

# Improving Tactile Gesture Recognition with Optical Flow

Shaohong Zhong<sup>1,\*</sup>, Alessandro Albini<sup>1,\*</sup>, Giammarco Caroleo<sup>1</sup> Giorgio Cannata<sup>2</sup> and Perla Maiolino<sup>1</sup>

**Abstract**— Tactile gesture recognition systems play a crucial role in Human-Robot Interaction (HRI) by enabling intuitive communication between humans and robots. The literature mainly addresses this problem by applying machine learning techniques to classify sequences of tactile images encoding the pressure distribution generated when executing the gestures. However, some gestures can be hard to differentiate based on the information provided by tactile images alone.

In this paper, we present a simple yet effective way to improve the accuracy of a gesture recognition classifier. Our approach focuses solely on processing the tactile images used as input by the classifier. In particular, we propose to explicitly highlight the dynamics of the contact in the tactile image by computing the dense optical flow. This additional information makes it easier to distinguish between gestures that produce similar tactile images but exhibit different contact dynamics. We validate the proposed approach in a tactile gesture recognition task, showing that a classifier trained on tactile images augmented with optical flow information achieved a 9% improvement in gesture classification accuracy compared to one trained on standard tactile images.

## I. INTRODUCTION

Research on Human-Robot Interaction (HRI) aims to provide smooth cooperation between humans and robots, allowing operators to interact with robots in the most natural way [1]. In this respect, a significant body of literature studies techniques to send commands to robots using gestures [2], [3]. These gesture recognition systems predominantly rely on the recognition of motions of the human body or hands and are used to trigger specific robot actions. Most of these systems are based on visual feedback obtained with the use of RGB or RGB-D cameras [4]–[8]. Additionally, gesture recognition based on information acquired from Inertial Measurement Unit (IMU) and/or Electromyography (EMG) sensors has also been considered in the literature [9]–[11].

Beyond the classification of body motions, another possibility is to directly interact with robots through physical contacts [12]. This type of communication is usually preferred when close HRI is required and where physical interaction is the most natural way to interact with the robot (e.g. to move the robot arm or to teach a movement [13]). In this respect, researchers have started taking advantage of tactile sensors to recognise human gestures applied to the robot [14]. In the literature, two main strategies have

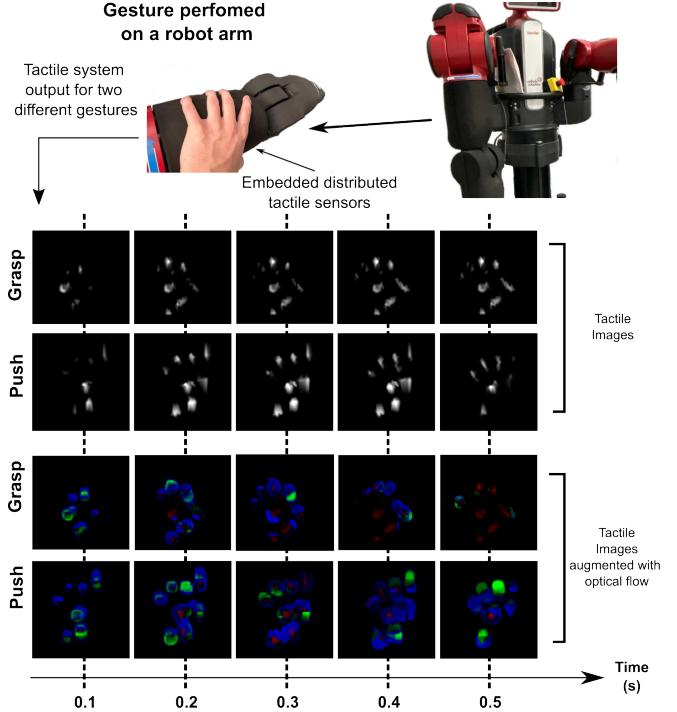


Fig. 1. The gesture applied by the human to the robot arm generates a sequence of tactile images at each time instant (each 0.1 s as shown in the Figure). The contact shapes captured by the tactile images are similar between the two different gestures. However, with the proposed representation, which highlights the contact dynamics in the green and blue channels of the image, the differences become much clearer — after 0.4 s, the dynamic fades for the Grasp gesture, while it is still present in the Push.

been proposed. The first is based on the processing of time series tactile data [15]–[17], some of which also contains multimodal information [18], [19]. The second is based on the processing and classification of the pressure distribution, encoded as sequences of tactile images that describe how the contact shape evolves over a fixed time window [20]–[23].

However, some gestures are not easily recognisable by analysing the contact shape alone. An example is given in Figure 1, showing two sequences of tactile images corresponding to *Grasp* and *Push* actions performed by two different users. Each tactile image shows a greyscale representation of the pressure applied by the human hand at each time frame. As can be observed, both gestures involve the same parts of the hand (fingertips, palm and thumbs are present in both sequences), making it hard to distinguish them solely on the basis of the contact shape. A possibility to make the gestures more distinguishable is to consider additional sensory inputs or the effects of shear forces [24]. However, this requires additional complexity from the hardware point

<sup>1</sup> is with the Oxford Robotics Institute (ORI), University of Oxford, UK.

