Applied Intelligence (2022) 52:16605–16620
https://doi.org/10.1007/s10489-022-03181-4

Deep multi-model fusion network based real object tactile understanding from haptic data

Joolekha Bibi Joolee[1] · **Md Azher Uddin**[2] · **Seokhee Jeon**[1]

Accepted: 2 January 2022 / Published online: 24 March 2022

Abstract

The tactile information of an object is one of the crucial features which define the impression of that object. This paper presents a novel multi-model fusion network for real object's tactile understanding from haptic data. Furthermore, a low-cost 3D printed artificial finger-based tactile sensing system is designed for capturing haptic information in the form of acceleration profile, angular velocity, and normal force. Our proposed multi-model fusion network includes three different networks. First, we introduce a novel ensemble 2D convolutional neural network, namely SpectroNet, which captures the spatial features from the spectrogram of acceleration profile. Second, we design a 1-D convolutional neural network (CNN) with residual connection for extracting detailed spatial information from each segment of collected data. Third, we design bi-directional gated recurrent unit networks (BiGRU) to capture temporal dynamics. Moreover, the attention mechanism is utilized in all three proposed networks to assign weights to the features according to their contributions, which enhance the performance further. Finally, extensive experimental analysis is conducted on our dataset (i.e., 60 real objects, which cover both planner and non-planner surfaces) as well as the TUM surface material database. Empirical evaluations demonstrate that the proposed method significantly outperformed state-of-the-art methods in terms of accuracy, precision, recall and F1-score. Furthermore, we also found that the proposed multi-model fusion network substantially improves the performance compared to the single network.

Keywords Tactile understanding · Multi-model fusion network · SpectroNet · 1-D convolutional neural network · Bi-directional gated recurrent unit networks

1 Introduction

Recently, tactile understanding has become very popular due to its wide range of applications in the manufacturing industry, robotics, and so on. Object recognition using tactile information is a powerful human capability. Based on the sense of touch, human beings can easily adjust their movements throughout object handling [1, 2]. However, a robot may find it very difficult to handle objects without haptic sensing. For instance, how to grip a fragile object, avoiding wet or slippery objects, and how to express the tactile properties of an unknown object to humans.

Object understanding and recognition have been largely investigated in the area of computer vision and huge advancements have been made. In [3], 2D and 3D features were computed from images to perform the object and material identification. Recently, deep learning-based methods have also been adopted for material understanding from images in [4–6]. However, object and material understanding based on visual information require vision sensors and the camera field of view, which may be restricted by lighting constraints and occlusion issues [7]. Haptic (i.e., proprioception and touch) information can help to overcome the limitation related to visual-based recognition. Besides, haptics can also provide further

✉ Seokhee Jeon
jeon@khu.ac.kr

Joolekha Bibi Joolee
julekhajulie@gmail.com

Md Azher Uddin
azher006@yahoo.com

[1] Department of Computer Science and Engineering, Kyung Hee University, 446-701, Seocheon-dong, Giheung-gu, Yongin-si, Gyeonggi-do, Republic of Korea

[2] Department of Mathematical and Computer Sciences, Heriot-Watt University Dubai, Dubai, United Arab Emirates

information about the object, such as texture, shape, and size, which can surely enhance the understanding.

Numerous recent works have focused on recognizing the surface materials and objects based on haptic information. SynTouch based BioTac sensors were employed for tactile object recognition in [1, 2, 7–11]. These works only record the pressure, normal force, thermistor data and overlook the acceleration signals. In their works, vibrations are measured from dynamic pressures. However, SynTouch based BioTac sensors require higher computation power to record the tactile data [12]. Furthermore, they are also costly [27]. On the other hand, pen-type tool-tips were utilized for analyzing the surface materials in [13–17]. However, these works do not consider the diversity of real-world objects, for example, vibrations due to varying bumpiness and roughness.

Deep learning based approaches have become quite common for signal analysis and recognition [1, 13, 30]. Moreover, recently, several hybrid deep models have been introduced to improve the performance further. For instance, 1D ensemble networks were proposed for real-time user recognition in [18]. Heo et al. [19] designed a multimodal and ensemble-based deep learning (ME-DeepL) framework using 1D-CNN and gated recurrent units (GRU) for forecasting wastewater influent loads. A CNN-LSTM-based hybrid deep learning model was employed to automatically detect myocardial infarction from ECG data in [20]. Huang et al. [21] introduced a deep ensemble capsule network for fault diagnosis. Later on, in-vehicle network anomaly detection is achieved using the Conv1D and Bi-LSTM based ensemble model, namely CLAM (CNN-LSTM with Attention model) [22].

Motivated by the aforementioned works, in this paper, we propose a novel hybrid deep learning framework for real-world object tactile understanding from haptic data. Furthermore, we design a low-cost 3D printed artificial finger-based tactile sensing system for capturing haptic information from objects. Since we aim to develop a tactile sensing system for robot fingers, our sensor is designed in finger shape. The proposed tactile sensor records the acceleration signals, angular velocity, and force data at the point when the user moves the tactile sensor over an object. Afterwards, the proposed deep multi-model fusion network captures the spatial and temporal features as follows. First, we introduce a novel ensemble network, SpectroNet, which captures the multi-scale spatial features from the spectrogram images of acceleration profiles. Second, we design a residual 1-D convolutional neural network (CNN) to extract detailed spatial information from acceleration, angular velocity, and force data. Third, we develop a bi-directional gated recurrent unit (BiGRU) network for capturing the temporal dynamics. Furthermore, the attention mechanisms are adopted in all three networks to assign weights to the features according to their contributions, which boosts the performance further. Later on, joint tuning layers and softmax function are applied to combine the obtained features and classify them. Finally, to demonstrate the effectiveness of the proposed approach, an extensive experimental analysis is conducted on our dataset (i.e., have 60 real objects, which cover both planner and non-planner surfaces) as well as TUM surface material database (i.e., public dataset) [13].

The major contributions of this work are summarized as follows.

- We propose a novel deep multi-model fusion network for real-world object tactile understanding from haptic data.
- A cost-effective 3D printed artificial finger-based tactile sensing system is designed for capturing haptic information from objects.
- A new attention-aware ensemble network, namely SpectroNet, is developed for capturing the multi-level spatial information from the spectrogram of acceleration profile.
- An attention-aware residual 1-D convolutional neural network is introduced to obtain detailed spatial information from the acceleration, angular velocity, and force data.
- A three-layered bi-directional gated recurrent unit (BiGRU) network is designed to capture the temporal dynamics from the collected haptic data. This network also includes an attention mechanism that assigns weights to the features based on their contributions.

This paper is ordered as follows. First, we review the related works in Section 2. Section 3 explains the proposed deep learning based framework for real object's tactile understanding. Datasets and experimental results are demonstrated in Section 4. Finally, the conclusion of our work is presented in Section 5.

2 Related work

Numerous works have been proposed for tactile understanding based on biomimetic sensors, haptic stylus, and tactile-enabled fingertip. In this section, we will discuss the relevant studies that are closely related to our work.

2.1 Biomimetic sensors based approaches

SynTouch BioTac sensors [12] have become very popular for collecting tactile information from objects. For instance, Chu et al. [8] utilized the SynTouch BioTac sensors to identify the meaning of haptic adjectives. In their work, a support vector machine was employed as a classifier. Later

on, Gao et al. [1] introduced convolutional neural networks (CNN) for object recognition based on tactile information. They also explored Long Short Term Memory (LSTM) network for classification. In addition to haptic data, visual information is also considered in [1]. A dictionary learning-based approach and an extreme kernel sparse learning approach were presented in [2, 9] for haptic-based object recognition, where Penn Haptic Adjective Corpus 2 (PHAC-2) dataset [8] was employed to evaluate the performance. Later on, Liu et al. [10] discussed the learning problem of identifying untouched tactile objects. Abderrahmane et al. [11] extended the haptic Zero-Shot learning approach, which was initially proposed in [7] for classifying the objects. Hand-crafted feature extractors [7] is replaced with CNN for attribute learning in [11]. Moreover, they considered both tactile and visual information. All the aforementioned works considered the SynTouch BioTac sensors for capturing the tactile signals, which represent the pressure, normal force, and thermistor data, and ignore the acceleration signal. In SynTouch BioTac sensor vibrations are computed from dynamic pressure. However, SynTouch based BioTac sensors require higher computation power to record the tactile information [12].

