Human Activities Recognition using 3D Pose Based Body Joint Angles

Md Matiqul Islam

Department of Information and Communication Engineering, University of Rajshahi, Rajshahi-6205, Bangladesh

Abstract. This paper represents, the classification of user activities such as standing, walking and running, based on right shoulder, left shoulder, right elbow, left elbow, right knee, left knee and heap 3d angles data. Four supervised classification techniques namely, k-Nearest Neighbor (k-NN), Support Vector Machines (SVM), Gaussian Naive Bayes (GNB), and Linear Discriminant Analysis (LDA) are compared in terms of correct classification rate, F-measure, recall, precision, and specificity. Based on our experiments, the results obtained show that the k-NN classifier provides the best performance compared to other supervised classification algorithms.

Keywords: k-NN, SVM, GNB, LDA.

1. Introduction

The capability of classifying the physical activity performed by a human is highly attractive for many applications in the field of computer vision such as healthcare monitoring, transportation mode recognition [1], indoor positioning, navigation, location-based services, context-aware behaviors,targeted advertising, and mobile social networks [2], customer behavior analysis in shopping mall and in developing advanced human-machine interfaces.Recent years have witnessed a significantincrease in the variety of consumer devices, which are not only equipped with traditional sensors like GPS, camera, Wi-Fi and Bluetooth but also newly-developed sensors like accelerometer, gyroscope,and barometer. These sensors can capture the intensity and duration of the activity and are even ableto sense the activity context. This can help consumers assess their activity levels and change theiractivity behaviors to keep fit and healthy.

Equipped with a variety of sensors, smartphones, on-body devices are more attractive for activity recognition compared to our proposed model 3D joint angle data because our model do not disturbusers' normal activities [3].

In this paper the most general approaches to automatic classification of human physical activity such as standing, walking and running are introduced and discussed. With regards to this problem, the main steps considering as generating 3D pose of human, calculating the 3d joint angles (right shoulder, left shoulder, right elbow, left elbow, right knee, left knee and heap), feature selection, extraction and classification are reexamined by following the diagram of Figure 1.

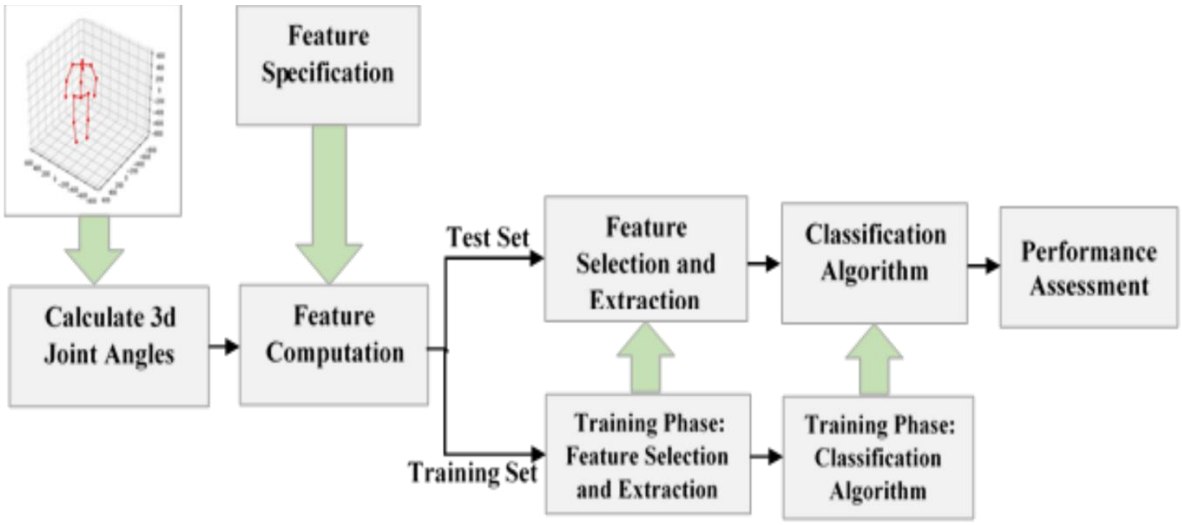

Figure 1. Conceptual scheme of a generic classification system with supervised learning.

1.1 Related Works

In the last years, many solutions for human activity recognition have been proposed, some of them aimed to extract features from depth data, such as [4], where the main idea is to evaluate spatiotemporal depth sub volume descriptors. A group of hyper surface normals (polynormal), containing geometry and local motion information, is extracted from depth sequences. The polynormals are then aggregated to constitute the final representation of the depth map, called Super Normal Vector (SNV). This representation can include also skeleton joint trajectories, improving the recognition results when people move a lot in a sequence of depth frames. Depth images can be seen as sequence features modeled temporally as subspaces lying on the Grassmann manifold [5]. This representation, starting from the orientation of the normal vector at every surface point, describes the geometric appearance and the dynamic of human body without using joint position. Other works proposed holistic descriptors: the HON4D descriptor [6], which is based on the orientations of normal surfaces in 4D, and HOPC descriptor [7], which is able to represent the geometric characteristics of a sequence of 3D points.

