VISION2TOUCH: IMAGING ESTIMATION OF SURFACE TACTILE PHYSICAL PROPERTIES

*Jie Chen, Shizhe Zhou**

College of Computer Science and Electronic Engneering
Hunan University, Changsha, Hunan, China.
{chenjie98, shizhe}@hnu.edu.cn

ABSTRACT

Similar to the human's multiple perception system, the robot can also benefit from cross-modal learning. The connection between visual input and tactile perception is potentially important for automated operations. However, establishing an algorithmic mapping of the visual modal to the tactile modal is a challenging task. In this work, we use the framework of GANs to propose a cross-modal imaging method for estimating the tactile physical properties values based on the Gramian Summation Angular Field, combined with visual-tactile embedding cluster fusion and feature matching methods. The approach estimates 15 tactile properties. In particular, the task attempts to predict unknown surface properties based on "learned knowledge". Our results surpass the state-of-the-art approach on most tactile dimensions of the publicly available dataset. Additionally, we conduct a robustness study to verify the effect of angle and complex environment on the network prediction performance.

Index Terms— Visual-Tactile, Physical Properties Estimation, Generative Adversarial Network, Cross-Modal

1. INTRODUCTION

People live in a world full of a wide variety of modal information. In order for artificial intelligence to be advanced enough to understand the world around us, it needs to be able to reason and interpret such multimodal signals together and to enable cross-modal learning[1]. Among these modalities, vision and touch are two important and interrelated perceptual channels. Cross-modal connections between vision and touch can enable robots to more effectively handle various objects and environments in both industrial settings and our daily lives, furthermore, can improve the ability of robots to interact with unknown environments and objects. In previous work, vision-based sensing technologies have been widely used in various robotic work scenarios, such as object recognition[2] and tracking[3], object detection[4] and driving navigation[5]. In addition, tactile sensor-based robots can perform tasks like object grasping and surface vibrotactile recognition. Some recent work[6, 7] has attempted to use visual and tactile sensors together to integrate and transform visual and tactile signals in order to mimic human cross-modal perception. Unlike the indirect visual-tactile perception tasks in existing robotics, our work attempts to generate physical tactile properties of the surface from visual images directly.

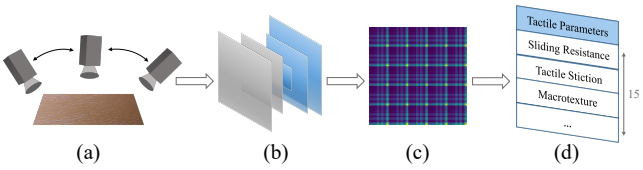

Fig. 1. **Workflow.** The main concept of imaging estimation using GAN-based methods. (a) : Visual data acquisition under the standard RGB camera. (b) : Cross-modal visual-tactile generation model. (c) : Generated results with tactile information. (d) : Predicted values reduced from the generated results.

The main objective of our work is to find a tighter mapping between vision and touch so that we can create a network that estimates tactile physical properties directly from visual information as shown in Fig. 1. In particular, our study expects the network to be able to estimate the properties of unknown textures based on "learned knowledge", which is a challenging task. To this end, we introduced a deep-learning-based framework that learns the complex relationship between visual perception and the tactile physical properties of surfaces such as friction, roughness, compliance, thermal conductance, and so on. The presented framework is built upon the base of the generative adversarial networks(GANs) with the GASF-based encoding framework, and trained with the additional WGAN-GP, feature-matching losses and visual-tactile embedding cluster fusion module. Our encoding framework can help reduce the scale discrepancy between visual images and tactile vectors, and the inclusion of the cluster fusion module and optimization strategy improves the generative performance and robustness of the network.

This work was funded by the National Science Foundation of China No.62076090, Huxiang Youth Talent Support Program, Hunan Province China, No.2020RC3014, Natural Science Foundation of Hunan Province, China, No.2022JJ30173.

