Active Exploration for Tactile Texture Perception

Aktive Exploration zur Taktilen Texturwahrnehmung
Studiengang Cognitive Science
Bachelor thesis by Alina Böhm
Date of submission: March 14, 2023

1. Review: Tim Schneider, M.Sc.
2. Review: Dr.-Ing. Boris Belousov
3. Review: Alap Kshirsagar, Ph.D.
4. Review: Prof. Jan Peters, Ph.D.
5. Review: Prof. Constantin Rothkopf, Ph.D.

Darmstadt

TECHNISCHE
UNIVERSITÄT
DARMSTADT

Erklärung zur Abschlussarbeit gemäß § 22 Abs. 7 und § 23 Abs. 7 APB der TU Darmstadt

Hiermit versichere ich, Alina Böhm, die vorliegende Bachelorarbeit ohne Hilfe Dritter und nur mit den angegebenen Quellen und Hilfsmitteln angefertigt zu haben. Alle Stellen, die Quellen entnommen wurden, sind als solche kenntlich gemacht worden. Diese Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen.

Mir ist bekannt, dass im Fall eines Plagiats (§ 38 Abs. 2 APB) ein Täuschungsversuch vorliegt, der dazu führt, dass die Arbeit mit 5,0 bewertet und damit ein Prüfungsversuch verbraucht wird. Abschlussarbeiten dürfen nur einmal wiederholt werden.

Bei der abgegebenen Thesis stimmen die schriftliche und die zur Archivierung eingereichte elektronische Fassung gemäß § 23 Abs. 7 APB überein.

Bei einer Thesis des Fachbereichs Architektur entspricht die eingereichte elektronische Fassung dem vorgestellten Modell und den vorgelegten Plänen.

Darmstadt, 14. März 2023

A.Böhm

Abstract

Tactile perception is an essential modality that humans have, but robots still lack. Importantly, when interacting with a novel object, humans can quickly identify its properties within a few touches [1]. A key research question is how we can make robots behave in a similar way. In this thesis, we develop an information-theoretic method for active exploration of textures and evaluate it using the vision-based tactile sensor GelSight Mini on different types of fabrics. Using sensor images, we can view the task of recognising textures as an image classification task. We provide ablation studies of the state-of-the-art Convolutional Neural Network (CNN) architecture Inception v3 [2], showing that different model variants can classify fabrics with high accuracy and that, in general, Neural Networks (NNs) can successfully process tactile information. We then propose a novel algorithm for active exploration of fabrics that leverages uncertainty in the predictions of NNs. Within this algorithm, we show that different exploration strategies can potentially affect the performance of our classifier, but need to be further examined for tasks that are more difficult for the robot to solve. We also investigate similarities with human tactile perception using two different techniques. Firstly, we design an experiment in which human participants are asked to discriminate fabrics based on tactile information, using the same trials as the robot. We then compare how the fabrics are explored and how accurately they can be identified. This technique shows that the robot can distinguish fabrics more accurately than humans, even without exploration. Secondly, we use the visualisation of saliency methods on NNs. The resulting explanations do not exhibit clear human-interpretable patterns that would support the understanding of the depicted textures.

Zusammenfassung

Die taktile Wahrnehmung ist eine wichtige Eigenschaft, über die Menschen verfügen, die Robotern jedoch fehlt. Wenn Menschen mit einem neuen Objekt interagieren, können sie dessen Eigenschaften innerhalb weniger Berührungen schnell erkennen [1]. Eine wichtige Forschungsfrage ist, wie wir Roboter dazu bringen können, sich auf ähnliche Weise zu verhalten. In dieser Arbeit entwickeln wir eine informationstheoretische Methode zur aktiven Exploration von Texturen und evaluieren sie mit Hilfe des visuotaktilen Sensors GelSight Mini auf verschiedenen Arten von Stoffen. Durch die Verwendung von Sensorbildern können wir die Aufgabe, Texturen zu erkennen, als eine Bildklassifizierungsaufgabe betrachten. Wir stellen Ablationsstudien der State-of-the-Art CNN-Architektur Inception v3 [2] zur Verfügung, die zeigen, dass verschiedene Modellvarianten Stoffe mit hoher Genauigkeit klassifizieren können und, dass neuronale Netze im Allgemeinen erfolgreich taktile Informationen verarbeiten können. Wir erstellen anschließend einen neuen Algorithmus für die aktive Exploration von Stoffen. Mit diesem Algorithmus zeigen wir, dass unterschiedliche verwendete Erkundungsstrategien potenziell die Leistung unseres Algorithmus beeinflussen können, aber noch in für den Roboter schwierigeren Aufgaben stärker geprüft werden müssen. Zusätzlich untersuchen wir Gemeinsamkeiten zum menschlichen Tastsinn mit zwei verschiedenen Ansätzen. Als erstes erstellen wir ein Experiment, in dem menschliche Teilnehmende Stoffe auf Basis von taktilen Informationen in den gleichen Versuchen wie ein Roboter unterscheiden müssen. Damit vergleichen wir, wie die Stoffe erkundet werden und wie genau sie erkannt werden können. Dieser Ansatz zeigt uns, dass der Roboter Stoffe sogar ohne Erkunden besser unterscheiden kann als Menschen. Der zweite Vergleichsansatz ist die Visualisierung von sogenannten Saliency Methoden von neuronalen Netzen. Die resultierenden Erklärungen zeigen keine klaren, für Menschen interpretierbaren Muster, die das Verständnis der dargestellten Stoffe unterstützen würden.

Contents

Figures and Tables

List of Figures

List of Tables

Abbreviations, Symbols and Operators

List of Abbreviations

Notation	Description
CNN	Convolutional Neural Network
EP	Exploratory Procedure
MLBP	Multi-Local Binary Pattern
NN	Neural Network
RNN	Recurrent Neural Network
UQ	Uncertainty Quantification
XAI	Explainable Artificial Intelligence

1. Introduction and Motivation

In the rapidly evolving field of robotics, new challenges are being solved every day. An important goal is to give robots the ability to not only solve specific tasks, but to do so regardless of changing conditions. In order to adapt to such changes in their environment, robots need to be able to perceive their surroundings. We want to take a step in this direction by implementing an active exploration algorithm. With this algorithm, we let a robotic arm actively explore its environment consisting of different textures. Our goal is to enable a robotic agent to learn to recognise these textures by autonomously deciding how often to touch them. Possible applications of our research could be the in-hand material classification for adjusting the handling of different objects according to their category or tasks related to recognising pieces of clothing.

We let the robotic arm perceive textures by translating tactile sensations into features of images. We can make this domain transfer by using vision-based tactile sensors that are able to represent surface textures in images. Using these images, we consider texture recognition as an image classification task. This task allows us to verify how well the agent understands that touches in a different position or orientation or on a different object, still belong to the same texture.

In addition, we investigate whether there are parallels between our results on tactile perception in robots and humans. In both cases, touching an object with a finger or a sensor reveals only partial and local information. We compare the mechanisms behind combining the pieces of information from exploring the textures to get a good "mental picture" in both humans and robots. We let human participants solve a texture recognition task and compare the resulting exploration strategies with those obtained by running our active exploration method on the robot. At the same time, we exploit techniques from the Explainable Artificial Intelligence (XAI) field to visualise the decision process of our classifier. The resulting features give further insight into how an agent perceives textures and whether its representation of these textures helps humans to understand both the agent's decision process and the textures themselves better.

This thesis is structured as follows. In Chapter 2 we introduce the foundations of this thesis, including concepts and tools around classifying images with few data points, and estimating the uncertainty and understanding the decision process of such a classifier. In Chapter 3 we summarise related research in the field of tactile sensing. In Chapters 4 and 5, we explain our methodology and present our results. Finally, we draw conclusions, discuss our research and propose future research directions in Chapters 6 and 7.

2. Foundations

In this section, we lay the foundation for the methods and techniques we use in this thesis, including image and video classification using NNs and transfer learning, as well as uncertainty quantification and sampling. We also introduce saliency methods from the field of XAI.

2.1. Image Classification

In this thesis, textures are explored using images from a vision-based tactile sensor. We can therefore use an image classification network to label these textures.

2.1.1. Inception v3

The architecture of the Inception v3 network is visualised in Figure 2.1. In the figure, the architecture is simplified by grouping blocks of layers that appear multiple times. The first block is InceptionA, appearing three times. The second block, appearing four times, is InceptionB. Finally, the last block, appearing two times is InceptionC. It is also worth mentioning that the Convolutional Layers are all followed by Batch Normalization Layers. While the exact reason why is still under discussion [3, 4], it is clear that these types of layers can speed up the learning process of NNs significantly.

2.1.2. Transfer Learning

The intuition behind transfer learning is that trained image classification networks are good domain-independent feature extractors [5]. Therefore, it is possible to use pretrained networks and fine-tune their weights instead of training from scratch. Only the last fully

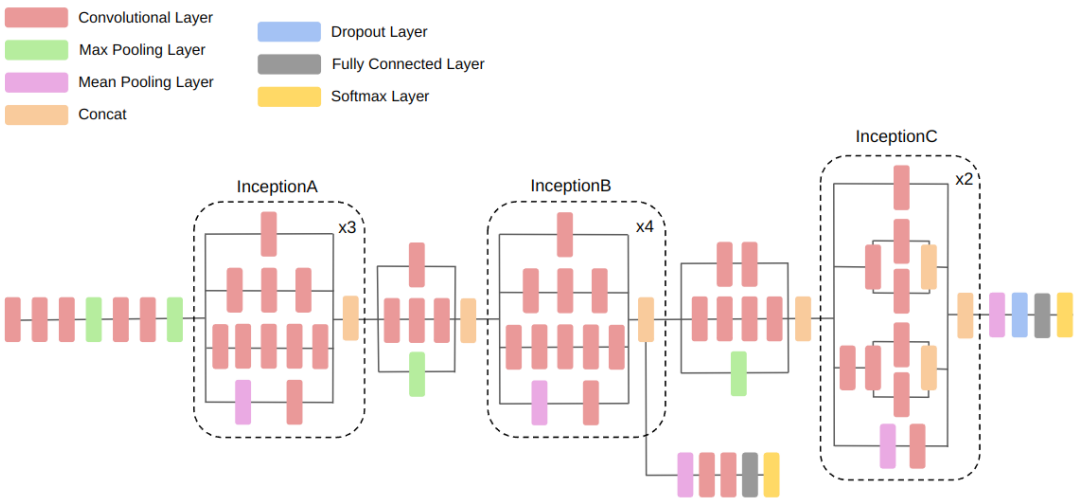


Figure 2.1.: The architecture of Inception v3.

connected layer that predicts a label for these features is reset. Because of this ability to flexibly adjust the last layer, we can use the CNN Inception v3 [2] trained on 1000 classes from the ImageNet dataset [6]. Additionally, it is shown that Inception v3 is suitable for transfer learning in [5].