Other works exploit both depth and skeleton data; for example, the *3.5D representation* combines the skeleton joint information with features extracted from depth images, in the region surrounding each node of interest [8]. The features are extracted using an extended Independent Subspace Analysis (ISA) algorithm by applying it only to local region of joints instead of the entire video, thus improving the training efficiency. The depth information makes it easy to extract the human silhouette, which can be concatenated with normalized skeleton features, to improve the recognition rate [9]. Depth and skeleton features can be combined at different levels of the activity recognition algorithm. Althloothi et al. [10] proposed a method where the data are fused at the kernel level, instead of the feature level, using the Multiple Kernel Learning (MKL) technique. On the other hand, fusion at the feature level of spatiotemporal features and skeleton joints is performed in [11]. In such a work, several spatiotemporal interest point detectors, such as Harris 3D, ESURF [12], and HOG3D [13], have been fused using regression forests with the skeleton joint features consisting of posture, movement, and offset information. Skeleton joints extracted from depth frames can be combined also with RGB data. Luo et al. [14] proposed a human action recognition framework where the pair wise relative positions of joints and Center-Symmetric Motion Local Ternary Pattern (CS-Mltp) features from RGB are fused both at feature level and at classifier level.

1.2 Generating 3d Pose of Human

Figure 1 illustrates the main contribution of our approach, a new multi-stage CNN architecture that can be trained end-to-end to estimate jointly 2D and 3D joint locations. Crucially it includes a novel layer, based on a probabilistic 3D model of human pose, responsible for lifting 2D poses into 3D and propagating 3D information about the skeletal structure to the 2D convolutional layers. In thisway, the prediction of 2D pose benefits from the 3D information encoded. Section 4 describes the new probabilistic 3D model of human pose, trained on a dataset of 3D mocap data. Section 5 describes all the new components and layers of the CNN architecture. Finally, Section 6 describesexperimental evaluation on the Human3.6M dataset wherewe obtain state-of-the-art results. In addition, we show qualitative results on images from the MPII and Leeds datasets

One fundamental challenge in creating models of humanposes lies in the lack of access to 3D data of sufficient variety to characterize the space of human poses. To compensate for this lack of data we identify and eliminate confounding factors such as rotation in the ground plane, limb length, and left-right symmetry that lead to conceptually similar poses being unrecognized in the training data. Simple preprocessing eliminates some factors. Size variance is addressed by normalizing the data such that the sum of squared limb lengths on the human skeleton is one; while left-right symmetry is exploited by flipping each pose in the x-axis and re-annotating left as right and vice-versa.

1.3 Aligning 3D Human Poses in the Training Set

Allowing for rotational invariance in the ground-planeis more challenging and requires integration with the datamodel. We seek the optimal rotations for each pose suchthat after rotating the poses they are closely approximatedby a low-rank compact Gaussian distribution.We formulate this as a problem of optimization over aset of variables. Given a set ofNtraining 3D poses, each represented as a (3×L) Matrix Pi of 3D landmark locations, where i∈{1,2,..,N} and L is the number of humanjoints/landmarks; we seek global estimates of an average3D poseμ, a set ofJorthonormal basis matrices1eandnoise varianceσ, alongside per sample rotationsR I and basis coefficientsaito minimize the following estimate

Where ai·e=∑jai,jej is the tensor analog of a multiplication between a vector and a matrix, and||·||22 is the squared Frobenius norm of the matrix. Here the y-axis is assumed to point up, and the rotation matrices Ri considered are ground plane rotations. With the large number of 3D pose samples considered (of the order of 1 million when training on the Human3.6M dataset [15]), and the complex interdependencies between samples for e and σ , the memory requirements mean that it is not possible to solve directly as a joint optimization over all variables using a non linear solver such as Ceres. Instead, we carefully initializeand alternate between performing closed-form PPCA [38] to update μ,a, e,σ; and updating Ri using Ceres [2] to minimize the above error. As we do this, we steadily increase the size of the basis from 1 through to its target size J. This stops apparent deformations that could be resolved through rotations from becoming locked into the basis at an early stage, and empirically leads to lower cost solutions.

To initialize we use a variant of the Tomasi-Kanade [39]algorithm to estimate the mean 3D pose μ. As theycomponent is not altered by planar rotations, we take as our estimate of theycomponent ofμ, the mean of each point intheydirection. For thexandzcomponents, we interleavethexandzcomponents of each sample and concatenate them into a large 2N×L matrix M, and find the rank twoapproximation of this such thatM≈A·B. We then calculateˆAby replacing each adjacent pair of rows of A with the closest orthonormal matrix of rank two, and takeˆA†Mas our estimate2of the x and z components of μ.

The end result of this optimization is a compact lowrank approximation of the data in which all reconstructed poses appear to have the same orientation (see Figure 2). Inthe next section we extend the model to be described as a multi-modal distribution to better capture the variations in the space of 3D human poses.