<sup>2</sup> is with the Department of Informatics, Bioengineering, Robotics and Systems Engineering (DIBRIS), University of Genoa, IT.

\*Equal contribution.

This work was supported by the SESTOSENSE project (HORIZON EUROPE Research and Innovation Actions under GA number 101070310).

We would like to acknowledge the use of the SCAN facility in carrying out this work.

of view. Moreover, distributed sensors capable of capturing shear forces can hardly scale over large areas.

It must be noted that, although the contact shape may be similar among various gestures, the contact dynamics may be completely different. As an example, while a *Grasp* is a quasi-static action, a *Push* involves higher contact dynamics, i.e., the value of the pixels in the image significantly changes between two frames. This is not immediately visible from the sequences of grayscale tactile images in Figure 1.

In this paper, we propose a simple and effective method to significantly improve the accuracy of a classifier trained for tactile gesture recognition. This approach does not require any additional hardware and it is based on a processing of tactile images only. Similarly to what is commonly done in computer vision for camera-based gesture recognition [25], [26], we exploit optical flow information to extract the dynamics of tactile gestures. Therefore, we propose to augment the tactile image by adding information on the variation of pixel intensities and displacement between consecutive time frames. The proposed tactile image consists of 3 channels: the red channel represents the pressure distribution at the given time instant; the green and blue channels encode the optical flow in the form of a dense flow field [27].

A sequence of augmented tactile images for *Grasp* and *Push* gestures can be seen in Figure 1. Compared with a standard tactile image, the contact dynamics are much more visible through augmentation, and we hypothesise that a classifier trained using augmented tactile images would obtain significantly better performance compared to the use of standard tactile images alone. In this respect, we first collect a dataset of tactile gestures from a number of human users interacting with the robot. Then, we test our hypothesis by designing a classifier based on a Convolutional Neural Network (CNN) and Recurrent Neural Network (RNN) architecture [28] trained on both types of input tactile image sequences.

The paper is structured as follows. Section II first describes the steps to obtain the proposed augmented tactile image from the sensors' raw data. Then, a description of the architecture of the gesture classifier follows. Section III reports details on the experimental setup, the data collection procedure, and the training and evaluation details for the classifier. Results are presented and discussed in Section IV. Conclusion follows.

## II. AUGMENTED TACTILE IMAGES FOR GESTURE RECOGNITION TASKS

This section describes how the proposed tactile data representation is built, starting from raw tactile sensor measurements, and provides details on the network architecture used for gesture classification.

### A. Augmented Tactile Image

The complete processing pipeline used to create the tactile image augmented with dense optical flow information is shown in Figure 2.

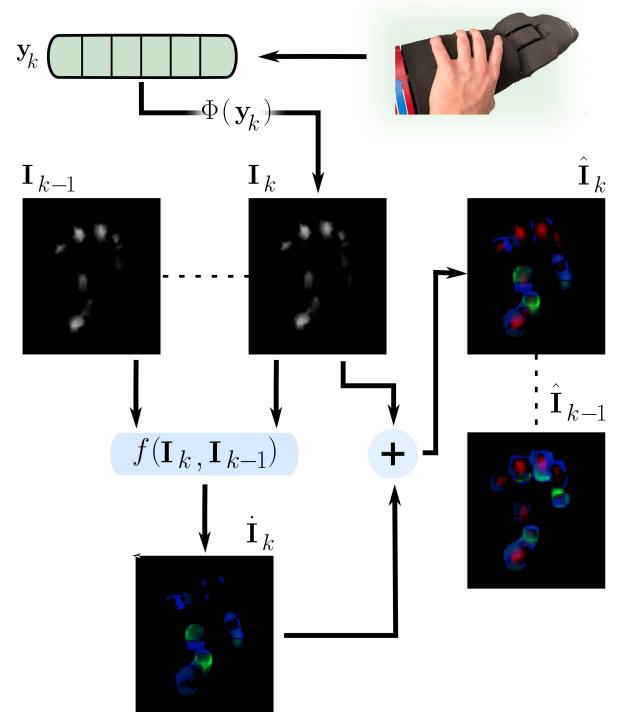


Fig. 2. Processing steps required to augment a tactile image using dense flow information. As a gesture is performed on the robot, the distributed tactile sensors embedded in its body capture the pressure distribution, generating an array of measurements. The measurements are processed to create a 3-channel tactile image. The red channel represents the contact shape captured at the given time instant. The green and blue channels encode respectively the magnitude and direction of the optical flow computed between consecutive time frames.

It is assumed that the physical gesture is performed on an area of the robot equipped with distributed tactile sensors that can capture the physical interaction. In the following, we refer to large-area tactile sensing technologies composed of distributed transducers, namely *taxels*, capable of providing information on the pressure applied over a certain region [29]–[33]. At each sampling time instant  $k$ , the sensors produce a set of responses  $\mathbf{y}_k = \{y_{1,k}, y_{2,k}, \dots, y_{i,k}, \dots, y_{N,k}\}$ , where  $y_{i,k} \in \mathbb{R}$ ,  $i = \{0, 1, \dots, N\}$ , and  $N$  is the total number of taxels. These responses contain raw measurements which are a function (typically non-linear) of the pressure applied on each taxel.

