2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 25-29, 2020, Las Vegas, NV, USA (Virtual)

Shape reconstruction of CCD camera-based soft tactile sensors

Gabor Soter[1], Helmut Hauser[1], Andrew Conn[2], Jonathan Rossiter[1] and Kohei Nakajima[3]

***Abstract*— CCD camera-based tactile sensors provide high-resolution information about the deformation of soft and elastic interfaces. However, they have poor scalibility as it is difficult to sense a large surface area without increasing the distance between the camera and the interface or using multiple processing chips. For example, using such tactile sensors for a whole robotic arm is not yet possible. In this work, we demonstrate a data driven method that can reconstruct the high-resolution information about deformation of the soft interface while keeping the space requirements and power consumption relatively low. Our modified tactile sensor incorporates two independent sensing techniques, one low- and one high-resolution, and we learn to map to the latter from the former. As a low-resolution sensor, we use liquid-filled channels that transmit the information from the location of the tactile interaction to a rigid display, where the liquid displacements are tracked by a CCD camera. Simultaneously, the same interaction is measured by tracking the markers on the bottom of the sensor using a second CCD camera. After data collection, we train two different machine learning models to reconstruct the time series of the high-resolution sensor. By training a convolutional autoencoder (CAE) and attaching it to the recurrent neural network (RNN), we demonstrate the reconstruction of high-resolution video frames using only the time series of the low-resolution sensor.**

I. INTRODUCTION

Over the last two decades, CCD camera-based optical tactile sensors have gained significant traction. They are used to detect high-resolution force distribution caused by an external object while being low-cost [1], robust [2] and small in size [3]. The areas of application include robotic grasping [3], slip and edge detection [4], [5] and human-machine interaction [6], [7]. Several camera-based sensors have been developed simultaneously by different research groups, however, their designs are fairly similar and differ only in small details. A CCD camera-based tactile sensor consists of a soft, thin and dark interface, a CCD camera placed at a certain distance from the interface and some kind of internal illumination. The interface consists of two- or three-dimensional coloured markers that are tracked by the camera. The space between the interface and the camera is filled typically with air, but transparent gels can also be used [8], [9]. The response of the sensor is determined by the combined material properties of the gel and the interface.

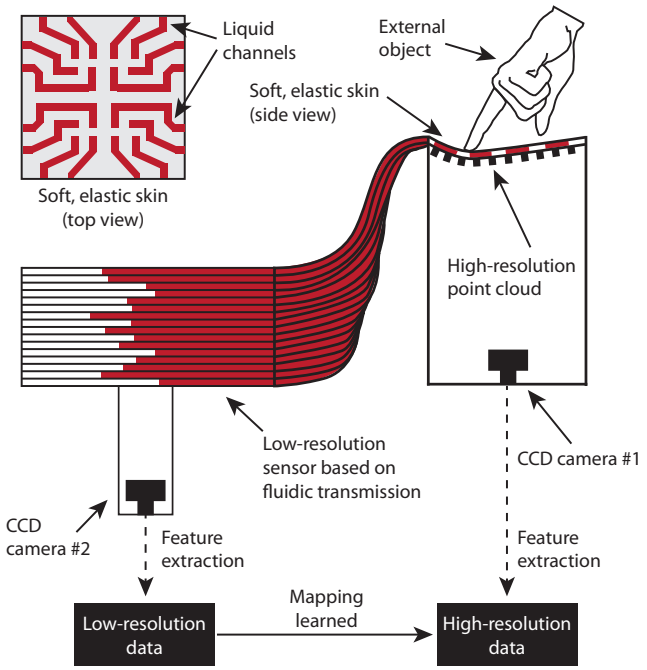


Fig. 1: The concept of the paper. Our goal was to create a scalable soft tactile sensor. We integrated two independent sensing technologies in the same sensor and learned a mapping between the two. The first one used a CCD camera to track markers on the elastic interface and to provide high-resolution information on the deformation of the sensor. Simultaneously, another CCD camera-based low-resolution sensor was used to measure fluid displacement.

