Visual-Haptic-Kinesthetic Object Recognition with Multimodal Transformer ⋆

Xinyuan Zhou[1], Shiyong Lan*[1], Wenwu Wang[2], Xinyang Li[1], Siyuan Zhou[1], and Hongyu Yang[1]

[1] College of Computer Science, Sichuan University, Chengdu 610065, China
[2] University of Surrey, Guildford, GU2 7XH, United Kingdom

Abstract. Humans recognize objects by combining multi-sensory information in a coordinated fashion. However, visual-based and haptic-based object recognition remain two separate research directions in robotics. Visual images and haptic time series have different properties, which can be difficult for robots to fuse for object recognition as humans do. In this work, we propose an architecture to fuse visual, haptic and kinesthetic data for object recognition, based on the multimodal Convolutional Recurrent Neural Networks with Transformer. We use Convolutional Neural Networks (CNNs) to learn spatial representation, Recurrent Neural Networks (RNNs) to model temporal relationships, and Transformer's self-attention and cross-attention structures to focus on global and cross-modal information. We propose two fusion methods and conduct experiments on the multimodal AU dataset. The results show that our model offers higher accuracy than the latest multimodal object recognition methods. We conduct an ablation study on the individual components of the inputs to demonstrate the importance of multimodal information in object recognition. The codes will be available at https://github.com/SYLan2019/VHKOR.

Keywords: Object Recognition · Multimodal Deep Learning · Multimodal Fusion · Attention Mechanism.

1 Introduction

In the real world, object recognition is fundamental to many of the cognitive and interactive capabilities of robots. With the development of sensor technology, machine vision performs well in terms of object appearance recognition [12] and object detection [29], as does machine haptics in texture recognition [19,5] and material classification [20]. These methods are often relying on only one type of sensory information. However, the information from a single modality may not be sufficiently reliable for object recognition. For example, the quality of the visual data can be affected by the quality of the camera, the presence of object occlusion and illumination, while the haptic data can be affected by the type of the haptic

⋆ This work was funded by 2035 Innovation Pilot Program of Sichuan University, China. * Corresponding author. E-mail:lanshiyong@scu.edu.cn.

sensor used, the area of the sensor placed, and the background noise presented in the environment. Even with the best hardware and ideal scene conditions, there are other issues that can cause significant challenges for object recognition with a single modality, for example, when recognising the glass materials, or objects with the same appearance but different content. To address these limitations, we consider robotic object recognition by using multimodal information.

Recent studies have explored methods for fusing visual and haptic data [28,8,9,6,13,27,18,21,24,3], but several challenges remain. Firstly, data in different modalities have different characteristics and representations. For example, image data is static, with a single image containing a wealth of visual information. Tactile data, on the other hand, is time-series and has a high sampling rate. Consequently, how to extract features from them effectively is still an important issue. Secondly, it is challenging to make accurate connections between data in more than two different modalities. Finally, multimodal fusion methods often require a large amount of computational resources and time, and how to improve computational efficiency with expected recognition accuracy is also a practical challenge. To address these issues, we design feature extraction networks for data in three different modalities: visual, haptic and kinesthetic, where the kinesthetic represents the kinematics information (more details can be seen in Section 4.1) of the robot's wrist, fingers and palm. Then, we propose two fusion methods based on Transformer's attention mechanisms and further improve the accuracy and efficiency of robotic object recognition with multimodal information.

In this paper, we design a Convolutional Recurrent Neural Network (CRNN) to extract features for data in three different modalities: visual, tactile and kinesthetic. We use Convolutional Neural Networks (CNNs) to extract the features of each modality, and use a Bi-directional Long Short Term Memory Network (Bi-LSTM) [15] to model the temporal relationships of the tactile sequences. Then, we use Transformer's attention mechanism to fuse unaligned signals, thus further improving the accuracy and efficiency of multimodal fused robotic object recognition. The proposed method has an advantage of using fewer Transformer encoders to achieve better performance than existing transformer based fusion methods. We conducted experiments on the latest AU dataset [4] and compared our method with popular methods in the field of robotic multimodal object recognition.