The array  $\mathbf{y}_k$  can be converted to a tactile image through a process of resampling and interpolation of the taxels' spatial distribution [34], [35]. In Figure 2, we refer to this transformation as  $\Phi(\mathbf{y}_k)$ . It must be noted that, if taxels are integrated over a non-planar robot body part (as in our experimental setup described in Section III), the generation of the tactile image is still possible as described in [36]. The resulting tactile image is a 1-channel image, whose pixel values are related to the pressure applied on a specific area of the robot. As visible in Figure 2, the single image  $\mathbf{I}_k$  captures the *shape* of the hand in contact with the robot body when the user is performing the gesture.

The next step is to consider a sequence of two consec-

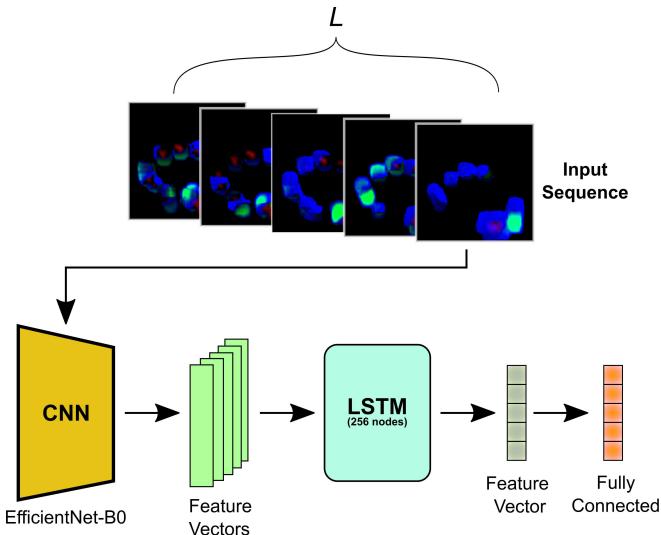


Fig. 3. Classification model for gesture recognition based on a CNN-LSTM architecture. The network takes a sequence of tactile images of length  $L$  as input.

utive tactile images and compute the dynamic information. The optical flow is computed using the Gunnar Farnebäck algorithm [37]. The algorithm can be used to compute a dense optical flow, estimating the flow at each pixel of the image. Compared to other popular methods, such as the Lucas-Kanade algorithm [38], which assumes constant displacement of the pixels, it is computationally more expensive but more suitable for dealing with complex displacement patterns. Due to the relatively small resolution of tactile images, compared to high-resolution images acquired with cameras, the higher computational time is not critical for our application.

The dense optical flow  $\hat{\mathbf{I}}_k = f(\mathbf{I}_k, \mathbf{I}_{k-1})$  is then computed between two consecutive frames, using the tactile image  $\mathbf{I}_{k-1}$  generated in the previous sampling step. The first step is initialised with an image  $\mathbf{I}_0$  whose values are null.  $\hat{\mathbf{I}}_k$  consists of a 2-channel image — the first encodes the magnitude of the flow, while the second encodes its direction in polar coordinates. The last operation consists of joining  $\mathbf{I}_k$  and  $\hat{\mathbf{I}}_k$  into a 3-channel image  $\hat{\mathbf{I}}_k$ , encoding, at time instant  $k$ , both the contact shape (red channel) and its dynamics, expressed as a dense flow — green for the magnitude and blue for the direction.

### B. Classification Architecture

To evaluate the effect of the proposed approach, we perform classification of the sequences of tactile images associated with different gestures. Since a single data sample consists of a time series of tactile images, this resembles a video classification problem. Therefore, inspired by prior works on video classification [39], we leverage a combination of a CNN and a RNN. For each input sequence, the CNN is first used to extract spatial image features from each tactile image frame. Then, the sequence of image features is passed through the RNN to capture the temporal dependencies between the frames. The output representation

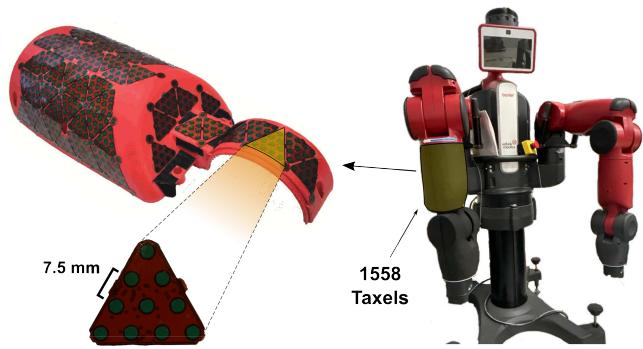


Fig. 4. Experimental setup: distributed tactile sensors are integrated into the forearm of the Baxter robot. The tactile system is composed of 1558 capacitive-based taxels, which acquire tactile measurements at 10 Hz. This platform is used to collect a dataset of 5 gestures from 38 people.

that captures both the spatial and temporal information is passed through a final layer of a fully-connected network to perform classification.

