UnSkEm: Unobtrusive Skeletal-based Emotion Recognition for User Experience

Muhammad Asif Razzaq[1], Jaehun Bang[1], Sunmoo Svenna Kang[2], and Sungyoung Lee[1]
Department of Computer Science and Engineering,
Kyung Hee University, 446-701, Republic of Korea
Email: {[1]asif.razzaq, [1]jhb, [1]sylee}@oslab.khu.ac.kr, [2]etxkang@khu.ac.kr

***Abstract*—In this paper, the proposed framework utilizes body joint movement patterns extracted from skeletal joint features from Kinect v2 sensor in order to recognize emotions. Instead of using traditional methods for feature learning such as feature clustering, we proposed two methods *Mesh Distance Features* and *Mesh Angular features* to represent highly accurate body postures. For these methods, we only considered upper body joints which were 15 in number. Recognition of human emotion is performed using Support Vector Machine (SVM) which is train with Sequential Minimal Optimization (SMO). The contribution of this paper is two-fold. Firstly it uses a limited set of skeletal joints instead of tracking whole-body joint coordinates. Secondly, it uses the proposed methods of MAD and MAF for feature extraction. The proposed framework recognizes six emotions (Anger, Happiness, Sadness, Neutral, Surprise, and Fear) over the dataset collected for evaluating the User Experience platform. The experimental results show promising higher accuracies for emotional state recognition in real-time.**

***Index Terms*—Emotion recognition; skeletal joint data; SVM; SMO; Kinect v2;**

I. Introduction

Indoor Human action recognition (HAR) in a smart environment is a popular research trend as it supports various types of application domains such human-computer interaction, sports, gaming, personalized recommender systems, surveillance systems, and so on. Using different visual recognition tools and technique various advances has been made. However, Microsoft Kinect sensors offer great benefits to such techniques for monitoring healthcare status, occupant detection, or HAR. The skeleton-based features have proved to be much robust while determining human actions. The modeling of 3D human skeletal joint positions opens up new research directions in the field of human tracking and HAR [16].

The presented study is used for detecting human's emotional and mental states by detecting their pose. For this person's joint movements belonging to the upper part of the body are tracked. Detection of emotions is based on body parts movement but not on the basis of facial or auditory expressions. Emotion detection is basically measured through the head, shoulder, hands, etc. expressions. The psychological investigators find a strong relationship and prove that everybody can reflect different expressions [4], [15] while having different emotions.

Here, we study the different body poses and estimate emotions using body joints movements [7] while performing different tasks. The picture of the person's state of mind can be estimated by associating the results for hand gestures, arm motions, head turnings and shoulders movements together. Through the coordinated patterns, we can identify some bodily gestures like Joyful, Sadness, Anger, Neutral, Surprise, etc. The 3D geometric and kinematic features extracted from the raw body joints data are used to infer behavioral pattern for headshake, body forward, backward or sideways movement, arm retraction, hand movements represented as some emotion. A machine learning classifier is trained based on extracted features to recognize emotional states for any of the systems or applications discussed earlier.

In this paper, we focus on providing an accurate representation of emotions through human exposed to Kinect-based sensor device, by illustrating an efficient way to generate features for training Support Vector Machines (SVM). The evaluation results proved to be robust and efficient while achieving considerably higher accuracy and better overall performance.

The rest of the paper is structured as follows. Section II presents the related work for this study. Section III describes an overview and discusses our proposed approach. Section IV compares experimental evaluations. Finally, Section V describes conclusions and suggests future work.

II. Related Work

"Emotion is the mental experience with high intensity and high hedonic content (pleasure/displeasure)" [2], which deeply affects our daily behaviors by regulating an individual's motivation [9], social interaction [11] and cognitive processes. A person's well-known human expressions are interlinked with the body postures, which can represent emotions or affective states such as neutral, anger, happiness, fear, surprise or sadness. In visual perception, body motion patterns and skeleton information, in particular, are considered to be a reflection for emotional states or a person's inner feelings [6]. For instance, in an angry situation, people may have aggressive hands or head gestures. Recent advancement in Computer Vision (CV) based on Artificial Intelligence (AI) has effectively recognized emotion expressions through voices, faces, images, or other objects. However, recognition of body gestures or expressions has received little investigation.