Compared with previous work, our method can generate results closer to the ground truth on multiple tactile attributes, achieving better average prediction results. We also conduct robustness experiments to consider the effects of angle and complex environments on the network generation performance. The experimental results show that our modal can output accurate predicted values from a single random angle image of a material with robustness against illumination, angle changes and noise.

2. RELATED WORK

Our experience of the world is multimodal. Multimodel learning is a vibrant field. For example, OpenAI kicks off a big year in multimodal learning with CLIP[8], which matches images and text. Dall·E[9] generates images that correspond to input text. Our work falls under the umbrella of cross-modal translation, which is an important research task in multimodal learning. Cross-modal translation can be classified into two types, example-based, and generative, where generative translation are considered to be a more challenging problem due to its need of the ability to generate signal or symbol sequences. Several prior research works[10, 11, 12] show generative adversarial networks with effective generation performances on cross-modal translation.

In the field of robotics, the addition of tactile property information can enrich the physical properties of the perceived object and help the robot decide in advance how to interact with the environment. Liu et al.[13] constructed a cross-modal perception from ground images to tactile signals based on the CycleGAN framework, which helps visually impaired people sensing the ground and brings a better traveling experience for them. TactGAN[14] is based on the dataset created by Strese[15] and attempts to learn the mapping relationship between vision and tactile by synthesizing real tactile signals from visual inputs. The goal of our research goes beyond the general task and hopes to estimate tactile properties directly from visual information. Our study use the dataset presented in the work[16] and compare our proposed method with it. In our work, we model visual to tactile cross-modal prediction as an image-to-image generation problem based on the bijective property of GASF coding. The 15 tactile property predictions can be obtained from the output. More details will be presented in Section 3.1.

3. METHODOLOGY

We study the problem of translating from the visual to the tactile domain, which can be modeled as a conditional image generation framework with a GASF-based encoding module. Fig. 2 shows the structure of our network. In particular, we focus on the construction of the imaging coding framework and the process of outputting physical tactile vectors from the generated results.

3.1. Imaging Coding Framework

Researchers are paying attention to how to reformulate sequential features as visual clues. Donner et al.[17] construct adjacency matrices from predefined recursive functions to convert time series into complex networks. Wang et al.[18] propose the Gramian Summation Angular Field(GASF) coding framework to encode time series into images. Motivated by these works, our work tries to represent the tactile data as a 2D image/matrix with the GASF-based imaging encoding framework. The GASF framework preserves the temporal and spatial information of the sequence allowing the machine to learn the structure and patterns of the sequence visually. At the same time, the GASF coding has an accurate inverse map on the [0, 1] interval. The particular bijective property lays the foundation for GAN-based imaging estimation.

We first connect the tactile vectors at five locations to obtain the tactile sequence $X = \{x_1, x_2, \ldots, x_n\}$, and rescale it so that all values can fall within the interval $[0, 1]$. Thus we can use a polar coordinate system to represent this tactile sequence X with the equation below:

$$\phi = \arccos(x_i), 0 \leq x_i \leq 1, x_i \in X. \tag{1}$$

Rescaled data in interval $[0, 1]$ corresponds to the cosine angle $\in \left[0, \frac{\pi}{2}\right]$. After converting to the polar coordinate system, we take the tactile sequence as a 1-D metric space, and by defining the inner product $\langle x, y\rangle = x \cdot y - \sqrt{1-x^2} \cdot \sqrt{1-y^2}$ we can define the GASF as follows:

