Jointly Modeling Spatio-Temporal Features of Tactile Signals for Action Classification

Jimmy Lin[1,2], **Junkai Li**[1,2], **Jiasi Gao**[1], **Weizhi Ma**[1*], **Yang Liu**[1,2*]

[1] Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China
[2] Department of Computer Science and Technology, Tsinghua University, Beijing, China
jimmy.lin9@gmail.com, li-jk23@mails.tsinghua.edu.cn, bnancy@sina.com,
{mawz, liuyang2011}@tsinghua.edu.cn

Abstract

Tactile signals collected by wearable electronics are essential in modeling and understanding human behavior. One of the main applications of tactile signals is action classification, especially in healthcare and robotics. However, existing tactile classification methods fail to capture the spatial and temporal features of tactile signals simultaneously, which results in sub-optimal performances. In this paper, we design **S**patio-**T**emporal **A**ware tactility **T**ransformer (**STAT**) to utilize continuous tactile signals for action classification. We propose spatial and temporal embeddings along with a new temporal pretraining task in our model, which aims to enhance the transformer in modeling the spatio-temporal features of tactile signals. Specially, the designed temporal pretraining task is to differentiate the time order of tubelet inputs to model the temporal properties explicitly. Experimental results on a public action classification dataset demonstrate that our model outperforms state-of-the-art methods in all metrics.

Introduction

Similar to visual and acoustic signals, tactile signals are important for modeling and understanding humans. In recent years, various wearable electronics have been designed to collect tactile signals, which are widely used in multiple scenarios, especially in healthcare and robotics (Zhu et al. 2019; Fan et al. 2020; Lou et al. 2020; Okunevich et al. 2021).

The collected tactile signals can be utilized for different purposes, and one of their main applications is the action classification task. Sundaram et al. (2019) propose to identify hand actions by tactile signals with sensors in gloves. Luo et al. (2021) and Wicaksono et al. (2022) use wearable electronic socks to collect tactile signals for feet action classification. Figure 1 is an example, where the continuous tactile signals are collected by e-textile sensors in socks, and then used to classify the action (e.g., walking, etc.).

Tactile signals are spatially and temporally sensitive, hence utilizing their spatio-temporal features is important for action classification. Firstly, tactile signals are spatially sensitive as they are not *translation invariant*. The same signals in different positions (i.e., collected by various sensors) indicate distinct actions. For example, the same signals collected by sensors located in different positions should be classified as standing on toes or heels, respectively. Secondly, tactile signals are temporally sensitive as they are collected regularly with high frequency, e.g., in 10HZ (10 data points per second), and the time order of these signals is informative. For example, if ignoring the order of collected signals, two signal sequences collected by the same sensor from distinct actions may be seen as identical actions (i.e., same elements with different orders), which becomes useless in classification.

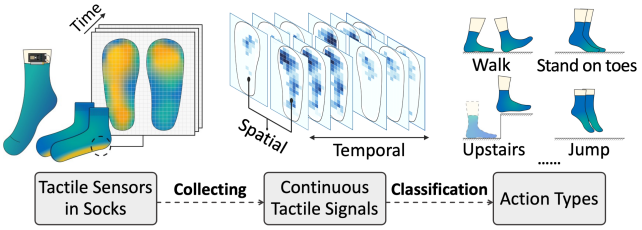


Figure 1: An overview of action classification based on tactile signals collected by wearable electronic socks.

Furthermore, we want to point out that jointly modeling spatial and temporal features is essential for tactile signals in action classifications. We conduct an empirical study on a real-scenario dataset (Luo et al. 2021), and draw the spatial and temporal features of different actions in Figure 2. The heatmap of each action shows the averaged results of all samples, which indicates spatial features. The temporal change of a specific sensor shows the averaged sequence data of all samples collected by this sensor, which indicates temporal features. As shown in Figure 2(a), two actions, stand on toes and lean left, have similar temporal features but different spatial features. However, in Figure 2(b), two actions, upstairs and walk fast, have similar spatial features but different temporal features. These observations verify tactile signals' spatio-temporal features, and further indicate that only using one of them is inadequate for classification.

However, existing tactile methods lack the ability to capture the mentioned temporal and spatial nature of tactile signals simultaneously. On the one hand, most previous tactile-related studies adopt CNN-based methods to model

*Corresponding authors.

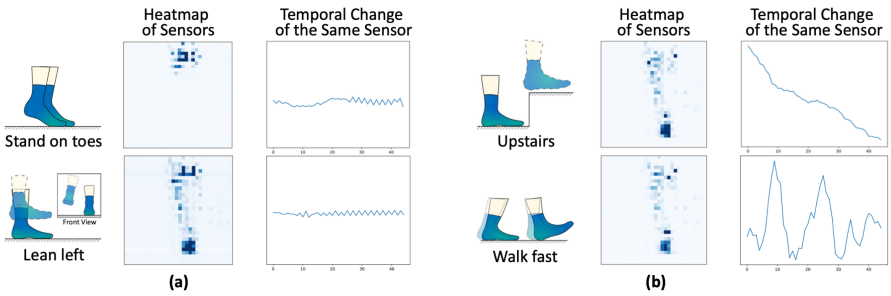


Figure 2: Empirical study of actions in a tactile dataset. Heatmaps are the averaged results of all samples collected by sensors in the left foot, and the tactile sensor of Figure 2(a) and 2(b) is located at positions (5,20) and (28,19) of the left foot, respectively.

the tactile signal frames and then combine them by concatenating or sequential models, which fail to jointly capture their translation variance and temporal properties (Luo et al. 2021; Sundaram et al. 2019; Gao et al. 2020). On the other hand, various transformer models have been designed to handle different continuous signals. But most of them (Zerveas et al. 2021; Tong et al. 2022; Amiridi, Darnell, and Jewell 2022) focus on temporal features, which are inadequate to model tactile signals' spatial nature, especially the *translation variant* property.

