Artificial SA-I, RA-I and RA-II/vibrotactile afferents for tactile sensing of texture

Nicholas Pestell and Nathan F. Lepora

Department of Engineering Mathematics and Bristol Robotics Laboratory, University of Bristol, Bristol BS8 1QU, UK

Robot touch can benefit from how humans perceive tactile textural information, from the stimulation mode to which tactile channels respond, then the tactile cues and encoding. Using a soft biomimetic tactile sensor (the TacTip) based on the physiology of the dermal–epidermal boundary, we construct two biomimetic tactile channels based on slowly adapting SA-I and rapidly adapting RA-I afferents, and introduce an additional sub-modality for vibrotactile information with an embedded microphone interpreted as an artificial RA-II channel. These artificial tactile channels are stimulated dynamically with a set of 13 artificial rigid textures comprising raised-bump patterns on a rotating drum that vary systematically in roughness. Methods employing spatial, spatio-temporal and temporal codes are assessed for texture classification insensitive to stimulation speed. We find: (i) spatially encoded frictional cues provide a salient representation of texture; (ii) a simple transformation of spatial tactile features to model natural afferent responses improves the temporal coding; and (iii) the harmonic structure of induced vibrations provides a pertinent code for speed-invariant texture classification. Just as human touch relies on an interplay between slowly adapting (SA-I), rapidly adapting (RA-I) and vibrotactile (RA-II) channels, this tripartite structure may be needed for future robot applications with human-like dexterity, from prosthetics to materials testing, handling and manipulation.

1. Introduction

Human tactile texture sensation is probably governed by complex relationships between the physical dimensions and properties of the texture and the nature of the mode of stimulation. Researchers have proposed a range of mechanisms for encoding properties such as roughness [1,2] in peripheral neural activity, encompassing different tactile channels (afferent types) and representations within spike trains (rate codes [3], population codes [4] and temporal codes [4]). That said, a combination of afferent types and encoding schemes are probably employed to form a complete tactile portrait of texture, as has been observed with other tactile dimensions [5].

Tactile texture perception has been examined widely in robotics with primary focus on classification of discrete texture classes, commonly with bioinspired sensors employing a vibrotactile modality [6–9]. The trend has been towards using complex natural textures [10–12] that vary across a range of dimensions, such as roughness, hardness and stickiness [8]. The focus in robotics has been on achieving high-accuracy on natural textures, for example by using large datasets and high-resolution data. Using natural textures does have its drawbacks though. For example, it opens a disconnect with investigations of human touch where there is a long tradition of using artificial stimuli that vary systematically in a single perceptual dimension such as roughness.

This approach towards high-accuracy texture classification is exemplified in [13], where a range of techniques were examined for classifying 23 natural textures using the iCub’s tactile forearm; the best performing method was a combined convolutional neural network–long short-term memory (CNN–LSTM) neural network model trained on spatio-temporal pressing and sliding
data. Another example of texture classification is with the GelSight optical tactile sensor, where fine-grained spatial information is encoded in high-resolution tactile images [10]. While these methods are clearly capable of achieving high accuracy on raw data from the sensors, they were not intended to relate to how humans perceive texture or be data efficient.

These robotic approaches are also orthogonal to investigations of the fundamental aspects of texture perception being considered in biology (e.g. [14–16]). For artificial tactile sensing, as in nature, one might expect that the ability of systems to transduce textural information depends on a complex relationship between the stimulus scale, the mode of stimulation, the properties of the tactile channel used to transduce information and the dimensionality of the tactile data. Thus, to achieve artificial texture perception with human-like capabilities, it may be necessary to instantiate artificial tactile sensing that mimics the various biological modalities that comprise natural touch sensing.

To examine this proposal, we use an established soft biomimetic optical tactile fingertip, the TacTip (figure 2) [17,18], from which we extract three distinct feature sets proposed to model natural tactile channels in human touch; these features are then used to classify a set of 13 artificial textures via dynamic stimulation. Two of these feature sets are here referred to as artificial SA-I and RA-I afferents, because a related paper demonstrates that they give viable models of the corresponding natural afferents for representing spatial information such as grating orientation [19]. This study considers a fuller set of artificial SA-I, RA-I and RA-II tactile afferents, with the main novel contributions:

(i) the introduction of a novel vibrotactile channel to the TacTip, proposed to mimic natural RA-II tactile afferents, by using a microphone embedded within the soft inner gel of the tactile sensor;
(ii) a method for speed-insensitive texture perception based on using frictional cues encoded within spatial modulation of artificial afferent response, tested with textured stimuli comprising raised-dot patterns similar to those used to vary roughness in perceptual psychophysics experiments [4]; and
(iii) a method for speed-insensitive texture perception in the vibrotactile channel based on a theory of speed-invariant texture perception in humans, in that texture is signalled by the harmonic structure of induced vibrations encoded within the subsequent harmonic structure of afferent firing [20,21].

The results from our experiments identify new principles for improving robot touch, particularly in terms of data-efficient methods for transducing tactile sensory information and extracting biologically inspired tactile features that could both transcend the particular type of tactile sensor and extend to tactile dimensions beyond texture classification.

2. Methods

2.1. Sensor design and fabrication

The relevant operating principles of the Bristol Robotics Laboratory (BRL) TacTip are described below in relation to the customization of the design for texture sensing. For manufacturing and detailed explanations of the design concepts, we refer the reader to a 2018 paper on ‘The TacTip Family’ [17] and a recent (2021) review of ‘Soft Biomimetic Optical Tactile Sensing with the TacTip’ [18].