The first camera-based soft tactile sensors was proposed by *Hristu et al.* [10]. The shape and the size of this sensor was comparable to a human finger, and it had a precisely located 5×5 marker grid on the interface. Both the camera and the illumination were placed inside the metal housing and the sensor has been used for single and multi-contact localisation. An improved version of the sensor had a finer grid including 10×10 markers and a pinhole to decrease the distance between the camera and the interface [11]. One of the limitations of the sensor was that its performance has been significantly affected by the ambient light. Recognising

[1]Gabor Soter, Helmut Hauser and Jonathan Rossiter are with Department of Engineering Mathematics, University of Bristol, BS8 1TH, United Kingdom, and with Bristol Robotics Laboratory, BS16 1QY, United Kingdom (email: gabor.soter@bristol.ac.uk; helmut.hauser@bristol.ac.uk and jonathan.rossiter@bristol.ac.uk).

[2]Andrew Conn is with Department of Mechanical Engineering, University of Bristol, BS8 1TH, United Kingdom, and with Bristol Robotics Laboratory, BS16 1QY, United Kingdom (email: a.conn@bristol.ac.uk).

[3]Kohei Nakajima is with Graduate School of Information Science and Technology, The University of Tokyo, 113-8656 Tokyo, Japan (email: k_nakajima@mech.t.u-tokyo.ac.jp).

This research was partially funded by EPSRC grants EP/M026388/1, EP/M020460/1 and EP/R02961X/1, by Leverhulme Trust research project RPG-2016-345, by the New Energy and Industrial Technology Development Organization (NEDO), by JSPS KAKENHI Grant Numbers JP18H05472, and by JSPS International Research Fellow (Graduate School of Information Science and Technology, The University of Tokyo).

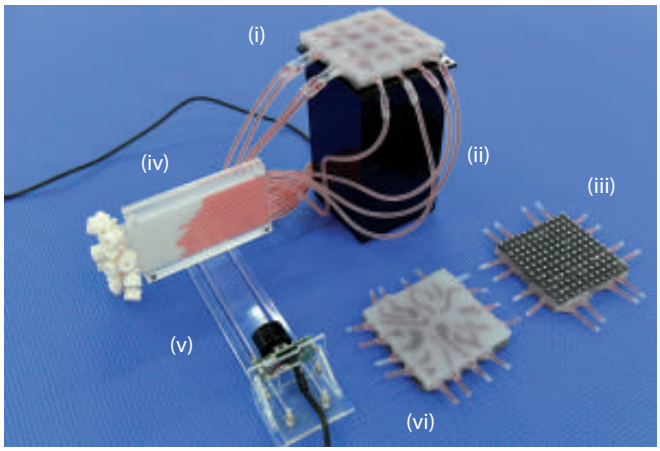


Fig. 2: The experimental setup: (i) soft tactile interface, (ii) rigid frame with a camera inside, (iii) coloured markers on the bottom of the sensor, (iv) the display and (v) the camera of the fluidic sensor, (vi) elastic interface with different internal channel arrangement.

that this as a major issue, the GelForce sensor introduced by *Kamiyama et al.* [12] had a non-transparent, dark interface to block the ambient light. The interface had a two layer marker grid with different colours. It has been integrated with a robotic hand to control grasping [3]. The sensor had light emitting diodes (LEDs) with different colours that were precisely located in the housing of the sensor. These exposed the soft interface from different directions and the reflection was tracked by the camera using different colour channels [13]. The sensor has been applied to detect object hardness [14], texture recognition [15], [16], geometry measurement and slip detection [17]. The markers on the sensor interface were painted manually and this introduced measurement error. This problem has been solved by the TacTip, a soft robotic fingertip developed by *Chorley et al.* [9]. The tactile interface had three-dimensional markers that amplified the mechanical deformation and they were fabricated using the same mold as the interface ensuring that the markers are precisely positioned. A comprehensive review on the TacTip can be found in [1].

One of the disadvantages of these sensors is that they are difficult to scale. Currently, there are two methods for scaling: enlarging the surface area of the interface or using multiple cameras. A sensor with larger surface area would require the camera to be placed at a larger distance from the interface. This is simply not practical because this distance is already in the centimeter range [15]. Enlarging the surface area would require a larger distance, however, the tactile sensors of a robot are usually placed where the space is already limited (e.g. hands of a humanoid robot). The other scaling method is to place multiple tactile sensors next to each other. However, this would increase the computational power that is required to process the camera images. Both the available space and computational power are limited in physical robots and, for these reasons, scaling CCD camera based tactile sensors remained an open problem.