$$GASF = [\cos(\phi_i + \phi_j)] \tag{2}$$

$$= X^{'} \cdot X - \sqrt{I - (X')^2} \cdot \sqrt{I - X^2}. \tag{3}$$

I is the row vector $[1, 1, \ldots, 1]$. In this work we are particularly interested in the potential spatial information retained by the encoded tactile vector, which can help mitigate the scale discrepancy between the visual and tactile domains. In fact, this way of defining the inner product $k(x_i, x_j)$ increases the dimensions of the original data to implement data augmentation, which is usually equivalent to a kernel trick. As mentioned above, the mapping functions of 0/1 rescaled data are bijections. Given a tactile series, the proposed map produces one and only one result in the polar coordinate system with a unique inverse map. Actually, The main diagonal $GASF_{i,i}$ is the special case that contains the original value information. From the main diagonal $\{GASF_{i,i}\} = \{\cos(2\phi_i)\}$, we are allowed to precisely reconstruct the original sequences by:

$$\cos(\phi) = \sqrt{\frac{\cos(2\phi) + 1}{2}}, \phi \in \left[0, \frac{\pi}{2}\right]. \tag{4}$$

Thus, we can estimate tactile physical property vectors by recovering the underlying sequence information in the generation results.

3.2. Clustering Fusion Module

Previous work on cross-modal generation has demonstrated that strong supervision by adding auxiliary classification targets[19] can guide the output of more reasonable results. Residue-Fusion GAN[12] obtain better generative results by

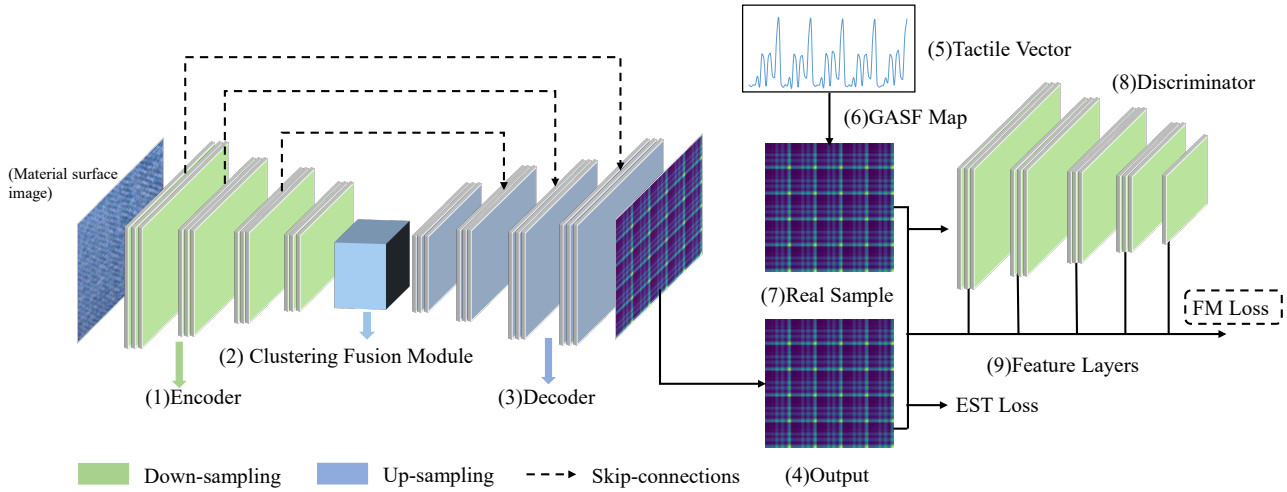

Fig. 2. **The overview of the proposed framework.** Our model passes the input images to the encoder(1) and the decoder(3). The clustering Fusion module(2) is constructed with 4 residual blocks. The real image obtained from the input physical tactile vector(5) processed by the Gramian Summation Angular Field(GASF) coding framework(6), with the generated image(4) is input to the discriminator(8) for conditional adversarial training. We also extract the output from the feature layers of the discriminator(8) for the computation of the feature matching loss(9).

