MSc Project - Specification and Proposed Design

Texture Recognition and Feature Sharing for Visual-Tactile Cloth using Deep Maximum Covariance Analysis Model

Written by: Hong Liu

Student ID: 201441283

Course Code: COMP702 MSc Project

Primary Supervisor: Shan Luo

Secondary Supervisor: Yannis Goulermas

Contents

1. Introduction

Humans use multiple sensing to observe the physical world and store the observation in memory with multi-sensorial modalities. These modalities contain different sort of human feelings, such as vision, touch, hearing, smell and so on. The most common modalities we use to recognise object are vision and touch. We can employ our eyes to see and capture the static states and dynamics of environment around us to form visual imagery. The visual image has representation of appearance, shape, colour of the object. For touch feeling, when we try to grasp an object through our hand, the detail future of object, such as magnitude, texture, shape can be obtained. Meanwhile, the visual experience become infeasible as our hand block out part of surface of the object and thus hand sensation helps us to 'see' the hidden visual features.

Relate researches in this filed has deep insight into the correlation between single sensor system and multiple sensors system. Modern robots capture the complementary information of the environment through multiple sensors that implemented in their body [1]. Machine vision has also gained a great deal of attention in recent years, object recognition can be achieved well and are able to usually acquire better accuracy than human [2]. In addition, machine touch also be widely used to texture recognition and considerable efforts have been made to use tactile sensing to establish system for object recognition [3,4]. Robotic systems are asked to achieve many complex objectives including environment exploring, object recognition and manipulation. Integrating multiple modalities of sensors has capability to overcome the drawback of single sensor that has necessarily shortage in power. Multiple sensors systems enable the acquisition of a coherent and comprehensive perception of the environment through providing redundant and complementary information that are important for construing confidence measures in the scene of objection recognition [5]. In this process, the fusion of touch and vison is a very powerful approach. This kind of fusion is interested in finding sharing features between multiple sensing modalities.

In this project, we attempt to address texture recognition from clothes or fabrics that have various physical properties. Tactile sensing about clothes describe direct experience including many detail information such as thickness, fold, stiffness, as well as yarn distribution. In deeded, visual sensing are also able to receive the identical feature about that clothes since

similar physical parameters should exists in same clothes. we expect to get more accurate classification of all clothes samples by combing the visual sensing and touch sensing instead of based on only one modality. Moreover, it is preferred that getting a joint latent space for cloth sharing feature. The dataset of this project come from the collection of both visual image and tactile image that taken by a Canna camera and a GelSight reading sensor [6].

This project aims to employ the association between visual sensing and touch sensing, and achieve cloth texture recognition and then categorise these clothes into accurate classes by adopting an appropriate type of method that relying on the association. We expect to construct a model that will learn a joint latent space from both the two modalities. The model will be composed of two separate convolutional neural networks and a maximum covariance analysis layer for sharing space of two modalities [6]. The two neuron networks dedicate to study respectively the features of visual images and touch images and then will learn a share representation space through maximum covariance analysis. The dataset to be used in this model derive from dual-modalities as noted before, visual images taken by a camera sensor and haptic image generated from a high-resolution GelSight sensor. We feed the datasets to our two neural networks architectures for training and determine whether different species of fabrics can be accurately recognised and classified to ground truth types of the clothes. In addition, we are also interested in compare the performance of cross-modality training and single training, and look for the joint space that boost our texture recognition task.

2. Related work

2.1 Vision and touch object recognition

There are many studies about how visual and touch information can be used to make judgments about the characteristics of an object. One response to the question is that we can use visual data to recognise an object more efficient than using touch data when we try to identify an object with our hands. A 50 matching-to-sample experiment lead to an obvious evidence that visual judgments are more accurate than haptic judgments when make judgement on the form of object [7]. Also, visual information can get heavier weight in cross-modality condition [8]. A special model of with weighted averaging was created for width measurement of two cubes and both visual and touch identify are involve in the model, the result of this experiment shows that higher weights appeared in visual information when

people are allowed to see and touch the object [9]. However, it is possible that touch sensing is appropriate for texture recognition rather than form recognition and the dominance of visual data may not reflect in the field of texture recognition. There is no clear difference between visual sensing and haptic sensing for distinguishing smoothness of an object's surface and the former shows a less accuracy than the latter if difference clearly exists [10].

