Mechatronics and Automation Technology
J.-Y. Xu (Ed.)

doi:10.3233/ATDE231142

An End-to-End Haptic Adjective Recognition Method with Self-Attention Mechanism

Yuanpei ZHANG [a,b], Zhuojun ZOU [a,c], Lin SHU [a,c], and Jie HAO [a,c,1]

[a] *Institute of Automation, Chinese Academy of Sciences, Beijing, China*
[b] *School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing, China*
[c] *Guangdong Institute of Artificial Intelligence and Advanced Computing, Guangzhou, China*

Abstract. Human's ability to describe the tactile sensation of objects has piqued the interest of numerous researchers seeking to augment the dexterity of robots in delicate tasks. However, most existing approaches are limited by their two-stage framework, resulting in low inference efficiency and unsatisfactory performance. To address this challenge, we propose the first end-to-end framework for haptic adjective classification. Specifically, our framework leverages the Space Encoding Module to capture long-term dependencies, and the Order Encoding Module to learn order information explicitly. We conduct experiments on the public PHAC-2 Dataset and the result shows that our method achieves F1 score of 0.759, outperforming previous work in a significant way.

Keywords. Haptic; Adjective Recognition; End-to-End Framework.

1. Introduction

Haptic perception is an essential component of human interaction with the external environment and can serve as a crucial modality in extreme scenarios described as dark, rainy, or foggy [1] . In recent years, tactile sensors have been increasingly integrated into robotic systems to perform various tasks, including texture classification [2] , slip detection [3] , and object recognition [4] . However, some researchers have shifted their focus towards more abstract haptic learning tasks, such as adjective recognition [5] . Adjective recognition aims to quantify the haptic properties that humans use to describe objects, including roughness, hardness, and compressibility. By processing tactile data, adjective recognition algorithms provide a feature description of an object, which can help robots perform practical downstream tasks such as grasping and classifying. For example, robots can classify previously unseen objects by identifying their properties and judging whether their surfaces are smooth or bumpy. Thus, adjective recognition is a

[1] Corresponding Author, Jie HAO, Institute of Automation, Chinese Academy of Sciences, Beijing, China; Guangdong Institute of Artificial Intelligence and Advanced Computing, Guangzhou, China; Email: jie.hao@ia.ac.cn.

practical and challenging research topic with numerous potential applications in the field of robotics.

Previous studies on adjective recognition have largely relied on hand-crafted features [5] or have been limited to specific tasks [6] . Richardson *et al.* [7] avoided these limitations by using unsupervised feature learning methods such as K-SVD and Spatio-Temporal Hierarchical Matching Pursuit (ST-HMP), but their method is based on a two-stage framework, which is not efficient during the inference process. To the best of our knowledge, although the end-to-end learning method has been successfully applied to multiple tactile applications such as texture image recognition [2] and slip detection [3] , it has not been used in adjective classification yet. Therefore, we introduce the first end-to-end haptic adjective recognition framework in this paper. Motivated by the success of Transformer [8] , we propose a novel attention-based Long Short-Term Memory (Atten-LSTM) model to evaluate haptic adjectives.

Our main contributions can be summarized as follows.

- We introduce the first end-to-end framework for haptic adjective recognition, which exhibits high inference efficiency and satisfactory performance. Specifically, in our proposed Atten-LSTM model, the Space Encoding Module is to capture long-term dependencies, while the Order Encoding Module is to learn order information.
- We conduct experiments on the Atten-LSTM model with the aim of evaluating its performance. The experimental results clearly demonstrate the effectiveness and strong generalization capability of the proposed network.

2. Related Work

Certain studies have placed a significant emphasis on guiding robots to learn intricate concepts and representations. The work of Chu *et al.* was pioneering in adjective recognition where they introduced a predefined set of adjectives and trained SVMs on tactile data to match adjectives with various objects [5] . In a similar vein, Strese *et al.* used acceleration sensors to capture vibrations when tapping onto an object with a tool, and developed a classification system that uses perception-related features such as hardness, roughness, and friction [9] . These approaches rely on hand-crafted features, which often require expertise and have a strong dependence on the data format. In contrast, Madry *et al.* proposed an unsupervised feature learning algorithm to avoid designing features [6] . However, their learned properties were only evaluated in some particular tasks such as grasp stability assessment and object recognition.