In this paper, we present a new method that addresses this issue. As shown in Figure 1, our sensors incorporate two measurement techniques: the bottom layer consists of coloured markers with a non-transparent, dark background, whereas the top layer has liquid-filled chambers [18]. Due to physical interaction, the interface undergoes large deformation and this can be measured by the two sensing methods simultaneously. The deformation dislocates the markers and this is tracked by a CCD camera #1, whereas the volume change of the internal chambers displaces the incompressible coloured liquid (red in Figure 2) that can be tracked by another camera ((v) in Figure 2). We use data driven approaches to reconstruct the high-resolution sensory signals from the low-resolution data. In this work, we present two possible neural network architectures to solve this problem. The first one maps the time-series of the fluid displacements of the low-resolution sensor to the time-series of the marker positions of the high-resolution sensor. This model is based on a long short-term memory (LSTM) type recurrent neural network [19]. The second model maps the time-series of the extracted features of the low-resolution sensor to the camera images of the high-resolution sensor. Here, a recurrent neural network is combined with a stacked convolutional autoencoder that encodes the camera images [20]. This way, the high-resolution deformation of the soft tactile interface can be reconstructed using only the time series of the low-dimensional sensor [21]. As the processing unit of the low-resolution sensor can be placed anywhere in the robot's body, space constraints are not an issue.

In particular, the contributions of this paper are as follows:

- We created a new type of soft sensor by combining liquid transmission and CCD camera-based marker tracking methods.
- We collected and processed experimental data by reading both cameras simultaneously and extracting a set of features of the video frames.
- We trained two different machine learning algorithms that allowed us to reconstruct the high-dimensional deformation of the elastic interface.

The remainder of this paper is as follows: In Section II, we discuss the design of the physical system, the fabrication processes and the data processing methods. In Section III, we present the results of the time series and the video frame reconstructions. Finally, in Section IV, we discuss the challenges and future potential of this work. The entire code and data is available on `https://github.com/gaborsoter/shape_ml`.

II. System Design

A. Hardware

The experimental setup is shown in Figure 2 and consists of two separate sensing techniques. The low-resolution measurement method includes internal chambers embedded in the soft interface filled with coloured liquid (i), a rigid display (iv) and a CCD camera (v). Each channel is closed

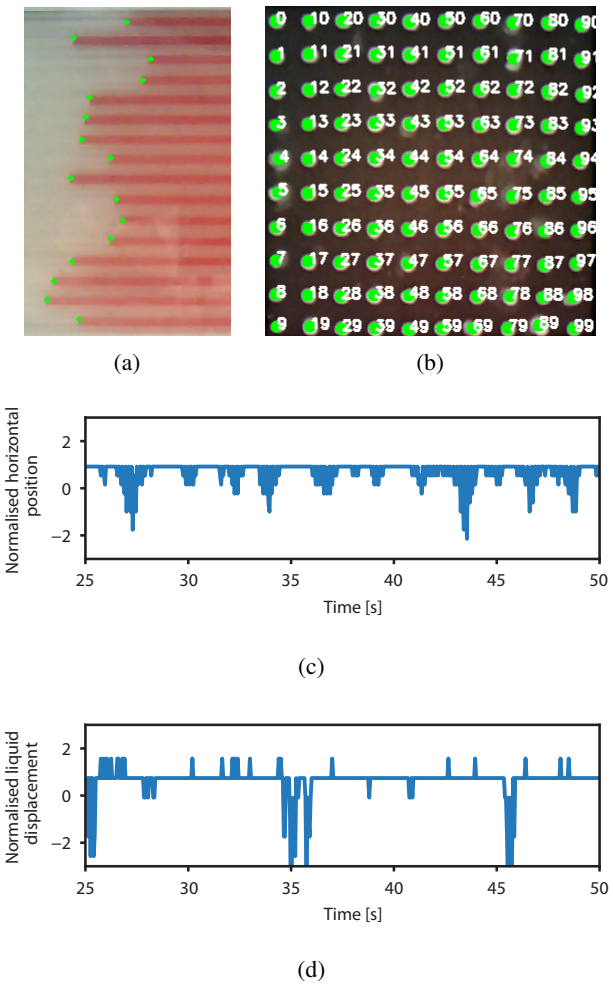


Fig. 3: The result of the image processing algorithms are shown in (a) and (b), and the time series after feature extraction are shown in (c) and (d) for only one feature of the low-resolution and the high-resolution sensor, respectively. The other features have similar characteristics. In case of low-resolution sensor we measured the horizontal position of the coloured liquid in each channel (16 green markers shown in (a)), whereas on the frames recorded by the high-resolution sensor the horizontal and vertical position of the markers were measured (horizontal and vertical position of 100 markers shown in green in (b)).

