Low-latency Classification of Social Haptic Gestures Using Transformers

Qiaoqiao Ren[*]
Qiaoqiao.Ren@ugent.be
AIRO - IDLab - Ghent University - IMEC
Ghent, Belgium

Yuanbo Hou[*]
Yuanbo.Hou@ugent.be
Waves - Ghent University
Ghent, Belgium

Tony Belpaeme
Tony.Belpame@ugent.be
AIRO - IDLab - Ghent University - IMEC
Ghent, Belgium

ABSTRACT

Social touch, and its recognition and classification, is increasingly important in human-robot interaction. We present a Transformer-based model trained and evaluated on an open-source dataset. The dataset, the Human-Animal Affective Robot Touch (HAART) dataset, was collected for the 2015 Recognition of Touch Gesture Challenge (RTGC 2015) and contains different haptic actions directed at a robotic animal. The actions are recorded using a multi-resolution pressure sensor. We feed the output, containing the touch type to the Nao robot to make the robot sense the touch type. The proposed transformer-based gesture classification model achieved 72.8% classification accuracy in 2.67 seconds, which outperforms the best-submitted algorithm of the RTGC 2015 which has a test classification accuracy of 70.9 % and needed 8 seconds.

CCS CONCEPTS

• **Computing methodologies** → **Neural networks**; • **Computer systems organization** → *Robotics*; • **Human-centered computing** → User centered design.

KEYWORDS

Social touch interaction, Gestures classification, Convolutional neural networks, Transformer, Attention mechanism

ACM Reference Format:
Qiaoqiao Ren, Yuanbo Hou*, and Tony Belpaeme. 2023. Low-latency Classification of Social Haptic Gestures Using Transformers. In *Companion of the 2023 ACM/IEEE International Conference on Human-Robot Interaction (HRI '23 Companion), March 13–16, 2023, Stockholm, Sweden.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3568294.3580059

1 INTRODUCTION

Non-verbal communication is a crucial element of human-human interaction, and within non-verbal behaviour, tactile interaction takes up a unique position through its use of physical force to convey social signals. In human-human interaction, tactile interaction is considered very important [3, 15], but in HRI –often due to technical limitations— tactile interaction has been given less attention than other forms of interaction.

Early work already showed the benefits of including gentle physical touch in interactions with elderly users [10]. Beyond that, haptic interaction, and specifically interaction that conveys emotions, can have therapeutic power [17]. Affective touch helps with communication and evokes arousal [6]. Key to haptic interaction is the ability to recognise and correctly classify different touch events. Touch can express a wide range of meanings and intent, such as emotions, affection, support, care, and agreement [1], and while people can readily recognise these, robots' ability to do so is currently fairly limited. Correctly recognising touch also allows the robot to respond appropriately and is an important contributing factor to the perceived agency of the robot [7]. Flagg and MacLean [8], for example, demonstrated a system that could recognise 9 key affective touch gestures with the aim to create an emotionally intelligent system. To support the comparison of different haptic touch classification methods, a dataset was collected that covered various social touch gestures by artificial skin: the Human-Animal Affective Robot Touch (HAART) database consists of seven touch gestures: no touch, constant, pat, rub, scratch, stroke, and tickle [14].

Several classification methods, some specific to classifying haptic gestures and some more generic, have been trained and evaluated on this dataset. Examples are multi-boost, logistic regression [12], random forest, and support vector machines classifiers [4, 9, 16]. Recently, a variety of Deep Learning algorithms have been used for classification with this dataset. A Convolutional Neural Network (CNN) achieved 83.2% classification accuracy [1] and a 3D CNN methods achieved 76.1% classification accuracy [19]. Since we aim to use the classification results for deciding which feedback to give to a user through a Nao, we have to find a balance between the classification accuracy and the response time, as sometimes it is better to sacrifice accuracy for a faster response.

In this paper, we present a Transformer-based Touch Gesture Classification model (TGC) to achieve a high accuracy that outperforms other classification algorithms on the HAART dataset. In addition, it has low latency and as such fits the requirements of our application.

[*] Equal contribution and joint first authors. Qiaoqiao Ren is corresponding author.

HRI '23 Companion , March 13–16, 2023, Stockholm, Sweden.

ACM ISBN 978-1-4503-9970-8/23/03…$15.00
https://doi.org/10.1145/3568294.3580059

2 METHODOLOGY AND MATERIALS

2.1 Datasets

The HAART dataset[1] consists of seven collected touch gestures from ten participants in twelve conditions performed on the same sensor that was installed on a robotic animal. Each touch lasts 10 seconds, and the data is capture at 54 frames per second (FPS). This dataset includes an 8 by 8 frame with pressure values ranging from 0 to 1023 as a CSV file.