In this paper, we design a **S**patio-**T**emporal **A**ware tactility **T**ransformer (**STAT**) to utilize tactile signals for action classification, which utilizes their temporal and spatial features simultaneously. We design spatial and temporal embeddings to explicitly model the translation variant and sequential features of tactile signals, respectively. Additionally, we introduce a temporal pretraining task to enhance the modeling of temporal features by distinguishing the time order of signal tubelets. After pretraining the STAT transformer, the embedding of the [CLS] token is utilized for action classification. Experimental results on tactile show that our model outperforms all baseline methods in all metrics, including state-of-the-art multivariate and video classification models. Further analyses verify the effectiveness of the proposed pretraining task and embeddings. To the best of our knowledge, this is the first transformer model designed for tactile signals by jointly modeling spatio-temporal features, which can be applied to various tactile-related scenarios.

Related Work

Action Classification with Tactile Signals

In recent years, various wearable electronics have been designed to model user actions based on tactile signals in different scenarios. Luo et al. (2021) design wearable electronic socks to classify user walking actions. Noh et al. (2021) use tactile signals in healthcare scenarios, which predict the fall risk of users. Robotic studies also point out that modeling tactile signals are important in understanding humans (Kragic et al. 2018; Negre et al. 2018).

Despite the importance of tactile signals in various scenarios, we find that previous tactile classification models are unable to capture the spatial and temporal properties of tactile signals simultaneously. Sundaram et al. (2019) use CNN to capture the embedding of each frame and simply concatenate them for action classification. Recent studies enhance this method by adopting a GRU/LSTM model rather than concatenation to model the sequential features (Luo et al. 2021; Okunevich et al. 2021; Gao et al. 2020). Cao et al. (2020) introduces temporal attention operation combined with spatial features in separate phases. However, as CNNs are designed to utilize the translation invariance features, they fail to capture tactile signals' translation variance.

Different from previous studies, we design a new transformer model to jointly capture the spatio-temporal features of tactile signals for action classification.

Transformers for Continuous Signals

Transformer models (Vaswani et al. 2017) have achieved great success in continuous signal classification tasks, e.g., videos and multivariate continuous signals (Tong et al. 2022; Zerveas et al. 2021; Zhao et al. 2022). We briefly review related transformers here, especially video transformers, as the input shape of videos is similar to tactile signals.

Existing transformer models fail to utilize the spatial and temporal features of tactile signals simultaneously. On the one hand, most video transformers use the visual transformer (Dosovitskiy et al. 2021) as a backbone model, and further propose new masking or input strategies (Arnab et al. 2021a; Bertasius, Wang, and Torresani 2021; Yan et al. 2022a; Liu et al. 2022; Yan et al. 2022b). Recent models propose to capture the spatio-temporal features of videos, e.g., VideoMAE (Tong et al. 2022), SSTSA (Alfasly et al. 2022). However, as video transformers aim to model the translation invariant of videos, they only use position embeddings in encoding, which fail to model the translation variant spatial property of tactile signals. On the other hand, most transformer methods proposed for multivariate continuous signals focus on modeling their temporal features, while ignoring the spatial relations among different signals (i.e., where the signals are collected), such as TST (Zerveas et al. 2021) and the transformer proposed by Hannan et al (2021). Furthermore, most transformer models rely on the masking and reconstruction pretraining task (Devlin et al. 2019; Bao et al. 2022; Tong et al. 2022), which cannot explicitly capture the temporal/spatial features of continuous signals.

Although these transformers are not designed for tactile signals, we will use them to verify the effectiveness of STAT.

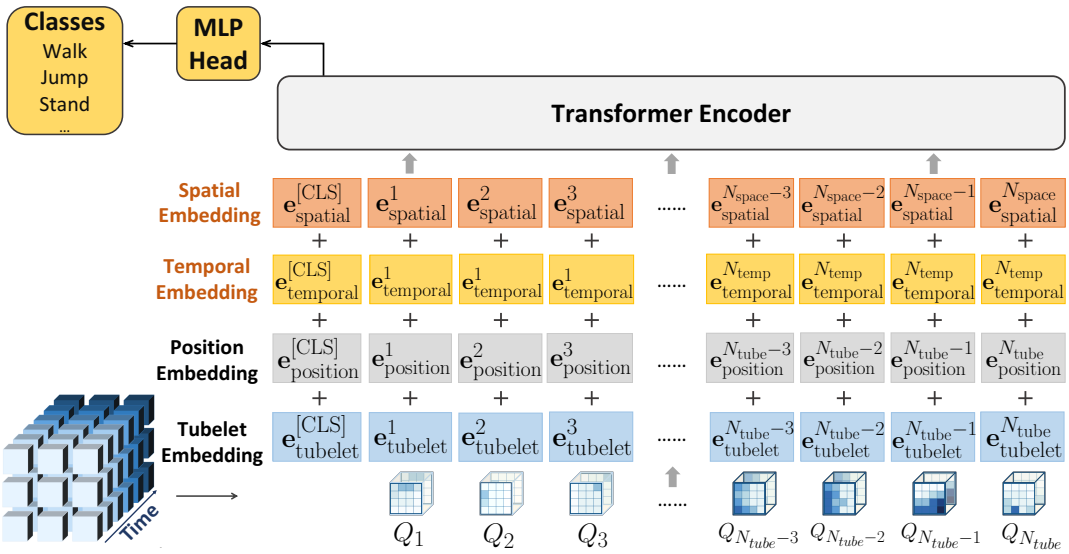


Figure 3: An overview of STAT model. Spatial and temporal embeddings are designed to jointly capture both properties.