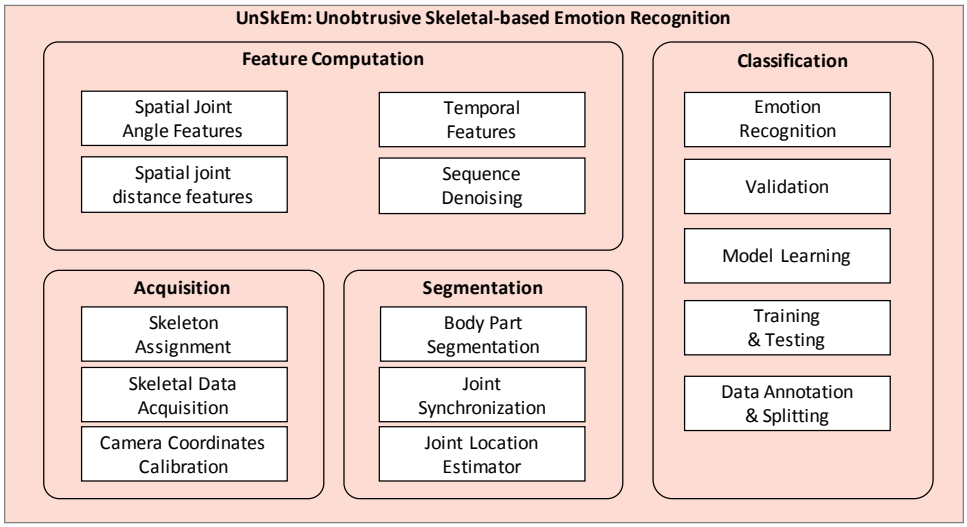

Fig. 1. Overall Architecture diagram for an Unobtrusive Skeletal-based Emotion Recognition (UnSkEm).

Body expressions in conjunction with joint motion patterns and postures have played an important role in formulating human emotions. For this, Coulson et al. [5] showed the positive influence of body postures, classified still image frames into human emotions through the systematized analysis. The analysis highlighted the importance of observer's head inclination and elbow's flexion during sadness and anger states respectively. They concluded that the movements of head, hands, shoulder, and elbows individual or jointly can represent body expression in the form of emotions. Manzi et al. [12] performed dynamic clustering over skeleton joint data by applying the x-mean ML algorithm for HAR. Li et al. [10] recognized the emotional states using Kinect sensors by analyzing human gaits. They evaluated a set of ML classifiers through the collected dataset and provided enough feasibility for automatically recognition of anger, happiness or neutral emotional states. Ahmed et al. [1] utilized different sets of features named "Joint Relative Angle", which were derived from the joint angles obtained by encoding the related motion patterns of skeletal body joints from the dataset in the Kinect skeletal gait database. The use of the skeletal joint, bone coordinates and an angle between them was studied by Garrido et al. [7]. Their study suggested that movement based interaction has ample effect to monitor patients based on joint coordinates and postures. Additionally, body movements and motion patterns have also been stated and compared in human-like robots [13] for identifying human emotions. These emotions were recognized based on joint rotations, arm, and head movements. Similarly, Cicirelli et al. [3] proposed a gesture recognition system using the Kinect sensor, which provided a real-time human skeleton. They provided a solution to resolve the spatial and temporal challenges of gestures using the Neural Network (NN).

We deployed a novel method for an unobtrusive [14] emotion recognition using Microsoft Kinect v2, portable, a low-cost, and camera-based sensor system. The sensor is able to detect human presence by identifying 25 body joints such as head, hands, etc., which are used to recognize the meanings of gestures in a pre-defined popular gesture scenario. The key contributions in this study include: methods for proposing feature extraction based on limited upper body skeletal joints; and accurate emotion detection from a limited feature set collected from selected set of joints.