The main contributions of this paper are summarized as follows:

- We design two new multimodal object recognition methods based on visual, haptic, and kinesthetic signals. A holistic neural network structure is used for multi-input single-output classification with unaligned multimodal data. Ablation experiments are performed to demonstrate the importance of complementary multimodal signals for object recognition.
- We design different feature extraction networks based on the characteristics of each signal, and combine the transformer with CNN and RNN. We compare the effects of different fusion methods and the attention mechanisms on multimodal classification networks. Our methods offer a higher accuracy than the mainstream methods in the field of robotic multimodal object recognition, and use fewer Transformer modules than the latest Transformer-based

multimodal fusion methods, which result in a model of fewer parameters and higher training speed.

2 Related Work

The study of multimodal fusion methods for robotic object recognition, grasping, and other operations is an emerging field. Initially, vision and haptics were used jointly to generate descriptions of object surfaces [1], and later extended to object recognition tasks [2]. Early works explored different ways of representing and encoding visual images and haptic sequences, using Dynamic Time Warping (DTW), K Nearest Neighbor (K-NN), and the Extreme Learning Machine (ELM) for classification [27,18,21]. However, more complex, higher-dimensional real-world data cannot be described by fixed equations, and the computational cost of the model designed for each specific task is prohibitive.

Since 2010s, Deep Learning (DL) has made outstanding contributions to various tasks, thanks to its ability in learning abstract and high-level representations of the data with a layered structure of the network. In recent years, CNNs have been widely used for multimodal object recognition tasks. The vast majority of work chooses CNNs or RNNs to extract features for each modal information, and then fuse them in a connected layer [28,13]. Some of these methods require a large amount of strictly aligned multimodal data to achieve good recognition results [23], while others may use a CNN-only or RNN-only network, which has disadvantages of a single network being difficult to effectively integrate different modal information characteristics [17]. As for CNN-only network, the performance of CNN is affected by the window size, where a small window may lead to loss of information over long distances, while a large window may lead to data sparsity problems and difficulties in training. As for RNN-only network, although LSTM as a typical RNN network is a natural choice for understanding haptic time series signals, it has been shown to be inferior to CNNs for haptic classification [13]. The DL methods do not explicitly translate from one modality to the other, as this is often very challenging.

Recent studies have demonstrated the effectiveness of attention mechanisms for sequential and spatially distributed inputs. Recently proposed Multimodal Transformer architecture (MulT [25]) in the field of Emotion Understanding uses Transformer based models for cross-modal representation of language, audio and visual modalities. Subsequently, some studies have extended attention mechanisms to visual-haptic setting, focusing on the ability to extract fused features of two modalities simultaneously. In [9], the self-attention mechanism is used, while ignoring the modality-to-modality connection. In [8], the integration of the self-attention mechanism and the cross-modal attention mechanism in a single Transformer encoder may result in a limited expressive capability of the model. The recent work Visuo-Tactile Transformers (VTT) [6] is a variant of Vision Transformer (ViT) [11] where the inputs are sliced into many patches. It breaks the internal structure of each modal information and requires a large amount of training data and computational resources, which limits its practical

applications. Although these studies explore the application of attention mechanisms to multimodal fusion, they only consider visual and haptic modalities and neglect other possible modalities. Arguably, little work has been done using the fusion structure of CNN, RNN with Transformer in the field of robotic multimodal object recognition.