Approach

Problem Statement

Our goal is to utilize tactile signals to classify user actions, where the data can be collected by various wearable electronics. The wearable devices often arrange sensors as a matrix, so we define the data matrix collected in a specific time point as a frame. Then, the task is defined as follows.

Given a tactile signal tensor $\mathcal{X} \in \mathbb{R}^{C \times T \times H \times W}$, where C represents the number of wearable devices, T represents the length of signal sequences (i.e., T frames), H and W mean the number of sensors in each column/row (i.e., the shape of frames), respectively. An example of tactile signals is shown in Figure 4(a). $\mathcal{X}_{c_i,t_j,h_k,w_l}$ represents the value collected by the sensor of device c_i in position (h_k, w_l) at time t_j. Each tactile segment $\mathcal{X}$ has an activity label y, and the total number of activity types is M. Our target is to accurately classify the given tactile signal $\mathcal{X}$ to its label y.

Overview

We propose a spatio-temporal aware tactility transformer for the action classification task based on tactile signals, which is named STAT. A new pretraining task and two extra embeddings are designed to capture the temporal and spatial features of tactile signals jointly in STAT.

Firstly, the designed spatio-temporal aware transformer encoder is introduced. We convert the tactile signal tensor to a tubelet sequence. Besides the widely used tubelet and position embeddings, we propose to add spatial and temporal embeddings to capture each tubelet's temporal and spatial features, respectively. Then, multi-layer transformer encoders are adopted to calculate the representations of tactile signals. Then, the adopted pretraining tasks are defined. Aside from the common masking and reconstruction loss, we designed a temporal pretraining task to explicitly discriminate the time order of tubelet pairs. Finally, we show how to adopt our model for action classifications.

Spatio-Temporal Aware Transformer Encoder

We will introduce the designed spatio-temporal aware transformer encoder shown in Figure 3. To simplify the notations, we only show the process for handling tactile tensor collected from one wearable device (i.e., $\mathcal{X} \in \mathbb{R}^{T \times H \times W}$), as we can easily expand our model to C-channel transformers to utilize signals collected by C devices.

Tubelet Inputs As the spatial and temporal dimensions of the tactile signals can be redundant, directly adopting the whole data in classification may result in reduced efficiency. Motivated by previous video transformer models that convert the video clip into tubelets to alleviate the spatio-temporal redundancy, we follow these studies by transferring the tactile signals into a tubelet sequence (Arnab et al. 2021b; Liu et al. 2021; Fan et al. 2021; Tu et al. 2022). We define a tubelet as $\mathcal{Q} \in \mathbb{R}^{L \times P \times P}$, where L represents its sequence length (i.e., the number of frames) and P represents the patch size (i.e., height and width). Figure 4(b) shows some examples of the converted tubelets, and the total number of tubelets for a tactile signal tensor $\mathcal{X}$ is $N_{\text{tube}} = THW/(LP^2)$.

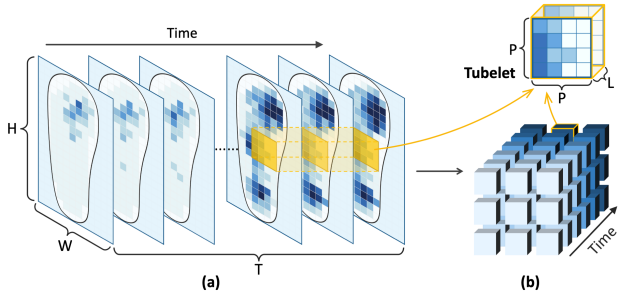


Figure 4: (a) Visualization of tactile signal $\mathcal{X} \in \mathbb{R}^{T \times H \times W}$. (b) Tubelet inputs, where each tubelet $\mathcal{Q} \in \mathbb{R}^{L \times P \times P}$.

Spatio-Temporal Enhanced Tubelet Embeddings Most video transformer models adopt the tubelet embeddings and position embeddings as the input of transformer encoders (Dosovitskiy et al. 2021; Tong et al. 2022). However, due to the fact that tactile signals do not have the translation invariance property as images/videos, simply adopting these settings cannot capture the spatial features of tactile signals. Additionally, jointly modeling spatial and temporal features are also essential in distinguishing actions, as shown in Figure 2. Thus, we propose to add spatial embeddings and temporal embeddings for each tubelet to capture the spatio-temporal features of tactile signals jointly.

Spatial Embeddings. Each tubelet is collected from a patch of sensors, and the sensors are located in certain positions. We use a spatial embedding $\mathbf{e}_k^{\text{spatial}}$ to represent where the tubelet signal is collected from, so that the spatial features will be encoded to explicitly model the translation variance. Tubelets collected by the same batch of sensors will get the same spatial embeddings, and the number of spatial embedding types is $N_{\text{space}} = HW/P^2$.

Following the traditional calculation of position embeddings (Vaswani et al. 2017), we utilize the sinusoidal positional encoding table to calculate the spatial embedding $\mathbf{e}_k^{\text{spatial}}$, where k represents the spatial position and $k \in \{1, 2, .., N_{\text{space}}\}$. The calculation is defined in Equation (1):

$$\begin{aligned} \mathbf{e}_{(k,2d)}^{\text{spatial}} &= \sin\left(\frac{k}{10000^{\frac{2d}{D}}}\right) \\ \mathbf{e}_{(k,2d+1)}^{\text{spatial}} &= \cos\left(\frac{k}{10000^{\frac{2d}{D}}}\right) \end{aligned} \tag{1}$$

Where D represents the embedding dimensions, and $\mathbf{e}_{(k,d)}^{\text{spatial}}$ refers to the d-th dimension of $\mathbf{e}_k^{\text{spatial}}$ ($d \in \{0, 1, 2, 3..., D\}$). Through this encoding process, the spatial embeddings can provide the transformer encoder with spatial knowledge of the tactile signals, which contributes to modeling the translation variant features.

