TOUCH GESTURE CLASSIFICATION USING TACACT DATASET

EXPLORATORY ANALYSIS ON SPATIAL AND TEMPORAL PROCESSING OF TOUCH GESTURE DATA

DORIS ADRIANA MARIA WEZENBERG

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF SCIENCE IN COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE

DEPARTMENT OF
COGNITIVE SCIENCE & ARTIFICIAL INTELLIGENCE
SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
TILBURG UNIVERSITY

STUDENT NUMBER

462228

COMMITTEE

dr. M. Jung
dr. T. Wiltshire

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

June 24, 2022

ACKNOWLEDGMENTS

I would like to thank Merel Jung for guiding me through this research project. Merel has provided me with the building blocks to conduct this study, showed me where to exploit research opportunities and has given me valuable insights on how to approach different challenges. Furthermore, I would like to thank the authors of the TacAct dataset, for creating such an elaborate dataset and hereby allowing for novel research opportunities in the field of touch gesture recognition. Conclusively, I would like to thank my family and friends for their valuable support and encouragement.

TOUCH GESTURE CLASSIFICATION USING TACACT DATASET

EXPLORATORY ANALYSIS ON SPATIAL AND TEMPORAL PROCESSING OF TOUCH GESTURE DATA

DORIS ADRIANA MARIA WEZENBERG

Abstract

Recent developments in the field of natural language processing (Miller, Bernstein, & McDaniel, 2021) and image recognition (Bianchini, Verma, & Salisbury, 2021) have increased the feasibility of the implementation of social robots in society. Touch gesture recognition is one of the modalities in physical human robot interaction (pHRI) that still demands improvement. Current recognition accuracy for touch gesture data on datasets with more than ten different gesture types is relatively low. For the Corpus of Social Touch (Jung, Poel, Poppe, & Heylen, 2014), a dataset containing 7805 instances of 14 different touch gestures, accuracy scores do not exceed 64% (Albawi & Ucan, 2018) (Li, Meng, & Zhang, 2022). Wang et al. (2021) created the TacAct dataset containing 24000 instances of 12 different touch gestures. To asses in which manner touch gesture data is processed and classified most effectively using a larger dataset, touch gesture data of the TacAct dataset is processed spatially and temporally in this study. Spatial processing by the extraction of keyframes has shown to be an effective method to represent touch gesture data. An accuracy of 81.5% is achieved using a Random Forest classifier, being relatively low in computational complexity compared to other methods. The accuracy of this model beats accuracy scores on comparable datasets with a similar number of gesture types and competes with the state-of-the-art results on the TacAct dataset. In summary, spatial processing shows a larger potential compared to temporal processing in classification accuracy compared to computational complexity. Further research opportunities addressing class distinction of touch gestures should be exploited to make direct inference on the validity of a specific approach.

1 INTRODUCTION

Recent years have seen major developments for the field of social robotics. The field of social robotics has emerged only recently, and it refers to the field in which the robot's goal is to "engage in an affective or otherwise helpful interaction" (Sheridan, 2020, p. 7) with another human. Several modalities in social robotics have seen great improvements in the past decade. Recent developments in image recognition have increased the possibilities for the visual modality (Bianchini et al., 2021) and advances in the field of natural language processing (NLP) have contributed greatly to the possibilities for the auditory modality (Miller et al., 2021). Consequently, physical human-robotics interaction (pHRI) is becoming more natural and implementation of robots in social interaction is becoming increasingly feasible. One of the modalities in social interaction between humans and robots that still demands improvement is the touch modality. Social touch (or: touch gesture, or: tactile emotion) is defined as interpersonal touch that is used to convey emotional or functional intents (Cascio, Moore, & McGlone, 2019). Touch gesture recognition presents itself as a prospective modality for pHRI. Research shows that touch gestures are essential in communicating emotional and functional intentions (Hertenstein, Holmes, McCullough, & Keltner, 2009). Furthermore, emotional valence may be detected more effectively by interacting via social touch compared to other modalities in pHRI (Horii, Nagai, & Asada, 2018). From a psychological perspective, several studies have shown the importance of physical touch for human wellbeing (Lee & Cichy, 2020), (Field, 2010), (Garcini et al., 2019). Consequentially, as technology continues to become increasingly intertwined with the daily interactions with our environment, touch gesture recognition could be highly valuable for the field of pHRI. It could provide opportunities for applications in smart homes (Miller et al., 2021) serving healthcare, education and other purposes. Although promising, touch gesture recognition presents itself as a difficult problem in social robotics. The lack of data due to the difficulty to collect touch data and the complexity of the data due to its spatiotemporal nature are two of the reasons why research in this field is not yet as advanced as for other modalities related to pHRI.

