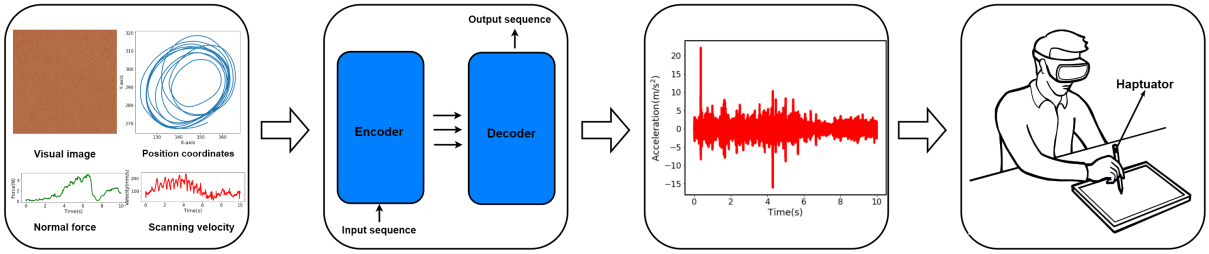
Multi-modal Transformer-based Tactile Signal Generation for Haptic Texture Simulation of Materials in Virtual and Augmented Reality

Shaoyu Cai*
City University of Hong Kong

Kening Zhu†
City University of Hong Kong



(a) Multi-modal visual-tactile input (b) Transformer-based generative model (c) Generated dynamic acceleration signals (d) Stylus-based texture interaction

Figure 1: The main concept of our multi-modal tactile signal generation framework using the transformer-based network. In our system, the (b) transformer-based generative model takes the RGB visual image, and users' scanning parameters (i.e., the position coordinates of the input stroke, the applied normal forces and the scanning velocities) as (a) multi-modal visual-tactile input, and accordingly generates (c) the sequence of dynamic contact acceleration signals for (d) stylus-based texture interaction in VR/AR.

Abstract

Current haptic devices can generate haptic texture sensations through replaying the recorded tactile signals, allowing for texture interaction of different materials in virtual reality (VR) and augmented reality (AR). As humans enable to feel different texture sensations under various scanning parameters (i.e., applied normal forces, scanning velocities and stroking directions/positions) on the material surface towards the same texture, such methods cannot support rendering natural haptic textures under various scanning parameters. To this end, we proposed a deep-learning-based approach for multi-modal tactile signal generation leveraging the framework of a transformer-based network. Our system takes the visual image of a material surface as the visual data and the acceleration signals with the scanning parameters induced by the pen-sliding movement on the surface as tactile data through a transformer-based generative model with the multi-modal feature embedding module for acceleration signals synthesis. We aim to synthesize dynamic acceleration signals based on the images of material surfaces and the users' scanning states to create natural and realistic texture sensations in VR/AR.

Index Terms: Human-centered computing—Human computer Interaction (HCI)—Interaction devices—Haptic devices

1 Introduction

Recently, some haptic devices have shown promising ability to render realistic haptic textures, such as tool-based haptic device [5] or barehand-based touchscreen [1], which potentially allows for texture simulation of materials in VR/AR to improve the realism of virtual scenes. Typically, when a user seeks to render the virtual texture for a real object, its surface haptic information (e.g., roughness) should be acquired to control haptic devices for texture simulation. However, it is still challenging to acquire high-quality tactile signals for materials texture simulation in VR/AR.

*e-mail: shaoyu.cai@my.cityu.edu.hk
†e-mail: keninzhu@cityu.edu.hk

While providing virtual textures through recorded tactile signals (e.g., contact acceleration signals or frictional-coefficient signals) can achieve realistic texture interaction to a large extent, such the approach ignores those scanning parameters (i.e., scanning velocities, applied forces, and directions), which limits to reproduce natural texture sensations under different scanning conditions. Culbertson et al. [5] presented a data-driven method to generate the acceleration signals varied with users' scanning velocities and forces based on autoregressive model sets for texture simulation. However, such an approach may not be practical for those anisotropic material surfaces. On the other hand, considering visual information of textured surfaces, Cai et al. [2] proposed a GAN-based image-to-friction generation framework to generate positional-based frictional signals from images for haptic texture simulation on the electrovibration tactile display. However, the unimodal visual input (i.e., images) excludes the temporal information of users' input, which may harm the interactive experience in VR/AR.