2.2 Haptic stylus based approaches

A stainless steel tool-tip based haptic stylus is commonly used for haptic texture modeling and rendering task [23, 24]. Similarly, several works utilized tool-tip based haptic stylus for capturing tactile information and surface material classification. For example, Zheng et al. [13] equipped a three-axes accelerometer with a tool-tip based haptic stylus for collecting the tactile information, where CNN is employed for surface texture classification. Later on, Strese et al. [14] used the same haptic stylus along with the sound and force information, however, they employed handcrafted features and a Naive Bayes classifier for surface texture analysis. Content-based surface material retrieval was presented in [15]. Afterward, a dictionary learning approach was introduced in [16] to classify the surface material. Tsuji et al. [17] designed a pen-type tactile sensor and a 2D convolutional neural network for capturing and understanding the tactile information. Hand-crafted features were extracted for surface material classification in [17]. However, these existing works only analyzed the surface materials and overlook the diversity of real-world objects.

2.3 Fingertip based approaches

Recent few works considered fingertip-based tactile understanding and tactile sensation assessing [25–27]. In [25], a biomimetic fingertip was designed utilizing polydimethylsiloxane elastomer to capture the tactile signals and k-nearest neighbors (kNN) classifier was employed for tactile analysis. In contrast, Tanaka et al. [26] designed an artificial finger for assessing tactile sensations, where they wrapped the skin vibration sensor with the artificial finger. Later on, in order to address the issues of SynTouch based BioTac sensors, Strese et al. [27] introduced a framework for collecting the tactile information and they used hand-crafted features for classification. However, a deep learning-based approach can further improve the performance of tactile understanding.

3 Proposed framework

In this paper, we introduce a novel deep learning-based framework for real object's tactile understanding from haptic data. Besides, we design a low-cost 3D printed artificial finger-based tactile sensing system for capturing haptic information from objects. Our proposed tactile sensor records the acceleration signals, angular velocity, and force data at the point when the user moves the tactile sensor over an object. The proposed deep multi-model fusion network includes three different networks. First, SpectroNet, which captures the spatial features from the spectrogram of acceleration profile. Next, we design a 1-D convolutional neural network (CNN) with residual connection for extracting the detailed spatial information from each segment of the collected data. Third, we develop a bi-directional gated recurrent unit (BiGRU) network for capturing the temporal dynamics. Afterward, we introduce joint tuning layers and softmax function to combine the captured features and classify them respectively. In the joint tuning layers, we constructed two fully connected layers with varying numbers of hidden nodes (i.e., 750 and 500 nodes, respectively), which concatenates the features of the SpectroNet, residual 1-D CNN, and BiGRU. Figure 1 presents the proposed framework for real object's tactile understanding from haptic data.

3.1 Tactile sensing system and data collection process

In this work, we design a 3D printed artificial finger-based tactile sensing system for capturing real object's tactile information. The embedded system includes a force sensitive resistor (FSR) (FSR400, Interlink) and an MPU-9250 sensor, which are equipped with the 3D printed artificial finger, as demonstrated in Fig. 2(a). The FSR captures the laterally applied normal force while the MPU-9250 sensor [29] records the acceleration signal and angular velocity in the three-axes. A USB data acquisition card (NI USB-6002), connected to a computer, MPU-9250 sensor, and pressure sensor records the tactile signals. Vibrations

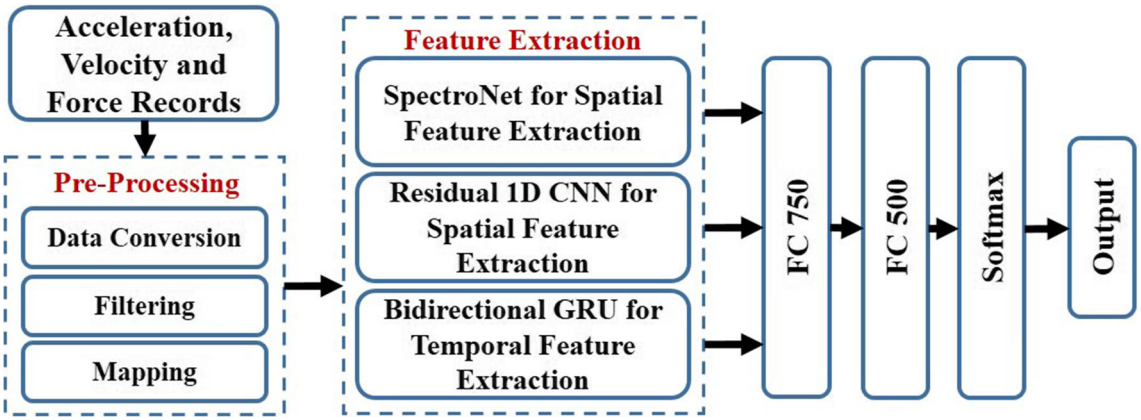

Fig. 1 Proposed framework for real object's tactile understanding from haptic data

between the object and the proposed tactile sensing system are recorded in the form of the acceleration signal.

We recorded the tactile information by employing the proposed tactile sensing system when the user moves the tactile sensor along the object. A total of 20 participants (15 males and 5 females) took part in this data collection process. Their mean age was 28.5 years (range from 24 to 35 years). They were informed of the experimental procedure beforehand. Each user collects the tactile signal for each object. Figure 2(b) shows the data collection process. We have collected 3s long tactile signals for each object. The sampling frequency of the recorded acceleration signal, angular velocity, and force data is 2 kHz. Afterward, the acceleration and angular velocity signals are band-limited within 10 Hz and 1,000 Hz to eliminate the effects of human hand motion and sensor noise. Furthermore, the DFT321 algorithm [23] is employed to map the three-axes signals into a single axis (see Figs. 3 and 4). In contrast, similar to [27], we convert the raw FSR sensor values to force data. After collecting and preprocessing the data, we have segmented the signals (i.e., acceleration, angular velocity and force) to extract the local spatial information and temporal dynamics by employing proposed residual 1-D convolutional neural network and Bi-directional gated recurrent unit network, respectively. In contrast, full acceleration profile of each trial is used as input to produce the spectrogram image using short-time Fourier transform [28] (see Fig. 5), which is then fed into SpectroNet.

3.2 SpectroNet

Convolutional Neural Networks (CNN) based approach is one of the most commonly studied deep learning methods in different domains of computer vision, and it showed outstanding performance. Therefore, in our work, we adopt CNN for extracting the spatial features from the spectrogram of acceleration profile. We design a novel ensemble 2D convolutional neural networks, namely SpectroNet, which is composed of three CNNs with different filter sizes. The architecture of the proposed SpectroNet is demonstrated in Fig. 6. Each CNN model includes three convolutional layers, three max-pooling layers, and an attention layer. The convolutional layers are responsible for capturing the features, while the max-pooling layers perform the dimensionality reduction of the

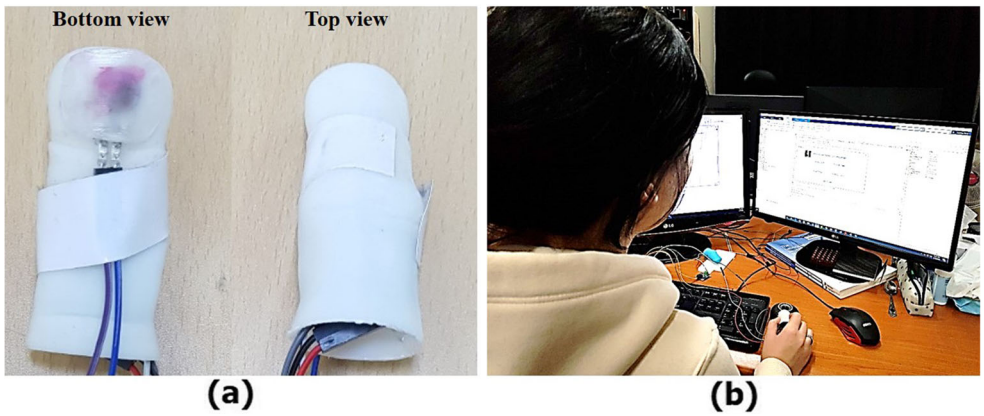

Fig. 2 (a) Artificial finger-based tactile sensing system, and (b) Data collection process

