STAM: An Attention Model for Tactile Texture Recognition

Guanqun Cao[1] and Shan Luo[1]

***Abstract*— In the past few years, tactile sensing has attracted great interest in robotics. A detailed understanding of the surface textures via tactile sensing is essential for tasks like exploration and manipulation. Previous works on texture recognition have been limited to treating all the regions in one tactile image or all the samples in one tactile sequence equally, which includes much irrelevant information. In this paper, we propose a novel Spatio-Temporal Attention Model (STAM) for tactile texture recognition, which pays attention to both spatial focus of each single tactile texture and the temporal correlation of a tactile sequence. The improved tactile texture perception can be applied to facilitate robot tasks like grasping and manipulation.**

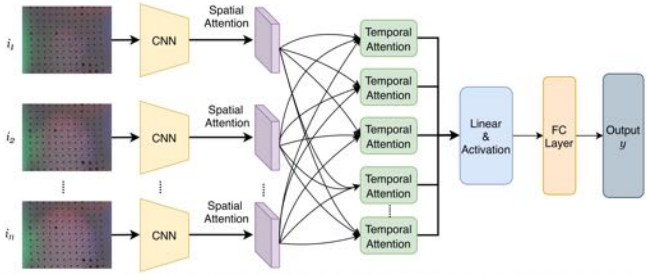

Fig. 1: ***The proposed STAM framework.***

I. Introduction

The sense of touch is one of the important information sources for both humans and robots to perceive the object properties in the physical world. One of the key object properties is the surface texture that is important for object recognition and dexterous manipulation of objects.

Tactile sensors have been used to discriminate surface textures to embody robots the sense of touch [1]–[3]. Similar to video sequences collected by a camera, information is also accumulated over a period of time by a tactile sensor, a GelSight tactile sensor [4] used in this paper. As the sensor scans a surface, the surface texture is recorded and tactile data is collected in a temporal order. The temporal patterning and correlation of tactile sequences are crucial to interpreting the stimulus of surface textures.

Humans perceive the surface textures temporally with tactile spatial events presented in sequences as well. When we use our fingers to scan an object surface, both spatial and temporal changes in skin deformation provide important cues for fine texture perception. In this exploratory procedure, we experience the ***tactile selective attention*** [5]: in the perceptual area of fingers, humans focus their attention on the points that give more excitement rather than treating the whole contacting region equally. On the other side, the perception is an accumulation of cognition that the previous contact events enable a prior knowledge for the perception and later contacts verify the previous judgement. In this paper, we propose a novel Spatio-Temporal Attention Model (STAM) to extract the tactile texture features spatially and temporally. We implement the attention model in a task of fabric texture recognition, with 100 pieces of fabrics and a GelSight sensor.

All the authors are at the Department of Computer Science, University of Liverpool, Liverpool L69 3BX, U.K. Emails: {g.cao, shan.luo}@liverpool.ac.uk.

II. Methodologies

As illustrated in Fig. 1, the STAM model consists of three parts: 1) CNNs that extract spatial features from each input tactile image; 2) A spatial attention module which highlights the salient features in each tactile texture; 3) Temporal attention modules which are used to model the correlation of salient features in different tactile images in one sequence.

A. CNN module

Each of the tactile images in the tactile sequence $I = \{i_1, i_2, ..., i_n\}$ is first fed into a pre-trained AlexNet architecture [6] to extract the spatial features. We take the output feature map $\boldsymbol{F} \in \mathbb{R}^{h \times w \times c}$ from the last max-pooling layer as the input to the spatial attention module, where h, w, c refer to the height, width and the number of channels respectively.