Temporal Embeddings. For tactile signal tubelets, their temporal features are important in distinguishing various actions. We propose temporal embeddings to represent the location of the tubelet in the time sequence, which refers to when the tubelet is collected.

Similar to the spatial embeddings, we use the sinusoidal positional encoding Equation (1) to generate the temporal embedding $\mathbf{e}_k^{\text{temporal}}$. The number of temporal embedding types is $N_{\text{temp}} = T/L$, and tubelets collected in the same frames have the same $\mathbf{e}_k^{\text{temporal}}$, where $k \in \{1, 2, ..., N_{\text{temp}}\}$.

These new embeddings are used to enhance the model with additional information about the spatial and temporal properties of tactile signals jointly, and Figure 3 shows an example. Then, we aggregated the proposed two embeddings with tubelet and position embeddings to calculate the input matrix $\mathbf{E}^{\text{input}}$ of transformer encoders by Equation (2). Ultimately, the aggregation of embeddings allows for the simultaneous embedding of spatial and temporal properties.

$$\mathbf{E}^{\text{input}} = \mathbf{E}^{\text{tubelet}} + \mathbf{E}^{\text{position}} + \mathbf{E}^{\text{spatial}} + \mathbf{E}^{\text{temporal}} \tag{2}$$

As shown in Figure 3, we append a [CLS] token at the beginning of the tubelet sequence, which is often used to represent the whole embedding sequence in transformer models. The tubelet, position, temporal, and spatial embeddings of this token are randomly initialized and optimized during training. Hence, we have $\mathbf{E}^{\text{input}} = [\mathbf{E}_{[\text{CLS}]}^{\text{input}}, \mathbf{E}_{\mathcal{Q}_1}^{\text{input}}, ..., \mathbf{E}_{\mathcal{Q}_{N_{\text{tube}}}}^{\text{input}}] \in \mathbb{R}^{(N_{\text{tube}}+1)\times D}$.

Transformer Encoders We utilize the classical transformer encoder (Vaswani et al. 2017) as the backbone network, whose effectiveness has been verified in various domains and tasks (Devlin et al. 2019; Arnab et al. 2021b; Dosovitskiy et al. 2021).

Our transformer encoder takes in $\mathbf{E}^{\text{input}}$ defined in the previous subsection. As the transformer encoder often consists of K transformer layers, we note the primary input $\mathbf{E}^{\text{input}}$ as $\mathbf{E}^{(0)}$, and $\mathbf{E}^{(k)} = \text{Transformer}_{\text{k}}(\mathbf{E}^{(k-1)})$, where $k \in \{1, 2, ..., K\}$. The output of the final layer $\text{Transformer}_{\text{K}}$ is the encoded representation of input tokens, and $\mathbf{E}_{[\text{CLS}]}^{(K)}$ is the final embedding of tactile signals.

Pretraining Tasks

Pretraining has been verified to be an effective technique to enhance transformer models in various scenarios, e.g., BERT for text (Devlin et al. 2019), BEIT for image (Bao et al. 2022), and VideoMAE for video (Tong et al. 2022). To achieve better classification performances, we choose to pretrain our STAT model before applying it to action classifications. We propose to use two pretraining tasks here, as shown in Figure 5. The first one is the masked tubelet reconstruction (MTR) task, which aims to reconstruct the masked input tubelets, which is also used in previous video transformers (Tong et al. 2022). The other one is our designed temporal pretraining task to explicitly model the temporal features of tactile signal tubelets. Although temporal embeddings are helpful in capturing temporal properties, we prefer to add a specific pretraining task due to the importance of temporal features in distinguishing different actions.

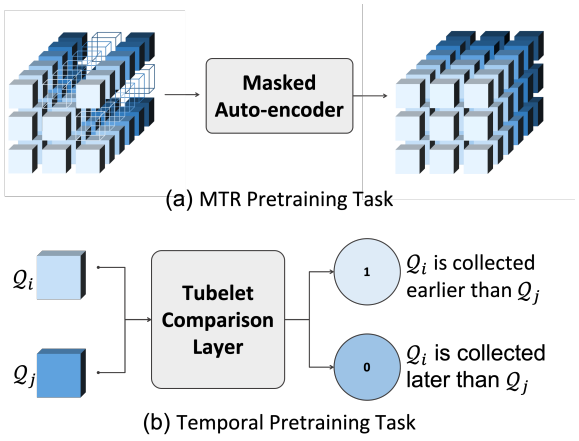


Figure 5: Illustrations of the adopted two pretraining tasks.

Masked Tubelet Reconstruction As shown in Figure 5(a), we use a masked auto-encoder to reconstruct the input signals, which is the MTR task in previous studies (Tong et al. 2022; Arnab et al. 2021a). MTR randomly masks tubelets from videos and reconstructs them in pretraining, and its loss function is defined as follows:

$$\mathcal{L}_{\text{MTR}} = \frac{1}{|\mathbb{M}|} \sum_{t \in \mathbb{M}} |V_t - \hat{V}_t|^2 \tag{3}$$

Where $\mathbb{M}$ is the set of masked tubelets' indexes, V is the input video, and $\hat{V}$ is the reconstructed one. Specially, we adopt spatial-based random masking instead of randomly masking all tubelets. In this strategy, we randomly select sensor groups from the N_{space} types for masking. All signals collected by the chosen masking sensors (i.e., tubelets with the same spatial embeddings) will be masked. The motivation is that this masking strategy will contribute to better utilizing the spatial features among sensors. For the mask ratio, we leave it as a hyper-parameter study.

Temporal Pretraining Our self-supervised temporal pretraining task enhances transformer encoders by training to distinguish the time order of two randomly selected tubelets, so that the temporal features can be maintained in the model, as shown in Figure 5(b).

