Toward Vision-Based Object Compliance Estimation

Malte Kuhlmann[1], Ziteng Li[1], and Nicolás Navarro-Guerrero[1]

***Abstract*— Estimating compliance is an important skill in many areas, such as agriculture and the biomedical field. Traditional methods are highly specialized for specific use cases and often suffer from high costs and low portability, making them unsuitable for robotic applications. Due to the increased popularity of vision-based tactile sensors, modern solutions are shifting towards utilizing tactile images for compliance estimation. However, these neural network based solutions still suffer insufficient prediction accuracy, utilize additional analytical features and do not employ state-of-the-art techniques. We propose two new model architectures based on Recurrent Convolutional Networks and Transformers. Our proposed models display improved prediction accuracy across multiple metrics and show that analytical input features are unnecessary to achieve state-of-the-art performance.**

I. Introduction

The crucial role of compliance estimation in various domains, such as agriculture [1], the biomedical field [2], [3] or grasping [4], [5], has been extensively documented in the literature. In this context, compliance is understood as the general concept of stiffness of an object [6], as defined by various metrics, including Young's Modulus or the Shore Hardness.

Traditionally, object compliance estimation required specialized equipment that is only usable in specific use cases [7], [8], [9], [10], [11], [12]. Such solutions are often unsuitable for stiffness sensing regarding affordability and portability, an important ability in robotic dexterous hands. In recent years, vision-based tactile sensors have gained traction in many tasks [13]. These sensors capture tactile information by taking RGB images of their gel surface deformation. Due to their low cost and small size, they are very suitable for robotic compliance and stiffness estimation.

Recently, research has been done on compliance estimation utilizing tactile sensors. For instance, Yuan et al. [14] proposed a recurrent neural network that utilizes five images from a GelSight sensor [15] to predict the Shore 00 hardness of objects. Their dataset did not cover the entire range of hardness because the Shore 00 hardness scale was used as their target value. Burgess and Zhao [16] similarly proposed a model based on convolutional neural networks to estimate compliance. They reported a $\log_{10}$ accuracy of 0.804 but did not utilize state-of-the-art techniques, leaving room for improvements. Therefore, we propose two new frameworks for detecting object compliance based on vision-based tactile sensors, applied force and gripper width.

[1]L3S Research Center, Leibniz Universität Hannover, Hanover, Germany {malte.kuhlmann, ziteng.li, nicolas.navarro}@l3s.de

II. Method

A. Dataset

We use the dataset released by Burgess and Zhao [16]. The dataset covers 285 objects with Young's Modulus ranging from $1e3$ to $1e12$, containing soft and hard materials. Each data point in the dataset consists of three tactile images captured using a GelSight sensor [15]. They represent the start, middle and end of a touch trajectory. The normal force, the width of the gripper and analytical estimations of the Young's Modulus, which were calculated through Hooke's Law [17] and Hertzian contact theory [18], were also captured.

B. Model frameworks

As our baseline, we utilize the model Top10NN proposed by Burgess and Zhao [16]. It utilizes Convolutional Neural Networks (CNN) to extract features from the three images. These features are then concatenated with the optional information (normal force, the width of the gripper and the analytical estimation of the Young's Modulus) and used as the input into a decoder, which then returns the estimated Young's Modulus. The architecture of this model does not employ a pre-trained image feature extractor, which has been proven to extract information from images more efficiently [19]. Furthermore, other architectures exist, such as Transformer [20], that have been proven to work well with time-series tasks [21].

The first model we propose, VGG-LSTM, is based on Long-term Recurrent Convolutional Networks [22]. Inspired by Yuan et al. [14], we utilize the CNN features from the $fc7$ layer of the pre-trained VGG16 model [19]. These features and, optionally, the three additional inputs contained in the dataset are used as the input for the LSTM blocks. Since we have three different images, our model consists of three LSTM blocks, each LSTM block also integrating the information from the previous image. The outputs of each LSTM block are then averaged using learnable weighted averaging, which returns the estimated Young's Modulus.

The second model we propose, Res-Tf, is based on Residual Networks [23] and incorporates a Transformer encoder [20]. We employ a pre-trained ResNet model to extract features from the three tactile images. These features are then concatenated with their respective positional data, which is then encoded by a Transformer encoder. This encoding is combined with the optional input and decoded by a final decoder block to estimate the Young's Modulus.