III. Proposed Approach

We propose an Unobtrusive Skeletal-based Emotion Recognition (UnSkEm) framework for learning emotions from 3D skeleton data representing upper body motion patterns from the Kinect v2 sensor. We considered six emotions to be detected from the body joints movements and their orientation as mentioned below. The overall architectural view for the proposed framework is shown in Figure 1, which is further decomposed into 4 sub-modules such as skeletal joint *Acquisition*, skeletal frame *Segmentation*, *Feature Computation*, and emotion *Classification*. The following are the emotions recognized in this study along with their definitions in terms of body movements and pose consideration.

- Anger: User having clenched fists in front of the chest with head tilts, arms crossed across to chest.
- Happiness: arms opened with joy with raised head or both thumbs up.
- Sadness: Drooping shoulder or one palm on the head or lowering the head or head in the user's hands.
- Neutral: both open hands at rest on the computer desk sitting straight on the chair.
- Surprise: shrugging of shoulders with sudden backward movements or both palms towards a user with raised elbows.
- Fear: Crossed arms.

A. Skeletal Joint Acquisition

The kinect sensor device can track and collect information from the movements associated with the user's joints at the

rate of 30 frames per second(fps) for maximum 6 number of a user. The proposed architecture is capable to handle and process the skeleton joints information for multiple users. We developed Java-based API to implement the proposed UnSkEm framework to detect 3D skeletal joint data from Kinect for inferring different body positions of detected users. The 3D coordinates for 15 upper body joints were collected at the frame rate of 30fps. The selected joints are used to recognize users along with their actions. In our study, we considered the user's in a sitting position working in front of a computer desk for UX evaluations on electronic devices [8]. We developed body language dictionary for UnSkEm using 15 joints as mentioned in Figure 3.

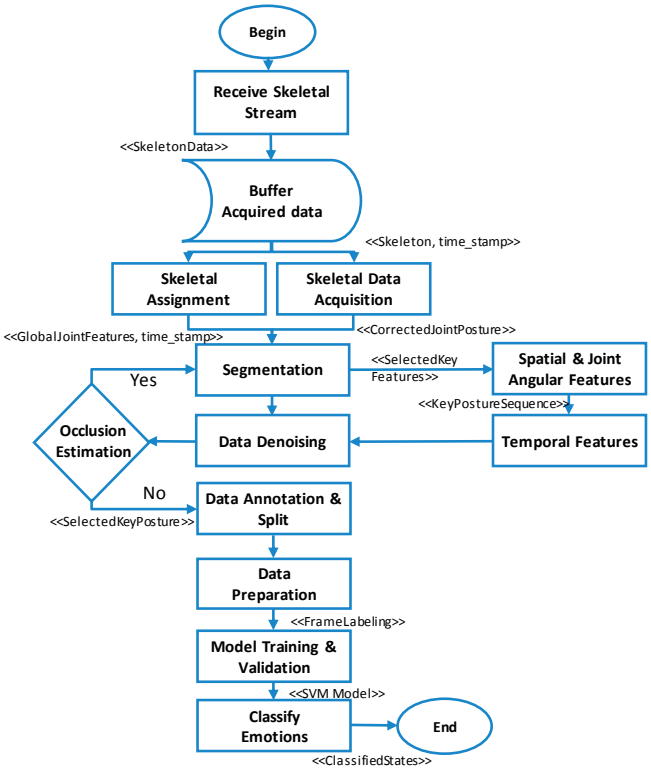

Fig. 2. Workflow illustrating deployment UnSkEm.

B. Skeletal Frame Segmentation

In this study, we considered a 3-sec sliding window to record arbitrary user actions for pre-defined emotions. During this time, Kinect captured 90 frames for the skeletal body joints. The set of joint's 3D coordinates used to calculate the inter-joint distance and angular feature calculations. The overall workflow details are discussed further 2. Since we collect 30 fps for skeletal joints and in a 3-sec window they become 90, the acquired joint coordinates may be repeated, not accurate, or missing. We ensured a unique stride in an individual segment by eliminating [10] the repeated frames that are most likely to be representing noise for connected body joints in a stream.