Specifically, for each sample, we perform simple preprocessing on individual image frames, including resizing and normalisation. Then, we use an EfficientNet pretrained on ImageNet to extract the image features from individual frames [40], [41], and an LSTM network to capture the temporal information [42]. The choice of the CNN and RNN architecture is modular and can be adjusted depending on the size and type of the dataset. During training, the weights of all of the neural networks, including the pre-trained networks, are updated.

As visible in Figure 3, the input to the network is a sequence of images of fixed length  $L$ . The effect of changing the length  $L$  on the classification accuracy is discussed in Sections III and IV.

## III. EXPERIMENTS DESCRIPTION

### A. Experimental Setup

The dataset has been collected on the platform shown in Figure 4, consisting of a Baxter robot equipped with 1558 distributed tactile sensors on the forearm. Sensors are covered by a conductive black fabric, and their placement (for the upper half of the forearm) can be seen in Figure 4. The tactile sensing technology, namely *CySkin*, is an improved version of that presented in [43]. It is composed of triangular modules hosting up to 11 capacitive-based taxels, whose pitch is 7.5 mm. Measurements are collected through a CAN bus at 10 Hz with 16-bit resolution.

### B. Tactile Dataset Collection

Users were asked to perform gestures on the robot's forearm, similar to those described in [36]:

- *Grasp*: the user firmly grasps the robot forearm with one hand.
- *2hGrasp*: the user firmly grasps the robot forearm using both hands.
- *Twist*: the user grabs and twists the robot forearm along its rotational axis.

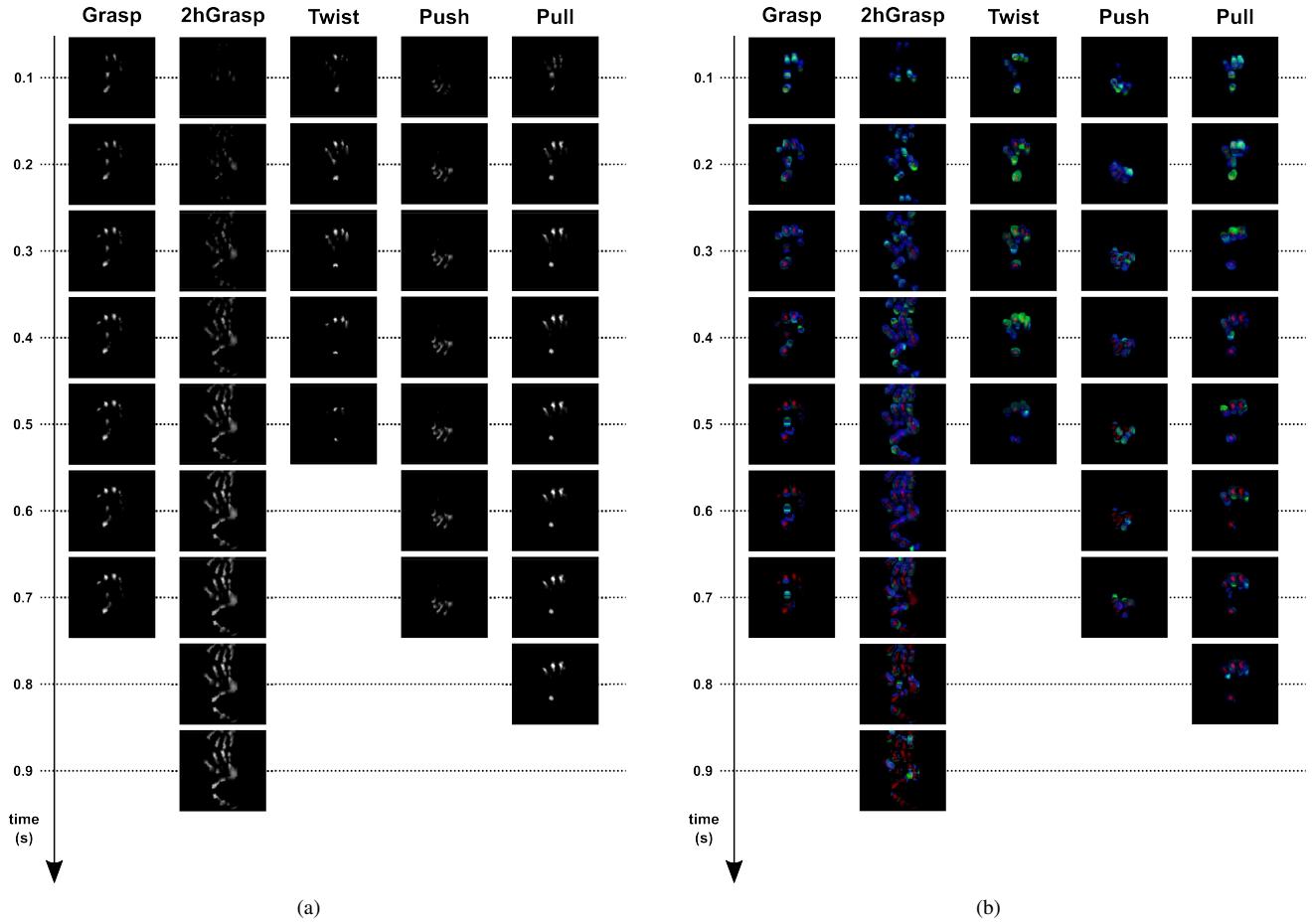


Fig. 5. Tactile images corresponding to 5 different gestures performed by a single user. An image is generated at each sampling step of the tactile system (0.1 s). As visible, using standard tactile images, Grasp, Twist and Pull look similar in this case. However, when using the proposed 3-channel encoding, their differences become more visible. (a) Sequences of standard 1-channel tactile images; (b) Sequences of tactile images embedding optical flow information.