C. Experimental Design

Burgess and Zhao [16] preprocessed the dataset by augmenting the dataset with randomized image flipping and

TABLE I
THE COMPARISON OF OVERALL PERFORMANCE AMONG ALL MODELS.

Method	Input				Random				Stratified			
					Seen-Object		Unseen-Object		Seen-Object		Unseen-Object	
	RGB	F	W	E	$\log_{10}$ accuracy	N-MSE	$\log_{10}$ accuracy	N-MSE	$\log_{10}$ accuracy	N-MSE	$\log_{10}$ accuracy	N-MSE
Top10NN	✓	-	-	-	0.765 ± 0.0123	0.0165 ± 0.0009	0.455 ± 0.0233	0.0374 ± 0.0020	0.777 ± 0.0191	0.0136 ± 0.0009	0.541 ± 0.0229	0.0379 ± 0.0020
	✓	✓	-	-	0.785 ± 0.0367	0.0151 ± 0.0016	0.426 ± 0.0366	0.0385 ± 0.0026	0.806 ± 0.0165	0.0123 ± 0.0008	0.510 ± 0.0314	0.0360 ± 0.0014
	✓	✓	✓	-	0.783 ± 0.0185	0.0153 ± 0.0011	0.427 ± 0.0434	0.0332 ± 0.0028	0.808 ± 0.0089	0.0124 ± 0.0004	0.530 ± 0.0260	0.0346 ± 0.0011
	✓	✓	✓	✓	0.733 ± 0.0088	0.0182 ± 0.0010	0.447 ± 0.0433	0.0326 ± 0.0033	0.789 ± 0.0069	0.0142 ± 0.0005	0.526 ± 0.0533	0.0358 ± 0.0025
VGG-LSTM	✓	-	-	-	0.888 ± 0.0078	0.0082 ± 0.0005	$\mathbf{0.596} \pm 0.0153$	0.0221 ± 0.0014	0.904 ± 0.0056	0.0060 ± 0.0004	0.638 ± 0.0158	$\mathbf{0.0209} \pm 0.0011$
	✓	✓	-	-	0.884 ± 0.0075	0.0078 ± 0.0004	0.577 ± 0.0320	0.0231 ± 0.0025	0.905 ± 0.0059	0.0061 ± 0.0003	0.621 ± 0.0518	0.0237 ± 0.0050
	✓	✓	✓	-	0.883 ± 0.0065	0.0074 ± 0.0004	0.580 ± 0.0401	0.0265 ± 0.0037	0.903 ± 0.0082	0.0059 ± 0.0003	0.611 ± 0.0332	0.0236 ± 0.0025
	✓	✓	✓	✓	0.883 ± 0.0055	0.0081 ± 0.0006	0.592 ± 0.0196	$\mathbf{0.0218} \pm 0.0014$	0.892 ± 0.0071	0.0065 ± 0.0005	0.631 ± 0.0333	0.0245 ± 0.0043
Res-Tf	✓	-	-	-	0.903 ± 0.0086	0.0080 ± 0.0009	0.594 ± 0.0618	0.0243 ± 0.0034	0.898 ± 0.0074	0.0071 ± 0.0007	0.612 ± 0.0368	0.0261 ± 0.0024
	✓	✓	-	-	0.903 ± 0.0041	0.0079 ± 0.0007	0.586 ± 0.0415	0.0253 ± 0.0036	0.908 ± 0.0051	0.0064 ± 0.0007	$\mathbf{0.649} \pm 0.0181$	0.0253 ± 0.0019
	✓	✓	✓	-	$\mathbf{0.907} \pm 0.0056$	0.0078 ± 0.0005	0.574 ± 0.0432	0.0256 ± 0.0030	$\mathbf{0.911} \pm 0.0071$	$\mathbf{0.0058} \pm 0.0009$	0.607 ± 0.0327	0.0264 ± 0.0028
	✓	✓	✓	✓	0.902 ± 0.0047	$\mathbf{0.0070} \pm 0.0006$	0.551 ± 0.0666	0.0272 ± 0.0036	0.905 ± 0.0061	0.0060 ± 0.0003	0.629 ± 0.0403	0.0255 ± 0.0045

adding noise to thegripper's width. The dataset is then randomly split into train, validation and test sets. They utilized two different splitting methods to create different validation and test sets. In the first method, the validation and test set contain seen objects (Seen-Objects); in the second method, the objects contain unseen objects (Unseen-Objects).