2.2 Transition of texture information across modalities

Some research conducted in the transition of vision and touch information at texture discrimination task [11]. A test about texture recognition of some fabric was allocated to two groups (5- and 8-years old children) by using intra-modal and cross-modal. When vary degrees of softness and thickness were observed in the texture samples, both the two children group reached the identical performance across vision and touch. Nevertheless, visual identifying prior than touch when the test objects have similar haptic properties, which means that the texture recognition with cross modalities was mainly finished through bottom-up perceptual processing on the test object [12]. An experiment that surveying the affection of bimodal exploration on the perception dimension of people compare to unimodal exploration argue that bimodal condition obtained better fit than unimodal when recognition scenes are spatial arrangement task (SAT) and free classification task (FCT) [13].

2.3 Integrated system of vision and touch

A new visual-tactile integrated model was proposed to recognise object. In this model, the dataset of dual-modal was gathered respectively, using multivariate-time-series and covariance descriptor describe tactile data and visual image and then executing classification task though employing a fusion algorithm-joint group kernel sparse coding (JGKSC). The experiment with this model proved that the performance of the integrated system is expressively pass beyond than those of single modality approaches [14].

3. Project scope

3.1 Project rationality and relate question

Object recognition has been widely used in the modern industries such as computer vision, mechanical intelligence and automation. Numerous methods for object recognition are of using the data derive from various camera to make image classification. In addition, tactile sensing can be also used to obtain multiple characteristics of objects and thus applied in

shape recognition, manipulate objects, as well as surface reconstruction. Therefore, we attempt to give deep insight into improving the performance of object recognition. Although many research progresses dedicated to the task of object recognition by employing visual and touch information independently. However, due to the two modalities share similar and complementary parameters [6], it is still an attractive issue that how to enable the fusion of visual and tactile information to facilitate recognition and classification of objects.

3.2 Research scopes

This project takes into consideration texture recognition of 100 types of usual clothing as our experiment scenario. We collect clothing images from both visual and tactile modalities and then form using these imagery data to form a set of pair samples in which every pair consist of one vision image and one tactile image. Then we create a neural network classifier to classify the clothing and compare to the ground truth labels. Though, due to the reason of any inaccuracy moving on the gathering process and the cloth source are not totally same, our training image samples cannot be strongly paired [6]. Thus, in order to against the difference and weak pairing between the dual modalities, we propose a new framework, deep maximum covariance analysis for the visual tactile integration in fabric texture recognition.

To achieve the framework, the main objectives are listed as follows:

- Creating two special convolutional neural networks (AlexNet) that have independent learning parameters to achieve deep learning of clothing texture.
- Feeding visual images and haptic images to the networks in order to extract feature.
- Developing a maximum covariance descriptor (WMCA) to learn a joint hidden space for sharing feature.
- Analysing the classification results and make evaluation for the performance of the fusion model.

3.3 AlexNet

Our project has large amount of quality and labelled images to be trained on, which required a more complex neural network in order to best result. As a tremendously capable deep learning model, AlexNet was widely used in the scenarios that have large datasets such as

ImageNet classification. Thus, we also introduce AlexNet to our project for texture recognition task.

The architecture of AlexNet is made up of multiple layers including five convolution layers, three max pooling layers and three full connected layers.

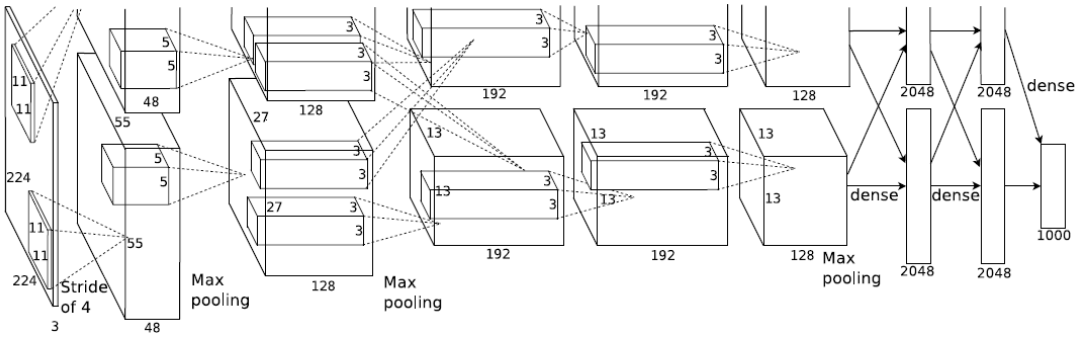


Figure 1. The architecture of AlexNex [15].