Firstly, two tubelets $\mathcal{Q}_i$ and $\mathcal{Q}_j$ are randomly selected from the whole set. Note that we should make sure $\mathcal{Q}_i$ and $\mathcal{Q}_j$ are collected at different times, so the temporal embeddings of them are different (i.e., $\mathbf{e}^{\text{temporal}}_{\mathcal{Q}_i} \neq \mathbf{e}^{\text{temporal}}_{\mathcal{Q}_j}$). Then, we use the encoded embeddings $\mathbf{E}^{(K)}_{\mathcal{Q}_i}$ and $\mathbf{E}^{(K)}_{\mathcal{Q}_j}$ of tubelet $\mathcal{Q}_i$ and $\mathcal{Q}_j$ to identify the time order of them. If tubelet $\mathcal{Q}_i$ is collected earlier than $\mathcal{Q}_j$, the identification result should be $y^{\text{temp}}_{i,j} = 1$, otherwise 0.

We choose a simple but effective way to optimize this task. We concatenate the two embeddings $\mathbf{E}^{(K)}_{\mathcal{Q}_i}$ and $\mathbf{E}^{(K)}_{\mathcal{Q}_j}$, and use a linear layer with a sigmoid activation function to predict the time order $\hat{y}^{\text{temp}}_{i,j}$. A binary cross-entropy loss is utilized to optimize this pretraining task. Moreover, for each tactile signal tensor, only randomly selecting one pair of tubelets for pretraining is not enough. So N_{comp} tubelet pairs are randomly selected and used in pretraining, i.e., $(\mathcal{Q}_{i_1}, \mathcal{Q}_{j_1})$, ..., $(\mathcal{Q}_{i_{N_{\text{comp}}}}, \mathcal{Q}_{j_{N_{\text{comp}}}})$. Finally, the loss function is formally defined in Equation (4).

$$\begin{aligned} \hat{y}^{\text{temp}}_{i_n,j_n} =& \sigma(\mathbf{W}_{\text{frame}}(\mathbf{E}^{(K)}_{\mathcal{Q}_{i_n}} \oplus \mathbf{E}^{(K)}_{\mathcal{Q}_{j_n}})^\top \\ \mathcal{L}_{\text{temp}} =& -(y^{\text{temp}}_{i_n,j_n} \log(\hat{y}^{\text{temp}}_{i_n,j_n}) \\ & + (1 - y^{\text{temp}}_{i_n,j_n}) \log(1 - \hat{y}^{\text{temp}}_{i_n,j_n})) \end{aligned} \tag{4}$$

Where $\oplus$ means vector concatenation, σ is a sigmoid activation function, $n \in \{1, 2, ..., N_{\text{comp}}\}$ and $\mathbf{W}_{\text{frame}} \in \mathbb{R}^{1 \times 2D}$.

To simultaneously utilize these two tasks in pretraining, we add an extra setting that only randomly selects unmasked tubelets, so we can optimize them together. Specifically, we aggregate the MTR loss with our temporal loss through a weight coefficient β, which is a hyper-parameter. The final pretraining loss is defined as follows:

$$\mathcal{L}_{\text{pretrain}} = \mathcal{L}_{\text{MTR}} + \beta \mathcal{L}_{\text{temp}} \tag{5}$$

Actions	#Samples	Actions	#Samples
Downstairs (C1)	4,942	Stand_toes (C6)	3,978
Jump (C2)	3,090	Upstairs (C7)	5,025
Lean_left (C3)	5,047	Walk (C8)	6,078
Lean_right (C4)	5,011	Walk_fast (C9)	5,360
Stand (C5)	5,024		

Table 1: Statistics and short-names of each type of action.

With the spatial-based random mask strategy in the MTR task and the designed temporal pretraining task, we enhance the representation ability of transformer encoders to better capture the spatial and temporal properties of tactile signals jointly. Similar to the pretraining of video transformers, we only use the training set of tactile signals for pretraining, as there lacks large scale open tactile datasets.

Fine-Tuning for Action Classification

After introducing our STAT model and pretraining tasks, we will show how to train STAT for action classification.

We follow the approach of other transformer models by using the embedding of [CLS] token to represent the entire signal sequence. Firstly, we take the embedding of [CLS] token from the last transformer layer block (i.e., $\mathbf{E}^{(K)}_{[\text{CLS}]}$), which represents the whole input signal. Then, we add a linear layer on the top of this embedding to classify it into action types (shown in Figure 1). The loss function is:

$$\begin{aligned} \hat{\mathbf{y}}_i =& \delta(\mathbf{W}_{\text{c}}(\mathbf{E}^{(K)}_{[\text{CLS}]})^\top + \mathbf{b}_{\text{c}}) \\ \mathcal{L} =& \text{CrossEntropy}(\hat{\mathbf{y}}_i, \mathbf{y}_i) \end{aligned} \tag{6}$$

Where δ is the softmax activation function, $\mathbf{W}_{\text{c}} \in \mathbb{R}^{M \times 2D}$ and $\mathbf{b}_{\text{c}} \in \mathbb{R}^{1 \times M}$, and $\mathbf{y}_i$ is a one-hot vector where only the index of the true label is 1.