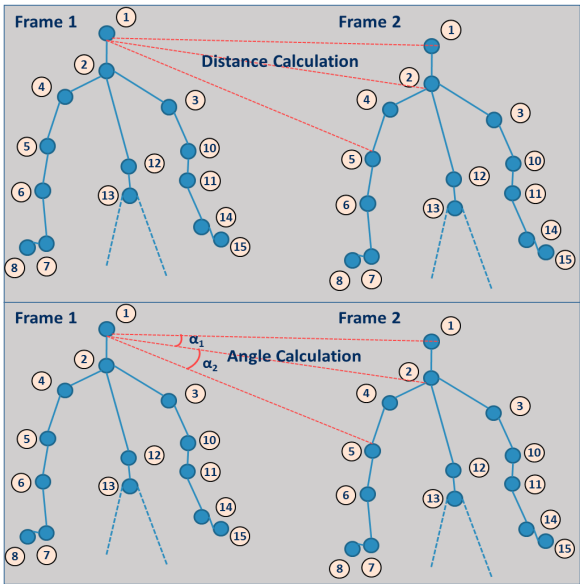

Fig. 3. Workflow illustrating *Mesh* Distance and Mesh Angular Feature Extraction.

C. Feature Computation

We introduce unique features extraction methods from the associated joints and their angles in between. We considered emotion recognition based on the seated subject so only the joint information relevant to the upper body such as head, neck, shoulders, elbow, wrists, and hands were considered for feature extraction. However, we discarded lower body joints information. Due to the higher rate of skeletal frames and the subject's natural body movements, it may involve overlapping body segments such as hands, arms movements in front of head, shoulders or spine joints leading to noisy frames. In order to handle noisy skeletal joint information, we propose *Mesh Distance Features* and *Mesh Angular Features*. The proposed feature extraction methods have proved to effective in terms of features correlations with resulting classification accuracies.

1) Mesh Distance Features: The joint distance-based features vary over the time significantly and it is difficult to extract consistent features. However, in order to measure accurate motion patterns, and record motion parameters using joint information, we introduce Mesh-based distance features (MDF). MDF provides rich connecting joint distance features while the subject is performing naturally. This also gave variance in the body parts segments lengths. We used Euclidean distance function for distance measurement amongst each referral joint with all other joints under consideration to record features and measure movements meticulously as shown in Figure 3a.

$$d(Jt_i, Jt_j) = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2} \quad (1)$$

where i is the referral joint and $j = \{1, 2, \ldots 15\}$.

2) Mesh Angular Features: As mentioned previously joint distance-based features are very dynamic, additionally angles

TABLE I
CONFUSION MATRIX USING TRAINED SMO (SVM) MODEL ON THE TEST DATA

Mean Classification Acccuracy: (**96.73%**) & Classification Error: (**3.27%**)							
Type of Emotions		Emotion Recognition Rate (%)					
		Anger	Happiness	Sadness	Neutral	Surprise	Fear
Ground Truth	Anger	96.45	0	2.5	0	0.76	0
	Happiness	0	99.7	0	0	1.1	0
	Sadness	2.4	0	97.2	0	6.1	0
	Neutral	0	0	0.8	100	0	0
	Surprise	0.60	0.3	0.3	0	92.04	5
	Fear	0.55	0	0	0	0	95

formed between these joints also provide important information for studying body motion patterns. We introduced *Mesh Angular Feature* (MAF) to measure the angle for each joint. The directional angular feature is computed for each joint connecting the neighboring joints as well as the other joints as shown in Figure 3b.

$$\theta_{i,j} = \cos^{-1}\left(\frac{\overrightarrow{u_i} \cdot \overrightarrow{v_j}}{||\overrightarrow{u_i}|| \cdot ||\overrightarrow{v_j}||}\right) \quad (2)$$

where i is the referral joint and $j = \{1, 2, \ldots 15\}$.

We collected MDF and MAF to study the motion patterns while the subject is having a different sets of emotions. The captured features are concatenated into larger feature vector showing how upper body joints behave in terms of distance and angles over time.