1.4 3D Joint Angles Calculation

To compute the angle between two vectors in 3D space:

- Calculate the magnitude of the vectors
- Divide each vector by its vector magnitude to compute its unit vector
- Compute the dot product of the unit vectors
- Take the arcosine of the dot product of the unit vectors to get the angle between the vectors in radians.

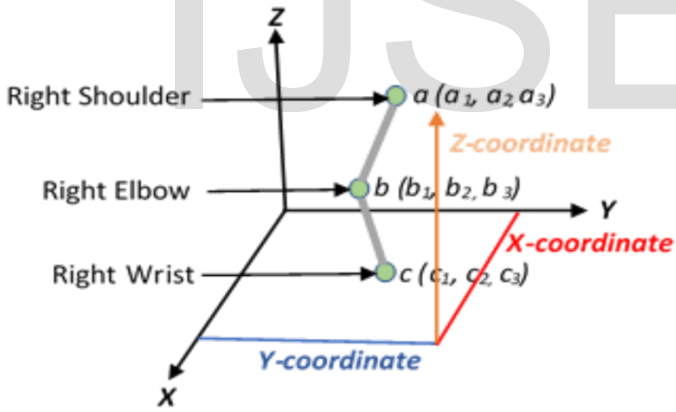

Figure2: Angle calculation of right elbow joint from 3D coordinate space.

Forexample, if we calculate the angle of right elbow as shown in Figure 1 we have to generate the vector A and$B$as

$$A = ab = (a_1 - b_1, a_2 - b_2, a_3 - b_3) = [A_1, A_2, A_3]$$
$$B = bc = (b_1 - c_1, b_2 - c_2, b_3 - c_3) = [B_1, B_2, B_3]$$

The dot product of vector A and B is defined as

$$A.B = A_1B_1 + A_2B_2 + A_3B_3$$

The magnitude of a vector A is denoted by ||A||. The dot product of vector A with itself is

$$A.A = ||A||^2 = A_1^2 + A_2^2 + A_3^2$$

Which gives

$$||A|| = \sqrt{A.A} = \sqrt{A_1^2 + A_2^2 + A_3^2}$$

Similarly, we can calculate

$$||B|| = \sqrt{B.B} = \sqrt{B_1^2 + B_2^2 + B_3^2}$$

The dot product of two non-zero Euclidean vectors A and B is given by

$$A.B = ||A||||B|| \cos\theta$$

Where θ is the angle between A and B.

Similar way we calculate the right shoulder, left shoulder, left elbow, heap, right knee and left knee joint angles labeled in Figure 3. The graphical plot of different joint angles to classify the standing, walking and running are shown in Figure 4, 5 and 6 respectively.