The BRL TacTip has two main sections (figures 1 and 2): a rigid body housing the electronic components, which is three-dimensionally printed in ABS, and a soft tip fabricated using a multi-material fused deposition modelling printer. The tip comprises two materials printed as a single part: a rubber-like ‘skin’ (Tango Black Plus, Stratasys) and rigid rim (VeroWhite, Stratasys). The inside of the tip is filled by injecting a clear silicone gel (RTV2790S, Techsil UK). The gel provides stiffness to the tip which helps to minimize hysteresis while enabling compliance with a stiffness difference between the skin (Shore A 26–28) and gel (Shore OO 10). The gel is optically clear so that three-dimensionally printed white markers on the inside of the compliant skin can be imaged by an internal camera.

The TacTip body’s primary function is to house electronic components and to serve as a mount for auxiliary robotic systems. To capture images of the white markers, a USB camera module (Ailipu Technology) is used at resolution 640 × 480 pixels and frame rate 90 fps. The length of the TacTip body is chosen to provide full view of the markers through the wide-angle lens. At the base of the body where it joins with the tip is a printed circuit board ring containing six white LEDs that illuminate the markers for imaging by the camera.

The sensor modifications for this work are as follows.

Microphone transduction. In accordance with the aforementioned motivations, multi-modality is facilitated by using acoustic vibrations occurring within the gel during dynamic stimulation. To detect these vibrations, a small electret microphone of diameter 6 mm (Kingstate, KEKG3642TF-A) is fitted to the inside of the TacTip’s VeroWhite rim (figure 2). During assembly, the microphone is press fitted into a cavity prior to filling with gel. The gel is in contact with the front surface of the microphone and the cabling is routed through a hole in the side of the tip. The sensor is designed so that physical interaction between the TacTip’s ‘skin’ and a stimulus will induce vibrations that propagate through the silicone gel, which can then be picked up and converted to a voltage by the microphone. The signal is amplified to line-level ($V_{pp} \approx 2$ V) using an inverting op-amp circuit.

Tip shape. In this work, a flat skin is used to facilitate contact with textured surfaces. The rationale behind this modification is to reduce the physical distance between the point of interface with stimuli and the microphone. Acoustic vibrations will attenuate as they travel through the silicone. In practice, after the tip has been filled with gel the skin is no longer completely flat, even though efforts were taken to maintain a flat profile during fabrication. In particular, gravity causes the skin surface to slightly ‘bulge’ owing to the weight of the gel when the sensor is oriented downwards.

Marker layout. The TacTip design here uses 361 pins/markers that is about three-times finer than the conventional TacTips used in other studies [17]. This design was both to improve spatial resolution and to have a common pin array with [19]. The pins are arranged in a simple 19x19 square array with separation 1.2 mm (figure 1). A regular square array is used so that information encoded in the relationship between movements of neighbouring markers (spatial coding) can be more easily identified. The marker density of approximately 70 cm$^{-2}$ is approximately half the innervation density of type-I afferents in the human fingertip [22,23].

Artificial fingerprint. Scheibert et al. suggest that the human fingerprint may amplify the vibrations produced at the interface of the skin and stimulus, and so benefit texture perception [24]. Our preliminary investigations also confirmed that this effect occurs with our proposed artificial fingertip design. Here we
use the design concepts realized by Cramphorn et al. [25] to similarly modify our multi-modal design to mimic the function of an external fingerprint and internal epidermal ridges. Specifically: (i) the outside of the skin is augmented with small nodules (diameter 2 mm) that form the physical analogue of the papillary ridges that comprise a fingerprint; these nodules are made from the same material as the skin and are thus also compliant. The pattern of nodules on the skin exterior mirrors that of the pins on the interior, with the nodules located directly below one-in-two pins (figure 2); and (ii) each pin contains a rigid ‘core’ that is mechanically fused with its respective white marker. The function of these cores is to enhance the stiffness contrast between the pins and the silicone gel to mimic the function of the stiffer epidermal ridges [26].

2.2. Data collection

For experimental testing, a set of textured drum stimuli are used that are patterned with tetragonal arrays of raised bumps on their exterior circumferences (figure 3, \( N_{\text{tex}} = 13 \)). The drums are 3D-printed in a rigid plastic (VeroWhite, Stratasys) using a PolyJet printer (Objet, Stratasys), with the textured bumps raised 0.5 mm beyond the drum’s 80 mm diameter. Both the diagonal bump separation, \( x \), and the bump diameter, \( d \), are varied systematically to provide a linear variation in texture period while constraining the area of raised bumps across stimuli to be constant (figure 3). This constraint is important because it prohibits the availability of intensive cues owing to surface area that would be available from the contact, rather than the roughness of the texture.

The tactile sensor is held stationary against textured drums (figure 4), which are driven by a 60 W brushed motor connected to the drum spindle. This connection uses a 4.8 : 1 planetary gear head and a custom-built 3 : 1 belt drive. The drum velocity is controlled in closed loop.

During the experiments, the sensor is held with interpenetration depth close to 1 mm and pointing vertically downwards, with the apparatus designed so that this indentation is constant upon changing drums. This depth was tuned manually to give sufficient contact for reliable tactile signals while minimizing wear on the abrasive surface of the rotating drums. Each drum is rotated over a set of \( N_{\text{speeds}} = 10 \) linear speeds, linearly spaced from 10 to 100 mm s\(^{-1}\), giving a total of \( N_{\text{tex}} N_{\text{speeds}} = 130 \) stimulus conditions. During each experimental run, the TacTip is stimulated for 65 s per stimulus condition while simultaneously recording TacTip video and microphone data. The initial and final 2.5 s of each recording are thrown away to keep only the constant-speed stimulation.