at both ends, this ensures that the liquid always returns back to its initial state and helps avoid liquid separation. The high-resolution sensing method consists of a marker grid on the bottom layer of the sensor and a CCD camera placed in the box (ii). The interface was made of Ecoflex 10 (Smooth-On) by casting. First, we fabricated the top layer with the internal chambers using a laser cut mold. Then, we created the bottom layer with the three-dimensional markers using another laser cut mold. While there were no restrictions on the material of the first mold, the second mold was laser cut from wood to avoid thermal warping. During the casting process a vacuum chamber was used in order to remove the trapped air from the silicone. After curing, the end of markers were painted with white silicone paint. Finally, silicone tubes were glued to the elastic interface using silicone glue (Sil-Poxy, Smooth-On). The internal chambers are closed on one end and this makes it challenging to fill up the chambers as traditional syringe pump-based methods do not work. Therefore, we placed the sensor in a tank full of coloured liquid and squeezed out the air from the chambers manually. When the air left these chambers, the coloured liquid filled up its space. More detail on the fabrication of sensors with liquid transmission can be found in [18].

B. Data collection and preparation

We collected data on the interaction between a human finger and the soft interface. The interface was fixed to the rigid frame and we collected data on approximately 1000 interactions. Before each interaction, we let the interface return to its original shape, then pushed the interface with one finger with an intensity chosen by the user and then released it. We recorded the video frames of both cameras simultaneously using the same clock. In order to extract the features from both videos, we used a series of transformations on the video frames. First, we cropped and rotated both videos in order to correct for fabrication errors. Once the frames were aligned, we applied different filters to process the video frames. For the low-resolution, liquid transmission sensor our goal was to find the position of the very left end of the liquid in each channel. After grayscaling the video frames of the low-resolution sensor we applied an adaptive threshold operator to distinguish the red liquid from the white background. Then, the image was split into smaller subregions that were one pixel high and represent the individual channels. We defined a function that iterates through the image from the left and finds the very left black pixel—the equivalent of red after thresholding. In the case of the high-resolution sensor our goal was to track the horizontal and vertical position of the markers.

The video frames of the high-resolution sensor were blurred in order to decrease the noise caused by the reflection of the light on the silicone layer. After this, the frames were binary thresholded and we used morphological image processing operations to filter out any small areas that represented either fabrication error or light reflection. OpenCV's contour finding algorithm was used to find the contours of the individual markers.

III. Results

A. Time series prediction

In this section we train a machine learning model that maps the time series of the liquid displacements to the time series of the absolute positions of the markers (see Figure 4). The machine learning model is a three-layer deep long short-term memory (LSTM) type of recurrent neural network. Each hidden layer had 256 neurons, rectified linear unit (ReLU) activation functions and a forget bias of 1.0. We trained the network using the Adam optimiser algorithm with a learning rate of 0.001 for 100 epochs.