Since our model will be trained on 100 classes from both visual and tactile images, we can substitute the 1000 neurons with 100 neurons in the last full connected layer. We need to crop the original image to the size of 227x227x3 and then feed them into the network to learn the representation of clothing. Then the output of the last full connected layer will experience a computing with Softmax activation function that generate a distribution over 100 classes corresponding to 100 types of fabric in this project

What made AlexNet special compare to common convolutional neural network are a few new features as following [15]:

- Relu Non-linearity

  AlexNet use Rectified linear unit function as activation instead of *Sigmoid* or *Tanh* function in traditional neuron network in early stage. Relu function has capability of getting good performance by less computation and learning acceleration.

- Overlapping pooling

  Standard pooling operation is located in the neighbouring groups neurons and thus there is no overlap between adjacent operations. However, AlexNet introduce a polling operation that executing pooling computing in an area that overlap each other. This overlapping technique has been proved to hardly get overfit and error reduction by 0.4%.

- Multiple GPU

AlexNet mode can be trained on GPU by placing half of neurons on GPU and another half was computed on CPU, which lead to that larger model are allowed to be trained and spending less time to train the network.

- Local response normalization
  A local normalization scheme, also known as “brightness normalization” was introduced to generalise our dataset and reduce error-rate.

3.4 Weakly-Paired Maximum Covariance Analysis

The datasets in this project come from two sensing modalities and we attempt to find joint latent space between them. We can choose maximum covariance analysis (MCA) to extract share information through multimodal dimensionality reduction that projecting data from high dimension to lower dimension of feature spaces.

MCA algorithm employ singular value decomposition (SVD) on the covariance matrix to acquire a lower-dimension representation about that matrix. SVD can decompose any matrix to below form:

$$A_{mxn} = U_{mxm} S_{mxn} V^T_{nxn}$$

Where U and V are left and right singular matrices respectively. S is a diagonal matrix consist of singular values elements in diagonal positions and zero values on the rest. Basically, the sum of the first 10% even 1% singular values is approximately equivalence of total sum of all singular values and thus we usually extract the first q singular values.

We assume X and X’ are the two modalities of clothing dataset and thus XX’ is the covariance matrix about the clothing. Then we apply SVD on the matrix and obtain the projection matrices W and W’ which are q-dimension subspace which be made of the first q columns from U and V. Finally, a lower-dimensional representation can be attained as

$$X_{new} = W^T X,\ X'_{new} = W'^T X'.$$

MCA aims to maximise the covariance between two lower-dimension matrices as large as possible and its objective function can be given as

$$\max_{w,w'} \mathrm{tr}[W^T X\, X'^T W']$$

where tr symbol means matrix trace computation which sum of all diagonal elements of the matrix [16].

Traditional multi-modal algorithms rely on strong correlation and dependencies between paired samples [16]. However, as mentioned, we gather the visual and tactile images by using

different collecting methods and thus slightly different source. For this reason, we assume a weakly-paired data. Thus, we adopt a variant MCA, weakly-paired maximum covariance analysis (WMCA).