2.2. Uncertainty Quantification and Active Learning

In the field of active learning [7], the idea is that allowing a learning agent to actively choose the data on which to train can lead to improved performance while requiring less data. In this thesis, we make use of the uncertainty sampling framework for data querying. In this framework, instances are queried that the learner is most uncertain about.

In order to use uncertainty sampling, we want to obtain not only the outputs of a network, but also a measure of the certainty of those outputs. The field of UQ is concerned with how to obtain this kind of information. A comprehensive review of UQ methods and their advantages and disadvantages is given in [8].

2.3. Data Augmentation

Since our goal is to distinguish objects based on a few touches, we use data augmentation to get more information from each data point. We use geometric transformations, specifically adding ten random rotations of each image. In other domains, such as digit classification, rotating images could be problematic as it could change their true label [9]. For example, an image of the digit 6 becomes an image of the digit 9 when rotated 180°. With textures, however, the labels are invariant to rotation. In the context of sensor data, image rotation can be compared to pressing the sensor on an object from different angles. Thus, data augmentation by adding random rotations is close to sampling more data and is uncritical for our texture classification task.

2.4. Explainability

In this thesis we use methods from the field of XAI to visualise features of NNs. Rather than merely looking at end results such as predictions of NNs, XAI helps to gain insights into why these predictions were made [10].

We use the saliency methods *Input X Gradient* and *Grad-CAM*, both of which work with gradients of input images. Note that it is common practice to ignore negative gradients, as they are likely to indicate the choice of a different class label than the one actually predicted [11].

2.4.1. Input X Gradient

The *Input X Gradient* method [12], also known as Input Gradient Attribution, is a technique for explaining the predictions of machine learning models. It is based on the idea that the importance of a feature for a given prediction can be approximated by the product of the input value of that feature and the gradient of the model's output with respect to that feature.

More formally, given a machine learning model f(x), a given input x, and an output y=f(x), the *Input X Gradient* method calculates the attribution of each input feature i as the product of x_i and the gradient of y with respect to x_i, mathematically represented as

$$\text{attribution}_\text{i} = x_i \frac{\delta y}{\delta x_i}$$

This method is often used to understand the factors that contribute to a particular prediction and can be applied to a wide range of models, including neural networks. It has been shown to be effective in providing insight into the decision-making process of complex models and can be used as a tool to identify and correct potential biases in the model.

It is also important to note that this method is model-specific and the interpretation of results is highly dependent on the model architecture and dataset used [13].

2.4.2. Grad-CAM

Gradient-weighted Class Activation Mapping (*Grad-CAM*) is a visualization method for understanding the decision-making process of CNNs [11]. *Grad-CAM* provides a way to generate heatmaps that highlight the regions of the input image that are most important for the model's prediction.

Grad-CAM works by calculating the gradient of the model output with respect to the last activation layer for an input image. The gradients are then backpropagated through the model to the last convolutional layer to get the relevance of each feature. The resulting relevance scores are used to weight the activations of each feature in the final activation layer, effectively highlighting the most important features for prediction.

Mathematically, this can be expressed as:

$$\alpha_k^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^k}$$
$$L_{\text{Grad-CAM}}^c = \text{ReLU}(\alpha_k^c A^k)$$

In this definition, 1/Z is a proportionality constant, y^c are the outputs for class c, and A^k is the k-th feature map of the last convolutional layer. α_k^c describes the importance weights extracted by global average pooling over the width and height (i, j). The heatmap $L_{\text{Grad-CAM}}^c$ is the result of applying the ReLU activation function to the linear combination of the importance weights α and the activation maps A [11].

Grad-CAM can be applied to any CNN architecture and is suitable for a wide range of tasks such as image classification, object detection and segmentation. The method has been shown to provide high quality explanations for the model's predictions and has been used in many applications to provide insight into the model's decision-making process.

3. Related Work

In this section, we give an overview of research in the fields of human and robotic tactile perception and active exploration, and therefore provide an understanding of what has already been achieved in relation to this thesis.

3.1. Tactile Sensing in Humans

Understanding the way in which humans explore objects is an issue of high complexity.

In [14], it is demonstrated that humans can distinguish objects accurately based on touch only. The authors infer that it plays a crucial role whether the objects in question are familiar, and create a performance baseline for human tactile sensing. In this thesis, we do not only compare the accuracy of human and robotic tactile sensing but investigate the underlying exploration strategies for recognising textures. One related aim is to find a formalisation of Exploratory Procedures (EPs) - the distinctive hand movements that humans use to learn more about object properties such as shape and texture [15]. In [16], it is shown that the task participants have to solve and the material properties influence the choice of movement. The authors of this paper also successfully infer which material category is being perceived from the EPs used. We pick up this notion but look more deeply into individual movements. Our focus lies on investigating how much time is spent on each revisit of an object and which objects are explored the most when given multiple options.

3.2. Vision-Based Tactile Sensors

In the following, we discuss related research where tactile sensors are used to enable robots to perceive objects in a similar way to humans.

In this thesis, the vision-based sensor GelSight Mini[1] is used. Inside the sensor, there is a camera that records the deformation of an elastomer. This allows the sensor to visually represent surface properties. GelSight sensors provide a high-resolution 2D and 3D understanding of objects.

As shown in [17], the height map produced by the GelSight sensors provides important information that can then be used to extract global and local patterns using a Multi-Local Binary Pattern (MLBP) operator. The use of this operator leads to accurate surface texture recognition in 99% of cases.

An application very close to ours, combining GelSight images with uncertainty quantification, can be found in [18]. Their motivation is to let robots take over the task of sanding surfaces. The challenge is to find a common solution for working both with large objects and small spaces. Whether human or not, the agent solving this task must be able to identify the surface roughness of various different objects, each of which has its own peculiarities. There are two different subtasks, judging whether a surface has less, the same or more roughness than another (relative classification) and directly predicting the roughness (absolute classification). The performance baseline for both variants is the level of accuracy that humans achieve by touching the objects. This is compared to the performance of a NN and humans classifying GelSight images. The network is a Bayesian CNN, which means that it outputs not only the predictions but whole distributions over them, giving a measure of uncertainty. This estimate is used to weight the network's outputs for smaller subsections of an image to predict its label. In the end, the tactile sensor outperforms the human baseline for both touch and vision in both relative and absolute classification.

3.2.1. Active Exploration Using Tactile Sensors

Combining active perception and vision-based tactile sensors is explored in [19]. The objective of this work is to allow a robot to actively perceive clothing material. The setup consists of a robotic arm with a GelSight sensor that can squeeze fabrics. The positions of

[1] `https://www.gelsightmini.com/`

items of clothing are captured by a Kinect sensor. The authors use two different CNNs, one to extract material properties and the other to select points of interest to explore next based on the Kinect scans. With this setup, an accuracy of over 80% is achieved in extracting the properties of materials on which the models are trained. For unseen materials, performance is lower but still above chance in all categories.

4. Methodology

In this section, we describe our methods for conducting experiments, including classifying textures, measuring the uncertainty of such a classifier, active exploration, and analysing tactile sensing in humans.

Figure 4.1.: The 25 different fabrics used in texture classification. Fabrics marked with an X are used in the human experiment and active exploration setting as well.

4.1. Texture Classification

We gather a set of 25 fabrics that have a perceptually similar texture when judged by humans. Utilising such challenging data samples allows us to investigate the limits of the performance of vision-based tactile sensors. An overview of the fabrics we use can be seen in Figure 4.1.

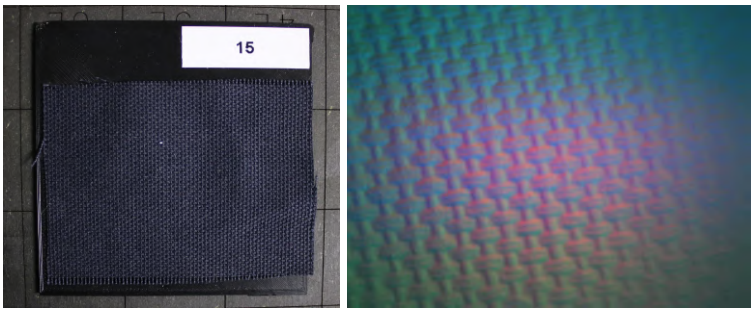


Figure 4.2.: A piece of fabric and the image created by pressing a Gelsight Mini sensor on it.

For translating the surface properties of the fabrics into images we use the GelSight Mini sensor as shown in Figure 4.2. This sensor is mounted on a Franka Panda 7-DOF robotic arm. The whole setup can be seen in Figure 4.3.

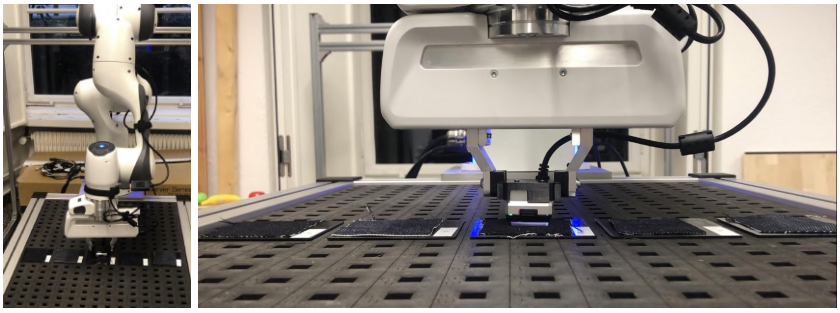


Figure 4.3.: The setup for our experiments consisting of a Franka Panda robotic arm grasping a GelSight Mini sensor. The sensor is pushed down on fabrics to capture texture images.

Our dataset for this ask contains ten training images, 20 for validation, and 20 for testing. The classification experiment is crucial for all the following tasks since it shapes the data

set. When picking the fabrics for the active exploration task we want them to be really similar and thus hard to distinguish. At the same time, they can't be too similar, rendering the task of classifying them correctly unrealistic. For instance, there might be fabrics that have different colours but almost identical textures.