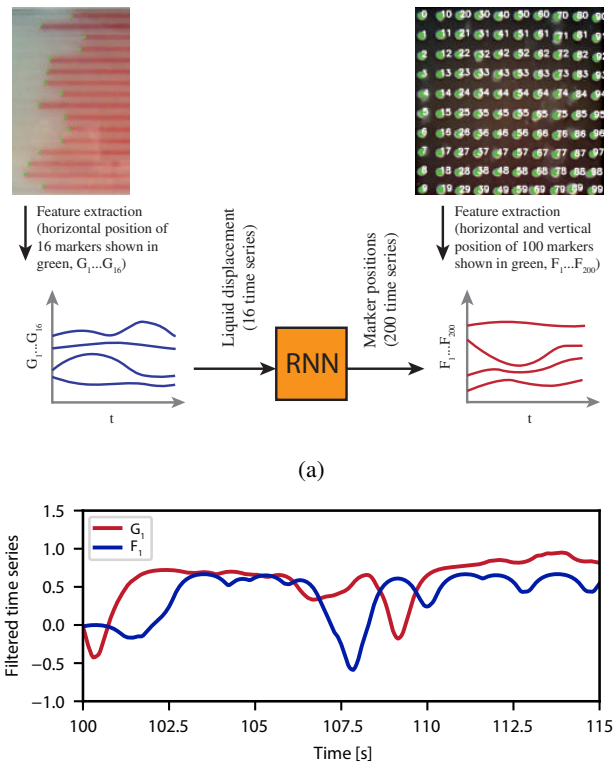


Fig. 4: (a) The architecture of the marker position prediction. First, the one-dimensional liquid positions (left) and the two-dimensional marker positions (right) are extracted from the video frames generating time series data for both the low-resolution and the high-resolution sensors. Next, these time series are prepared to be fed into the RNN by a series of transformation functions, such as filtering, scaling and baseline shifting. An example for both the input and output time series are shown in (b).

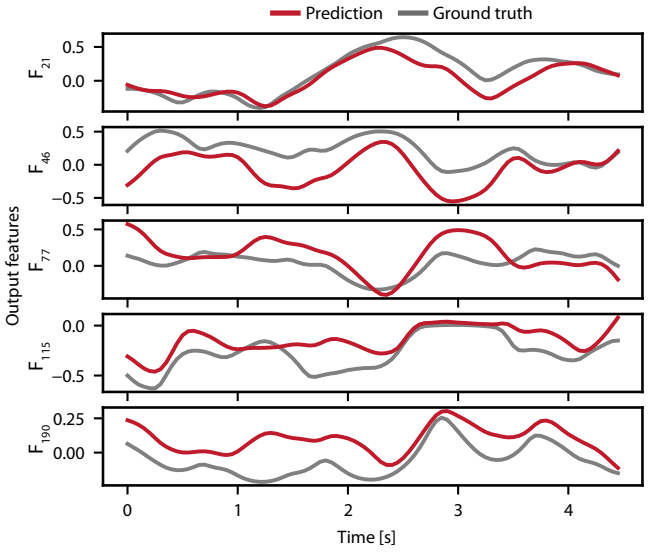


Fig. 5: The result of the time series prediction for five illustrative markers. The recurrent neural network was able to predict the trend of the position of the markers, however, some error between the prediction and the ground truth is noticeable.

Before the training, we used a first-order Butterworth low-pass filter with a cutoff frequency of f_{cutoff}=0.5 Hz on the datasets to clean the input and output time series. The result of the cleaning is shown in Figure 4. At each step of the training, we used a dataset that started at a random time step and contained the data of each liquid displacement channel for the previous 10 time steps. The algorithm's goal was to minimise the root mean square error at the 11th time step between the predicted and ground truth output time series (absolute positions of the markers). The schematic of this process is shown in Figure 4. On the input side, we fed the recurrent neural network with 16 $\times$ 10 datapoints (number of liquid channels times the time dependency parameter) and it predicted 200 features (two-dimensional position of the 100 markers) on the output side at the next time step.

As shown in Figure 5, the algorithm was able to predict the trend of the time series, however, some error between the ground truth and the prediction is noticeable. This, did not change substantially when we changed the number of layers, their sizes, the scaling algorithm (instead of compressing the time series into the [-1,1] range, transforming the dataset so that is has zero mean and unit variance), or the cutoff frequency of the filter or the optimisation algorithm.

B. Video frame reconstruction

In this section, we introduce an algorithm that learns to map the low-resolution time series to the video frames. Predicting video frames using the architecture shown in Figure 4 is challenging as the output data has typically two or three orders of magnitude higher dimensionality than the input data. In our case, with a one channel (grayscaled) video frame with the size of 68 x 68 pixels we have 4624 datapoints at each time step on the output side. For this reason, we trained a stacked convolutional autoencoder that could encode the video frames and decrease the dimensionality of the output data at the same time. The RNN+CAE combined model is shown in Figure 6. First, the CAE was trained using the first 3780 video frames to find an encoded representation of each frame. At this point the encoded representation of each frame was a three-dimensional array with a size of 9$\times$9$\times$4 = 324. Note that many of these features were zero and did not change over time, which made the RNN difficult to train. For this reason, a filtering algorithm was implemented that automatically detected these zero features and removed them from the dataset. After removing these features the size of the encoded representation dropped to 306. This array was used as an output of the recurrent neural network whereas the input was the same as in the previous section. The result of the prediction and the image reconstruction is shown in Figure 7. Although, the RNN is not always able to reproduce the large frequency oscillations