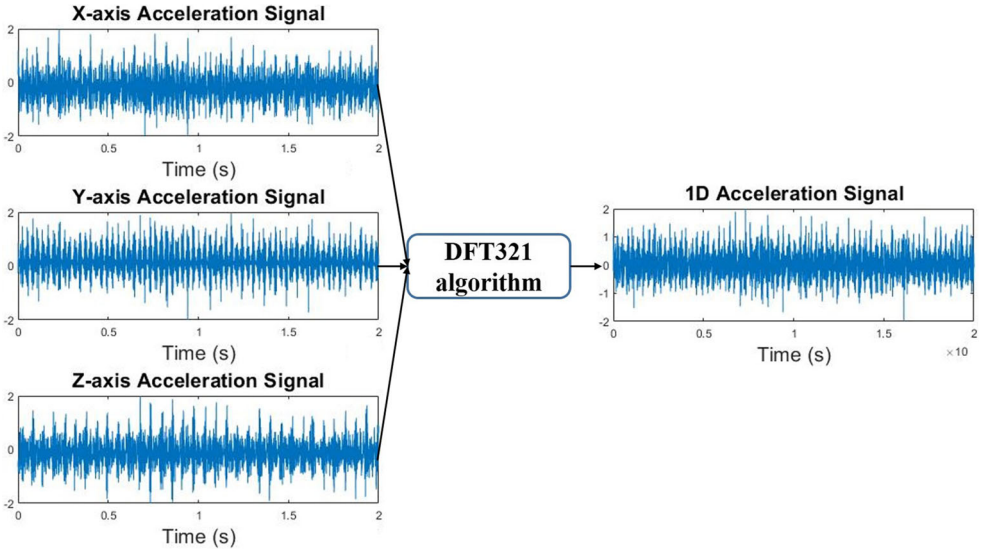

Fig. 3 Three-axes to 1D acceleration signal conversion using DFT321 algorithm

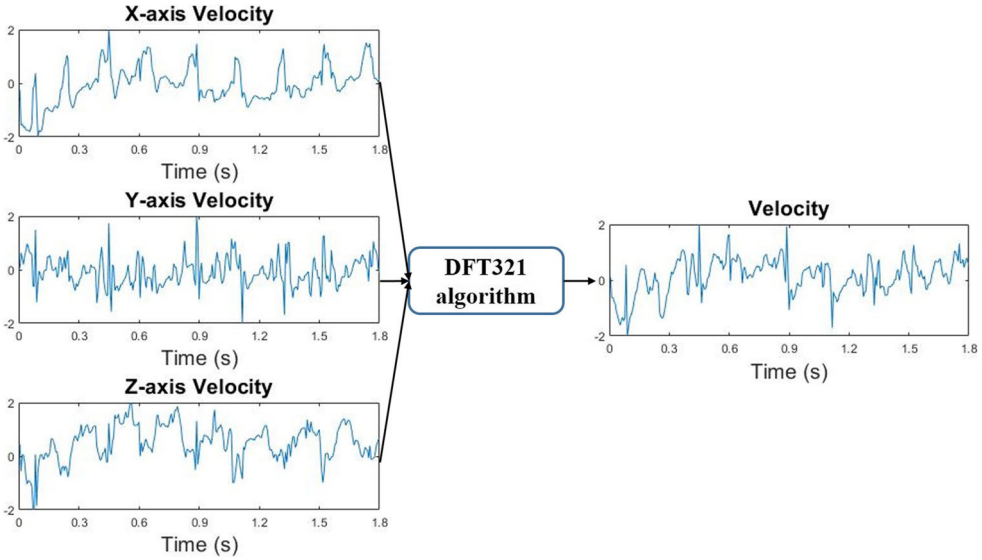

Fig. 4 Three-axes to 1D velocity conversion using DFT321 algorithm

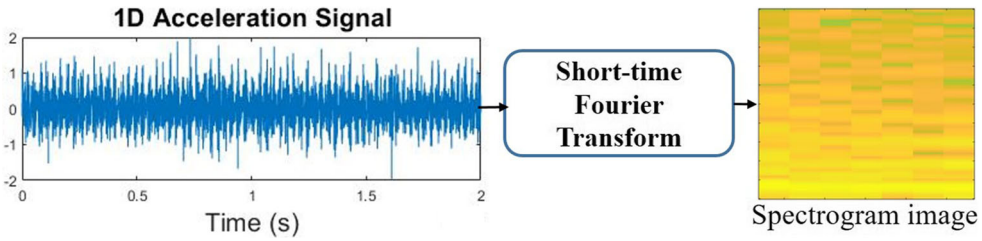

Fig. 5 Generation of spectrogram image from 1D acceleration signal using short-time Fourier transform

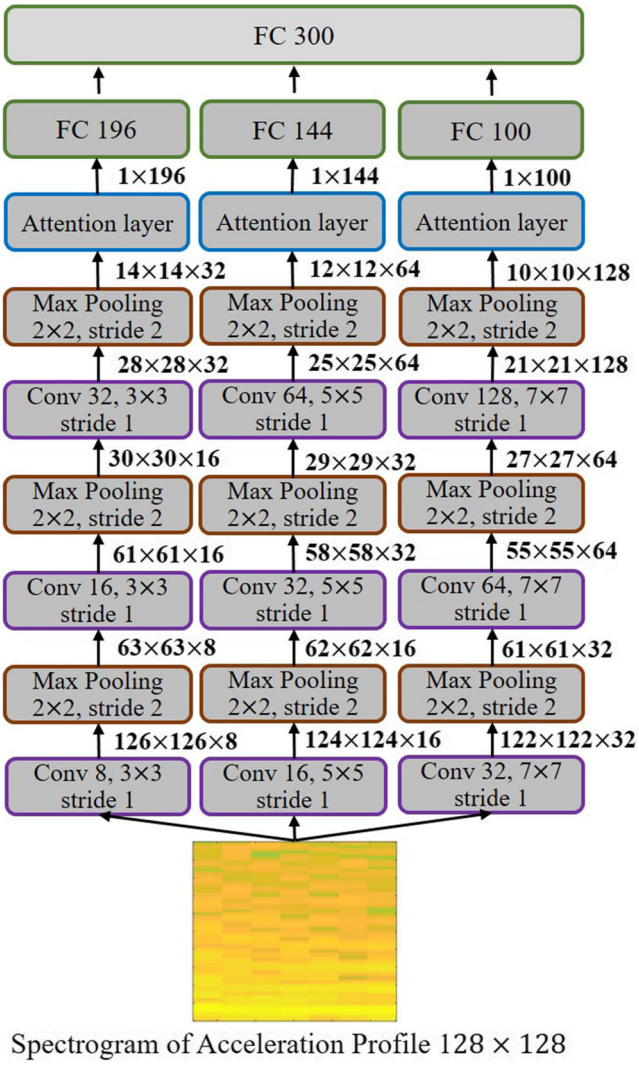

Fig. 6 Architecture of the proposed SpectroNet

individual feature map. Moreover, by applying different sizes of filters, different scales of local features can be efficiently obtained. In the convolution operation, we employed 3 × 3, 5 × 5, and 7 × 7 sizes of filters, whereas the max-pooling operation is done on 2 × 2 blocks. Besides, in the convolution operation, we utilize multiple filters to extract diverse aspects of each scale of local information. Afterward, an attention mechanism is adopted in each CNN model to improve the performance and reduce overfitting issues. The number of parameters in the fully connected (FC) layer can be huge and easily overfitted if all the feature maps are directly merged. Additionally, different feature maps have different contributions. The attention layer includes an FC layer along with the softmax function. Moreover, in the attention layer, the element-wise multiplication is executed between the vector before activation in FC and the vector after activation. The computation can be represented as,

$$h_n = \sum w_{ni} x_i + b_i \tag{1}$$

$$y_n = \phi(h_n) x_n = \frac{exp(h_n)}{\sum_n exp(h_n)} x_n \tag{2}$$

Where x is the input vector, y_n is the output vector, w_{ni} is the weight matrices, b_i is the bias, h_n is the vector of n neurons in FC layer, and $\phi(h_n)$ represents the attention weighted vector. The features produced from the attention layer are fed into the fully connected layer. Finally, we design a joint tuning FC layer that combines the features from three CNNs respectively.

3.3 Residual 1D convolutional neural network (CNN)

CNN's are commonly investigated to capture the features from the images. However, these days CNNs are also become popular for 1D signal analysis and recognition [1, 13, 30]. In this work, to capture the spatial features from each segmented signal, we proposed a novel deep residual 1D convolutional neural network (CNN). The architecture of the proposed deep residual 1D CNN is presented in Fig. 7. Our proposed 1D CNN with residual connection is consists of five 1D convolutional layers, five max-pooling layers, one attention layer, and two fully connected layers.