Motivated by these previous works, one potential idea is to adopt multi-modal information that includes spatial and temporal contents (i.e., visual texture images and users' scanning parameters) to generate tactile signals for realistic haptic texture rendering. In addition, some deep-learning-based approaches, especially the transformer-based generative methods [7], achieve considerable performance on multi-modal data generation (e.g., images, texts, and time-series signals). Hence, for the multi-modal tactile signal generation task, the visual textured images and users' scanning parameters can be treated as the visual input data and the tactile input data, respectively, and the contact acceleration signals as the tactile output data. Our goal is to build the mapping between the multi-modal visual-tactile input and the tactile output data.

In this work, we developed an end-to-end transformer-based framework for tactile signal synthesis, to create realistic texture interaction in virtual and augmented reality. As shown in Fig. 1, our model takes the multi-modal visual-tactile input: visual images, position coordinates, normal forces and scanning velocities, and outputs the accordingly acceleration signals for vibrotactile texture rendering. We modified the basic transformer model [7] to adapt our multi-modal data sequence input and added an additional multi-modal embedding layer for multi-modal feature fusion. We built an augmented visual-tactile haptic data set based on the HaTT database [4] for training our proposed model and conducted a preliminary test to validate the

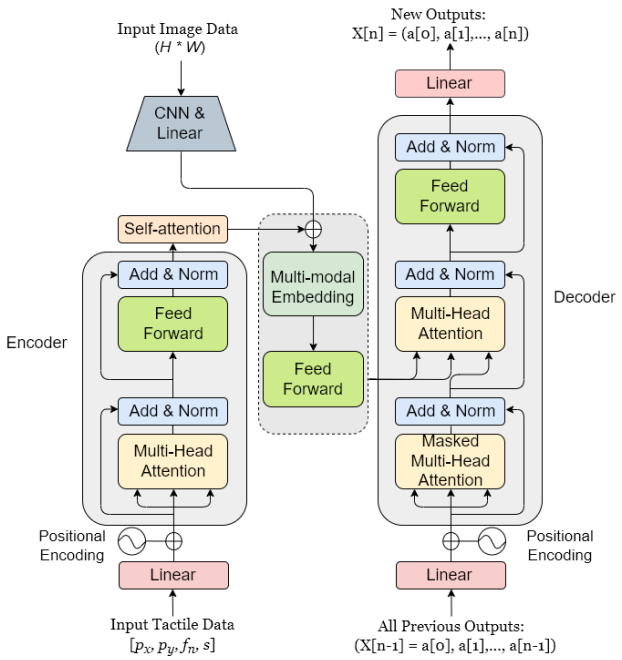


Figure 2: The framework of our tactile signal generation model.

feasibility of multi-modal dynamic tactile signal generation. The generated dynamic acceleration signals can further control the Haptuator embedded into the stylus to produce vibrotactile textures of virtual objects on a display surface (e.g., the tablet in Fig. 1 (d)) in VR/AR.

2 Data Representation and Network Description

In our work, we aim to build the algorithmic mapping between the multi-modal visual-tactile input (visual images x, scanning velocities v, positions p and normal forces f) and the tactile output (acceleration signals a), which can be simplified as: $G(x,v,p,f) \rightarrow a$ where G represents the generative model. In HaTT database [4], the stroke data in any direction has been recorded as a list of positional points with a 10 kHz sampling rate, where the scanning speed is implicitly included in the sequence of the points, so we adopted the similar data format of Sketch-RNN [6] to represent our users' input stroke data. Specifically, we represented the users' scanning input data as a 4-d vector: $[p_x,p_y,f_n,s]$, where the first two elements (p_x, p_y) represented the coordinate offsets in the 2-d coordinate system and the third element f_n indicated the applied normal force at the current coordinates; the last element s was a binary value representing whether the pen was lifted away from the display.

