CovP3DJ: Skeleton-parts-based-covariance Descriptor for Human Action Recognition

Hany A. El-Ghaish[1], Amin Shoukry[1,3] and Mohamed E. Hussein[2,3]
[1]CSE Department, Egypt-Japan University of Science and Technology, New Borg El-Arab City, Alexandria, Egypt
[2]Information Sciences Institute, Arlington, Virginia, U.S.A.
[3]Computer and Systems Engineering Department, Faculty of Engineering, Alexandria University, Alexandria, Egypt

Keywords: Hand-crafted Features, Covariance Descriptor, Skeleton-based Human Action Recognition.

Abstract: A highly discriminative and computationally efficient descriptor is needed in many computer vision applications involving human action recognition. This paper proposes a hand-crafted skeleton-based descriptor for human action recognition. It is constructed from five fixed size covariance matrices calculated using strongly related joints coordinates over five body parts (spine, left/ right arms, and left/ right legs). Since covariance matrices are symmetric, the lower/ upper triangular parts of these matrices are concatenated to generate an efficient descriptor. It achieves a saving from **78.26** % to **80.35** % in storage space and from **75** % to **90** % in processing time (depending on the dataset) relative to techniques adopting a covariance descriptor based on all the skeleton joints. To show the effectiveness of the proposed method, its performance is evaluated on five public datasets: MSR-Action3D, MSRC-12 Kinect Gesture, UTKinect-Action, Florence3D-Action, and NTU RGB+D. The obtained recognition rates on all datasets outperform many existing methods and compete with the current state of the art techniques.

1 INTRODUCTION

Human action recognition is continually evolving to cope with the challenges facing computer vision applications such as surveillance systems, robotics, interactive games etc. Therefore, there is a vital need to provide a highly discriminative, compact, robust, and fast action descriptor. Datasets for human action recognition can be collected using RGB cameras, motion capture systems (Mocap) and Kinect cameras. Data collected using RGB cameras (Aggarwal and Ryoo, 2011) are affected by different factors such as occlusion, variation in camera views, illumination change, and background clutter. RGB videos suffer from their inability to capture the body motion in the 3-D space. However, Mocap systems are capable of monitoring human motion at specific 3-D locations using accurate and expensive sensors. Mocap data is acquired as 3-D depth data from which skeleton 3-D joint locations are derived. HDM05 (Müller et al., 2007) is an example of a Mocap dataset. Recently, cheap but less accurate Kinect cameras are capable of estimating the 3D skeleton joint locations. They are being used with many applications such as interactive games (XBOX), robots vision, surveillance, human action and sign language recognition. Estimation of the rigid articulated joints of a human skeleton is much easier using Mocap systems or Kinect cameras than using RGB cameras (Vemulapalli et al., 2014).

There are intensive researches in recognizing human actions based on skeleton data. Some of them use hand-crafted features while others use deep learning methods to automate the feature extraction process. Despite the progress achieved by deep learning techniques, they have the following disadvantages: (1) they need large datasets to train their models and don't fit well with small datasets, (2) they require high performance computing platforms for model training, and (3) they remain black-boxes for a human modeler (Schmidhuber, 2015). Accordingly, there is still a vital need to use hand-crafted features especially in situations where a large dataset is not available.

In this paper, a hand-crafted compact, robust and discriminative descriptor is proposed for human action recognition. It is a skeleton-part-based descriptor as shown in Figure 1. Specifically, a covariance matrix descriptor is constructed for each of the body parts (spine, left/right arms, and left/right legs) that a human can use to compose an action. Since the covariance matrix is symmetric, only the lower/upper

El-Ghaish, H., Shoukry, A. and Hussein, M.
CovP3DJ: Skeleton-parts-based-covariance Descriptor for Human Action Recognition.
DOI: 10.5220/0006625703430350
In *Proceedings of the 13th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2018) - Volume 5: VISAPP*, pages 343-350
ISBN: 978-989-758-290-5

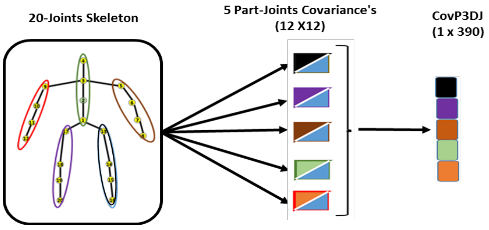


Figure 1: The construction of the proposed **CovP3DJ** descriptor. The five body parts (spine, left/right arms, and left/right legs) consists of 4 joints each. Each joint has 3 coordinates, which makes a total of 12 degrees of freedom. Therefore, five covariance matrices each of size 12×12 can be constructed, reduced (because of symmetry) and concatenated in a row vector of size 1×390 to represent the whole body motion.

triangular part of each matrix is needed to represent the motion of that part. Finally, these triangular parts are concatenated into one vector descriptor which represents the corresponding body motion. The proposed descriptor is compacted while remaining discriminant, as will be shown in the experimental results. The proposed descriptor is called **CovP3DJ** which means **Cov**ariance computation over the skeleton **P**arts of **3D J**oints coordinates.