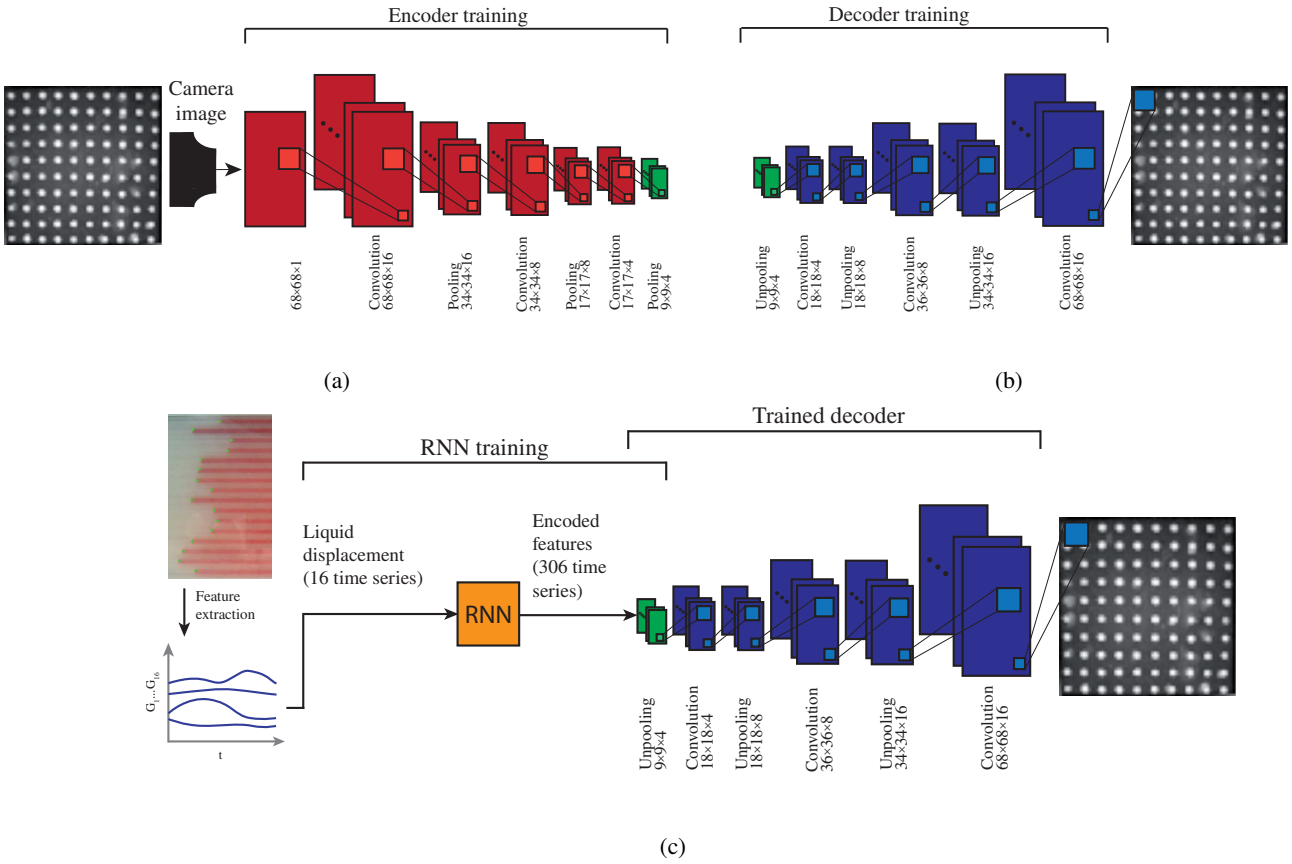


Fig. 6: The system architecture and the learning process. First, the dimensionality of the video frames of the high-resolution sensor were reduced by the stacked convolutional autoencoder. During the training we used the Adam optimiser algorithm, 100 iterations and a batch size of 50 to minimise the binary cross-entropy loss between the input and output of the CAE. Both the encoder (a) and decoder (b) components had five hidden layers. (c) After training the CAE, the recurrent neural network was trained using the time series of the low-dimensional sensor as input and the encoded representation of the video frames of the high-resolution sensor as output. Last, the prediction of the RNN was fed into the decoder component in order to reconstruct the video frames.