A method for investigating if individual fabrics are hard to distinguish is computing the confusion matrix. The entry (i,j) of the quadratic confusion matrix represents the probability that a sample from class i is given the label j. Ideally, it should be the identity matrix indicating that each object is always assigned the correct label. When looking at the overall accuracy of the model we get an idea of its general performance but the confusion matrix gives much more insights. It tells us not only if the model is struggling but also where the problems are.

As another method to gain insight into how a model processes the texture images, we compute saliency maps using *Input X Gradient* and *Grad-CAM*, which are described in Chapter 2, after training.

4.2. Model Architectures

To get a picture of how different model architectures influence the performance of a classifier on our fabrics, we use three different NNs.

The first two model variants are both Inception v3, but for one of them we use the weights obtained by pretraining the model on the ImageNet dataset ("pretrained network") and for the other variant we use randomly initialised weights ("random weights network"). We hypothesise that pretrained visual features can be useful for texture recognition and use both of these model variants to test this hypothesis.

The third NN consists of all the layers of Inception v3 before the first InceptionA block and after the last InceptionC block ("small inception network"), as can be seen in Figure 4.4. We use this network to verify whether a simpler model is able to solve the texture classification task too.

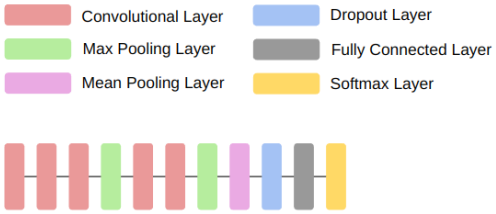


Figure 4.4.: The architecture of our small inception network.

4.3. Uncertainty Representation

4.3.1. Dropout

Our dropout method makes use of dropout layers to estimate the uncertainty of NN outputs. The small inception and Inception v3 architectures after the addition of dropout layers can be seen in Figure 4.5 and Figure 4.6. For dropout layers, there are no weights that change during training and they only have one parameter. This parameter describes the probability with which connections are zeroed out. As these connections are chosen at random, a model with dropout layers produces different outputs for the same input when called multiple times. Because of this randomness, we get to compute an average and variance on the outputs giving us a sense of uncertainty of that NN regarding different objects. After the exploration is done, average pooling is used to get a prediction for the reference object. To each input, we assign the label with the highest average model output.

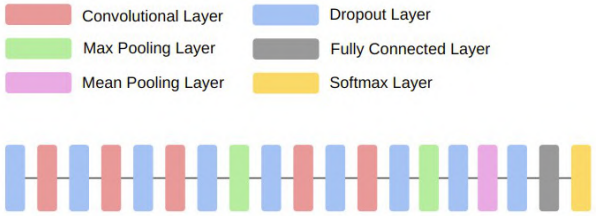


Figure 4.5.: The small inception architecture with added dropout layers for UQ.

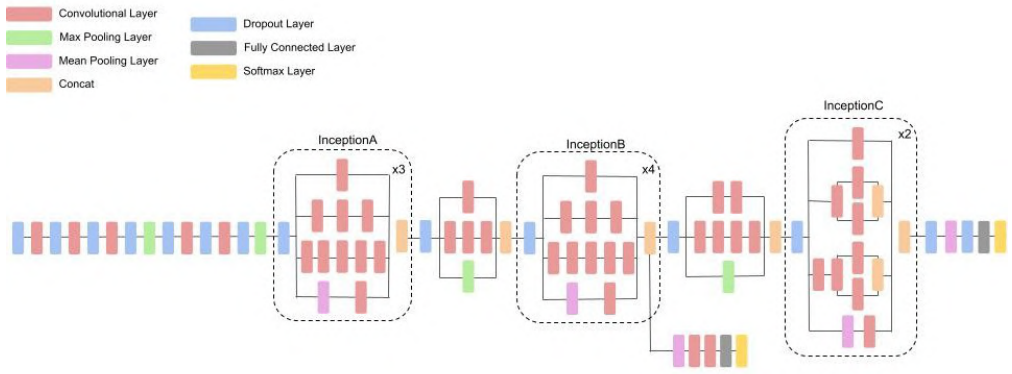


Figure 4.6.: The Inception v3 architecture with added dropout layers for UQ.

4.3.2. Metrics for Uncertainty

Using the dropout method, we get multiple outputs for the same input. To be able to conclude the uncertainty using these outputs there are different metrics. We formulate the output of the NN given the reference object o_{ref} and the model parameters θ as $p(c|o_{ref}, \theta)$. In order to be able to interpret the logits returned by a fully connected network layer as probabilities we apply the Softmax function. It squashes the outputs into the [0, 1] range and ensures $\sum_{i=1}^{n} p(c_i|o_{ref}, \theta) = 1$ for n classes c_i. In the following we are writing p_i instead of $p(c_i|o_{ref}, \theta)$.

The first metric is the expected variance of the classifiers outputs $\mathbb{E}[\text{Var}[p_i]]$. The assumption is that a high discrepancy between the models' outputs suggests that there is a high uncertainty. This way we get an estimated uncertainty for each class c_i.

The second metric is the Entropy $H(p) = -\sum_{i=1}^{n} p_i \log p_i$. As the goal is to minimise the entropy we want to find the class c_i that contributes the most to the term. Thus, we are looking for $\max_i -\{p_i \log p_i\} = \max_i p_i h(p_i)$ with $h(p_i) = -\{\log p_i\}$ denoting the information of class c_i occurring.

4.4. Tactile Sensing

We evaluate the performance of both humans and a robotic system on distinguishing textures using the same setup, which we describe here.

Our setup consists of five platforms with fabrics mounted on top of them. One of those platforms is the reference object and the other four are comparison objects. One of the comparison objects has the same fabric as the reference object and the goal is to find out which one. The agent solving this task, whether human or robotic, can only use tactile information about the texture of the fabrics. In total, we use eight of our 25 fabrics and test for consistent results among subsets of these eight fabrics. In the experiment, one trial consists of finding out which comparison object has the same fabric as the reference object for one subset of fabrics.

4.4.1. Human Tactile Sensing

For our human experiment, we use the data of ten participants between the ages of 23 and 35. One of them is female and the rest are male and all of them are either undergraduate students or doctoral candidates. In order to investigate how humans distinguish fabrics based only on tactile information, the subjects are blindfolded and may only use their fingertips to explore fabrics. We leave the option to choose between using the index finger or middle finger but the participants have to use the finger they choose in the beginning throughout the whole experiment and also use the same finger with both hands. The left finger rests on the reference object while the right finger is used to gather information about the four comparison objects. Additionally, the participants are advised to touch each object at least once before making a prediction. To avoid learning from trials we do not provide any feedback on a participant's performance during the experiment. The setup for a trial can be seen in Figure 4.7.

We use each of the fabrics as the reference object four times resulting in a total of 32 trials. We hypothesise that the positions of the comparison objects influence the predictions of human participants. Therefore, we place the correct fabric once in each of the four possible locations as shown in Tables 4.1 and 4.2.

During the trials, we record the hand movements of the subjects with a camera. With the resulting videos, we analyse the number of revisits per object and the time spent on that object for each revisit. We count each time a participant switches between two objects as a new revisit. We provide small poles next to the objects which are used for locating the texture platforms. After the experiment, we use the predictions made by each participant to compute their confusion matrix of all eight fabrics.

	reference fabric	fabric 1	fabric 2	fabric 3	fabric 4
1	23	4	8	23	7
2	18	18	23	21	17
3	4	18	11	17	4
4	21	21	11	17	7
5	7	18	7	23	11
6	8	21	4	8	11
7	17	8	21	17	23
8	18	21	4	18	11
9	11	4	8	23	11
10	18	17	18	8	21
11	11	8	17	11	7
12	7	4	8	7	23
13	17	11	18	21	17
14	4	23	8	4	21
15	23	23	4	21	17
16	23	21	17	4	23

Table 4.1.: The fabrics used for the first 16 trials.

	reference fabric	fabric 1	fabric 2	fabric 3	fabric 4
17	17	17	18	8	4
18	21	4	21	7	8
19	11	11	18	17	7
20	18	4	7	11	18
21	11	23	11	4	8
22	21	23	17	21	4
23	8	8	21	7	18
24	8	23	8	17	4
25	4	4	17	23	21
26	4	23	4	7	18
27	23	4	23	18	17
28	7	21	8	17	7
29	8	17	23	7	8
30	7	7	11	8	18
31	17	4	17	23	18
32	21	4	7	11	21

Table 4.2.: The fabrics used for the last 16 trials.

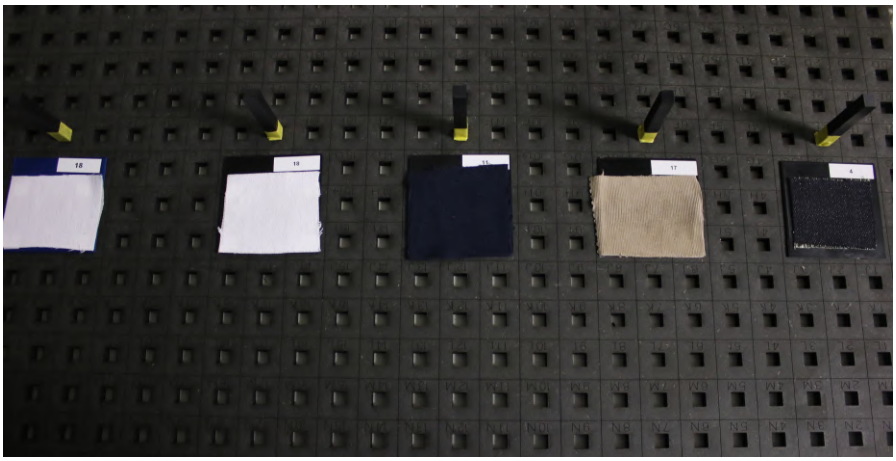


Figure 4.7.: The setup of a trial in the human experiment. The reference object is the one on the left and the correct comparison object which participants need to find is the one in position 1, which means that it is right next to the reference object. The small poles next to the platforms ought to be used for locating the texture platforms since the participants are blindfolded.

4.4.2. Active Exploration

In the active exploration setting, the three models are trained on sensor data collected by a robotic arm equipped with a GelSight Mini sensor.

A trial starts by letting the robot create a baseline for classification by collecting one sensor image for each object. The image corresponding to the reference object is stored as the test object and the other images are training images. On all of these images, we apply data augmentation. We store each image ten times and apply random rotations to the input data and thus simulate touching a fabric from multiple angles only using one image of that fabric. We then train the model on the resulting training data for ten epochs. After that, a new round starts. We evaluate the model's uncertainty by calculating the variance and entropy on 20 different predictions for the test images.