The TacAct dataset (Wang et al., 2021) is a recently published dataset containing tactile data of twelve different touch gestures. The publication of this dataset allows for novel research opportunities, as it contains roughly three times as much data as the Corpus of Social Touch (Jung et al., 2014), a dataset on which most previous research on touch gesture recognition has focused. The aim of this research is to explore two different approaches

to classify touch gestures using the TacAct dataset, and to compare the results to previous research on the CoST and the TacAct dataset.

Computational complexity creates difficulties for spatio-temporal processing of touch gesture data, particularly considering the scope of this study and the instrumental capacity. A spatio-temporal approach to classify touch gesture data could potentially be the most effective as it combines spatial as well as temporal information, but it brings high computational costs. If spatial or temporal processing would independently be sufficient to process touch gesture data, this could have positive implications for the field of touch gesture recognition as it would reduce computational complexity for touch gesture processing. As a result, this study focuses on classifying touch gestures by exploiting the temporal and the spatial aspect of the data respectively. The following research question is approached in this study:

> *To what extend are the spatial and the temporal aspect of touch gesture data respectively predictive for touch gesture classification?*

Spatial processing is assessed by considering only a single frame per gesture, which is the frame with the highest sum of the pressure values. Temporal processing is analysed by using different metrics over a number of frames per gesture, in combination with spatio-temporal feature extraction. Two classification algorithms are selected based on previous research in the field and taking into account the scope of this study. The performance of the two different approaches is evaluated based on accuracy and the performance across classes is considered by evaluating the f1 scores. Computational complexity is also taken into account in a qualitative manner. This study should give researchers in the field of touch gesture recognition more insight in where to lay the focus for data collection, representation and processing by exploring the predictive value of the spatial and the temporal aspect of touch gesture data respectively and comparing the results to the state-of-the art results on the CoST and the TacAct dataset.

The Random Forest classifier using the raw data from the keyframes as input outperforms the accuracy scores that are achieved with the temporal approach with an accuracy of 81.5%. This demonstrates the effectiveness of spatial processing of touch gesture data. The computational complexity is relatively low, as no pre-processing technique is applied and only one frame per gesture is used. Findings point out that the information from temporal data on the mean and the sum of the pressure values is less predictive, as accuracy scores do not exceed 60%. This could indicate that classification accuracy is higher when more importance is given to the spatial aspect of the data in contrast to the temporal aspect of the data,

and that this is also beneficial for computational complexity. The accuracy scores of the spatial approach for touch gesture classifiaction of the TacAct dataset surpass the state of the art results on the CoST dataset and compete with the results on the TacAct dataset of the LeNet-5 model of Li et al. (2022).

In summary, a spatial and a temporal approach to classify a novel larger data set of touch getures have been explored and compared to each other and to previous research. The spatial approach appears to be an effective solution for touch gesture classification as it achieves a relatively high accuracy with a lower computational complexity compared to a temporal approach. In addition to this, the classification accuracy of the spatial approach is relatively high compared to previous research on the CoST dataset and the state of the art results for the TacAct dataset.

2 RELATED WORK

2.1 *Different types of social touch*

Social touch (or: touch gesture, or: tactile emotion) is defined as interpersonal touch that is used to convey emotional or functional intents (Cascio et al., 2019). Research shows that basic emotions such as fear, anger and disgust, as well as more complex messages such as trust can be communicated through social touch (Hertenstein, Keltner, App, Bulleit, & Jaskolka, 2006). Due to the qualitative nature of social touch, it can be stated that an infinite number of messages can be communicated through this channel. Similarly, the way in which a message can be communicated can also take many different forms. There exists variability along two axes; the variability in the types of emotions that are being communicated which influence how the gesture is executed, and the variability across humans in the way in which the same emotions are communicated through a physical action. Consequentially, the large variability related to touch gestures makes it difficult to distinguish and group different types of social touch. For the field of pHRI, distinguishing as well as grouping touch gestures based on certain characteristics is a prerequisite for the classification of touch gestures. While for the TacAct dataset 12 different gesture types are considered (Li et al., 2022), other researchers take a different approach and consider a different number of gesture types (e.g. 7 different gesture types for the HAART dataset (Cang et al., 2015) versus 14 different gesture types for the CoST dataset (Jung et al., 2014).

Comparing the defined gesture types for the CoST and the TacAct dataset (table 1), it shows that both authors distinguished individual categories for the gesture types 'poke', 'scratch', 'hit' and 'squeeze'. In

Table 1: Overview of the different gesture types for the CoST and the TacAct dataset respectively.