Recently, Richardson *et al.* used unsupervised feature learning methods, specifically K-SVD and Spatio-Temporal Hierarchical Matching Pursuit (ST-HMP) to perform adjective recognition tasks [7] . Similarly, Wu *et al.* proposed a multi-label classification approach to capture potential relationships among different adjectives [10] . Compared with the method of Chu *et al.* [5] , their method has stronger generalization performance.

In the preceding literature, the work of [5] and [7] is most closely related to our research. However, they both rely on a two-stage framework that results in low inference efficiency and limited generalization ability. In this work, we attempt to use an end-to-end approach to learn complex haptic concepts and representations.

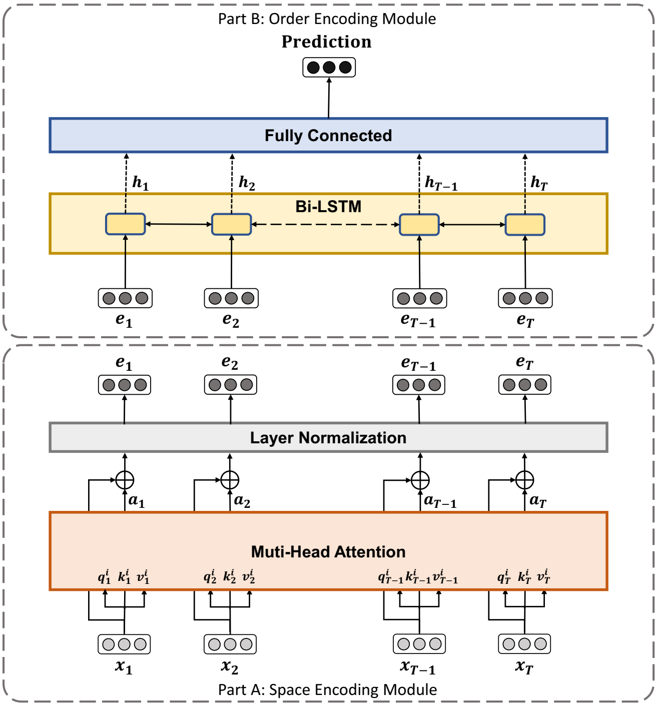

Figure 1. Overall architecture of the Atten-LSTM model.

3. Methodology

Our goal is to perform adjective recognition; in other words, we expect to obtain all the adjective properties corresponding to a given tactile sequence from the material. We adopt the adjective set defined by Chu *et al.* [5] and transform the adjective perception task into a multi-label classification problem.

Tactile data sequences typically comprise hundreds of frames and convey strong order information. While LSTM [11] have shown remarkable capabilities in modeling sequence data and capturing order information, their effectiveness is limited when encountering long sequences. Thus, processing tactile data that may contain hundreds of frames cannot be handled by these models alone. Conversely, the self-attention mechanism can associate different frames within a sequence by computing correlations without regard to their distances [8] . However, since the correlation computation is performed globally, it cannot take advantage of the order information in the sequence. Consequently, models relying solely on the self-attention mechanism are inadequate for processing tactile sequences.

In this scenario, we propose a novel attention-based LSTM network (Atten-LSTM) as illustrated in figure 1. The network consists of two modules: Space Encoding Module and Order Encoding Module. The former incorporates the multi-head self-attention mechanism to capture long-term dependencies adequately, while the latter utilizes either the LSTM layers to learn order information. By leveraging the complementary strengths of these modules, we argue that the Atten-LSTM model exhibits a robust modeling capability for tactile data sequences with hundreds of frames and strong order information.