2.2.1. Tactile image data

The TacTip camera output for each stimulus condition is truncated to 5000 frames (55 s) to ensure an equal number of samples per class. Samples are generated from individual frames. To examine the capacity for these methods to enable speed-invariant texture classification, tactile image data are separated into training, validation and testing sets using a 10-fold leave-one-speed-out
cross-validation procedure, where in each of the 10 individual sets a different stimulus speed is ‘held out’. Data from the held out speed are randomly split (50:50) to produce validation and test sets and the remaining nine speeds used for training.

Spatial encoding. Samples are constructed from individual frames, where \( N_{\text{train}} = 45,000 \), \( N_{\text{val}} = 2500 \) and \( N_{\text{test}} = 2500 \) samples per class, respectively. Over all textured stimuli, \( N_{\text{tex}}N_{\text{train}} = 585,000 \), \( N_{\text{tex}}N_{\text{val}} = 32,500 \) and \( N_{\text{tex}}N_{\text{test}} = 32,500 \) samples, respectively. Tactile spatial encoding arises from the spatial structure within the image frames captured by the camera.

Spatio-temporal encoding. Samples are constructed from sets of 10 adjacent tactile images, yielding a total of 500 samples per stimulus condition (in analogy with human perception, this is a 0.1 s encoding window). Thus, \( N_{\text{tex}}N_{\text{train}} = 58,500 \), \( N_{\text{tex}}N_{\text{val}} = 3,250 \) and \( N_{\text{tex}}N_{\text{test}} = 3,250 \) samples, respectively. Each sample is now a \( 640 \times 480 \times 10 \) three-dimensional array, where the decoding model can extract features contained within this full three-dimensional spatio-temporal structure.

With a finite amount of data, there is a trade-off between encoding window length and the number of samples. We experimented with different duration of encoding windows and found that a window of 0.1 s (10 time samples) provided sufficient numbers of samples from our training data while also enabling the validation accuracy to reach a reasonable value.

### 2.2.2. Microphone data

**Temporal encoding.** Overlapping samples of 1 s with stride 0.5 s are generated from sets of 88,200 audio data points, giving 58 samples per stimulus condition. As with the tactile image data, microphone data are separated into train, validation and test sets using a similar 10-fold leave-one-speed-out cross-validation procedure. In each set, a different stimulus speed is ‘held out’ and randomly split 29:30 for validation and test sets. The remaining nine speeds are used for training, giving \( N_{\text{tex}}N_{\text{train}} = 69,030 \), \( N_{\text{tex}}N_{\text{val}} = 3,770 \) and \( N_{\text{tex}}N_{\text{test}} = 3,900 \) samples respectively. These data samples permit only a purely temporal encoding over the one-dimensional array of temporal data.

### 2.3. Feature extraction

The methods below for generating artificial SA-I and RA-I afferents are shown in a related paper to model many aspects of firing rate codes of their natural counterparts, e.g. adaptation rates, spatial modulation and sensitivity to edges, and have been shown to provide biologically plausible population codes for grating orientation discrimination [19]. Furthermore, the RA-I model is similar to other transduction models where the first derivative of pressure is used as the primary input to a biological neuron model [27–29].

Likewise, the processing of the vibrotactile channel is based on the prevailing theory of human texture perception that proposes the observed invariance to scanning speed is based on mechanisms analogous to those that give timbre invariance in hearing [21]. The harmonic structure of induced vibrations and resulting neuron firing is consistent across scanning speeds [20,30], with somatosensory cortex extracting these textural properties from the harmonic structure of neural activity.

| Figure 3. Patterns of 13 textured stimuli, from completely smooth (a) to coarsest texture (m). The textures comprise raised bumps of diameter \( d \) and bump-to-bump spacing \( x \) from one centre to another along the diagonals. The values \( x, d \) are chosen to maintain the same overall bump area for all non-smooth patterns, so that contact area is not a stimulus cue. |
|---|
| Figure 4. Experimental set-up for dynamic touch experiments with textured stimuli. The tactile sensor is held stationary with its skin against a rotating textured drum whose speed is modulated by a DC motor and drive mechanism. |
2.3.1. Artificial SA-I afferents
SA afferents refer to a class of afferents with slow adaptation rates. That is, they respond to sustained pressure or deformation of the skin as their associated mechanoreceptors (Merkel cells) fire tonically. Type-I implies that the afferents densely innervate the skin regions and exhibit small receptive fields with well-defined borders [22]. They are therefore associated with high spatial acuity.

Artificial SA-I activity for each channel is modelled by Euclidean distance of a marker from its at-rest position. Considering a sustained stimulus, the deformation of the tip will remain consistent and therefore so will the positions of the markers. In practice, marker positions are extracted from tactile image data with a simple blob detection algorithm implemented using OpenCV in Python. The artificial SA-I response at frame \( i \) for marker \( n \), \( \text{SA}_{n,i} \), is computed as the Euclidean distance between the marker positions \( (x_{n,i}, y_{n,i}) \) and an initial undisturbed at-rest frame \( (i = 0) \):

\[
\text{SA}_{n,i} = \sqrt{(x_{n,i} - x_{n,0})^2 + (y_{n,i} - y_{n,0})^2}.
\]  

Thus, the artificial SA-I spatial encoding and spatio-temporal encoding samples are of dimensions \( 19 \times 19 \) and \( 19 \times 19 \times 10 \), respectively.

2.3.2. Artificial RA-I afferents
By contrast, RA-I afferents are rapidly adapting, meaning their firing tends to decrease rapidly when subject to sustained stimulus; they also densely innervate the skin [23] and have small receptive fields that we again associate with physical movement of the TacTip markers. It is believed that these afferents are particularly sensitive to the velocity of the skin within the receptive field [22,31] and thus tend to fire whenever skin is moving. We model this behaviour as marker speed, which is inherently transient; i.e. it will be zero when the stimulus is sustained and positive when the stimulus changes.