The object which is explored next is dependent on the chosen strategy. For the variance and entropy strategies, the agent chooses the object with the highest variance or entropy, respectively. In the random strategy, any object is chosen according to a uniform distribution. This strategy provides information about whether sampling according to uncertainty provides an advantage in our task. The fourth strategy is called YOTO ("you

only touch once"). With this strategy, the objects are not revisited at all after the baseline is created. This strategy gives us insights into whether the agent even needs additional data to recognise the textures successfully. The classifier is then trained further with all the available data for ten epochs. This process is repeated every round.

For the evaluation of the different strategies and models in general, we run four different trials for 20 rounds. As the dataset can get unbalanced we apply class-weighting in our loss function according to the number of available images. After each round, a confusion matrix is computed for the training images to help gain insight into what the models learn each round. However, the evaluation of these matrices is left for future research.

For direct comparison to human tactile sensing, we only focus on the small inception network. We compute the average amount of touches the participants need for each trial and shorten the trials for the robot to that same number of touches to compare the performance of humans and the robot after the same amount of exploration. We use the final predictions of the model to compute a confusion matrix on the eight fabrics using the four different strategies. For an insight into the effect of exploration on the model's decision process, we visualise the saliency maps using *Grad-CAM* and *Input X Gradient* after each round on a validation dataset. This dataset only consists of one image per class and no random rotations are applied, thus ensuring that we can directly see differences in the saliency maps as they are always applied to the same images.

5. Experiments and Results

In this chapter, we report the results of our experiments. In the first section, we present the results of our basic texture classification task using three different models. After that, we evaluate the experiment with humans and our active exploration algorithm, and compare the resulting accuracies and chosen strategies. Lastly, we show the results of using the saliency methods *Input X Gradient* and *Grad-CAM* on our texture images.

5.1. Texture Classification

In the task of classifying all 25 textures, we see significant differences within the performance of the three models, pretrained Inception v3, Inception v3 with random weights, and the small inception model. As can be seen in Figure 5.1, pretraining helps the model to learn to recognise the textures. This advantage proves that the network has learned to extract domain-independent visual features while training on a different domain of images in ImageNet. It can be seen that the small inception model needs more training time to reach the same performance as the other two models, which is expected since it is missing all the inception blocks. At the same time, the performance on the validation data is close to the random weights model, proving that the small inception network is still able to generalise well.

In Table 5.1 we report the performance of the models on the test data, averaged over five runs with different random seeds, and show that all three models achieve an accuracy of over 94%. From these high levels of performance, we conclude that each of the networks is capable of recognising the fabrics sufficiently well and that none of the fabrics need to be excluded.

	pretrained model	random weights model	small inception model
mean	**96.82**%	94.32%	94.26%
std	1.16%	2.21%	1.58%

Table 5.1.: The accuracy of three different models on the test dataset of 25 classes of fabrics.

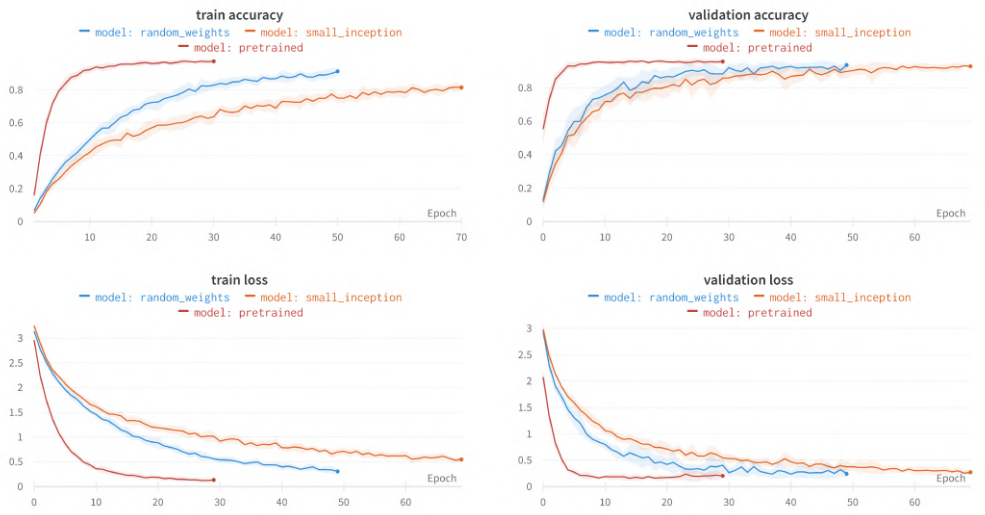


Figure 5.1.: The performance of three different models on the training and validation data of 25 classes of fabrics, averaged over ten runs. Each model is trained until the performance on the validation data converges.

5.2. Human Tactile Sensing

In the human experiment, the average accuracy in discriminating the fabrics is 66.88%, ranging from 53.13% to 75%. In Figure 5.2, we summarise the average performance and how it changes when we look at trials according to the position of the correct comparison object.

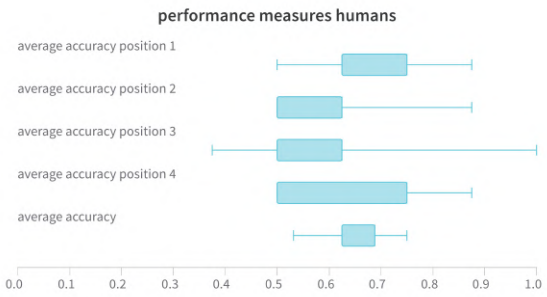


Figure 5.2.: The average accuracy of all participants in general and for each position of the correct comparison object, summarised in a box plot. Each box ranges from the first to the third quartile of the data. The whiskers (horizontal lines) extend from the minimum to the first quartile and from the third quartile to the maximum.

The average number of revisits before making a prediction in each trial is shown in Figure 5.3, ranging from five to nine revisits needed per trial.

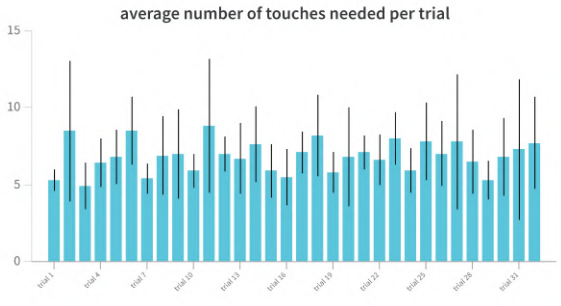


Figure 5.3.: The average number of touches required for people to make a prediction in each trial.

5.3. Active Exploration

First, we examine the performance of the three different models in the active exploration algorithm. For all three models, we collect the results of five subsets of fabrics using the four different strategies and average the performance of each model. In Figure 5.4, it can be seen that the larger Inception v3 models still have an advantage over the small inception model, similar to the texture classification task. However, the differences have become smaller, and the random weights model initially even gives an advantage over the pretrained model. The small inception model can solve the task sufficiently well, while at the same time being the least computationally expensive model variant. Hence, we continue to look at its performance using different strategies.

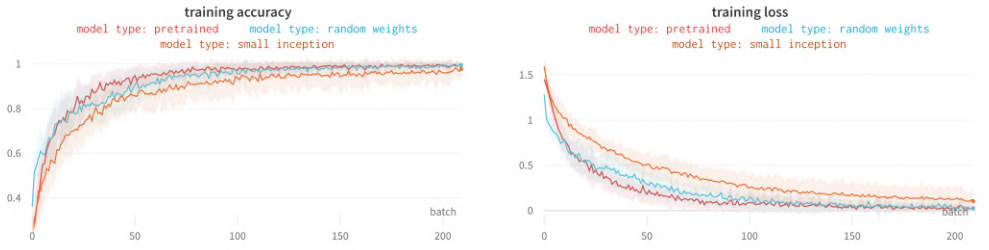


Figure 5.4.: The performance of the three models trained for 20 rounds. The metrics are averaged over five runs with different fabrics for each of the four strategies.

Figure 5.5 shows the influence of the different strategies on the performance of the small inception model. When we run the experiment for 20 rounds, sampling generally offers an advantage, as the YOTO strategy has the lowest performance on the training data. On average, the model performs best using the variance strategy, closely followed by the entropy strategy and random sampling. Still, the standard deviation of the performance shows that the differences are not significant. When it comes to uncertainty, using entropy or variance as a metric yields slightly different results. In both cases the sampling strategies outperform YOTO.

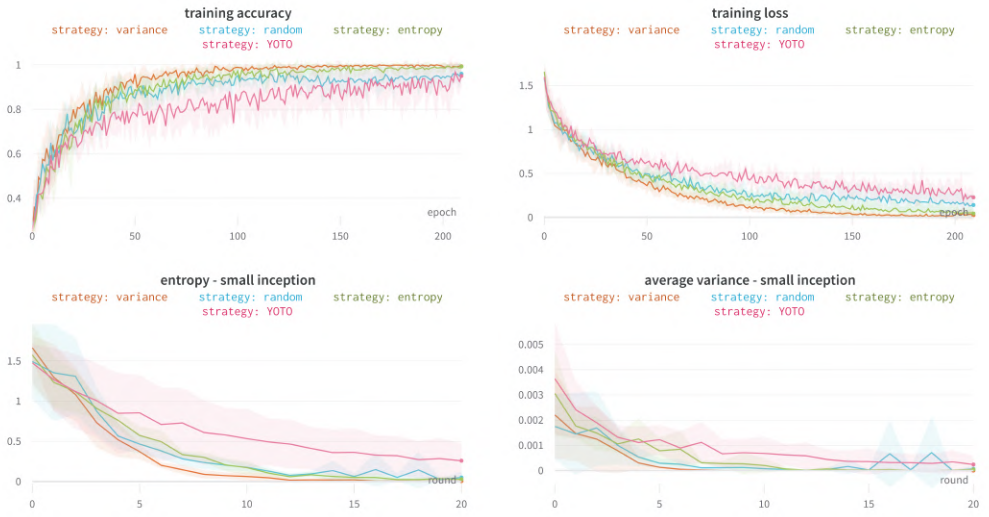


Figure 5.5.: The performance of the small inception network trained for 20 rounds. The metrics are averaged over five runs with different fabrics.

5.4. Comparison

5.4.1. Accuracy

In Table 5.2, we compare how accurate the predictions of the human participants and the robot are using the four different strategies. On average, the robot outperforms the humans by more than 10%, regardless of the strategy. The YOTO strategy also results in lower accuracy than the other strategies where new data is sampled. However, when it comes to the sampling strategies, the differences are very small. We also notice that the standard deviation is high in all cases, which prompts us to take a closer look at the individual trials.