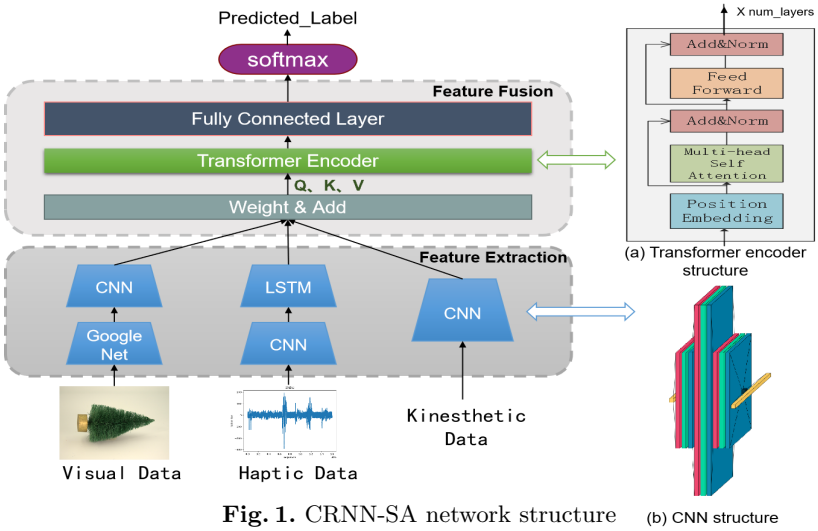

Fig. 1. CRNN-SA network structure

3 Model Architectures

In general, our Multimodal CRNN with Transformer structure includes four modules. (i) Tactile time series module: features are extracted using CNN and Bi-LSTM networks, which preserve the spatial information and temporal relationships of the time series. (ii)Visual image module: due to the small number of images, features are extracted using pre-trained GoogLeNet and CNNs. (iii) Kinesthetic sequence module: features are extracted using CNNs. All the above three modules are the same in the subsequent methods. (iv) Transformer encoder-based fusion module: for this we propose two different structures. The first method (CRNN-SA) fuses the features of all modalities by computing the weights and enhances the fusion by the self-attention mechanism (Figure 1). The second method (CRNN-CA) obtains the cross-attention of each modality with the fused modality and then concatenates them (Figure 2). Finally, the features are classified based on the fused features. These two methods explore the effect of different attention mechanisms and fusion methods on the object recognition. Our approach belongs to feature-level fusion among information fusion methods (the remaining two are data-level and decision-level fusion) and has been proven to be more effective than other types of fusion methods in similar tasks [3].

3.1 Feature Extraction

Kinesthetic CNN Model The kinesthetic data collected in AU dataset [4] consist of readings from the robot's wrist and the infra-red proximity sensor,

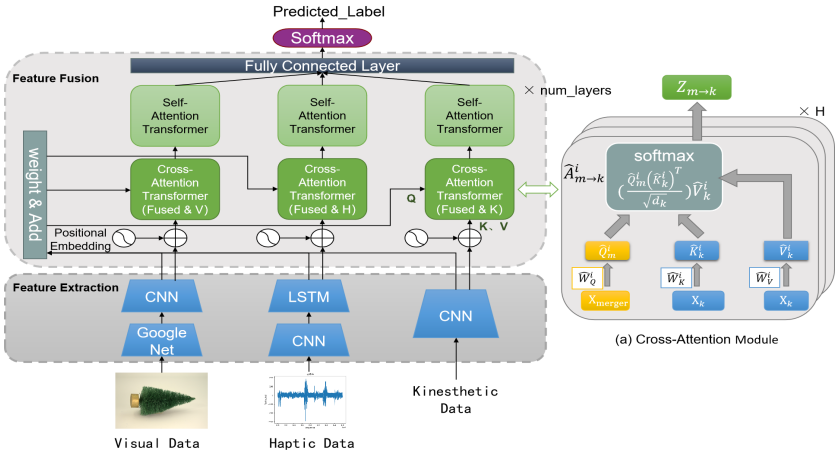

Fig. 2. CRNN-CA network structure

and the positions of each of the hand 's five fingers. It is formally represented as sequence data of short length and does not include temporal information. Therefore, a three-layer one-dimensional convolutional neural network is used to extract the features of the information. Each layer includes a batch normalization and a rectified linear unit (ReLU) function as the activation function. Finally, a global average pooling layer is added to replace the fully connected layer to retain the global information. The CNN of the same structure is used for each modality, as shown in Figure 1.