adding a residual fusion module to the generative model, which allows the network to extract more label information from the input visual modality. However, in general tasks, semantic labels contain information about material categories and do not always provide rich tactile information, e.g. plastics come in many forms with distinct surface properties but fall under one label. Therefore, in our work, we extend the unsupervised clustering fusion module to the network by creating visual-tactile embedding labels instead of semantic labels. Considering the excellent performance of deep clustering algorithms in image clustering tasks, we use a classical deep clustering approach, the Deep Adaptive Clustering framework (DAC)[20], to create our clustering target. We concatenate visual data and tactile information to obtain visual-tactile representations and input them into the adaptive clustering framework. In DAC, similarities are computed based on the cosine distance between label features generated by the deep convolutional network. We remove the final fully connected layer, extract the feature representation of the label information, and then upsample the label feature information through 4 layers of residual blocks to connect it with the feature vector provided by the generator downsampling process as a clustering fusion module.

3.3. Cross-Modal Network

Our cross-modal learning model consists of a generator and a discriminator. For the generator, we use the U-Net model[21] structure as the network backbone. Prior work adopts a strategy of jumping connections between the upsampling and downsampling layers to let this information pass directly through the network and achieve effective results. Thus in the generator model we skip connections between each layer i and the layer $n-i$, where n is the total number of layers.

The discriminator uses the structure of patchGAN[10], which effectively models the image as a Markov random field. Regular GANs map from an image to a single scalar output, which signifies "real" or "fake", whereas the patchGAN maps from to an $N \times N$ array of outputs X, where each $X_{i,j}$ signifies whether the $patch_{i,j}$ in the image is real or fake. Such structure fuses local image features with overall image features, allowing for more preservation of the image in generation tasks. In our model, the final receptive fields of the discriminator turn out to be 70×70 patches in the input image.

A common approach to improving the stability of the network is to adopt WGAN-GP loss[22] as L_{adv-gp} to apply a gradient penalty to each sample independently. Another way is to add feature matching loss L_{fm} during the training process[23]. We extract the feature outputs from multiple layers of the discriminator and match these feature representations from the real and the generated outputs with $L1$ distance. In addition, we calculate the MSE distance as L_{est} between the generated result and the ground truth. Hence, our final objective functions of the proposed network are shown below:

$$L_{fm} = \mathbb{E}_{y\sim p_{(y)},\tilde{y}\sim p_{(\tilde{y})}} \sum_{i=1}^{T} \frac{1}{N_i} \parallel D^{(i)}(y) - D^{(i)}(\tilde{y}) \parallel_1 . \tag{5}$$

$$L = L_{adv-gp} + \alpha L_{fm} + \beta L_{est}. \tag{6}$$

Here, y and $\tilde{y}$ denote the real samples and generated samples, $p_{(y)}$ and $p_{(\tilde{y})}$ denote the distributions of the real and generated data. And T is the total number of layers in the discriminator

D, $D^{(i)}$ represents the features in the i-th layer, and N_i is the number of elements in $D^{(i)}$.

4. EXPERIMENTS

4.1. Dataset

We use the Surface Property Synesthesia dataset[16] as our dataset. The dataset provides a collection of surface images of 421 materials taken at different angles by a standard RGB camera under a diffuse light source. 15 surface tactile physical properties are measured with the Biotac Tactile sensing device. The Biotac Tactile Sensor has three sets of independent sensors. When the Biotac core is in contact with a surface, multiple tactile signals are recorded and calibrated to output 15 tactile physical properties.

4.2. Comparison and Ablation Study

To ensure a fair comparison with the method validated in [16], we follow the experimental setup described in[16] and randomly divide the 400+ visual-tactile pairs into 90/10 training/validation splits. For the comparison experiments, the image from the lowest viewing point,i.e.the -45° along the roll axis of the surface is selected as input. It can be seen from the comparison Table 1 that our proposed single-image estimation method outperformed the state-of-the-art results obtained in [16] on several tactile dimensions, obtaining the best average R^2 in the single-image estimation comparison task, and the lowest MAE score in all single/multiple image experiments. In addition, we implement three different models in the ablation experiment(Table 2) to investigate the effectiveness of each optimization module. In our ablation study, the visual-tactile cluster fusion module has the greatest impact on the generation results. By adding pre-trained visual-tactile embedding clustering fusion information to the latent space of the generator, the ability to generate cross-modal results based on unknown input types can be improved.