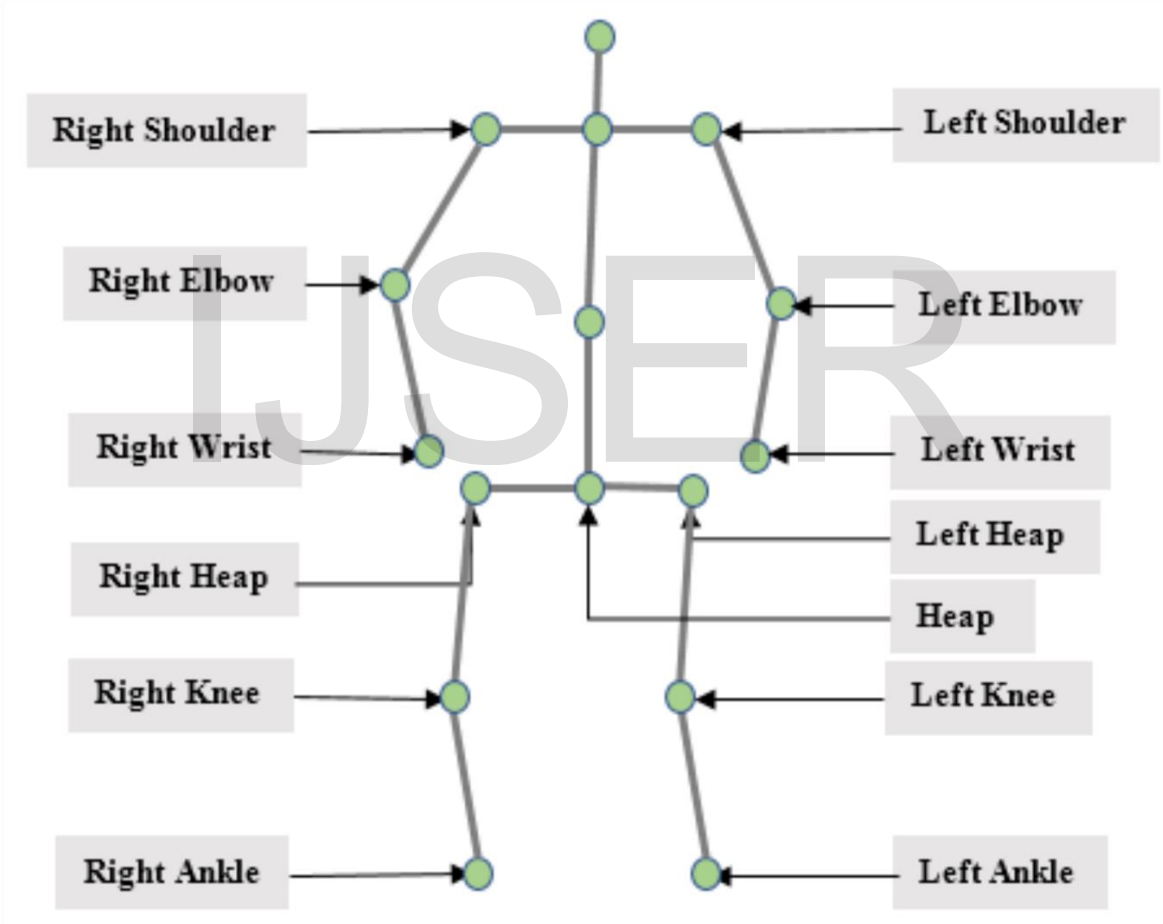

Figure 3. Graphical illustration of different joint angles

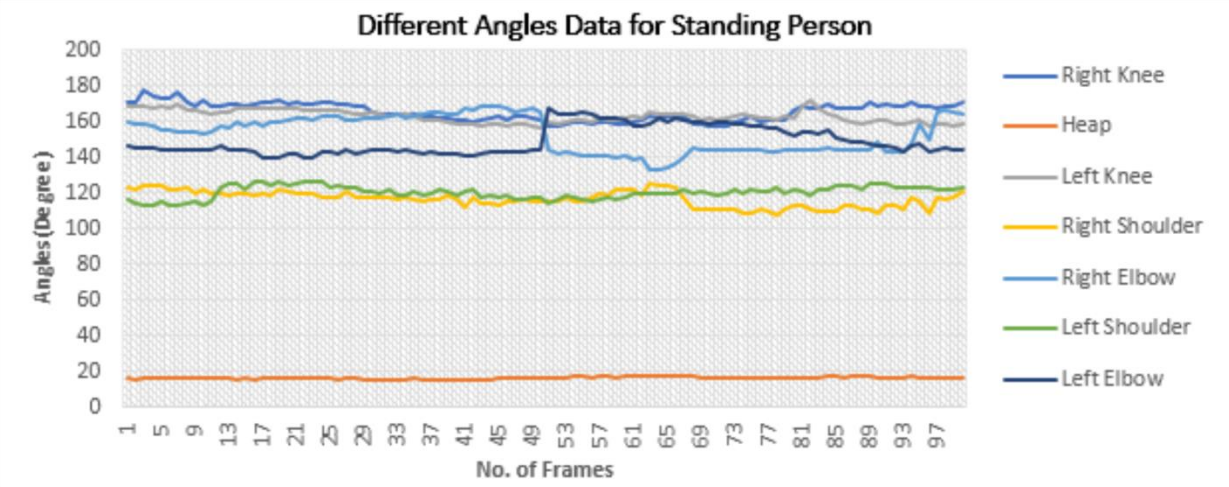

Figure4: Plot of the degree of 7 joint angles used for classification.

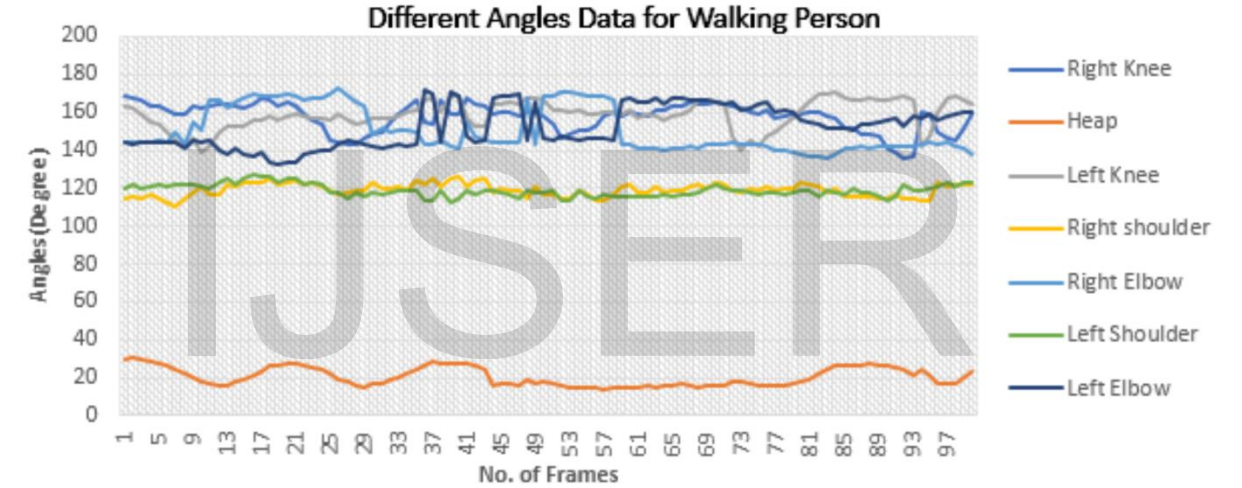

Figure5: Plot of the degree of 7 joint angles used for classification.