Organization: A review of the related work is introduced in Section 2. The proposed recognition framework and its key components are presented in Section 3. The experimental results are analysed in Section 4. Conclusion and future work are summarized in Section 5.

2 RELATED WORK

Many ideas have emerged during the past few years for solving the challenges imposed by human action recognition using different modalities (skeleton, RGB images, and depth images), descriptors and classifiers. In this section, we provide a brief discussion about the work done on skeleton-based and Covariance-based approaches for human action recognition as they are the most related to our work.

In general, skeleton-based human action recognition methods can be classified into two major categories which are joint-based and part-based (Vemulapalli et al., 2014). The joint-based methods deal with the skeleton joints as a set of points in 3D space from which features are extracted to describe the temporal evolution of their motions. On the other hand, the part-based category view the skeleton body as a set of parts which are connected using line segments that form the human skeleton then extract features that are based on the human geometry.

Joint-based Methods: The method in (Evangelidis et al., 2014a) proposes a Fisher kernel descriptor to represent the skeleton quads of the action sequence and its split sub-sequences. The concatenation of all the generated Fisher vectors is considered the action descriptor. Action recognition is based on SVM. In (Li and Leung, 2017a) a graph-based representation is used to describe the spatial structure of the skeleton joints, and the top-k Relative Variance of Joint Relative Distance (RVJRD) decides the joints pairs that should be selected according to the activity level. Pyramids of covariances are used to extract the temporal features. Classification is performed by matching the similarities among graph kernels.

Part-based Methods: In (Vemulapalli et al., 2014), a geometry descriptor between the body parts using a rotation and a transformation matrix is represented in a Special Ecludian (SE) space descriptor. Also, the Dynamic Time Warping (DTW), and Fourier transformation are used for the alignment and temporal features extraction. The SVM is used as a classifier. The approach in (Chaudhry et al., 2013) divides the skeleton into small parts hierarchically, then bio-inspired features are extracted from each part. The obtained features are modeled by a linear dynamic system. In (Ohn-Bar and Trivedi, 2013b), skeleton sequences were represented by estimating the joints angles, trajectories, and the classification is done using SVM. According to (Gavrila et al., 1995), the angles between 3D joints are measured, and DTW is used for temporal alignment.

Covariance-based Methods: The covariance is the measure of how two random variables change concerning each other. It is positive when variables tend to show similar behavior and negative otherwise. The covariance matrix descriptor computation can be applied to the skeleton coordinates, RGB images, and depth images. It has been used in many computer vision applications. For instance, method in (Ma et al., 2014) proposed a verification and person re-identification recognition systems, gBiCov. Biologically inspired features (BIF) are extracted at different scales using Gabor filters then encoded by covariance matrix descriptor to represent the whole action. KISSME (Koestinger et al., 2012) is used as a metric for measuring the equivalence between descriptor pairs. A covariance descriptor, Cov3DJ, is proposed in (Hussein et al., 2013) to represent the motion of the skeleton 3D joint coordinates over time. The temporal dependencies of the skeleton joints are obtained by a hierarchy of overlapped covariances. The Support Vector Machine (SVM) is used as a classifier.

Like the above Covariance-based approaches, we use the covariance matrix but in a different way that

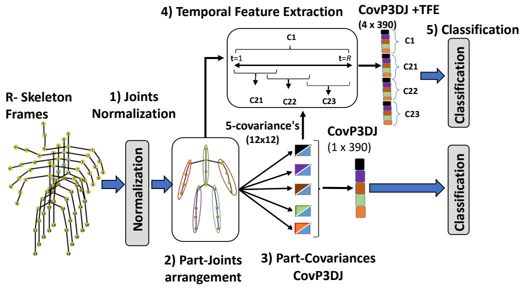


Figure 2: The overall architecture of the proposed recognition model. It receives frames sequence of length R. The recognition model works in 5 steps: 1) Joints Normalization, 2) Part-Joints arrangement, 3) Part-Covariances **CovP3DJ**, 4) Temporal Features Extraction (TFE), and 5) Classification.

enables the **CovP3DJ** descriptor to be more compacted and robust in representing the action classes. The skeleton 3D joints coordinates are grouped into 5 body parts to construct five covariance matrices. Constructing the covariance matrix for each body part (spine, arms, and legs) separately, then combining them after selecting only the lower/upper triangular part of each one has three main advantages. They are our contribution: 1) Reducing the size of the descriptor to about five times less than computing it over the whole skeleton joints as in (Hussein et al., 2013). 2) Making the training and testing time shorter. 3) Integrating five covariance matrices to represent the whole body motion over the sequences of the action frames.