2.2 Transformer

The Transformer is a novel architecture first proposed in [18], which adopts the self-attention mechanism, differentially weighting the significance of each part of the input data. It is heavily used in sequence-to-sequence tasks, such as machine translation. We use self-attention to capture the contextual relationship between one frame's tactile data in the total frame length, which is called the multi-head attention block, as we use multiple attention vectors. Then a simple feed-forward neural network is applied to every attention vector to transform the attention vectors into a form acceptable for the next encoder layer. Neural attention mechanisms can allow networks to focus on a subset of their inputs (or features) and through this select specific inputs. The attention mechanism can be applied to any input, regardless of its shape. In cases of limited computational power, the attention mechanism is a resource allocation scheme that is the main means of solving the information overload problem by allocating computational resources to more important tasks. The attention mechanism is used to dynamically generate weights for different connections, which is called the self-attention model. Since the weights of the self-attention model are generated dynamically, it can handle different length sequences of information.

2.3 Proposed model

The HAART dataset contains 432 different recordings of seven types of touch gestures. Each recording is sampled at a rate of 54 Hertz and each data point has a duration of 8 seconds. As the sensor is a 2D-matrix, together with the time dimension, the data is three-dimensional. Since the input for the model has to be of an equal size, we can use the raw sensor data as input for touch gesture classification. We propose to use the encoder block in Transformer for touch gesture classification, as the encoder can learn physical feature embeddings from the sensor data. After being processed by the encoder block, we use a multi-layer perception as a classifier. The outcome of the model is fed to the Nao robot, to give the robot the ability to sense different forms of touch gestures.

Compared to Convolutional Recurrent Neural Networks (CRNN), the Transformer-based models can capture long-term dependencies across the image sequences and process the spatial-temporal features simultaneously. We first benchmark the Transformer-based models on a HAART dataset for classification. Following that, we show that the proposed Transformer-based Touch Gesture Classification model (TGC) outperforms the existing models in the RTGC 2015 with regard to classification accuracy and computational efficiency. Our final proposed TGC model is a Transformer-based model with an encoder followed by three dense layers. Our code is publicly available on GitHub -https://github.com/Yuanbo2020/TGC.

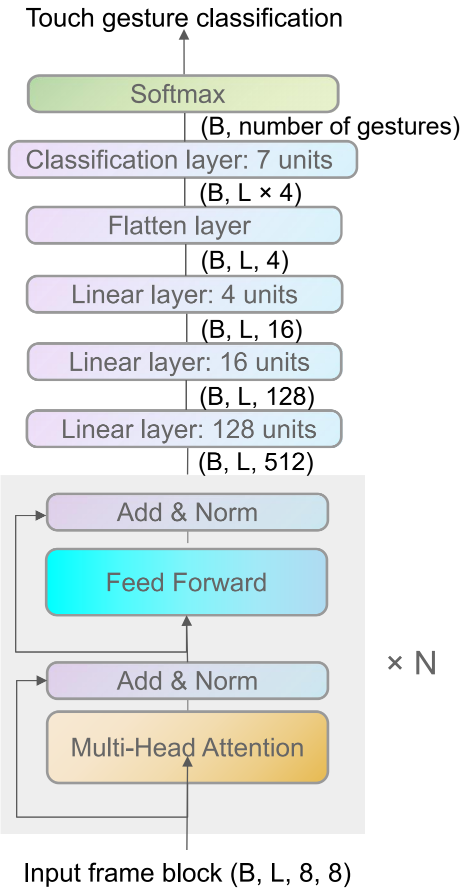

Figure 1: The proposed model GTC.

The main challenge is to classify the touch gesture in a short time while still maintaining an acceptable classification accuracy. Each recording of a touch gesture lasts 8 seconds at a sampling rate of 54 Hz. However, waiting until the end of the recording before deciding on which touch category is being sensed would mean that the robot's response is unnaturally slow. So ideally we prefer a fast response over an accurate response, which allows to robot to quickly react to haptic events. So rather than using the full recording (8 ×8 ×432) input to the Transformer, we study whether shorter recordings would still provide sufficient accuracy and low latency.