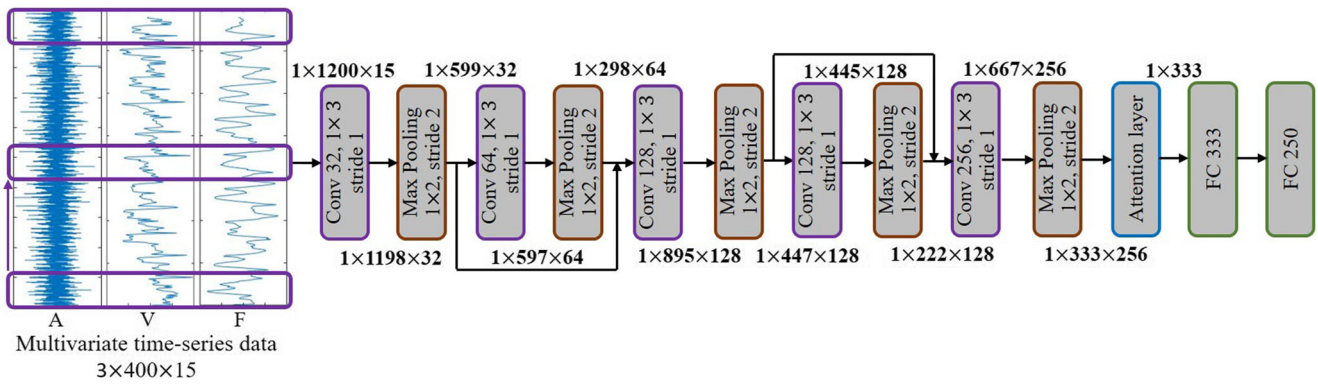

Fig. 7 Architecture of the proposed residual 1D convolutional neural network

Similar to the 2D CNNs, the convolutional layers extract the appearance information, while the max-pooling layers reduce the dimensionality of the individual feature map. The first convolution layer employs 32 filters, the second convolution layer employs 64 filters, the third and fourth apply 128 filters, while the fifth convolution layer employs 256 filters. All the filters have the same kernel size 1×3. The max-pooling operation is done over a window size 1×2. Furthermore, each convolution process is followed by Rectified Linear Unit (ReLU), which is a nonlinearity function. In the residual connection, firstly, the output of the first max-pooling layer is concatenated with the output of the second max-pooling layer. Secondly, the output of the third max-pooling layer is concatenated with the output of the fourth max-pooling layer. Afterward, in order to assign different weights to the different feature maps according to their contributions, an attention layer is employed. The attention layer computation procedure is based on the similar principle as in [31], which can be represented as follows,

$$x_i = \sum_{j=1}^{m} \alpha_j s_{ij} \tag{3}$$

$$\alpha_j = \frac{exp\ a(s_{ij}, \varphi)}{\sum_m exp\ a(s_{ij}, \varphi)} \tag{4}$$

Where, x_i is the attention vector, α_j represents the attention weight, s_{ij} is the feature map, and $a(s_{ij}, \varphi)$ indicates the auxiliary neural network with one-layer, m nodes and parameters φ. Finally, the 1D CNN network concludes with two fully connected layers having 333 and 250 neurons, respectively. Each fully connected layer includes a dropout of 0.5 to counter the overfitting issues. The proposed deep residual 1D CNN produces a 1×250 size feature vector.

3.4 Bi-directional gated recurrent unit (BiGRU) network

Existing statistical approaches such as ARIMA (Autoregressive integrated moving average) can handle the time series data, however, the performance is not good enough like deep neural networks. Because they do not take long-term temporal dependence into account. In contrast, the recurrent neural network (RNN) can effectively process the time-series data via hidden states. However, if the inputs are long sequences, then RNN based approaches can experience gradients vanishing and gradient explosion issues. To resolve this issue, the Long- Short Term Memory (LSTM) was proposed to learn long sequences through input, forget and output gates [32]. Nevertheless, the LSTM model requires a large number of training samples. In contrast, GRUs (gated recurrent units) train faster and perform better than LSTMs on less training data. The GRU is a variant of an LSTM, which can learn both short-term and long-term sequences.

In our work, Bi-directional gated recurrent unit (BiGRU) network with an attention mechanism is employed to further strengthen the performance of the proposed framework. The proposed BiGRU model includes three bidirectional GRU layers with forward and backward GRU, an attention layer, and two fully connected layers, as demonstrated in Fig. 8. GRU improves the three-gate structure of LSTM by excluding the cell state and forget gate. GRU includes two gates: update gate z_t and reset gate r_t. The update gate z_t controls the information to be kept in the current state, while the reset gate r_t assures the important sequences to be

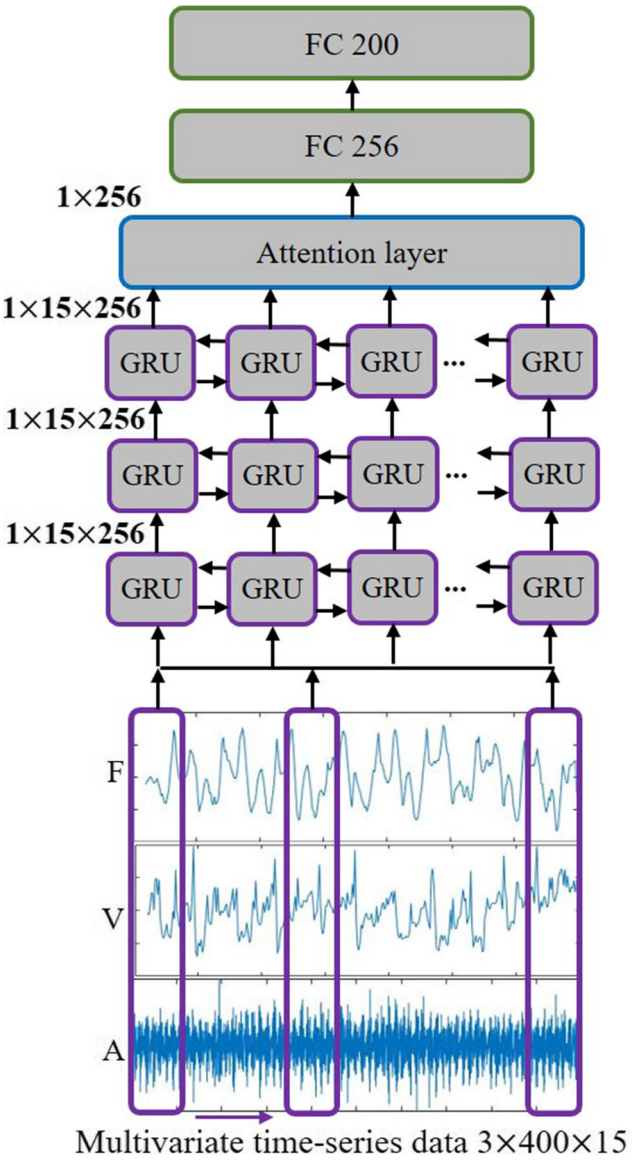

Fig. 8 Architecture of the proposed BiGRU network

carried to the subsequent step. The computation of GRU can be represented as follows.

$$z_t = \sigma(w_z[h_{t-1}, x_t]) \tag{5}$$

$$r_t = \sigma(w_r[h_{t-1}, x_t]) \tag{6}$$

$$\tilde{h_t} = tanh(w[r_t h_{t-1}, x_t]) \tag{7}$$

$$h_t = (1 - z_t)h_{t-1} + z_t\tilde{h_t} \tag{8}$$

Where, w_z, w_r, and w are the weight matrices. h_{t-1} represents the output of the previous hidden state, $\tilde{h_t}$ denotes the intermediate memory, σ and $tanh$ denote the sigmoid and hyperbolic tangent function, respectively. x_t is the input sequence at time t and h_t indicates the hidden state at time step t. BiGRU is designed by including a forward GRU layer and a backward GRU layer. The backward GRU capture the backward temporal features by taking the input sequence x in reverse. Subsequently, hidden states from forward $(\overrightarrow{h_t})$ and backward $(\overleftarrow{h_t})$ direction at time t are combined as follows and form the final output H_t.

$$H_t = con[\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{9}$$

After the final time step, we compute the final feature vector of BiGRU layer-1, $H = (H_1, H_2, \ldots H_t)$. In order to obtain further deeper information, the output of BiGRU layer-1 is fed into BiGRU layer-2 and the output of BiGRU layer-2 is fed into BiGRU layer-3. The BiGRU layer-2 and BiGRU layer-3 have the same composition as BiGRU layer-1. The attention layer takes the output of BiGRU layer-3 as input. The weighting vector $\sigma = (\sigma_1, \sigma_2, \ldots \sigma_t)$ is computed from the output sequence $H = (H_1, H_2, \ldots H_t)$ of BiGRU layer-3, while the attention vector a is calculated as a weighted sum of these hidden states as follows,

$$a = \sum_{t=1}^{m} \sigma_t H_t \tag{10}$$

$$\sigma_t = \frac{exp({o_t}^T o_w)}{\sum_t exp\ ({o_t}^T o_w)} \tag{11}$$

$$o_t = tanh(W_w H_t + b_w) \tag{12}$$

Where o_w and W_w are the weight matrices, and b_w is the bias. The representation o_t is calculated by (12). Finally, the output of the attention layer are fed into the fully connected layer, which is responsible for producing the final temporal feature vector.

4 Experiments

In this section, we assess the performance of the proposed framework on our dataset and TUM surface material database (publicly available) [13]. Here at first, we describe the datasets and the experimental settings. Later, experimental findings are reported along with an ablation study. Finally, we compare state-of-the-art approaches with the proposed deep multi-model fusion network.

