Abstract—In 3D human motion pose-based analysis, the main problem is how to classify multi-class label activities based on primitive action (pose) inputs efficiently for both accuracy and processing time. Because, pose is not unique and the same pose can be anywhere on different activity classes. In this paper, we evaluate the effectiveness of Extreme Learning Machine (ELM) in 3D human motion analysis based on pose cluster. ELM has reputation as eager classifier with fast training and testing time but the classification result originally has still low testing accuracy even by increasing the hidden nodes number and adding more training data. To achieve better accuracy, we pursue a feature selection method to reduce the dimension of pose cluster training data in time sequence. We propose to use frequency of pose occurrence. This method is similar like bag of words which is a sparse vector of occurrence counts of poses in histogram as features for training data (bag of poses). By using bag of poses as the optimum feature selection, the ELM performance can be improved without adding network complexity (Hidden nodes number and training data).

I. INTRODUCTION

Nowadays, the most challenging task in computer vision based human motion analysis is to understand and recognize human action in semantic interpretation (high level vision) rather than intermediate-level (human tracking) or low-level vision (Human detection) [11]. Semantic meaning of action recognition is actually a task to label or classify an activity motion as belongs to one of some meaningful action classes by using machine learning algorithms. The classifier should perform efficiently to give both well-acceptance accuracy and fast processing time especially for real time application.

Extreme Learning Machine (ELM) has reputation as “extreme” in processing speed machine learning [3], however, the effectiveness of ELM in human motion analysis is largely unknown. How ELM, as eager classifier with fast processing time reputation, can deal with human activity classification based on pose cluster with semantic meaning. Such problem has reputation as hard classification problem due to pose cluster characteristic. Pose cluster in motion activity has feature characteristic for pose position in time sequence (spatiotemporal) and the frequency of pose occurrence in motion sequence. Both features are not unique and could be anywhere distributed on different activity classes. The distribution of features depends on the regularity of motion activity. For example, dance actions can be viewed as more regular than badminton sport actions. Dance action has a more strict rule followed by the dancer during her performance. In this paper, we evaluate the effectiveness of ELM classifier to deal with Balinese traditional dance and badminton sport. Both have different feature characteristic and it is difficult to find a deterministic function to select the optimal features and optimal classifier structure.

ELM itself has two major issues need to be addressed to improve its accuracy [2],[4] :

1) The structure size of the number of hidden nodes. The optimal number is still unknown with trial-and-error.
2) Whether the computation complexity can be further reduced when given large number of training data and when large number of hidden nodes required.

Our contribution is how the effective learning method of ELM to deal with 3D human motion pose-based classification problem to give not only processing speed but also well-acceptance accuracy by using the efficient network structure.

II. RELATED WORKS

The taxonomy of Human actions distinguishes actions into action primitive (called pose), action and activity [9]. Pose is a set of features of body part location and can be interpreted as meaningful string of symbol to describe the activity. Different activities may have more than one similar poses anywhere determined by pose location and the frequency of pose occurrence.

Some papers explained how ELM algorithm used in human action recognition but not in pose-based recognition. Minhas [8] has introduced Incremental ELM in human action recognition based on snippets of shape of human actor by approximately changing intensity histograms to extract pyramid histograms of oriented gradient features. Venkatesh [10] has introduced Fully complex-valued ELM classifiers for human action recognition using the fully complex-valued hyperbolic secant as an activation function. Optical flow-based features extracted from the video sequences are utilized to recognize 10 different human actions. Nizami [5] has presented a multi-view gait recognition algorithm for identification at a distance used two well known and effective gait representations namely Motion Silhouette Image (MSI) and gait energy image (GEI). The features for MSI and GEI images are extracted using Independent Component Analysis (ICA). Extreme Learning Machine (ELM) classifier is then used for classification. ELM is focused as fast classifier and the performance is increased simply by tuning the network parameter using hidden nodes number, training methods or activation function improvement.

Some papers in multi classification study with non ELM classifier, however, emphasize the importance of feature selection to reduce the dimension of data. The objective is to reduce the complexity and computation time of non ELM which significantly different with ELM classifier. One of the feature selection method is known as bag of words or bag of visual words (computer vision) or bag of correlated poses [1] and various ensemble techniques like boosting and bagging.
TABLE I. ACCURACY PERCENTAGE FOR ALL CLASS IN BADMINTON MOTION.