Visual CNN Model Due to the high cost of data collection in real-world tasks, visual data in practical applications usually consists of a limited number of images. As a result, the parameters of a large network may not be sufficiently optimized with limited data. We adopt the idea of migration learning, using the pre-trained Inception-v3 [22] model on the ImageNet [10] dataset as the basic model, and then add the same convolutional neural network and an average pooling layer as mentioned above to enhance feature extraction.

Haptic CRNN Model The haptic data in the AU dataset [4] are time series collected by five microphones set on the robot's hands. Firstly, the spatial features of the haptic data are extracted using the same convolutional neural network. Then the temporal features are extracted with a bi-directional LSTM (Bi-LSTM) [14]. Similarly, a global average pooling layer is added at the end.

Let $\{X_k, X_v, X_h\}$ represent the kinesthetic, visual and haptic raw inputs. We use the function F_g to represent Inception-v3, the function F_c to represent the above three-layer CNN, and the function F_r to represent the Bi-LSTM. For each input, the above process can be expressed as follows:

$$\begin{aligned} D_k &= F_c(X_k) \\ D_v &= F_c(F_g(X_v)) \\ D_h &= F_r(F_c(X_h)) \end{aligned} \tag{1}$$

3.2 Feature Fusion

In order to compare the fusion methods based on the Transformer's self-attention mechanism with the cross-attention mechanism, we designed two different networks. The two network structures differ in the fusion of multimodal features but are identical in the feature extraction part. In structure (i) (Figure1), the features of each modality are weighted and fused to obtain the kinesthetic-visual-tactile fusion vector D_{merger}, where w_k, w_v, w_h are learnable parameters. The internal connections of the fused features are then reinforced by the self-attention module of the Transformer encoder as

$$D_{merger} = [w_k \cdot D_k^T + w_v \cdot D_v^T + w_h \cdot D_h^T]^T \tag{2}$$

In structure (ii) (Figure 2), we draw on the idea of MulT for modal fusion using the encoder module of the Transformer. The potential connections between different modalities are represented using cross-modal attention, followed by a sequence model using fused features for prediction. The difference is that we only use three cross-modal attention modules and three self-attention module (six in total), whereas MulT uses six cross-modal attention modules and three self-attention modules (nine in total). This is because we first fused the modal features initially by the way of Equation (2), and then used the fusion vector D_{merger} as the common modality to perform cross-modal attention with each individual modal feature. This allows the potential representation of each modality for the common modality to be learned and reduces the number of parameters that need to be trained.

Before sending features to the transformer encoder to compute self or cross attention, positional embedding needs to be added to the inputs, otherwise the distance dependencies in the features would be lost. Let the input feature be D, let its max length be L, its dimension be d_k, and the position embedding vector of element $D_{i,j}$ be $PE_{i,j}$, given as

$$PE_{i,j} = \begin{cases} \sin(\frac{i}{10000^{j/d_k}}), & \text{if } j \text{ is even} \\ \cos(\frac{i}{10000^{(j-1)/d_k}}), & \text{if } j \text{ is odd} \end{cases} \tag{3}$$

Then, the final input to the encoder, X, is

$$X = D + PE \tag{4}$$

It is necessary to note that we do not perform word embedding on the input, as here each element of the sequence feature corresponds to its native meaning and has no higher dimensional meaning.