- *Push*: the user pushes the robot forearm away.
- *Pull*: the user grabs and pulls the robot's forearm.

It must be noted that when collecting these kinds of gesture datasets, data are affected by user variability [15], [24], [36]. Physical characteristics such as hand size or grasping strength may affect the way the user interacts with the robot. Since we wanted to let the users interact with the robot in the most natural way possible, participants were informed only by describing the gesture to be applied. No instructions were given about the force to be applied, the area of the forearm to be touched, or the length of the gesture. In order to take into account possible user variations and capture them in the dataset, we involved 38 users in the experiments<sup>1</sup>. They differ in terms of gender (68.4% Male, 31.6% Female), handedness (92.1% Right, 7.9% Left), age (21-34 years), weight (48-95 kg), and hand size (16-20 cm)<sup>2</sup>. Compared to other studies on touch gesture recognition, this dataset has a number of gesture classes similar to [19]–[23], while the number of people involved in the study is in line

with [21], [23] and much higher than [15], [17]–[20], [22].

Each gesture was repeated 5 times in two different positions—standing in front of and on the side of the robot. Considering all 5 gestures, this led to 50 samples for each user, corresponding to a total of 1900 samples. During the whole experiment, the robot was commanded to maintain its pose. Each recorded sample consists of a sequence of tactile measurements collected every 0.1 seconds. The length of the gestures varies from a minimum of 0.5 s to a maximum of 2.7 s. Such a difference in length is due to the fact that after performing the gesture, users were still holding the robot for a while. In order to avoid the classifier being biased by the length of the gesture, we set the input size of the network to  $L = 5$ , corresponding to 0.5 s, i.e., the shortest duration in our dataset. Furthermore, when testing the classifier, only the first  $L$  tactile frames are considered. Additional analysis on the input length  $L$  is reported in Section IV.

From the raw measurements, two datasets of tactile images were generated. The first dataset is composed of 3-channel tactile images generated as described in Figure 2. The second dataset contains standard 1-channel tactile images and is needed as a comparison to analyze the advantage of including optical flow information in the input data to the classifier. The

<sup>1</sup>Users involved in the experiments signed an informed consent form.

<sup>2</sup>The hand size is measured from the base of the palm to the tip of the middle finger.

size of each image is  $357 \times 334$ . They have been created by resampling the distribution of the taxels as described in [36] with a spatial sampling step of 1 mm.

Finally, Figure 5(a) reports tactile images of the 5 gestures performed by a single user and sampled at  $0.1\text{ s}^3$ . As visible from the contact shape, this user uses her/his whole hand to perform the *2hGrasp* gesture, while mostly the fingers were involved in the *Push* operation. Regarding *Grasp*, *Twist*, and *Pull*, the contact shape is similar among the three gestures, since the participant mostly involved the fingertips in these three gestures. Figure 5(b) shows the same data processed as described in Section II-A. As shown, when superimposing the dense flow over the contact shape, *Grasp*, *Twist*, and *Pull* become much more distinguishable. In particular, in the *Grasp* case, the dynamic component quickly fades, and the static component of the pressure distribution becomes dominant at the fourth sample. In the *Twist* action, the green channel (corresponding to the magnitude of the optical flow) is most dominant and is concentrated at the fingertips where the user is increasingly applying force to twist the robot forearm. Similarly, for the *Pull* gesture, the dynamic part is much more evident than in the *Grasp* and with a different pattern than the *Twist*. However, this is just an example - as previously discussed, the way users interact with the robot may vary. For instance, the *Push* gestures in Figures 1 and 5 are completely different - in Figure 1 fingertips, thumb and palm are involved, whereas only the fingers are visible in Figure 5.

### C. Model Train and Test Details

Both datasets were split into train and test. To ensure the validity of the average accuracy result, we randomly sample a fixed number of 30 samples ( $\sim 10\%$ ) from each gesture class to form the test dataset, and use the rest ( $\sim 90\%$ ) as the training dataset. Furthermore, we performed augmentation on the training dataset, whereby for each sequence longer than  $L$  frames, we use a sliding window approach to generate additional sequences for training by slicing the full sequence into smaller sequences of  $L$  consecutive frames, thereby generating more training samples. This results in  $\sim 1800$  samples for each gesture in the training dataset.

The CNN-LSTM network in Figure 3 based on EfficientNet-B0 (see Section II-B) was pretrained on ImageNet to extract the image features. We apply a dropout of 0.3 before passing the output of the LSTM network to a fully connected layer of 256 neurons for classification. For training, we use the standard cross entropy loss, and use stochastic gradient descent for gradient updates with a learning rate of 0.001. We train the models for 100 epochs.