Experiments

Experimental Settings

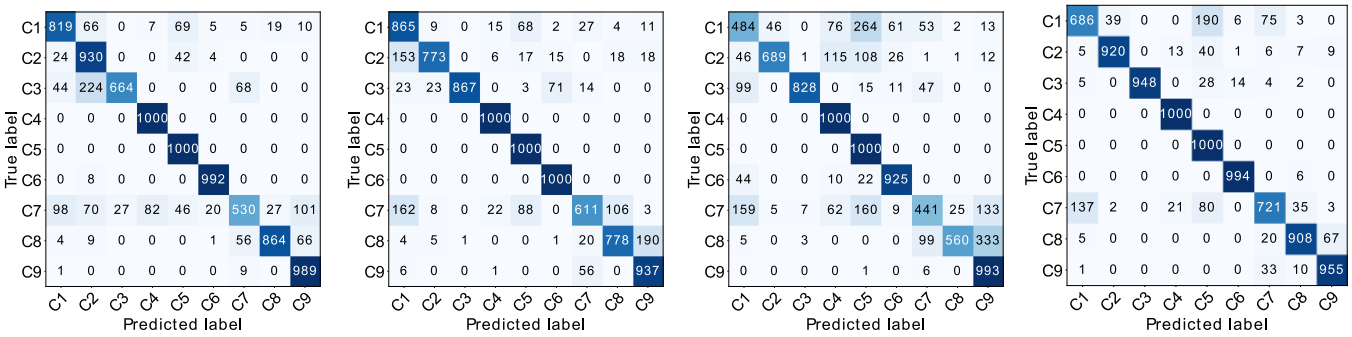
Dataset As tactile action classification is a promising new application scenario that is under development, there is only one large-scale open dataset by far. So our experiments are conducted on the public tactile signal dataset[1], which is collected by individuals with two wearable electronic socks to perform specific actions. The dataset consists of tactile signals with 9 labeled actions, namely walking, leaning on the left foot, leaning on the right foot, climbing downstairs, climbing upstairs, jumping, standing on toes, fast walking, and standing upright. The statistics are shown in Table 1. T, H, and W are set to 45, 32, and 32, respectively.

As the sampling frequency is 15HZ, each piece of data is collected in 3 seconds. Following the providers' settings, 500 and 1,000 samples of each action are used in validation and testing, respectively, and the other samples are used in training (each action type will be sampled to 4,000 samples). Only the training set will be adopted for model pretraining to avoid data leakage.

[1] http://senstextile.csail.mit.edu/

Models	ACC@1	ACC@3	Macro-F1
CNN&GRU (Luo et al. 2021)	0.8794±0.0280	0.9497±0.0183	0.8743±0.0319
TST (Zerveas et al. 2021)	0.8701±0.0252	0.9637±0.0147	0.8660±0.0272
VideoMAE (Tong et al. 2022)	0.7705±0.0906	0.9287±0.0177	0.7521±0.1027
STAT w/o pretraining	0.8050±0.0549	0.9528±0.0225	0.7946±0.0652
STAT	**0.9033**±0.0098	**0.9830**±0.0081	**0.9015**±0.0104

Table 2: Overall performances of all models.



(a) CNN&GRU (b) TST (c) VideoMAE (d) STAT

Figure 6: Confusion Matrices of all models.

Hyper-parameters	Values
#Comparison Pairs N_{comp}	10, 20, 30, 40, 50
Loss Weight β	0.5, 0.75, 1, 1.5, 2, 2.5
Masking Ratio	0.1 - 0.9 with step length 0.1
Adam Learning Rate	1e-3, 5e-3, 1e-2
Transformer Layer K	3, 6, 9, 12

Table 3: Summarization of tuned hyper-parameters.

Metrics We use accuracy and macro-F1 as evaluation metrics. As there are multiple classes, we report both Top-1 & Top-3 accuracy as in previous studies (Luo et al. 2021). Besides, we add Macro-F1 to show the comprehensive performance on the imbalanced dataset of all models.

Baselines Several state-of-the-art baselines are adopted:

- CNN&GRU (Luo et al. 2021): This method adopts convolution and recurrent networks for action classification;
- TST (Zerveas et al. 2021): TST is a state-of-the-art transformer-based model for continuous multivariate signal classification with pretraining;
- VideoMAE (Tong et al. 2022): This is a state-of-the-art video classification model with masked auto-encoders.

Implementation Details We tune hyper-parameters as shown in Table 3. The tubelet parameters L and P are set to 5 and 4, and the pretraining and fine-tuning epoch is set to 60. The embedding dimension D is set to 768, in which batch size is 64 and weight decay is 1e-4. For baselines, we employ their public implementations and tune them with hyperparameters suggested by their authors.

All experiments are implemented by Pytorch 1.7 and executed on 4 Tesla V100 or GeForce RTX 3090 GPUs. Experiments are repeated 5 times with different random seeds. Besides, the total training time of STAT is similar to VideoMAE (10 hours). The code is available at https://github.com/Aressfull/sock_classification. Note user consent is necessary before applying it in real scenarios.

Overall Performances

Experimental results of our STAT and baselines are reported in Table 2. TST, VideoMAE, and STAT models are only pretrained with the training set to avoid data leakage, and STAT w/o pretraining is directly trained for the classification task.

Firstly, our pretrained STAT outperforms all baseline models in all metrics, showing that jointly modeling spatial and temporal features contributes to better action classification results. STAT achieves 2.7%, 2.0%, and 3.1% improvements than the best baseline in ACC@1, ACC@3, and Macro-F1, respectively. Secondly, STAT without pretraining performs worse than most baselines, showing that our pretraining provides significant improvements for STAT in the action classification task. Thirdly, for the baseline models, the widely used tactile CNN&GRU model achieves comparable results as TST, showing that modeling spatial features and temporal features are both important in action classification. However, VideoMAE model performs the worst, which indicates that simply reusing the video transformer will get worse performance. The reason should be videoMAE cannot

#	TE	SE	TPT	ACC@1	ACC@3	Macro-F1
1	✓			0.8764	0.9678	0.8715
2		✓		0.8417	0.9506	0.8337
3	✓	✓		0.8947	**0.9854**	0.8957
4			✓	0.8299	0.9507	0.8247
5	✓	✓	✓	**0.9033**	0.9830	**0.9015**

Table 4: Experimental results of various ablation strategies. TE: Temporal Embeddings, SE: Spatial Embeddings, and TPT: Temporal Pretraining Task.

capture the translation variant of tactile signals, and tactile signals are more dense than videos (VideoMAE only uses 8 frames but uses 45 frames here).