WMCA algorithm is able to conclude strong pairings instances from poor paired dataset by adding a *n x n'* pairing matrix π in which one instance in *i*th row represent a sample in visual dataset and it corresponds to a unique *j*th sample in tactile dataset and there are total n sample pairings [16]. According to this rule, $\pi \in \{0, 1\}^{n \times n'}$, which indicates that the elements of the pairing matrix are either 0 (non-pair) or 1 (pair). In addition, if we order the samples according to their location in weakly-paring group, a pair matrix that all diagonal elements are 1 can be obtained to guarantee accurate match. As noted early, standard MCA ask a strong paired dataset but it is impractical in our datasets and thereby WMCA will be used in this project. We change the objective function to

$$\max_{w,w',\pi} \mathrm{tr}[WTX\ \pi\ X'TW']$$

WMCA pseudo code like below:

WEAKLY-PAIRED MAXIMUM COVARIANCE ANALYSIS

INPUT:
Weakly-paired data from sensors one $\mathbf{X}$ and two $\mathbf{X}'$
Desired output dimensionality $q \leq \min(\{n, n', d, d'\})$
INITIALIZATION:
$\eta = 1$
$\hat{\mathbf{\Pi}} = \mathrm{diag}(\hat{\mathbf{\Pi}}^1, \ldots, \hat{\mathbf{\Pi}}^g)$ and $\mathbf{\Pi} \rightarrow \hat{\mathbf{\Pi}}$ wherein
$[\hat{\mathbf{\Pi}}^h]_{i,j} = \min(n_h, n'_h)^{-1} \forall i = 1, \ldots, n_h, j = 1, \ldots, n'_h$
ANNEALING WMCA:
while $\eta \geq 0$
Run *Alternating Maximization*
Reduce η
ALTERNATING MAXIMIZATION:
while trace value of $\mathbf{W}^t\mathbf{X}\mathbf{\Pi}\mathbf{X}''\mathbf{W}'$ increases
Step 1) Maximize with respect to $\mathbf{W}$ and $\mathbf{W}'$:
Obtain $\mathbf{W}$ and $\mathbf{W}'$ from MCA($\mathbf{X}\mathbf{\Pi}\mathbf{X}'^T, q$)
Step 2) Maximize with respect to $\mathbf{\Pi}$:
Set all elements of $\mathbf{\Pi}$ to zero
for $h = 1$ to g
Compute the cost matrix $\mathbf{C} = [\mathbf{X}_h''\mathbf{W}'\mathbf{W}^t\mathbf{X}_h]^t$
Solve linear assignment problem for $\mathbf{C}$
Set elements of $\mathbf{\Pi}$ to 1 for assigned pairings
Anneal) Relax pairings:
$\mathbf{\Pi} \rightarrow \eta\hat{\mathbf{\Pi}} + (1 - \eta)\mathbf{\Pi}$
OUTPUT:
Projection matrices $\mathbf{W}$ and $\mathbf{W}'$

Figure 2. An illustration of WMCA algorithm [16]

Because of adding the n x n' pairing matrix π, we need to optimize three parameters w, w' and π, so we introduce a new maximization mechanism that combining SVD and Hungarian algorithm [17].

3.5 Deep Maximum Covariance Analysis Model (DMCA)

This model aims to learn representations of visual images and touch images and then pass them to a nonlinear transformation to learn a joint latent space.

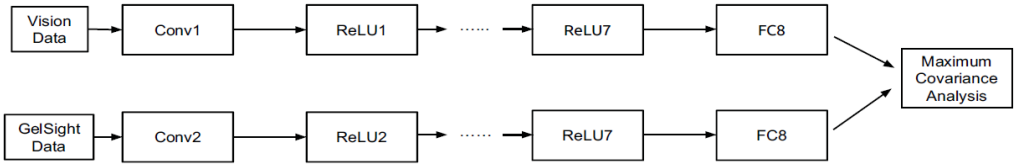


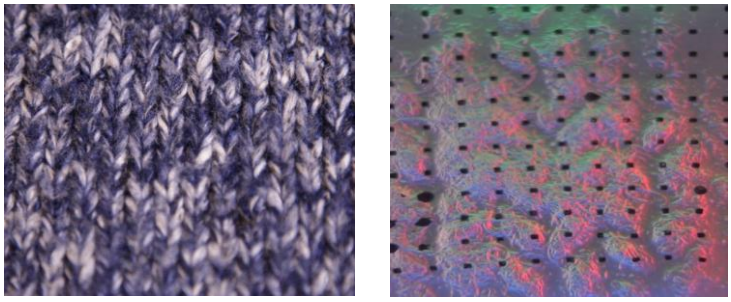
Figure 3. An illustration of DMCA model [6]

There are some steps have to be done in this model:

1) Giving pre-processed visual and tactile images to the two separate convolutional neural networks with AlexNet architecture and extract significant features of these images.
2) Passing the extracted feature embeddings to WMCA layers and then learn a joint latent space.
3) Computing the covariance between visual and tactile inputs in the joint latent space according to the objective function mentioned before.
4) Train our model by continuously maximize w, w' and π and optimize the parameters of two networks in order to better classification.