The proposed TGC model architecture consists of a Transformer with an encoder; there are N identical blocks in the encoder, which consist of two sub blocks. The input frame block has a batch size B, frame length of the input data L, and 8×8 width and height for each sample. The encoder consists of N identical blocks with a multi-head attention layer and a feed-forward layer (FF) with layer normalization. The first sub block implements a multi-head self-attention mechanism which refers to the *Multi-Head Attention*

[1] https://www.cs.ubc.ca/labs/spin/data/

(MHA) in Fig. 1 and the second sublayer is a fully connected feed-forward network which is identical to *Feed Forward* (FF). Each of these two sublayers has a residual connection around it and a normalisation layer, abbreviated as *Add & Norm* in Fig. 1, these normalise the sum computed between the sublayer input and the output generated by the sublayer itself. All the parameters in MHA and FF use Transformer's default setting [18]. As shown in Fig. 1, there are four fully connected layers followed by the encoder. And the final classification layer uses softmax as the activation function. We first optimised the model using a grid search for the dimensions of the different layers. The TGC model with 6 blocks achieved the best performance, as shown in Tab. 2. In addition, we compared different frame lengths to get an optimal model balancing accuracy and classification response time.

We tried different frame lengths (9, 27, 54, 72, 108, 144, and 216 frames), which means that we will have a recording with a length of $(8 \times 8 \times frame\_len)$ frames or $(8 \times 8 \times frame\_len/54)$ seconds, which in essence makes the task similar to event recognition in short video sequences. Previous research aimed to achieve more accurate classification results, while ignoring the response time.At worst, we have to wait for the whole recording to be fed through the network. This takes 8 seconds in addition to the computing time needed for the classifier, which is too long for responsive human-robot interaction. We set ourselves the goal of achieving acceptable classification accuracy within 5 seconds.

DETAILS ON COMPUTER HERE: The model ran on an Intel(R) Xeon(R) CPU E5-2680 processor with a card Tesla V100 GPU for 100 epochs.

3 EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Earlier approaches

The random forest algorithm was the most popular algorithm in the RTGC 2015 challenge [4, 9, 16]. The highest accuracy obtained on the test set is 70.9% by using random forest. Other machine learning methodologies have also been explored, such as Support Vector Machines (SVMs) with 68.5% accuracy on the test dataset and multi-boosting with 64.5%. Deep learning methods, like CNN and RNN, had been applied too and got comparable results: CNNs obtained 56.1% test accuracy, and CRNN model and Autoencoder-RNN model both achieved 61% test accuracy [13].

Cang et al. [5] report up to 90.3% accuracy by using 20-fold cross-validation when including subject and condition labels as features. In addition, earlier studies used leave-one-subject-out cross-validation to evaluate their model performance, with an accuracy of up to 83.2% [2]. However, due to the various data divisions, conditions, and/or labels used for subject information, direct comparisons between the accuracies reported for the RTGC 2015 and accuracies reported on the HAART test set are not meaningful [14]. Therefore, we use the HAART Data Set in RTGC 2015, including the training data set, and test dataset provided by RTGC 2015 allowing us to make meaningful comparisons.

3.2 Experiment results

The input of our network received raw gesture sequences with 8×8 resolution with different frame lengths (9, 27, 54, 72, 108, 144, and 216 frames). The output reports one of 7 gestures corresponding to the gestures label. Our experiments are conducted on the HAART dataset of the RTGC 2015. We first explored the number of blocks (1 to 9) to find the optimal model architecture, the accuracy is shown in the Tab. 2. The optimal number of blocks in the encoder is 6. Based on this, we explored the test accuracy in function of the different lengths of recordings.

Tab. 3 shows that the proposed model classification accuracy is 72.79% with a frame length of 144, which means that the proposed model could give out the classification result in 2.67 seconds from the start of the gesture, which is much quicker than the typically reported results which need at least 8 seconds. Specifically, the optimal proposed model could get 70.8% and 71.4% test accuracy in 1 second (54 frames) and 1.5 seconds (72 frames), respectively, while the optimal classification accuracy (70.9%) obtained in the RTGC 2015 in 8 seconds.

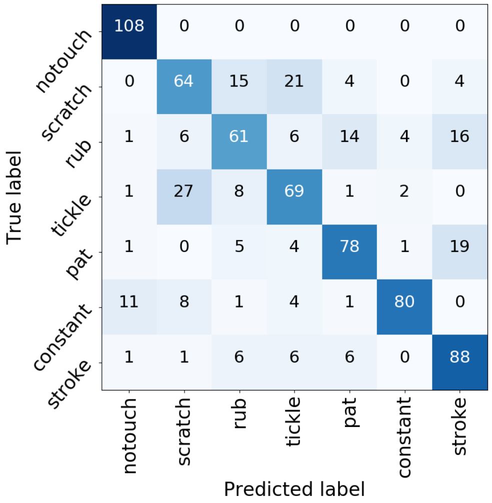

Figure 2: The confusion matrix output by the TGC model on the test data.