CoST	TacAct
grab	pull
hit	squeeze
massage	push
pat	hold
pinch	grasp
poke	poke
press	static drag
rub	strongly hit
scratch	soft slide
slap	scratch
squeeze	soft hit
stroke	sliding drag
tap	
tickle	

addition to this, both considered some form of grabbing, 'grabbing' vs. 'grasping'. Other than this, both studies consider relatively different types of gestures. These two studies show that there is little conformity with regards to defining different gesture types in the field of touch gesture recognition. The variability in defining classes should be taken into account when evaluating and comparing research in the field of touch gesture recognition.

2.2 *Classifying social touch*

To implement touch gesture recognition in social robots, it is necessary for computers to be able to classify touch gestures. In order to apply a machine learning approach for this classification problem, data must be collected. Different methods have been used to collect touch gesture data, such as tactile sensors (Jung, Poel, Poppe, & Heylen, 2016), (Altun & MacLean, 2015), wearable devices (Blumrosen, Sakuma, Rice, & Knickerbocker, 2020) and touch screen (Ghosh, Hiware, Ganguly, Mitra, & De, 2019). In 2015, Cang et al. published the HAART dataset, containing touch data on 7 different gestures. Using the Random Forest algorithm, 90% accuracy was achieved. Jung et al. (2014) created the Corpus of Social Touch, a dataset containing 7805 instances of 14 different touch gestures using a tactile sensor. This dataset enabled researchers to analyze touch gesture data with more gestures and improved the state of the art in the field of touch recognition. In table 2, an overview is presented of the existing

classification models for the CoST dataset. The accuracy varies from 42% (Altun & MacLean, 2015) to 64% (Albawi & Ucan, 2018) and shows a great variety in the number of features extracted (e.g. 273 by Ta, Johal, Portaz, Castelli, and Vaufreydaz (2015) compared to none by Albawi and Ucan (2018)) and the applied classification method. Contrasting methods, such as machine learning models with complex feature techniques (Ta et al., 2015) and deep learning models without preprocessing (Albawi & Ucan, 2018), are both able to achieve accuracy scores slightly above 60%. For the HAART dataset, generalizability showed to be insufficient due to the few different gestures, while for the CoST dataset accuracy scores are relatively low (table 2). To provide researchers in the field with a larger dataset and a sufficient number of gesture types, Wang and Chen created the TacAct dataset. This dataset consists of 24000 instances of 12 different touch gestures, containing roughly three times as much data as the CoST dataset. Up to this point, the only research that has been published using the TacAct dataset is presented by the authors of the dataset. The Convolutional Neural Network (CNN) named LeNet-5 including an input layer, three convolutional layers, two pooling layers, a fully connected layer and an output layer proved most effective (Wang et al., 2021). The highest accuracy of 95.46% was achieved by using a frame length of 80 frames, although this model brought high computational costs.

3 METHOD

3.1 *Dataset description*

The TacAct dataset (Wang & Chen, 2021) contains tactile data on twelve different touch gestures, namely pull, squeeze, push, hold, grasp, poke, static drag, strong hit, soft slide, scratch, soft tap, and sliding drag (table 1). The dataset is balanced, meaning that frequencies are evenly distributed for all classes. The data was collected using a pressure sensor on a robotic arm. The data was recorded on a 32 by 32 tactile sensor grid at 100 Hz. 50 participants were represented in the dataset and for each gesture type the action was repeated 40 times, resulting in 480 actions per participant and a total of 24.000 actions. Each gesture recording includes a variable number of frames and frames are captured only when a threshold is met to indicate that the action has started. The data are spatiotemporal, meaning that the data relate to both space and time. The dataset also includes extracted keyframes for all gesture recordings, representing the frame with the maximum total pressure value.

Table 2: Chronological comparison of existing classification methods applied on CoST dataset. Adapted from (Li et al., 2022) and (Albawi & Ucan, 2018).

Reference	Features	Classifier	Accuracy (%)	SD (%)
(Jung, 2014)	28	Bayesian	53	N/A
		SVM[1]	46	N/A
(Jung et al., 2014)	28	Bayesian	54	12
		SVM	53	11
(van Wingerden, Uebbing, Jung, & Poel, 2014)	45	Neural Network	54	15
(Altun & MacLean, 2015)	42	Random Forest	56	13
(Gaus et al., 2015)	5 set	Random Forest	59	N/A
		Boosting	58	N/A
(Hughes, Farrow, Profita, & Correll, 2015)	7	Deep Autoenc.	56	N/A
(Ta et al., 2015)	273	SVM	61	N/A
		Random Forest	61	N/A
(Jung, 2017)	54	Bayesian	57	11
		Decision Tree	48	10
		SVM	60	11
		Neural Network	59	12
(Hughes, Krauthammer, & Correll, 2017)	Raw Data	CNN	42	11
	Raw Data	CNN-RNN[2]	53	N/A
	7	Deep Autoenc.	34	N/A
(Albawi & Ucan, 2018)	**Raw Data**	**CNN**	**64**	**12**
(Wei, Liu, Wang, & Sun, 2019)	2 set	ELM[3]	61	N/A
		CNN	42	N/A