3.1. The Space Encoding Module

Firstly, the tactile sequences are encoded by the Space Encoding Module. Following the method of Transformer [8] , the Space Encoding Module utilizes the multi-head self-attention mechanism to transform input sequences into abstract and high-dimensional embedding representations. Suppose the input to the network is a sequence expressed as

$$X = \{x_1, x_2, ..., x_{T-1}, x_T\} \in \mathbb{R}^{T\times d}, \tag{1}$$

where $x_t \in \mathbb{R}^d$ is the input signal at time step t, d is the dimension of the signal, and T is the length of the sequence. Multi-head attention allows the model to simultaneously focus on information among different frames in multiple representation subspaces. To conduct multi-head self-attention, X is projected h times using different linear projection matrices to obtain queries Q_i, keys K_i and values V_i for h different subspaces, or heads:

$$Q_i = XW_i^Q,\ \ K_i = XW_i^K,\ \ V_i = XW_i^V, \tag{2}$$

where $W_i^Q, W_i^K, W_i^V \in \mathbb{R}^{d\times(d/h)}$ are linear projection matrices for the i-th head, and $i = 1, 2, ..., h$ is the index of heads. Next, we compute scaled dot-product attention for each head:

$$head_i = \text{Softmax}(\frac{Q_i K_i^T}{\sqrt{d_k}})V_i, \tag{3}$$

where $\frac{1}{\sqrt{d_k}} = \sqrt{\frac{h}{d}}$ is the scaling factor, and $\mathrm{head_i} \in \mathbb{R}^{T\times(d/h)}$ represents the output attention for the i-th head. Then, results from each head are concatenated and projected to the initial dimensional space:

$$A = \{a_1, a_2, ..., a_{T-1}, a_T\} = \text{Concat}(head_1, ..., head_h)W^O, \tag{4}$$

where $W_O \in \mathbb{R}^{d\times d}$ is the linear projection matrix. Finally, we employ a residual connection [12] for the multi-head attention layer, followed by layer normalization [13] :

$$E = \{e_1, e_2, ..., e_{T-1}, e_T\} = \text{LayerNorm}(X + A) \in \mathbb{R}^{T\times d}. \tag{5}$$

3.2. The Order Encoding Module

The Order Encoding Module utilizes the LSTM layers to model the temporal dynamics of the encoded sequence E. Composed of a cell and several gates, the LSTM unit stores historical information through the memory cell and implements the long-term memory by adding or removing the cell state using different gates. In our implementation, we employ multiple bidirectional LSTM layers to capture the context of the sequence in both forward and backward directions. The hidden states h_t generated by the final LSTM layer are sampled every f frames and fed into a fully connected network which makes the prediction:

$$\begin{aligned} H &= \text{LSTM}(E) = \{h_1, h_2, ..., h_{T-1}, h_T\}, \\ Y &= \text{FC}(\text{Sample}(H)) = \text{FC}(\{h_{fi}\}),\ \ i \in \mathbb{N}. \end{aligned} \tag{6}$$

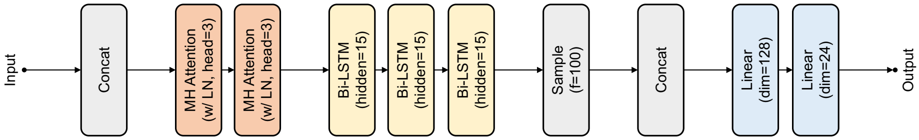

Figure 2. Detailed Parameters of the Atten-LSTM model.

3.3. Loss Function

We adopt the multi-label weighted binary cross-entropy loss function [14] for our training. The total loss is defined as the average of the losses for each sample in a batch, and the loss l_n for each sample is defined as:

$$l_n = -\frac{1}{C}\sum_{c=1}^{C} p_c y_{n,c} \log\left(\sigma(x_{n,c})\right) + (1 - y_{n,c}) \log\left(1 - \sigma(x_{n,c})\right), \tag{7}$$

where C is the number of adjectives, $y_{n,c}$ is the binary label for the c-th adjective, $x_{n,c}$ is the prediction for the c-th adjective, and $\sigma(\cdot)$ is the sigmoid function. Due to the imbalanced distribution of positive and negative samples in the dataset, we apply weight p_c to achieve a trade-off between precision and recall. The value of p_c is calculated based on the proportion of positive samples in the c-th adjective. Specifically, when considering an adjective with N_p positive samples and N_n negative samples, the value of p_c is determined as:

$$p_c = \frac{N_n}{N_p}. \tag{8}$$

3.4. Implementation Details

A detailed description of the model parameters for both models is presented in figure 2. The Atten-LSTM model includes 2 multi-head attention layers, 3 Bi-LSTM layers, and 2 linear layers, with a sampling interval of $f = 100$.