The individual accuracies per trial are shown in Figure 5.6. We can see that the robot reaches 100% accuracy in 12 out of 32 trials, regardless of the chosen strategy. At the same time, humans never reach an average of 100%, i.e. there was no trial in which all participants predicted the correct object. The figure also shows that the best-performing strategy is very trial dependent. In trial 11, no sampling provides better results than random sampling. On the other hand, in trials 12, 15, 16, and 31, random sampling

	humans	variance	entropy	random	YOTO
mean	66.88%	**90.00**%	88.13%	89.38%	80.63%
std	16.93%	15.24%	14.24%	14.35%	22.42%

Table 5.2.: The average accuracy of humans compared to the small inception network using different strategies using the same number of touches.

outperforms all other strategies. In addition, the human participants show higher accuracy than the robot in trial 3 and outperform at least some strategies in eight trials. It is also worth noting the difference between trials 15 and 16. These two trials consisted of the same set of fabrics and the same reference object; only the order is changed. For the active exploration algorithm, the performance does not vary between these two trials since there is no concept of order for our classifier. For humans, however, the accuracy changes, indicating that the same set of fabrics does not always lead to the same prediction.

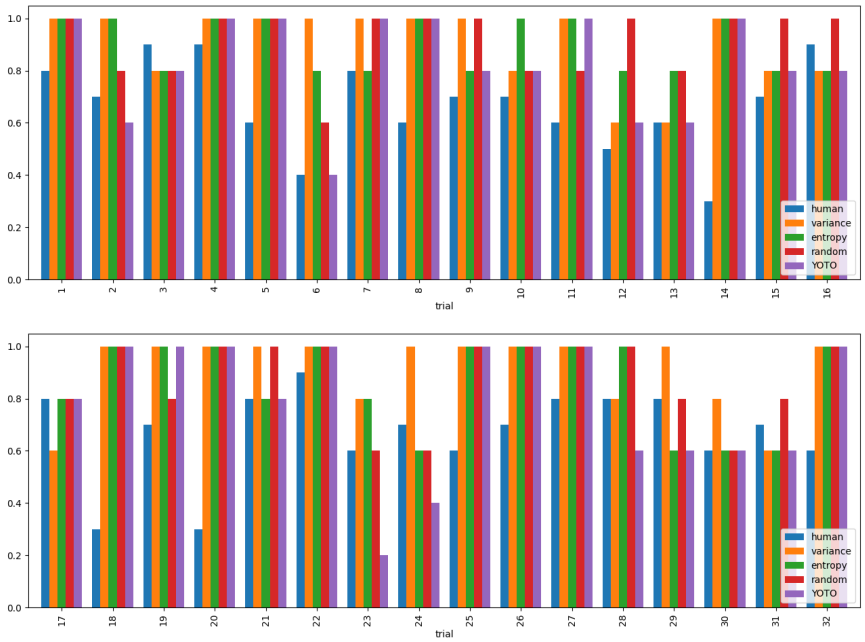
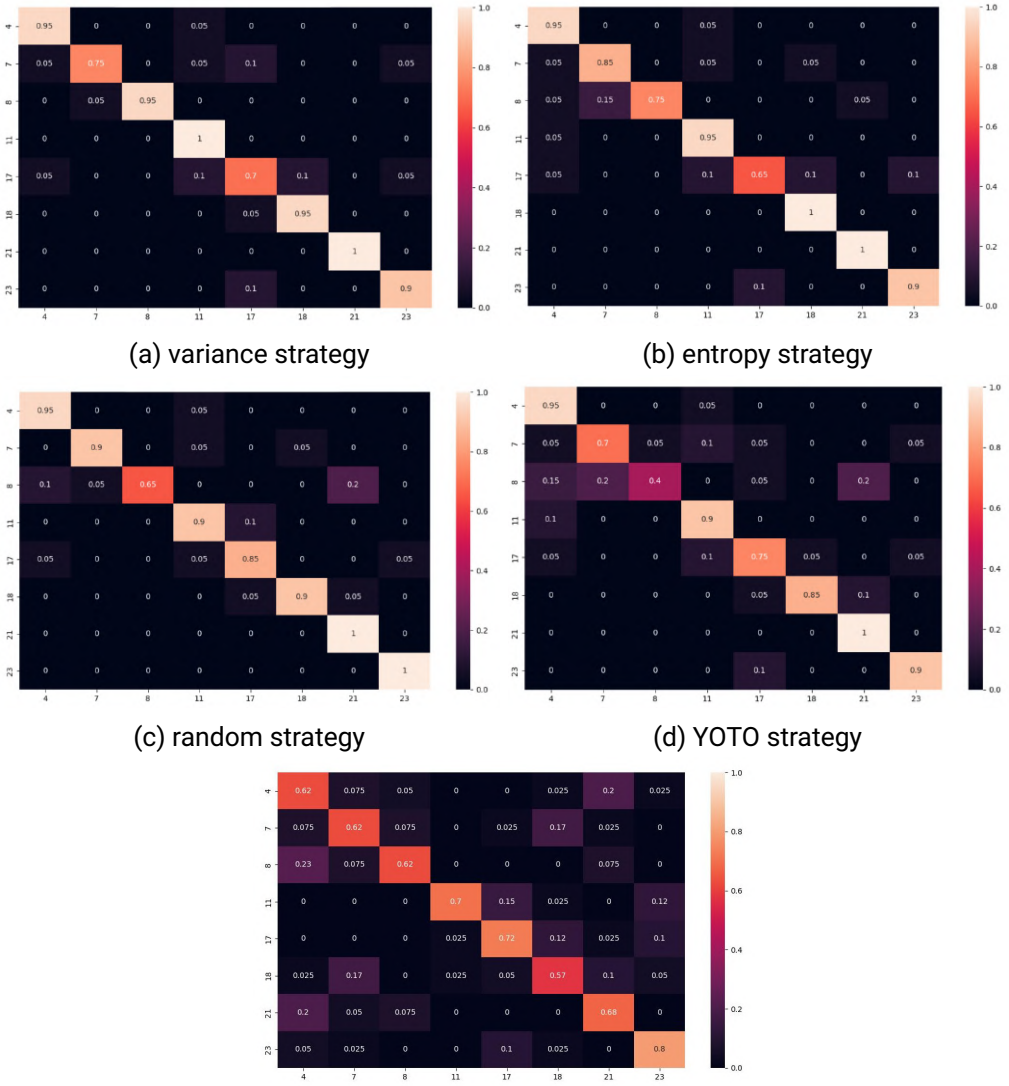


Figure 5.6.: The average accuracy of humans and the robot for each trial.



(a) variance strategy

(b) entropy strategy

(c) random strategy

(d) YOTO strategy

(e) humans

Figure 5.7.: The confusion matrices of the eight fabrics included in the experiment of the robot using different strategies (a - d), and averaged over all participants (e).

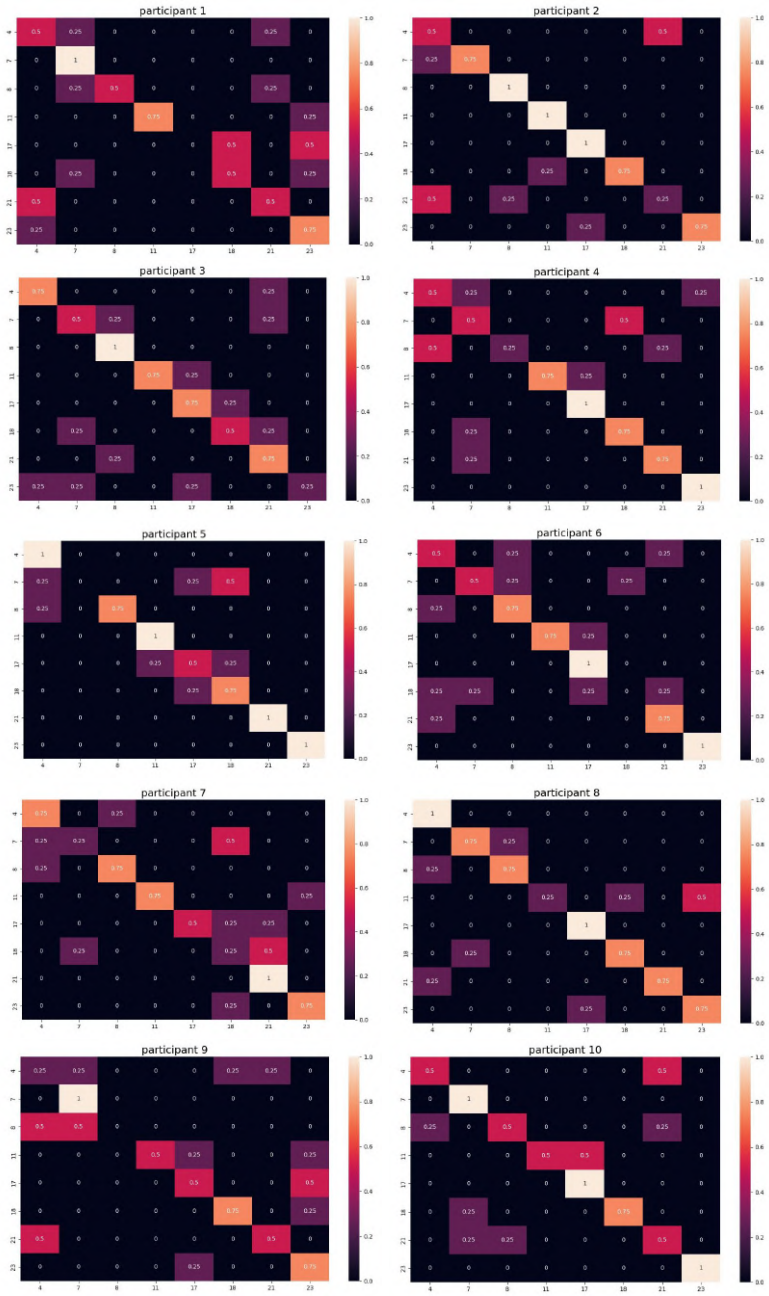


Figure 5.8.: The confusion matrices of each participant individually.