3 THE FRAMEWORK

In this paper, our goal is to create a compact, robust, and discriminative descriptor to represent the dynamics of the body pose and motion of the skeleton joints 3D coordinates over a sequence of action frames. Towards this goal, we propose a covariance-based descriptor that is constructed from the skeleton 3D joint coordinates of the body parts which we call **CovP3DJ**. Basically, the overall proposed framework is illustrated in Figure 2, and summarized in the following steps: (1) Skeleton 3D joints coordinates normalization, (2) Part-joints arrangement (Spatial features), (3) Part-covariance (**CovP3DJ**) (4) Temporal Features Extraction (TFE), and (5) Classification.

(1) Joints Coordinates Normalization: the skeleton coordinates are normalized to force all the skeleton coordinates (X, Y, Z) of all joints to be in the interval [0,1] to become scale invariant. Equation 1 indicates how the X coordinates are normalized. The same equation is applied for the normalization of Y and Z coordinates.

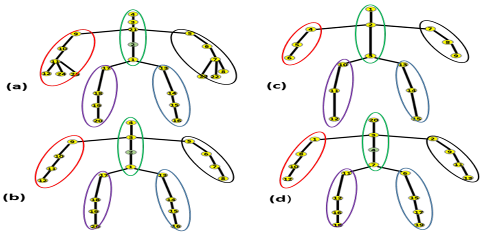


Figure 3: Skeleton arrangement: (a) 25-Joints skeleton of NTU RGB+D datast, (b) 20- Joints skeleton of MSR-12 Kinect Gesture and MSR-Action3D datasets, (c) 15-Joints skeleton of Florence3D Action dataset. and (d) 20-Joints skeleton of -Action dataset (it differs from 20-joint skeleton of (b) only in the body model of joints numbering).

$$X = (X - min(X))/(max(X) - min(X)) \qquad (1)$$

(2) Part-Joints Arrangement:, the skeleton 3D joints coordinates are divided into five groups of the body parts (spine, left/right arms, and left/right legs). The benefit of this arrangement is to extract the part motion among joints of the same body part that are strongly correlated and ignores the weak correlation. This step is implemented by grouping each part joints in a separate structure and horizontally stacking their X's, Y's and Z's coordinates as shown in Figure 3 .

(3) Part-Covariance, CovP3DJ: to represent the dynamic motion behavior of each body part over time, the covariance matrix is evaluated over part joints to measure how strong is the relation among them. Five covariance matrices are constructed for body parts; then the lower/upper triangular part of each matrix is selected and converted to a feature vector. All the 5 feature vectors are concatenated into one vector which we call CovP3DJ that represents the whole body motion.

Covariance Descriptor, C(**M**), measures how strong is the relation among **M** random variables by encoding their joint probability distribution. C(**M**) dimension is $\mathbf{M} \times \mathbf{M}$ which is of fixed size regardless of the number of frames composing the action sequence, as shown in Equation 3. Since C(**M**) is a symmetric positive definite matrix, we can select the lower/upper triangle part of the matrix to encode the entire information into $1 \times ((\mathbf{M} \times (\mathbf{M}+1))/2)$ dimension.

Assume that we have an action sequence of length R and N is the number of skeleton Joints of 3D coordinates (x,y,z) each. Then, $\mathbf{M} = 3\times$ N random variables as shown in Equation 2. Also, C(**M**) can be calculated over the action sequence of length R frames by the Equation 3 over all skeleton joints.

$$\mathbf{M} = [x_1, \ldots, x_N, y_1, \ldots, y_N, z_1, \ldots, z_N]' \qquad (2)$$

$$C(\mathbf{M}) = \frac{1}{R-1} \sum_{t=1}^{R} (\mathbf{M} - \bar{\mathbf{M}})(\mathbf{M} - \bar{\mathbf{M}})' \qquad (3)$$

For our **CovP3DJ**, C(**M**) is constructed over each body part to extract the spatial relation of the adjacent part joints. In this case, **M** has $(3 \times (N/5))$ random elements, assuming that all joints have the same number of joints. We denote **M** here by $\mathbf{M}_p$ (short for random variable of each part joints) and the number of joints N in each body part as N_p, see Equation 4. The covariance matrix, in this case, is called C($\mathbf{M}_p$), where p is the body part, as indicated in Equation 5.

$$\mathbf{M}_p = [x_1, ..., x_{N_p}, y_1, ..., y_{N_p}, z_1, ..., z_{N_p}]' \tag{4}$$

$$C(\mathbf{M}_p) = \frac{1}{R-1} \sum_{t=1}^{R} (\mathbf{M}_p - \bar{\mathbf{M}}_p)(\mathbf{M}_p - \bar{\mathbf{M}}_p)' \tag{5}$$