1. Support Vector Machine 2. Recurrent Neural Network 3. Extreme Learning Machine

3.2 *Classifiers*

Previous research shows that out of the existing classification methods for the CoST dataset, the Convolutional Neural Network (CNN) of Albawi and Ucan (2018) achieved the highest accuracy (64%) out of all classifiers. In addition to this, the Random Forest model of Ta et al. (2015) also achieved a relatively high accuracy of 61%. Hereby, it can be concluded that a Random Forest algorithm as well as CNN are suitable architectures for this type of classification problem. While Random Forest is a tree-based classifier, the CNN is a specific type of artificial neural network. Both algorithms are of a different type and both are suitable for touch gesture classification. By taking these two characteristics into account and also considering the scope of this study, a Random Forest and a multi-layer perceptron (MLP) algorithm using back-propagation will be used in this study to classify touch gestures and assess the effectiveness of the different approaches. By choosing two different classifiers to evaluate the approaches, any induced bias related to the selected classification algorithm should be avoided as much as possible.
The parameter grid that is used for optimization for the Random Forest includes:

- **bootstrap**: True, False
- **max_depth**: 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, None
- **max_features**: 'auto' and 'sqrt'
- **min_samples_leaf**: 1, 2, 4
- **min_samples_split**: 2, 5, 10
- **n_estimators**: 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800, 2000

For the MLP architecture, the search is performed over the following parameters:

- **hidden_estimators layer_estimatorssizes**: 10,20,30, 100, 200
- **activation**: 'tanh' and 'relu'
- **solver**: 'sgd' and 'adam'
- **alpha**: 0.0001, 0.05
- **learning_estimatorsrate**: 'constant' and 'adaptive'

3.3 *Spatial processing*

To evaluate the predictiveness of the spatial aspect of the data, a single frame per gesture is considered. This frame is referred to as the keyframe of a gesture and it is selected based on the highest total pressure value. To process the raw data, the 32x32 two-dimensional pressure grid of the keyframe is converted to a two-dimensional 1x1024 array. The data are split into a train (80%) and a test (20%) set, using stratified sampling to ensure class balance is maintained. Splitting is done user-independently, to promote generalizability of results. Consequently, both a Random Forest classifier and a MLP are used to classify the data. The models are optimized using grid search with cross validation.

3.4 *Temporal processing*

The importance of the temporal aspect of touch gesture data is assessed by considering the mean and sum of the pressure values of 20 frames, corresponding to 200 milliseconds. Both the mean and the sum are considered in this approach as both metrics capture information from the entire set of data and the use of two different metrics helps to avoid bias related to the selected metric. A fixed number of 20 frames per gesture is chosen to analyze the gestures, which is motivated by the trade-off between information and computational complexity, as explored in (Wang et al., 2021). The model of (Wang et al., 2021) with a frame length of 20 proves effective and achieves an accuracy of approximately 88%, while the slope of the function of the accuracy compared to the number of frames decreases strongly after 20 frames. Zero padding is applied to process gestures with frame sequences of less than 20 frames. Spatiotemporal features are extracted using the TSFresh library. TSFresh is a Python package that provides several methods to extract time series features from data (Henderson & Fulcher, 2021), such as for example the kurtosis, the permutation entropy or whether there exists duplicates in the time-series data. The data is split into a train (80%) and test (20%) set. Stratified sampling is applied to ensure class balance is maintained. Splitting is done user-independently, to promote generalizability of results. The Random Forest and the MLP are both optimized using grid search with cross validation.

3.5 *Evaluation*

Evaluation of the two approaches is based on classification accuracy, as classes are balanced. To break down the accuracy scores and asses performance across classes, f1 scores are evaluated for individual classes for

the best performing classifier, considering both precision and recall. Computational complexity is also considered and evaluated in a qualitative manner. The accuracy scores of the two classifiers for the spatial as well as the temporal approach will be compared to the state of the art results on the CoST dataset, the TacAct dataset and to each other.