4.1 Dataset

Our dataset contains the three-axes acceleration signals, three-axes angular velocities, and force data from 60 real-world objects. The objects have either planner or non-planner surfaces, which are demonstrated in Fig. 9. There are 40 non-planner objects, while 20 are planner surfaces. A total of 20 users move the proposed tactile sensor over each

Fig. 9 Overview of all 60 objects used in the experiments

Fig. 10 Overview of all 69 surface objects included in the TUM haptic texture database. Note that 1-7 are meshes, 8-16 are stones, 17-22 are glossy surfaces, 23-31 are wooden surfaces, 32-35 are rubber type surfaces, 36-43 are fibers, 44-53 are foams, 54-60 are foils and papers, 61-69 are textiles

object with different velocities and forces for three seconds. Therefore, we obtained data for 1200 trials.

TUM haptic texture database [13] includes the three-axes acceleration signals for 69 planner surface objects (see Fig. 10). The sampling frequency of the recorded acceleration data is 10 kHz, and they recorded data for 25 seconds. For each object, acceleration data are collected ten times. Therefore, the TUM haptic texture database has 690 trials data. All the three axes acceleration signals were combined to one using DFT321 [23]. The acceleration data is downsampled to 2 kHz in our work, and we used 3 seconds recordings for each trial.

4.2 Experimental settings

In our work, we used a five-fold cross-validation procedure to evaluate the performance of the proposed approach [33]. During each cross-validation, 80% data is used for training, and 20% data is used for testing [33]. Hyperparameters are a fundamental part of model training because they straightforwardly control the network performance. In cross-validation, after trying different hyperparameters and performing rigorous model training, we chose the best hyperparameters. As hyperparameters, we employed Adam optimizer to train the model with a learning rate of 0.001, momentum of 0.9, decay of 0.0005 and batch size of 32. The number of epochs is set to 600. Besides, Dropout [34] regularization is applied to enhance the generalization capability and reduce the overfitting issues. We collected and pre-processed the data using Matlab 2018b, while Tensor Flow and Keras framework are used for the deep learning part. In the experiment, we applied four different types of performance evaluation metrics, i.e., Accuracy, Precision, Recall and F1 score. The values of each performance index are calculated from the confusion matrix. The confusion matrix consists of four-test metrics, i.e., TP (true positive), TN (true negative), FP (false positive), and FN (false negatives) [35]. Accuracy, Precision, Recall and F1 Score are calculated by using the following equations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

$$Precision = \frac{TP}{TP + FP} \tag{14}$$

$$Recall = \frac{TP}{TP + FN} \tag{15}$$

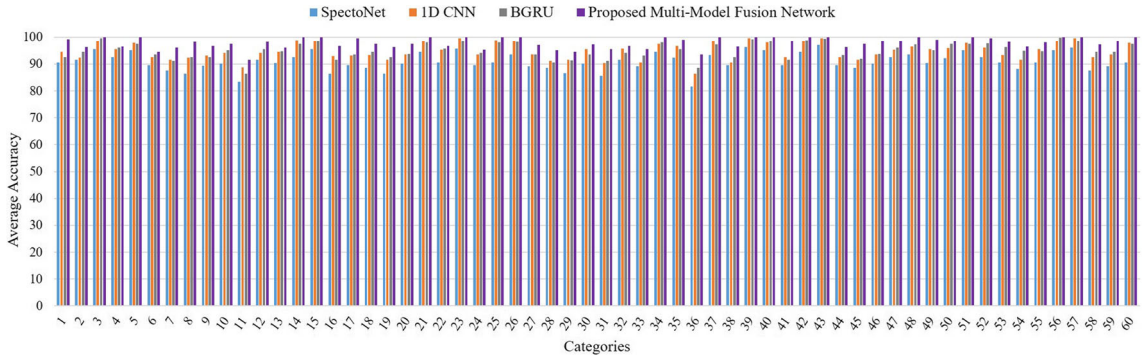

Fig. 11 Accuracy comparison for each category by employing SpectoNet, residual 1D CNN, BiGRU and multi-model fusion network on our dataset

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{16}$$

4.3 Overall performance of the proposed multi-model fusion network

We evaluated the performance of each proposed component on our dataset. The average accuracy on our dataset were 90.9%, 94.8%, 95.1%, and 97.9% for SpectoNet, residual 1D CNN, BiGRU, and multi-model fusion network, respectively. For the proposed multi-model fusion network, ceramic bowl, steel spoon, Fan, pencil bag, tissue box, concrete, leather, and Carpet2 categories show the best accuracy, whereas paper cup, notebook, lotion bottle, and helmet categories exhibit lower performance. This is because they have less intra-class variance and represent similar kinds of tactile information. Soap case, concreate, cloth cover categories show the best accuracy for the residual 1D CNN model. In contrast, paper plate, concreate, and Carpet2 objects demonstrate the higher performance for the BiGRU network. Figure 11 presents the accuracy comparison for each category on our dataset.

We also estimated the accuracy of the proposed fusion network as well as each proposed component on the TUM haptic texture database. The average accuracy on the TUM haptic texture database were 92.02%, 94.93%, 94.05%, and 98.18% for SpectoNet, residual 1D CNN, BiGRU, and multi-model fusion network, respectively. Figure 12 demonstrates the accuracy comparison for each category on the TUM haptic texture database. Our approach showed lower performance for the wooden surfaces and textiles compared to the other categories. However, overall accuracy is reasonably competent.

4.4 Ablation study

In this section, we investigate the effectiveness of different components from our multi-model fusion network. At first, we measure the performance (i.e., accuracy, precision, recall and F1 score) of the joint tuning layer, attention mechanism, and different sizes of the filters in the proposed SpectoNet. Table 1 presents the effectiveness of different components in SpectoNet on our dataset. From this study, we can observe that the joint tuning layer improves the accuracy by 2.1%, while the attention technique increases the accuracy by 2.7%. On the other hand, 3×3 filter based CNN exhibits the best accuracy among 3×3, 5×5, and 7×7 filter based CNN. Furthermore, we also compared the performance of our proposed SpectoNet with the state-of-the-art CNN models, i.e., AlexNet [36], VGG16 [37], and ResNet18 [38]. In terms of accuracy, the proposed SpectoNet outperforms the

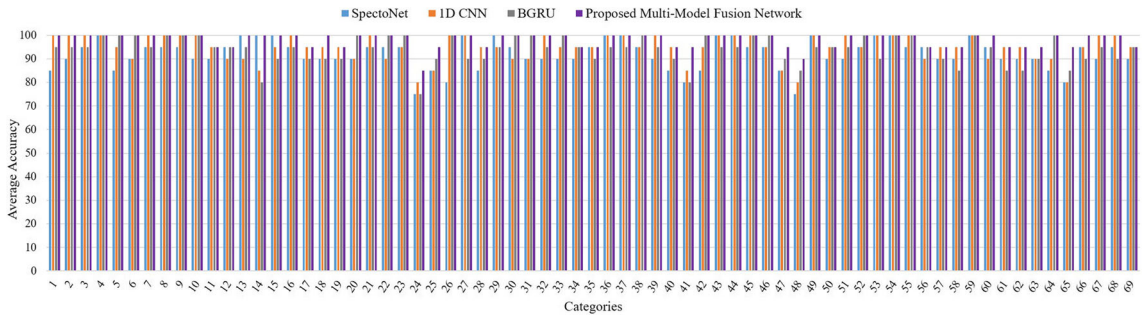

Fig. 12 Accuracy comparison for each category by employing SpectoNet, residual 1D CNN, BiGRU and multi-model fusion network on TUM haptic texture database

Table 1 The effectiveness of different components in SpectoNet using our dataset

Approach	Accuracy	Precision	Recall	F1 Score
AlexNet	81.4%	80.3%	79.1%	79.7%
VGG16	84.6%	83.5%	82.3%	82.9%
ResNet18	86.7%	85.2%	84.3%	84.7%
CNN with 3×3 filter	84.1%	82.9%	81.5%	82.2%
CNN with 5×5 filter	83.5%	81.7%	80%	80.8%
CNN with 7×7 filter	82.6%	81.2%	79.5%	80.3%
SpectoNet without joint tuning layer	88.8%	86.9%	85%	85.9%
SpectoNet without attention mechanism	88.2%	87.1%	86.3%	86.7%
Proposed SpectoNet	90.9%	89.7%	88.4%	89.04%

AlexNet, VGG16, and ResNet18 models by 9.5%, 6.3%, and 4.2%, respectively.

