *et al.*, 2011] Volodymyr Mnih, Hugo Larochelle, and Geoffrey Hinton. Conditional restricted boltzmann machines for structured output prediction. In *Proceedings of the International Conference on Uncertainty in Artificial Intelligence*, 2011.

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

[Nowozin and Shotton, 2012] Sebastian Nowozin and Jamie Shotton. Action points: A representation for low-latency online human action recognition. Technical Report MSR-TR-2012-68, Microsoft Research Cambridge, July 2012.

[Ofli *et al.*, 2012] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 8–13, 2012.

[Sanin *et al.*, 2013] A. Sanin, C. Sanderson, M. Harandi, and B.C. Lovell. Spatio-temporal covariance descriptors for action and gesture recognition. In *Workshop on the Applications of Computer Vision (WACV)*, 2013.

[Shotton *et al.*, 2011a] 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 *IEEE Conference on Computer Vision and Pattern Recognition, 2011*, pages 1297–1304, 2011.

[Shotton *et al.*, 2011b] 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 *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1297–1304, 2011.

[Tuzel *et al.*, 2006] Oncel Tuzel, Fatih Porikli, and Peter Meer. Region covariance: a fast descriptor for detection and classification. In *Proceedings of the 9th European conference on Computer Vision - Volume Part II*, pages 589–600, 2006.

[Tuzel *et al.*, 2008] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on riemannian manifolds. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 30(10):1713 –1727, oct. 2008.

[Wang *et al.*, 2012a] Jiang Wang, Zicheng Liu, Jan Chorowski, Zhuoyuan Chen, and Ying Wu. Robust 3d action recognition with random occupancy patterns. In *European Conference on Computer Vision (ECCV)*, 2012.

[Wang *et al.*, 2012b] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2012.

[Xia *et al.*, 2012] L. Xia, C.C. Chen, and JK Aggarwal. View invariant human action recognition using histograms of 3d joints. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 20–27, 2012.

[Yao *et al.*, 2011] Angela Yao, Juergen Gall, Gabriele Fanelli, and Luc Van Gool. Does human action recognition benefit from pose estimation? In *Proceedings of the British Machine Vision Conference*, pages 67.1–67.11, 2011.