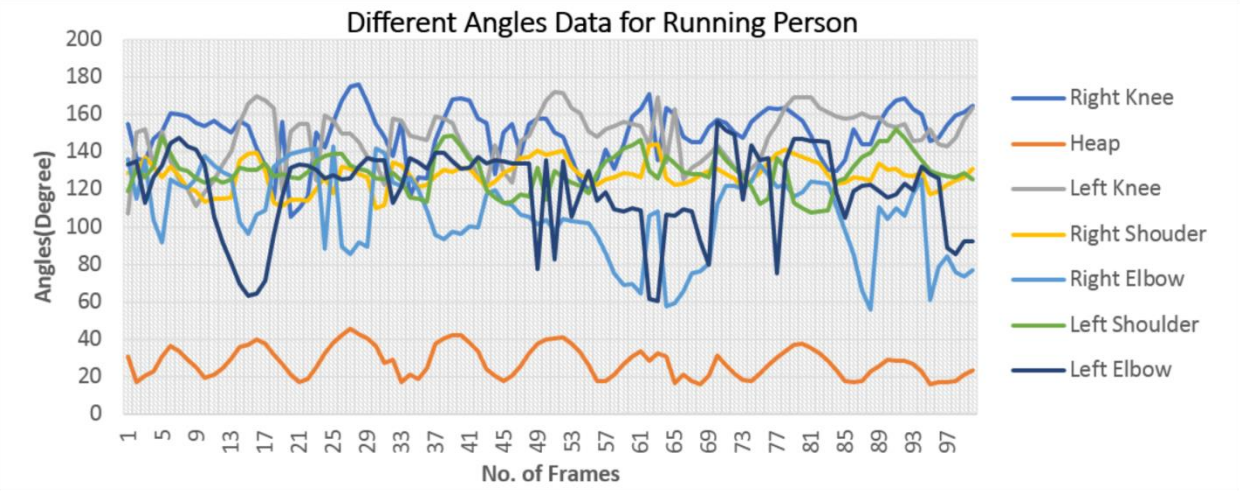

Figure6: Plot of the degree of 7 joint angles used for classification.

2. Performance Metrics

To verify the performance of the proposed models, we employed five widely used evaluation metrics for multi-class classification.

i. Precision

The precision or positive predictive value (PPV) is defined as the proportion of instances that belongs to a class (TP: True Positive) by the total instances, including TP and FP (False Positive) classified by the classifier as belong to this particular class.

$$\boldsymbol{Precision = TP/(TP + FP)}$$

ii. Recall

The recall or sensitivity is defined as the proportion of instances classified in one class bythe total instances belonging to that class. The total number of instances of a class includes TP and FN (False Negative).

$$\boldsymbol{Recall = TP/(TP + FN)}$$

iii. Accuracy

Measures the proportion of correctly predicted labels over all predictions:

$$\boldsymbol{Over\ all\ accuracy = (TP + TN)/(TP + TN + FP + FN)}$$

iv. F1 measure:

A weighted harmonic means of precision and recall. The F1 score can be interpreted as a weighted average of the precision and recall, where an F1 score reaches it best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score equal. The formula for the F1 measure is:

$$\boldsymbol{F1\ measure = (2 * Precision * Recall)/(Precision + Recall)}$$

v. Support:

The support is the number of samples of the true response that lie in that class.

4. Experimental Results

In this section, we review and compare the performance of different supervised machine learning approaches to recognize the different human activities.

The classification result obtained from our experiment are shown in Tables 1,2,3 and 4. It is seen from tables that k-NN classifier gives the highest accuracy and averagely it is 94%. Gaussian Naïve Bayes (GNB) and Linear Discriminant Analysis (LDA) classifiers accuracy are approximately same and it is averagely 89% and 88% respectively.

Activity(knn)	Precision	Recall	F1-Score	Support
Standing	0.89	0.98	0.93	200
Walking	0.96	0.89	0.92	200
Running	0.99	0.95	0.97	200
Avg/Total	0.94	0.94	0.94	600

Table1: Recognition accuracy for k-NN classifier.

Activity(gnb)	Precision	Recall	F1-Score	Support
Standing	0.82	0.95	0.88	200
Walking	0.87	0.83	0.85	200
Running	0.99	0.87	0.93	200
Avg/Total	0.89	0.89	0.89	600

Table2: Recognition accuracy for GNB classifier.

Activity(lda)	Precision	Recall	F1-Score	Support
Standing	0.74	0.94	0.83	200
Walking	0.89	0.74	0.81	200
Running	1.00	0.88	0.94	200
Avg/Total	0.88	0.86	0.86	600

Table3: Recognition accuracy for LDA classifier.

Activity(svm)	Precision	Recall	F1-Score	Support
Standing	0.96	0.78	0.86	200
Walking	0.96	0.52	0.67	200
Running	0.61	1.00	0.75	200
Avg/Total	0.84	0.77	0.76	600

Table4: Recognition accuracy for SVM classifier.