Multi-Head Self-Attention We implement the self-attention mechanism by modifying the multi-head self-attention method in the Transformer architecture [26]. The multi-headed self-attention mechanism is used to extract the internal connections of the features of the fused modalities. Its structure is shown in

Figure 1. The Transformer encoder consists of N identical self-attention layers. Each self-attentive layer has two parts: (i) the Multi-Head Self-Attention and Normalisation, (ii) the Feed Forward network. The number of attention heads is H. Taking the n-th self-attention layer and the i-th attention head as an example, and let the output of the previous layer be Z_{n-1}. For $n = 1$, $Z_{n-1} = X$, where X is defined in Equation (4). In this part, $Q, K, V \in \mathbb{R}^{m \times d_k}$ are the projected queries, keys, and values respectively, and $W_Q, W_K, W_V \in \mathbb{R}^{d_{model} \times d_k}$ are the learned weight matrices for the projection, where d_k is defined in the position embedding and m is the number of samples in the input data.

$$Q_n^i = Z_{n-1} W_Q^i, \quad K_n^i = Z_{n-1} W_K^i, \quad V_n^i = Z_{n-1} W_V^i \tag{5}$$

Then the scaled dot-product attention is computed for each set, where $A_n^i \in \mathbb{R}^{m \times d_k}$ is the output for the set, and $\sqrt{d_k}$ is a scaling factor to stabilize the gradients during training.

$$A_n^i = \text{softmax}\left(\frac{Q_n^i (K_n^i)^T}{\sqrt{d_k}}\right) V_n^i \tag{6}$$

Finally, the attention outputs corresponding to the H sets are concatenated and projected using another learned weight matrix, where $W_n^O \in \mathbb{R}^{H d_k \times d_{model}}$ is the learned weight matrix for the final projection,

$$A_n = \text{concat}(A_n^1, A_n^2, ..., A_n^H) W_n^O \tag{7}$$

The layer also uses residual connectivity and layer normalization. After a feed-forward network, $Z_n \in \mathbb{R}^{m \times d_{model}}$ is the final output of the n-th multi-head self-attention encoder,

$$\begin{gathered} Z'_n = \text{LayerNorm}(Z_{n-1} + A_n) \\ Z_n = \text{FFN}(Z'_n) = \text{LayerNorm}(Z'_n + f_n(Z'_n)) \\ \text{where} \quad f(Z'_n) = W_1 max(0, W_0 Z'_n + b_0) + b_1 \end{gathered} \tag{8}$$

Multi-Modal Cross-Attention We uses multi-headed cross-modal attention to obtain the potential adaptation of a single modality to a multimodal fused signal, with the structure shown in Figure 2. For each modality, only one layer of cross-modal attention is used, which helps reduce over-parameterization of the model. The fused features D_{merger} are shown in Equation (2). For clarity, we follow the deductive process in Multi-Head Self-Attention because the structure within the two is similar. The difference lies in the calculation of the Q, K, V. The multi-modal cross-attention of kinesthetic features $\hat{A}_k^i$ can be described as follows, and other modalities can be generalized in the same way.

$$\begin{gathered} \hat{Q}_m^i = X_{merger} \hat{W}_Q^i, \quad \hat{K}_k^i = X_k \hat{W}_K^i, \quad \hat{V}_k^i = X_k \hat{W}_V^i \\ \hat{A}_{m \to k}^i \sim \hat{Q}_m^i (\hat{K}_k^i)^T \hat{V}_k^i \\ \hat{A}_{m \to k} = \text{concat}(\hat{A}_{m \to k}^1, \hat{A}_{m \to k}^2, ..., \hat{A}_{m \to k}^H) W^O \\ Z'_{m \to k} = \text{LayerNorm}(X_k + \hat{A}_{m \to k}) \\ Z_{m \to k} = \text{FFN}(Z'_{m \to k}) \end{gathered} \tag{9}$$

The cross-modal attention of each modal feature is then fed into a Transformer encoder with the self-attention module, and the resulting three vectors are concatenated and passed through fully-connected layers for object class prediction.

$$Z_{merger} = [(Z_{m\to k}); (Z_{m\to v}); (Z_{m\to h})] \tag{10}$$

4 Experiments and Results

In this section, we compare the performance of the two methods in this paper with popular multimodal fusion techniques on the latest AU dataset used for multimodal object recognition. Next, we perform a set of ablation studies to evaluate the impact of multiple modal combinations on object recognition. Finally, we analyse the reasons for false object prediction results.