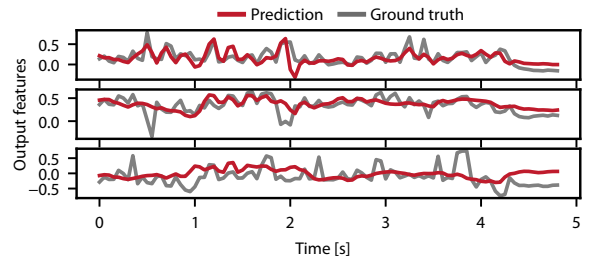
of the encoded features, it is able to follow the lower frequency trend of these features. This prediction is then fed into the decoder component of the previously trained CAE and the video frames are reconstructed. Due to the filtering effect of the prediction, some tactile interactions are lost in the reconstructed video, but many the interactions are correctly predicted.

IV. Discussion

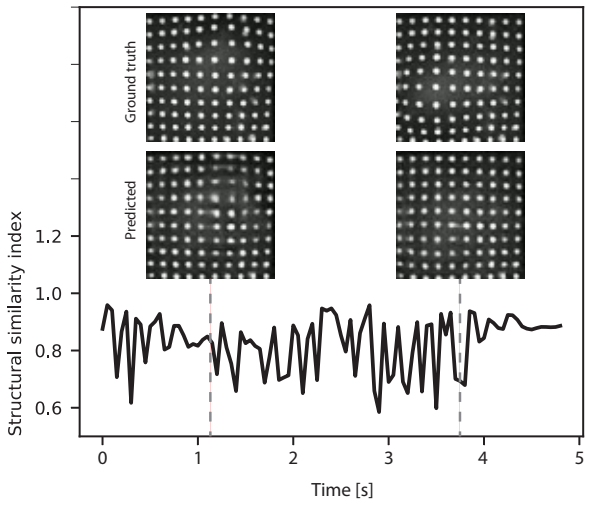
In this paper, we present a data-driven approach for scaling CCD camera based tactile sensors. We designed, fabricated and tested soft interfaces that incorporate two independent measurement methods. The low-resolution sensing method is based on fluidic displacement, whereas the high-resolution sensor uses marker tracking in order to measure the deformation of the elastic interface. We collected data of the interaction between the sensor and a human finger. We preprocessed this data and trained two different machine learning models. The first one used the extracted features of the liquid displacement sensor as input and the two-dimensional position of the markers on the bottom layer of the interface. In the second architecture, the input remained the same, whereas the output of the model was the reconstructed video frames of the marker tracking CCD camera. Due to the dimensionality mismatch, we used a stacked convolutional autoencoder in order to decrease the dimensionality of the video frames and then we used the encoded features as the output of the recurrent neural network. This way, we could reconstruct the high-resolution information of the tactile interface using only the time-series of the low-resolution sensor.

In both cases we observed similar characteristics: the algorithm was able to follow the low frequency trend of the time series, but often failed to predict target with a small error. One solution is to collect more data, but since the data collection is currently a manual process, it might not be feasible. Automating the data collection would help not only create more data, but to create more data with higher quality.

During the data processing it was observed that both



(a)



(b)

Fig. 7: (a) Prediction of the encoded features. Three representative examples of the 306 non-zero time series are shown. (b) The structure similarity index was used to quantify the end-to-end error of the reconstruction. The original and predicted frames are shown at t = 1.15 s and t = 3.75 s.

sensors produced drift, which caused problems during the training of the machine learning models. This was due to the softness of the interface as it produced a slight, permanent deformation during the data collection. We corrected this manually by transforming the datasets to have the same mean and variance. There are three potential solutions for this: (1) using a stiffer interface, (2) automating the dataset correction by constantly measuring the mean and the variance of the distribution and (3) using the relative -rather than absolute - features of the sensors (i.e. the relative positions of the markers).

The internal channels with the liquid transmission-based sensing encode the deformation of the tactile sensor. This means that the geometry and structure of channels have an effect on training data and, therefore, on the learning process. We expect to optimise the design using fewer channels while preserving the maximum amount of information during the interaction. This will further increase the scalability of the CCD camera-based soft tactile interfaces and the accuracy of our method [22].

References

[1] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora. The TacTip family: Soft optical tactile sensors with 3D-printed biomimetic morphologies. *Soft Robotics*, 5(2):216–227, 2018. PMID: 29297773.