Artificial RA-I activity is derived from the same tactile image data as the SA-I activity. The response at frame \( i \) for marker \( n \), \( \text{RA}_{n,i} \), is computed as the absolute difference between the artificial SA-I responses on adjacent frames, representing a radial speed of marker motion relative to the marker positions in the initial at-rest frame:

\[
\text{RA}_{n,i} = |\text{SA}_{n,i} - \text{SA}_{n,i-1}|.  
\]  

As with artificial SA-I samples, RA-I spatial encoding and spatio-temporal encoding samples are of dimensions \( 19 \times 19 \) and \( 19 \times 19 \times 10 \), respectively.

One should recognize that equations (2.1) and (2.2) are simplifications of the afferent activity and may not encode some important skin dynamics; for example, neither the sign nor circumferential component of the marker velocity is represented in the artificial RA-I activity. Our intent here is to provide a parsimonious representation of afferent activity from simple transformations of the optical tactile output, as a foundation for improved models in the future.

2.3.3. Vibrotactile channel
A second rapidly adapting population, the RA-II afferents, have innervation densities considerably lower than that of type-I afferents [32]. Each RA-II afferent is terminated by a single Pacinian corpuscle, which are sparsely located deep in the subcutis with large receptive fields having obscure boundaries. They are sensitive primarily to high-frequency vibrations (50–500 Hz) with extremely low-amplitude thresholds (1 \( \mu \mbox{m} \)) in the range of 100–300 Hz [33].

In design, the vibrotactile channel is analogous to the form of the natural RA-II channel in two ways: firstly, the microphone innervates the TacTip sparsely relative to markers; secondly, it is located deep within the soft silicone gel analogous to the dermis. The information acquired by the electret microphone is also similar to the dynamics of the natural RA-II channel in being: (i) responsive to physical vibrations (acoustics), which are transient by definition; (ii) having an operating frequency range 10 Hz–20 kHz (sampled at 44.1 kHz) that contains that of the Pacinian system in its lower range; and (iii) exhibiting extreme sensitivity to low-amplitude vibrations. Note also that the front sensitive area of the microphone is embedded within the supporting visco-elastic gel of the tactile sensor, which appears to have attenuated frequencies above approximately 150 Hz (e.g. figure 7). Hence, the effective range of the embedded microphone is more similar to that of natural RA-II afferents than an auditory range.

The method of feature extraction for the artificial vibrotactile channel is based on extracting its harmonic structure. From each sample of raw audio data in the time domain, a one-dimensional vector of amplitudes at discrete frequencies is constructed with a fast Fourier transform (FFT). These feature vectors are truncated to a maximum frequency of 200 Hz, corresponding to a discrete feature vector size of \( 200 \times 1 \). The amplitudes are scaled to fall between 0 and 1 with a division by the largest amplitude across the entire training dataset.

2.4. Tactile encoding models
All tactile encoding models are trained in the same way using Keras with TensorFlow backend. Training data are fed through the networks in batches of 64 samples. The entire set is propagated for a maximum of 150 epochs. An adaptive momentum estimation optimizer (ADAM) on a categorical cross-entropy loss is used for updating the weights after each batch. Training may stop early if validation accuracy plateaus with the use of a patience factor of 30 epochs. For the spatial, spatio-temporal and temporal encoding/decoding methods, a separate model is trained for each of the 10 held-out speeds.

Overall, there are 10 spatial encoding SA-I and RA-I texture classification models, one per speed \( v \): SA-SE-\( v \) and RA-SE-\( v \), respectively, and likewise 10 spatio-temporal models, SA-STE-\( v \) and RA-STE-\( v \). There are also 10 temporal encoding vibrotactile texture classification models, vib-TE-\( v \).

2.4.1. SA-I and RA-I afferents: spatial encoding models
Spatial encoding models of artificial SA-I and RA-I afferents are constructed with a two-dimensional convolutional neural network (schematic shown in figure 5a). This network permits \( 19 \times 19 \) tactile images as input and outputs a 13 x 1 vector with each element corresponding to an individual texture class. All hidden layers use ReLU activation functions and the output layer uses softmax activations on each neuron. Regularization techniques include drop-outs of 0.4, 0.2 and 0.2 prior to layers FC-1, FC-2 and the output, respectively, and L2 regularization (factor 0.005) on each dense layer. Batch normalization is used after each convolutional layer. This architecture was chosen by manually tuning network hyper-parameters for good validation performance over all textures using a single set of hyper-parameters for all models for each of the leave-one-out speeds. In practice, this tuning was straightforward and many similar architectures would have worked well.

2.4.2. SA-I and RA-I afferents: spatio-temporal encoding models
Spatio-temporal decoding models of artificial SA-I and RA-I afferents are constructed with a convolutional-LSTM (ConvLSTM) built over the two-dimensional convolutional network described above (schematic shown in figure 5b).

In essence, a frame at each timestamp, \( t_p \), of a sample is passed through a feature detector ‘CNN’ taken directly from...
that used in the spatial decoding model (figure 5a; layers Conv2D-1 to MaxPool-2). The output feature maps are flattened before being passed to a multi-layer LSTM unit, consisting of three layers with 10 LSTM blocks each. The output of the LSTM unit is passed to a fully connected unit (FC) also taken directly from the spatial decoding model (figure 5a: layers FC-1 to FC-2). An output layer, consisting of 13 neurons, provides a prediction for each class.