We additionally employ stratified sampling [24] based on shape and material to ensure balanced splits. In our case, the test set comprises 20% of the dataset. The rest is divided into 10 equally sized folds, where one fold is selected as the validation set. During the evaluation, we reported the average of 10 runs, where a different fold was used as the validation set each time to ensure robust results.

We employ a combination of mean squared error (MSE) and $l2$ regularization during model training. Furthermore, learning rate decay was employed. The model architecture and hyperparameter, such as learning and decay rates, were optimized using SMAC3 [25]. We evaluate the models on log normalized MSE (N-MSE) and $\log_{10}$ accuracy. The $\log_{10}$ accuracy describes how often a prediction is within a magnitude of the target value [16].

III. Result and Discussion

The results for the experiments can be found in Table I. Our proposed models outperform the Top10NN baseline model in all scenarios. For Seen-Object with random sampling, the Res-Tf model with all optional inputs performs best with an N-MSE of 0.0070. For Unseen-Object with random sampling, we observe that VGG-LSTM achieves the best N-MSE of 0.0218 with all optional inputs.

Utilizing stratified sampling, we observe a visible improvement in the $\log_{10}$ accuracy metric across all models. However, that is not always the case for N-MSE. The best-performing model for stratified sampling for Seen-Object is the Res-Tf model without the estimation feature with an N-MSE of 0.0058. For Unseen-Objects the VGG-LSTM model with only images performs best in N-MSE with 0.0209.

During the comparison of the results, we observed better performance for soft objects up to the Young's Modulus of the GelSight sensor, which could suggest that sensor compliance impacts compliance estimation performance. Additionally, we observed a strong variance in performance between dataset splits when utilizing the Unseen-Object split strategy, indicating problems with the dataset distribution.

The results for stratified sampling indicate that analytical compliance is not needed to achieve state-of-the-art performance. When comparing the results for VGG-LSTM with the results for Res-Tf, we observe that the $\log_{10}$ accuracy is often worse even if the N-MSE is better. This is caused by the wide range in which estimation is counted as accurate in the $\log_{10}$ accuracy.

IV. Conclusion

We proposed two new model frameworks for estimating the Young's Modulus of an object. Both models outperformed the baseline considering the baseline's metric $\log_{10}$ accuracy and the most standard N-MSE. We showed that features other than tactile information are unnecessary to reach state-of-the-art compliance estimation performance.

In future research, we want to study the possible effect of the sensor's compliance on estimating objects' compliance. Similarly, a larger, more balanced dataset could be created to allow further insights.

ACKNOWLEDGMENT

This research was partially funded by the Niedersächsisches Ministerium für Wissenschaft und Kultur via the Volkswagen Foundation under the Programme zukunft.niedersachsen: Forschungskooperation Niedersachsen – Israel project No. 15-76251-5616/2023 (ROMEO).

References

[1] M. Baietto and A. D. Wilson, "Electronic-Nose Applications for Fruit Identification, Ripeness and Quality Grading," *Sensors*, vol. 15, no. 1, pp. 899–931, 2015.

[2] J. T. Iivarinen, R. K. Korhonen, P. Julkunen, and J. S. Jurvelin, "Experimental and Computational Analysis of Soft Tissue Stiffness in Forearm Using a Manual Indentation Device," *Medical Engineering & Physics*, vol. 33, no. 10, pp. 1245–1253, 2011.

[3] M. Cianchetti, C. Laschi, A. Menciassi, and P. Dario, "Biomedical Applications of Soft Robotics," *Nature Reviews Materials*, vol. 3, no. 6, pp. 143–153, 2018.

[4] A. J. Spiers, M. V. Liarokapis, B. Calli, and A. M. Dollar, "Single-Grasp Object Classification and Feature Extraction with Simple Robot Hands and Tactile Sensors," *IEEE Transactions on Haptics*, vol. 9, no. 2, pp. 207–220, 2016.

[5] S. Toprak, N. Navarro-Guerrero, and S. Wermter, "Evaluating Integration Strategies for Visuo-Haptic Object Recognition," *Cognitive Computation*, vol. 10, no. 3, pp. 408–425, 2018.