Table 1. **The tactile estimation results(R^2).** The comparison experiment results are displayed. Red and blue text correspond to the first and second best scores respectively.

Properties	Baseline(single)[16]	DEC (single)[16]	NVS (multi)[16]	VB-NVS (multi)[16]	Ours
fRS	0.07	0.54	0.62	0.65	0.09
cDF	0.49	0.52	0.53	0.50	0.58
tCO	0.50	0.62	0.63	0.61	0.76
cYD	0.44	0.64	0.55	0.57	0.70
aTK	−0.46	−0.07	0.02	−0.05	0.53
mTX	0.43	0.43	0.56	0.58	0.59
cCM	0.13	0.47	0.53	0.57	0.63
cDP	0.35	0.67	0.54	0.64	0.56
cRX	0.11	0.44	0.49	0.45	0.48
mRG	0.46	0.47	0.55	0.57	0.21
mCO	0.56	0.54	0.68	0.70	0.73
uRO	0.32	0.44	0.51	0.47	0.69
tPR	0.57	0.65	0.54	0.68	0.36
uCO	0.57	0.59	0.63	0.66	0.52
fST	0.53	0.59	0.64	0.61	0.54
MeanR2	0.34	0.50	0.53	0.55	0.53
MeanMAE	6.17	5.53	5.34	5.28	4.97

4.3. Robustness Study

Considering the possible light angle deviation and complex environment during robot operation, we conduct a robustness study to verify the effect of angle and complex environment on the network prediction performance. To test the effect of angle on our model, we randomly selected images with angular deviations from the training data in the range of 1°-45° and 1°-25° to form the test set.

For the interference test, we simulate different luminance environments, and add Gaussian noise, salt noise, pepper noise, and salt & pepper noise to the test set respectively. The results are shown in Table 2. In general, the experiments show that our model is robust to illumination, viewing angles and noisy environments.

Table 2. **The tactile estimation results(R^2).** The ablation and robustness experiment average results are displayed. For R^2 metric, higher values are better. For MAE and FID, lower values are better. Bolded text indicates the best score.

	MeanR2	MeanMAE	FID
w/o L_{fm}	0.44	5.60	33.84
w/o L_{est}	0.48	5.47	32.05
w/o Cluster fusion	0.09	7.81	54.80
lighting↑	0.49	5.11	**24.28**
lighting↓	0.49	5.16	27.91
Gaussian noise	0.44	5.83	30.70
salt noise	0.44	5.65	25.80
pepper noise	0.44	5.59	30.12
s&p noise	0.49	5.34	30.33
angle(1°-45°)	0.39	6.18	27.11
angle(1°-25°)	0.47	5.54	24.45
Ours	**0.53**	**4.97**	**24.28**

5. CONCLUSIONS

In this work, we propose a method of imaging predictive physical tactile properties based on a GASF encoder. Finding and understanding the connection between visual and tactile information is a challenging task. The experiment results show that the prediction ability of our model is superior to all comparison methods. Ablation research is conducted to reveal the effectiveness of cluster fusion and feature matching methods. The robustness study shows that our method is robust to changing environments and noise disturbances. Our method can be potentially applied to a variety of robot operation tasks with the adaptation to the disturbed working environment.

6. REFERENCES

[1] Tadas Baltrusaitis, Chaitanya Ahuja, Louis Philippe Pattern Analysis Morency, and Machine Intelligence, "Multimodal machine learning: A survey and taxonomy," vol. PP, no. 99, pp. 1–1, 2017.