To clarify how the **CovP3DJ** reduces the descriptor size by **78.68 %** on 20-joints skeleton datasets compared with the one that is constructed over all skeleton joints. For **Cov3DJ** (Hussein et al., 2013), assume that we have a 20-joints skeleton ($N = 20$ joints) and 5 body parts (p) of ($N_p = N/5 = 4$) joints of 3D coordinates (x,y,z). Then, **M** will be ($N \times 3 = 60$) random elements, the C(**M**) dimension over the sequence of any action is ($\mathbf{M} \times \mathbf{M} = 60 \times 60 = 3600$). After the upper triangle part of C(**M**) is selected, its dimension becomes ($\mathbf{M}(\mathbf{M}+1)/2 = 1830$) as in (Hussein et al., 2013). While our **CovP3DJ** as shown in Figure 1 is slightly different, we have to construct the $C(\mathbf{M}_p)$, for each body part (p) of 4-joints of 3D coordinates ($N_p = 4$ joints). Then, $\mathbf{M}_p$ will have ($3 \times 4 = 12$) random variables, the generated part covariances, $C(\mathbf{M}_p)$, will be of size (12×12) , and the lower/upper triangular part of them has a dimension of $((12 \times 13)/2 = 78)$. Afterward, all part-covariance matrices are concatenated to generate one compacted and robust feature vector of size ($5 \times 78 = 390$). This means that our descriptor size is reduced by **78.68 %** than the descriptor used in (Hussein et al., 2013) over all skeleton joints. Similarly, the descriptor size is reduced by **78.26**% and **80.35** % for the 15-joints skeleton and 25-joints skeleton datasets respectively.

(4) Temporal Feature Extraction (TFE): To manage the action frames ordering, we use a temporal hierarchical overlapped covariances over the action sequence of the body parts as in (Hussein et al., 2013). Number of covariances (Nb_{cov}) at different levels (L) with overlap (OL) is calculated by $Nb_{cov} = 2^{(L-1)} * 2 - 1$. For example, at L=1 (over all sequence) the number of generated covariances is one (C1) and it is 3 (C21, C22, and C23) at L=2, see Figure 2 step (4). If TFE is used for two levels, both the proposed and Cov3DJ (Hussein et al., 2013) descriptors size are multiplied by 4. It means that our descriptor (**CovP3DJ**) is still reduced by **78.68 %** in the case of using temporal hierarchical covariances.

(5) Classification: We used the SVM to train the classification model on MSR-12 Kinect Gesture, MSR-Action3D, UTKinect-Action, and Florence3D datasets. The Random Forest (RF) is used to train the classifier on the NTU RGB+D dataset.

4 EXPERIMENTAL RESULTS AND EVALUATION

In order to evaluate the effectiveness of the proposed recognition system when **CovP3DJ** descriptor is used, five public datasets: MSR-Action3D (Li et al., 2010), MSR-12 Kinect Gesture (Fothergill et al., 2012), UTKinect-Action (Xia et al., 2012), Florence3D Action (Seidenari et al., 2013), and NTU RGB+D(Shahroudy et al., 2016) are tested. These datasets are different in the number of samples, subjects, the type of Kinect camera used, the number of action classes and the input modalities (skeleton, RGB and Depth map images) that are provided in each dataset. In the next subsections datasets description, experimental results, and analysis are discussed.

4.1 MSR-Action3D Dataset

MSR-Action3D (Li et al., 2010) dataset is collected using Kinect V1 which captures depth sequences of actions using depth sensors. The collected skeleton data consists of 20 joints. This dataset has twenty actions. The dataset actions are performed by ten subjects twice with a total number of samples 567. It is divided into three action sets: AS1, AS2, and AS3. Each set consists of 8 classes without overlap among sets. The dataset is evaluated using cross subject evaluation protocol in which half the subjects are used for training, and the rest is used for testing.

Comparative results against the state of the art methods on the MSR Action3D dataset are summarized in Table 1. The **CovP3DJ** average (over the three subsets) recognition accuracy is 90.98 %. It outperforms the majority of existing methods, competes with the approaches that got the highest recognition rates by a low margin difference; less than 1% and it outperforms the deep learning methods in (Veeriah et al., 2015) and (Martens and Sutskever, 2011) by a large margin.

For the sake of completeness, Table 2 records a set of approaches that reported their recognition accuracy's with the same assessments over the action subsets. The fifth column in Table 2 monitors the average accuracy of the three action subsets (AS1, AS2, and AS3). It is clear from the last row of Table 2 that our recognition accuracy 90.98 % is slightly lower than the other methods on AS2. This occurs because

Table 1: Comparative results with the state of the art on the MSR-Action3D dataset. The average accuracy (%) is measured on the three action subsets (AS1, AS2, and AS3).