Another method to gain more insight into how humans perceive textures compared to the robot is to examine the confusion matrices of the fabrics. In Figure 5.7, we show the confusion matrices for each strategy and averaged over all participants. For humans, there does not seem to be any fabric that is inherently difficult to recognise for all participants. Looking at the individual confusion matrices in Figure 5.8, there are some cases where participants had 0% accuracy on a fabric that others were able to classify correctly every time, meaning that the fabrics have varying degrees of difficulty, but also that it depends on each participant whether a fabric is hard for them to recognise.

5.4.2. Strategies

If we normalise the time spent on each object per trial, we get a distribution of relative times per fabric. Since these relative times sum to one, we can compute the Jensen-Shannon distance comparing the time spent on the fabrics by humans and the active exploration algorithm using different strategies. The YOTO strategy is excluded from this calculation as there is no exploration after creating a baseline with one touch per object. The relative times for the YOTO strategy are, therefore, uniformly distributed.

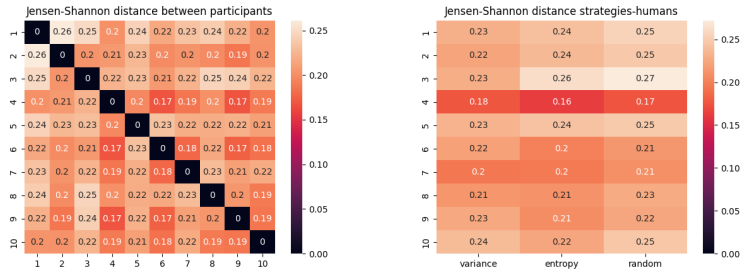


Figure 5.9.: The Jensen-Shannon distance between the time spent on each object comparing participants with each other and with the strategies, computed for each trial and then averaged.

The Jensen-Shannon distance can take values in the range [0, 1], with lower values indicating a greater similarity between two distributions. The Jensen-Shannon distance of a distribution to itself is zero. What we can infer from the results shown in Figure 5.9 is that comparing the robot with different strategies to the participants results in a similar range of values as comparing the participants' exploration to each other. There is one participant that stands out as having a relatively small distance to the robot's three strategies. If we

average the distances of each participant, we get a mean Jensen-Shannon distance of 0.219 for the variance strategy, 0.218 for the entropy strategy, and 0.231 for the random sampling strategy, meaning that the uncertainty-based strategies are on average slightly more similar to human exploration under the Jensen-Shannon distance. On the other hand, we again observe a high variance between trials. There are some trials where the same two participants follow a very similar exploration strategy and others where they choose different approaches with a higher Jensen-Shannon distance. See Figure 5.10 for a detailed view of participant 1's strategy compared to the other participants in each trial. We observe the same problem with the variance between trials when comparing humans and the robot. We also compare the strategy chosen by each participant for trials 15 and 16, which contain the same objects in a different order. The distances shown in Table 5.3 show that participants do not necessarily follow the same strategy in these two trials.

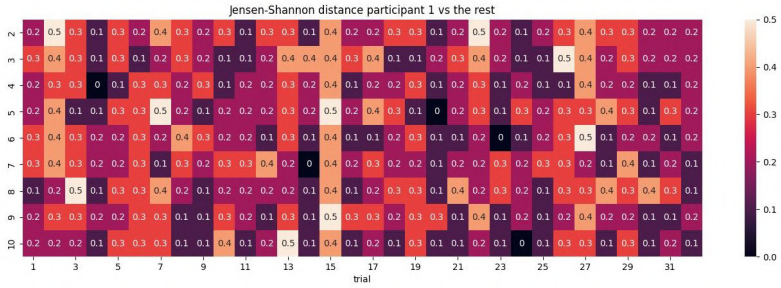


Figure 5.10.: The Jensen-Shannon distance between the time spent on each object comparing participant 1 with the other participants for each individual trial.

1	2	3	4	5	6	7	8	9	10
0.27	0.15	0.45	0.3	0.44	0.23	0.1	0.16	0.1	0.4

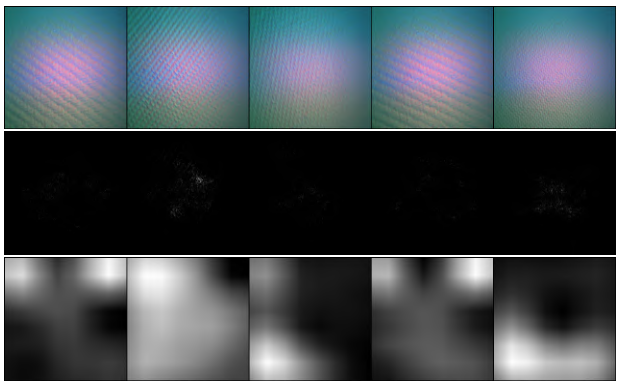
Table 5.3.: The Jensen-Shannon distance between the relative time spent on each object for trials 15 and 16 within the same participant.

Finally, we observe a similar aspect when it comes to the variance strategy of the robot and the human participants. In 63.75% of the trials, the participants touched the object most that they predicted to be the same as the reference object. With the variance strategy we get 56.25% and with the entropy strategy only 32.5% of the trials where the predicted object is touched most.

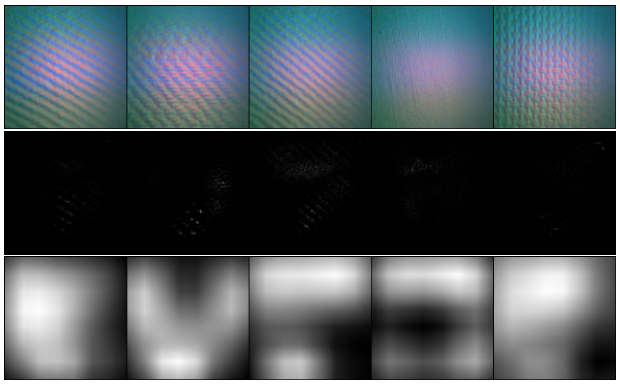
5.5. Explanations

The results of the two saliency methods *Input X Gradient* and *Grad-CAM* for the texture classification task on all 25 fabrics are shown in Figure 5.11. Note how *Grad-CAM* is computed on the last convolutional layer, for which the input size varies between the models, resulting in saliency maps with higher resolution for the small inception model.

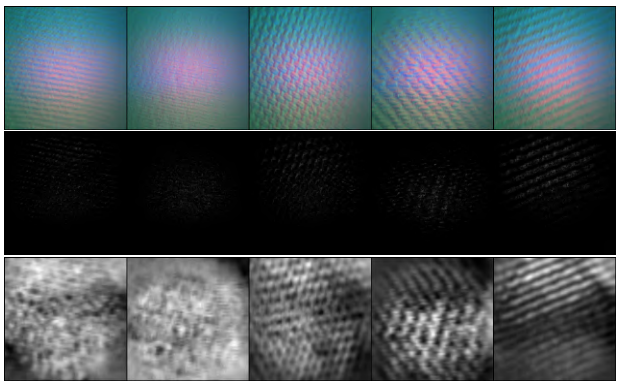
As for the active exploration experiment, Figure 5.12 shows the saliency maps at the beginning and at the end of a trial consisting of five rounds. The changes in the input images compared to the texture classification task result from the need to scale up after adding random rotations. For the *Input X Gradient*, there appear to be more image regions with higher gradients compared to the saliency maps before exploration. The *Grad-CAM* visualisation is very noisy and does not allow us to identify patterns of changes during exploration.



(a) pretrained model



(b) random weights model



(c) small inception model

Figure 5.11.: Input images, *Input X Gradient*, and *Grad-CAM* for the three models from the texture classification task, in that order.

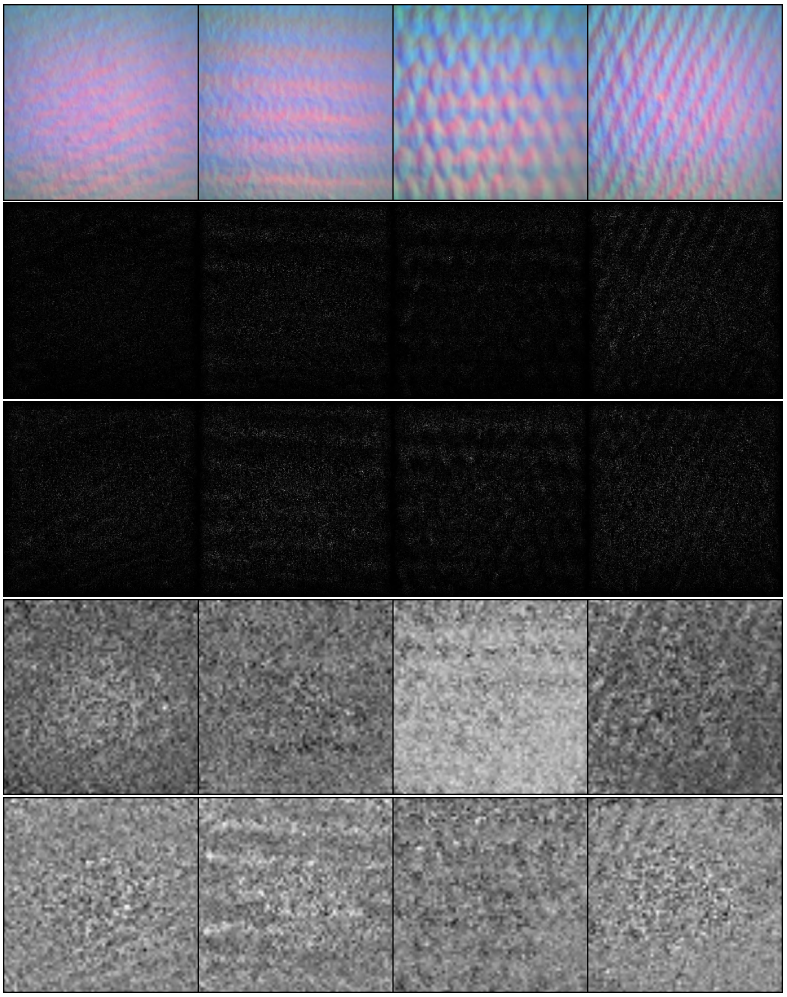


Figure 5.12.: Input images, *Input X Gradient*, and *Grad-CAM* for the small inception model, each shown before (top) and after (bottom) exploration.