To further analyze the performances of different models in various classes, we show the confusion matrices of all models on the test set in Figure 6. From the figures, we have the following observations: Firstly, CNN&GRU performs worse in temporal and spatial sensitive classes, i.e., upstairs and lean left, showing the weaknesses of current tactile classification models. Specifically, CNN&GRU classifies many upstairs samples as walk fast due to their similar spatial features (as shown in Figure 2(b)), while our STAT can distinguish these actions more accurately due to the modeling of temporal features. Secondly, TST performs even worse than CNN&GRU in many actions, indicating that focusing on modeling temporal features is not enough for tactile signals. For example, TST mistakes a number of lean left samples as stand on toes because they have similar temporal features (as shown in Figure 2(a)). Our STAT rarely makes mistakes on these actions as they are distinct in spatial features. Thirdly, our STAT model performs the best in most classes, as we jointly capture both the spatial and temporal properties of tactile signals. Meanwhile, due to the translation variant property of tactile signals, VideoMAE, which is designed for video classifications, is unsuitable for this task.

Analyses

Ablation Study To verify the effectiveness of the designed pretraining task and embeddings, we conduct ablation studies. Table 4 shows our ablation strategies and their performances. Note that the MTR pretraining task, position, and tubelet embeddings are used in all experiments, as we focus on analyzing the newly designed models here.

We have the following observations from the results: Firstly, all designed modules contribute to the classification task, as STAT (Strategy 5) achieves the best performance with all modules in ACC@1 & Macro-F1, and comparable results in ACC@3. Secondly, by comparing Strategies 1,2,3 in pairs, we find that removing any one of the two designed embeddings will result in a large drop in performance. Besides, temporal embeddings are more important than spatial embeddings, as Strategy 1 performs better. Thirdly, STAT with both embeddings (Strategy 3) outperforms STAT with only the temporal task (Strategy 4) in all metrics. This indicates that only adopting the proposed pretraining task cannot make full use of its ability.

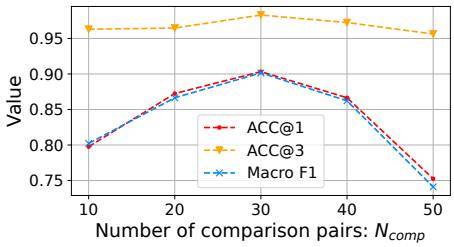


Figure 7: Effect of the number of comparison pairs.

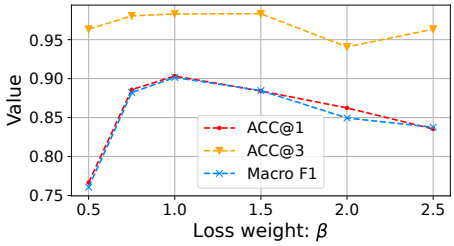


Figure 8: Effect of the weight of temporal pretraining loss.

Hyper-parameter Analyses Due to the space limit, we only show two conducted hyper-parameter experiments.

Effect of the Number of Comparison Pairs N_{comp}. To verify the effect of N_{comp} in the temporal pretraining task, we conduct analyses experiments and summarize the results in Figure 7. The best performance is achieved when $N_{\text{comp}} = 30$. Fewer comparison pairs perform worse may be caused by insufficient training, while more pairs will not contribute to better results either.

Effect of the Loss Weight β. We adjust the weight β for our temporal pretraining task in Equation (5) in different values, and the results are shown in Figure 8. It indicates that a too-low or too-high value of β will hurt the performance of our STAT model, and $\beta = 1$ performs the best.

Conclusions

Tactile signals are essential in modeling and understanding user behavior in various scenarios. However, neither previous tactile classification models nor transformer models for other continuous signals fail to simultaneously capture the spatial and temporal features of tactile signals. In this study, we propose a spatio-temporal aware tactility transformer to jointly model both spatial and temporal properties of tactile signals for action classification tasks. Spatial and temporal embeddings are designed to capture the translation variance and sequential features, respectively. Additionally, the proposed temporal pretraining task explicitly models the time order features. Experimental results show that our model outperforms all baseline models in all metrics.

Our model shows promising performance and can contribute to better utilizing tactile signals in other scenarios. In the future, we plan to introduce side information about tactile signals to achieve better performance.

Acknowledgements

This work is supported by the National Key R&D Program of China (2022ZD0160502) and the National Natural Science Foundation of China (No. 62372260). We appreciate all the reviewers for their insightful suggestions.

References

Alfasly, S.; Chui, C. K.; Jiang, Q.; Lu, J.; and Xu, C. 2022. An Effective Video Transformer With Synchronized Spatiotemporal and Spatial Self-Attention for Action Recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Amiridi, M.; Darnell, G.; and Jewell, S. 2022. Latent Temporal Flows for Multivariate Analysis of Wearables Data. *arXiv preprint arXiv:2210.07475*.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lucic, M.; and Schmid, C. 2021a. ViViT: A Video Vision Transformer. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 6816–6826. IEEE.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021b. ViViT: A Video Vision Transformer.

Bao, H.; Dong, L.; Piao, S.; and Wei, F. 2022. BEiT: BERT Pre-Training of Image Transformers. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Bertasius, G.; Wang, H.; and Torresani, L. 2021. Is Space-Time Attention All You Need for Video Understanding? In Meila, M.; and Zhang, T., eds., *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, 813–824. PMLR.