3.6 Dataset

The dataset used in this project are the visual images and tactile images of the 100 species of clothing texture. Visual images are generated by Keeping a Cannon Camera parallell to the surface of the cloth and then capture the its texture with multiple times of rotations [18]. When it comes to haptic images, we use a GelSight sensor to gather cloth texture. The GelSight sensor is a plastic cube with a layer of elastomer or rubber and the four walls of the cube have light of different colours. When we place the sensor on cloths and then press sensor, the rubber will be deformed and light bounces off of the cube and is captured by the camera straddling on the surface. The images of two modalities can be seen from Figure 4.



(a) (b)

Figure 4. Texture image of clothing gathered from Camera (a) and GelSight Sensor (b)

By using these two devices, this project has substantial number of ViTac sample set including 1,000 visual images and 96,536 touch images [6]

For the purpose of ethically using data, the ethical source of data should be carefully considered. Our project gains the images set of two modalities from academic papers recognised by 1st supervisor and they are definitely synthetic data rather than human data.

3.7 Experimental evaluation

For our project, a key aspect of determining whether our DMCA model is fit for texture classification is assessing its predictive performance. DMCA expect to correctly classify the label of every sample of clothing datasets to corresponding ground truth label as much as possible.

To evaluate the predictive performance of the model, we can choose multi-class performance metrics for our supervised learning. In a common multi-class classification problem, we need to classify each sample into 1 of N different classes. We build a confusion matrix where the row elements of the matrix represent predicted labels and the columns give the actual label. Assuming that the samples are ordered according to their location in weakly-paring group

	Actual_1	Actual_2	Actual_3	……	Actual_n
Predicted_1	num	num	num	num	num
Predicted_2	num	num	num	num	num
Predicted_3	num	num	num	num	num
……	num	num	num	num	num
Preducted_n	num	num	num	num	num

Table 1. Confusion table of multi-class performance evaluation

There are many metrics for measuring the performance of our model, including accuracy (A), precision (P), recall (R).

$$A = (TP+TN)/(TP+TN+FP+FN)$$

$$P = TP / (TP + FP)$$

$$R = TP / (TP + FN)$$

Accuracy means the performance accuracy of the model by computing the sum of diagonal divided by the total sum of all elements in the table; Precision (P) and recall (R) need to be considered on a per-class basis. We can choose appropriate metrics in this project.

4. Project Plans

4.1 Project outputs

The project will produce the outputs as below:

- Project specification: a design document that describes what the project is about and what question the project wish to solve, as well as available solutions that is appropriate to the question.
- Python source code: describes the implementations of algorithms and analysis methods of the proposed model.
- Experiment results and analysis: the output of texture classification and relate analysis.
- Evaluation of the proposed model: the result of measuring the proposed model with multi-class performance metrics.
- Project Report: Summary of the project.

4.2 Milestones and main tasks

Milestone 1 - Project starting and write the specification and design of the project. (04/06/2020 – 14/07/2020)

1) Search and review academic literatures relate to the theme of the project
2) Understand the work principle and algorithm implementation of proposed model
3) Collect dataset including vision images and touch images
4) Arrange project meeting and discuss project progress and questions with supervisor
5) Write project specification and presentation
6) Oral presentation

Deliverables: Project Specification, Design Presentation

Milestone 2 – Model implementation and training and test. (22/06/2020 – 30/07/2020)

7) Import the pre-trained AlexNet architecture to establish our ConvNets
8) Rescale images and extract its centre part of 227x227x3
9) Data augmentation (rotation, colour shifting)
10) Training the AlexNet model and revise the hyperparameters
11) Test the model with test dataset

Deliverables: Python code and experiment outputs

Milestone 3 – Result analysis and performance evaluation (31/07/2020 – 10/08/2020)