Besides, we set the batch size as 16 and employ Adam optimizer [15] during our training process. The learning rate is initially set as 5e-4 and decays to 5e-5 halfway through training. The model is implemented on the PyTorch platform and trained on an Nvidia Tesla V100 GPU.

4. Experiments and Results

Our research is closely related to the studies conducted by Chu *et al.* [5] and Richardson *et al.* [7], where they employ the PHAC-2 Dataset to assess the performance of their methods. The PHAC-2 Dataset is collected using a robot that is equipped with two BioTac tactile finger sensors capable of measuring vibration, pressure, temperature, and fingertip deformation. During the data collection process, the robot executed four Exploratory Procedures (EPs) of *Squeeze*, *Hold*, *Slow Slide*, and *Fast Slide*.

Hence, we validated the performance of our Atten-LSTM model on the PHAC-2 dataset. We used the same training and testing sets as Chu *et al.* and Richardson *et al.* so that we could directly compare our results. Specifically, we used different training

Table 1. F1 Scores across Adjectives and EPs Using the Atten-LSTM Model. $\gg$ and $>$ represent relative increases in performance from Richardson *et al.* [7] of more than 0.10 and 0.03, respectively. $\approx$ represents a difference of no more than 0.03. $\ll$ and $<$ represent relative decreases of more than 0.10 and 0.03, respectively. Darker shadings indicate higher performance.

	Squeeze	*Hold*	*Slow Slide*	*Fast Slide*	PE*
smooth	0.818 $\gg$	0.511 $<$	0.600 $\approx$	0.600 $\approx$	25
solid	1.000 $\approx$	1.000 $\approx$	1.000 $\approx$	1.000 $\approx$	22
squishy	1.000 $>$	0.968 $>$	0.906 $>$	0.867 $>$	21
compressible	1.000 $>$	1.000 $\approx$	1.000 $>$	0.967 $>$	20
hard	0.983 $\approx$	1.000 $\approx$	0.984 $\approx$	0.967 $<$	20
textured	0.444 $\ll$	0.778 $\gg$	0.667 $>$	0.667 $\gg$	16
soft	0.930 $\gg$	0.976 $>$	0.857 $>$	0.792 $>$	13
absorbent	0.952 $>$	0.976 $\gg$	0.930 $\gg$	0.927 $>$	9
rough	0.870 $\gg$	0.800 $\gg$	0.571 $>$	0.750 $>$	9
thick	0.457 $\ll$	0.615 $>$	0.541 $\approx$	0.553 $\gg$	9
cool	0.800 $\gg$	0.900 $\gg$	0.750 $>$	0.783 $\gg$	8
slippery	0.778 $<$	0.889 $>$	0.462 $\ll$	0.588 $<$	8
fuzzy	0.444 $\gg$	0.254 $\approx$	0.625 $\gg$	0.154 $\ll$	6
porous	1.000 $\approx$	0.690 $\approx$	0.952 $\gg$	0.778 $\gg$	6
springy	0.741 $\gg$	0.432 $\approx$	0.667 $\gg$	0.824 $\gg$	6
scratchy	0.889 $\gg$	0.696 $\gg$	0.040 $\ll$	0.483 $\gg$	5
hairy	0.316 $<$	1.000 $\gg$	0.609 $\gg$	0.571 $\gg$	4
bumpy	0.625 $\gg$	0.952 $\gg$	0.267 $\ll$	0.818 $\ll$	2
metallic	1.000 $\gg$	1.000 $\gg$	0.800 $\gg$	0.857 $\gg$	2

*PE indicates the number of positive examples in the training set.