Methods	Accuracy (%)
Rec. Neural Net. (Martens and Sutskever, 2011)	42.5
Hidden Markov Model (Xia et al., 2012)	78.97
Histograms of 3D joints (Xia et al., 2012)	78.97
EignJoints (Yang and Tian, 2012)	82.3
Space and Temporal part-sets (Wang et al., 2013)	90.22
Cov3DJ (Hussein et al., 2013)	90.53
Random Forest (Zhu et al., 2013)	90.90
Lie-Group (Vemulapalli et al., 2014)	89.48
HON4D (Oreifej and Liu, 2013)	88.89
DCSF (Xia and Aggarwal, 2013)	89.3
LSTM (Veeriah et al., 2015)	87.78
DMM-HOG (Yang et al., 2012)	85.52
LTBSVM (Slama et al., 2015a)	91.21
Pose-base (Wang et al., 2013)	90.2
HODG (Gowayyed et al., 2013)	91.3
Skeletal shape trajectories (Amor et al., 2016)	90
JSG(top-K RVJRD)+ JSGK (Graph) (Li and Leung, 2017b)	92.2
Bio-inspired Dynamic (Chaudhry et al., 2013)	90.0
FV of Skeleton Quads (Evangelidis et al., 2014a)	89.86
CovP3DJ, L=3, OL	**90.98**

Table 2: Recognition accuracy's over the three action sets: AS1, AS2, and AS3 of MSR Action3D dataset.

Method	AS1	AS2	AS3	Avg(%)
Cov3DJ (Hussein et al., 2013)	88.04	89.29	94.29	90.53
Bag of 3d points (Li et al., 2010)	72.9	71.9	79.2	74.7
(Chen et al., 2016)	96.2	83.2	92	90.47
HOD (Gowayyed et al., 2013)	92.39	90.18	91.43	91.26
Lie in Group (Vemulapalli et al., 2014)	95.29	83.87	98.22	91.26
CovP3DJ, L=3, OL	**93.48**	**84.82**	**94.29**	**90.98**

the Cov3DJ descriptor does not distinguish well between the opposite actions. The results of the proposed method are obtained with the number of levels (L) in the temporal hierarchy is 3 (L=3), and the overlap (OL) is enabled. It is mentioned in the Table 2 as (**CovP3DJ, L=3, OL**).

Table 3: Time comparison of a model training and testing in Seconds when the proposed **CovP3DJ** and **Cov3DJ** (Hussein et al., 2013) descriptors are used on the MSR-12 Kinect Gestures and MSR-Action3D datasets.

Datasets	**Experiments**	**CovP3DJ**	**Cov3DJ**	**Reduced Time** (%)
MSR-12 Kinect Gesture L=2, OL= True (Seconds)	**LOO**	**71.23**	322.06	**71.33**
	50% Split	28.53	144	**80.55**
	1/3 Training	25	103	**75.73**
	2/3 Training	50	189	**73.57**
MSR- Action3D L=3, OL=True (Seconds)	**AS1**	0.58	7.20	**94.3**
	AS2	0.57	6.33	**90.22**
	AS3	0.58	7	**91.7**

Table 3 monitors a time comparison between the CovP3DJ and Cov3DJ on the MSR-Action3D dataset. It is obvious that there is a significant time saving (about **91.5 %**) when our proposed descriptor is used. The obtained results in the table validate our hypothesis (regarding the significant reduction in processing time when the proposed descriptor is used).

4.2 MSR-12 Kinect Gesture Dataset

MSR-12 Kinect Gesture (Fothergill et al., 2012) is considered a relatively large dataset that is collected by Kinect V1 for action detection. It is annotated by (Hussein et al., 2013) to be used for action and gestures recognition. It consists of 594 sequences representing 12 classes collected by 30 subjects. Each sequence contains a gesture that is performed several times by a single subject. The dataset is evaluated using four different scenarios (Hussein et al., 2013): Leave one subject out (LOO), 50% subject split (cross subject), $1/3$ split, and $2/3$ split. The proposed descriptor, **CovP3DJ** is compared with the Cov3DJ (Hussein et al., 2013) since it gives the most recent results for the MSR-12 Gesture dataset.

Table 4 shows the recognition accuracies of Cov3DJ (Hussein et al., 2013) and the proposed **CovP3DJ**. Experiments in Table 4 are performed when the number of levels (L) in the temporal hierarchy is chosen to be one (L=1) and two (L=2) with/without overlapping among sub-sequences of the action. The proposed method delivers its highest score over the compared method in the 50 % subject split setup which means that **CovP3DJ** is robust against the reduced number of samples. On the other hand, the difference is small between the compared methods for the other experiments. Also, the obtained recognition rates of both methods can be ordered descendingly starting from the highest score for the $2/3$ training setup, then the $1/3$ training setup, the Leave one out, and finally the 50 % subject split. Results of the two compared methods are similar in all experimental setups, however ours has the advantages of reducing both the descriptor size and the time to 78.68 % and about 75% (See Table 3 for time comparison), respectively.