12) Experiment result analysis
13) Evaluation of the model

Deliverables： Analysis and evaluation result

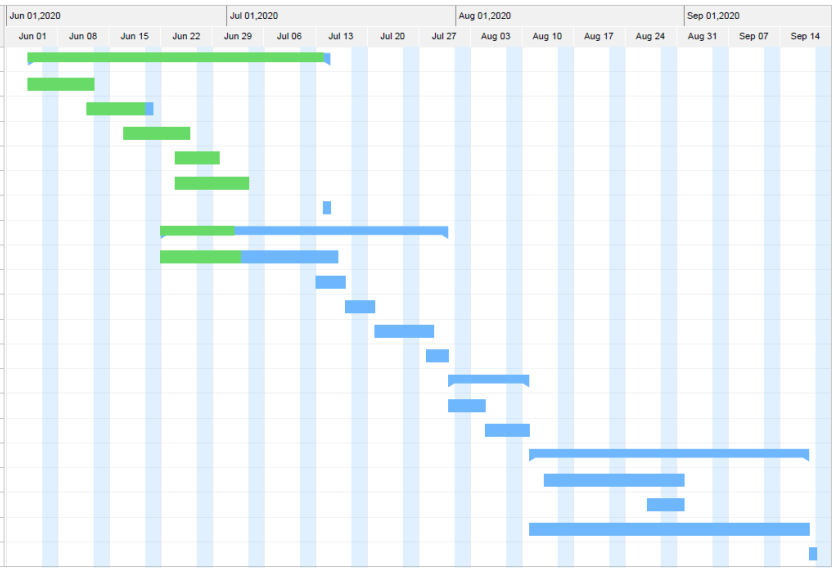
Milestone 4 – Completion of the project (11/08/2020 – 17/09/2020)

14) Prepare presentation
15) Final presentation
16) Write final project report

Deliverables: Project dissertation

4.3 Gantt Chart

ID	Task	Start	Finish	Duration	Progress	Priority
1	Project Starting	2020-06-04	2020-07-14	29 d	97%	None
2	Review academic literatures	2020-06-04	2020-06-12	7 d	100%	None
3	Investigate model and method	2020-06-12	2020-06-20	6 d	100%	None
4	Collect dataset	2020-06-17	2020-06-25	7 d	100%	None
5	Project meeting	2020-06-24	2020-06-29	4 d	100%	None
6	Project specification	2020-06-24	2020-07-03	8 d	100%	None
7	Oral presentation	2020-07-14	2020-07-14	1 d	0%	None
8	Model Traning and Test	2020-06-22	2020-07-30	29 d	27.3%	None
9	Model importing	2020-06-22	2020-07-15	18 d	50%	None
10	Data preprocessing	2020-07-13	2020-07-16	4 d	0%	None
11	Data augmentation	2020-07-17	2020-07-20	2 d	0%	None
12	Model training	2020-07-21	2020-07-28	6 d	0%	None
13	Model test	2020-07-28	2020-07-30	3 d	0%	None
14	Analysis and Evaluation	2020-07-31	2020-08-10	7 d	0%	None
15	Result analysis	2020-07-31	2020-08-04	3 d	0%	None
16	Model evaluation	2020-08-05	2020-08-10	4 d	0%	None
17	Final Review	2020-08-11	2020-09-17	28 d	0%	None
18	Prepare presentation	2020-08-13	2020-08-31	13 d	0%	None
19	Final presentation	2020-08-27	2020-08-31	3 d	0%	None
20	Final dissertation	2020-08-11	2020-09-17	28 d	0%	None
21	Finish Project	2020-09-18	2020-09-18	1 d	0%	None