6. Conclusion and Discussion

In this thesis, we enable robotic tactile sensing of textures by collecting a dataset of sensor images of 25 different fabrics. We show that three different models can successfully discriminate these fabrics. However, learning to recognise the textures takes less time with the larger model, Inception v3. The advantage of the pretrained Inception v3 confirms that transfer learning is effective in our task. On the other hand, the advantage decreases in the active exploration experiment. One possible conclusion is that dropout layers render pretraining less helpful, as many network connections are zeroed out. In this setting, the small inception model performs only slightly worse than the larger models. When comparing strategies, using the variance of the model outputs as a metric for uncertainty and sampling accordingly provides the best performance on average, but the advantage is very small. Once we reduce the trial length to the average time observed in human participants, the sampling strategy becomes even less important. There are trials where it does not matter whether the robot revisits any objects, and all four strategies, including YOTO, can predict the correct fabric 100% of the time. We conclude that these trials are too easy for our purposes and that the baseline of ten rotated images from one touch already provides a lot of information.

The results visualised in Figure 5.5 show that all strategies lead to a reduction in uncertainty, regardless of whether entropy or variance is used as a metric. Sampling generally provides an improvement, but as with accuracy, there is no significant improvement when using uncertainty-based strategies.

The results of the individual experiments regarding accuracies and strategies allow us to conclude that, for our experiment, there is no clear answer to the question of which strategy performs best and which is most similar to human strategies. This inconclusiveness is also influenced by the fact that our participants do not show a clear pattern when it comes to recognising and revisiting textures.

6.1. Limitations

In this section, we discuss the limitations of our method and highlight that our findings raise new research questions that need to be explored.

6.1.1. Human Variance

Our results suggest that viewing people as a single group is problematic. They do not generally confuse the same fabrics or follow the same strategies. For instance, participant 4 chooses a strategy that is on average more similar to the robot using the entropy strategy than to the exploration of any other human participant.

In addition, there are factors for each participant that are not considered in this thesis. Aspects such as attention or the weight participants give to the conflicting goals of being certain before making a prediction and minimising the number of touches used are not measured, but may play an important role.

6.1.2. Comparability

We acknowledge that the overall comparability of human and robotic tactile perception is limited in some respects.

One fundamental aspect worth discussing is the use of supervised learning. The robotic agent is trained to recognise textures using an image classifier, which falls into the category of supervised learning. In this type of learning, all data points have labels. The agent then learns by predicting the class labels and adjusting the network weights according to a loss calculated with respect to the actual labels. As a result, it is always implicitly clear which comparison object is the same as the reference object. The goal of the classifier is to be able to reproduce these labels, especially the one for the reference object. On the other hand, humans are not asked to assign labels to each object, but only to find the one comparison object that they judge to be the same as the reference object, without having access to any true labels.

Secondly, there are some limiting aspects related to our method of exploration. A strategy often observed in participants is to move their fingers around an object. The robot can only press the sensor straight down on the object, take a picture and then move up again.

Room for other exploratory movements is not included in our algorithm, as this thesis is about creating a baseline for active tactile sensing. In addition, participants are allowed to use two fingers at the same time. During the experiment, many of them move their fingers synchronously. Their two fingers then provide temporal information about the reference and comparison objects that can be directly compared. This procedure also leads to the caveat that the humans and the robot get different information from revisiting an object. For the robot, a revisit is equivalent to taking an image. This image is then augmented to produce ten inputs for our classifier, but there is no temporal relationship between them and they are treated as individual images that could just as well be produced by different revisits of the same object. Another piece of information not available to the robot is the degree of rotation and the position at which it touches a fabric each time. After all, processing a rotated image and knowing how much the sensor was rotated to get that image are two different things. Humans have access to this kind of information because of their proprioception. They can feel the position of their fingers and relate which movements lead to which tactile sensations.

A related aspect is that we can only compare relative results when it comes to the amount of time spent on each object. For the robot there is no real sense of time, each revisit is about taking one image. Therefore, the relative time spent per object can only be calculated as the relative number of revisits per object. For humans, the actual time spent touching an object varies with each revisit. Because of this variation, we calculate the relative time spent on each object using the actual time spent on each revisit. The results can therefore only give us a rough understanding of how similar the emerging strategies are.

Another problem we are aware of is that for humans, trials are not independent. There are many factors that we have not accounted for in this thesis. Participants could learn from each trial and transfer their knowledge to create a mental image of all the fabrics presented in the experiment, thereby improving their performance. The robot has no memory of fabrics that do not appear in the current trial and is initialised with random weights each time. We try to counteract the human learning process to increase the comparability between the settings by not giving any feedback to the participants and by not letting them know the total number of fabrics. Still, we cannot prevent learning completely.

6.1.3. Strategy Selection

Our results indicate that the task of discriminating four fabrics is not very challenging for our selected models compared to humans. Even without any exploration, the accuracy of our algorithm using the YOTO strategy is on average 14% higher than human performance. The results for the different exploration strategies do not indicate a significant advantage of uncertainty sampling over random sampling. Accordingly, it does not seem to matter for this task which object the agent collects new data for. As long as there is new data, the agent can learn and improve. It is also possible that the differences between the strategies become significant only after several training rounds.

6.1.4. Explainability

From our point of view, an explanation is considered good if it helps people to understand the images. In our case, the explanations should highlight aspects of the images that help viewers improve in recognising the texture depicted in them. However, compared to looking at the images without explanations, we do not gain new information by looking at the saliency maps of the input images for recognising the fabrics. Therefore, *Input X Gradient* and *Grad-CAM* are not effective methods for our dataset when applied directly to the input images.

Our conclusion is that our data may be an illustration of the limitations of saliency methods. For example, if we were in the domain of animal images, heat maps can show important aspects of the animal, such as its head or paw. These aspects are essential for humans to classify the animal. In our case, our data does not contain concepts such as body parts, or even foreground and background, but only regular texture patterns. In these patterns, spatial relationships between features would be the most important concepts for texture recognition. These spatial relations cannot be directly visualised by *Input X Gradient* or *Grad-CAM*, and we cannot infer whether the network uses these concepts based on these two explanatory methods.

We see the Fourier transform as a way of visualising the concepts of texture images. On transformed images, saliency methods can reveal the importance of certain frequencies and provide helpful explanations. However, our results are inconclusive and there is a need for further research into concept-based explanations of texture images.

7. Outlook

In this thesis, we implement an active exploration method for a robot equipped with a tactile sensor and present a new dataset of sensor images of different fabrics. In addition, we conduct an experiment with human participants who perform a similar texture recognition task as the robot. With our results, we lay the foundation for possible further research in the field of robotic tactile sensing and active exploration. We find that using a small part of the Inception v3 model architecture results in more accurate texture recognition than humans.

For a comprehensive overview of the capabilities and limitations of different exploration strategies, experiments with more difficult tasks are needed. A task that cannot be solved accurately by the robot without exploration could provide more insights into the effects of different strategies. In our task, we can see that sampling data according to uncertainty can speed up learning, but when we look at just a few training rounds, the effect is barely noticeable. It would also be interesting to take another step towards autonomous perception of the environment. If the robotic arm had the ability to decide not only which object to touch and how often to touch it, but also how to touch it, it would be possible to compare emerging exploratory procedures in robots and humans. An important research question is whether movements provide the same information gain in both types of agents.

Another aspect worth investigating is how our results show a different training process within a model trained on all 25 classes of fabrics versus only four. The main difference, apart from the number of classes to be identified, is the addition of dropout layers. This raises the question of the exact role of dropout in CNNs. We hypothesise that dropout renders the pretraining of a model ineffective, and that this is the reason why the advantage of the pretrained Inception v3 over other networks is smaller in our active exploration setting. To verify this hypothesis, further tests should be performed with networks with and without dropout attempting to solve the same task. The same effect can be seen with saliency methods. We see the potential for explanations that help humans understand how the same NN understands data before and after the addition of dropout.

In general, the relationship between human and robotic tactile sensing is only briefly explored in this thesis. There is a need for more extensive experimentation with human texture discrimination. Possible factors that could influence the outcome of such an experiment, such as age, background and time elapsed within an experiment, are not yet covered by our research. Our results also suggest an influence of the positioning of the objects, which is not sufficiently explored in this work to draw firm conclusions.

As our preliminary results suggest a significant advantage of robotic tactile perception over human tactile perception in the texture classification task, another exciting direction for future research is to investigate whether this advantage also holds for other tactile perception tasks, such as inferring the shape or stiffness of objects.

Bibliography

[1] K. Drewing, A. Lezkan, and S. Ludwig, "Texture discrimination in active touch: Effects of the extension of the exploration and their exploitation," in *2011 IEEE World Haptics Conference*, pp. 215–220, 2011.

[2] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818–2826, 2016.

[3] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*, pp. 448–456, PMLR, 2015.

[4] S. Santurkar, D. Tsipras, A. Ilyas, and A. Madry, "How does batch normalization help optimization?," *Advances in neural information processing systems*, vol. 31, 2018.

[5] M. Hussain, J. J. Bird, and D. R. Faria, "A study on cnn transfer learning for image classification," in *Advances in Computational Intelligence Systems: Contributions Presented at the 18th UK Workshop on Computational Intelligence, September 5-7, 2018, Nottingham, UK*, pp. 191–202, Springer, 2019.

[6] L. Fei-Fei, H. Su, M. Do, K. Li, and J. Deng, "Construction and analysis of a large scale image ontology," 2009.

[7] B. Settles, "Active learning literature survey," 2009.

[8] M. Abdar, F. Pourpanah, S. Hussain, D. Rezazadegan, L. Liu, M. Ghavamzadeh, P. Fieguth, X. Cao, A. Khosravi, U. R. Acharya, *et al.*, "A review of uncertainty quantification in deep learning: Techniques, applications and challenges," *Information Fusion*, vol. 76, pp. 243–297, 2021.

[9] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[10] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, and G.-Z. Yang, "Xai—explainable artificial intelligence," *Science robotics*, vol. 4, no. 37, p. eaay7120, 2019.

[11] R. R. Selvaraju, A. Das, R. Vedantam, M. Cogswell, D. Parikh, and D. Batra, "Grad-cam: Why did you say that? visual explanations from deep networks via gradient-based localization," *CoRR*, vol. abs/1610.02391, 2016.

[12] A. Shrikumar, P. Greenside, and A. Kundaje, "Learning important features through propagating activation differences," *CoRR*, vol. abs/1704.02685, 2017.

[13] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," *Advances in neural information processing systems*, vol. 31, 2018.

[14] R. L. Klatzky, S. J. Lederman, and V. A. Metzger, "Identifying objects by touch: An "expert system"," *Perception & psychophysics*, vol. 37, pp. 299–302, 1985.

[15] S. J. Lederman and R. L. Klatzky, "Extracting object properties through haptic exploration," *Acta psychologica*, vol. 84, no. 1, pp. 29–40, 1993.