4.3 UTKinect-Action Dataset

UTKinect (Xia et al., 2012) dataset has been collected by Kinect V1 sensors. The captured skeleton has 20 joints with 3D coordinates. Also, 199 sequences are available in this dataset. It has 10 classes and each action is performed twice by 10 subjects. Although this dataset is small, it has a view and intra-class variation challenges. The experiments are conducted using cross subject evaluation protocol (Zhu et al., 2013); half subjects are used for training and the other for testing. The average recognition rate on the UTkinect dataset is collected after 10 different subjects splits.

Table 5 reports the obtained recognition accuracy's on the UTKinect dataset. The highest and lowest accuracies are 98.3% in (Li and Leung, 2017a) and

Table 4: Recognition rate (%) on MSR-12 Gesture dataset using different experiments and descriptor setup, where L=number of levels for computing the descriptor in hierarchy and OL equivalent to overlap between sub-sequences of the action.

	Cov3DJ (Hussein et al., 2013) (%)			**CovP3DJ** (%)		
	L=1	L=2	L=2, OL	L=1	L=2	L=2, OL
Leave One Out	92.7	93.6	93.6	92.33	92.88	93.19
50 % Subject split	90.3	91.2	91.7	**90.62**	**91.63**	**92.15**
1/3 Training	97.7	97.8	97.9	97.0	97.34	97.34
2/3 Training	98.6	98.7	98.7	98.14	98.39	98.45

Table 5: Comparative results of the state of the art on UTKinect dataset.

Method	Accuracy %
(Xia et al., 2012)	90.92
(Devanne et al., 2015)	91.5
(Vemulapalli et al., 2014)	97.08
(Slama et al., 2015b)	88.5
(Luvizon et al., 2017)	98.00
(Zhu et al., 2013)	87.90
(Ding et al., 2016)	94.5
(Li and Leung, 2017a)	98.3
CovP3DJ, L=2, OL=True	97.02

87.90% in (Zhu et al., 2013) respectively. Our recognition rate is 97.02% which is very near to the highest recorded accuracy on the UTKinect dataset. Figure 4 (a) indicates the output confusion matrix of our system on the dataset.

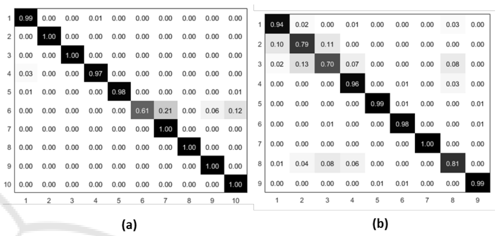
4.4 Florence3D-Action Dataset

Florence3D-Action (Seidenari et al., 2013) dataset has 215 action sequences of 9 actions. The dataset was collected using Kinect sensors and consists of 15-joints skeletons. Ten subjects performed each action three times. Intra-class variation challenges exist in this dataset because the same action may be performed by the left or right hand. We followed the cross subject evaluation protocol in (Zhu et al., 2013) to test the performance of the proposed model.

Table 6: Comparative results of the state of the art on Florence3D-Action dataset.

Method	Accuracy %
(Seidenari et al., 2013)	82.0
(Devanne et al., 2015)	87.04
(Vemulapalli et al., 2014)	90.88
(Slama et al., 2015b)	94.39
CovP3DJ, L=2, OL=True	91.00

Table 6 records comparative results with the state of art methods against the proposed method, **CovP3DJ**. We faced a problem regarding the order of action frames even if the temporal feature extraction is used. Our recognition rate is 91% which is considered the second highest score in the table. Figure 4 (b) illustrates the obtained confusion matrix of the **CovP3DJ** on the Folerence 3D dataset.



(a)

	1	2	3	4	5	6	7	8	9	10
1	0.99	0.00	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
3	0.00	0.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
4	0.03	0.00	0.00	0.97	0.00	0.00	0.00	0.00	0.00	0.00
5	0.01	0.00	0.00	0.00	0.98	0.00	0.00	0.00	0.00	0.01
6	0.00	0.00	0.00	0.00	0.00	0.61	0.21	0.00	0.06	0.12
7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00	0.00
8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00
10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	1.00

(b)

	1	2	3	4	5	6	7	8	9
1	0.94	0.02	0.00	0.01	0.00	0.00	0.00	0.03	0.00
2	0.10	0.79	0.11	0.00	0.00	0.00	0.00	0.00	0.00
3	0.02	0.13	0.70	0.07	0.00	0.00	0.00	0.08	0.00
4	0.00	0.00	0.00	0.96	0.00	0.01	0.00	0.03	0.00
5	0.00	0.00	0.00	0.00	0.99	0.01	0.00	0.00	0.01
6	0.00	0.00	0.01	0.00	0.00	0.98	0.00	0.00	0.01
7	0.00	0.00	0.00	0.00	0.00	0.00	1.00	0.00	0.00
8	0.01	0.04	0.08	0.06	0.00	0.00	0.00	0.81	0.00
9	0.00	0.00	0.00	0.00	0.01	0.01	0.00	0.00	0.99

Figure 4: Confusion Matrices of the proposed **CovP3DJ** on (a) UTKinect Aciton dataset and (b) Florence3D-Action dataset.