Figure7 shows the confusion matrix for the recognition result of k-NN classifier. It can be found that only 35 out of 600 samples are misclassified. The activities “standing” and “walking” are less discriminative than running in this case.

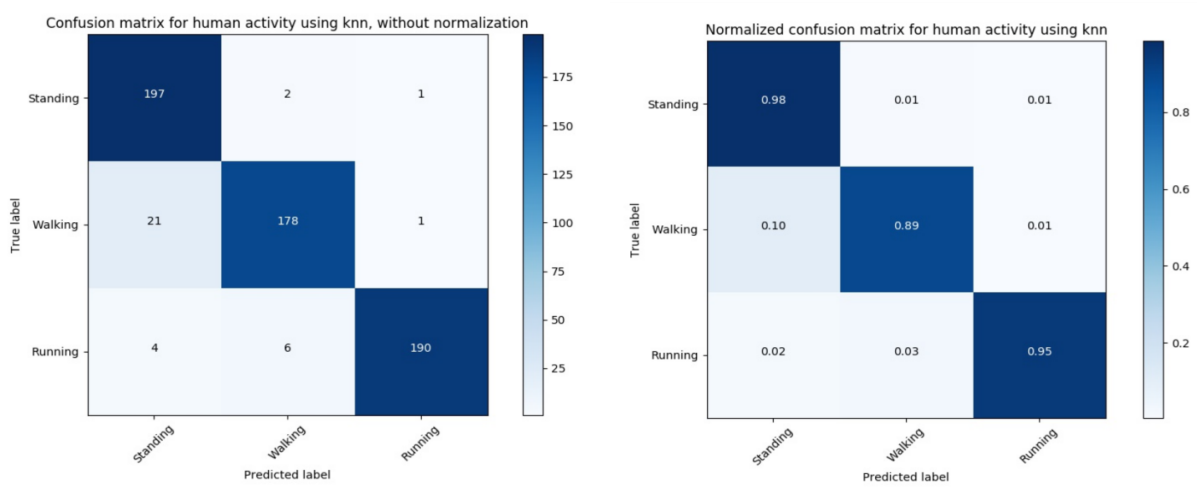

Figure7: Confusion matrix for k-NN classifier.

Figure8 shows the confusion matrix for the recognition result of GNB classifier. It can be found that only 69 out of 600 samples are misclassified. The activities "standing" and "walking" are less discriminative than running in this case.

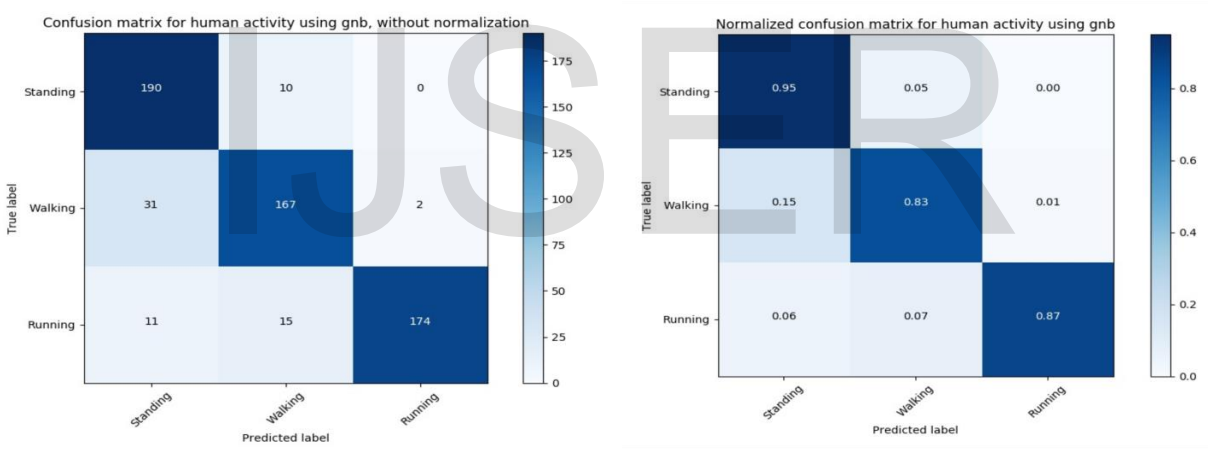

Figure8: Confusion matrix for GNB classifier.

Figure9 shows the confusion matrix for the recognition result of LDA classifier. It can be found that only 86 out of 600 samples are misclassified. The activities "standing" and "walking" are less discriminative than running in this case.