We also evaluate the effectiveness of the residual connections, attention mechanism, number of fully connected layers, and convolutional layers in the residual 1D CNN. Table 2 shows the effectiveness of different components in residual 1D CNN using our dataset. Residual connections increase the accuracy by 1.9%, while the attention mechanism improves the accuracy by 2.3%. In contrast, the three convolutional layers-based residual 1D CNN model shows the lowest accuracy of 91.6%, while residual 1D CNN with one fully connected layer demonstrates an accuracy of 93.7%.

Table 3 demonstrates the performance of BiGRU by varying the number of layers on our dataset. This experiment shows that BiGRU with an attention mechanism outperforms the GRU with an attention mechanism. The attention mechanism improves the accuracy of the proposed BiGRU by 1.8%. From this experiment, we can also observe that increasing the number of layers may not guarantee the best performance. In our work, 3-layer BiGRU outperforms the 4-layer based BiGRU.

Finally, we compared the accuracy of residual 1D CNN, BiGRU, and multi-model fusion network with the different number of segments as input, as presented in Fig. 13. From this experiment, we can see that the proposed multi-model fusion network performs best when the number of segments is set to 15. The performance of the proposed framework reduces with the increase in the number of segments. This is because the proposed multi-model fusion network was unable to extract the spatiotemporal feature effectively when the number of segments increased.

4.5 Comparison with state-of-the-art approaches

In this section, we compared the performance between the proposed multi-model fusion network and state-of-the-art approaches on our dataset and TUM haptic texture database, as demonstrated in Tables 4 and 5. Five-fold cross-validation procedure is used to produce the results. The existing approaches were implemented again to compare them with the proposed approach. The input layer and the network structure are designed the same way as in the literature. However, in ME-DeepL [19] and CLAM [22], final regression layer is replaced with the classification layer, since ME-DeepL [19] and CLAM [22] were initially designed for forecasting tasks. After trying various hyperparameters and performing rigorous training of all the models, we choosed the best hyperparameters. Table 6 reports the optimization hyperparameters used in this work.

On our dataset, the proposed multi-model fusion network significantly outperformed the Haptic CNN [1], Haptic LSTM [1], HapticNet [13], 1D CNN [30], and handcrafted multi-modal features [14, 27]. In contrast, the accuracy of our framework surpasses the state-of-the-art deep learning approaches, i.e., ME-DeepL [19], CNN-LSTM [20] and

Table 2 The effectiveness of different components in residual 1D CNN using our dataset

Approach	Accuracy	Precision	Recall	F1 Score
1D CNN with 3 Conv layers	91.6%	90.2%	89.5%	89.8%
1D CNN with 1 FC layer	93.7%	92.5%	91.3%	91.9%
1D CNN without residual connections	92.9%	91.5%	90%	90.7%
1D CNN without attention mechanism	92.5%	91.7%	90.5%	91.1%
Proposed residual 1D CNN	94.8%	93.5%	92.2%	92.8%

Table 3 Performance of BiGRU by varying the number of layers on our dataset

Approach	Accuracy	Precision	Recall	F1 Score
2 layer GRU with attention mechanism	91.3%	90%	88.5%	89.2%
3 layer GRU with attention mechanism	92.7%	91.5%	89.7%	90.6%
1 layer BiGRU with attention mechanism	91.5%	90.2%	88.9%	89.5%
2 layer BiGRU with attention mechanism	93.6%	92.5%	90%	91.2%
3 layer BiGRU and without attention mechanism	93.3%	92.6%	91.4%	91.9%
4 layer BiGRU with attention mechanism	92.6%	91.3%	90.2%	90.7%
Proposed BiGRU	95.1%	93.7%	92.8%	93.2%

CLAM [22] by 4.7%, 7.2%, and 5.5%, respectively. From this experiment, we can also notice that, our approach beats the existing approaches not only in terms of accuracy but also in terms of precision, recall and F1 score.

Similarly, on TUM haptic texture database, our approach shows better accuracy, precision, recall and F1 score than the Haptic CNN [1], Haptic LSTM [1], HapticNet [13], 1D CNN [30], and handcrafted multi-modal features [14, 27]. However, the ME-DeepL [19], and CLAM [22] demonstrate close performance to the proposed multi-model fusion network.

To present the variability information (min, max, variation) in detail, we also showed the accuracy comparison in terms of box plot in Figs. 14 and 15 for our dataset and the TUM haptic texture database, respectively. Afterwards, we computed the average rankings of the methods (see Fig. 16). According to this experiment, our proposed approach achieved the lowest average ranking among all the existing methods examined in this work. Therefore, we can conclude that our approach exhibited the best performance, followed by the ME-DeepL [19], CLAM [22], CNN-LSTM [20] and HapticNet [13]. Handcrafted multimodal features [14] showed the worst performance. Later on, the Friedman test is conducted to test for significant differences in the performance of different approaches. The test results were significant at $p = 1.09 \times 10^{-7}$ and $p = 1.31 \times 10^{-5}$ for our dataset and TUM haptic texture database, respectively. As significant differences existed among the methods, we continue with the Nemenyi posthoc test. The results of the Nemenyi posthoc test are demonstrated in Figs. 17 and 18 for our dataset and the TUM haptic texture database, respectively. From these results, we can observe that our approach is statistically significant over Haptic CNN [1], Haptic LSTM [1], HapticNet [13], Handcrafted multimodal features [14], CNN [17], and CNN-LSTM [20]. However, in both datasets, ME-DeepL [19] results were not significantly different from our approach.

Table 7 presents the comparison of parameters and computational cost between the proposed approach and existing approaches. The main drawback of our work is that compared to existing methods, our framework requires higher computational cost and needs a larger number of parameters to compute. However, our approach is still able to classify the object in near real-time with higher accuracy.

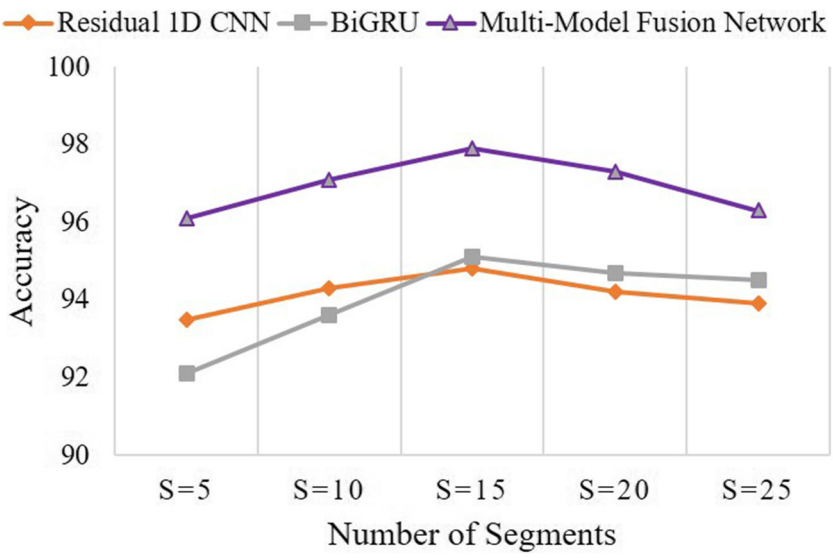

Fig. 13 Accuracy for different number of segments

Table 4 Comparison between proposed multi-model fusion network and state-of-the-arts on our dataset

Method	Accuracy	Precision	Recall	F1 Score
Haptic CNN [1]	87.2%	85.1%	83.2%	84.1%
Haptic LSTM [1]	88.6%	86.5%	84.8%	85.6%
HapticNet [13]	89.5%	87.3%	85.5%	86.4%
Handcrafted multimodal features [14]	67.4%	62.5%	60%	61.2%
CNN [17]	88.3%	86.2%	84.1%	85.1%
Handcrafted multi-modal features [27]	71.1%	68.5%	65%	66.7%
1D CNN [30]	86.4%	84.2%	82.4%	83.3%
ME-DeepL [19]	93.2%	91.2%	90%	90.6%
CNN-LSTM [20]	90.7%	88.9%	87.5%	88.2%
CLAM [22]	92.4%	90.3%	88.5%	89.3%
Proposed Multi-Model Fusion Network	97.9%	95.8%	94.6%	95.2%

Table 5 Comparison between proposed multi-model fusion network and state-of-the-arts on TUM haptic texture database

Method	Accuracy	Precision	Recall	F1 Score
Haptic CNN [1]	88.5%	86.3%	85%	85.6%
Haptic LSTM [1]	89.2%	87.5%	85.7%	86.6%
HapticNet [13]	91%	89.5%	87.3%	88.4%
Handcrafted multimodal features [14]	75%	72.6%	70%	71.3%
CNN [17]	90%	88.2%	86.5%	87.3%
Handcrafted multi-modal features [27]	90.5%	89.1%	87.3%	88.2%
1D CNN [30]	88.6%	85%	83.2%	84.1%
ME-DeepL [19]	94.5%	92.7%	90.3%	91.5%
CNN-LSTM [20]	92.6%	90.3%	88.5%	89.4%
CLAM [22]	94.1%	92.4%	90%	91.2%
Proposed Multi-Model Fusion Network	98.18%	96.5%	95.2%	95.8%