B. Spatial Attention Module

In order to emphasize informative areas in each texture frame, we develop a spatial attention module to assign higher weights to more crucial areas, whereas lower weights are assigned to the areas that contain less information. We apply two pooling operations, i.e., max-pooling and average-pooling, to the spatial feature $\boldsymbol{F}$ obtained from the CNN module along channel axis to form spatial context descriptors. The average-pooling is applied to learn tactile information effectively while max-pooling is adopted to maintain prominent features. These two pooling layers generate two spatial features $\boldsymbol{F}^S_{max}$ and $\boldsymbol{F}^S_{avg}$ respectively. After that, the $\boldsymbol{F}^S_{max}$ and $\boldsymbol{F}^S_{avg}$ are concatenated and convolved by a 7×7 convolution and a sigmoid layer to produce a 2D spatial attention map $\boldsymbol{A_S(F)}$:

$$\begin{aligned} A_S(F) &= \sigma(f^{7\times7}([MaxPool(\boldsymbol{F}); AvgPool(\boldsymbol{F})])) \\ &= \sigma(f^{7\times7}([\boldsymbol{F}^S_{max}; \boldsymbol{F}^S_{avg}])), \end{aligned} \tag{1}$$

where σ denotes the sigmoid function. Then we get the output feature map $\boldsymbol{F^S} = \boldsymbol{A_S(F)} * \boldsymbol{F}$ from the spatial attention module, where $*$ refers to the element-wise multiplication.

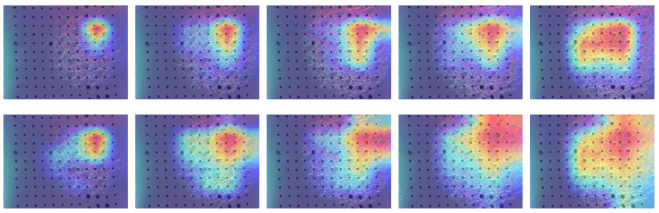

Fig. 2: ***Spatial Attention Distribution.*** The first row represents the non-attention heat maps while the second row represents the spatial attention heat maps. As can be seen, more regions are activated by spatial attention mechanisms.

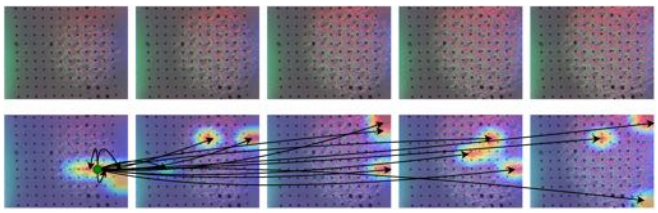

Fig. 3: ***Temporal Attention Distribution.*** The first row represents the raw tactile textures in a sequence. The highlighted regions in each image in the second row are 3 most related clues of the green dot region to support the recognition.

TABLE I: Recognition results of different models while different numbers of tactile images are used in a sequence.

Models	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$
CNNs	67.23%	72.04%	75.26%	78.06%	79.56%	81.29%
CNNs+Spatial Attention	72.12%	73.97%	78.60%	80.43%	80.43%	80.86%
STAM	**76.50%**	**79.35%**	**80.00%**	**80.64%**	**81.72%**	**81.93%**

TABLE II: Recognition results while some noisy data is included in the dataset.

Models	$n=2$	$n=3$	$n=4$	$n=5$	$n=6$	$n=7$
CNNs	53.20%	58.20%	59.60%	61.23%	64.60%	69.40%
CNNs+Spatial Attention	55.40%	60.80%	62.60%	62.80%	65.40%	71.00%
STAM	**72.00%**	**72.20%**	**75.80%**	**76.61%**	**80.80%**	**80.20%**