4 RESULTS

4.1 *Spatial processing*

The Random Forest model scored a training accuracy of 1, indicating a possible risk of over-fitting. Parameter tuning with cross validation resulted in a test accuracy of 81.5% (table 3). With regards to the number of estimators, 600 estimators were used for optimal results. The minimum samples required for a split was set to 5 and 2 minimum samples were required at a leaf. The maximum depth was set at 70 and no bootstrapping was applied. The MLP classifier achieved a test accuracy of 76.9%. The Rectified Linear Unit (ReLu) function was used as activation function, the optimal learning was found to be 0.0001 and set to constant, the hidden layer sizes were set to 200 and the Adaptive Moment Estimation (Adam) algorithm was used for optimization. The Random Forest classifier scored the lowest F1 scores for the classes 'strong hit', 'soft hit' and 'soft slide' (table 4).

4.2 *Temporal processing*

Using TsFresh, 372 spatiotemporal features were extracted from the data. As depicted in table 3, an accuracy of 58.2% was achieved with Random Forest using the mean pressure value per frame. Parameters were optimized using grid search with cross validation. For optimal results, 600 estimators were used. The minimum number of samples required for each split was set to 5, 2 minimum samples were required at a leaf, maximum depth was set at 70 nodes and no bootstrapping was used. For the sum of the pressure values, Random Forest achieved an accuracy of 58.8%. The best parameters were: 100 estimators, 5 minimum samples required to split, 1 minimum sample for a leaf, a max depth of 70 and no bootstrapping. For the MLP architecture, an accuracy of 48.6% was achieved using the mean of the pressure values for each frame. The ReLu activation function was found to be optimal out of the search grid. The learning rate was constant and set to 0.001, hidden layers size was set to 30, and the weights were optimized using the Adam optimization algorithm for stochastic gradient decent. Using the sum of the pressure values per frame, an accuracy score

Table 3: Overview of the different pipelines used to classify touch gestures of the TacAct dataset

Data	Pre-processing	Feature extraction	Features	Model	Accuracy (%)
Keyframe	**None**	**Raw data**	**1024**	**Random Forest**	**81.5**
Keyframe	None	Raw data	1024	MLP	76.9
20 frames	Mean	TsFresh	372	Random Forest	58.2
20 frames	Mean	TsFresh	372	MLP	48.6
20 frames	Sum	TsFresh	372	Random Forest	58.8
20 frames	Sum	TsFresh	372	MLP	31.2

Table 4: F1 scores per gesture type for the Random Forest classifier for the spatial and the temporal approach respectively

	Spatial	Temporal
Gesture type	F1 score	F1 score
Pull	0.81	0.37
squeeze	0.86	0. 56
push	0.87	0.55
hold	0.81	0.56
grasp	0.79	0.43
poke	0.87	0.78
static drag	0.82	0.40
strongly hit	0.74	0.92
soft slide	0.77	0.56
scratch	0.86	0.51
soft hit	0.72	0.94
sliding drag	0.86	0.36

of 31.2% was achieved using a MLP architecture. Grid search with cross validation resulted in ReLu as the activation function, a constant learning rate of 0.0001, a hidden layer size of 30 and the Adam algorithm for optimization. The Random Forest model scored the lowest F1 score for the gesture types 'pull' and 'sliding drag' (table 4).

5 DISCUSSION

The Random Forest model using keyframes achieved the highest accuracy out of all compared methods. In addition to this, the computational complexity is relatively low, as only one frame per gesture is considered. Comparing these results with the LeNet-5 CNN of Wang et al. (2021) that considered spatio-temporal information from the TacAct dataset, it shows

that the CNN of Wang et al. (2021) is able to achieve an accuracy of around 88% using 20 frames per gesture and accuracy improved up to 95.46% using 80 frames. Using keyframes and hereby only considering the spatial aspect of the data, an accuracy of 81.5% is achieved by using only a single frame and a Random Forest classifier. By using keyframes, a high accuracy is achieved with less frames compared to previous research on the TacAct dataset. In addition to this, the results of the spatial approach surpass the state of the art results on the CoST dataset. While for the CoST dataset accuracy scores did not exceed 64%, an accuracy of 81.5% is achieved on the TacAct dataset. In contrast to most spatio-temporal approaches for the CoST dataset, this approach only considered spatial information and did not apply any deep learning techniques. These findings indicate a positive effect for classification accuracy that can be attributed either to the spatial approach or to the TacAct dataset. Further research should investigate whether the spatial approach using keyframes is as effective for other datasets as it is for the TacAct dataset, and wether techniques that were used on the CoST dataset perform even better on the TacAct dataset. Hereby, direct inference can be made on the value of a larger dataset and the potential of extracting keyframes respectively.