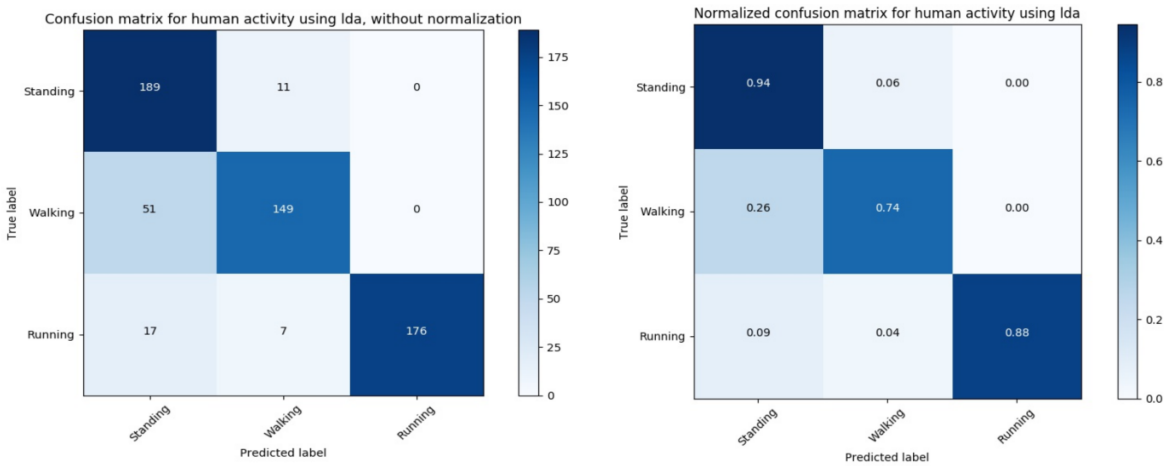

Figure9: Confusion matrix for LDA classifier.

Figure 10 shows the confusion matrix for the recognition result of SVM classifier. Each result of the confusion matrix gives the number of samples that are classified to certain activity classes labeled by the columns. Each diagonal element in the matrix gives the number of samples belonging to one activity that are correctly classified. It can be found that only 141 out of 600 samples are misclassified. The activities "standing" and "walking" are less discriminative than running in this case.

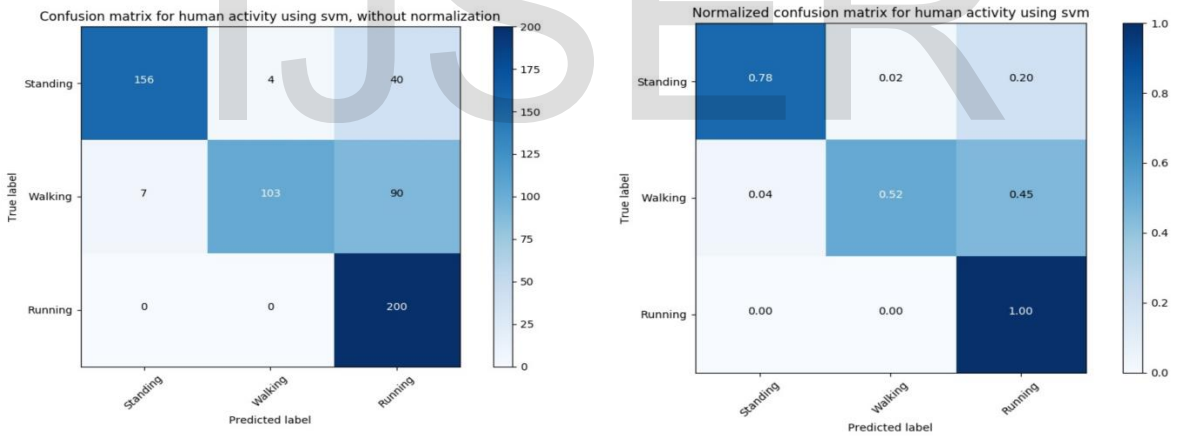

Figure10: Confusion matrix for SVM classifier.

Classifier	Accuracy on training set	Accuracy on validation set
SVM	1.00	0.77
k-NN	0.96	0.94
LDA	0.85	0.86
GNB	0.87	0.89

Table5: Accuracy on training set and validation set on different classifier.

Figure 11 shows the learning curve of different classifier. The learning curve shows the validation and training score of an estimator for varying numbers of training samples. If both the validation score and

the training score converge to a value that is too low with increasing size of the training size of the training set, we will not benefit much from more training data