Reusing the feature detection and fully connected parts of the spatial decoding model provides a more controlled comparison between encoding mechanisms; i.e. differences in performance between static and dynamic touch can more easily be attributed to the availability of temporal features. The LSTM hyper-parameters were chosen by manual tuning for good validation performance over all textures using a single set of hyper-parameters for all models.

### 2.4.3. Vibrotactile channel: temporal encoding models

Temporal decoding models of the vibrotactile channel are constructed with a one-dimensional convolutional neural network (schematic shown in figure 6). This network takes $200 \times 1$ harmonic feature vectors as input and outputs a $13 \times 1$ vector corresponding the texture classes. All hidden layers use ReLU activation functions and the output layer uses softmax activations on each neuron. Regularization techniques included a drop-out of 0.4 prior to layer FC-1 and L2 regularization (factor 0.005) on layer FC-1. Batch normalisation is used after each convolutional layer. Again, this architecture was chosen by manually tuning hyper-parameters for good validation performance over all textures for all models.

A data augmentation procedure is implemented to improve the accuracy of these temporal encoding/decoding methods and encourage generalization across speeds and textures. Informed by the hypothesis that a speed-invariant harmonic structure of induced vibration is a viable cue for textural properties [21,30], we propose that data collected at different speeds can be simulated by ‘stretching’ and ‘compressing’ the FFT samples in the frequency domain (figure 7). For each original training sample, six augmented samples are generated: three by stretching in the frequency domain by a factor uniformly sampled.
between scaling factors of [1,2] to represent higher speeds and the remaining three compressed by a factor uniformly sampled between scaling factors of [0.5,1] to represent lower speeds.

Overall, $6 \times N_{\text{train}} = 714$ augmented training samples per texture and speed are generated, where $N_{\text{train}} = 119$ is the original number of training samples per texture class and speed. For each of the 10 temporal encoding models, $6 \times 9 \times 13 \times N_{\text{train}} = 83538$ augmented training samples are available over all 13 textures trained in the leave-one-speed-out classification model.

3. Results

3.1. Artificial SA-I afferents

3.1.1. Visual inspection of data

Example SA-I tactile images are shown for the smoothest $a$, middle $g$ and coarsest $m$ textured stimuli at the slowest, middle and fastest speeds ($10 - 100$ mm s$^{-1}$) (figure 8a). The 10 example tactile images for each texture were collected at the speed indicated on the image label, e.g. $g$-10 is for texture $g$ at 10 mm s$^{-1}$.

From visual inspection, we could not observe any features in the SA-I tactile images that robustly signified texture class. For example, images $m$-20 and $g$-40 in figure 8a for different texture classes look similar, while there is a considerable variation with speed within both texture classes $m$ and $g$.

To eliminate the possibility that the variation in figure 8a is owing to stochastic variation in time with contacting each texture, figure 8b shows tactile images of artificial SA-I on stimulus class $g$ (class with the median bump spacing). Rows are separated according to speed (10, 50 and 100 mm s$^{-1}$) and consist of time series of 10 consecutive frames (increasing times $t_1$ to $t_{10}$). Tactile images appear similar in time (rows) and dissimilar in speed (columns).

Considering figure 8a,b in combination, we conclude that the spatial modulation of artificial SA-I firing over the TacTip is highly dependent on both texture and speed. The dependence is such that changes in speed can compensate changes in texture (and vice-versa).

Note also that many of the SA-I tactile images exhibited asymmetry. For example, a light crescent shape to the left or right of the image (e.g. $g$-40, $g$-30, $m$-20), suggesting that when stimulated via sliding, the TacTip’s deformation was dominated by shearing effects.

3.1.2. Texture classification

Accuracies of all texture classification models over artificial SA-I, RA-I and vibrotactile afferents averaged across the hold-out speeds are reported in table 1. Here, we focus on the RA-I afferents, and show the corresponding confusion matrices using temporal (TE) and spatio-temporal encoding (STE) schemes in figure 9a,b, respectively.

For both the spatial and spatio-temporal encoding models (SA-SE, SA-STE), each model performed better than chance suggesting that there is coherence within texture classes in terms of speed-invariant spatial features (10 panels in each of figure 9a,b). The overall accuracy averaged across all 10 models for the distinct speeds was approximately 46% in both cases (table 1), which were the lowest of all models considered.

For all hold-out speeds, near-perfect predictions were made for stimulus class $a$ for a completely smooth texture (top-left of all panels in figure 9a,b). Overall, there is a general trend towards better performance at either end of the range of stimuli, which may be because textures at the extremes of the roughness scale have fewer neighbouring classes, making prediction easier. In addition, the extremes may be more distinctive; for example, the completely smooth texture is easily recognizable.

The best-performing SA-I model was for spatial encoding with hold-out speed of 80 mm s$^{-1}$ (SA-SE-80), attaining a classification accuracy of 63%. There appears to be a trend between increasing hold-out speed and improved performance, i.e. more salient data being created when the stimulation speed is increased. However, this relationship seems to be attenuated at the extremes of the speed range, which we attribute to the network being trained only on data collected either at faster or slower speeds for 10 and 100 mm s$^{-1}$, respectively, leading to poorer generalization.

Overall, the spatial and spatio-temporal encoding models (SA-SE, SA-STE) performed similarly, suggesting that models based on SA-I afferents do not benefit from the addition of a temporal dimension; i.e. artificial SA-I afferents do not encode texture temporally.

3.2. Artificial RA-I afferents

3.2.1. Visual inspection of data

Example RA-I tactile images are shown for the smoothest $a$, middle $g$ and coarsest $m$ textured stimuli (figure 10a) made from frames corresponding to those for the SA-I tactile images shown in figure 8a. Likewise, figure 10b shows example RA-I tactile images just for the middle stimulus class $g$ for three speeds (columns: $v = 10, 50$ and $100$ mm s$^{-1}$) and 10 consecutive frames (rows: times $t_1$ to $t_{10}$).