The confusion matrix output by the TGC model for recordings of length 2.67 seconds is given in Fig. 2. Unsurprisingly, *no touch* is classified not confused with any other label. In contrast, the *tickle* and *rub* gestures were harder to classify than other gestures. Most notably, *scratch* and *tickle* is easily confused in short-time gesture performance, even for human annotators. In addition, *rub* and *pat* were often misclassified as a *stroke.* Both classification errors can be explained by the gestures are quite similarly in the first two seconds. The sensor would output higher values on some point both for *rub*, *pat*, and *stoke.* In addition, gesture *scratch* is often misclassified as *rub.*

Gesture *constant* is a touch gesture occurring continuously over time, so it is easy to distinguish from other touch gestures. The distribution of the learned representation on the HAART dataset is visualized in Fig. 3 using t-distributed stochastic neighbor embedding (t-SNE). It is easy to observe that gesture *constant* is easily distinguished from other gestures. Gesture *tickle* is a light touch or prod to a part of the body, and as there are moments when there is no force during *tickle*, it is sometimes classified as *no touch.* The

Table 1: Touch gesture classification accuracy comparison of the proposed model against existing classification using HAART dataset

Researchers[reference]	Classification algorithm	Accuracy(%)
Balli Altuglu et al. [4]	Random forest	61.5
Ta et al. [16]	Random forest	70.9
Ta et al. [16]	SVM	68.5
Gaus et al. [9]	Random forest	66.5
Gaus et al. [9]	Multiboosting	64.5
Hughes et al. [13]	CNN	56.1
Hughes et al. [13]	CRNN	61.4
Hughes et al. [13]	Autoencoder-RNN	61.4
Proposed TGC	Transformer-based	**72.8**

Table 2: Test accuracy comparison for different numbers of blocks

N	1	2	3	4
Accuracy	65.6 ± 2.3%	65.9 ± 4.8%	67.9 ± 1.6%	68.6 ± 2.0%
N	5	6	7	8
Accuracy	68.9 ± 1.8%	**69.8±2.2%**	69.5 ± 0.8%	69.4 ± 0.9%

Table 3: Test accuracy comparison for different frame length

Input length (second)	0.17 second	0.5 second	1 second	1.5 seconds	2 seconds	2.7 seconds	4 seconds
TGC model test accuracy	57.0±1.9%	66.6±1.3%	70.8±1.5%	71.1±1.4%	72.3±1.2%	**72.8±1.8%**	71.6±1.0%

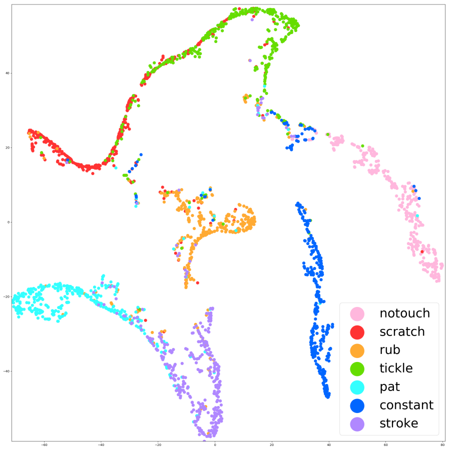

Figure 3: Visualization of the distribution of the learned representation by the TGC on the HARRT dataset using t-SNE [11].

stroke gesture is easy to confuse with *pat*, as in a *stroke* the hand is placed on the sensor surface and slowly moved along the surface, and in *pat* the hand is tapped quickly on the sensor, there is a fair bit of overlap between the two, as seen in the t-SNE plot. Similarly, *scratch* and *tickle* are easily confused. Finally, *no touch* is sometimes confused with *constant* as the sensor sometimes reports non-zero values which have not been calibrated away.

4 DISCUSSION

We propose a system for classifying touch gestures using a TGC model which generalises across different subjects. The proposed TGC system yields an accuracy of 72.8% within 2.7 seconds. The proposed TGC model has a comparative classification accuracy without any preprocessing or manual feature extraction, and as such implements an end-to-end classifier. When faster classification response are needed, our model obtained 70.77% and 71.14% classification accuracy within 1 second and 1.5 seconds, respectively. This is very competitive when compared to the optimal result from RTGC 2015 challenge, which achieved 70.9% classification accuracy in 8 seconds. In conclusion, the proposed model outperforms the state-of-the-art algorithms from the RTGC 2015 and holds promise for building responsive and accurate haptic interaction with robots. Furthermore, our model allows for a fast and perhaps inaccurate classification, allowing the robot to provide an early response, while still allowing for the classifier to correct to a more accurate response as more data frames come in. We believe that a low latency response will lead to higher perceived agency and eventually lead to a higher quality interaction.