<table>
<thead>
<tr>
<th>Pose Cluster</th>
<th>K-NN</th>
<th>FK-NN</th>
<th>C-NN</th>
<th>ELM1</th>
<th>ELM2</th>
<th>ELM3</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>97.96</td>
<td>98.994</td>
<td>89.145</td>
<td>60.111</td>
<td>64.203</td>
<td>68.045</td>
</tr>
<tr>
<td>15</td>
<td>88.125</td>
<td>98.327</td>
<td>89.506</td>
<td>53.896</td>
<td>54.497</td>
<td>95.831</td>
</tr>
<tr>
<td>25</td>
<td>79.777</td>
<td>98.293</td>
<td>88.829</td>
<td>35.244</td>
<td>44.073</td>
<td>97.797</td>
</tr>
</tbody>
</table>

III. METHODS

The observation data of human gestures are captured using a static mounted Kinect depth camera positioned in front of the performer at the rate of 30 fps with distance at 2-3 meters in closed area. It is working based on pattern grid reflection of near-to-infrared light emitter to the object [6]. Using openNI framework, kinect can generate skeleton information from human pose instantly in 3D coordinate \((x,y,z)\). Modifying Raptis [7] and Heryadi [12], skeleton joint features are represented by parameter of spherical 3D coordinate \(\theta\) as inclination and \(\varphi\) as azimuth. We used all joints except NECK (Small variance) with total 14 \(\theta\) and 14 \(\varphi\) multiplied by N frame sequences (Different for each motion). We used some basic motion from Badminton sport action and traditional dance (Balinese dance) to define the motion classes. The motion class for Badminton are Long Service, Short Service, Overhead Loop, Forehand Stroke, Backhand Stroke, Drop Shot, Underhand lob, and Smash (Total 8 class). The motion class for traditional dance are Agem kanan, Agem kiri, Piles, Ngeseh, Luk Nerudut, Malpal, Ngegoel, Mungkahlawang, and Nayog (Total 11 class). We used Kmeans algorithm to cluster 28x32N features data observation into a group of key poses. Each motion is represented as a sequence of cluster pose labels as string representation then converted using arithmetic series within range -1 to 1 because it is not a monotonic function. For validation, we used cross validation with Kfold N=5.

IV. EXPERIMENT AND EVALUATIONS

Comparison with another statistical classifier Nearest Neighbor family (K-NN, Fuzzy K-NN and CNN), ELM is the fast in processing time even with complex network structure to approach the same level of generalization accuracy (In Table 1, ELM3 with 61 inputs, 30x61 Hidden layer Nodes and concatenated 30X235 training rows has testing time 0.12s for 60 testing rows; K-NN required 4.1s, Fuzzy K-NN required 7.16s and CNN required 0.62s per testing row). However, this learning method by increasing the network structure to improve the performance not always reliable and giving the best solution all the time.

A. Increasing The Network Structure

In Figure 2, Badminton motion, the increase of hidden nodes number is not giving any contribution significantly to the accuracy (Less than 10% fluctuated) unless we add more training data with good result in big pose clusters only. But, in certain pose, adding more hidden nodes and training data, makes the accuracy fluctuated (unstable). Comparing with Figure 3, dance motion which is more regular, the increase of hidden node numbers not giving the best accuracy performance even it tends to be unstable with 20% variance except for certain pose cluster=20. Bigger pose clusters tend to have lower accuracy performance. It seems for both regularity and no regularity data, the structure network has a congestion limit. Increasing the hidden nodes number and training data, not even giving contribution significantly to the performance.

B. Selecting The Training Data

Another easy learning method is by training data selection to get only qualified training data into learning process. We select the training data with intra class variance \(\geq 0.25\) and with inter class p-values for testing the hypothesis of no correlation with value \(0.75 \leq p < 1\). The overall accuracy is decreased because certain class has no training data left at all, thus the accuracy is 0. In another class, no data filtered at all, thus the accuracy is still the same. For certain class in dance motion, the performance can be improved by reducing data similarity. But not adequate training data, will make the generalization performance is not well performed (Figure 4).
C. Feature Selection

This paper proposes to use the frequency of pose occurrences as features rather than time sequence of pose. Similar to bag of words, we called it bag of pose. It significantly reduces the dimension of training data because pose sequence in time domain has sequence location and the occurrence frequency. The frequency of pose occurrence has more distinction between the class, so the accuracy can be improved. Both badminton and dance motion in frequency domain showed better accuracy (15-20% increase) using single ELM classifier and has stable accuracy for big pose clusters (See Figure 5 and Figure 6). Processing time reduced because of matrix input size is equal with fixed pose cluster number, not dependent to frame length.

V. CONCLUSIONS AND DISCUSSIONS

The way to improve ELM classifier generalization accuracy performance can be obtained by selecting the feature of training data from more complex dimension size in time sequence to be more simple and distinguished in frequency of occurrence (bag of poses). Addressing accuracy performance on ELM, we can improve without adding network structure complexity by selecting the optimum feature of training data first. So, we can make sure the ELM performance well improved with just simple network structure. We evaluate its effectiveness in classifying motions with different degree of regularity (freedom). Even though statistically, the frequency feature selection might be better as the selected feature, but syntactic the pose location probably still has important meaning to be studied further in our future works.

REFERENCES