There is less structure within these artificial RA-I tactile images compared with their SA-I counterparts; for example,
the geometry of the contact region is no longer as apparent in figure 10a,b. This difference in apparent structure is probably a consequence of the way the samples were constructed: each RA-I tactile image is in effect constructed from the difference in adjacent SA-I tactile images, which look similar (figure 8b). However, even though there is less geometric structure, the information may still be informative about the texture, as will be examined below.

As with the artificial SA-I tactile images, RA-I images do not exhibit any clear visual features which robustly signify texture class (figure 10a), except for an apparent difference in intensity between the textured g, m and smooth a classes. Contrastingly, however, the artificial RA-I tactile images do not show the same degree of consistency between adjacent images (figure 10b), indicating there may be information about texture encoded in the time-sequence of these tactile images.

### 3.2.2. Texture classification

Texture classification with the artificial RA-I afferents is now considered by showing the corresponding confusion matrices using spatial (SE) and spatio-temporal encoding (STE) models in figure 11a,b, respectively. Accuracies averaged across all hold-out speeds are reported in table 1.

Both the spatial and spatio-temporal encoding models (RA-SE, RA-STE) tended to perform better than the corresponding SA-SE models for the SA-I afferents, as seen from confusion matrices of RA-I models (figure 11a,b) compared to SA-I models (figure 9a,b). However, the average accuracy of spatial encoding RA-SE models across held-out speeds (53%) was only slightly higher than that of the SA-SE models (table 1). This difference is owing partly to the poorer accuracy of the RA-SE-10 model (28%) than the SA-SE-10 model (40%). Artificial RA-I afferent performance also exhibited improved performance for higher hold-out speeds, in a similar trend to that seen for artificial SA-I performance.

Overall, the spatio-temporal models (RA-STE) performed best with a mean accuracy of 70% (table 1) and the largest individual model accuracies (RA-STE-5; figure 11b). This suggests that artificial RA-I afferents can encode texture in the spatio-temporal modulation of their responses; i.e. there is benefit to extracting spatio-temporal over purely spatial features. This result contrasts with SA-I artificial afferents, a result predicted

---

**Table 1.** Comparative accuracies of each encoding mechanism, averaged across hold-out speeds (for comparison, the random chance over 13 textures is about 8%).

| model   | afferent | encoding      | accuracy |
|---------|----------|---------------|----------|
| SA-SE   | SA-I     | spatial       | 40%      |
| RA-SE   | RA-I     | spatial       | 53%      |
| RA-STE  | RA-I     | spatio-temporal | 46%      |
| vib-TE (real data) | RA-II/vibrotactile | temporal | 50%      |
| vib-TE (augmented) | RA-II/vibrotactile | temporal | 90%      |

---

Figure 8. Tactile images of artificial SA-I responses according to textured stimulus and speed (denoted by the letter and number; see figure 3). (a) Examples from speed 50 mm s\(^{-1}\) on the smooth a, middle-coarseness g and coarsest stimuli m. (b) Examples from the middle-coarseness g stimulus at the minimum 10 mm s\(^{-1}\), middle 50 mm s\(^{-1}\) and maximum 100 mm s\(^{-1}\) speeds. For each case, 10 frames spaced every 10 time samples \((t_0, t_0, \ldots, t_{10})\) are shown to indicate the response variance. Note that the texture pattern is not visible in the tactile image, even for the coarsest stimulus m. We attribute this effect to the stimulus motion and the relaxation time of the skin, rather than the image resolution.
from inspection of the tactile images (figure 10b versus 8b), where more variation was seen within a single dynamic sample (row) of artificial RA-I responses compared to artificial SA-I responses. That artificial RA-I afferents were generally better predictors of texture than SA-I afferents when using a held out speed, suggests that artificial RA-I afferents offer better generalization to previously unobserved speeds. Equivalently, SA-I models tend to over-fit, responding to spatial structures that vary with speed rather than speed-invariant features. This hypothesis is supported by comparing figures 8b and 10b for the SA-I and RA-I responses: within a given speed there appears to be little coherence in terms of spatial structure for RA-I firing, unlike with SA-I tactile images. Thus, RA-I data makes the model less likely to learn undesirable features; i.e. the data are naturally regularized.

3.3. Artificial RA-II afferents

3.3.1. Visual inspection of data

Examples of vibrotactile channel samples are generated from the harmonic structure of induced vibrations; i.e. FFT shape (examples in figure 12). Visual inspection suggests that the chosen resolution (200 features at 1 Hz resolution from 1 s of time-series data) provides sufficient detail to visually distinguish textures based on the shape of their FFTs up to 200 Hz, above which there was very little spectral power in all cases. Consequently, there is a 1 s processing delay for this channel, compared with just 0.1 s used for the artificial SA-I and RA-I channels.

The hypothesis that the harmonic structure of induced vibrations is a viable cue for texture discrimination (§2.4.3) is supported by the FFTs shown in figure 12: there is a consistent shape within each texture, with speed only affecting the scaling rather than structure. For example, at higher speeds the feature vectors are ‘stretched-out’ with peak frequencies increased and at lower speeds the feature vectors are ‘compressed’ with peak frequencies reduced. This structure led us to adopt an augmentation procedure to generate data samples for training, by ‘stretching’ and ‘compressing’ the original collected data accordingly. This was visualized with original and augmented vibration samples collected on stimulus i (figure 7).