In attempting to assess the predictiveness of the spatial and the temporal aspect of touch gesture data respectively, the two approaches are compared in this study. However, caution must be taken in this comparison for several reasons, as multiple decisions are made in this study that could influence the results and hereby induce bias in the study. These decisions mainly include the selection of 20 frames per gesture and the decision to use TsFresh for spatio-temporal feature extraction. Although these decisions are based on thorough evaluation, it might be the case that different decisions could have lead to different results. Further research should point out the validity of these decisions and the results.

It is unclear to what extend the different characteristics of the TacAct and the CoST dataset influence the results. As there exists a large variability in how touch gesture data is recorded, how classes are defined and how touch gestures are executed, there may exists large differences between the two datasets. These difference might have consequences that influence recognition accuracy. Therefore, further research should asses the generalizability of the results on the TacAct dataset. This should be taken into account when comparing the results to previous research in the field.

Classification on the keyframes of the TacAct dataset shows the lowest f1-score for classes 'strong hit' and 'soft hit' for the spatial approach. By contrast, for the temporal approach these classes show by far the highest f1-scores (table 4). These are the only two classes in the dataset that are not of

fixed length and contain substantially less frames. However, classification on keyframes should not be affected by the number of frames for a gesture, while the temporal approach could be negatively affected by a shorter time period. These results are counter intuitive and further research should point out which aspect of these gesture types results in their deviating f1-scores compared to other classes.

6 CONCLUSION

For the spatial approach, the optimized Random Forest classifier achieves an accuracy of 81.5% and the MLP achieves a slightly lower accuracy of 76.9%. The computational complexity is relatively low, as no pre-processing is applied and only one frame per gesture is used, resulting in 1024 features using the raw data. Generally, it can be stated that a spatial approach is an effective technique to classify touch gestures, as it is relatively low in computational complexity compared to other techniques, and accuracy scores are relatively high using machine learning models. The Random Forest classifier shows more effective than the MLP architecture.

The accuracy scores for the temporal approach are lower compared to the accuracy scores for the spatial approach. The Random Forest classifier achieves an accuracy of 58.2% considering only the temporal aspect of the data using the mean of the pressure values, and a comparable accuracy of 58.8% using the sum of the pressure values. For the MLP, accuracy scores were also lower for the temporal approach compared to the spatial approach. Moreover, while for the temporal approach only 372 features are extracted, extensive pre-processing is applied to extract the spatio-temporal features from the frames, which is a process that is relatively high in computational complexity. By this, it can be concluded that considering a computational acceptable number of frames and the mean or sum as a metric, the temporal approach is outperformed by the spatial approach and the spatial characteristics show a higher predictive value. By only considering a spatial approach, accuracy scores are relatively high for the TacAct dataset, indicating that a spatial approach to touch gesture data by extracting keyframes is an effective technique when computation is constrained. Further research should point out to what extend the spatial approach of using keyframes is also effective on comparable datasets and hereby evaluate to what extend the results of this study are generalizable.

It is difficult to conclude this research with a definitive answer to the extend to which the spatial and temporal aspect of touch gesture data are respectively predictive for touch gesture classification. Findings have shown that for the TacAct dataset, the spatial aspect is predictive up to a classification accuracy of 81.5% using a Random Forest classifier. The

temporal approach showed that gestures can be classified by focusing on the temporal aspect of a gesture, but only up to an accuracy of 58.2% using the mean and 58.8% using the sum with a Random Forest classifier. However, it should be taken into account that the temporal approach required more processing steps and more decisions, by which bias could be induced in the approach and which could have influenced the results. Generally, findings point out that the spatial aspect is more predictive for touch gesture recognition on the TacAct dataset compared to the temporal approach, as a higher accuracy is achieved with less pre-processing and by using less information from the gestures. To what extend these findings are generalizable, is up for discussion. Comparing the results of the spatial approach to other approaches for the TacAct dataset, shows that the spatial approach is indeed effective compared to the spatio-temporal approach of (Li et al., 2022). Comparable accuracy scores are achieved using a considerably lower number of frames and a machine learning model. Based on accuracy, results are also valuable compared to the CoST dataset. However, due to the lack of conformity in the definition of touch gestures, it is difficult to draw a conclusion on the effectiveness of the spatial approach using keyframes on touch gesture recognition in general.