REFERENCES

[1] Mohammed Ibrahim Ahmed Al-Mashhadani, Theyazn HH Aldhyani, Mosleh Hmoud Al-Adhaileh, Alwi M Bamhdi, Mohammed Yahya Alzahrani,

Fawaz Waselallah Alsaade, and Hasan Alkahtani. 2021. Human-Animal Affective Robot Touch Classification Using Deep Neural Network. *Comput. Syst. Sci. Eng.* 38, 1 (2021), 25–37.

[2] Saad Albawi, Oguz Bayat, Saad Al-Azawi, and Osman N Ucan. 2018. Social touch gesture recognition using convolutional neural network. *Computational Intelligence and Neuroscience* 2018 (2018).

[3] Hasan Alkahtani, Theyazn HH Aldhyani, and Mohammed Al-Yaari. 2020. Adaptive anomaly detection framework model objects in cyberspace. *Applied Bionics and Biomechanics* 2020 (2020).

[4] Tugce Balli Altuglu and Kerem Altun. 2015. Recognizing touch gestures for social human-robot interaction. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 407–413.

[5] Xi Laura Cang, Paul Bucci, Andrew Strang, Jeff Allen, Karon MacLean, and HY Sean Liu. 2015. Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 147–154.

[6] Angela Chan, Francis Quek, Haard Panchal, Joshua Howell, Takashi Yamauchi, and Jinsil Hwaryoung Seo. 2020. The Effect of Co-Verbal Remote Touch on Electrodermal Activity and Emotional Response in Dyadic Discourse. *Sensors* 21, 1 (2020), 168.

[7] Jonathan Chang, Karon MacLean, and Steve Yohanan. 2010. Gesture recognition in the haptic creature. In *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. Springer, 385–391.

[8] Anna Flagg and Karon MacLean. 2013. Affective touch gesture recognition for a furry zoomorphic machine. In *Proceedings of the 7th International Conference on Tangible, Embedded and Embodied Interaction*. 25–32.

[9] Yona Falinie A Gaus, Temitayo Olugbade, Asim Jan, Rui Qin, Jingxin Liu, Fan Zhang, Hongying Meng, and Nadia Bianchi-Berthouze. 2015. Social touch gesture recognition using random forest and boosting on distinct feature sets. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 399–406.

[10] Marcel Heerink, Ben Krose, Vanessa Evers, and Bob Wielinga. 2006. The influence of a robot's social abilities on acceptance by elderly users. In *ROMAN 2006-The 15th IEEE International Symposium on Robot and Human Interactive Communication*. IEEE, 521–526.

[11] Geoffrey E Hinton and Sam Roweis. 2002. Stochastic neighbor embedding. *Advances in neural information processing systems* 15 (2002).

[12] Dana Hughes, Nicholas Farrow, Halley Profita, and Nikolaus Correll. 2015. Detecting and identifying tactile gestures using deep autoencoders, geometric moments and gesture level features. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 415–422.

[13] Dana Hughes, Alon Krauthammer, and Nikolaus Correll. 2017. Recognizing social touch gestures using recurrent and convolutional neural networks. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2315–2321.

[14] Merel M Jung, Xi Laura Cang, Mannes Poel, and Karon E MacLean. 2015. Touch challenge'15: Recognizing social touch gestures. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 387–390.

[15] Merel M Jung, Lisa Van der Leij, and Saskia M Kelders. 2017. An exploration of the benefits of an animallike robot companion with more advanced touch interaction capabilities for dementia care. *Frontiers in ICT* 4 (2017), 16.

[16] Viet-Cuong Ta, Wafa Johal, Maxime Portaz, Eric Castelli, and Dominique Vaufreydaz. 2015. The Grenoble system for the social touch challenge at ICMI 2015. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. 391–398.

[17] Puja Varsani. 2018. *The power of affective touch within social robotics*. Ph. D. Dissertation. Middlesex University.

[18] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[19] Nan Zhou and Jun Du. 2016. Recognition of social touch gestures using 3D convolutional neural networks. In *Chinese Conference on Pattern Recognition*. Springer, 164–173.