4.1 Data Description and Preprocessing

Publicly available datasets in multimodal object recognition field are still rare and of small size. We compared the PHAC-2 dataset (2015) [7], VHAC dataset (2022) [28], Toprak S' dataset (2018) [24] and the AU dataset (2021) [4]. Finally, we chose the AU dataset because it is open source, has the largest number of objects and contains three types of modal data. This dataset presents multimodal data for 63 objects with some visual and haptic ambiguity, which contains visual, kinesthetic and haptic (audio/vibrations) data. The data for each modal are not collected simultaneously and are therefore unaligned.

Visual Data The visual data in AU dataset consist of four RGB images, a background image for each object and three images of different faces of the object. Because the amount of visual data in the AU dataset is small, we use image enhancement techniques to make the training samples richer and more diverse and to ensure there is no duplicated images in the training and test sets. The methods are as follows: (i) Adjust the brightness and contrast of the images to 0.7-1.3 times of their original values. (ii) Flip the image horizontally and vertically. (iii) Rotate the image by 180°. The final visual data of each object is expanded from 3 images to 50 images. During training and testing, each image used as input is resized to 256x256 pixels, normalized, and the object is segmented using background subtraction.

Kinesthetic Data The kinesthetic data includes the current readings of the robot's wrist, the positions of the five fingers and the readings of the IR proximity sensor at the center of the palm for each exploration process ("unsupported holding", "enclosure", and "pressure-squeeze"). We did not perform complex preprocessing of the kinesthetic data, but simply concatenated them of each sample in the same dimension.

Haptic Data The haptic data comprises vibration data collected by the five channels/microphones during each exploration ("feel" and "pressure-poke") with a sampling rate of 400 kHz. Firstly, to compensate for the noise generated by the robot actuators, cooling fans and other moving parts changes during data

collection, we subtracted background noise from the raw haptic data. In addition, since the sampling rate of the haptic data is much higher than that of other modal data, we downsampled it to 2500 Hz to save the time and space cost of computing. Finally, we normalized the haptic data as follows,

$$\hat{S} = \frac{S - \bar{S}}{\sigma} \tag{11}$$

where $\bar{S}$ and σ are the mean and standard deviation of the data in each microphone channel, respectively.

4.2 Experimental Settings

Baseline Structures To evaluate the performance of the two methods in this paper, we use two popular multimodal object recognition methods as baselines: the concatenation method similar to Zhang et al. (MMM) [28] and the method adapted from MulT [25]. For feature extraction, we compare the CNN-only method (CNN-T) and the RNN-only method (RNN-T) with the CRNN method in this paper. In fusion level, we compared data-level fusion (Early), decision-level fusion (Late) with the feature-level fusion of this paper.

Implementation Detail Adam [16] is used as the optimizer of the model, and the learning rate lr is initially 0.0001. When the evaluation metric no longer improves after 10 epochs, the learning rate is reduced to $lr = lr * 0.5$. The size of batch is 8, and the number of training epochs is 200. To quantitatively evaluate our model, we use the classification accuracy and the weighted F1-score as our evaluation metrics. The experiments were deployed on a host computer configured with an NVIDIA GeForce RTX 3090 (24GB) GPU, and the GPU was used for training throughout.

Loss Function We use the categorical crossentropy shown below as the loss function,

$$Loss = \sum_{i}^{\text{output size}} y_i \cdot \log \hat{y}_i \tag{12}$$

where y_i is the desired output and $\hat{y}_i$ is the actual output.

4.3 Experimental Results

Table 1 shows the results of our methods against other popular methods on AU dataset, where the red numbers are the best results and the blue numbers are the second best results.