For evaluation, we use three different seeds for each classification experiment and compare the results using the augmented 3-channel tactile images versus the original standard 1-channel tactile images. Furthermore, when analysing the effect of the length of the input sequence  $L > 5$  in

<sup>3</sup>The contact shape in the *2hGrasp* appears to be split due to the transformation of the 3D taxel distribution (wrapped around the forearm) into a 2D image [36].

		Standard Tactile Image				
		Grasp	0	15	3	2
True	Grasp	78	0	15	3	2
	2hGrasp	1	98	0	0	0
Twist	Grasp	78	0	15	3	2
	2hGrasp	1	98	0	0	0
Push	Grasp	78	0	15	3	2
	2hGrasp	1	98	0	0	0
Pull	Grasp	78	0	15	3	2
	2hGrasp	1	98	0	0	0
		Grasp	0	15	3	2

(a)

		Tactile Image + Optical Flow				
		Grasp	0	15	3	2
True	Grasp	95	0	3	1	0
	2hGrasp	1	95	0	3	0
Push	Grasp	95	0	3	1	0
	2hGrasp	1	95	0	3	0
Pull	Grasp	95	0	3	1	0
	2hGrasp	1	95	0	3	0
		Grasp	0	15	3	2

(b)

Fig. 6. Normalised confusion matrices for the classification experiments. The results are averaged across three seeds. The proposed representation boosts the classification accuracy by about 9%. (a) Confusion matrix obtained by training the model with standard tactile images - mean accuracy  $80.7 \pm 0.4\%$ . (b) Confusion matrix obtained by training the model with augmented tactile images - mean accuracy  $89.1 \pm 0.2\%$ .

the second part of Section IV, we pad the samples with insufficient frames with additional blank frames.

## IV. RESULTS

Figure 6 show the confusion matrices of the classifiers evaluated on the first  $L = 5$  frames of sequences belonging to the test set.

Using the original standard tactile images, the CNN-LSTM classifier achieves a mean classification accuracy of  $80.7 \pm 0.4\%$  across the 5 gestures. As shown in Figure 6(a), the classifier trained on standard tactile images mainly confuses *Grasp* with *Twist* and *Twist* with *Pull*, while *Push* is mainly confused with *Grasp* and *Twist*.

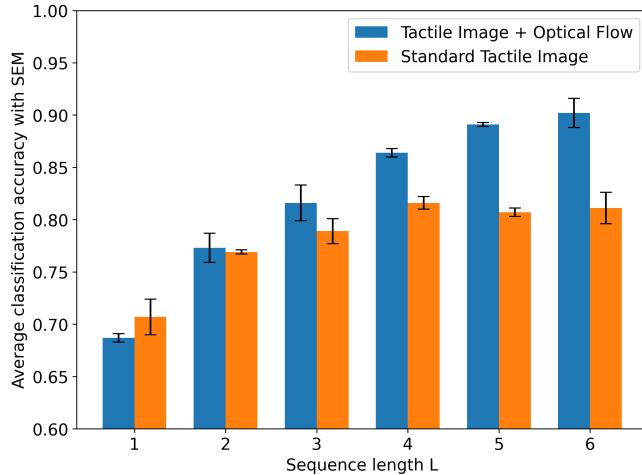


Fig. 7. Trend of the mean classification accuracy by varying the input size  $L$  from 1 to 6.

On the contrary, the proposed augmented tactile image allows the network to improve the classification accuracy by about 9%, achieving  $89.1 \pm 0.2\%$ . This improvement is achieved solely by performing the simple preprocessing described in Section II-A. From the confusion matrix in Figure 6(b) it is clear that *Grasp*, *Twist*, *Push*, and *Pull* are much more distinguishable. The recognition of *Grasp* improved by 17%, reaching 95%. Similarly, the accuracy for *Twist* increased by 16%. Compared to other gestures, *Pull* only gains an increment of 6% and is mostly confused with *Twist*. We argue that this is mainly due to the fact that both gestures primarily involve the fingertips. Although the dynamics are different for most users, a larger dataset may be required to properly train the model to distinguish between these two gestures.

Finally, we also performed additional experiments by changing the input length  $L$  and evaluating the effects on classification accuracy. In particular, smaller values of  $L$  allow for classifying the gesture in a shorter time, thus reducing the delay between the user command and a possible action triggered by the robot in response to the specific gesture. On the contrary, larger values of  $L$  increase the delay but allow for collecting more tactile information, possibly improving classification accuracy. In this respect, both classifiers have been retrained by considering  $L = \{1, 2, 3, 4, 6\}$ .

As shown in Figure 7, in the case of tactile images augmented with optical flow,  $L = 4$  frames already provide a reasonably good classification accuracy. With  $L < 4$ , the mean accuracy is below 85%, with lower repeatability among the seeds for  $L = 2$  and  $L = 3$ . For  $L = 6$ , as expected, the accuracy slightly improves to more than 90%. Regarding the model trained on standard tactile images, it is clear that performance is always lower than the model trained with augmented images. Furthermore, for  $L \geq 3$ , the accuracy reaches a plateau, and even for  $L = 6$ , it remains lower than what was obtained with the augmented tactile images using half of the input length.