[6] D. R. H. Jones and M. F. Ashby, *Engineering Materials 1: An Introduction to Properties, Applications and Design*, 5th ed. Butterworth-Heinemann, 2019, vol. 1.

[7] O. A. Juiña Quilachamín and N. Navarro-Guerrero, "A Biomimetic Fingerprint for Robotic Tactile Sensing," in *International Symposium on Robotics (ISR Europe)*. Stuttgart, Germany: VDE Verlag GmbH, Sept. 2023, pp. 112–118.

[8] K. Shimizu, R. H. Purba, K. Kusumoto, X. Yaer, J. Ito, H. Kasuga, and Y. Gaqi, "Microstructural Evaluation and High-Temperature Erosion Characteristics of High Chromium Cast Irons," *Wear*, vol. 426–427, pp. 420–427, 2019.

[9] K. Inoue, S. Okamoto, Y. Akiyama, and Y. Yamada, "Effect of Material Hardness on Friction Between a Bare Finger and Dry and Lubricated Artificial Skin," *IEEE Transactions on Haptics*, vol. 13, no. 1, pp. 123–129, 2020.

[10] M. Britton, E. Parle, and T. J. Vaughan, "An Investigation on the Effects of in Vitro Induced Advanced Glycation End-Products on Cortical Bone Fracture Mechanics at Fall-Related Loading Rates," *Journal of the Mechanical Behavior of Biomedical Materials*, vol. 138, p. 105619, 2023.

[11] R. Lu and Y. Peng, "Hyperspectral Scattering for Assessing Peach Fruit Firmness," *Biosystems Engineering*, vol. 93, no. 2, pp. 161–171, 2006.

[12] S. J. Lederman and R. L. Klatzky, "Haptic Perception: A Tutorial," *Attention, Perception, & Psychophysics*, vol. 71, no. 7, pp. 1439–1459, 2009.

[13] N. Navarro-Guerrero, S. Toprak, J. Josifovski, and L. Jamone, "Visuo-Haptic Object Perception for Robots: An Overview," *Autonomous Robots*, vol. 47, no. 4, pp. 377–403, 2023.

[14] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-Independent Hardness Estimation Using Deep Learning and a Gelsight Tactile Sensor," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2017, pp. 951–958.

[15] W. Yuan, S. Dong, and E. H. Adelson, "GelSight: High-Resolution Robot Tactile Sensors for Estimating Geometry and Force," *Sensors*, vol. 17, no. 12, p. 2762, 2017.

[16] M. Burgess and J. Zhao, "Learning Object Compliance via Young's Modulus from Single Grasps with Camera-Based Tactile Sensors," 2024.

[17] J. Rychlewski, "On Hooke's law," *Journal of Applied Mathematics and Mechanics*, vol. 48, no. 3, pp. 303–314, 1984.

[18] A. C. Fischer-Cripps, "The Hertzian Contact Surface," *Journal of Materials Science*, vol. 34, no. 1, pp. 129–137, 1999.

[19] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," in *International Conference on Learning Representations (ICLR)*, vol. 3rd. arXiv.org, 2015, p. 14.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems (NIPS)*, vol. 30, 2017, p. 11.

[21] A. Arnab, M. Dehghani, G. Heigold, C. Sun, M. Lučić, and C. Schmid, "ViViT: A Video Vision Transformer," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 6816–6826.

[22] J. Donahue, L. A. Hendricks, M. Rohrbach, S. Venugopalan, S. Guadarrama, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PP, no. 99, pp. 1–1, 2016.

[23] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[24] K. Sechidis, G. Tsoumakas, and I. Vlahavas, "On the Stratification of Multi-Label Data," in *Machine Learning and Knowledge Discovery in Databases*, D. Gunopulos, T. Hofmann, D. Malerba, and M. Vazirgiannis, Eds. Springer, 2011, vol. 6913, pp. 145–158.

[25] M. Lindauer, K. Eggensperger, M. Feurer, A. Biedenkapp, D. Deng, C. Benjamins, T. Ruhkopf, R. Sass, and F. Hutter, "SMAC3: A Versatile Bayesian Optimization Package for Hyperparameter Optimization," *Journal of Machine Learning Research*, vol. 23, no. 54, pp. 1–9, 2022.