Table 6 Optimization hyperparameters used for training the models

Model	Optimizer	Learning Rate	Momentum	Decay	Batch size	Epochs
Haptic CNN [1]	SGD	0.001	0.9	0.0005	32	400
Haptic LSTM [1]	SGD	0.001	0.9	0.0005	32	350
HapticNet [13]	Adam	0.001	0.9	0.0005	32	800
CNN [17]	Adam	0.001	0.9	0.0005	32	700
1D CNN [30]	Adam	0.001	0.9	0.0005	64	600
ME-DeepL [19]	Adam	0.001	0.9	0.0005	32	800
CNN-LSTM [20]	Adam	0.001	0.9	0.000001	64	500
CLAM [22]	Adam	0.001	0.9	0.0005	32	600
Multi-Model Fusion Network	Adam	0.001	0.9	0.0005	32	600

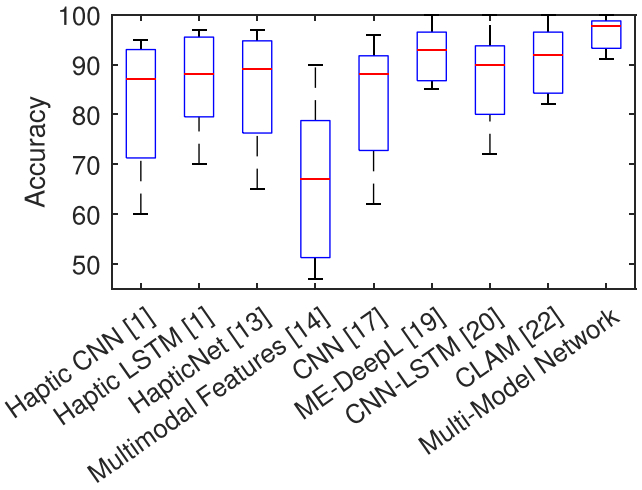

Fig. 14 Accuracy comparison between proposed multi-model fusion network and state-of-the-arts on our dataset using box plot

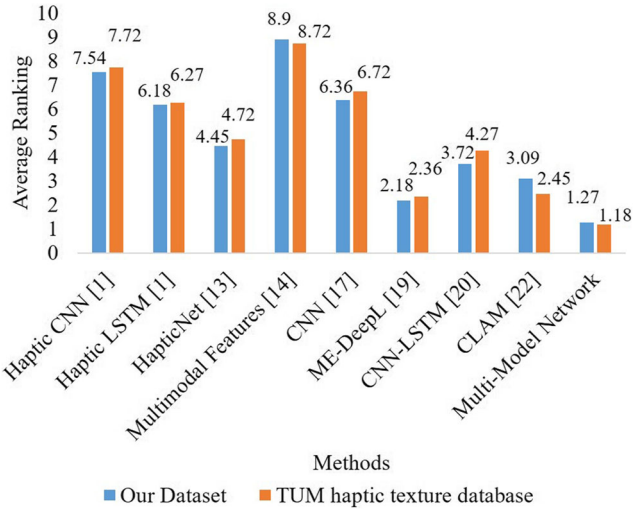

Fig. 16 Average rankings obtained from Friedman test on our dataset and TUM haptic texture database

The main advantages of our work are as follows:

- Our proposed finger-shaped tactile sensor is simple and cost-effective compared to the existing tactile sensor.
- The proposed SpectroNet is capable of extracting multi-scale diverse features.
- The multi-model fusion network obtains the state-of-the-art accuracy, precision, recall and F1 Score on our dataset as well as on the TUM haptic texture database.
- There is a significant improvement in the individual network with the proposed attention mechanism.

5 Conclusion

In this work, we introduced a novel multi-model fusion network for real object's tactile understanding from haptic data. Furthermore, we designed a finger-shaped tactile sensing system for capturing haptic information from objects. Our proposed sensor is composed of a force-sensitive resistor and an MPU-9250 sensor. Hence, the

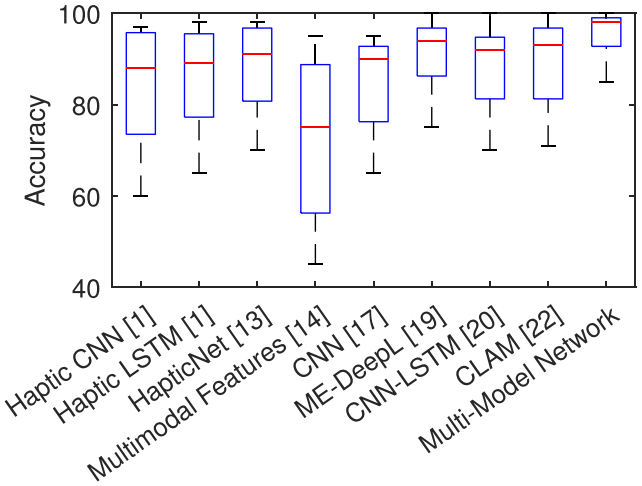

Fig. 15 Accuracy comparison between proposed multi-model fusion network and state-of-the-arts on TUM haptic texture database using box plot

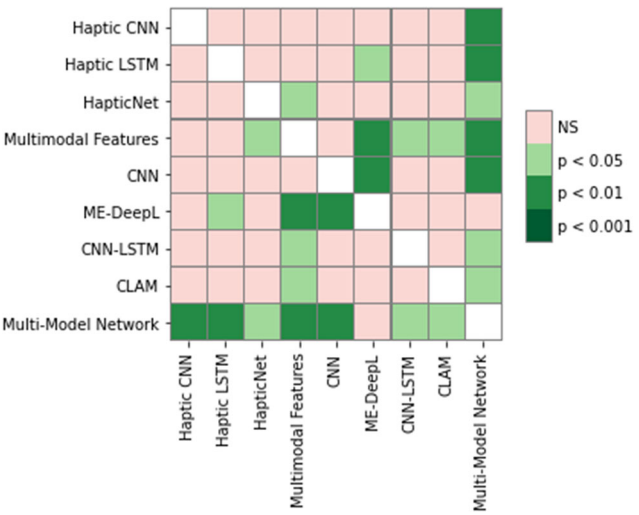

Fig. 17 Nemenyi post-hoc test results on our dataset

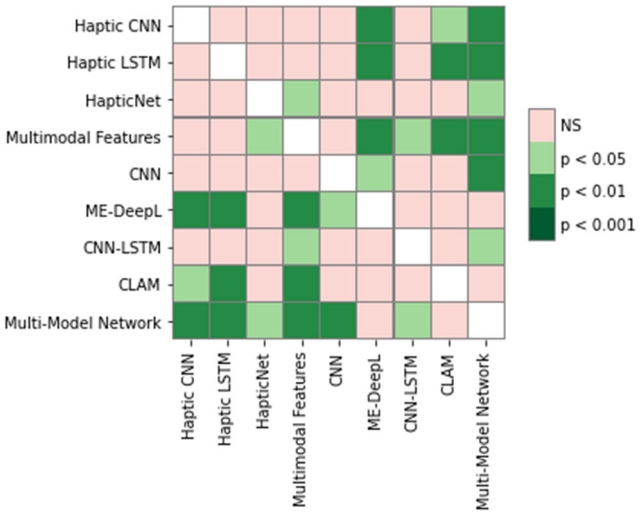

Fig. 18 Nemenyi post-hoc test results on TUM haptic texture database

prototype is simple and cost-effective. The proposed multi-model fusion network consists of an ensemble 2D convolutional neural network, a 1-D convolutional neural network with residual connection, and bi-directional gated recurrent unit networks. Besides, attention mechanisms are applied in all three models to allocate weights to the features according to their contributions, which further improves performance. We also introduced a dataset that contains acceleration signals, angular velocities, and force data from 40 non-planner objects and 20 planner surfaces. We achieved 97.9% accuracy, 95.8% precision, 94.6% recall and 95.2% F1 Score on our dataset and 98.18% accuracy, 96.5% precision, 95.2% recall and 95.8% F1 Score on the TUM haptic texture database, respectively. Besides, experimental evaluations confirm that the proposed method significantly outperformed state-of-the-art approaches.

Acknowledgements This research was supported by the Preventive Safety Service Technology Development Program funded by the Korean Ministry of Interior and Safety under Grant 2019-MOIS34-001.