## V. CONCLUSION

In this paper, we analyse the effect of incorporating dynamic contact information into tactile data to improve the performance of gesture recognition tasks. Specifically, we propose a representation of tactile data as a 3-channel image, which integrates both the contact shape and its variation over consecutive frames through dense optical flow.

We validated this approach using a gesture classification task with a dataset collected from a large number of participants interacting with a robot. The results demonstrate that including optical flow information provides a simple yet effective solution to improve classification performance. Our experimental findings reveal a significant improvement in classification accuracy, highlighting the effectiveness of this method without necessitating additional hardware or complex machine learning architectures.

Future works will be dedicated to extending the proposed representation by also including additional sensors commonly available on robots (such as torque sensors at the joints), and analysing how these can be exploited to further improve classification accuracy even while considering more complex gestures.

## REFERENCES

- [1] R. Gervasi, L. Mastrogiacomo, and F. Franceschini, “A conceptual framework to evaluate human-robot collaboration,” *The International Journal of Advanced Manufacturing Technology*, vol. 108, no. 3, pp. 841–865, May 2020.
- [2] H. Liu and L. Wang, “Gesture recognition for human-robot collaboration: A review,” *International Journal of Industrial Ergonomics*, vol. 68, pp. 355–367, 2018.
- [3] Z. Xia, Q. Lei, Y. Yang, H. Zhang, Y. He, W. Wang, and M. Huang, “Vision-based hand gesture recognition for human-robot collaboration: a survey,” in *2019 5th International Conference on Control, Automation and Robotics (ICCAR)*. IEEE, 2019, pp. 198–205.
- [4] M. Hasanuzzaman, V. Ampornaramveth, T. Zhang, M. Bhuiyan, Y. Shirai, and H. Ueno, “Real-time vision-based gesture recognition for human robot interaction,” in *2004 IEEE International Conference on Robotics and Biomimetics*, 2004, pp. 413–418.
- [5] J. Lee-Ferng, J. Ruiz-del Solar, R. Verschae, and M. Correa, “Dynamic gesture recognition for human robot interaction,” in *2009 6th Latin American Robotics Symposium (LARS 2009)*, 2009, pp. 1–8.
- [6] M. Sigalas, H. Baltzakis, and P. Trahanias, “Gesture recognition based on arm tracking for human-robot interaction,” in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2010, pp. 5424–5429.
- [7] R. C. Luo and Y.-C. Wu, “Hand gesture recognition for human-robot interaction for service robot,” in *2012 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2012, pp. 318–323.
- [8] W. Qi, S. E. Ovur, Z. Li, A. Marzullo, and R. Song, “Multi-sensor guided hand gesture recognition for a teleoperated robot using a recurrent neural network,” *IEEE Robotics and Automation Letters*, vol. 6, no. 3, pp. 6039–6045, 2021.
- [9] S.-O. Shin, D. Kim, and Y.-H. Seo, “Controlling mobile robot using imu and emg sensor-based gesture recognition,” in *2014 Ninth International Conference on Broadband and Wireless Computing, Communication and Applications*, 2014, pp. 554–557.
- [10] A. Carfi, C. Motolese, B. Bruno, and F. Mastrogiovanni, “Online human gesture recognition using recurrent neural networks and wearable sensors,” in *2018 27th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2018, pp. 188–195.
- [11] S. Singhvi and H. Ren, “Comparative study of motion recognition with temporal modelling of electromyography for thumb and index finger movements aiming for wearable robotic finger exercises,” in *2018 3rd International Conference on Advanced Robotics and Mechatronics (ICARM)*, 2018, pp. 509–514.