The results show that our method CRNN-SA has the best performance, followed by our method CRNN-CA. Both have less number of parameters than the MulT method. Based on the analysis of the results, we can draw the observations that (i) The complex modal fusion method does not improve the accuracy of classification. (ii) The higher the degree of retaining the original features of each modality, the more improvements it will lead to in terms of classification results. (iii) Enhancing the attention within the fused features can help improve classification performance.

Table 1. Comparison of the multimodal fusion methods on AU dataset.

Models	Feature Extraction	Fusion	Accuracy	F1_score	Parameters
MMM[28]	CNN+RNN	concatenation	0.8571	0.8317	1,060,482
MulT[25]	CNN+RNN	Transformer	0.8286	0.8026	1,110,842
CNN-T	CNN	Transformer	0.8632	0.8317	848,082
RNN-T	RNN	Transformer	0.8234	0.8101	715,474
Early	CNN	concatenation	0.7222	0.6866	273,855
Late	CNN+RNN	decision fusion	0.8535	0.8212	1,096,703
CRNN-SA(ours)	CNN+RNN	Transformer	0.9127	0.9061	1,070,128
CRNN-CA(ours)	CNN+RNN	Transformer	0.8746	0.8505	1,105,190

Ablation Study To further investigate the impact of individual modal data in a multimodal object recognition task, we conducted ablation studies on AU dataset and the results are shown in Table 2. Firstly, we compare the classification accuracy on our method (CRNN-SA) using only unimodal data (visual, haptic or kinesthetic) as input. Then, we compared the classification accuracy on our method (CRNN-SA) using a combination of two modal data as input.

Table 2. Comparison of the visual-haptic-kinesthetic inputs on AU dataset.

Inputs data	Accuracy	F1_score
visual	0.6195	0.6013
haptic	0.7635	0.7495
kinesthetic	0.5970	0.5903
visual+haptic	0.8889	0.8630
visual+kinesthetic	0.6540	0.6255
haptic+kinesthetic	0.8071	0.7781
visual+haptic+kinesthetic	0.9127	0.9061

Table 2 shows the comparison results on AU dataset using our method and different combinations of inputs, where the red numbers are the best results. Because this dataset contains objects that are visually or haptically ambiguous, the classification results for unimodal data are much lower than multimodal, which is consistent with our experience in life. And the results demonstrate the importance of the haptic data.

5 Conclusion

In this paper, we have presented two multimodal object recognition methods (CRNN-SA, CRNN-CA), where we have made improvements for existing methods in both the feature extraction and feature fusion steps. After extracting each modal feature using the CRNN method, we fuse the feaures with Transformer's self-attention mechanism and fully connected layers in CRNN-SA, and with the multi-modal cross-attention mechanism adapted from MulT in CRNN-CA. Both methods are applicable to unaligned multimodal data. Among them, the CRNN-SA method outperforms the most popular CNN-only with concatenation method in terms of classification accuracy, and the CRNN-CA method proposes a new cross-modal attention mechanism and uses fewer encoder modules than the MulT method. Future work aims to create a multimodal object

recognition dataset, and explore the integration of deep learning with reinforcement learning, with a view to deploying the results in real-world applications.