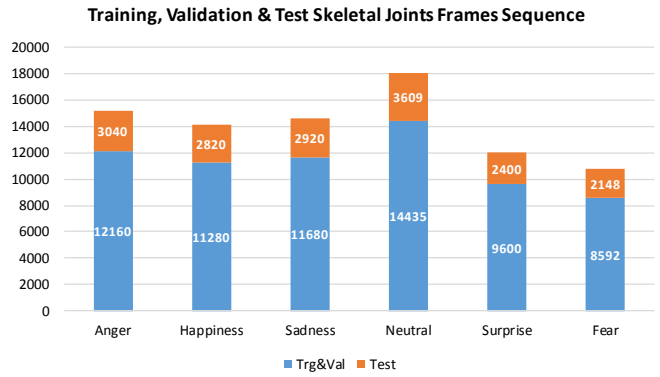

Fig. 4. Performance Measures: Training and evaluation emotion classification using UnSkEm platform

D. Skeletal Classification

In order to correctly classify six different emotions based on the skeletal joint sequences of similar actions, we used the Sequential Minimal Optimization (SMO) process for training multi-class Support Vector Machines (SVMs). SVMs works in a supervised learning fashion by providing models for binary classification, however, the multi-classification SVMs are implemented using several combinations of Binary SVMs. SVM models are trained using the skeletal joint data obtained for the evaluation of the User Experience platform. The skeletal joint data was split into two halves of training (80%) and test (20%) data. Features extracted from MDF and MAF are concatenated and grouped together by representing similar emotions in the data collected for the training phase. The trained SVM models were evaluated using 10-fold cross-validation.

IV. EXPERIMENTAL EVALUATIONS

The skeletal joint dataset collected, comprises of 6 subjects and 6 emotions to validate the proposed UnSkEm framework. Skeletal joint emotion samples include sequences of Anger, Happiness, Sadness, Neutral, Surprise and Fear using 15 upper body skeletal joints. The collected sequences are bifurcated into two halves, training, and test sets, which included skeletal joint frame sequences collected through Kinect v2 sensor. The Kinect is placed at the height 1.70m on a tripod stand above the ground level and at distance 1.5m away from subjects who were in a seated position in front of the computer desk to capture only upper body joints sequence. The extracted feature vector through MDF and MAF were annotated with the respective emotion class. The SMO based SVM models are trained using 80% of the skeletal joint frame sequences as mentioned in Figure 4.

Evaluation matrix for test split based on the trained model for 6 emotions from the upper body skeletal frame sequence are analyzed which are shown in Table I. The results show that Neutral emotion was detected 100% as the subject sits on the computer chair with both arms resting on the computer desk. Emotion *Surprise* (92.04%) was most often classified as *Sadness* (6.1%), *Fear* (95.0%) misclassified as *Surprise* and *Anger* (96.45%) as *Sadness*. This leads to a fact that while performing body motions for these emotions, subject body movements and patterns are of low intensity and highly indistinguishable.

The presented confusion matrix clearly depicts that the subject's emotional states can be described through the body motion patterns such as movement of head, shoulders, elbows, wrists, etc. These emotional states when recorded through the Kinect sensor, the obtained dataset could be used to recognize the subject's emotional states based on skeletal joint movements. For this, conventional machine learning or deep learning methods could be applied.

V. CONCLUSIONS AND FUTURE WORK

The presented study proposes a framework, which utilizes body joint movement patterns extracted from skeletal joint features from Kinect v2 sensor in order to express body movement-based emotions. Instead of using traditional

methods for feature learning from full-body movements, we proposed two methods *Mesh Distance Features* and *Mesh Angular features* to represent highly accurate body postures using upper body movement. Recognition of human emotion is performed using SMO based SVM trained models with 10-fold cross-validation. The experiments were performed to improve the emotion recognition accuracies using the dataset collected from 6 subjects for evaluating the User Experience Platform. The proposed framework recognizes six emotions (Anger, Happiness, Sadness, Neutral, Surprise, and Fear). The promising accuracies through UnSkEm framework make it feasible to use such an unobtrusive method for recognizing human emotion to other developed systems such as behavior trackers, user experience measurement, gaming evaluations in real-time.