3.3.2. Texture classification

Texture classification with the vibrotactile afferents is now considered by showing the corresponding confusion matrices using a temporal encoding model in figure 13 with either

![Figure 9. Confusion matrices for texture classification with artificial SA-I tactile afferents. (a) Spatial encoding model (SA-SE-v) tested on the 10 hold-out speeds \(v\) from 10 to 100 mm s\(^{-1}\). (b) Spatio-temporal encoding model (SA-STE-v) tested on the same speeds.](image-url)
We attribute this to two effects: first, fewer data can be analyzed at the lower speeds because no augmentation over speeds was used for the SA-I and RA-I models (and it is non-obvious how such augmentation could be done). Moreover, there are other differences such as window size and sampling rate that differ between the SA-I/RA-I and RA-II modalities that also make a direct comparison less meaningful. The appropriate interpretation is that the augmented temporal encoding for the vibrotactile modality and spatio-temporal encodings for the RA-I modality both give good performance compared with other alternatives for those modalities.

4. Discussion

The artificial SA-I and RA-I afferents used in this study are shown in a related paper to offer viable models for their natural counterparts [19]. Here we extended the perceptual test space to dynamic stimulation and enabled an additional temporal encoding dimension. In addition, we endowed a soft biomimetic optical tactile sensor (the TacTip) with a novel and complementary modality for that sensor type: the vibrotactile channel, which draws inspiration from RA-II afferents in human touch.

4.1. Frictional cues enable salient spatial codes for texture roughness

Artificial SA-I and RA-I tactile images were dominated by global shearing as a consequence of frictional forces between the dynamic stimulus and the TacTip’s skin, rather than informative about texture than tactile images, however, because no augmentation over speeds was used for the SA-I and RA-I models and it is non-obvious how such augmentation could be done. Moreover, there are other differences such as window size and sampling rate that differ between the SA-I/RA-I and RA-II modalities that also make a direct comparison less meaningful. The appropriate interpretation is that the augmented temporal encoding for the vibrotactile modality and spatio-temporal encodings for the RA-I modality both give good performance compared with other alternatives for those modalities.
Figure 11. Confusion matrices for the accuracy of texture classification with artificial RA-I tactile afferents. (a) Spatial encoding model (RA-SE-v) tested on the 10 hold-out speeds $v$ from 10 to 100 mm s$^{-1}$. (b) Spatio-temporal encoding model (RA-STE-v) tested on the same speeds.

Figure 12. Confusion matrices of dynamic texture prediction using leave-one-speed-out cross validation with models trained on augmented vibration channel data. Each matrix refers to a different model trained and tested with the held-out speed on its individual title.
spatial aspects seen in related work on static stimuli [19]. In particular, the texture pattern is not visible in the tactile image, which we attribute to the stimulus motion and relaxation time of the skin rather than the image resolution. Very light contacts with indentation less than the texture height (0.5 mm) may have enabled greater use of local spatial features; however, our tactile sensing surface bulges slightly, so this would be difficult to test with the current design. Even so, when using only a spatial code, artificial SA-I afferents performed much better than chance (figure 9a) and artificial RA-I afferents provided accurate predictions of texture (figure 11a). It is interesting that biological RA-I afferents also have the highest bandwidth for textural information, in being informative about the broadest range of textures, which is in accord with our results.

These results highlight a distinct difference between the present study, where sliding was used, and others using spatial encoding for texture classification in robotics that employed static or pressing stimulation [34–37]. Although there are parallels in terms of the encoding mechanism, the tactile cue is distinct because of the difference in stimulation technique. Here, we have realized a frictional cue, visible as global shear in the tactile images, whereas static or pressing is likely to encode spatial cues only. Although it has been shown that humans may use spatial cues for coarse texture perception, e.g. for subjective roughness judgements [1,2,38] and discrimination [38], studies have shown that human surface discrimination may employ a ‘sticky–slippery’ dimension which relies on frictional cues [39]. This subjective sticky–slippery property has previously been used in robot texture discrimination for successful classification of real-world textures [8], although this was encoded in a frictional force measurement rather than spatially within tactile images.

4.2. Artificial RA-I afferents benefit from temporal encoding

Artificial RA-I afferents exhibited greater temporal variation than artificial SA-I afferents; for example, the difference in tactile images between adjacent frames was more significant (figures 10b versus 8b). In human studies, it is understood that RA-I afferents are particularly responsive to vibrations (5–40 Hz) [40], which are usually low amplitude compared with deformations associated with static touch. The

Figure 13. Confusion matrices of dynamic texture prediction using leave-one-speed-out cross validation with models trained on (a) non-augmented and (b) augmented vibration channel data. Each matrix refers to a different model trained and tested with the held-out speed on its individual title.
After more, the frequency composition of the RA-I response and peripheral fibres respond more robustly than SA-I afferents availability has so far provided little improvement [43].

Neurophysiological experiments have shown that RA-I peripheral fibres respond more robustly than SA-I afferents to dynamic stimulation with a range of natural textures. Furthermore, the frequency composition of the RA-I response reflected that of oscillations in skin induced by scanning [41], a finding which has informed many to believe that natural RA-I afferents are used for mediating information about texture through a temporal representation of spike patterns [41]. Here we demonstrated that by simply taking the temporal derivative across pairs of tactile images to construct artificial RA-I activity, temporal codes can be used more effectively than on single tactile images.

While other studies have employed spatio-temporal encoding for robotic texture classification [11,13,42], to our knowledge ours is the first where the velocity component of raw tactile features have been used. Our finding that this encoding gave an effective discriminator of texture may have implications for these other studies, in particular to improve the effectiveness of temporal coding where its availability has so far provided little improvement [43].