C. Temporal Attention Module

After obtaining the extracted features of spatial attention module from each texture frame, we concatenate all the features together to get a sequence of spatial features $\boldsymbol{F^{S(n)}}$. In order to model the long-distance dependency in tactile sequence, we develop a temporal attention module on top of the spatial attention layer. This temporal attention mechanism aims to estimate the salience and relevance of all the locations through the time regardless of distance, which takes the information from global observations into consideration. $\boldsymbol{F^{S(n)}}$ is first converted into two feature spaces $q(\boldsymbol{F^{S(n)}})$ and $k(\boldsymbol{F^{S(n)}})$ by two sets of $1\times1\times1$ convolutions, where $q(\boldsymbol{F^{S(n)}}) = W_q\boldsymbol{F^{S(n)}}$ and $k(\boldsymbol{F^{S(n)}}) = W_k\boldsymbol{F^{S(n)}}$ (W_q and W_k are trainable weight matrices). The attention map $\boldsymbol{A_T(F^{S(n)})}$ is given as follows:

$$\boldsymbol{A_T(F^{S(n)})}_{j,i} = \frac{\exp\left(s_{ij}\right)}{\sum_{i=1}^{m}\exp\left(s_{ij}\right)}, \tag{2}$$

where $s_{ij} = q\left(F_i^{S(n)}\right)k\left(F_j^{S(n)}\right)^T$ and $m = n\times h\times w$. $\boldsymbol{A_T(F^{S(n)})}_{j,i}$ demonstrates the extent to which temporal attention refers to the i^{th} feature while updating the j^{th} feature of any locations. The output feature map of the temporal attention is $\boldsymbol{F^T} = (F_1^T, F_2^T, ..., F_j^T, ..., F_m^T)$ and $\boldsymbol{F_j^T}$ can be given as follows:

$$\boldsymbol{F_j^T} = \sum_{i=1}^{m}\boldsymbol{A_T(F^{S(n)})}_{j,i}\boldsymbol{v}\left(\boldsymbol{F_i^{S(n)}}\right) + \boldsymbol{F_j^{S(n)}} \tag{3}$$

where $\boldsymbol{v}\left(\boldsymbol{F_i^{S(n)}}\right) = W_v\boldsymbol{F_i^{S(n)}}$ is a linear transformation of $\boldsymbol{F_i^{S(n)}}$. $\boldsymbol{F_j^{S(n)}}$ is added back to avoid gradient vanishing.

III. Experiments

We first conduct an ablation study with different neural network frameworks to learn how the spatial attention and temporal attention help to improve the recognition. Moreover, to study the impact of the length of each input sequence on our methods, we vary the length n of a sequence from 2 to 7. As illustrated in Table I, with the help of spatial attention and temporal attention, the recognition accuracy improves step by step under most of the cases as the length of sequence increases. Note that some tactile images are collected before the GelSight sensor contacts the objects that cannot provide useful information for the recognition. We include these tactile images as noisy data to verify the robustness of the models, shown in Table II. It can be seen that the performance of our proposed STAM model still maintains at the same level while the other models cannot sustain the recognition accuracy compared with the accuracy in Table I, which shows the strong robustness of the STAM model against the noisy data. We also visualize the attention maps, illustrated in Fig. 2 and Fig. 3, to demonstrate the effectiveness of spatial and temporal attention modules.

IV. Conclusion

In this paper, we investigate the attention mechanism in robotic tactile perception, for the first time. Our proposed methods has resulted in significant improvement of the recognition accuracy, by up to 18.8%, compared to the non-attention based models. The improved tactile texture perception can be applied to facilitate robot tasks like grasping specific objects and manipulation.

References

[1] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *Mechatronics*, 2017.

[2] S. Luo and et al., "ViTac: Feature sharing between vision and tactile sensing for cloth texture recognition," in *ICRA*, 2018.

[3] J.-T. Lee, D. Bollegala, and S. Luo, ""touching to see" and "seeing to feel": Robotic cross-modal sensory data generation for visual-tactile perception," in *ICRA*, 2019.

[4] W. Yuan, S. Dong, and E. Adelson, "Gelsight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors*, 2017.

[5] K. Sathian and H. Burton, "The role of spatially selective attention in the tactile perception of texture," *Perception & Psychophysics*, 1991.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *NeurIPS*, 2012.