ACKNOWLEDGMENT

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2017-0-01629) supervised by the IITP (Institute for Information & communications Technology Promotion)" and IITP-2017-0-00655 and NRF-2016K1A3A7A03951968 and NRF-2019R1A2C2090504.

REFERENCES

[1] Faisal Ahmed, Padma Polash Paul, and Marina L Gavrilova. Kinect-based gait recognition using sequences of the most relevant joint relative angles. 2015.

[2] Michel Cabanac. What is emotion? *Behavioural processes*, 60(2):69–83, 2002.

[3] Grazia Cicirelli, Carmela Attolico, Cataldo Guaragnella, and Tiziana D'Orazio. A kinect-based gesture recognition approach for a natural human robot interface. *International Journal of Advanced Robotic Systems*, 12(3):22, 2015.

[4] Ross A Clark, Yong-Hao Pua, Adam L Bryant, and Michael A Hunt. Validity of the microsoft kinect for providing lateral trunk lean feedback during gait retraining. *Gait & posture*, 38(4):1064–1066, 2013.

[5] Mark Coulson. Attributing emotion to static body postures: Recognition accuracy, confusions, and viewpoint dependence. *Journal of nonverbal behavior*, 28(2):117–139, 2004.

[6] Nourhan Elfaramawy, Pablo Barros, German I Parisi, and Stefan Wermter. Emotion recognition from body expressions with a neural network architecture. In *Proceedings of the 5th International Conference on Human Agent Interaction*, pages 143–149. ACM, 2017.

[7] Juan E Garrido, Victor MR Penichet, Maria Dolores Lozano, Alberto Mora Plata, and Jose AF Valls. The use of joint coordinates to monitor patients in a movement-based interaction system. *Universal Access in the Information Society*, 18(1):3–16, 2019.

[8] Jamil Hussain, Wajahat Ali Khan, Taeho Hur, Hafiz Syed Muhammad Bilal, Jaehun Bang, Anees Ul Hassan, Muhammad Afzal, and Sungyoung Lee. A multimodal deep log-based user experience (ux) platform for ux evaluation. *Sensors*, 18(5):1622, 2018.

[9] Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. Emotion, motivation, and anxiety: Brain mechanisms and psychophysiology. *Biological psychiatry*, 44(12):1248–1263, 1998.

[10] Shun Li, Liqing Cui, Changye Zhu, Baobin Li, Nan Zhao, and Tingshao Zhu. Emotion recognition using kinect motion capture data of human gaits. *PeerJ*, 4:e2364, 2016.

[11] Paulo N Lopes, Peter Salovey, Stéphane Côté, Michael Beers, and Richard E Petty. Emotion regulation abilities and the quality of social interaction. *Emotion*, 5(1):113, 2005.

[12] Alessandro Manzi, Paolo Dario, and Filippo Cavallo. A human activity recognition system based on dynamic clustering of skeleton data. *Sensors*, 17(5):1100, 2017.

[13] Derek McColl and Goldie Nejat. Recognizing emotional body language displayed by a human-like social robot. *International Journal of Social Robotics*, 6(2):261–280, 2014.

[14] Muhammad Asif Razzaq and Sungyoung Lee. Mmou-ar: Multimodal obtrusive and unobtrusive activity recognition through supervised ontology-based reasoning. In *International Conference on Ubiquitous Information Management and Communication*, pages 963–974. Springer, 2019.

[15] Muhammad Asif Razzaq, Claudia Villalonga, Sungyoung Lee, Usman Akhtar, Maqbool Ali, Eun-Soo Kim, Asad Khattak, Hyonwoo Seung, Taeho Hur, Jaehun Bang, et al. mlcaf: multi-level cross-domain semantic context fusioning for behavior identification. *Sensors*, 17(10):2433, 2017.

[16] Lei Wang, Du Q Huynh, and Piotr Koniusz. A comparative review of recent kinect-based action recognition algorithms. *arXiv preprint arXiv:1906.09955*, 2019.