References

1. Allen, P.K.: Surface descriptions from vision and touch. In: IEEE International Conference on Robotics Automation. pp. 394–397 (1984)
2. Allen, P.K.: Integrating Vision and Touch for Object Recognition Tasks, p. 407–440. Ablex Publishing Corp., USA (1995)
3. Bednarek, M., Kicki, P., Walas, K.: On robustness of multi-modal fusion—robotics perspective. Electronics **9**, 1152 (07 2020)
4. Bonner, L.E.R., Buhl, D.D., Kristensen, K., Navarro-Guerrero, N.: Au dataset for visuo-haptic object recognition for robots (2021)
5. Cao, G., Zhou, Y., Bollegala, D., Luo, S.: Spatio-temporal attention model for tactile texture recognition. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 9896–9902 (2020)
6. Chen, Y., Sipos, A., Van der Merwe, M., Fazeli, N.: Visuo-tactile transformers for manipulation. In: 2022 Conference on Robot Learning (CoRL). Proceedings of Machine Learning Research, vol. 205, pp. 2026–2040 (2022)
7. Chu, V., Mcmahon, I., Riano, L., Mcdonald, C.G., He, Q., Perez-Tejada, J.M., Arrigo, M., Darrell, T., Kuchenbecker, K.J.: Robotic learning of haptic adjectives through physical interaction. Robotics and Autonomous Systems **63**, 279–292 (2015)
8. Cui, S., Wei, J., Li, X., Wang, R., Wang, S.: Generalized visual-tactile transformer network for slip detection. IFAC-PapersOnLine **53**(2), 9529–9534 (2020)
9. Cui, S., Wang, R., Wei, J., Hu, J., Wang, S.: Self-attention based visual-tactile fusion learning for predicting grasp outcomes. IEEE Robotics and Automation Letters **5**(4), 5827–5834 (2020)
10. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. pp. 248–255 (2009)
11. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale (2020)
12. Fanello, S.R., Ciliberto, C., Noceti, N., Metta, G., Odone, F.: Visual recognition for humanoid robots. Robotics and Autonomous Systems **91**, 151–168 (2017)
13. Gao, Y., Hendricks, L.A., Kuchenbecker, K.J., Darrell, T.: Deep learning for tactile understanding from visual and haptic data (2015)
14. Graves, A., Schmidhuber, J.: Framewise phoneme classification with bidirectional lstm and other neural network architectures. Neural networks : the official journal of the International Neural Network Society **18**, 602–10 (07 2005)
15. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Computation **9**(8), 1735–1780 (1997)
16. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (12 2014)
17. Le, M., Rathour, V., Truong, Q., Mai, Q., Brijesh, P., Le, N.: Multi-module recurrent convolutional neural network with transformer encoder for ecg arrhythmia classification. pp. 1–5 (2021)

18. Liu, H., Yu, Y., Sun, F., Gu, J.: Visual–tactile fusion for object recognition. IEEE Transactions on Automation Science and Engineering **14**(2), 996–1008 (2017)
19. Luo, S., Yuan, W., Adelson, E., Cohn, A.G., Fuentes, R.: 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 (2018)
20. Strese, M., Brudermueller, L., Kirsch, J., Steinbach, E.: Haptic material analysis and classification inspired by human exploratory procedures. IEEE Transactions on Haptics **13**(2), 404–424 (2020)
21. Sun, F., Liu, C., Huang, W., Zhang, J.: Object classification and grasp planning using visual and tactile sensing. IEEE Transactions on Systems, Man, and Cybernetics: Systems **46**(7), 969–979 (2016)
22. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2818–2826. Los Alamitos, CA, USA (jun 2016)
23. Tatiya, G., Sinapov, J.: Deep multi-sensory object category recognition using interactive behavioral exploration. In: 2019 International Conference on Robotics and Automation (ICRA). pp. 7872–7878 (2019)
24. Toprak, S., Navarro-Guerrero, N., Wermter, S.: Evaluating integration strategies for visuo-haptic object recognition. Cognitive Computation **10**, 408–425 (06 2018)
25. Tsai, Y.H., Bai, S., Liang, P., Kolter, J., Morency, L.P., Salakhutdinov, R.: Multimodal transformer for unaligned multimodal language sequences. vol. 2019, pp. 6558–6569 (07 2019)
26. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6000–6010. NIPS'17, Curran Associates Inc., Red Hook, NY, USA (2017)
27. Yang, J., Liu, H., Sun, F., Gao, M.: Object recognition using tactile and image information. In: 2015 IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 1746–1751 (2015)
28. Zhang, P., Zhou, M., Shan, D., Chen, Z., Wang, X.: Object description using visual and tactile data. IEEE Access **10**, 54525–54536 (2022)
29. Zhao, Z.Q., Zheng, P., Xu, S.T., Wu, X.: Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems **30**(11), 3212–3232 (2019)