[2] Sinapov, J., Sukhoy, V., Sahai, R., Stoytchev, A. Robotics: A publication of the IEEE Robotics, and Automation Society, "Vibrotactile recognition and categorization of surfaces by a humanoid robot," vol. 27, no. 3, pp. 488–497, 2011.

[3] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad preprint arXiv:.00831 Schindler, "Mot16: A benchmark for multi-object tracking," 2016.

[4] Andreas Geiger, Philip Lenz, Christoph Stiller, and Raquel Urtasun, "Vision meets robotics: The kitti dataset," *International Journal of Robotics Research (IJRR)*, 2013.

[5] Adam Ligocki, Ales Jelinek, and Ludek Zalud, "Brno urban dataset-the new data for self-driving agents and mapping tasks," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2020, pp. 3284–3290.

[6] Y. Li, J. Y. Zhu, R. Tedrake, and A. Torralba, "Connecting touch and vision via cross-modal prediction," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[7] N. Heravi, W. Yuan, A. M. Okamura, and J. Bohg, "Learning an action-conditional model for haptic texture generation," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*.

[8] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, and Jack arXiv preprint arXiv:.00020 Clark, "Learning transferable visual models from natural language supervision," 2021.

[9] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya "Zero-shot text-to-image generation," 2021.

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134.

[11] Shaoyu Cai, Yuki Ban, Takuji Narumi, and Kening Zhu, "Frictgan: Frictional signal generation from fabric texture images using generative adversarial network," in *ICAT-EGVE*, pp. 11–15.

[12] Shaoyu Cai, Kening Zhu, Yuki Ban, Takuji Automation Letters, "Visual-tactile cross-modal data generation using residue-fusion gan with feature-matching and perceptual losses," vol. 6, no. 4, pp. 7525–7532, 2021.

[13] Huaping Liu, Di Guo, Xinyu Zhang, Wenlin Zhu, Bin Fang, Fuchun Transactions on Automation Science Sun, and Engineering, "Toward image-to-tactile cross-modal perception for visually impaired people," vol. 18, no. 2, pp. 521–529, 2020.

[14] Yusuke Ujitoko and Yuki Ban, "Vibrotactile signal generation from texture images or attributes using generative adversarial network," in *International Conference on Human Haptic Sensing and Touch Enabled Computer Applications*. pp. 25–36, Springer.

[15] Matti Strese, Clemens Schuwerk, Albert Iepure, and Eckehard transactions on haptics Steinbach, "Multimodal feature-based surface material classification," vol. 10, no. 2, pp. 226–239, 2016.

[16] Matthew Purri and Kristin Dana, "Teaching cameras to feel: Estimating tactile physical properties of surfaces from images," in *European Conference on Computer Vision*. Springer, 2020, pp. 1–20.

[17] R. V. Donner, Y. Zou, J. F. Donges, N. Marwan, and J. Kurths, "Recurrence networks - a novel paradigm for nonlinear time series analysis," vol. 12, no. 3, pp. 129–132, 2010.

[18] Zhiguang Wang and Tim "Imaging time-series to improve classification and imputation," 2015.

[19] B. Duan, W. Wang, H. Tang, H. Latapie, and Y Yan, "Cascade attention guided residue learning gan for cross-modal translation," in *2020 25th International Conference on Pattern Recognition (ICPR)*.

[20] Jianlong Chang, Lingfeng Wang, Gaofeng Meng, Shiming Xiang, and Chunhong Pan, "Deep adaptive image clustering," in *Proceedings of the IEEE international conference on computer vision*, pp. 5879–5887.

[21] Olaf Ronneberger, Philipp Fischer, and Thomas Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. pp. 234–241, Springer.

[22] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville, "Improved training of wasserstein gans," 2017.

[23] Anders Boesen Lindbo Larsen, Sren Kaae Snderby, Hugo Larochelle, and Ole "Autoencoding beyond pixels using a learned similarity metric," 2015.