7 LIMITATIONS AND FUTURE DIRECTIONS

For the field of pHRI and particularly for touch gesture recognition, it will be highly valuable to further investigate all dimensions along which humans interpret social gestures. In this research, the focus was centered on the physical aspect of touch gestures expressed in pressure values over time. However, the visual modality could also contribute to the interpretation of touch gestures in the form of facial expression or eye gaze. In addition to this, the auditory modality could intensify or change the meaning of a touch gesture in the form of shouting for example. By breaking down social interaction into different modalities, a bottom-up approach is taken and all components can be analyzed respectively. However, in distinguishing different gesture types, their meaning is essential and the physical channel is often not the only channel by which meaning is communicated. Integration of the different modalities that play a role in conveying meaning in touch gesture interactions could result in more concrete and wholesome definitions of touch gesture categories (even along a continuous spectrum instead of constrained by discrete categories). As a consequence, touch gesture data sets can be labeled in such a way that class definitions become more universal and data sets become more compatible. This could boost research in the field of touch gesture recognition as it improves generalizability of results. In addition to its potential for the definition

of touch gesture classes, the possibility to use information of different modalities to classify touch gestures could greatly benefit the classification accuracy as more information can be considered.

Exploring all the dimensions along which meaning of social gestures is communicated and integrating this information, could have great potential to advance the field of touch gesture recognition. More universal classes of touch gestures and a spectrum on which these categories can be plotted, along with the possibility to consider multidimensional information can improve the ability for social robots to recognize touch gestures and consequently enhance the interaction between humans and robots. Conclusively, this can result in a more desirable implementation of social robots in society. Besides the lack of a universal approach to the definition of touch gesture classes, the fact that a spatio-temporal approach could not be considered in this research due to instrumental constrains is limiting for this study. A potential research direction would be to explore a spatio-temporal and a spatial approach respectively and compare the two.

REFERENCES

Albawi, S., & Ucan, O. (2018, 10). Social touch gesture recognition using convolutional neural network. *Computational Intelligence and Neuroscience, 2018*. doi: 10.1155/2018/6973103

Altun, K., & MacLean, K. E. (2015). Recognizing affect in human touch of a robot. *Pattern Recognition Letters, 66*, 31-40. Retrieved from https://www.sciencedirect.com/science/article/pii/S016786551400333X (Pattern Recognition in Human Computer Interaction) doi: https://doi.org/10.1016/j.patrec.2014.10.016

Bianchini, B., Verma, P., & Salisbury, J. K. (2021). Towards human haptic gesture interpretation for robotic systems. In *2021 ieee/rsj international conference on intelligent robots and systems (iros)* (p. 7334-7341). doi: 10.1109/IROS51168.2021.9636015

Blumrosen, G., Sakuma, K., Rice, J. J., & Knickerbocker, J. (2020). Back to finger-writing: Fingertip writing technology based on pressure sensing. *IEEE Access, 8*, 35455-35468. doi: 10.1109/ACCESS.2020.2973378

Cang, X., Bucci, P., Strang, A., Allen, J., Maclean, K., & Liu, H. Y. (2015, 11). Different strokes and different folks: Economical dynamic surface sensing and affect-related touch recognition. In (p. 147-154). doi: 10.1145/2818346.2820756

Cascio, C. J., Moore, D., & McGlone, F. (2019). Social touch and human development. *Developmental Cognitive Neuroscience, 35*, 5-11. Retrieved from https://www.sciencedirect.com/science/article/

pii/S1878929317301962 (Social Touch: A new vista for developmental cognitive neuroscience?) doi: https://doi.org/10.1016/j.dcn.2018.04.009

Field, T. (2010). Touch for socioemotional and physical well-being: A review. *Developmental Review*, *30*(4), 367-383. Retrieved from https://www.sciencedirect.com/science/article/pii/S0273229711000025 doi: https://doi.org/10.1016/j.dr.2011.01.001

Garcini, L., Chen, M., Brown, R., LeRoy, A., Cano, M. , Peek, K., & Fagundes, C. (2019, 12). "abrazame que ayuda" (hug me, it helps): Social support and the effect of perceived discrimination on depression among us- and foreign-born latinxs in the usa. *Journal of Racial and Ethnic Health Disparities*, *7*. doi: 10.1007/s40615-019-00676-8

Gaus, Y. F. A., Olugbade, T., Jan, A., Qin, R., Liu, J., Zhang, F., ... Bianchi-Berthouze, N. (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* (p. 399–406). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2818346.2830599 doi: 10.1145/2818346.2830599

Ghosh, S., Hiware, K., Ganguly, N., Mitra, B., & De, P. (2019). Emotion detection from touch interactions during text entry on smartphones. *International Journal of Human-Computer Studies*, *130*, 47-57. Retrieved from https://www.sciencedirect.com/science/article/pii/S1071581918304889 doi: https://doi.org/10.1016/j.ijhcs.2019.04.005

Henderson, T., & Fulcher, B. D. (2021). An empirical evaluation of time-series feature sets. *CoRR*, *abs/2110.10914*. Retrieved from https://arxiv.org/abs/2110.10914

Hertenstein, M. J., Holmes, R., McCullough, M. E., & Keltner, D. (2009). The communication of emotion via touch. *Emotion*, *9 4*, 566-73.