4.4 Risk Assessment

Identified Risks	**Likelihood**	**Impact**	**Risk Symptoms**	**Risk Management/ Mitigating Factors**
Potential hazards	Very low (1) Low (2) Med (3) High (4) Very High (5)			
Research personnel				
Getting sick	1	2	Delay project proceeding	Keep good lifestyle
Unreasonable Time management	3	3	Cannot complete the report on time	Allocate your time wisely Negotiate with supervisor for delay
Research Integrity				
Academic sources are insufficient to meet research	3	4	Poor evidence support argument Digress from the subject	Acquire valuable materials from supervisor Reading a lot
Algorithm implement difficulty	3	5	Programs cannot run and outputs	Refer to the other' work Seek assistance Tune algorithm
Loss of file and code	2	5	Loss motivation Waste time	Keep backup regularly Upload to online
Uncontrollable Factors				
Library and material access difficulty	5	3	Cannot to access University's library and Google scholar	Purchase VPN Download material in advance
Coronavirus	5	3	Negative impact on work efficiency	Keep personal health Seek assistance from supervisor and University
Poor computation power	3	4	Slow training speed of model	Try to use Google Colab GPU

5. References

[1] Gorges N, Navarro SE, Göger D, Wörn H. Haptic object recognition using passive joints and haptic key features. In: Proceedings—IEEE international conference on robotics and automation; 2010. p. 2349–55.

[2] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. In: 2015 IEEE international conference on computer vision (ICCV). IEEE; Dec 2015. p. 1026–34

[3] Madry M, Bo L, Kragic D, Fox D. ST-HMP: unsupervised spatio-temporal feature learning for tactile data. In: Proceedings—IEEE international conference on robotics and automation; 2014. p. 2262–9.

[4] Soh H, Su Y, Demiris Y. Online spatio-temporal Gaussian process experts with application to tactile classification. In: IEEE international conference on intelligent robots and systems; 2012. p. 4489–96

[5] Kim JK, Wee JW, Lee CH. Sensor fusion system for improving the recognition of 3D object. In: IEEE conference on cybernetics and intelligent systems, 2004, vol 2; 2004. p. 1207–12

[6] Luo, S., Yuan, W., Adelson, E., Cohn, A.G. and Fuentes, R., 2018, May. Vitac: Feature sharing between vision and tactile sensing for cloth texture recognition. In 2018 IEEE International Conference on Robotics and Automation (ICRA) (pp. 2722-2727). IEEE.

[7] JONES, B. (1981). The developmental significance of cross-modal matching. In R. D. Walk & H. L. Pick, Jr., *Intersensory perception and sensory integration.* New York: Plenum Press.

[8] Jones, B. and O'Neil, S., 1985. Combining vision and touch in texture perception. *Perception & Psychophysics*, *37*(1), pp.66-72.

[9] JONES, B. (1983). Psychological analyses of haptic and haptic-visual judgments of extent. *Quarterly Journal of Experimental Psychology: Human Experimental Psychology*, 35, 597-606.

[10] HELLER, M. A. (1982). Visual and tactual texture cooperation: Intersensory cooperation. *Perception & Psychophysics*, 31, 339-344.

[11] Whitaker, T.A., Simões-Franklin, C. and Newell, F.N., 2008. Vision and touch: Independent or integrated systems for the perception of texture?. *Brain research*, *1242*, pp.59-72.

[12] Picard, D., 2007. Tactual, visual, and cross-modal transfer of texture in 5- and 8-year-old children. Perception 36, 722–736.

[13] Ballesteros, S., Reales, J.M., Poncé de Leon, L., García, B., 2005. The perception ofecological textures by touch: does the perceptual space change under bimodal visual

and haptic exploration? Proceedings of the First Joint Eurohaptics Conference and Symposium on Haptic Interfaces for Virtual Environment and Teleoperator Systems (World Haptics). Barlow, H.B., 1978. The efficiency of detecting changes of density in random dot patterns. Vis. Res. 18, 637–650.

[14] Liu, H., Yu, Y., Sun, F. and Gu, J., 2016. Visual–tactile fusion for object recognition. *IEEE Transactions on Automation Science and Engineering*, *14*(2), pp.996-1008.

[15] Krizhevsky, A., Sutskever, I. and Hinton, G.E., 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097-1105).

[16] Kroemer, O., Lampert, C.H. and Peters, J., 2011. Learning dynamic tactile sensing with robust vision-based training. *IEEE transactions on robotics*, *27*(3), pp.545-557.

[17] H. W. Kuhn, "The hungarian method for the assignment problem," Naval Research Logistics Quarterly, vol. 2, 1955.

[18] Yuan, W., Wang, S., Dong, S. and Adelson, E., 2017. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5580-5588).