4.5 NTU RGB+D Dataset

NTU-RGB+D (Shahroudy et al., 2016) is the largest available dataset. This dataset consists of 56880 samples of 60 action classes. The actions are performed by 40 subjects and three Kinect cameras with different viewing angles and distances from the subjects are used. We followed the evaluation protocol suggested by the authors of this dataset(Shahroudy et al., 2016), which consists of two scenarios: Cross-View (CV) and Cross-Subject (CS).

Table 7: A comparative result of the state of the art on NTU RGB+D dataset.

	Method	CS (%)	CV (%)
1	HOG^2 (Ohn-Bar and Trivedi, 2013a)	32.24	22.27
2	Super Normal Vector (Yang and Tian, 2014)	31.82	13.61
3	HON4D (Oreifej and Liu, 2013)	30.56	7.26
4	Lie Group (Vemulapalli et al., 2014)	50.08	52.76
5	Skeletal Quads (Evangelidis et al., 2014b)	38.62	41.36
6	**CovP3DJ, L=1, OL=False**	**51.4**	**52.88**

Table 7 reports results related to some hand-crafted methods that have been conducted on NTU RGB+D dataset. It is shown in Table 7 that our **CovP3DJ** got the highest results (51.4% for CS and 52.88 % for CV) when L=1 and OL=False against the used hand-crafted approaches on the NTU RGB+D dataset. These results are obtained after normalizing the skeleton coordinates and camera views in order to make the model skeleton and views invariant.

5 CONCLUSIONS AND FUTURE WORK

We presented a skeleton part based covariance descriptor (**CovP3DJ**) for human action recognition. **CovP3DJ** is the concatenation of the lower/upper triangular parts of five applied covariance matrices over body parts (spine, arms, and legs). The descriptor achieves a saving of 78.36 %, 78.68%, and 80.35 for 15-joints, 20-joints, and 25-joints skeleton datasets in storage space and from 75% to 90% in processing time (depending on the dataset) compared with techniques adopting a covariance descriptor based on all the skeleton joints. The obtained average recognition accuracy's are 90.98 %, 97.02% and 91.00% on MSR-Action3D dataset, UTKinect and Florence3D datasets respectively. It is between 92.33 and % 93.19 % using Leave One Out method and lies between 90.62 % and 92.15 % using the 50% split, 97.34% using 1/3 Traing split and about 98.45 % using 2/3 Training split. The **CovP3DJ** acheived recognition rates of 51.4 % in CS and 52.88 % in CV on the NTU RGB+D dataset. Meanwhile, all the obtained recognition rates of the proposed framework on all datasets outperform the majority of existing methods and compete with the state of the art. However, **CovP3DJ** is simple and efficient in both space and time consumption compared with other methods.

For the future work, we will integrate other features/modalities beside the **CovP3DJ** to enhance the power of the proposed framework in detecting not only the pose and motion but also the direction of the motion to reduce the miss-classification rates.

ACKNOWLEDGEMENTS

We would like to acknowledge Egypt-Japan University of Science and Technology (E-JUST), specially the Cyber Physical System Laboratory (CPS) which provided us with all the required hardware and software tools to achieve this piece of research in Human Action Recognition. Also, We attribute credit to Nanyang Technology Institution for providing us with their NTU RGB+D dataset.

REFERENCES

Aggarwal, J. K. and Ryoo, M. S. (2011). Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16.

Amor, B. B., Su, J., and Srivastava, A. (2016). Action recognition using rate-invariant analysis of skeletal shape trajectories. *IEEE transactions on pattern analysis and machine intelligence*, 38(1):1–13.

Chaudhry, R., Ofli, F., Kurillo, G., Bajcsy, R., and Vidal, R. (2013). Bio-inspired dynamic 3D discriminative skeletal features for human action recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 471–478.

Chen, C., Liu, K., and Kehtarnavaz, N. (2016). Real-time human action recognition based on depth motion maps. *Journal of real-time image processing*, 12(1):155–163.

Devanne, M., Wannous, H., Berretti, S., Pala, P., Daoudi, M., and Del Bimbo, A. (2015). 3-d human action recognition by shape analysis of motion trajectories on riemannian manifold. *IEEE transactions on cybernetics*, 45(7):1340–1352.

Ding, W., Liu, K., Cheng, F., and Zhang, J. (2016). Learning hierarchical spatio-temporal pattern for human activity prediction. *Journal of Visual Communication and Image Representation*, 35:103–111.

Evangelidis, G., Singh, G., and Horaud, R. (2014a). Skeletal quads: Human action recognition using joint quadruples. *Ieee Icpr*, pages 4513–4518.