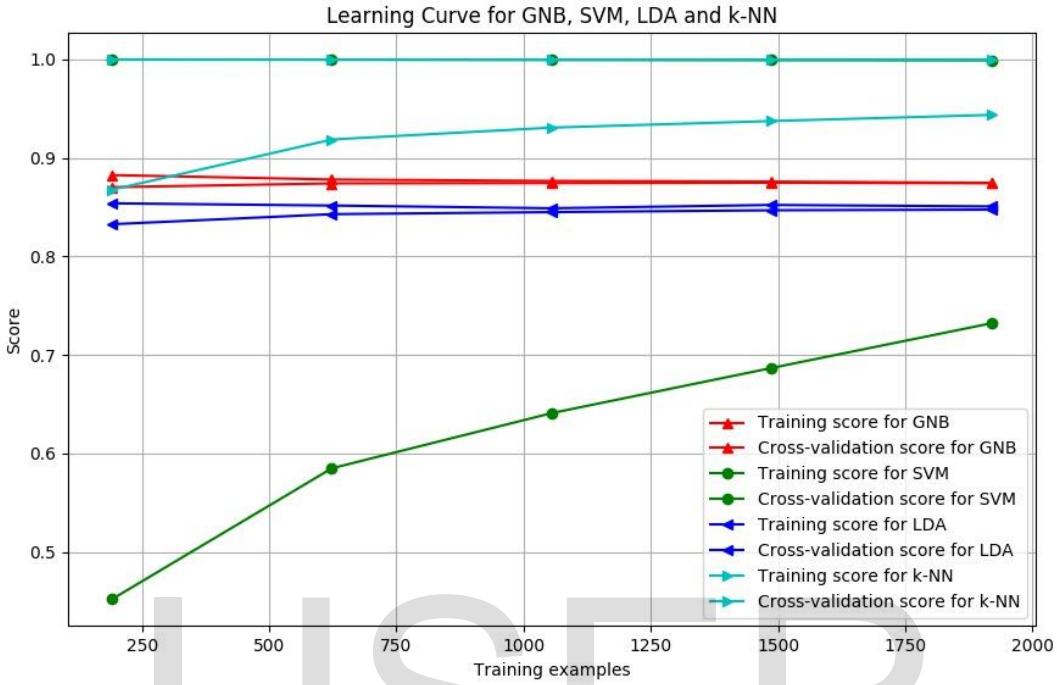

Figure 11: Learning curve for GNB, SVM, LDA and k-NN classifier.

Conclusion

We have presented a review of different classification techniques that were used to recognize human activities from 3d joint skeletal data. This paper describes the whole structure of the recognition detection process, from data acquisition to classification.

References

[1] Incel, O.D.; Kose, M.; Ersoy, C. A review and taxonomy of activity recognition on mobile phones.BioNanoScience2013,3, 145–171.
[2] Lockhart, J.W.; Pulickal, T.; Weiss, G.M. Applications of mobile activity recognition. In Proceedings of the2012 ACM Conference on Ubiquitous Computing, Pittsburgh, PA, USA, 5–8 September 2012; pp. 1054–1058.
[3]Tome D, Russell C, Agapito L. Lifting from the deep: Convolutional 3d pose estimation from a single image. CVPR 2017 Proceedings. 2017 Jul 26:2500-9.
[4] X. Yang and Y. Tian, "Super normal vector for activity recognition using depth sequences," in *Proceedings of the 27th IEEE Conference on Computer Vision and Pattern Recognition (CVPR'14)*, pp. 804–811, IEEE, Columbus, Ohio, USA, June 2014.
[5] R. Slama, H. Wannous, and M. Daoudi, "Grassmannian representation of motion depth for 3D human gesture and action recognition," in *Proceedings of the 22nd International Conference*

on Pattern Recognition (ICPR '14), pp. 3499–3504, IEEE, Stockholm, Sweden, August 2014.
[6] O. Oreifej and Z. Liu, "HON4D: histogram of oriented 4D normals for activity recognition from depth sequences," in *Proceedings of the 26th IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*, pp. 716–723, IEEE, June 2013.
[7] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian, "HOPC: histogram of oriented principal components of 3D pointclouds for action recognition," in *Computer Vision—ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds., vol. 8690 of *Lecture Notes in Computer Science*, pp. 742–757, Springer, 2014.
[8] G. Chen, D. Clarke, M. Giuliani, A. Gaschler, and A. Knoll, "Combining unsupervised learning and discrimination for 3D action recognition," *Signal Processing*, vol. 110, pp. 67–81, 2015.
[9] A. A. Chaaraoui, J. R. Padilla-Lopez, and F. Fl ´ orez-Revuelta, "Fusion of skeletal and silhouette-based features for human action recognition with RGB-D devices," in *Proceedings of the 14th IEEE International Conference on Computer Vision Workshops (ICCVW '13)*, pp. 91–97, IEEE, Sydney, Australia, December 2013.
[10] S. Althloothi, M. H. Mahoor, X. Zhang, and R. M. Voyles, "Human activity recognition using multi-features and multiple kernel learning," *Pattern Recognition*, vol. 47, no. 5, pp. 1800–1812, 2014.
[11] Y. Zhu, W. Chen, and G. Guo, "Fusing spatiotemporal features and joints for 3D action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW '13)*, pp. 486–491, IEEE, June 2013.
[12] G. Willems, T. Tuytelaars, and L. Van Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *Computer Vision—ECCV 2008*, D. Forsyth, P. Torr, and A. Zisserman, Eds., vol. 5303 of *Lecture Notes in Computer Science*, pp. 650–663, Springer, Berlin, Germany, 2008.
[13] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal ¨ descriptor based on 3D-gradients," in *Proceedings of the 19th British Machine Vision Conference (BMVC '08)*, pp. 995–1004, September 2008.
[14] J. Luo, W. Wang, and H. Qi, "Spatio-temporal feature extraction and representation for RGB-D human action recognition,"*Pattern Recognition Letters*, vol. 50, pp. 139–148, 2014.