[2] N. F. Lepora, A. Church, C. de Kerckhove, R. Hadsell, and J. Lloyd. From pixels to percepts: Highly robust edge perception and contour following using deep learning and an optical biomimetic tactile sensor. *IEEE Robotics and Automation Letters*, 4(2):2101–2107, April 2019.

[3] K. Sato, K. Kamiyama, N. Kawakami, and S. Tachi. Finger-shaped gelforce: Sensor for measuring surface traction fields for robotic hand. *IEEE Transactions on Haptics*, 3(1):37–47, Jan 2010.

[4] J. W. James, N. Pestell, and N. F. Lepora. Slip detection with a biomimetic tactile sensor. *IEEE Robotics and Automation Letters*, 3(4):3340–3346, Oct 2018.

[5] C. Chorley, C. Melhuish, T. Pipe, and J. Rossiter. Tactile edge detection. In *SENSORS, 2010 IEEE*, pages 2593–2598, Nov 2010.

[6] K. Kamiyama, H. Kajimoto, M. Inami, N. Kawakami, and S. Tachi. Development of a vision-based tactile sensor. *IEEJ Transactions on Sensors and Micromachines*, 123:16–22, 01 2001.

[7] R. B. N. Scharff, R. M. Doornbusch, E. L. Doubrovski, J. Wu, J. M. P. Geraedts, and C. C. L. Wang. Color-based proprioception of soft actuators interacting with objects. *IEEE/ASME Transactions on Mechatronics*, 24(5):1964–1973, 2019.

[8] S. Dong, W. Yuan, and E. H. Adelson. Improved gelSight tactile sensor for measuring geometry and slip. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 137–144, Sep. 2017.

[9] C. Chorley, C. Melhuish, T. Pipe, and J. Rossiter. Development of a tactile sensor based on biologically inspired edge encoding. In *2009 International Conference on Advanced Robotics*, pages 1–6, June 2009.

[10] D. Hristu, N. Ferrier, and R. Brockett. The performance of a deformable-membrane tactile sensor: Basic results on geometrically-defined tasks. *Proceedings - IEEE International Conference on Robotics and Automation*, 1:508 – 513 vol.1, 02 2000.

[11] N. J. Ferrier and R. W. Brockett. Reconstructing the shape of a deformable membrane from image data. *The International Journal of Robotics Research*, 19(9):795–816, 2000.

[12] K. Kamiyama, K. Vlack, T. Mizota, H. Kajimoto, K. Kawakami, and S. Tachi. Vision-based sensor for real-time measuring of surface traction fields. *IEEE Computer Graphics and Applications*, 25(1):68–75, Jan 2005.

[13] W. Yuan, S. Dong, and E. H. Adelson. GelSight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12), 2017.

[14] W. Yuan, M. A. Srinivasan, and E. H. Adelson. Estimating object hardness with a gelSight touch sensor. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 208–215, Oct 2016.

[15] S. Luo, W. Yuan, E. Adelson, A. Cohn, and R. Fuentes. ViTac: Feature sharing between vision and tactile sensing for cloth texture recognition. 01 2018.

[16] R. Li and E. H. Adelson. Sensing and recognizing surface textures using a gelSight sensor. In *2013 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1241–1247, June 2013.

[17] S. Dong, W. Yuan, and E. Adelson. Improved gelSight tactile sensor for measuring geometry and slip. pages 137–144, 09 2017.

[18] G. Soter, M. Garrad, A. T. Conn, H. Hauser, and J. Rossiter. Skinflow: A soft robotic skin based on fluidic transmission. In *2019 2nd IEEE International Conference on Soft Robotics (RoboSoft)*, pages 355–360, April 2019.

[19] S. Hochreiter and J. Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.

[20] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber. Stacked convolutional auto-encoders for hierarchical feature extraction. In *Artificial Neural Networks and Machine Learning – ICANN 2011*, pages 52–59, Berlin, Heidelberg, 2011. Springer Berlin Heidelberg.

[21] G. Soter, A. Conn, H. Hauser, and J. Rossiter. Bodily aware soft robots: Integration of proprioceptive and exteroceptive sensors. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2448–2453, May 2018.

[22] E. J. K. Judd, K. M. Digumarti, J. M. Rossiter, and H. Hauser. Neatskin: A discrete impedance tomography skin sensor. In *Robosoft: IEEE International Conference on Soft Robotics*, 4 2020.