Evangelidis, G., Singh, G., and Horaud, R. (2014b). Skeletal quads: Human action recognition using joint quadruples. In *Pattern Recognition (ICPR), 2014 22nd International Conference on*, pages 4513–4518. IEEE.

Fothergill, S., Mentis, H., Kohli, P., and Nowozin, S. (2012). Instructing people for training gestural interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1737–1746. ACM.

Gavrila, D. M., Davis, L. S., et al. (1995). Towards 3-d model-based tracking and recognition of human movement: a multi-view approach. In *International workshop on automatic face-and gesture-recognition*, pages 272–277.

Gowayyed, M. A., Torki, M., Hussein, M. E., and El-Saban, M. (2013). Histogram of oriented displacements (hod): Describing trajectories of human joints for action recognition. In *IJCAI*, pages 1351–1357.

Hussein, M. E., Torki, M., Gowayyed, M. A., and El-Saban, M. (2013). Human action recognition using a temporal hierarchy of covariance descriptors on 3d joint locations. In *IJCAI*, volume 13, pages 2466–2472.

Koestinger, M., Hirzer, M., Wohlhart, P., Roth, P. M., and Bischof, H. (2012). Large scale metric learning from equivalence constraints. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2288–2295. IEEE.

Li, M. and Leung, H. (2017a). Graph-based approach for 3D human skeletal action recognition. *Pattern Recognition Letters*, 87:195–202.

Li, M. and Leung, H. (2017b). Graph-based approach for 3d human skeletal action recognition. *Pattern Recognition Letters*, 87:195–202.

Li, W., Zhang, Z., and Liu, Z. (2010). Action recognition based on a bag of 3d points. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*, pages 9–14. IEEE.

Luvizon, D. C., Tabia, H., and Picard, D. (2017). Learning features combination for human action recognition from skeleton sequences. *Pattern Recognition Letters*.

Ma, B., Su, Y., Ma, B., and Su, Y. (2014). Covariance Descriptor based on Bio-inspired Features for Person Reidentification and Face Verification To cite this version : Covariance Descriptor based on Bio-inspired Features for Person re-Identification and Face Verification.

Martens, J. and Sutskever, I. (2011). Learning recurrent neural networks with hessian-free optimization. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 1033–1040.

Müller, M., Röder, T., Clausen, M., Eberhardt, B., Krüger, B., and Weber, A. (2007). Mocap database hdm05. *Institut für Informatik II, Universität Bonn*, 2:7.

Ohn-Bar, E. and Trivedi, M. (2013a). Joint angles similarities and hog2 for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 465–470.

Ohn-Bar, E. and Trivedi, M. M. (2013b). Joint angles similarities and HOG2 for action recognition. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 465–470.

Oreifej, O. and Liu, Z. (2013). Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 716–723.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Seidenari, L., Varano, V., Berretti, S., Bimbo, A., and Pala, P. (2013). Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 479–485.

Shahroudy, A., Liu, J., Ng, T.-T., and Wang, G. (2016). Ntu rgb+ d: A large scale dataset for 3d human activity analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1010–1019.

Slama, R., Wannous, H., Daoudi, M., and Srivastava, A. (2015a). Accurate 3d action recognition using learning on the grassmann manifold. *Pattern Recognition*, 48(2):556–567.

Slama, R., Wannous, H., Daoudi, M., Srivastava, A., Slama, R., Wannous, H., Daoudi, M., Srivastava, A., and Action, A. (2015b). Accurate 3D Action Recognition using Learning on the Grassmann Manifold Accurate 3D Action Recognition using Learning on the Grassmann Manifold. 48:556–567.

Veeriah, V., Zhuang, N., and Qi, G.-J. (2015). Differential recurrent neural networks for action recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4041–4049.

Vemulapalli, R., Arrate, F., and Chellappa, R. (2014). Human action recognition by representing 3d skeletons as points in a lie group. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 588–595.

Wang, C., Wang, Y., and Yuille, A. L. (2013). An approach to pose-based action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 915–922.

Xia, L. and Aggarwal, J. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2834–2841.

Xia, L., Chen, C.-C., and Aggarwal, J. (2012). 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. IEEE.

Yang, X. and Tian, Y. (2014). Super normal vector for activity recognition using depth sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 804–811.

Yang, X. and Tian, Y. L. (2012). Eigenjoints-based action recognition using naive-bayes-nearest-neighbor. In *Computer vision and pattern recognition workshops (CVPRW), 2012 IEEE computer society conference on*, pages 14–19. IEEE.

Yang, X., Zhang, C., and Tian, Y. (2012). Recognizing actions using depth motion maps-based histograms of oriented gradients. In *Proceedings of the 20th ACM international conference on Multimedia*, pages 1057–1060. ACM.

Zhu, Y., Chen, W., and Guo, G. (2013). Fusing spatiotemporal features and joints for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 486–491.