4.3. Harmonic structure of vibrotactile channel data provides a speed-invariant code for texture classification

It is widely believed that the natural Pacinian (RA-II) system is the primary tactile channel for mediating information about fine surface texture [41,44,45]. We demonstrated highly accurate texture classification using a vibrotactile channel that is inspired by the natural RA-II channel (figure 13b). For a given stimulus, the harmonic structure of induced vibrations (shape of the frequency spectra scaled by the scanning speed) is preserved across different scanning speeds; therefore, the harmonic structure of induced vibrations provides a speed-invariant code for texture discrimination.

This finding about the preservation of harmonic structure is interesting because it aligns with a compelling theory to explain speed-invariance of texture perception in humans. In particular, it has been proposed that cortical computations are capable of extracting harmonic structure from neuron firing [30,46], an idea that was inspired by the timbre-invariance observed in auditory perception [21]. Using the speed-invariance of harmonic structure, here we considered artificial decoding models trained with augmented datasets where the held-out speed was simulated by stretching or compressing training data in the frequency domain. This augmentation method provided both a novel data generation procedure for training predictive models of texture class and resulted in an effective generalization of those models across the speed of interaction.

Features derived from the frequency spectra of contact-induced vibrations have been used in robotics for texture classification experiments [7–9]. However, our approach of decoding information from temporal artificial tactile data by using convolution neural networks with filters applied in the frequency domain that extract features from the shape of the frequency spectra, appears to be novel. Furthermore, speed-invariance of texture perception has been examined in robotics in only one other study [12], where stimulation speed was applied as an additional feature to a support vector machine classifier. By using a method based on a theory of speed-invariant texture perception in nature, our approach does not require any specific speed information. One should also note that there are important differences from [12] in the dynamics of the contact, in that they used a rigid tool running over a hard texture, whereas we used a soft fingertip interacting with a hard texture.

4.4. Implications for robot touch

Using natural touch as inspiration, in this paper we have identified several principles that are beneficial for artificial texture perception, arising from a consideration of the tactile cues from dynamic stimulation, transduction methods and encoding schemes. These techniques include the use of frictional cues for dynamic texture perception, the extraction of artificial RA-I tactile features from image differences, harmonic structure in a vibrotactile channel and speed-invariance as a data augmentation procedure. These methods will probably extend to other tactile sensors that provide rich spatio-temporal data, have a vibrotactile channel and can measure shear.

The present work opens up several areas for future exploration. The artificial SA-I, RA-I and vibrotactile/RA-II channels were considered in isolation, but how would they combine to give effective texture recognition? It can be straightforward to combine feature sets within deep neural networks, but there are subtleties in how multi-modal representations are learned. The stimuli used here varied in a single perceptual dimension, roughness, but how would the results extend to natural textures? The use of artificial stimuli from human psychophysics studies of other perceptual dimensions is a way forward to maintain the relationship with human texture perception; however, methods not directly connected with human touch may be valuable too, depending on the application. Lastly, a multi-modal sense of touch involving slow-adapting, rapid-adapting and vibrotactile channels is valuable not just for texture perception, but also for other applications of artificial tactile sensing. Dexterous manipulation is one key application area, as it is known from human studies that the Pacinian RA-II system is an integral component of how we pick up, handle and manipulate objects.

Data accessibility. All code to process the data, test/train the deep learning models and generate result figures is available at https://github.com/niplepora/afferents-tactile-textures-jrsi2022. Data are available at the University of Bristol data repository, data.bris, at https://doi.org/10.5523/bris.3e1750jw0ck25sxcpe69yj2.

Authors’ contributions. N.P.: conceptualization, investigation, writing—original draft; N.F.L.: conceptualization, funding acquisition, resources, supervision, writing—review and editing.

All authors gave final approval for publication and agreed to be held accountable for the work performed therein.

Competing interests. We declare we have no competing interests.

Funding. This research was funded by a Leverhulme Research Leadership Award on ‘A biomimetic forebrain for robot touch’ (RL-2016-39) and an EPSRC DTP PhD Scholarship for N.P.

Acknowledgements. We thank the anonymous reviewers for their helpful and constructive advice. We also thank Stephen Redmond and Chris Kent for examining the PhD viva of N.P. upon which this paper is based.
temporal codes mediate the tactile perception of natural textures. Proc. Natl Acad. Sci. USA 110, 17 107–17 112. (doi:10.1073/pnas.1305509110)

42. Oddo CM, Controzzi M, Beccai L, Cipriani C, Carrozza MC. 2011 Roughness encoding for discrimination of surfaces in artificial active-touch. IEEE Trans. Rob. 27, 522–533. (doi:10.1109/TRO.2011.2116930)

43. Yuan W, Mo Y, Wang S, Adelson E. 2018 Active clothing material perception using tactile sensing and deep learning. In 2018 IEEE Int. Conf. on Robotics and Automation (ICRA), May 2018, Brisbane, Australia, pp. 2842–2849. New York, NY: IEEE.

44. Hollins M, Washburn S, Bensmaïa SJ. 2001 Vibrotactile adaptation impairs discrimination of fine, but not coarse, textures. Somatosens. Motor Res. 18, 253–262. (doi:10.1080/01421590120089640)

45. Bensmaïa S, Hollins M. 2005 Pacinian representations of fine surface texture. Percept. Psychophys. 67, 842–854. (doi:10.3758/BF03195337)

46. Manfredi L, Saal H, Brown K, Zielinski M, Dammann J, Polashock V, Bensmaïa S. 2014 Natural scenes in tactile texture. J. Neurophysiol. 111, 1792–1802. (doi:10.1152/jn.00680.2013)