Hertenstein, M. J., Keltner, D., App, B., Bulleit, B. A., & Jaskolka, A. R. (2006, August). Touch communicates distinct emotions. *Emotion (Washington, D.C.)*, *6*(3), 528—533. Retrieved from https://doi.org/10.1037/1528-3542.6.3.528 doi: 10.1037/1528-3542.6.3.528

Horii, T., Nagai, Y., & Asada, M. (2018). Modeling development of multimodal emotion perception guided by tactile dominance and perceptual improvement. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(3), 762-775. doi: 10.1109/TCDS.2018.2809434

Hughes, D., Farrow, N., Profita, H., & Correll, N. (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 (p. 415–422). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2818346.2830601 doi: 10.1145/2818346.2830601

Hughes, D., Krauthammer, A., & Correll, N. (2017). Recognizing social touch gestures using recurrent and convolutional neural networks. In *2017 ieee international conference on robotics and automation (icra)* (p. 2315-2321). doi: 10.1109/ICRA.2017.7989267

Jung, M. (2014). Towards social touch intelligence: Developing a robust system for automatic touch recognition. In *Proceedings of the 16th international conference on multimodal interaction* (p. 344–348). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2663204.2666281 doi: 10.1145/2663204.2666281

Jung, M. (2017). *Socially intelligent robots that understand and respond to human touch* (Doctoral dissertation). doi: 10.3990/1.9789036543644

Jung, M., Poel, M., Poppe, R., & Heylen, D. (2014). Touching the void – introducing cost: Corpus of social touch. In *Proceedings of the 16th international conference on multimodal interaction* (p. 120–127). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi-org.tilburguniversity.idm.oclc.org/10.1145/2663204.2663242 doi: 10.1145/2663204.2663242

Jung, M., Poel, M., Poppe, R., & Heylen, D. (2016, 10). Automatic recognition of touch gestures in the corpus of social touch. *Journal on Multimodal User Interfaces*, *11*, 81-96. doi: 10.1007/s12193-016-0232-9

Lee, J., & Cichy, K. (2020, 04). Complex role of touch in social relationships for older adults' cardiovascular disease risk. *Research on Aging*, *42*, 016402752091579. doi: 10.1177/0164027520915793

Li, Y.-K., Meng, Q.-H., & Zhang, H.-W. (2022). Touch gesture recognition using spatiotemporal fusion features. *IEEE Sensors Journal*, *22*(1), 428-437. doi: 10.1109/JSEN.2021.3090576

Miller, J., Bernstein, M., & McDaniel, T. (2021). Next steps for social robotics in an aging world. *IEEE Technology and Society Magazine*, *40*(3), 21-23. doi: 10.1109/MTS.2021.3101931

Sheridan, T. B. (2020). A review of recent research in social robotics. *Current Opinion in Psychology*, *36*, 7-12. Retrieved from https://www.sciencedirect.com/science/article/pii/S2352250X2030004X (Cyberpsychology) doi: https://doi.org/10.1016/j.copsyc.2020.01.003

Ta, V., Johal, W., Portaz, M., Castelli, E., & Vaufreydaz, D. (2015, 11). The grenoble system for the social touch challenge at icmi 2015.

van Wingerden, S., Uebbing, T., Jung, M., & Poel, M. (2014). A neural network based approach to social touch classification. In *Proceed-*

ings of the 2014 workshop on emotion representation and modelling in human-computer-interaction-systems (p. 7–12). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/2668056.2668060 doi: 10.1145/2668056.2668060

Wang, P., & Chen, D. (2021, July). *A Physical Human-Robot Interaction Dataset - TacAct.* Zenodo. Retrieved from https://doi.org/10.5281/zenodo.5138841 doi: 10.5281/zenodo.5138841

Wang, P., Liu, J., Hou, F., Chen, D., Xia, Z., & Guo, S. (2021). Organization and understanding of a tactile information dataset tacact during physical human-robot interactions. *CoRR*, *abs/2108.03779*. Retrieved from https://arxiv.org/abs/2108.03779

Wei, J., Liu, H., Wang, B., & Sun, F. (2019, 07). Lifelong learning for tactile emotion recognition. *Interaction Studies*, *20*, 25-41. doi: 10.1075/is.18041.wei