[16] M. Cavdan, K. Doerschner, and K. Drewing, "Task and material properties interactively affect softness explorations along different dimensions," *IEEE Transactions on Haptics*, vol. 14, no. 3, pp. 603–614, 2021.

[17] R. Li and E. H. Adelson, "Sensing and recognizing surface textures using a gelsight sensor," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1241–1247, 2013.

[18] A. Amini, J. I. Lipton, and D. Rus, "Uncertainty aware texture classification and mapping using soft tactile sensors," in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 4249–4256, 2020.

[19] W. Yuan, Y. Mo, S. Wang, and E. H. Adelson, "Active clothing material perception using tactile sensing and deep learning," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 4842–4849, 2018.

A. Appendix

A.1. Digit Sensing

In this section, we report preliminary experiments with a Digit tactile sensor[1]. The main experiments were carried out with a GelSight Mini sensor, which provides a higher resolution, improved contact surface properties, and a more reliable lighting.

A.1.1. Classification Setup

At first, a set of five objects was used for classification. We created a dataset of ten images per object by pressing the Digit sensor on the surfaces. This data collection was done with varying orientations and positions, mimicking multiple touches with a fingertip. We split our data into six training, two validation, and two test images. For classification, we fine-tuned ResNet18 for 30 epochs.

To get a better insight into the importance of the chosen EP for the classification performance we additionally used data from lateral motions. Specifically, the Digit sensor was slid across the objects for 3 seconds while recording a video with a resolution of 30fps which lead to 90 frames per data point. Each video was then passed to Inception v3 which operated on the individual frames without classifying them. This feature extraction can be done by omitting the last network layer and therefore just outputting the features of images. In the following steps, ten feature-extracted frames were considered as one data point. Since the original videos had 90 frames, this split resulted in a nine times bigger data set. These 10-frame-videos were passed to a Recurrent Neural Network (RNN) which then used the temporal relationship between the features to make a prediction.

[1] https://digit.ml/

A.1.2. Classification Results

In the image classification task using the Digit sensor, the network performed with 100% accuracy on the test data within 30 epochs of training. The dataset was then recollected to get more variability between images. After that, the test performance was 70% within 70 epochs. The overall performance on the images with higher variability can be seen in Figure A.1.

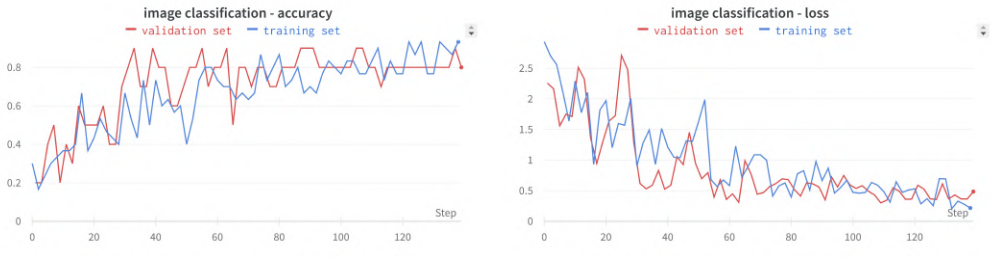


Figure A.1.: The network performance of classifying images using Digit.

The results of the video classification task can be seen in Figure A.2. The accuracy on the test data is 86%. The fact that the network structures used for image and video classification are very different prevents us from drawing direct conclusions about the underlying EPs pressing and sliding. However, when passing the Inception v3 features individually to the RNN instead of in groups of ten frames, the performance decreases to 75% on the test data. By changing this parameter, we see the value of temporal information in receiving multiple frames as input. These results support the notion that applying a sliding motion instead of a pressing motion helps the network's performance.

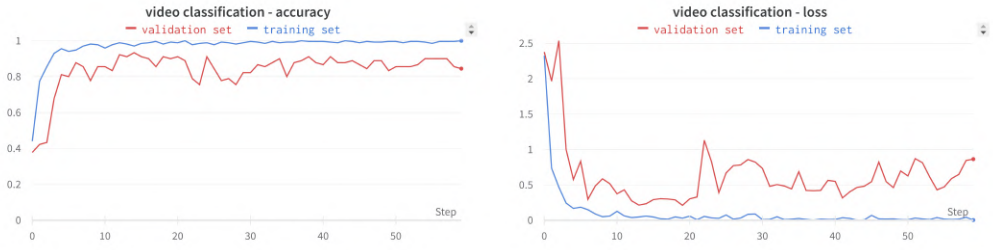
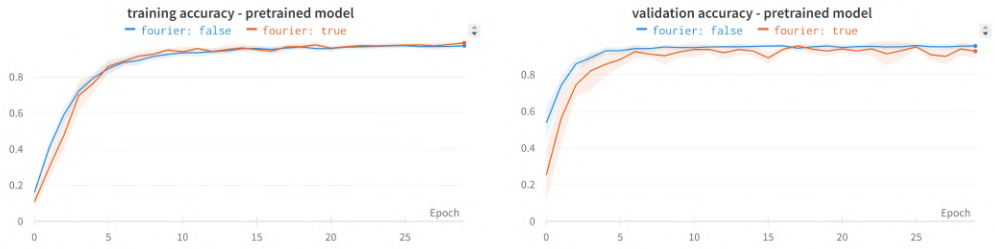


Figure A.2.: The network performance of classifying videos of lateral motion using Digit.

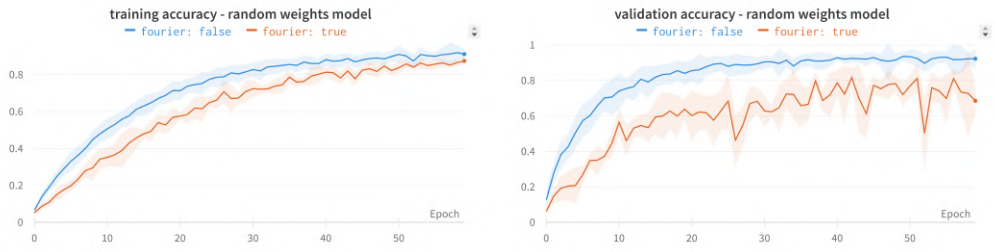
A.2. Concept-Based Explanations

Since saliency methods on the original input images do not provide helpful explanations, we investigate whether working with Fourier-transformed images can provide more helpful concept-based explanations for the texture classification task. We first train the three models to discriminate the 25 textures based on the Fourier-transformed inputs, and compare the performance with training on the original images. In Figure A.3 we show the accuracy on the training and validation data for each model. We can see that the use of pretrained network weights helps to classify the transformed images as well as the original images. For the random weights model, learning is slower, and for the small inception model, we need to increase training to 300 epochs to achieve approximately the same training accuracy as with the original images.

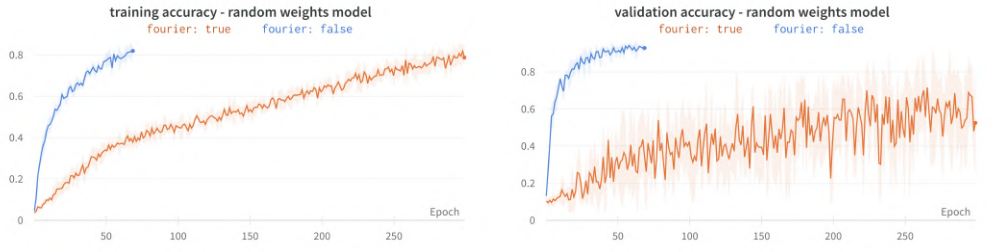
The results of applying saliency methods to the small inception model can be seen in Figure A.4. In addition to looking at the *Grad-CAM* and *Input X Gradient* images, we apply the inverse Fourier transform to the *Input X Gradient* results to get a better understanding of important concepts for the network. To map the transformed image back to the original, we adjust the *Input X Gradient* pixel values to be either zero if there is a negative or no influence, and one otherwise, resulting in the Fourier transformed image without the image parts that do not positively influence the classification. In terms of usefulness, we can see that the right-most *Grad-CAM* image shows the importance of the high-magnitude frequency components belonging to the regular patterns in the texture images. The explanation can therefore be considered to convey a helpful concept. However, this helpfulness cannot be observed as clearly in the other *Grad-CAM* heatmaps or the *Input X Gradient* results. We conclude that there is potential for good concept-based explanations using Fourier transforms of texture images, but more in-depth research is needed.



(a) pretrained model

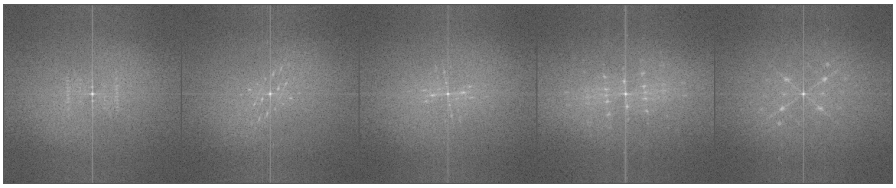


(b) random weights model

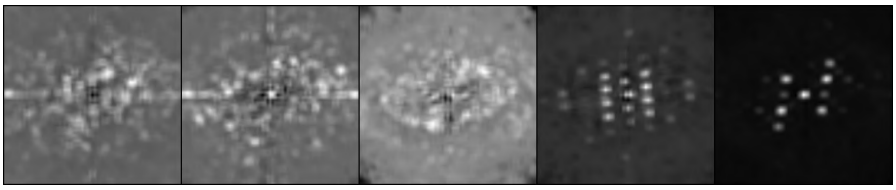


(c) small inception model

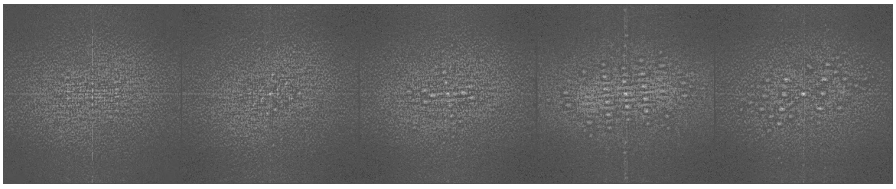
Figure A.3.: The network performances of classifying Fourier-transformed texture images of 25 fabrics.



(a) Fourier-transformed images



(b) *Grad-CAM*



(c) *Input X Gradient*

(d) inverse Fourier-transform on *Input X Gradient*

(e) original images

Figure A.4.: The results of saliency methods after training the small inception model on Fourier-transformed texture images.