Cao, G.; Zhou, Y.; Bollegala, D.; and Luo, S. 2020. Spatio-temporal attention model for tactile texture recognition. In *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 9896–9902. IEEE.

Devlin, J.; Chang, M.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Burstein, J.; Doran, C.; and Solorio, T., eds., *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, 4171–4186. Association for Computational Linguistics.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Fan, H.; Xiong, B.; Mangalam, K.; Li, Y.; Yan, Z.; Malik, J.; and Feichtenhofer, C. 2021. Multiscale Vision Transformers.

Fan, W.; He, Q.; Meng, K.; Tan, X.; Zhou, Z.; Zhang, G.; Yang, J.; and Wang, Z. L. 2020. Machine-knitted washable sensor array textile for precise epidermal physiological signal monitoring. *Science advances*, 6(11): eaay2840.

Gao, R.; Taunyazov, T.; Lin, Z.; and Wu, Y. 2020. Supervised Autoencoder Joint Learning on Heterogeneous Tactile Sensory Data: Improving Material Classification Performance. In *IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS 2020, Las Vegas, NV, USA, October 24, 2020 - January 24, 2021*, 10907–10913. IEEE.

Hannan, M. A.; How, D. N.; Lipu, M.; Mansor, M.; Ker, P. J.; Dong, Z.; Sahari, K.; Tiong, S. K.; Muttaqi, K. M.; Mahlia, T.; et al. 2021. Deep learning approach towards accurate state of charge estimation for lithium-ion batteries using self-supervised transformer model. *Scientific reports*, 11(1): 1–13.

Kragic, D.; Gustafson, J.; Karaoguz, H.; Jensfelt, P.; and Krug, R. 2018. Interactive, Collaborative Robots: Challenges and Opportunities. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 18–25. ijcai.org.

Liu, Z.; Feng, R.; Chen, H.; Wu, S.; Gao, Y.; Gao, Y.; and Wang, X. 2022. Temporal Feature Alignment and Mutual Information Maximization for Video-Based Human Pose Estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 10996–11006. IEEE.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2021. Video Swin Transformer.

Lou, M.; Abdalla, I.; Zhu, M.; Wei, X.; Yu, J.; Li, Z.; and Ding, B. 2020. Highly wearable, breathable, and washable sensing textile for human motion and pulse monitoring. *ACS applied materials & interfaces*, 12(17): 19965–19973.

Luo, Y.; Li, Y.; Sharma, P.; Shou, W.; Wu, K.; Foshey, M.; Li, B.; Palacios, T.; Torralba, A.; and Matusik, W. 2021. Learning human–environment interactions using conformal tactile textiles. *Nature Electronics*, 4(3): 193–201.

Negre, M.; Jorda, M.; Vardoulis, O.; Chortos, A.; Khatib, O.; and Bao, Z. 2018. A hierarchically patterned, bioinspired e-skin able to detect the direction of applied pressure for robotics. *Sci. Robotics*, 3(24).

Noh, B.; Youm, C.; Goh, E.; Lee, M.; Park, H.; Jeon, H.; and Kim, O. Y. 2021. XGBoost based machine learning approach to predict the risk of fall in older adults using gait outcomes. *Scientific reports*, 11(1): 1–9.

Okunevich, I.; Trinitatova, D.; Kopanev, P.; and Tsetserukou, D. 2021. DeltaCharger: Charging Robot With Inverted Delta Mechanism and CNN-Driven High Fidelity Tactile Perception for Precise 3D Positioning. *IEEE Robotics Autom. Lett.*, 6(4): 7604–7610.

Sundaram, S.; Kellnhofer, P.; Li, Y.; Zhu, J.; Torralba, A.; and Matusik, W. 2019. Learning the signatures of the human grasp using a scalable tactile glove. *Nat.*, 569(7758): 698–702.

Tong, Z.; Song, Y.; Wang, J.; and Wang, L. 2022. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. *CoRR*, abs/2203.12602.

Tu, D.; Sun, W.; Min, X.; Zhai, G.; and Shen, W. 2022. Video-based Human-Object Interaction Detection from Tubelet Tokens. *arXiv preprint arXiv:2206.01908*.

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

Wicaksono, I.; Hwang, P. G.; Droubi, S.; Wu, F. X.; Serio, A. N.; Yan, W.; and Paradiso, J. A. 2022. 3DKnITS: Three-dimensional Digital Knitting of Intelligent Textile Sensor for Activity Recognition and Biomechanical Monitoring. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2403–2409. IEEE.

Yan, S.; Xiong, X.; Arnab, A.; Lu, Z.; Zhang, M.; Sun, C.; and Schmid, C. 2022a. Multiview Transformers for Video Recognition. In *The IEEE / CVF Computer Vision and Pattern Recognition Conference (CVPR)*.

Yan, S.; Xiong, X.; Arnab, A.; Lu, Z.; Zhang, M.; Sun, C.; and Schmid, C. 2022b. Multiview Transformers for Video Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 3323–3333. IEEE.

Zerveas, G.; Jayaraman, S.; Patel, D.; Bhamidipaty, A.; and Eickhoff, C. 2021. A Transformer-based Framework for Multivariate Time Series Representation Learning. In Zhu, F.; Ooi, B. C.; and Miao, C., eds., *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, 2114–2124. ACM.

Zhao, Y.; Li, Z.; Guo, X.; and Lu, Y. 2022. Alignment-guided Temporal Attention for Video Action Recognition. *arXiv preprint arXiv:2210.00132*.

Zhu, M.; Shi, Q.; He, T.; Yi, Z.; Ma, Y.; Yang, B.; Chen, T.; and Lee, C. 2019. Self-powered and self-functional cotton sock using piezoelectric and triboelectric hybrid mechanism for healthcare and sports monitoring. *ACS nano*, 13(2): 1940–1952.