Table 7 Comparison of parameters and computational cost with different approaches

Method	Parameters (M)	Testing time (s)
Haptic CNN [1]	0.2	0.09
Haptic LSTM [1]	1.7	0.14
HapticNet [13]	4.4	0.19
CNN [17]	0.4	0.12
1D CNN [30]	0.1	0.08
ME-DeepL [19]	43.6	0.75
CNN-LSTM [20]	2.1	0.16
CLAM [22]	1.9	0.16
Multi-Model Fusion Network	62.13	0.92

References

1. Gao Y, Hendricks LA, Kuchenbecker KJ, Darrell T (2016) Deep learning for tactile understanding from visual and haptic data. In: Proceedings of IEEE international conference on robotics and automation 536–543
2. Liu H, Sun F, Guo D, Fang B, Peng Z (2017) Structured Output-Associated dictionary learning for haptic understanding. IEEE Trans Syst Man Cybern Syst 47(7):1564–1574
3. Degol J, Golparvar-Fard M, Hoiem D (2016) Geometry-Informed Material Recognition. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR) 1554–1562
4. Bell S, Upchurch P, Snavely N, Bala K (2015) Material recognition in the wild with the Materials in Context Database. In: Proceedings of IEEE conference on computer vision and pattern recognition (CVPR) 3479–3487
5. Zhang Y, Ozay M, Liu X, Okatani T (2016) Integrating deep features for material recognition. In: Proceedings of 23rd international conference on pattern recognition (ICPR) 3697–3702
6. Jiang X, Du J, Sun B, Feng X (2018) Deep dilated convolutional network for material recognition. In: Proceedings of eighth international conference on image processing theory, tools and applications (IPTA) 1–6
7. Abderrahmane Z, Ganesh G, Crosnier A, Cherubini A (2018) Haptic Zero-Shot learning: Recognition of objects never touched before. Rob Auton Syst 105:11–25
8. Chu V, McMahon I, Riano L, McDonald CG, He Q, Perez-Tejada JM, Arrigo M, Darrell T, Kuchenbecker KJ (2015) Robotic learning of haptic adjectives through physical interaction. Rob Auton Syst 63(3):279–292
9. Liu H, Qin J, Sun F, Guo D (2017) Extreme kernel sparse learning for tactile object recognition. IEEE Trans Cybern 47(12):4509–4520
10. Liu H, Sun F, Fang B, Guo D (2020) Cross-Modal Zero-Shot-Learning For tactile object recognition. IEEE Trans Syst Man Cybern Syst 50(7):2466–2474
11. Abderrahmane Z, Ganesh G, Crosnier A, Cherubini A (2018) Visuo-tactile recognition of daily-Life objects never seen or touched before. In: Proceedings of 15th international conference on control. Robotics and vision (ICARCV), Automation, pp 1765–1770
12. BioTac Product Manual [Online]. Available: https://www.syntouchinc.com/wp-content/uploads/2018/08/BioTac-Manual-V.21.pdf. Accessed 27 Apr 2021
13. Zheng H, Fang L, Ji M, Strese M, Özer Y, Steinbach E (2016) Deep learning for surface material classification using haptic and visual information. IEEE Trans Multimed 18(12):2407–2416
14. Strese M, Schuwerk C, Iepure A, Steinbach E (2017) Multimodal Feature-Based surface material classification. IEEE Trans Haptics 10(2):226–239
15. Strese M, Boeck Y, Steinbach E (2017) Content-based surface material retrieval. In: Proceedings of IEEE world haptics conference (WHC) 352–357
16. Liu H, Sun F, Fang B, Lu S (2018) Multimodal measurements fusion for surface material categorization. IEEE Trans Instrum Meas 67(2):246–256
17. Tsuji S, Kohama T (2019) Using a convolutional neural network to construct a pen-type tactile sensor system for roughness recognition. Sensors and Actuators A:, Physical 291:7–12

18. Kim MG, Pan SB (2019) Deep learning based on 1-D ensemble networks using ECG for Real-Time user recognition. IEEE Trans Industr Inform 15(10):5656–5663
19. Heo S, Nam K, Loy-Benitez J, Yoo C (2021) Data-Driven Hybrid model for forecasting wastewater influent loads based on multimodal and ensemble deep learning. IEEE Trans Industr Inform 17(10):6925–6934
20. Rai HM, Chatterjee K (2021) Hybrid CNN-LSTM deep learning model and ensemble technique for automatic detection of myocardial infarction using big ECG data. Appl Intell. https://doi.org/10.1007/s10489-021-02696-6
21. Huang R, Li J, Li W, Cui L (2020) Deep ensemble capsule network for intelligent compound fault diagnosis using multisensory data. IEEE Trans Instrum Meas 69(5):2304–2314
22. Sun H, Chen M, Weng J, Liu Z, Geng G (2021) Anomaly detection for In-Vehicle network using CNN-LSTM with attention mechanism. IEEE Trans Veh Technol 70(10):10880–10893
23. Culbertson H, Unwin J, Kuchenbecker KJ (2014) Modeling and rendering realistic textures from unconstrained Tool-Surface interactions. IEEE Trans Haptics 7(3):381–393
24. Abdulali A, Atadjanov IR, Jeon S (2020) Visually guided acquisition of contact dynamics and case study in Data-Driven haptic texture modeling. IEEE Trans Haptics 13(3):611–627
25. Zhengkun Y, Yilei Z (2017) Recognizing tactile surface roughness with a biomimetic fingertip: a soft neuromorphic approach. Neurocomputing 244:102–111
26. Tanaka Y, Hasegawa T, Hashimoto M, Igarashi T (2019) Artificial Fingers Wearing Skin Vibration Sensor for Evaluating Tactile Sensations. In: Proceedings of IEEE world haptics conference (WHC) 377–382
27. Strese M, Brudermueller L, Kirsch J, Steinbach E (2020) Haptic material analysis and classification inspired by human exploratory procedures. IEEE Trans Haptics 13(2):404–424
28. Fulop SA, Fitz K (2006) Algorithms for computing the time-corrected instantaneous frequency (reassigned) spectrogram, with applications. Journal of the Acoustical Society of America 119:360–371
29. MPU-9250 Product Specification. [Online]. Available: https://invensense.tdk.com/wp-content/uploads/2015/02/PS-MPU-9250A-01-v1.1.pdf. Accessed 27 Apr 2021
30. Bianco S, Napoletano P (2019) Biometric recognition using multimodal physiological signals. IEEE Access 7:83581–83588
31. Gao H, Kong D, Lu M, Bai X, Yang J (2018) Attention Convolutional Neural Network for Advertiser-level Click-through Rate Forecasting. In: Proceedings of the world wide web conference 1855–1864
32. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural computation 9(8):1735–1780
33. Li D, Fu Z, Xu J (2021) Stacked-autoencoder-based model for COVID-19 diagnosis on CT images. Appl Intell 51:2805–2817
34. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. J Mach Learn Res 15(1):1929–1958
35. Sokolova M, Lapalme G (2009) A systematic analysis of performance measures for classification tasks. Inf Process Manag 45(4):427–437
36. Krizhevsky A, Sutskever I, Hinton G (2012) ImageNet classification with deep convolutional neural networks. In: Proceedings of the 25th international conference on neural information processing systems 1097–1105
37. Simonyan K, Zisserman A (2015) Very deep convolutional networks for large-scale image recognition. In: Proceedings of the international conference on learning representations (ICLR) arXiv preprint arXiv:1409.1556
38. He K, Zhang X, Ren S, Sun J (2016) Deep Residual Learning for Image Recognition. In: Proceeding of the IEEE conference on computer vision and pattern recognition (CVPR) 770–778

Joolekha Bibi Joolee is currently a Ph.D. student of the Dept. of Computer Science and Engineering in Kyung Hee University, South Korea. She received her Masters degree in Computer Science and Engineering from kyung Hee University, South Korea and received her B.Sc. degree in Computer Science and Engineering from International Islamic University Chittagong, Bangladesh in 2015. Her research interests include HCI, haptic feedback modeling and rendering, and Deep Learning.

Md Azher Uddin received his B.Sc. degree in Computer Science and Engineering from International Islamic University Chittagong, Bangladesh in 2011 and the Masters leading to Ph.D. degree in Computer Science and Engineering from Kyung Hee University, South Korea, in August 2020. He is currently an Assistant Professor at the Department of Mathematical and Computer Sciences, Heriot-Watt University Dubai, United Arab Emirates. His research interests include Image Processing, computer vision, Machine Learning, and Big Data analytics.

Seokhee Jeon received the B.S. and PhD. degrees in computer science and engineering from the Pohang University of Science and Technology in 2003 and 2010, respectively. He was a postdoctoral research associate in the Computer Vision Laboratory at ETH Zurich. In 2012, he joined as an assistant professor the Department of Computer Engineering at Kyung Hee University. His research focuses on haptic rendering in an augmented reality environment, applications of haptics technology to medical training, and usability of augmented reality applications.