and test splits for each adjective, selecting 10% of both positively and negatively labeled materials to form the testing set for each adjective classifier, while the remaining materials were used for training. To avoid the classifier from mistakenly learning to classify materials instead of adjectives, all explorations for each material were placed in the same set. In addition, following the approach of Richardson *et al.*, we excluded adjectives with fewer than three positively labeled materials, analyzing 19 out of the original 24 adjectives used by Chu *et al.*

The F1 scores obtained from the Atten-LSTM are presented in table 1, with adjectives ordered by the number of positively labeled materials in the training set. We compared our results directly with those reported by Richardson *et al.*, as their findings have been proved to significantly outperform those of Chu *et al.*

The results demonstrate the superior performance of our proposed models compared to the studies conducted by Chu *et al.* and Richardson *et al.* Specifically, table 1 depict that our results outperform those reported by Richardson *et al.* for most adjectives during different EPs. Moreover, the mean of all individual F1 scores by Richardson *et al.* is *0.673*, whereas the mean of the best F1 scores across static and dynamic feature classifiers by Chu *et al.* is *0.371*. In contrast, our proposed models achieved an average F1 score of *0.759* for the Atten-LSTM, which exceed the F1 scores reported by Richardson *et al.* and Chu *et al.*

5. Conclusion and Future Work

In this paper, an end-to-end framework named Atten-LSTM is proposed to evaluate the haptic adjective attributes of various materials. The network leverages a combination of the self-attention mechanism and LSTM to extract the essential features of materials. The experimental results demonstrate the effectiveness of the proposed network.

Acknowledgements

This work is supported by the National Science and Technology Major Project from Minister of Science and Technology of China (Grant No. 2018AAA0103100), Guangzhou Basic Research Program (Grant No. 202201011389), and National Natural Science Foundation of China (Grant No. 62236007).

References

[1] LUO S, BIMBO J, DAHIYA R, et al. Robotic tactile perception of object properties: A review [J]. Mechatronics, 2017, 48: 54-67.

[2] FANG B, LONG X, SUN F, et al. Tactile-based fabric defect detection using convolutional neural network with attention mechanism[J]. IEEE Transactions on Instrumentation and Measurement, 2022.

[3] ZAPATA-IMPATA B S, GIL P, TORRES F. Learning spatio temporal tactile features with a convlstm for the direction of slip detection[J]. Sensors, 2019, 19(3): 523.

[4] KIRBY E, ZENHA R, JAMONE L. Comparing single touch to dynamic exploratory procedures for robotic tactile object recognition[J]. IEEE Robotics and Automation Letters, 2022, 7(2): 4252-4258.

[5] CHU V, MCMAHON I, RIANO L, et al. Robotic learning of haptic adjectives through physical interaction[J]. Robotics and Autonomous Systems, 2015, 63: 279-292.

[6] MADRY M, BO L, KRAGIC D, et al. St-hmp: Unsupervised spatio-temporal feature learning for tactile data[C]//2014 IEEE International Conference on Robotics and Automation (ICRA). IEEE, 2014: 2262-2269.

[7] RICHARDSON B A, KUCHENBECKER K J. Improving haptic adjective recognition with unsupervised feature learning[C]//2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019: 3804-3810.

[8] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[9] STRESE M, SCHUWERK C, STEINBACH E. Surface classification using acceleration signals recorded during human freehand movement[C]//2015 IEEE World Haptics Conference (WHC). IEEE, 2015: 214-219.

[10] WU H, LIU X, FANG S, et al. Leveraging multi-label correlation for tactile adjective recognition[C]//2020 3rd International Conference on Robotics, Control and Automation Engineering (RCAE). IEEE, 2020: 122-126.

[11] SUNDERMEYER M, SCHLÜTER R, NEY H. Lstm neural networks for language modeling [C]//Thirteenth annual conference of the international speech communication association. 2012.

[12] HE K, ZHANG X, REN S, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[13] BA J L, KIROS J R, HINTON G E. Layer normalization[A]. 2016.

[14] HO Y, WOOKEY S. The real-world-weight cross-entropy loss function: Modeling the costs of mislabeling[J]. IEEE access, 2019, 8: 4806-4813.

[15] KINGMA D P, BA J. Adam: A method for stochastic optimization[A]. 2014.