- [12] B. D. Argall and A. G. Billard, "A survey of tactile human–robot interactions," *Robotics and Autonomous Systems*, vol. 58, no. 10, pp. 1159–1176, 2010. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0921889010001375>
- [13] A. Albini, S. Denei, and G. Cannata, "Enabling natural human-robot physical interaction using a robotic skin feedback and a prioritized tasks robot control architecture," in *2017 IEEE-RAS 17th International Conference on Humanoid Robotics (Humanoids)*, 2017, pp. 99–106.
- [14] D. Silvera-Tawil, D. Rye, and M. Velonaki, "Artificial skin and tactile sensing for socially interactive robots: A review," *Robotics and Autonomous Systems*, vol. 63, pp. 230–243, 2015.
- [15] F. Naya, J. Yamato, and K. Shinozawa, "Recognizing human touching behaviors using a haptic interface for a pet-robot," in *IEEE SMC'99 Conference Proceedings. 1999 IEEE International Conference on Systems, Man, and Cybernetics (Cat. No.99CH37028)*, vol. 2, 1999, pp. 1030–1034 vol.2.
- [16] W. D. Stiehl and C. Breazeal, "Affective touch for robotic companions," in *Affective Computing and Intelligent Interaction*, J. Tao, T. Tan, and R. W. Picard, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2005, pp. 747–754.
- [17] G. Huisman, A. Darriba Frederiks, B. Van Dijk, D. Hevlen, and B. Kröse, "The tasst: Tactile sleeve for social touch," in *2013 World Haptics Conference (WHC)*, 2013, pp. 211–216.
- [18] M. Kaboli, A. Long, and G. Cheng, "Humanoids learn touch modalities identification via multi-modal robotic skin and robust tactile descriptors," *Advanced Robotics*, vol. 29, pp. 1411–1425, 11 2015.
- [19] S. yong Koo, J. G. Lim, and D. soo Kwon, "Online touch behavior recognition of hard-cover robot using temporal decision tree classifier," in *RO-MAN 2008 - The 17th IEEE International Symposium on Robot and Human Interactive Communication*, 2008, pp. 425–429.
- [20] D. S. Tawil, D. Rye, and M. Velonaki, "Interpretation of the modality of touch on an artificial arm covered with an eit-based sensitive skin," *The International Journal of Robotics Research*, vol. 31, no. 13, pp. 1627–1641, 2012. [Online]. Available: <https://doi.org/10.1177/0278364912455441>
- [21] A. Cirillo, P. Cirillo, G. De Maria, C. Natale, and S. Pirozzi, "A distributed tactile sensor for intuitive human-robot interfacing," *Journal of Sensors*, vol. 2017, p. 14, 04 2017.
- [22] D. Hughes, J. Lammie, and N. Correll, "A robotic skin for collision avoidance and affective touch recognition," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 1386–1393, 2018.
- [23] M. Salvato, S. Williams, C. Nunez, X. Zhu, A. Israr, F. Lau, K. Klumb, F. Abnousi, A. Okamura, and H. Culbertson, "Data-driven sparse skin stimulation can convey social touch information to humans," *IEEE Transactions on Haptics*, vol. PP, pp. 1–1, 11 2021.
- [24] H. Choi, D. Brouwer, M. A. Lin, K. T. Yoshida, C. Rognon, B. Stephens-Fripp, A. M. Okamura, and M. R. Cutkosky, "Deep learning classification of touch gestures using distributed normal and shear force," in *2022 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2022, pp. 3659–3665.
- [25] H. Cheng, L. Yang, and Z. Liu, "Survey on 3d hand gesture recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 9, pp. 1659–1673, 2016.
- [26] Y. Fang, K. Wang, J. Cheng, and H. Lu, "A real-time hand gesture recognition method," in *2007 IEEE International Conference on Multimedia and Expo*, 2007, pp. 995–998.
- [27] B. K. Horn and B. G. Schunck, "Determining optical flow," *Artificial intelligence*, vol. 17, no. 1-3, pp. 185–203, 1981.
- [28] V. Sharma, M. Gupta, A. Kumar, and D. Mishra, "Video processing using deep learning techniques: A systematic literature review," *IEEE Access*, vol. 9, pp. 139489–139507, 2021.
- [29] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing—from humans to humanoids," *IEEE Transactions on Robotics*, vol. 26, no. 1, pp. 1–20, 2010.
- [30] T. Someya, T. Sekitani, S. Iba, Y. Kato, H. Kawaguchi, and T. Sakurai, "A large-area, flexible pressure sensor matrix with organic field-effect transistors for artificial skin applications," vol. 101, no. 27, pp. 9966–9970, 2004.
- [31] G. Cannata, M. Maggiali, G. Metta, and G. Sandini, "An embedded artificial skin for humanoid robots," in *2008 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems*, Aug 2008, pp. 434–438.
- [32] P. Mittendorfer and G. Cheng, "Humanoid multimodal tactile-sensing modules," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 401–410, June 2011.
- [33] Y. Ohmura, Y. Kuniyoshi, and A. Nagakubo, "Conformable and scalable tactile sensor skin for curved surfaces," in *Proceedings 2006 IEEE International Conference on Robotics and Automation, 2006. ICRA 2006*, May 2006, pp. 1348–1353.
- [34] Z. Pezzementi, E. Plaku, C. Reyda, and G. D. Hager, "Tactile-object recognition from appearance information," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 473–487, 2011.
- [35] S. Luo, W. Mou, K. Althoefer, and H. Liu, "Novel tactile-sift descriptor for object shape recognition," *IEEE Sensors Journal*, vol. 15, no. 9, pp. 5001–5009, 2015.
- [36] A. Albini and G. Cannata, "Pressure distribution classification and segmentation of human hands in contact with the robot body," *The International Journal of Robotics Research*, vol. 39, no. 6, pp. 668–687, 2020.
- [37] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Image Analysis*, J. Bigun and T. Gustavsson, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2003, pp. 363–370.
- [38] B. Lucas and T. Kanade, "An iterative image registration technique," in *IJCAI'81*, pp. 674–679.
- [39] 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. 39, no. 4, pp. 677–691, 2017.
- [40] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri and R. Salakhutdinov, Eds., vol. 97. PMLR, 09–15 Jun 2019, pp. 6105–6114. [Online]. Available: <https://proceedings.mlr.press/v97/tan19a.html>
- [41] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
- [42] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
- [43] A. Schmitz, P. Maiolino, M. Maggiali, L. Natale, G. Cannata, and G. Metta, "Methods and technologies for the implementation of large-scale robot tactile sensors," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 389–400, 2011.