Fig. 2 shows the proposed network structure of our multi-modal transformer-based tactile signal generation model consisting of an encoder block and a decoder block. Unlike the traditional transformer structure applied in natural language processing (NLP) [7], we replaced the input/output embedding layers with linear layers to fit our input size of the encoder/decoder, and a linear projection layer to reshape the final output from the decoder into the target size. In particular, we added a self-attention layer, a multi-modal embedding layer and a feed forward layer in the latent space (the dotted line block in Fig. 2) to adapt multi-modal visual-tactile data fusion [3]. Specifically, we first built a 2-d CNN-based structure followed by a linear layer to extract the features of the image data, and then implemented the self-attention calculation for the encoder's output to produce another feature vector, both two feature vectors for the channel-wise concatenation in the multi-modal embedding layer. Then we passed the concatenation into a feed forward layer and connected to the decoder. Finally, the decoder generated the final output from the multi-modal fusion input from the latent space and the previous output.

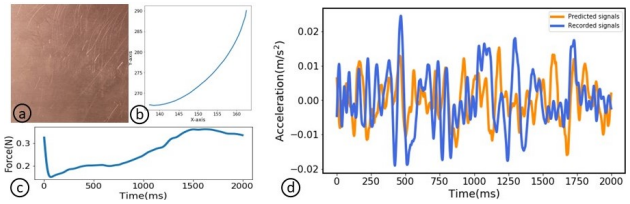


Figure 3: The preliminary results of tactile signal generation for aluminium material. The input data include (a) textured image, (b) positional sequence and (c) force sequence; the output data is (d) acceleration signals (The orange and blue lines represent the generated and recorded (real) signals, respectively).

3 Preliminary Results

We preliminary evaluated our proposed multi-modal transformer-based network for tactile signal generation. We selected the aluminium texture shown in Fig.3 (a) as our tested sample and adopted the sliding window-based data augmentation strategy [2] for the multi-modal data set creation with L_1 loss as our training loss function. To optimize the calculated performance, we split the temporal data sequence as the independent sequence with 200-ms long, which includes a maximum of 2000 data points for each generated acceleration signals sequence. Fig.3 (d) demonstrates the result of generated acceleration signals under the visual textured image as spatial input and sliding stroke data including the positional sequence and force sequence as temporal input. The Root-Mean-Squared-Error (RMSE) value of the selected sample was 0.0082 (SD = 0.0023).

4 Conclusion and Future Work

We presented a multi-modal transformer-based framework for stylus-based haptic texture modelling and rendering in virtual and augmented reality. In particular, our proposed network taking the multi-modal visual-tactile input, captures the temporal and spatial features from users' scanning parameters and material textured surfaces to generate dynamic contact acceleration signals for vibrotactile texture rendering. In our future work, we aim to quantitatively and qualitatively evaluate the performance of generated acceleration signals and explore the effectiveness of our added modules (e.g., multi-modal embedding layers) through several ablation studies. In addition, we will conduct user studies to evaluate the realness of generated virtual haptic textures and further integrate our haptic texture modelling and rendering framework with VR/AR devices to improve the user experience of texture simulation in virtual environments.

References

[1] O. Bau, I. Poupyrev, A. Israr, and C. Harrison. Teslatouch: electrovibration for touch surfaces. In *Proceedings of the 23nd annual ACM symposium on User interface software and technology*, pp. 283–292, 2010.

[2] S. Cai, L. Zhao, Y. Ban, T. Narumi, Y. Liu, and K. Zhu. Gan-based image-to-friction generation for tactile simulation of fabric material. *Computers & Graphics*, 102:460–473, 2022.

[3] S. Cai, K. Zhu, Y. Ban, and T. Narumi. Visual-tactile cross-modal data generation using residue-fusion gan with feature-matching and perceptual losses. *IEEE Robotics and Automation Letters*, 6(4):7525–7532, 2021.

[4] H. Culbertson, J. J. L. Delgado, and K. J. Kuchenbecker. One hundred data-driven haptic texture models and open-source methods for rendering on 3d objects. In *2014 IEEE Haptics Symposium (HAPTICS)*, pp. 319–325. IEEE, 2014.

[5] H. Culbertson, J. Unwin, and K. J. Kuchenbecker. Modeling and rendering realistic textures from unconstrained tool-surface interactions. *IEEE transactions on haptics*, 7(3):381–393, 2014.

[6] D. Ha and D. Eck. A neural representation of sketch drawings. In *International Conference on Learning Representations*, 2018.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.